Smooth hazards with multiple time scales
An application to transitions from cohabitation to marriage

Extended Abstract

Angela Carollo
carollo@demogr.mpg.de
Max Planck Institute for Demographic Research
Rostock, Germany

Jutta Gampe
gampe@demogr.mpg.de
Max Planck Institute for Demographic Research
Rostock, Germany

Hein Putter
h.putter@lumc.nl
LUMC, Leiden, The Netherlands

Paul H.C. Eilers
p.eilers@erasmusmc.nl
Erasmus MC, Rotterdam, The Netherlands

November 1, 2019

Abstract

Marriage is frequently preceded by a period of cohabitation. The transition rate from cohabitation to marriage can be considered over two time scales, the age of the individual (age-specific rates) or the duration of the cohabitation. Traditional approaches choose one time scale as the dominant one and model the other time scale as a, possibly time-varying, covariate. We propose an approach to model a hazard jointly over two time dimensions. The model assumes a smooth bivariate hazard function, and the function is estimated by two-dimensional P-splines. We use data from the German Family Panel (pairfam), and we demonstrate that considering the two time scales jointly provides additional insights about the transition from cohabitation to marriage.

KEYWORDS: Time scales; multidimensional hazard; P-splines; cohabitation; marriage

1 Introduction

Premarital cohabitation has become a common form of living arrangement and many individuals view cohabitation as an alternative to marriage, however, the latter is still an important institution in people's lives. Consequently, many couples will eventually marry after a period of cohabitation.

When modelling transitions from one state to another in the life course, hazard models are the most common approach. Age-specific transition rates are the most prominent choice, however, other time dimensions, such as the duration since entry in the current state, are also of interest. This also applies to the transition from cohabitation to marriage. Age certainly has a major role in the transition from cohabitation to marriage, however, previous research has also pointed out the important role of duration of the cohabitation in triggering the transition to marriage (Di Giulio et al., 2019; Hiekel et al., 2015). It is likely that the two time dimensions will interact with each other and that the hazard of marrying is determined by both time scales.

In this paper we will examine the hazard of moving from cohabitation to marriage along two time scales, age and duration of cohabiting, jointly. Rather than choosing one dominant time scale and incorporating the other time dimension as a (time-varying) covariate we consider the hazard as a smooth bivariate function over the two time scales. This allows a flexible interplay between the two time dimensions.

To demonstrate the approach we will analyse data from the German Family Panel (pairfam). Up to 85% of marriages, among West German couples, started with a cohabitation (Hiekel and Fulda, 2018). Married couples still are privileged over non-married couples, by the provision of social benefits, for example through more favourable taxation of married couples (Baizán et al., 2004). However, attitudes towards marriage differ significantly between West and East Germany, with a higher proportion of people cohabiting without marrying in East Germany (Hiekel et al., 2015) We therefore expect that the hazard of making the transition from cohabitation to marriage will assume different shapes in the West German and East German subpopulation.

The rest of this paper is organized as follows. First, we describe the model for the smooth two-dimensional hazard function and how it can be estimated by bivariate P-splines. A description of the data and some descriptive results for the study sample follows. We will then show the main results of our two-way model. Finally, we will discuss our analysis and the main findings and compare them with previous research on the same topic. Conclusions and outlook will follow.

2 Smooth bivariate hazard model

2.1 Univariate hazard smoothing with P-splines

To introduce notation and the P-spline approach of estimating smooth hazards we first exemplify the method in a one-dimensional setting. If X denotes the (continuous) random variable describing the time to the event of interest, then the hazard of X is

$$\lambda(x) = \lim_{\Delta s \downarrow 0} \frac{1}{\Delta s} P(x < X \leq x + \Delta s \,|\, X > x)\,.$$

If the hazard $\lambda(x)$ is piecewise-constant, then the estimation reduces to common occurrence-exposure rates. The time axis is divided into consecutive intervals $I_k = (\tau_{k-1}, \tau_k]$ with hazard

levels λ_k, and maximum likelihood estimation reduces to estimating constant hazards over each of the intervals I_k

$$\hat{\lambda}_k = \frac{\text{No. of events observed during } I_k}{\text{Total exposure time observed during } I_k} =: \frac{n_k}{r_k}\,.$$

This is equivalent to Poisson regression for $N_k \sim \text{Poi}(\mu_k)$ and expected values $\mu_k = \text{E}(N_k) = r_k\,\lambda_k$ (Holford, 1980). The finer the intervals I_k the more flexibly the hazard $\lambda(x)$ can be modelled, however, this flexibility comes at the price of increased variability and erratic behaviour in regions where few individuals are observed.

A standard solution to this problem is to require that the hazard $\lambda(x)$ is a smooth function. One expresses (the log of) this smooth function as a linear combination of suitable basis functions whose coefficients are restricted by a roughness penalty. This is the idea underlying P-splines smoothing (Eilers and Marx, 1996).

The time axis is split into a large number K of (rather) short bins, commonly of equal length (with midpoints c_k). The basis functions are a set of M equally spaced B-splines of degree p, and the log-hazard $\ln\lambda(x)$ is expressed as a linear combination of these B-splines

$$\ln\lambda(x) = \sum_{m=1}^{M} B_m(x)\,\alpha_m\,.$$

The coefficients α_m in the linear combination are restrained by a difference penalty (of order d) that guarantees that neighbouring coefficients will not differ strongly and hence smoothness of the resulting estimated log-hazard is implied (Eilers, 1998).

For the Poisson regression approach this leads to the following specification: Let $n = (n_1, \ldots, n_K)^{\text{T}}$ and $r = (r_1, \ldots, r_K)^{\text{T}}$ be the vector of observed number of events and total time at risk in the K bins, respectively, so that $\mu_k = r_k\,\lambda_k$. We denote $\eta_k = \ln\lambda_k$ so that $\mu_k = r_k\,e^{\eta_k}$.

For a chosen B-splines basis (degree p and number of basis functions M) the $K \times M$ matrix of regressors B is given by the elements $b_{km} = B_m(c_k)$, which is the m^{th} B-spline evaluated in the midpoint c_k of the k^{th} time interval I_k. Therewith we can express

$$\eta_k = \ln\lambda_k = \ln\lambda(c_k) = \sum_{m=1}^{M} B_m(c_k)\,\alpha_m = \sum_{m=1}^{M} b_{km}\alpha_m.$$

The m coefficients α_m have to be estimated to obtain the (log-)hazard. The Poisson log-likelihood for the K counts is

$$\ell(\alpha) = \sum_{k=1}^{K} n_k\eta_k - \sum_{k=1}^{K} r_k\,\exp\{\eta_k\}, \tag{1}$$

leading to the score equations

$$B^{\text{T}}\,(n - r * e^{\eta}) = 0\,,$$

which are solved by iteratively weighted least-squares (McCullagh and Nelder, 1989). The system

$$B^{\text{T}}\tilde{M}B\alpha = B^{\text{T}}(n - \tilde{\mu} + \tilde{M}B\tilde{\alpha}) \tag{2}$$

is solved repeatedly for α until convergence. Here $M = \text{diag}(\mu)$ and the tilde indicates the current value in the iteration.

To incorporate the smoothness assumption the coefficients α are penalized by a difference penalty $\alpha^{\mathrm{T}} D_d^{\mathrm{T}} D_d \alpha = \alpha^{\mathrm{T}} P \alpha$, where D_d is a matrix that builds differences of order d of the coefficients. The penalty becomes large if neighbouring coefficients differ strongly. The Poisson log-likelihood (1) is supplemented by this penalty term, with a smoothing parameter ρ added to tune the strength of the penalty, leading to the penalized log-likelihood

$$\ell_P(\alpha;\rho) = \left(\sum_{k=1}^{K} n_k \eta_k - \sum_{k=1}^{K} r_k \, \exp\{\eta_k\} \right) - \rho \, \alpha^{\mathrm{T}} P \alpha \, . \tag{3}$$

The system (2) changes to

$$(B^{\mathrm{T}} \tilde{M} B + \rho P) \, \alpha = B^{\mathrm{T}} (n - \tilde{\mu} + \tilde{M} B \tilde{\alpha}) \, . \tag{4}$$

The optimal value of the smoothing parameter ρ is chosen by optimizing a criterion that balances fidelity to the data and model complexity, such as AIC or BIC. The model is fitted for a sequence of ρ-values, equally spaced over $\log_{10} \rho$, and the model with the minimal AIC (or BIC) is selected as optimal.

2.2 Hazards with two time scales

Now we consider two time scales, which we call x_1 and x_2, simultaneously. For simplicity of presentation we assume that time is measured in the same unit for both axes so that an increment of Δs in x_1 corresponds to the same increment in x_2. In this way individuals move in a Lexis diagram along diagonal lines with slope 1 (Keiding, 1990).

In our application x_1 corresponds to the age of the individual (centered so that $x_1 = 0$ corresponds to age 15) and x_2 is the length of cohabitation. When individuals start cohabiting they are in a point $(x_1, x_2) = (x_1^{(0)}, 0)$, where $x_1^{(0)}$ is the age-at-entry into cohabitation, and from that point onward they will move along a diagonal line which represents their 'cohabitation history'. The line will terminate at $(x_1, x_2) = (x_1^{(0)} + s, s)$ either with a marriage (event) or when the observation is censored (drop-out or no marriage before end of observation period).

The hazard of the event of interest for an individual at point (x_1, x_2) is

$$\lambda(x_1, x_2) = \lim_{\Delta s \downarrow 0} \frac{1}{\Delta s} P(x_1 < X_1 \leq x_1 + \Delta s, x_2 < X_2 \leq x_2 + \Delta \,|\, X_1 > x_1, X_2 > x_2) \, , \tag{5}$$

where X_1 and X_2 denote the corresponding random variables of time to event on the two time scales.

Smoothing of bivariate hazards can be achieved by extending the approach described in Section 2.1 to two dimensions as proposed in Currie et al. (2004). The Lexis plane is divided in a tessellation of squares (most common choice, but rectangles would also work) and for each square the number of events n_{jk} and the total exposure time r_{jk} is determined from the individual observations.

If one time axis is age and the other is duration since entry into the current state (in our example: cohabitation), then the possible combinations of x_1 and x_2 are restricted to the positive half-plane where $x_2 < x_1$ (one cannot be cohabiting for longer than being alive or, as we consider here, alive after age 15). To overcome this restricted domain, we transform the points (x_1, x_2)

into new points (y_1, y_2) by

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 - x_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}. \tag{6}$$

The time scale y_1 corresponds to the age-at-entry into cohabitation and y_2 is the duration as before. In this system of time scales individuals move along vertical lines in their 'cohabitation history'. By redefining the time scales in this way, points (y_1, y_2) can be observed, in principle, in the full positive plane.

The (y_1, y_2)-plane is divided into small squares (which correspond to parallelograms in the (x_1, x_2) plane), and the event counts n_{jk} and total exposure times r_{jk} are calculated.

A bi-dimensional B-spline basis is obtained as a tensor product of the univariate B-spline bases constructed over the two time dimensions, that is

$$\mathbf{B} = \mathbf{B}_2 \otimes \mathbf{B}_1,$$

where 2 indicates the duration axis y_2, and 1 indicates the age-at-entry axis y_1. This B-spline basis is then used as regressor matrix in Poisson regression. Correspondingly, there is now a matrix $A = (\alpha_{lm})$ of coefficients.

The penalty matrix consists now of two terms, one for the row coefficients and one for the columns. The penalty matrix is

$$\mathbf{P} = \rho_1(\mathbf{I}_{K_2} \otimes \mathbf{D}_1^{\mathrm{T}}\mathbf{D}_1) + \rho_2(\mathbf{I}_{K_1} \otimes \mathbf{D}_2^{\mathrm{T}}\mathbf{D}_2). \tag{7}$$

Here, $\mathbf{I}_{K_1}$ and $\mathbf{I}_{K_2}$ are identity matrices of the same dimensions as the number of intervals in the age-at-entry direction (K_1) and duration direction (K_2), respectively. The matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ are difference matrices referring to the age-at-entry and duration (we leave out the difference order here for simplicity). Finally, ρ_1 and ρ_2 are the smoothing parameters for the two dimensions, chosen over a bivariate grid of values by minimizing the AIC of the model. The smoothing parameters can be different to allow different amount of smoothing in the row and column direction, if the data suggest different smoothness of the bivariate hazard in the two directions.

The estimated hazard surface over the (y_1, y_2) plane can then be back-transformed and plotted for the original (x_1, x_2) time axes. Details of this approach are presented for density estimation in Carollo and Gampe (2019).

3 Data

3.1 The German Family Panel

We use data from the first ten waves of the German Family Panel (pairfam), release 10.0 (Brüderl et al., 2019). A detailed description of the study can be found in Huinink et al. (2011).

The Panel Analysis of Intimate Relationships and Family Dynamics (PAIRFAM) is a longitudinal panel survey providing rich data on the formation and development of intimate relationships and families. The panel started with about 12,000 randomly selected individuals (anchors) of three different cohorts (1991-1993, 1981-1983, and 1971-1973). The first wave of interviews was conducted is 2008. In 2009, about 1,500 individuals living in East Germany were sampled as

part of the panel DemoDiff, which was initiated following the design of pairfam. Beginning with wave 5, the two studies have been integrated and they are now run in parallel.

For this analysis we used the generated dataset *biopart* which provides both retrospective and prospective information on individuals' relationships histories from age 14, including cohabitations and marriages, on a monthly basis. Details of the generated dataset, as well as of the study design can be found in the Data Manual (Brüderl et al., 2019). All variables selected for this analysis come from the dataset *biopart*, except for the information about residence at the time of cohabitation, which is extracted from the waves' specific questionnaires.

We included in our sample all individuals who have experienced at least one cohabitation with a partner of the opposite sex. Same-sex couples are excluded from the analyses because regulations regarding same-sex marriages have changed during the period of observations. We excluded all cohabitations where one of the partner is younger than 15, as well as marriages below 18 years of age. We considered only first cohabitations and we also excluded cohabitations which started directly with a marriage, or that followed a previous marriage. We excluded couples in which either the main individual or the partner dies. Finally, we included only cohabitations for which it was possible to identify whether the individual lived in East or West Germany at time when the cohabitation started.

3.2 Descriptives of the sample

We analyze a total of 7850 first cohabitations. Of these, 47.82% ended with a marriage. For the total of our sample, the mean age at marriage was 27.5 (sd=4.8) and on average individuals married 3.6 years after they moved in together (sd = 3.2).

Sample	N	% married	age at cohabitation	age at marriage	duration at marriage
Women West	2973	51.26	23.05 (4.3)	26.52 (4.6)	3.19 (2.9)
Men West	2310	47.27	24.84 (4.6)	28.60 (4.6)	3.17 (2.8)
Women East	1429	46.47	22.23 (4.4)	26.85 (4.9)	4.46 (3.7)
Men East	1138	41.65	24.71 (4.6)	29.40 (4.9)	4.41 (3.9)

Table 1: Descriptive statistics of the analysed sample; mean and standard deviation (in parentheses). Age at marriage and duration of cohabitation are calculated based only on events (marriages).

Table 1 provides some basic information about the composition of the four groups. Marriage is more common among West German couples than East German couples, and in both regions more women in the sample marry than men in the sample. Women start cohabiting and marry, on average, earlier than men. Women living in East Germany are on average one year younger than their West Germany counterpart when they move together with a partner, while the difference for men is minor. Finally, couples who eventually marry, do so one year earlier in West Germany than in East Germany.

Figure 1 depicts the cohabitation histories of the individuals in the four groups, by age and duration of the cohabitation. Each line represents a cohabitation, which either ended with a

marriage (red circles), or due to censoring (black dots). Since we are able to identify the date when the cohabitation started, each line starts at duration $x_2 = 0$.

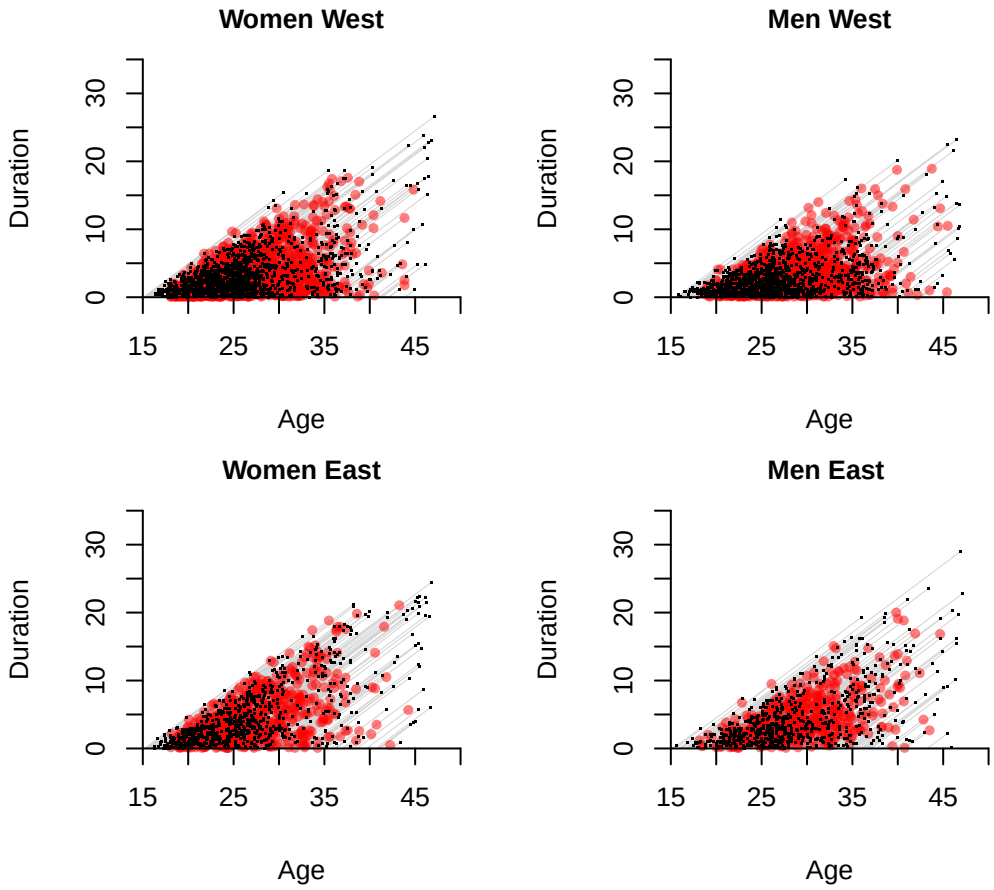


Figure 1: Cohabitation histories for the four sub-samples

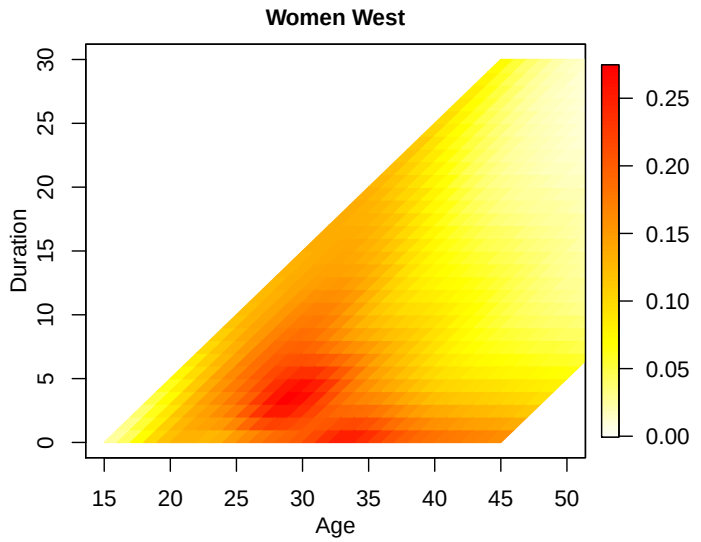
4 Bivariate smoothed hazards

We estimate the smooth hazard for each of the four groups separately, following the procedure described in Section 2.2. Details of the setup for the P-spline regression are presented in Table 2. K indicates the number of intervals in each direction, M is the number of B-spline basis functions, ρ are the smoothing parameters and the subscripts 1 and 2 refer to the age-at-entry and duration dimensions, respectively. We always used cubic B-splines (degree $p = 3$) and a second order difference-penalty. The optimal ρ_1 and ρ_2 are chosen from a linear grid of values for $\log_{10}(\rho)$, ranging from $\log_{10}(\rho) = -2$ to $\log_{10}(\rho) = 2$ for both rows and columns, as the combination which minimizes the AIC of the model.

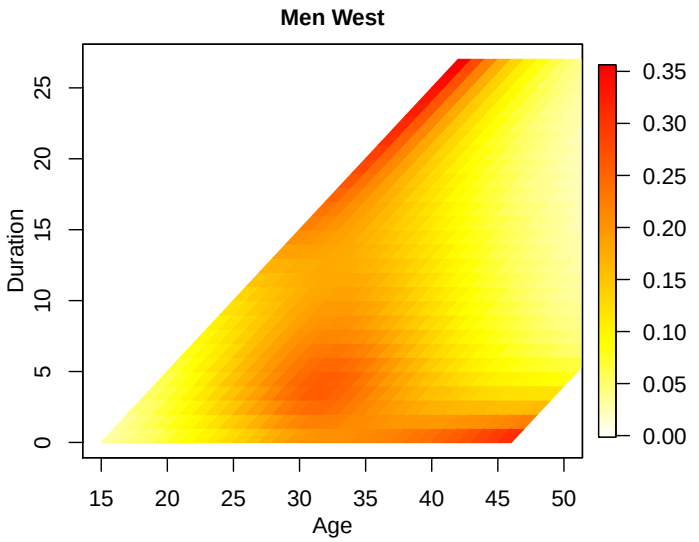
	K_1	K_2	M_1	M_2	ρ_1	ρ_2
Women West	31	31	15	15	0.63	2.51
Men West	32	28	15	14	10	3.98
Women East	30	29	15	14	10	3.98
Men East	33	34	16	17	2.51	3.98

Table 2: Specification of the P-spline hazard model and optimal smoothing parameters, separately for the four groups.

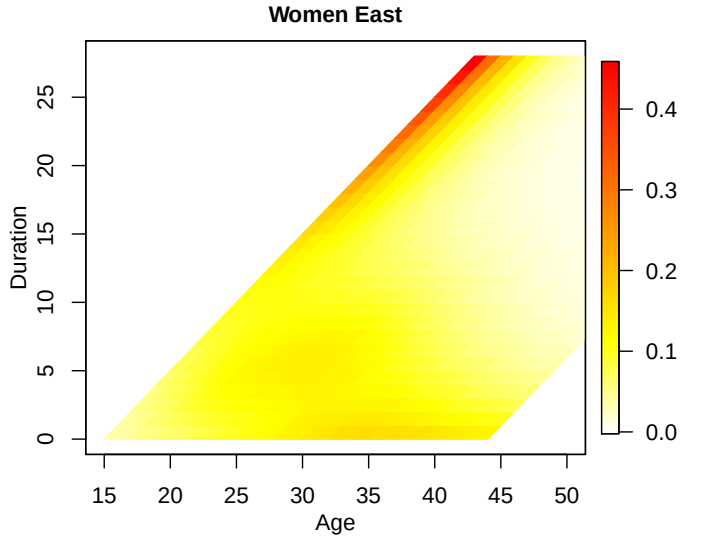
The estimated smooth hazards for the four samples as shown in Figure 2. The hazard of marrying after a period of cohabitation is higher among West German couples than among East German ones. The four hazards show some common features, for example a tendency for individuals who start a cohabitation in their twenties to marry within 10 years. This tendency seems to be stronger for men and women living in West Germany than for their East German counterpart. The hazard for East German men shows a peak at younger ages and longer durations when compared to the one of West German men. Another common feature of the hazards for West German men and East German women is a steep increase in the hazard for individuals who started cohabiting at really young ages but marry after a long cohabitation. Both West German and East German men who start a cohabitation in their forties tend to marry within a couple of years. Finally, West German women who start a cohabitation between 30 and 35 years of age have a higher hazard of marrying in the next couple of years than all the other groups.



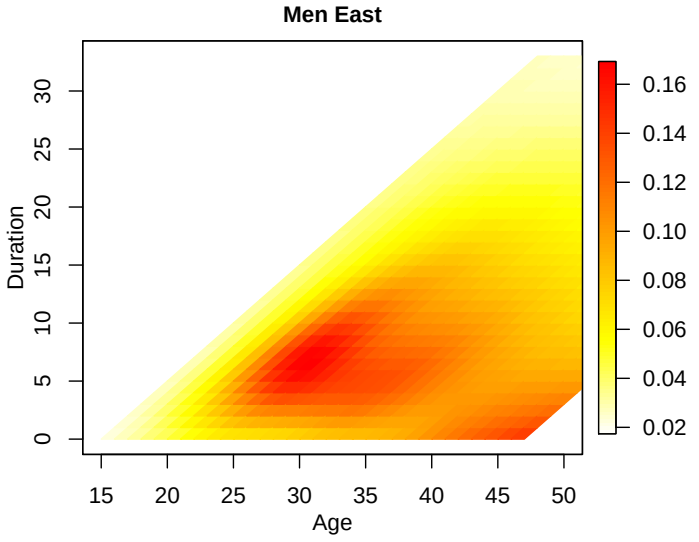
(a) Women living in West Germany at the time of cohabitation



(b) Men living in West Germany at the time of cohabitation



(c) Women living in East Germany at the time of cohabitation



(d) Men living in East Germany at the time of cohabitation

Figure 2: Bidimensional smooth hazard of marrying after a cohabitation by age and duration of the cohabitation

5 Discussion and Outlook

In this paper we use a new approach to analyse transitions from cohabitation to marriage jointly over two times scales, age and duration of the cohabitation. This approach allows a more flexible analysis of the hazard of an event registered over two time scales, without the necessity to prefer one scale over the other. Therefore, we are able to explore the differences and similarities of the estimated hazards as a results of the interplay between the two time dimensions.

The estimated surface can subsequently be used to determine the transition rate for a particular age or age-at-entry or duration by 'cutting' through the surface in an appropriate direction. While we estimated a smooth bivariate surface without further constraints it is of interest to explore whether a simpler form of interplay between the two time scales, such as a log-additive formulation, would fit the data equally well but more parsimoniously.

Currently the proposed method is limited to the estimation a bivariate hazard surface without any additional covariates that can operate on this baseline hazard. We plan to extend the approach by including covariates in a proportional hazards framework.

Acknowledgment: This paper uses data from the German Family Panel pairfam, coordinated by Josef Brüderl, Sonja Drobnič, Karsten Hank, Bernhard Nauck, Franz Neyer, and Sabine Walper. pairfam is funded as long-term project by the German Research Foundation (DFG).

References

Baizán, P., A. Aassve, and F. C. Billari (2004). The interrelations between cohabitation, marriage and first birth in germany and sweden. *Population and Environment 25*(6), 531–561.

Brüderl, J., S. Drobni, K. Hank, B. Nauck, F. J. Neyer, S. Walper, P. Alt, C. Bozoyan, P. Buhr, C. Finn, M. Garrett, H. Greischel, N. Gröpler, K. Hajek, M. Herzig, B. Huyer-May, R. Lenke, L. Minkus, B. Mller, T. Peter, C. Schmiedeberg, P. Schtze, N. Schumann, C. Thönnissen, M. Wetzel, and B. Wilhelm (2019). *The German Family Panel (pairfam).* ZA5678 Data file Version 10.0.0. Cologne: GESIS Data Archive.

Brüderl, J., K. Hajek, M. Herzig, R. Lenke, B. Müller, and P. Schütze (2019). *pairfam Data Manual.* (Release 10.0 ed.). LMU Munich.

Carollo, A. and J. Gampe (2019). Bivariate density estimation with restricted domains: working with oblique coordinates. In *Proceedings of the 34th International Workshop on Statistical Modelling (IWSM), Guimaraes, Portugal.*

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

Di Giulio, P., R. Impicciatore, and M. Sironi (2019, MAY 14). The changing pattern of cohabitation: A sequence analysis approach. *Demographic Research 40*, 1211–1248.

Efron, B. (2002). The two-way proportional hazards model. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 899–909.

Eilers, P. H. C. (1998). Hazard smoothing with B-splines. In B. D. Marx and H. Friedl (Eds.), *Proceedings of the 13th International Workshop on Statistical Modeling*, New Orleans, pp. 200–207.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89–102.

Hiekel, N. and B. E. Fulda (2018). Love. break up. repeat: The prevalence and stability of serial cohabitation among west german women and men born in the early 1970s. *Demographic Research 39*(30), 855–870.

Hiekel, N., A. C. Liefbroer, and A.-R. Poortman (2015). Marriage and separation risks among german cohabiters: Differences between types of cohabiter. *Population Studies 69*(2), 237–251. PMID: 26160505.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics 36*(2), 299–305.

Huinink, J., J. Brüderl, B. Nauck, S. Walper, L. Castiglioni, and M. Feldhaus (2011). Panel analysis of intimate relationships and family dynamics (pairfam): Conceptual framework and design. *Zeitschrift fr Familienforschung - Journal of Family Research 23*, 77–101.

Iacobelli, S. and B. Carstensen (2013). Multiple time scales in multi-state models. *Statistics in Medicine 32*(30), 5315–5327.

Keiding, N. (1990). Statistical inference in the lexis diagram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 332*(1627), 487–509.

McCullagh, P. and J. A. Nelder (1989). *Generalized linear models* (Second ed.). London: Chapman & Hall.