Ref: Austrian Journal of Statistics
35 (2006), no. 2&3, 253–260

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

Michal Friesl
University of West Bohemia in Pilsen, Czech Republic

Abstract: The paper deals with nonparametric Bayes estimators in the Koziol-Green model of random censorship. A gamma process is assumed as a prior distribution for cumulative hazard rate and the Bayes estimator incorporating the proportional hazards censorship property of the model is presented. The estimator is also applied to two data sets from literature.

Keywords: Nonparametric Bayes Estimator, Survival Function, Random Censorship, Informative Censoring, Channing House Data.

1 Introduction

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

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

where I is indicator of noncensored observation. In present paper we consider the proportional hazards censorship model of Koziol and Green (1976), in which survival functions $S(t) = \mathrm{P}(X > t)$, $t > 0$, of X and $S_Y(t) = \mathrm{P}(Y > t)$, $t > 0$, of the time censor are moreover assumed to satisfy

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

with some positive constant γ. Using cumulative hazard rate $\Lambda(t) = -\ln S(t)$ of X, condition (2) states that the cumulative rate of Y equals $\gamma\Lambda$. For continuous distributions (but this will not actually be our case) the model can similarly be defined by the proportionality assumption of noncumulative hazard rates ($\lambda(t) = -S'(t)/S(t)$ of X, e.g.) and the conditions are also equivalent to independence of random variables Z and I. Then the constant γ is linked with the probability that the uncensored time will be observed, the relation is $\mathrm{P}(I = 1) = \mathrm{P}(X \le Y) = (1+\gamma)^{-1}$. Csörgő (1988) reviews various implications of assumption (2) to inference.

The subject of the present paper is nonparametric Bayesian estimation of the survival function S. Using nonparametric Bayesian setup (introduced by Ferguson, 1973) we are not limited with possible shapes of S to certain parametric family. Instead, S (or Λ) is picked from a class of potentially all survival functions (cumulative hazard rates). Of course then the prior does not act on a space of values of several parameters of the family but describes distribution of the function S (or function Λ) considered as a stochastic process, see Walker, Damien, Laud, and Smith (1999) for a review.

In survival analysis, processes neutral to the right (Doksum, 1974), i.e. with corresponding cumulative hazard rate process Λ having independent increments, prove manageable as prior processes. Bellow we will use the gamma process. The next section introduces notation and prior distribution in detail, section 3 describes the Bayes estimator. In section 4 an application of the estimator to two data sets taken from literature is shown.

2 The model

We work with data formed by a random sample $(Z_1,I_1),\dots,(Z_n,I_n)$ from (1) assuming (2). The sample $Z_1,\dots,Z_n$ of the observed minima consists of say $N \leq n$ distinct times (we allow for ties) denoted by

$$T_1 < \cdots < T_N, \tag{3}$$

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, i.e. N_j is the number of items at risk at time T_{j+1}. Let

$$U_j = \sum_{k;Z_k=T_j} I_k \quad \text{and} \quad C_j = \sum_{k;Z_k=T_j} (1-I_k), \quad j = 1,\dots,N,$$

denote the number of uncensored and censored observations with times Z_k equal to T_j and let

$$i = i(s) = \max\{k; T_k \leq s\} + 1 \in \{1,\dots,N+1\} \tag{4}$$

indexes the interval $[T_{i-1},T_i)$ which contains s.

We will assume that prior distribution of the unknown parameter Λ is a gamma process

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

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

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

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

Recall that even if the 'mean' cumulative hazard rate Λ_0 is continuous, the realization of the gamma process Λ is with probability 1 a cumulative hazard rate of some discrete distribution (this is where positive probability of ties in data arises) and has infinitely many jumps in every interval on which Λ_0 increases. Note also that $\mathrm{E}S(t) = \mathrm{E}\exp(-\Lambda(t))$ considered as a survival function does not equal to $\exp(-\mathrm{E}\Lambda(t)) = \exp(-\Lambda_0(t))$ but rather to product integral $\prod_{(0,t]}(1-\mathrm{E}\Lambda^*(\mathrm{d}s))$ where $\Lambda^*(t) = -\int_{(0,t]}(S(\mathrm{d}s)/S(s-))$ is a modification of cumulative hazard rate related to S which is occasionally used.

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

If the censoring distribution was independent of Λ, standard formulae of Ferguson and Phadia (1979) would apply (the $Y's$ could be considered fixed) and yield the estimator

$$\widehat{S}_{\mathrm{FP}}(t)=\left(\prod_{j=1}^{i-1}\left(\frac{n_0+N_{j-1}}{n_0+N_{j-1}+1}\right)^{n_0\Lambda_0(T_{j-1},T_j)}J_j\right)\cdot\left(\frac{n_0+N_{i-1}}{n_0+N_{i-1}+1}\right)^{n_0\Lambda_0(T_{i-1},t)} \tag{5}$$

where $i=i(t)$ is defined in (4) and J_j is a term that generates a jump; if $U_j=0$, then $J_j=1$ and there is no jump at T_j. But regarding (2) we can find a better estimator that will utilize the additional information from Y.

3 Posterior distribution and estimators

The posterior distribution is described in terms of neutral to the right process again. We use the notation

$$\begin{aligned}
M_j(\gamma)&=n_0+N_j(1+\gamma), \quad j=0,\dots,N, \quad \text{and}\\
c_j(\gamma)&=\sum_{k=1}^{U_j}\sum_{\ell=1}^{C_j}(-1)^{k+\ell}\binom{U_j}{k}\binom{C_j}{\ell}\ln\frac{M_j(\gamma)+C_j}{M_j(\gamma)+C_j+k+\ell\gamma},\\
q_j(\gamma)&=\left(M_{j-1}(\gamma)\right)^{-n_0\Lambda_0(T_{j-1},T_j)}c_j(\gamma), \qquad j=1,\dots,N.
\end{aligned}$$

Proposition 1. *Given γ the process Λ a posteriori corresponds to a neutral to the right distribution which also has jumps at observation times (3). The increments of Λ over intervals not containing T_j's are (given γ) gamma distributed, specifically for $(s,t]\subset(T_{j-1},T_j)$ we have*

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

The size of the jump at T_j has probability density function

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

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

$$\pi(\gamma|\,data)\propto\left(\prod_{j=1}^{N}q_j(\gamma)\right)\pi(\gamma) \tag{6}$$

with respect to μ.

Proof. The posterior distribution follows by recognizing the alleged distributions in posterior moment generating functions of increments of Λ, similarly to the case with no ties in Friesl (2005). □

We may note that the conditional distribution of jump sizes is not affected by dislocation of the observed times T_k and does not depend on the assumed prior central rate Λ_0 either. It is given solely by configuration of counts of uncensored and censored observations. This is an implication of homogeneity of the Lévy measure of the prior process.

Using the independent increments property of conditional posterior distribution of the process Λ for given γ we can express the posterior conditional expected value of $S(s) = \exp(-\Lambda(s)) = \left(\prod_{j<i} \mathrm{e}^{\Lambda(T_j-)-\Lambda(T_{j-1})}\mathrm{e}^{\Lambda(T_j)-\Lambda(T_j-)}\right)\mathrm{e}^{\Lambda(s)-\Lambda(T_{i-1})}$ given γ, here $i = i(s)$ is from (4). Averaging further over γ with weights given by the right hand side of (6) we get

$$A(s) = \int \Big(\prod_{j<i} q_j^+(\gamma)\Big)\Big(\prod_{j\geq i} q_j(\gamma)\Big)\left(\frac{M_{i-1}(\gamma)}{M_{i-1}^+(\gamma)}\right)^{n_0\Lambda_0(T_{i-1},s)} \pi(\gamma\,|\,\mathrm{data})\,\mathrm{d}\mu(\gamma),$$

where $q_j^+(\gamma)$ is defined in the same way as $q_j(\gamma)$ with $M_.(\gamma)$'s in definitions of $q_j(\gamma)$ and $c_j(\gamma)$ replaced by $M_.^+(\gamma) = M_.(\gamma)+1$. The estimator of S follows.

Proposition 2. *The Bayes estimator of the survival function $S(t)$ taken as its posterior expected value reads*

$$\widehat{S}(t) = A(t)/A(0).$$

Proof. To get the expectation we divide $A(t)$ by normalizing constant of (6) which can be written as $A(0)$. □

The expression is explicit up to the integration with respect to γ. No (numerical) integration is needed provided the range of γ is finite.

4 Examples

As an illustration we consider two data sets from literature. Besides the above estimator $\widehat{S}$ we display the Bayes estimator without the Koziol-Green assumption (i.e. $\widehat{S}_{\mathrm{FP}}$ from (5))

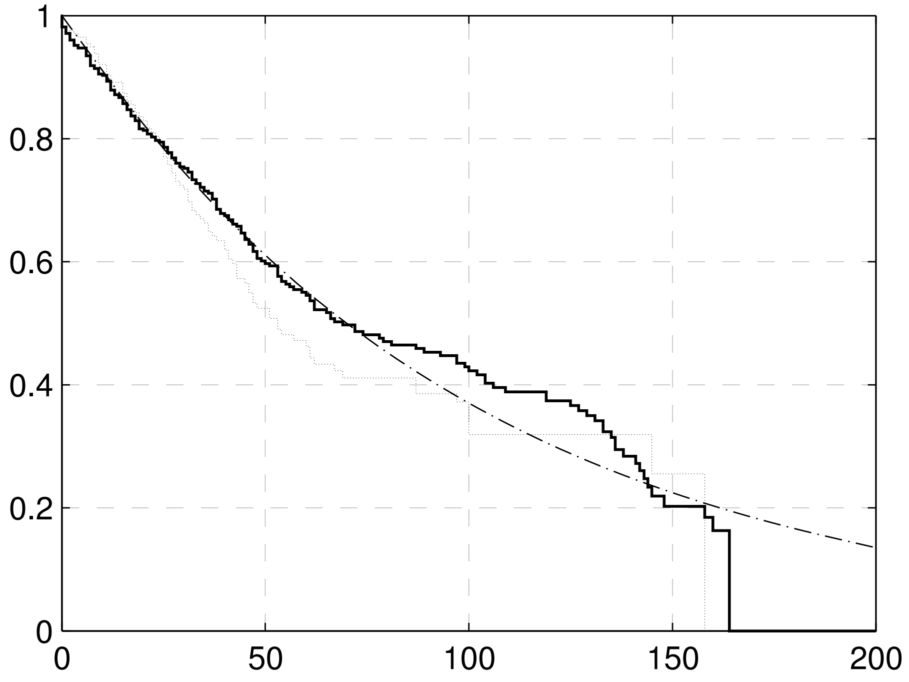

Figure 1: Prostate cancer data. ACL and KM estimators (thick solid and thin dotted stairs) of S together with Exp(100) survival function (smooth dashed-dotted curve).

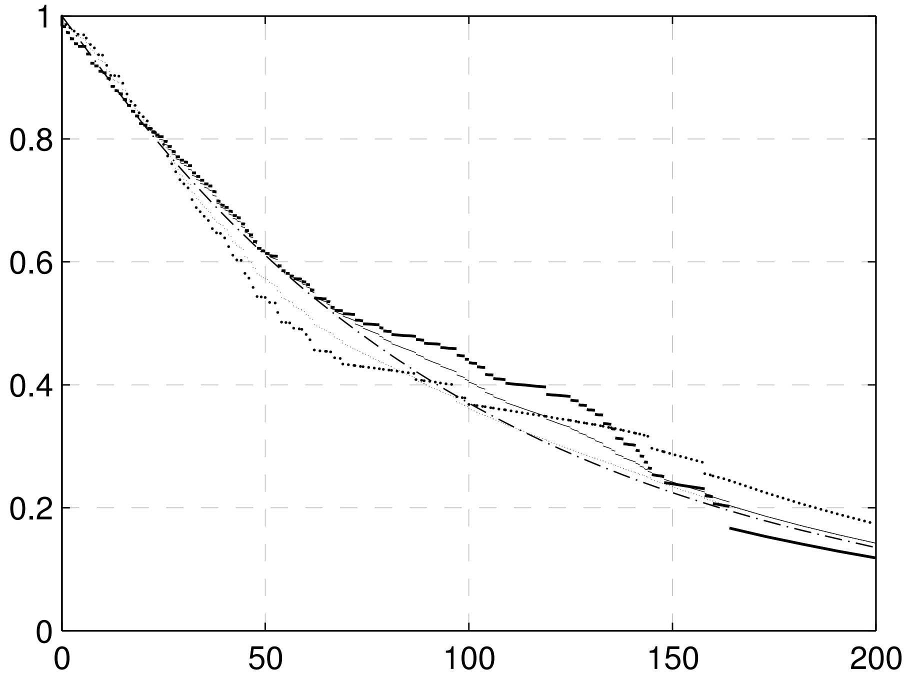

Figure 2: Prostate cancer data. Nonparametric Bayes estimators of S with (solid) and without (dotted) the Koziol-Green model assumption, together with $\exp(-\Lambda_0(t))$ (dashed-dotted). Using Λ_0 of $\mathrm{Exp}(100)$, $n_0 = 10$ (thick) and $n_0 = 100$ (thin).

and also the standard Kaplan-Meier (KM) nonparametric estimator

$$\widehat{S}_{\mathrm{KM}}(t) = \prod_{j=1}^{i(t)-1} \Big(1 - \frac{U_j}{N_{j-1}}\Big), \quad t < T_N,$$

and Abdushukurov, Cheng and Lin (ACL) estimator

$$\widehat{S}_{\mathrm{ACL}}(t) = \Big(\frac{\#\{k; Z_k > t\}}{n}\Big)^{\Sigma_1^n I_j/n} = \Big(\frac{N_{i(t)-1}}{n}\Big)^{\Sigma_1^N U_j/n}.$$

The ACL estimator reflects the proportionality property of the Koziol-Green model simply by taking a power (the exponent equals to estimated proportion of uncensored items) of the sample survival function estimator of variable Z.

Where applicable the prior distribution for γ is taken uniform on the set of nine values of γ yielding $(1+\gamma)^{-1} = 0.1, \ldots, 0.9$. Several other choices of prior on this set were also tested and except for very sharp prior knowledge of γ the choice of probabilities does not seem in both examples to affect the results much.

The first data set are survival times of 211 state IV prostate cancer patients treated with ostregon at V.A.C.U.R.G. as presented in Hollander and Proschan (1979). Among $n = 211$ observations 90 are uncensored, minimum is 0 and maximum 164 months. The number of distinct times among observations is 97, of which 53 are ties. In figure 1 we can see Kaplan-Meier and ACL estimators of the survival function S. The graph is completed by the survival function of the exponential distribution with mean 100 month, which we use as a centre of the prior gamma process when computing nonparametric Bayes estimators. This distribution was tested to fit the data in Koziol and Green (1976) using proportionality assumption, but it is rejected by other tests; the proportionality assumption for these data may not hold.

Figure 2 displays nonparametric Bayes estimators $\widehat{S}$ and $\widehat{S}_{\mathrm{FP}}$ corresponding to the case with (without, respectively) the proportionality assumption. As a prior sample size we select $n_0 = 10$ and $n_0 = 100$. While $\widehat{S}_{\mathrm{FP}}$ only jumps at times with at least one uncensored observation (at times with no uncensored observation we may however note nonsmooth changes in slope), the estimator $\widehat{S}$ jumps at all observation times.

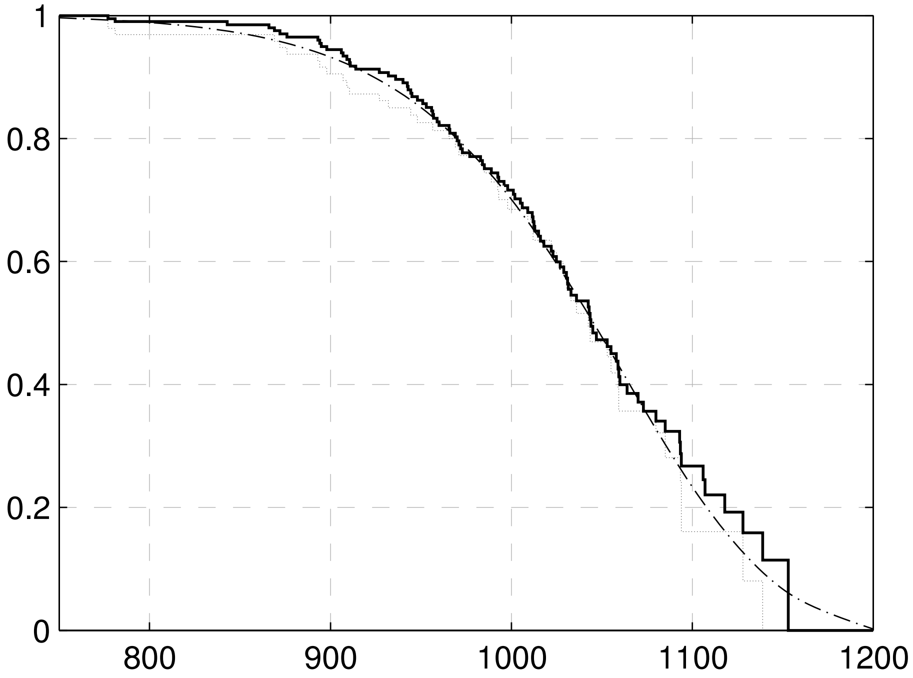

Figure 3: Channing House data. ACL (thick solid) and KM (thin dotted) estimators of S together with survival function of Weibull distribution with cumulative hazard rate $\Lambda_0(x) = (x/\theta)^b$, $x > 0$, $\theta = 1071$, $b = 15.9$ (dashed-dotted).

Figures 3–5 refer to the Channing House data (Hyde, 1977) on lifetimes of 97 men, ignoring left truncation, which satisfy the Koziol-Green model (Csörgő, 1988). The data consist of $n = 97$ observations out of which 46 are uncensored. Minimal observed lifetime is 775, maximal 1153 month. We find 10 duplicate observation times and 2 triplicated. Figure 3 shows $\widehat{S}_{\mathrm{KM}}$ and $\widehat{S}_{\mathrm{ACL}}$ estimators, and a reference prior mean we impose, namely the Weibull distribution with parameters obtained by transformation of quantile estimates of associated Weibull distribution of Z.

Figure 4 reflects influence of the choice of shape parameter b in the prior (twice the estimated value and half of it) on the nonparametric Bayes estimator when $n_0 = 50$. We can see a difference in the tails. Figure 5 illustrates the effect of the strength of belief in the prior to the results. We take some incorrect prior and confront $n_0 = 50$ with $n_0 = 10$.

A final note concerns practical computation of the estimator. One should consider a number of significant digits needed to evaluate $c_j(\gamma)$ and $c_j^+(\gamma)$ to get a sufficient accuracy

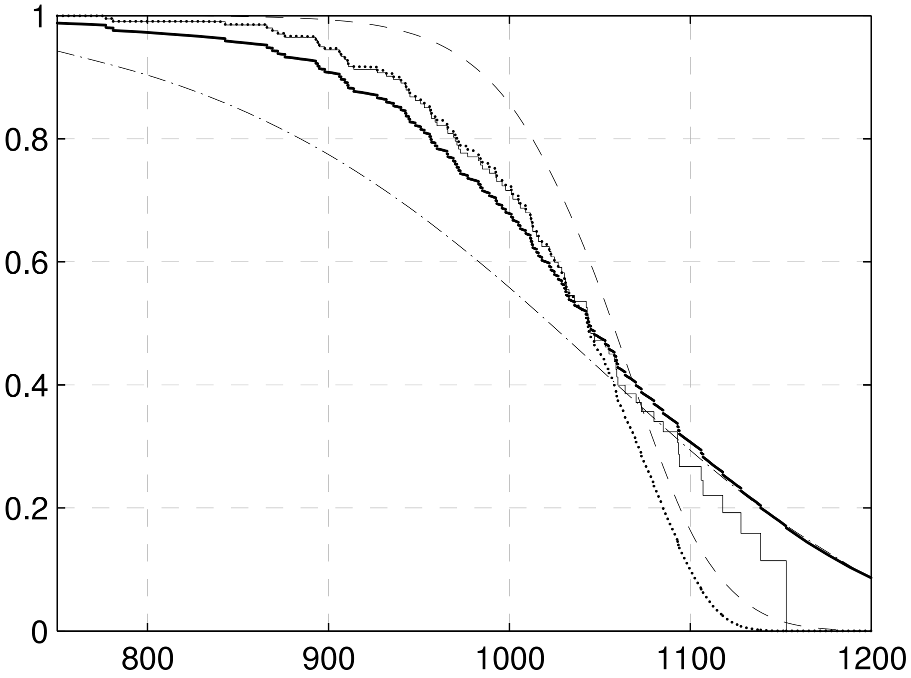

Figure 4: Channing House data. Nonparametric Bayes estimators in the Koziol-Green model (thick) using $n_0 = 50$ and Weibull shape parameter $b/2$ (solid) and $2b$ (dotted), together with ACL estimator (thin stairs in the middle) and prior mean survival functions (dashed-dotted and dashed).

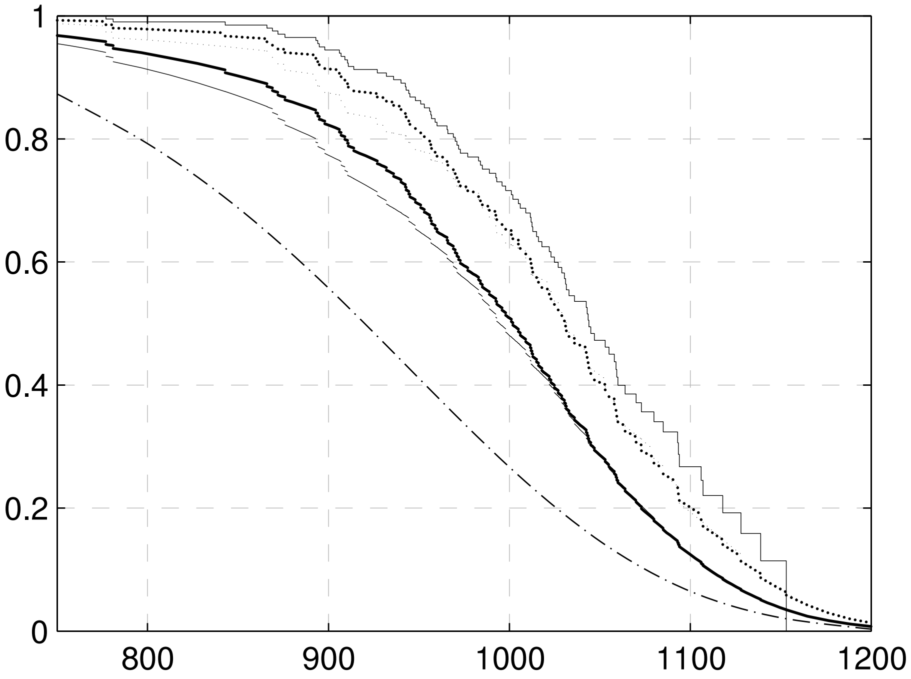

Figure 5: Channing House data. Nonparametric Bayes estimators of S with (thick) and without (thin) the Koziol-Green model assumption using $n_0 = 50$ (solid), $n_0 = 10$ (dotted) and Weib$(0.9\theta, b/2)$, together with mean prior (bellow) and ACL estimator (up).

for large values of U_j and C_j.

Acknowledgement

The author was supported by the Research Plan MSM 4977751301.

References

Csörgő, S. (1988). Estimation in the proportional hazards model of random censorship. *Statistics*, *19*(3), 437–463.

Doksum, K. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probability*, *2*, 183–201.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.*, *1*, 209–230.

Ferguson, T. S., Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.*, *7*(1), 163–186.

Friesl, M. (2005). Neparametrické bayesovské odhady v Koziolově-Greenově modelu náhodného cenzorování. In J. Antoch G. Dohnal (Eds.), *ROBUST 2004* (pp. 93–100). Praha: JČMF.

Hollander, M., Proschan, F. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics*, *35*(2), 393–401.

Hyde, J. (1977). Testing survival under right censoring and left truncation. *Biometrika*, *64*(2), 225–230.

Koziol, J. A., Green, S. B. (1976). A Cramér-von Mises statistic for randomly censored data. *Biometrika*, *63*(3), 465–474.

Walker, S. G., Damien, P., Laud, P. W., Smith, A. F. M. (1999). Bayesian nonparametric inference for random distributions and related functions. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, *61*(3), 485–527.

Author’s address:

Michal Friesl
Department of Mathematics
University of West Bohemia in Pilsen
Univerzitní 22, 306 14 Plzeň
Czech Republic

E-mail: friesl@kma.zcu.cz
http://www.kma.zcu.cz/Friesl