Journal Title
XX(X):1–8
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Supporting Material for: A Change-Point Model for Detecting Heterogeneity in Ordered Survival Responses

Olivier Bouaziz [1], Grégory Nuel [2]

1 The Expectation step in the EM algorithm

In this section we explicit formula (4) of the main paper. The (E-step) of the EM algorithm is defined by

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \mathbb{E}_{R_{1:n}|\text{data},\boldsymbol{\theta}_{\text{old}}}\left[\log \mathbb{P}(\text{data}, R_{1:n}|\boldsymbol{\theta})\right],$$

and the (M-step) corresponds of maximizing the previous quantity with respect to θ:

$$\hat{\theta} = \arg\max_{\theta} \mathbb{E}_{R_{1:n}|\text{data},\boldsymbol{\theta}_{\text{old}}}\left[\log \mathbb{P}(\text{data}, R_{1:n}|\theta)\right].$$

We then have:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \int_{R_{1:n}} \mathbb{P}(R_{1:n}|\text{data};\boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(R_{1:n}, \text{data};\boldsymbol{\theta}) dR_{1:n},$$

with $\mathbb{P}(R_{1:n}, \text{data};\boldsymbol{\theta}) = \mathbb{P}(\text{data}|R_{1:n};\boldsymbol{\theta}) \times \text{constant}$, where the constant does not depend on θ. Notice that $\mathbb{P}(\text{data}|R_{1:n};\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathbb{P}(\text{data}_i|R_i;\boldsymbol{\theta})$ since the distribution of data_i

[1] MAP5, Université Paris Descartes, Paris, France
[2] LPMA, CNRS 7599, Université Pierre et Marie Curie, Paris, France

Corresponding author:
Olivier Bouaziz, 45 rue des Saints Pères, 75270 Paris Cedex 06, France
Email: olivier.bouaziz@parisdescartes.fr

depends only on R_i. Therefore,

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) &= \sum_{i=1}^{n} \int_{R_{1:n}} \mathbb{P}(R_{1:n}|\text{data};\boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(\text{data}_i|R_i;\boldsymbol{\theta}) dR_{1:n} \\ &= \sum_{i=1}^{n} \int_{R_i} \left(\int_{R_{1:n}^{-i}} \mathbb{P}(R_{1:n}|\text{data};\boldsymbol{\theta}_{\text{old}}) dR_{1:n}^{-i} \right) \log \mathbb{P}(\text{data}_i|R_i;\boldsymbol{\theta}) dR_i, \end{aligned}$$

where $R_{1:n}^{-i}$ represents the sequence $R_1, \dots R_{i-1}, R_{i+1}, \dots, R_n$. Then, $\int_{R_{1:n}^{-i}} \mathbb{P}(R_{1:n}|\text{data};\boldsymbol{\theta}_{\text{old}}) dR_{1:n}^{-i} = \mathbb{P}(R_i|\text{data};\boldsymbol{\theta}_{\text{old}})$ and

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) &= \sum_{i=1}^{n} \int_{R_i} \mathbb{P}(R_i|\text{data};\boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(\text{data}_i|R_i;\boldsymbol{\theta}) dR_i \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{P}(R_i = k|\text{data};\boldsymbol{\theta}_{\text{old}}) \log \mathbb{P}(\text{data}_i|R_i = k;\boldsymbol{\theta}), \end{aligned}$$

which is equation (4) of the main paper.

2 The exponential and Weibull baseline hazards

In this model, we assume that the baseline hazard in the k^{th} segment index belongs to the Weibull family with shape parameter λ_k and scale parameter p_k. That is, $\lambda_k(t) = p_k(t/\lambda_k)^{p_k-1}/\lambda_k$, $\Lambda_k(t) = (t/\lambda_k)^{p_k}$ and $S_k(t) = \exp(-(t/\lambda_k)^{p_k})$.

Equation (2) of the main paper can then be written in the following way:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i \left(\log(p_k) - p_k \log(\lambda_k) + (p_k - 1)\log(T_i) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \left(\frac{T_i}{\lambda_k}\right)^{p_k} \exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

The exponential family is derived as a special case of the Weibull case by setting $p_k = 1$ for all $k = 1, \dots, K$. In that case, Equation (2) of the main paper reduces to:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i \left(-\log(\lambda_k) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \left(\frac{T_i}{\lambda_k}\right) \exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

Computation of the estimates through Equation (3) of the main paper is done via the **survreg** function in the `survival` R package. The gradient vector and Hessian matrix can directly be derived from the expression of the log-likelihood and the estimates can then be computed using the Newton-Raphson algorithm. A weight option is also available in the **survreg** function which allows to compute estimates that precisely maximize the log-likelihood $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ presented in Equation (3) of the main paper.

The models obtained under these families of baseline hazard functions have the nice property that they both belong to the class of parametric Cox models and of parametric Accelerated Failure Time models[1]. Moreover, the two parameters of the Weibull family make the baseline hazard quite flexible. As a matter of fact, the Weibull model will provide a fairly good fit to any true baseline hazard that is monotone with time. However, these families of model will not properly fit a model with true baseline hazard having a bathtub shape (i.e a $\cup$ shape) or an upside down bathtub shape (i.e. a $\cap$ shape) which are common types of baseline that can occur in practice.

The model introduced in the next section does not assume any specific shape for the baseline hazard and consequently will be able to fit any class of baseline hazard functions. However, this model requires to specify in advance a number of cutpoints and makes the approximation that the hazard is constant between each cutpoint.

3 The piecewise constant baseline hazard

In this model, the baseline hazard on each segment index is assumed to be piecewise constant on L cuts represented by $c_0, c_1, \dots, c_L$, with the convention that $c_0 = 0$ and $c_L = +\infty$. Let $I_l(t) = I(c_{l-1} < t \leq c_l)$. We suppose that

$$\lambda_k(t) = \sum_{l=1}^{L} I_l(t)\alpha_l^k,$$

$$\Lambda_k(t) = \alpha_1^k t I_1(t) + \sum_{l=2}^{L}(\alpha_1^k c_1 + \cdots + \alpha_{l-1}^k(c_{l-1} - c_{l-2}) + \alpha_l^k(t - c_{l-1}))I_l(t),$$

$$S_k(t) = \exp(\alpha_1^k t)I_1(t) + \sum_{l=2}^{L}\exp(\alpha_1^k c_1 + \cdots + \alpha_{l-1}^k(c_{l-1} - c_{l-2}) + \alpha_l^k(t - c_{l-1}))I_l(t).$$

Equation (2) of the main paper can then be written in the following form:

$$\log\left(e_i(k;\boldsymbol{\theta})\right) = \Delta_i\left(\log(\lambda_k(T_i)) + \boldsymbol{X}_i\boldsymbol{\beta}_k\right) - \int_0^{\tau} Y_i(t)\lambda_k(t)dt\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k).$$

For computational purpose, it is interesting to note that the log-likelihood can be written in a Poisson regression form. Introduce $R_{i,l} = \int_0^{\tau} Y_i(t)I_l(t)dt = I(T_i \geq c_{l-1})(c_l \wedge T_i - c_{l-1})$, the total time individual i is at risk in the lth interval and $O_{i,l} = \int_0^{\tau} I_l(t)dN_i(t) = I_l(T_i)\Delta_i$, the number of events for individual i in the lth subinterval. Then, we have $\Delta_i \log(\lambda_k) = \sum_l O_{i,l}\log(\alpha_l^k)$, $\int_0^{+\infty} Y_i(t)\lambda_k(t)dt = \sum_l \alpha_l^k R_{i,l}$ and the log-likelihood can be written again as:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{L} w_i(k;\boldsymbol{\theta}_{\text{old}}) \left\{ O_{i,l}(\log(\alpha_l^k) + \boldsymbol{X}_i\boldsymbol{\beta}_k) - \alpha_l^k R_{i,l}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k) \right\}.$$

This log-likelihood is proportional to the log-likelihood one would obtain in a Poisson regression, where the $O_{i,l}$ are the response variables and are assumed to follow, conditionally on the $\boldsymbol{X}_i$, a Poisson distribution with parameter equal to $\alpha_l^k R_{i,l}\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)$. Therefore, the estimates can easily be computed using the **glm** function in the R software and specifying $\log(R_{i,l})$ as "offsets" in the model. See for instance Aalen et al.[2] p.223-225 for more details on the connection between piecewise-constant hazard model and Poisson regression. A weight option is also available in the **glm** function. Finally, note that the exponential case could also be derived as a special case of the piecewise constant hazard family with $L = 1$.

As mentioned earlier, the piecewise constant hazard model is very useful when one does not know the shape of the baseline hazard a priori. However one must specify in advance the value of L in the model. Usually choosing an adequate number of cutpoints allows to provide a good balance between bias and variance estimation. However in our context, detection of the breakpoints is not very sensitive to the choice of L. This is discussed in more details in Section 5.3 of the main paper.

4 The nonparametric baseline hazard

In the absence of weights, this model has been widely used because of its great flexibility, the baseline hazard being estimated without making any assumption on its shape, and because it can easily be implemented in a straightforward manner. First, the regression parameter is estimated by maximizing the Cox partial likelihood which contains terms involving only the regression parameter (and not the baseline hazard). Secondly, the baseline hazard estimator is deduced by the martingale decomposition of the observed counting process. From Equation (1) of the main paper applied to the observed counting and at-risk processes, one gets the following decomposition: for $k = 1, \ldots, K$, $i = 1, \ldots, n$,

$$N_{ik}(t) - \int_0^t Y_{ik}(s)\exp(\boldsymbol{X}_i\boldsymbol{\beta}_k)d\Lambda_k(s) = M_{ik}(t),$$

where $N_{ik}(t) = N_i(t)I(R_i = k)$, $Y_{ik}(t) = Y_i(t)I(R_i = k)$ and $M_{ik}(t)$ is a martingale with respect to the filtration $\sigma(N_{ik}(s), Y_{ik}(s), \boldsymbol{X}_i : 0 \leq s \leq t)$. Taking the expectation conditionally on $\{N_{1:n}(t), Y_{1:n}(t), \boldsymbol{X}_{1:n} : 0 \leq t \leq \tau; \boldsymbol{\theta}_{\text{old}}\}$, summing over the n individuals and taking the differential of both sides of the equation shows that the expression

$$\sum_{i=1}^{n} \{dN_i(t) w_i(k; \boldsymbol{\theta}_{\text{old}}) - Y_i(t) \exp(\boldsymbol{X}_i \boldsymbol{\beta}_k) w_i(k; \boldsymbol{\theta}_{\text{old}}) d\Lambda_k(t)\} \tag{1}$$

is centered. A weighted Nelson-Aalen estimator is derived from this relation:

$$\tilde{\Lambda}_k(t, \boldsymbol{\beta}_k) = \sum_{i=1}^{n} \int_0^t \frac{w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(s)}{\sum_j Y_j(s) \exp(\boldsymbol{X}_j \boldsymbol{\beta}_k) w_j(k; \boldsymbol{\theta}_{\text{old}})}.$$

More details on the standard estimation procedure in the Cox model can be found for instance in Andersen et al.[3]. Now, plugging-in this quantity into $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{\text{old}})$ gives the following weighted Cox partial likelihood:

$$Q^{\text{PL}}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K | \boldsymbol{\theta}_{\text{old}})$$
$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^\tau \left\{ \boldsymbol{X}_i \boldsymbol{\beta}_k + \log(w_i(k; \boldsymbol{\theta}_{\text{old}})) - \log\left(\sum_{j=1}^{n} Y_j(t) \exp(\boldsymbol{X}_j \boldsymbol{\beta}_k) w_j(k; \boldsymbol{\theta}_{\text{old}}) \right) \right\} w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(t).$$

Introduce for $k = 1, \ldots, K$, $l = 0, 1, 2$, $S_k^{(l)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) = \sum_j Y_j(t) \boldsymbol{X}_j^{\otimes l} \exp(\boldsymbol{X}_j \boldsymbol{\beta}) w_j(k; \boldsymbol{\theta}_{\text{old}})$ and $E_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) = S_k^{(1)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) / S_k^{(0)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}})$. Then, on each stratum k, define the score function

$$U_k(\boldsymbol{\beta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n} \int_0^\tau \{\boldsymbol{X}_i - E_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}})\} w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(t),$$

such that $\widehat{\boldsymbol{\beta}}_k$ verifies the equality $U_k(\widehat{\boldsymbol{\beta}}_k|\boldsymbol{\theta}_{\text{old}}) = 0$.

Introduce $V_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) = S_k^{(2)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) / S_k^{(0)}(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) - E_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}})^{\otimes 2}$ and let

$$I_k(\boldsymbol{\beta}|\boldsymbol{\theta}_{\text{old}}) = \sum_{i=1}^{n} \int_0^\tau V_k(t, \boldsymbol{\beta}; \boldsymbol{\theta}_{\text{old}}) w_i(k; \boldsymbol{\theta}_{\text{old}}) dN_i(t),$$

represents minus the derivative of the score function with respect to $\boldsymbol{\beta}$. Then, computation of the estimator $\widehat{\boldsymbol{\theta}}$ can be performed using the iterative Newton-Raphson algorithm. The m^{th} iteration step writes as follows:

$$\widehat{\boldsymbol{\beta}}_k^{(m)} = \widehat{\boldsymbol{\beta}}_k^{(m-1)} + I_k(\widehat{\boldsymbol{\beta}}_k^{(m-1)}|\boldsymbol{\theta}_{\text{old}})^{-1} U_k(\widehat{\boldsymbol{\beta}}_k^{(m-1)}|\boldsymbol{\theta}_{\text{old}}).$$

At convergence, we get the estimator $\widetilde{\boldsymbol{\theta}} = (\tilde{\Lambda}_1, \ldots, \tilde{\Lambda}_K, \widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K)$ where $\tilde{\Lambda}_k(t) = \tilde{\Lambda}_k(t, \widehat{\boldsymbol{\beta}}_k)$ are plug-in Nelson-Aalen estimators of the cumulative hazard functions. Note that the $\widetilde{\boldsymbol{\theta}}$ estimator can be computed with the **coxph** function in the R `survival` library. The weights option can be directly specified in this function.

Finally, as for the parametric models, computation of the new weights is done through the EM algorithm (see Section (3) of the main paper). Then, a simple idea could be to use plug-in estimators again, i.e. to replace $\boldsymbol{\theta}$ by $\widetilde{\boldsymbol{\theta}}$ in the expression of the $e_i(k; \boldsymbol{\theta})$. However, although this is a relevant strategy for the parametric models it will not lead to a consistent estimator for the Cox model. Because of the shape of the Nelson-Aalen estimators, which are stepwise functions, the information in the estimated partial likelihood (or equivalently in $e_i(k; \widetilde{\boldsymbol{\theta}})$), at a given time point is limited. To stabilize the solution, smoothing is needed. In Section 5.2 of the main paper, new kernel type estimators of the Λ_ks and λ_ks are derived and are used as plug-in estimates in order to compute the weights.

5 Calibration of the censoring distribution in the simulations

We present here the parameter of the censoring distribution used in Section (6) of the main paper. In Scenario 1, the censoring was distributed as a uniform distribution with parameters 0 and 2.4, such that 24%, 65% and 60% of individuals were respectively censored in segments 1, 2 and 3. In Scenario 2, the censoring was distributed as a uniform distribution with parameters 0 and 1.8, such that 33%, 47% and 67% of individuals were respectively censored in segments 1, 2 and 3. In Scenario 3, the censoring was distributed as a uniform distribution with parameters 0 and 1.5, such that 38%, 54% and 58% of individuals were respectively censored in segments 1, 2 and 3. In Scenario 4, the censoring was distributed as a uniform distribution with parameters 0 and 0.9, such that 23%, 58% and 67% of individuals were respectively censored in segments 1, 2 and 3.

6 Additional comments for smooth change of the hazard rate

In the model from Section 7.2 from the main paper, we consider here the simulation of two different samples: the first one is simulated under the assumption that $\mathrm{RH} = 10$ and estimated with the one breakpoint model (which gave the smallest BIC value among other breakpoint models) and the other one is simulated under the assumption that $\mathrm{RH} = 50$ and estimated with the two breakpoints model (which again had the smallest BIC value among other models). Figure 1 gives the a posteriori marginal breakpoint distribution in each case along with their respective estimated weighted survival distributions. It should be noted that each of these plots is for a single sample and is not representative of the overall behaviour of the estimation method but is given as a simple illustration as what can be observed in a case of non abrupt changes of the survival distribution.

In both scenarios it was usually observed that the a posteriori breakpoint distribution tends to be widely spread such as illustrated by Figure 1. No pattern could be observed for the localisation of the maximums which seemed to occur at an arbitrarily position in the segment $[a, b]$. For the $RH = 10$ scenario, in the top-left panel of Figure 1, the maximum a posteriori of the one breakpoint model occurred approximately in 1968. For the $RH = 50$ scenario, in the top-right panel of Figure 1, the maximum a posteriori of the two breakpoints model occurred approximately in 1958 and 1966.

Note that values of the maximum of the probabilities are quite low here compared to what is obtained for the diabetic patients dataset (see Figure 3 of the main paper) for instance. This is due to the choice of continuous years of birth in the simulation setting. Simulating discrete years of birth instead will lead to very similar results on the overall (especially with results very close to the one obtained in Table 3 of the main paper), but with a posteriori probability maximums much closer to 1. This is a general behaviour (which is not restricted to the smooth change of hazard scenarios) due to the fact that in the continuous case there are as much years of birth as the number of individuals while there are only a small number of years of birth in the discrete case.

7 Implementation of a smoothing spline estimation method on the diabetes dataset

In order to model calendar year using regression splines, the hazard rate was nonparametrically estimated taking left truncation into account for all pair of diabetes onset/time since diabetes diagnosis in years (there are 40 different years of diabetes onset and 49 different years for the time since diabetes diagnosis variable). Then the resulting estimation was smoothed using the R function **bigtps** of the package `bigsplines`[4]. We present the result in Figure 2 on the log-hazard scale using 100 knots (this is the default value for bidimensional splines) with a smoothing penalty equal to 1 (this small penalty was chosen in order to avoid too much irregularity in the hazard estimation).

Next we applied our method using a piecewise constant hazard model on the diabetes dataset without adjusting with respect to the gender. Four cuts were chosen in the piecewise-constant-hazard model at times 10, 20, 30 and 40. Using the BIC criterion two breakpoints were found which occur at the years 1946 and 1962 with a posteriori probabilities of having a breakpoint at these locations equal respectively to 59% and 93%. The hazard rate estimation from our model is shown on Figure 3.

When we compare Figures 2 and 3 we see that the breakpoints found by our method cannot be seen from the regression spline method. Also, if we read Figure 2 on the "time since diagnosis" axis (that is, from left to right) we see that for a given year of diabetes onset the hazard rate has a bell shape (low values for small times then the hazard increases for medium times and finally has low values again for high times). On the opposite, our method gives an increasing hazard rate for a given cohort year of diabetes onset which is the tendency we observed when subsets of individuals with a common interval of years of diabetes onset are chosen and estimation of the hazard is performed on these subsets. Also the breakpoints can be clearly seen in Figure 3 at years of diabetes onset 1946 and 1962 and the parsimonious representation of the hazard gives an easy interpretation of the risk of death since diabetes diagnosis through the years of diabetes onset. For instance, one could comment that there has been a shift of the hazard rate for the cohort years of diabetes onset $1933 - 1945$ to $1946 - 1961$ which could be a sign of medical improvement over time: the hazard rate of death for $0 - 10$ years after diabetes onset in the cohort $1933 - 1945$ is similar to the hazard rate of death for $10 - 20$ years after diabetes onset in the cohort $1946 - 1961$ and the hazard rate of death for $10 - 20$ years after diabetes onset in the cohort $1933 - 1945$ is similar to the hazard rate of death for $20 - 30$ years after diabetes onset in the cohort $1946 - 1961$.

References

1. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. Wiley Series in Probability and Statistics. Wiley-Interscience (John Wiley & Sons), Hoboken, NJ; 2002.
2. Aalen OO, Borgan Ø, Gjessing HK. Survival and Event History Analysis. Statistics for Biology and Health. Springer Science; 2008.
3. Andersen PK, Borgan Ø, Gill RD, Keiding N. Statistical models based on counting processes. Springer Series in Statistics. New York: Springer-Verlag; 1993.
4. Helwig NE, Ma P. Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples. Journal of Computational and Graphical Statistics. 2015;24(3):715–732.

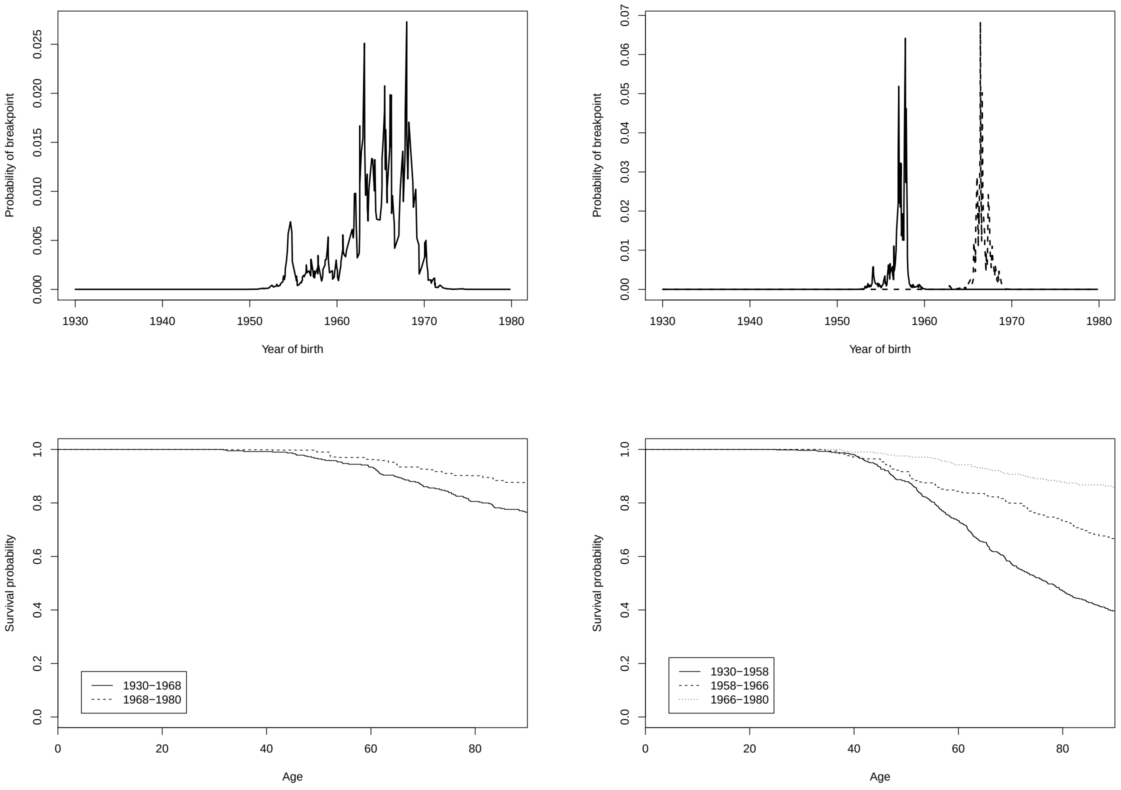

Figure 1. Marginal distribution of the breakpoints in the simulation setting of a smooth change of hazard rate. Left side: model with one breakpoint and $\mathrm{RH} = 10$. Right side: model with two breakpoints and $\mathrm{RH} = 50$.

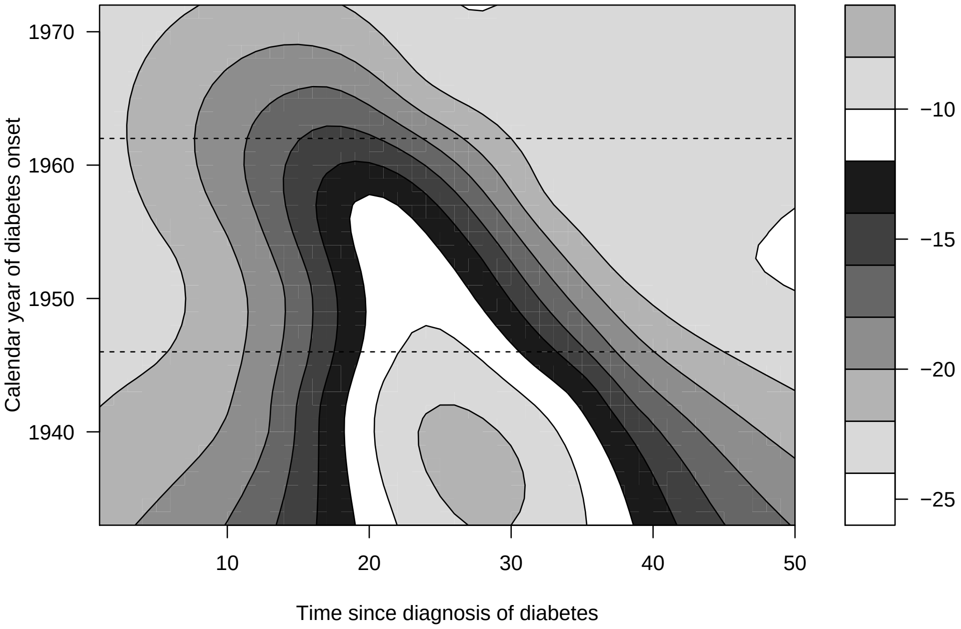

Figure 2. Hazard estimation from the regression spline model for the diabetes data on the log scale.

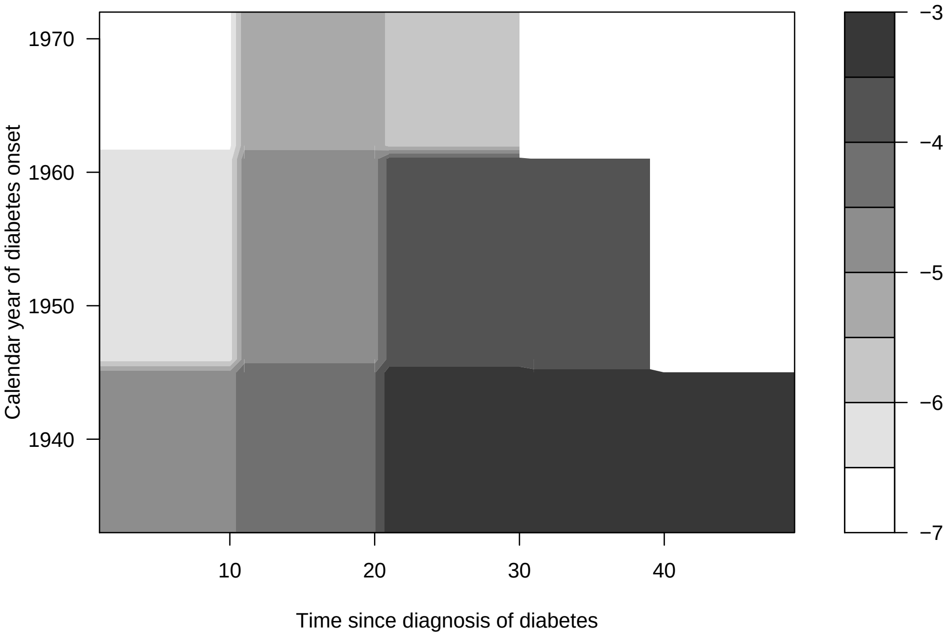

Figure 3. Hazard estimation from the two breakpoints model for the diabetes data on the log scale.