Models based on mixtures of hazard rates with application to statistical reliability analysis

Petr Volf, ÚTIA AV ČR

This contribution presents and explores methods of construction and estimation of models based on mixtures of hazard rates, and tries to demonstrate their usefulness and applicability.

O U T L I N E:

1. Models based on mixtures
2. Mixtures of hazard rates - properties, estimation
3. Artificial example – bath-tube shape of h.r.
4. Real data example

Mixture of distributions (densities, distr. functions) – a standard tool of modeling:

$$f(x) = p_1 \cdot f_1(x) + p_2 \cdot f_2(x). \tag{1}$$

Interpretation:

with probability p_1 an item belongs to first group having density f_1, with p_2 to the second group (compare with cluster analysis).

It is also the way how data representing mixture (1) could be generated.

- - - - -

Probability densities are used also as components for construction of regression models, in the same way as regression splines.

For instance a linear combination (now it need not be convex) of gaussians (in this context called also "radial basis functions") is used to create a curve or surface.

Example 1 – distribution of unemployment duration:

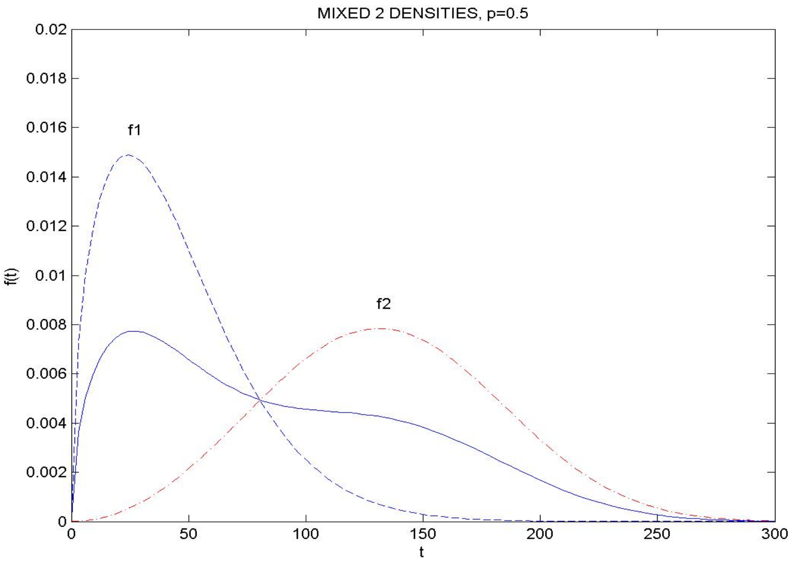

Figure 1: Example of mixture of 2 densities representing "movers", f_1, and "stayers", f_2, resulting $f = (f_1 + f_2)/2$ (thick curve).

f_1 is of Weibull distribution with $\alpha_1 = 50, \beta_1 = 1.5$,

$f_2 \sim$ Weibull with $\alpha_2 = 150, \beta_2 = 3$

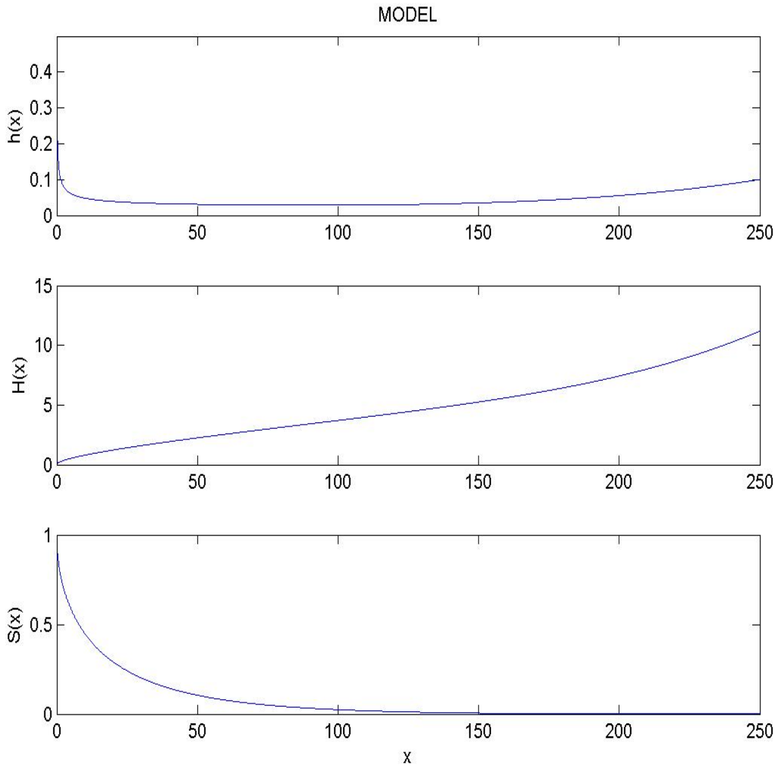

Figure 2: Example 2: Hazard rate of "bath-tube" shape.

Mixture of decreasing h.r., Weibull with $\alpha_1 = 1$, $\beta_1 = 0.5$, constant h.r., i.e. of exponential distribution, $\alpha_2 = 30$, $\beta_2 = 1$, and increasing h.r. of Weibull with $\alpha_3 = 150$, $\beta_3 = 5$.

$h = p_1\, h_1 + p_2\, h_2 + p_3\, h_3$ with $p_1 = 0.2$, $p2 = 0.5$, $p_3 = 0.3$.

A special property of Weibull distribution:

When its hazard rate is multiplied by a constant $c > 0$, we obtain the hazard rate of another Weibull distribution.

Weibull cumulated (integrated) hazard rate is

$$H(x) = a\,x^{\beta} \quad or \quad H(x) = \left(\frac{x}{\alpha}\right)^{\beta}.$$

Hence

$$c\,H(x) = c\,a\,x^{\beta} = a^{*}\,x^{\beta} \;\; or \;\; H(x) = \left(\frac{x}{\alpha^{*}}\right)^{\beta}, \;\; \alpha^{*} = \alpha/c^{1/\beta}.$$

Any mixture of Weibull h. rates is equivalent to simple sum of other Weibull h. rates, parameters of would-be combination are not identifiable.

For instance, in Figure 2, the model is equivalent to

$h = h_1^* + h_2^* + h_3^*$, with $\alpha_1^* = 25$, $\alpha_2^* = 60$, $\alpha_3^* = 190.8$, β-s the same.

Other distribution types have not such property, hence it has a sense to consider a model - linear combination of hazard rates.

Interpretation?

– sum of hazard rates $h = \sum h_j$ corresponds to the hazard rate of a serial system composed from independent items with h_j,

survival time of the system then to minimum of survival times of components, $S = \prod S_j$.

Multiplication $h^* = c \cdot h$ corresponds to proportional change of hazard rate (it is actually the first step to 'proportional hazard regression model' or to frailty models), then $S^* = (S)^c$.

Each approach to mixtures has its specific tools and methods,

often based on iterative optimization of certain criterion, e.g. the ML, Bayes (connected with MCMC),

or on some other distance (KS, CvM, AD, ...).

One traditional method of Weibull distribution fitting uses the (weighted) least squares:

Let CHR be $H(t) = a \cdot t^{\beta}$, data T_i, $i = 1, ..., N$,, estimated C.H.R. $\hat{H}(T_i)$. Then

$$\ln \hat{H}(T_i) \sim \ln(a) + \beta \cdot \ln(T_i).$$

Or, in case of additive hazard rate composed from two Weibull hazard rates, we can use that

$$\hat{H}(T_i) \sim a_1 \, T_i^{\beta_1} + a_2 \, T_i^{\beta_2}, \tag{2}$$

linear at least w.r. to a_j-s.

Artificial data example

Data $X_i, i = 1, .., N$, $N = 500$, were generated from the "bath-tube" model from above.

Hence, it was expected that its h.r. is given by the sum of two Weibull and one exponential components.

MLE:

– It is possible, for given parameters, to evaluate both the log-likelihood and its derivatives,

– while values maximizing the log-likelihood were found by a random search (more effective than an iterative way to solution).

	α_1	α_2	α_3	β_1	β_3
MLE	18.0308	78.1834	197.4708	0.5047	3.6122
90%	10.8970	42.4971	148.4805	0.4549	2.2910
CI	25.1646	113.8697	246.4611	0.5545	4.9334

Table 1: ML Estimates

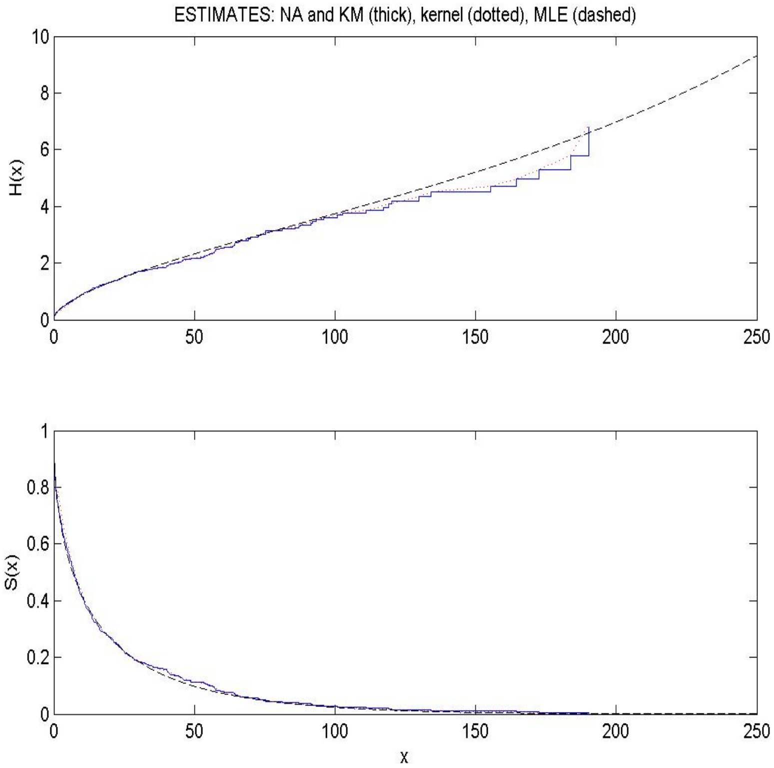

Figure 3: Stepwise: empirical cumulated h.r. and survival function, dotted are kernel-smoothed estimates, dashed curves are based on the MLE.

Real data example

The data are taken from Ruhi, S. (2015), *Application of mixture models for analyzing reliability data: A case study.*

They concern to damage of windshield of aircrafts, time is in 1000 hours.

153 observations: 88 times to failure, 65 are randomly right censored.

Hence the P.L.E. (Product Limit Estimator) of Kaplan and Meier will be used as nonparametric estimator of survival function,

cumulated hazard rate will be estimated by the Nelson-Aalen estimator (N.A.E.).

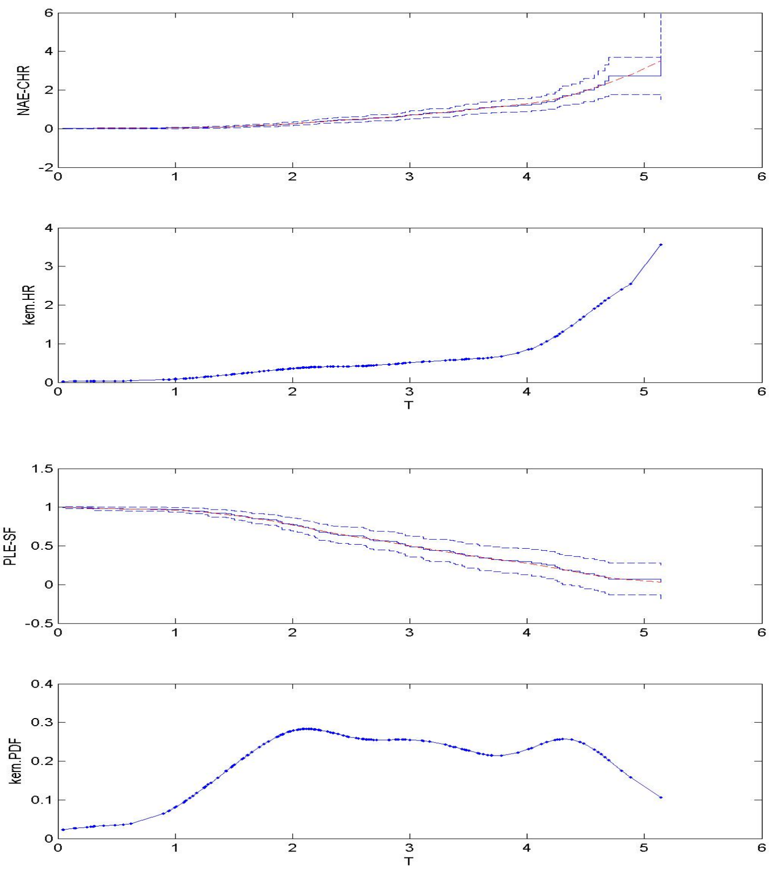

Figure 4: The N.A.E. with 95% CI-s, smoothed estimate of h.r., then the P.L.E. with 95% CI-s, smoothed estimate of density, data of Ruhi (2015).

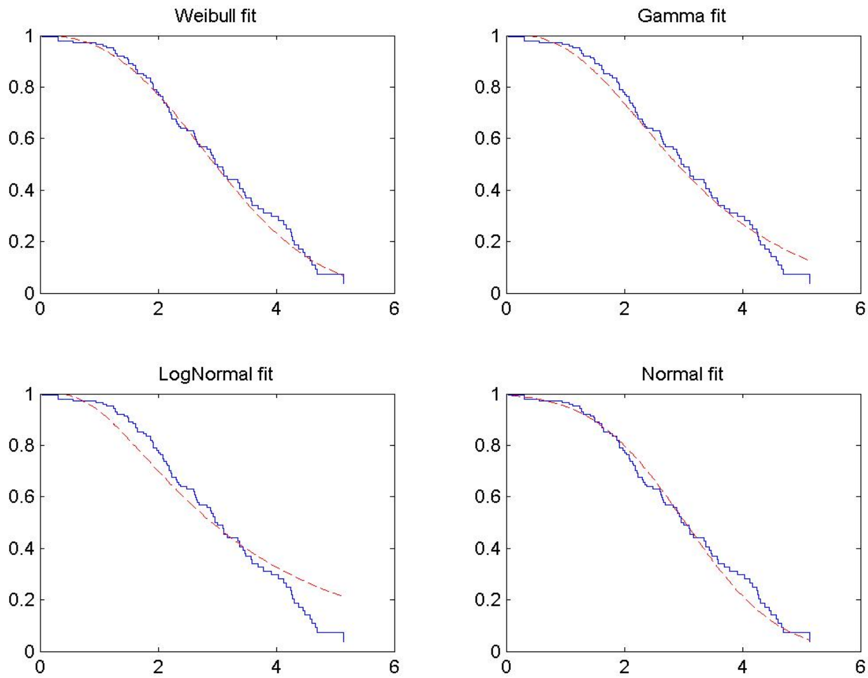

Figure 5: The P.L.E. from data and survival functions (dashed) of several parametric distributions with parameters obtained by the MLE. None of them covers the data sufficiently.

Ruhi (2015) and others tried to model the data by mixture of two distribution densities.

Success was checked by several criteria, e.g. likelihood or Kolmogorov-Smirnov distance of model survival function from nonparametric P.L.E.

His best KS result was 0.0374 obtained by mixture of Weibull and Normal distributions.

I used sum of two Weibull hazard rates.

Again, MLE was used, the best solution approached by a random search.

	α_1	α_2	β_1	β_3
MLE	5.5984	4.0581	1.7433	3.4518
90%	3.3633	2.2963	1.1495	2.6683
CI	7.8335	5.8199	2.3371	4.2353

Table 2: ML Estimates, KS distance = 0.0795.

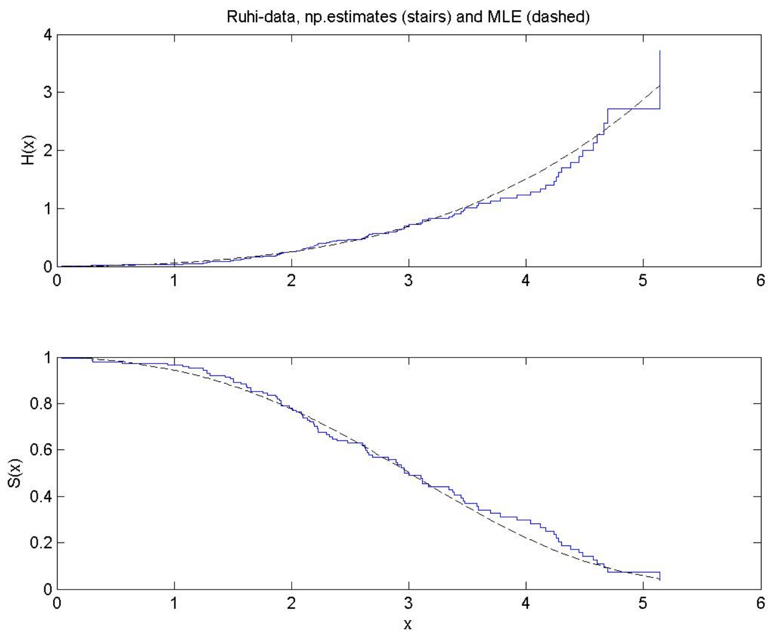

Figure 6: C.H.R. above, with stepwise N.A.E., survival functions are below, with stepwise P.L.E. Dashed curves correspond to mixed hazard rates model obtained by the MLE.

Figure 6 shows that the fit should be improved:

Other distributions? more components?

Change-point models for hazard rates

Change point detection - standard statistical techniques,

there are results dealing with changes of hazard rates, too.

Simplest approach: detecting when the residuals (i.e. reasonably defined deviation of data from actual model) are crossing a given border,

(e.g. statistical process control).

The problem could be viewed also as an incremental construction of a signal model.

"Two change-points incremental" model for Ruhi data:

1-st Weibull hazard rate starting from $t = 0$, two Weibull hazard rates added at two times of change, T_{chj}.

Then $h(t) = h_1(t) + h_2(t - T_{ch1}) \cdot I[t > T_{ch1}] + h_3(t - T_{ch2}) \cdot I[t > T_{ch2}]$ notation $H_j(t) = a_j \cdot t^{\beta_j}$ was used.

Parameters β_j, $j = 1, 2, 3$ and change points were proposed randomly, parameters a_j were obtained by the least squares method weighted by asymptotic variance of the N.A.E. $\hat{H}$ (see relation (2) from Introduction).

Param.:	a_1	a_2	a_3	β_1	β_2	β_3	T_{ch1}	T_{ch2}
Estim.:	0.0400	0.1825	1.4095	2.0210	1.2731	1.4176	1.3941	4.0767

Table 3: Estimates of incremental model parameters and 2 change-points, KS distance =0.0310, re-computed α-s: 4.9145, 3.8047, 0.7849, from the relation $\alpha = 1/a^{1/\beta}$.

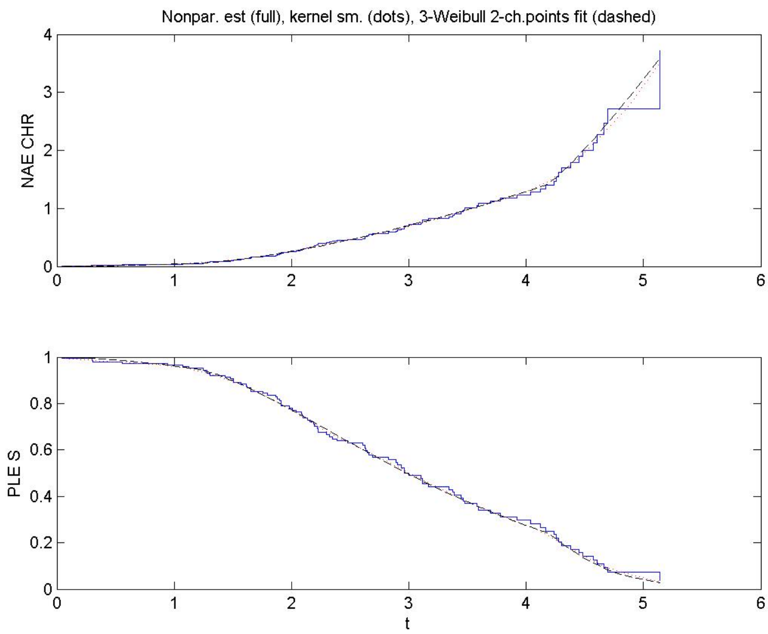

Figure 7: C.H.R. above, with stepwise N.A.E., survival functions are below, with stepwise P.L.E. Dashed curves correspond to change-point model.

No confidence intervals were computed. A variant – random search with the MCMC and Bayes framework, yielding Bayes credibility intervals.

Other attempts

The problem of construction of hazard rate from sum of components can be also interpreted as the problem of a regression model fitting the non-parametric (Nelson–Aalen) estimate of C.H.R.

To keep some interpretation, the model should be constructed from a small set of given parametric function.

A natural choice could be a polynomial model (without intercept),

again, it corresponds to a sum of Weibull cumulated hazard rates.

$$\hat{H}(T_i) \sim \sum_{j=1}^{K} a_j \cdot T_i^j.$$

The best polynomial model fitted to Ruhi data,

using weighted least squares.

Regression diagnostics led to full model with $K = 5$:

Estimated parameters a_j and 95% CI

$a_1 = \quad 0.1299 \quad (0.0264,\ 0.2334)$

$a_2 = -0.3212\,(-0.4811,\ -0.1613)$

$a_3 = \quad 0.2988 \quad (0.2115,\ 0.3862)$

$a_4 = -0.0863\,(-0.1063,\ -0.0663)$

$a_5 = \quad 0.0087 \quad (0.0071,\ 0.0103)$

Residual variance = 0.0021, BIC = -5.7892, min KS dist. = 0.0381.

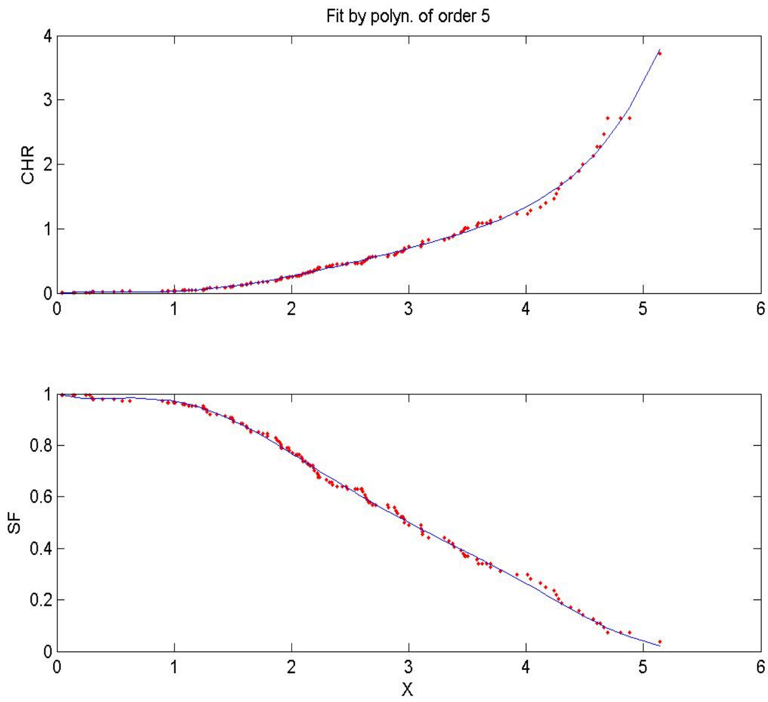

Figure 8: The N.A.E.,P.L.E. (dots) and the best polynomial fit.

Examle – "Strike Data"

In 1967, a strike at a Quebec aluminium smelter resulted in the uncontrolled shutdown of electrolytic cells. The company claimed that the shutdown caused the shorter operating lives of cells operating at the time. The case led to a legal action and initialized a need of a deep statistical analysis of the data, in order to confirm expected higher failure rate after the intervention (shutdown) and to estimate statistically the losses caused by this (eventual) higher rate.

The more details about the case, as well as the complete data were published in "Case Studies in Data Analysis" (CS, 1982), a section of Canadian Journal of Statistics, V. 10 (1982).

Together 572 cell, of 34 types, some passed this shut-down, some not.

Choice of the model for the hazard rate of failure of i-th cell

(reference time t is the age of cell, in days)

$$h_i(t) = h_0(t) \cdot \exp\left\{b(t - U_i) \cdot 1[t > U_i] + c(x_i)\right\}, \quad t \in [0, T_i],$$

where T_i is the survival time of i-th cell,

U_i is the age of i-th cell at the moment of intervention,

function b describes the intervention influence, $b(s) = 0$ for $s < 0$.

x_i is the type of cell i, from 1 to 34.

Cox model with a non-parametric regression function $b(s)$.

Results: The progress of iteration was controlled and its convergence observed from the changes of estimated parameters $c(m)$. Originally, function $b(s)$ was estimated at equidistant points, at each 10 (days). The full domain of s was from 0 to 1 837, so that we obtained 183 values. Then the estimate was secondary smoothed, i.e. the values were averaged (in a weighted way) in a moving window. The graphs of estimated functions b and c are displayed in Figures 1a, 1b.

After we decided to stop the iterations (when changes of values of $c(m)$ were less than 0.1%), we computed the estimate of cumulative baseline hazard function $H_0(t)$. From it, by a kernel smoothing of its increments, we obtained a graph of estimate of $h_0(t)$ in Figure 1c.

Estimated number of lost days - its expectation was about 64 000 days, more details of its distribution can be obtained by random generation from the model, keeping function $b \equiv 0$.

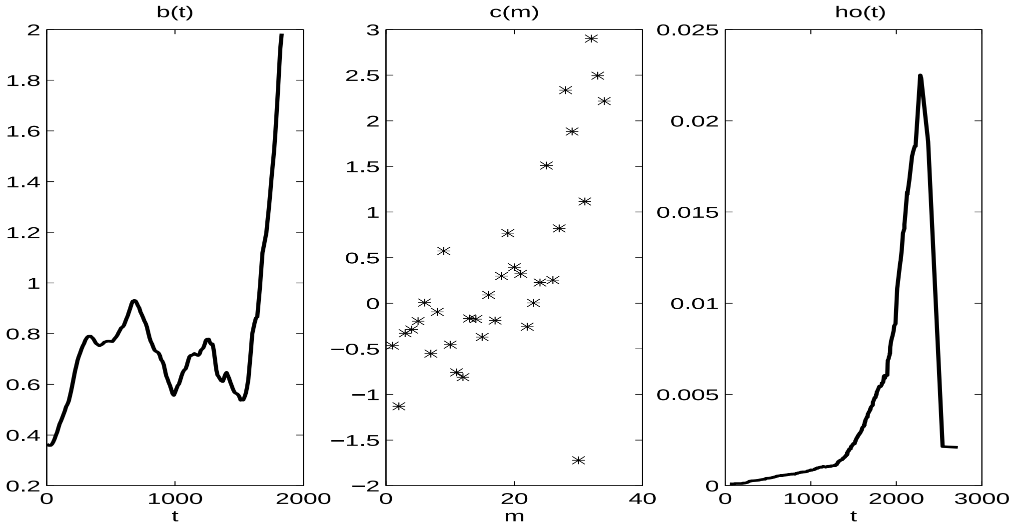

Figure 1: Estimates of functions $b(s)$, $c(m)$ and baseline hazard rate $h_0(t)$ smoothed from estimated $H_0(t)$.

Testing the goodness–of–fit

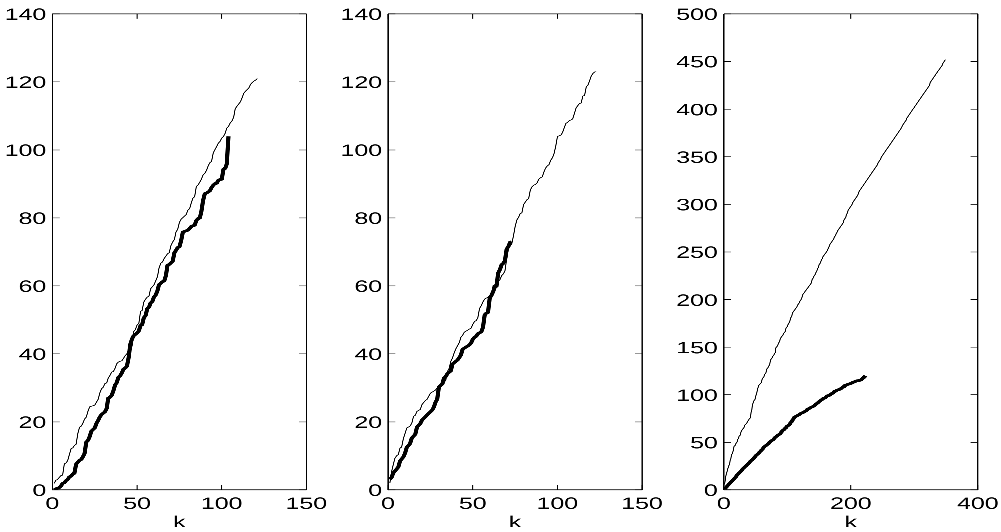

Figure 2: Graphical goodness-of-fit tests. Plots of $\hat{A}_S(T_k)$ for: a) thick - cell types $21-30$, thin - types $1-4$; b) thin - types $11-20$, thick - types $31-34$; c) thick - cells without intervention, thin - cells which passed intervention

Figure 2a shows the plots for cell types 1–4 (121 cells, high survival, mostly without intervention) and cell types 21–30 (104 cells, lower survival, mostly with intervention). Then, in Figure 2b there are the plots for cells of types 11–20 (121 cells with rather high survival, in spite of the intervention) and cells 31–34 (73 experimental cells with low survival, without intervention). It is possible to say that the model fits well for all types of cells.

Figure 2c compares cells which passed or not the intervention. There was a high positive correlation between the survival of a cell and the event that this cell passed the intervention. In other words, the cells which had higher survival (caused only by a random fluctuations) were more likely to survive to (and to pass) the moment of intervention. We tested this connection with the help of 2×2 contingency table (Table 1), with resulting chi-squared test highly significant (i. e. rejecting the hypothesis of independence).

	$T \leq 1\,300$	$T > 1\,300$	Totals
Int.	114	235	349
No int.	184	39	223
Totals	298	274	572

TABLE 1. Resulting value of test statistics $\chi^2_{(1)}$ is 135.46.

References

Ruhi, S. (2015): Application of mixture models for analyzing reliability data: A case study. Open Access Library Journal, 2: e1815. http://dx.doi.org/10.4236/o

Kalbfleisch J. D. and Struthers C. A. (1982): An analysis of the Reynolds Metals Company data. In Case Studies in Data Analysis, Canad. J. Statist. 10, 237–259.

Thomas D. C. (1982): Case analysis using Cox's model. In Case Studies in Data Analysis. Canad. J. Statist. 10, 237–259.

Volf P. (2004). An application of nonparametric Cox regression model in reliability analysis: A case study, Kybernetika vol.40, 5 (2004), 639-648.

Conclusion?

The case studied here was simplified just to sums of Weibull hazard rates. A next problem to explore is to use mixtures (of hazard rates) of other distributions and to study whether it is possible to estimate both parameters of distributions and coefficients of mixture and whether this task is unambiguous....

References

[1] Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure Time Data. New York, Wiley, 2002.

– and other on survival analysis

[2] Ruhi, S. (2015): Application of mixture models for analyzing reliability data: A case study. Open Access Library Journal, 2: e1815. http://dx.doi.org/10.4236/oalib.1101815