Estimation in the Koziol-Green model using a gamma process prior

MICHAL FRIESL

Department of Mathematics, University of West Bohemia in Pilsen, Czech Republic
Univerzitní 22, 306 14 Plzeň, friesl@kma.zcu.cz

ZÁPADOČESKÁ UNIVERZITA V PLZNI

1. The model of Koziol and Green

In survival analysis we deal with times to occurrence of an event. If a variable of interest X with survival function S is censored from the right by a random variable Y (independent of X) we get the random censorship model. We observe a pair

$$Z = \min(X,Y) \quad \text{and} \quad I = I_{[X\le Y]}, \tag{1}$$

where I is an indicator of noncensored observation.

Here we deal with the proportional hazards censorship model of Koziol and Green [7], in which survival function S_Y of the time censor is moreover assumed to satisfy

$$\boxed{S_Y(x) = (S(x))^\gamma, \quad x>0,} \tag{2}$$

with some positive constant γ. Or, using cumulative hazard rate $\Lambda = -\ln S$ of X, the rate of Y equals $\gamma\Lambda$. Equivalently, Z and I are independent, in continuous case $\mathrm{P}(X\le Y) = (1+\gamma)^{-1}$. See [3] for a review of implications to inference.

We work with a random sample of size n from (1) with (2). The sample $Z_1,\dots,Z_n$ of the observed minima consists of $N\le n$ distinct times (we allow for ties) denoted by

$$T_1 < \cdots < T_N,$$

we also define $T_0 = 0$ and $T_{N+1} = \infty$. Let

$$N_j = \#\{k; Z_k > T_j\}, \quad j = 0,\dots,N,$$

be the number of items failed or censored after T_j and let U_j and C_j denote the number of uncensored and censored observations with time $Z_k = T_j$.

2. A nonparametric Bayesian approach

Using nonparametric Bayesian setup (introduced by [2]) we are not limited with possible shapes of S to certain parametric family. Instead, S is chosen at random from a class of potentially all survival functions. Of course then the prior is not defined for several parameters of the family but describes distribution of the function S considered as a stochastic process, see [8] for a review.

In survival analysis, processes neutral to the right [1], i.e. with corresponding cumulative hazard rate process Λ having independent increments, prove manageable. Specifically, we will assume that Λ is a gamma process

$$\Lambda(0) = 0 \quad \text{and} \quad \Lambda(s,t) = \Lambda(t)-\Lambda(s) \sim \mathrm{G}(n_0, n_0\Lambda_0), \quad 0\le s\le t,$$

where Λ_0 is cumulative hazard rate of some continuous distribution, $n_0>0$, and $\mathrm{G}(a,p)$ denotes the gamma distribution with shape parameter p and scale parameter $1/a$. As we have

$$\mathrm{E}\,\Lambda(t) = \Lambda_0(t) \quad \text{and} \quad \operatorname{var}\Lambda(t) = 1/n_0, \quad t>0,$$

the parameters Λ_0 and n_0 represent a "central distribution" and accuracy of prior information, respectively.

Remind that although we assume continuous Λ_0, the realizations of Λ with probability 1 relate to discrete distributions and have infinitely many jumps in every interval. Also $\mathrm{E}S(t) = \mathrm{E}\exp(-\Lambda(t))$ does not exactly equal to $\exp(-\Lambda_0(t))$.

Let γ have a prior density $\pi(\gamma)$ with respect to some measure μ on $(0,\infty)$ and be independent of Λ.

Had the censoring distribution been independent of Λ, standard formulæ of [4] would immediately apply (the $Y's$ could be considered fixed). But due to (2) this is not the case and we develop an estimator that will utilize the additional information from Y.

3. Posterior distribution and estimators

We reflect (2) and derive the posterior distribution described bellow. For $m = 1,2$ denote

$$N_j^m(\gamma) = n_0 + N_j(1+\gamma) + m, \quad j=0,\dots,N, \quad \text{and}$$
$$c_j^m(\gamma) = \sum_{k=1}^{U_j}\sum_{\ell=1}^{C_j}(-1)^{k+\ell}\binom{U_j}{k}\binom{C_j}{\ell}\ln\frac{N_j^m(\gamma)+C_j}{N_j^m(\gamma)+C_j+k+\ell\gamma},$$
$$q_j^m(\gamma) = \big(N_{j-1}^m(\gamma)\big)^{-n_0\Lambda_0(T_{j-1},T_j)}c_j^m(\gamma), \qquad j=1,\dots,N.$$

Given γ the process Λ corresponds to a neutral to the right distribution which also has jumps at observation times. The increments of Λ over intervals not containing T_j's are (given γ) gamma distributed, for $(s,t]\subset(T_{j-1},T_j)$ it is

$$(\Lambda(s,t)\,|\,\text{data},\gamma) \sim \mathrm{G}(N_{j-1}^0(\gamma), n_0\Lambda_0(s,t)).$$

The jump at T_j has probability density function

$$x^{-1}\mathrm{e}^{-(N_j^0(\gamma)+C_j)x}(1-\mathrm{e}^{-x})^{U_j}(1-\mathrm{e}^{-\gamma x})^{C_j}/c_j^0(\gamma), \quad x>0,$$

where $c_j^0(\gamma)$ is a norming constant. Marginal posterior distribution of γ has density

$$\pi(\gamma\,|\,\text{data}) \propto \Big(\prod_{j=1}^N q_j^0(\gamma)\Big)\pi(\gamma). \tag{3}$$

We average posterior conditional expected value of $S(s) = \exp(-\Lambda_s)$ given γ with (3) to get

$$A(s) = \int\Big(\prod_{j<i}q_j^1(\gamma)\Big)\Big(\prod_{j\ge i}q_j^0(\gamma)\Big)\Big(\frac{N_{i-1}^0(\gamma)}{N_{i-1}^1(\gamma)}\Big)^{n_0\Lambda_0(T_{i-1},s)}\mathrm{d}\pi(\gamma|\text{data}),$$

where $i = i(s) = \max\{k; T_k\le s\}$ denotes the interval which contains s. The Bayes estimator of the survival function $S(t)$ taken as its posterior expected value then reads

$$\boxed{\widehat{S}(t) = A(t)/A(0),} \tag{4}$$

an expression explicit up to the integration with respect to γ. A warning—one should consider a number of significant digits needed to evaluate $c_j^m(\gamma)$ for larger values of U_j and C_j directly.

4. Examples

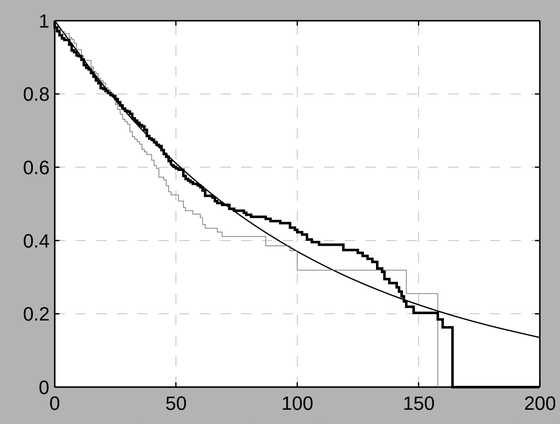
We illustrate the above estimator in two examples of data sets from literature. We also display estimators that do not use the Koziol-Green assumption (2).

Figures 1 and 2 correspond to 211 state IV prostate cancer patients treated with ostregon at V.A.C.U.R.G. as presented in [5]. We use the exponential distribution with mean 100 month (tested to fit in [7] using proportionality assumption) as a centre of the prior gamma process, although it is rejected by other tests and the proportionality assumption may not hold.

Figures 3–5 relate to the Channing House data [6] on lifetimes of 97 men, ignoring left truncation, for which the Koziol-Green model holds [3]. As a prior mean we impose the Weibull distribution with parameters obtained by transformation of quantile estimates of associated distribution of Z.

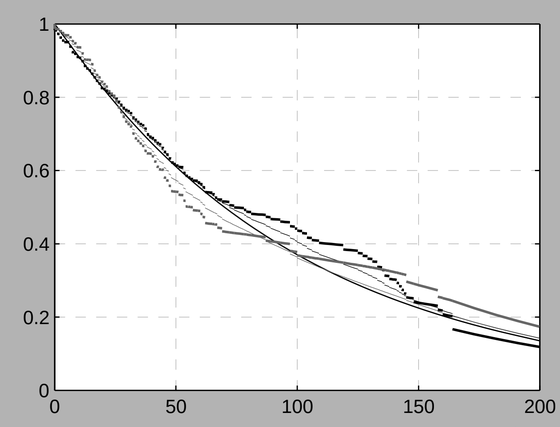
In both examples, the prior distribution for γ is uniform on the set of nine values of γ yielding $(1+\gamma)^{-1} = 0.1,\dots,0.9$. Except for very sharp prior knowledge of γ the choice of prior distribution for γ does not seem to affect the results much.

1



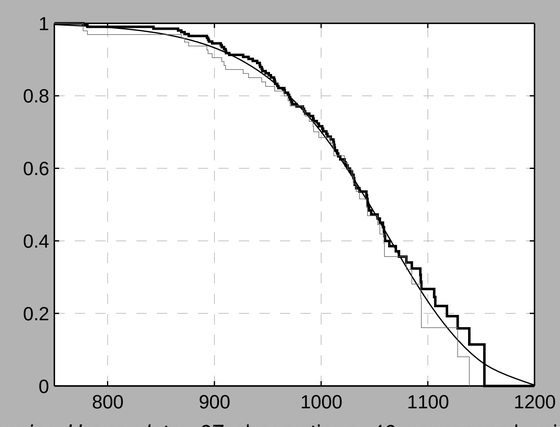
Prostate cancer data: 211 observations, 90 uncensored, minimum 0, maximum 164 months. Abdushukurov, Cheng and Lin (ACL) and Kaplan-Meier (KM) estimators (thick black and thin gray line) of S together with $\mathrm{Exp}(100)$ survival function.

2



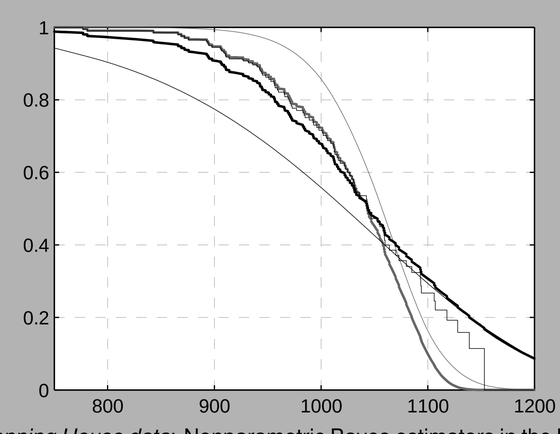
Prostate cancer data: Nonparametric Bayes estimators of S with and without (black and gray) the Koziol-Green model assumption, together with the prior centre. Using Λ_0 of $\mathrm{Exp}(100)$, $n_0 = 10$ (thick) and $n_0 = 100$ (thin).

3



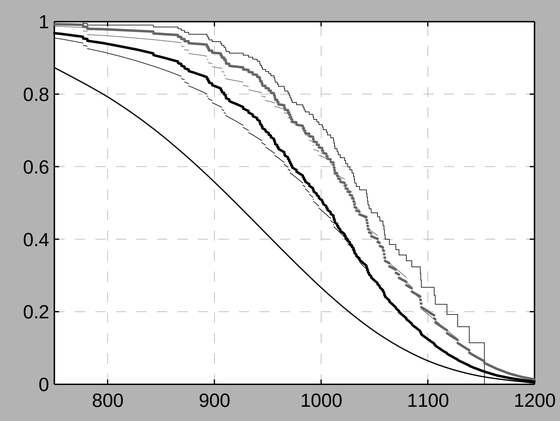
Channing House data: 97 observations, 46 uncensored, minimum 775, maximum 1153 month. ACL (thick black) and KM (thin gray) estimators of S together with survival function of Weibull distribution with cumulative hazard rate $\Lambda_0(x) = (x/\theta)^b$, $x>0$, $\theta = 1071$, $b = 15.9$.

4



Channing House data: Nonparametric Bayes estimators in the Koziol-Green model (thick) using $n_0 = 50$ and Weibull shape parameter $b/2$ (black) and $2b$ (gray), together with ACL estimator and prior mean survival functions (thin).

5



Channing House data: Nonparametric Bayes estimators of S with (thick) and without (thin) the Koziol-Green model assumption using $n_0 = 50$ (black), $n_0 = 10$ (gray) and $\mathrm{Weib}(0.9\theta, b/2)$, together with mean prior (below) and ACL estimator (up).

5. References

[1] DOKSUM, K., *Tailfree and neutral random probabilities and their posterior distributions*, Ann. Probability **2** (1974), no. 2, 183–201.

[2] FERGUSON, T. S., *A Bayesian analysis of some nonparametric problems*, Ann. Statist. **1** (1973), no. 2, 209–230.

[3] CSÖRGŐ, S., *Estimation in the proportional hazards model of random censorship*, Statistics **19** (1988), no. 3 , 437–463.

[4] FERGUSON, T. S., AND PHADIA, E. G., *Bayesian nonparametric estimation based on censored data*, Ann. Statist. **7** (1979), no. 1, 163–186.

[5] HOLLANDER, M., AND PROSCHAN, F., *Testing to determine the underlying distribution using randomly censored data*, Biometrics **35** (1979), no. 2, 393–401.

[6] HYDE, J., *Testing survival under right censoring and left truncation*, Biometrika **64** (1977), no. 2, 225–230.

[7] KOZIOL, J. A., AND GREEN, S. B., *A Cramér-von Mises statistic for randomly censored data*, Biometrika **63** (1976), no. 3, 465–474.

[8] WALKER, S. G., DAMIEN, P., LAUD, P. W., AND SMITH, A. F. M., *Bayesian nonparametric inference for random distributions and related functions*, with discussion and a reply by the authors, J. R. Stat. Soc. Ser. B Stat. Methodol. **61** (1999), no. 3, 485–527.