

NEPARAMETRICKÉ BAYESOVSKÉ ODHADY V KOZIOLOVĚ-GREENOVĚ MODELU NÁHODNÉHO CENZOROVÁNÍ

Michal Friesl

Klíčová slova: Funkce spolehlivosti, neparametrické bayesovské odhady, náhodné cenzorování, Koziolův-Greenův model, gama proces.

Abstrakt: V příspěvku je odvozen neparametrický bayesovský odhad funkce spolehlivosti při Koziolově-Greenově modelu náhodného cenzorování, jako apriorní rozdělení kumulativní intenzity poruch se předpokládá gama proces.

1 Úvod

Mezi bayesovskými metodami jsou jako neparametrické označovány ty, které počítají s apriorními rozděleními nikoli pro konečný počet parametrů určitého rozdělení, ale pro obvykle nekonečněrozměrné parametry, jako je např. distribuční funkce. Ferguson [2] představil jako apriorní model na prostoru pravděpodobnostních měr rozdělení Dirichletova procesu a od té doby byla zkoumána jako apriorní celá řada procesů (beta, gama, smíšené, zprava neutrální) s uplatněním v různých modelech analýzy dat o přežití či teorie spolehlivosti, včetně cenzorování. V následujícím textu se budeme věnovat odhadu funkce přežití v Koziolově-Greenově modelu náhodného cenzorování [7]. V tomto modelu je rozdělení cenzoru svázáno s rozdělením cenzorované veličiny a nelze proto použít tradiční odhady.

Ještě předtím ale stručně přiblížíme princip neparametrických bayesovských metod, některá apriorní rozdělení a uplatnění těchto metod v analýze dat o přežití obecně. Z přehledných článků shrnujících tuto problematiku jmenujme [4] či [11].

2 Neparametrická apriorní rozdělení

Uvažujme pozorování X s hodnotami v měřitelném prostoru $(\mathcal{X}, \mathcal{A})$, jejichž rozdělení Q neznáme. Při parametrickém přístupu předpokládáme, že rozdělení Q je určitého typu, že pochází z určité úzké rodiny rozdělení parametrizované konečným počtem parametrů, např. že je normální, $Q = Q_{\mu,\sigma^2} = \mathrm{N}(\mu, \sigma^2)$. V bayesovské statistice považujeme parametry za náhodné veličiny, apriorní informace o nich je vyjádřena apriorním rozdělením na množině možných hodnot parametrů (v uvedeném příkladu na $\mathbf{R} \times \mathbf{R}^+$).

Při "neparametrickém" přístupu se v úvahách o Q neomezujeme, třídou možných rozdělení pozorování mohou být potenciálně všechna rozdělení na $(\mathcal{X}, \mathcal{A})$, neznámým parametrem je sama pravděpodobnostní míra Q. Pravděpodobnosti $Q(A)$ jednotlivých množin z $\mathcal{A}$ považujeme za náhodné veličiny

a celou náhodnou pravděpodobnost $Q = (Q(A), A \in \mathcal{A})$ můžeme chápat jako náhodný proces indexovaný prvky z $\mathcal{A}$. Apriorní informace o něm je popsána apriorním rozdělením na množině pravděpodobnostních měr, resp. jejich charakteristik, např. distribučních funkcí v případě pozorování s reálnými hodnotami.

Asi nejznámějším neparametrickým apriorním rozdělením je rozdělení *Dirichletova procesu* [2].

Definice 2.1. *Řekneme, že proces Q je Dirichletův s parametrem $\alpha = n_0 Q_0$, kde $n_0 \in \mathbf{R}^+$ a Q_0 je pravděpodobnostní míra na $(\mathcal{X}, \mathcal{A})$, pokud pro libovolný rozklad $A_1, \dots, A_k \in \mathcal{A}$, $\bigcup A_i = \mathcal{X}$ disjunktně, je*

$$(Q(A_1), \dots, Q(A_k)) \sim \mathrm{D}(\alpha(A_1), \dots, \alpha(A_k)),$$

kde D *značí Dirichletovo rozdělení. Značíme $Q \sim \mathcal{D}(\alpha)$.*

Je $\mathrm{E}\, Q(A_i) = Q_0(A_i)$ a $\operatorname{var} Q(A_i) = Q_0(A_i)(1 - Q_0(A_i))/(n_0 + 1)$, rozdělení procesu je tedy soustředěno kolem pravděpodobnostní míry Q_0, zatímco n_0 udává stupeň koncentrace. Dirichletův proces pokrývá ve smyslu nosiče širokou třídu rozdělení na $(\mathcal{X}, \mathcal{A})$, na druhou stranu Q je s pravděpodobností 1 diskrétním rozdělením. Dirichletovým procesem je např. $Q(A) = \sum p_n \delta_{Y_n}(A)$ (δ_x značí Diracovu míru soustředěnou v bodě x), kde body $Y_1, Y_2, \dots \in \mathcal{X}$, na nichž je hodnota Q soustředěna, se generují jako náhodný výběr z rozdělení Q_0, a to nezávisle na příslušných pravděpodobnostech $p_n = \theta_n \prod_{j<n}(1 - \theta_j)$, $n \in \mathbf{N}$, kde $\theta_1, \theta_2, \dots$ jsou nezávislé s beta rozdělením $\mathrm{B}(1, n_0)$ [9]. [2] ukazuje jinou volbu bodů a skoků.

Mnohem širší třídou apriorních rozdělení pro rozdělení na $\mathcal{X} = \mathbf{R}$ jsou procesy *neutrální zprava* [1]. O zprava neutrálním procesu hovoříme, když normalizované přírůstky distribuční funkce $F(t) = Q(-\infty, t\rangle$

$$F(t_1), \quad \frac{F(t_2) - F(t_1)}{1 - F(t_1)}, \quad \dots \quad , \frac{F(t_n) - F(t_{n-1})}{1 - F(t_{n-1})}$$

jsou nezávislé pro libovolná $t_1 < t_2 < \dots < t_n$. Totéž můžeme vyjádřit také pomocí příslušné "kumulativní intenzity" $\Lambda(t) = -\ln(1 - F(t))$.

Definice 2.2. *Řekneme, že Q, resp. Λ je zprava neutrální proces, jestliže Λ je neklesající, zprava spojitý proces s nezávislými přírůstky a $\Lambda(-\infty) = 0$, $\Lambda(\infty) = \infty$.*

Nemá-li proces Λ nenáhodnou složku, s pravděpodobností 1 intenzita Λ opět přísluší diskrétnímu rozdělení. Tak jak intenzita Λ generuje na $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$ míru, značíme stručně např. $\Lambda(s, t) = \Lambda(t-) - \Lambda(s)$, $\Lambda\{t\} = \Lambda(t) - \Lambda(t-)$, apod. Speciálním případem zprava neutrálního procesu je gama proces.

Definice 2.3. *Mají-li přírůstky zprava neutrálního procesu Λ gama rozdělení, $\Lambda(s, t) \sim \mathrm{G}(n_0, n_0 \Lambda_0(s, t))$, kde $n_0 > 0$ a Λ_0 je nějaká kumulativní intenzita, nazveme ho gama procesem a značíme $\Lambda \sim \mathcal{G}(n_0, \Lambda_0)$.*

Parametry n_0 a Λ_0 mají podobný význam jako parametry Dirichletova procesu, je $\mathrm{E}\,\Lambda(s,t) = \Lambda_0(s,t)$, $\mathrm{var}\,\Lambda(s,t) = \Lambda_0(s,t)/n_0$. Rozdělením $\mathrm{G}(n_0, 0)$ rozumíme rozdělení degenerované v 0. Podobně jako Dirichletův proces, i gama proces může být definován na obecném prostoru, nejen na $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$.

Podívejme se nyní, jak vypadají neparametrické bayesovské odhady odpovídající uvedeným apriorním procesům. Jako pozorovaná data uvažujme jednak úplný náhodný výběr $X_1, \ldots, X_n$ z rozdělení Q a jednak výběr s cenzorováním. Připouštíme tedy, že i-tý čas X_i může být zprava cenzorován veličinou Y_i na něm nezávislou. Skutečně pozorované hodnoty pak tvoří náhodný výběr dvojic

$$\text{data} = (Z_1, \delta_1), \ldots, (Z_n, \delta_n), \quad \text{kde } Z_i = \min(X_i, Y_i) \text{ a } \delta_i = I_{[X_i \leqq Y_i]} \tag{1}$$

jsou pozorovaný čas (skutečné pozorování X_i nebo čas cenzorování) a indikátor cenzorování. Značme jako $N_t = \#\{i, X_i > t\}$, resp. $N_t = \#\{i, Z_i > t\}$, počet pozorování překračujících t.

Je-li apriorním rozdělením Q Dirichletův proces $\mathcal{D}(\alpha)$, $\alpha = n_0 Q_0$, pak aposteriorní rozdělením $(Q \mid X_1, \ldots, X_n)$ je opět Dirichletův proces, $\mathcal{D}(\alpha + \sum \delta_{X_i})$. V případě reálných pozorování tak např. při kvadratické ztrátové funkci máme jako bayesovské odhady distribuční funkce $F(x) = Q(-\infty, x\rangle$, resp. střední hodnoty rozdělení Q, příslušné aposteriorní střední hodnoty

$$\widehat{F}(x) = \frac{(\alpha + \sum \delta_{X_i})(-\infty, x\rangle}{n_0 + n} = \frac{n_0 F_0(x) + n F_n(x)}{n_0 + n}, \quad \widehat{\mathrm{E}\,Q} = \frac{n_0\,\mathrm{E}\,Q_0 + n\bar{X}}{n_0 + n},$$

kde F_0 je distribuční funkce příslušná Q_0 a F_n je empirická distribuční funkce pozorování. V případě cenzorovaných pozorování (1) už konjugovanost systému Dirichletových měr neplatí, aposteriorní rozdělení spadá mezi beta-Stacyovy procesy [12]. Odhad funkce spolehlivosti $S = 1 - F$ je ale dobře znám a má tvar [10]

$$\widehat{S}(t) = \frac{n_0 S_0(t) + N_t}{n_0 + n} \prod_{x \in M_t^0} \frac{n_0 S_0(x-) + N(x-) - u(x)}{n_0 S_0(x-) + N_x}$$

kde $S_0(t) = 1 - F_0(t)$, $u(t)$ je počet necenzorovaných pozorování v okamžiku t a M_t^0 jsou okamžiky s (alespoň jedním) cenzorovaným pozorováním, $M_t^j = \{x \leqq t, \exists_i\, Z_i = x, \delta_i = j\}$, $j = 0, 1$. Dokonce rozdělení Y_i se pro různá i mohou lišit a může jít také o degenerované náhodné veličiny, cenzorování časem. Při zmenšující se míře počáteční informace $n_0 \to 0$ je $\widehat{F} \to F_n$ a $\widehat{\mathrm{E}\,Q} \to \bar{X}$ a podobně v případě cenzorování dostaneme v limitě neparametrický Kaplanův-Meierův odhad, $\widehat{S}(t) \to \prod_{x \in M_t^1}(1 - u(x)/N(x-))$.

Podobná "konjugovanost" platí i pro zprava neutrální procesy, ve třídě zprava neutrálních procesů (při daných hodnotách cenzorů) zůstaneme i v případě cenzorování, viz [3]. Zajímavou vlastností je, že i když proces kumulativní intenzity Λ nemá za apriorního rozdělení skoky v pevných bodech, v aposteriorním rozdělení se takové skoky v okamžicích pozorování vždy objeví. Ve třídě zprava neutrálních procesů zůstaneme dokonce i při jiném cenzorování, např. $X_i \geqq Y_i$. Zápis odhadů je v obecné šíři komplikovaný, uveďme

aspoň speciální případ gama procesu. Je-li apriorně $\Lambda \sim \mathcal{G}(n_0, \Lambda_0)$, kde Λ_0 je spojitá, a jsou-li pozorování navzájem různá, pak

$$\widehat{F}(t) = 1 - \Big(\frac{n_0 + N_t}{n_0 + N_t + 1}\Big)^{n_0\Lambda_0(t)} \prod_{i, X_i \leqq t} \Big(\frac{n_i + 1}{n_i}\Big)^{n_0\Lambda_0(t)} \frac{\ln((n_i+2)/(n_i+1))}{\ln((n_i+1)/n_i)},$$

kde $n_i = n_0 + N_{X_i}$. Aposteriorní rozdělení Λ ale gama procesem není. V případě cenzorování (1) má (při Z_i navzájem různých a spojité Λ_0) odhad opět stejný tvar, součin nyní probíhá pouze přes necenzorovaná pozorování.

Existuje a používá se řada dalších apriorních rozdělení, např. směsi Dirichletových či gama procesů nebo naopak Dirichletův proces, jehož parametr je směsí rozdělení. Jiné apriorní procesy mohou vést, na rozdíl od popsaných, ke spojitým rozdělením, nebo přímo modelovat náhodné hustoty či (nekumulativní) intenzity poruch. V určitých situacích může být vhodnější zabývat se místo kumulativní intenzity Λ její variantou $\Lambda^*(x) = \int_{\langle 0,x\rangle}(dF(t)/S(t-))$, $x > 0$, viz [5].

3 Odhady v Koziolově-Greenově modelu

V Koziolově-Greenově modelu se u pozorování (1) předpokládá, že distribuční funkce F dob života X_i je s distribuční funkcí F_Y časů cenzorování Y_i svázána podmínkou $1-F_Y = (1-F)^\gamma$ s nějakou konstantou $\gamma > 0$, tzn. že kumulativní intenzity jsou si úměrné, $\Lambda_Y = \gamma\Lambda$. Rozdělení X_i a Y_i tak mají společný parametr a i v pozorování samotné cenzorující veličiny ($Y_i = t$ nebo $Y_i \geqq t$) je obsažena informace o kumulativní intenzitě Λ. S tím uvedené příklady odhadů nepočítají. Podobný problém řeší v modelu konkurujících si rizik [8] zavedením "vícerozměrného" Dirichletova procesu.

Budeme předpokládat, že "parametry" Λ a γ jsou při apriorním rozdělení nezávislé, Λ je gama proces $\mathcal{G}(n_0, \Lambda_0)$ a rozdělení γ má hustotu $\pi(\gamma)$, $\gamma > 0$. Pro jednoduchost zápisu budeme dále předpokládat, že Λ_0 je kumulativní intenzita spojitého rozdělení, že pozorování Z_i jsou navzájem různá a již uspořádaná vzestupně. Označme ještě dále jako $\mathrm{G}_0(a, b)$, $a > 0, b > 0$ rozdělení na $\mathbf{R}^+$ s hustotou a momentovou vytvořující funkcí

$$f(x) = \frac{1}{x}\frac{\mathrm{e}^{-ax} - \mathrm{e}^{-bx}}{\ln(a/b)},\ x > 0, \quad M(\theta) = \frac{\ln((a-\theta)/(b-\theta))}{\ln(a/b)},\ \theta < \min(a, b),$$

které vyplyne jako aposteriorní rozdělení skoků u Λ, a nakonec

$$C_j(\gamma) = \Big(\frac{n_0}{n_0 + N_{Z_{j-1}}(1+\gamma)}\Big)^{n_0\Lambda_0(Z_{j-1}, Z_j)}, \tag{2}$$

$$D_j(\gamma) = \begin{cases} -\ln \frac{n_0 + N_{Z_j}(1+\gamma)}{n_0 + N_{Z_j}(1+\gamma)+1}, & \delta_j = 1 \\ -\ln \frac{n_0 + N_{Z_j}(1+\gamma)+1}{n_0 + N_{Z_j}(1+\gamma)+1+\gamma}, & \delta_j = 0, \end{cases}$$

kde N_t stále značí počet pozorování Z_j překračujících t, za našich speciálních předpokladů tedy $N_{Z_j} = n - j$.

Nyní můžeme zformulovat tvrzení o aposteriorním rozdělení parametrů. Přestože při apriorním rozdělení proces Λ nemá skoky v pevně daných bodech, v aposteriorním rozdělení takové skoky jsou, a to v bodech pozorování. Skutečnost, že rozdělení těchto skoků nezávisí (v našem případě) na konkrétních časech pozorování je dána homogenitou gama procesu jakožto Lévyova procesu (viz [3]).

Tvrzení 3.1. *Je-li Λ apriorně rozdělena jako gama proces $\mathcal{G}(n_0, n_0\Lambda_0)$, kde Λ_0 je absolutně spojitá funkce, a γ má apriorní hustotu $\pi(\gamma)$, $\gamma > 0$, a je nezávislé s Λ, pak při daných pozorováních* (1) *s uspořádáním $Z_1 < \cdots < Z_n$ pro aposteriorní rozdělení platí*

- $(\Lambda \mid \text{data}, \gamma)$ *je rozdělení procesu s nezávislými přírůstky, přičemž pro $(s,t) \subset (Z_i, Z_{i+1})$, $i = 0, \ldots, n$ (volíme $Z_0 = 0$, $Z_{n+1} = \infty$) je*

$$(\Lambda(s,t) \mid \text{data}, \gamma) \sim \mathrm{G}(n_0 + N_s(1+\gamma), n_0\Lambda_0(s,t))$$

  a pro $t = Z_i$ má Λ v t skok s rozdělením

$$(\Lambda\{Z_i\} \mid \text{data}, \gamma) \sim \begin{cases} \mathrm{G}_0(N_{Z_i}(1+\gamma), N_{Z_i}(1+\gamma)+1), & \delta_i = 1, \\ \mathrm{G}_0(N_{Z_i}(1+\gamma)+1, N_{Z_i}(1+\gamma)+1+\gamma), & \delta_i = 0, \end{cases}$$

- $(\gamma \mid \text{data})$ *má rozdělení s hustotou*

$$\pi(\gamma \mid \text{data}) \propto \Big(\prod_{j=1}^{n} C_j(\gamma)\, D_j(\gamma)\Big)\pi(\gamma), \quad \gamma > 0.$$

Důkaz. Aposteriorní rozdělení získáme postupnými aktualizacemi po jednotlivých pozorováních (Z, δ). Každé takové pozorování zachycuje buď dvojici $X = t, Y \geqq t$, když $Z = t, \delta = 1$, nebo $X > t, Y = t$, když $Z = t, \delta = 0$. Naznačíme důkaz aktualizace po prvním pozorování pro případ s $\delta = 0$, tj. výpočet aposteriorního rozdělení po pozorování dvojice $X > t, Y = t$.

Zvolme libovolně dělení $0 = t_0 < t_1 < \cdots < t_k < \infty$ s dělícími body různými od t a nechť i je ten index, pro nějž $t \in (t_{i-1}, t_i)$. Postupujeme ve dvou krocích. Nejprve určíme aposteriorní rozdělení veličin γ a $\lambda = (\lambda_1, \ldots, \lambda_k)$, kde $\lambda_j = \Lambda(t_{j-1}, t_j)$, $j = 1, \ldots, k$, při daném $Y = t$. Jejich momentová vytvořující funkce má v bodě $(\theta_0, \theta_1, \ldots, \theta_k)$ hodnotu

$$M(t) = \mathrm{E}(U \mid Y = t) = \int_0^\infty \mathrm{e}^{\gamma\theta_0} M_{(\lambda|\gamma, Y=t)}(\theta_1, \ldots, \theta_k)\pi(\gamma \mid Y = t)\,\mathrm{d}\gamma, \quad (3)$$

$U = \mathrm{e}^{\gamma\theta_0}\mathrm{e}^{\sum \lambda_j\theta_j}$. Spočteme $I(x) = \int_x^\infty M(t) f_Y(t)\,\mathrm{d}t$, $x \in (t_{i-1}, t_i)$, kde f_Y značí nepodmíněnou hustotu Y, a odtud pak $M(t) = (-1/f_Y(t))(\,\mathrm{d}I(t)/\,\mathrm{d}t)$.

Z definice podmíněné střední hodnoty máme

$$\begin{aligned} I(x) &= \int_{Y>x} \mathrm{E}(U \mid Y)\,\mathrm{d}\,\mathrm{P} = \int_{Y>x} U\ \,\mathrm{d}\,\mathrm{P} = \mathrm{E}(U \cdot \mathrm{P}(Y > x \mid \Lambda, \gamma)) \\ &= \mathrm{E}(U\mathrm{e}^{-\gamma\Lambda(x)}) = \mathrm{E}\Big(\mathrm{e}^{\gamma\theta_0} \prod_{j<i} \mathrm{e}^{-\lambda_j(\gamma-\theta_j)} \cdot \mathrm{e}^{-\lambda_{i1}(\gamma-\theta_i)}\mathrm{e}^{\lambda_{i2}\theta_i} \cdot \prod_{j>i} \mathrm{e}^{\lambda_j\theta_j}\Big) \\ &= \int_0^\infty \mathrm{e}^{\gamma\theta_0} \prod_{j<i} \Big(\frac{n_0}{n_0+\gamma-\theta_j}\Big)^{n_0\lambda_j^0} \cdot \Big(\frac{n_0}{n_0+\gamma-\theta_i}\Big)^{n_0\lambda_{i1}^0(x)} \\ &\qquad \cdot \Big(\frac{n_0}{n_0-\theta_i}\Big)^{n_0\lambda_{i2}^0(x)} \prod_{j>i} \Big(\frac{n_0}{n_0-\theta_j}\Big)^{n_0\lambda_j^0} \pi(\gamma)\,\mathrm{d}\gamma \end{aligned}$$

při značení $\lambda_j^0 = \mathrm{E}\,\lambda_j = \Lambda_0(t_{j-1}, t_j)$ a podobně $\lambda_{i1}^0(x) = \Lambda_0(t_{i-1}, x)$, $\lambda_{i2}^0(x) = \Lambda_0(x, t_i)$. Ve výsledném výrazu pro $M(t)$ (po zmíněném zderivování) si při srovnání s (3) přečteme hustotu $\pi(\gamma \mid Y = t)$ a vytvořující funkci $M_{(\lambda|\gamma,Y=t)}$. Ta se rozpadá na součin vytvořujících funkcí veličin λ_j, veličiny $\lambda_1, \ldots, \lambda_k$ jsou tedy při daném γ (a $Y = t$) nezávislé. Kromě λ_i jde o vytvořující funkce gama rozdělení, veličina $\lambda_i = \Lambda(t_{i-1}, t_i)$ má vytvořující funkci jako součet nezávislých veličin s rozděleními $\mathrm{G}(n_0 + \gamma, n_0\Lambda_0(t_{i-1}, t))$, $\mathrm{G}_0(n_0, n_0 + \gamma)$ a $\mathrm{G}(n_0, n_0\Lambda_0(t, t_i))$. Ve skutečnosti popisuje tato trojice rozdělení veličin $\Lambda(t_{i-1}, t)$, $\Lambda\{t\}$ a $\Lambda(t, t_i)$.

Nakonec přidáme ještě pozorování $X > t$, aktualizovaná aposteriorní hustota např. pro $\lambda = (\lambda_j, j \neq i, \lambda_{i1}, \lambda_{\{t\}} = \Lambda(\{t\}), \lambda_{i2})$ a γ bude

$$\begin{aligned} \pi(\lambda, \gamma \mid X > t, Y = t) &\propto \mathrm{P}(X > t \mid \lambda, \gamma, Y = t)\pi(\lambda, \gamma \mid Y = t) \\ &= \mathrm{P}(X > t \mid \lambda, \gamma)\pi(\lambda, \gamma \mid Y = t) = \mathrm{e}^{-(\sum_{j<i}\lambda_j) - \lambda_{i1} - \lambda_{\{t\}}}\pi(\lambda, \gamma \mid Y = t). \end{aligned}$$

V případě s $\delta = 1$ bychom stejným postupem nejprve určili rozdělení parametrů při $X = t$ a následně přidali $Y \geqq t$. Podobně bychom aktualizovali aposteriorní rozdělení při dalších pozorováních, do dělení bychom zařadili také všechny časy předchozích pozorování. □

V důkazu jsme využili, jak radí [6], konkrétního tvaru apriorního rozdělení. Obecnější důkazovou techniku, založenou na Lévyově míře neklesajícího procesu Λ, nabízí [1]. Ke správné aposteriorní hustotě ale vede i jednoduchý intuitivní přístup — zkombinovat apriorní hustotu parametrů s věrohodnostní funkcí danou pozorováními. Do apriorní hustoty ovšem musíme pro (apriorně) nulové veličiny $\lambda_{\{j\}} = \Lambda\{t\}$, $t = Z_j$, formálně zapsat "hustotu" $\lambda_{\{j\}}^{-1}\mathrm{e}^{-n_0\lambda_{\{j\}}}$, $\lambda_{\{j\}} > 0$, jakožto hustotu "gama" rozdělení $\mathrm{G}(n_0, 0)$. Máme tedy

$$\pi(\lambda) \propto \prod \lambda_j^{\lambda_j^0 - 1}\mathrm{e}^{-n_0\lambda_j} \cdot \lambda_{\{j\}}^{-1}\mathrm{e}^{-n_0\lambda_{\{j\}}}$$

a do věrohodnostní funkce s každým pozorováním přidáváme člen typu

$$\mathrm{e}^{-\sum_{j\leqq i}\lambda_j(1+\gamma)}\mathrm{e}^{-\sum_{j<i}\lambda_{\{j\}}(1+\gamma)}p_i$$

pro pozorování v čase t_i, kde $p_i = 1 - \mathrm{e}^{-\lambda_{\{i\}}}$ u pozorování necenzorovaného a $p_i = \mathrm{e}^{-\lambda_{\{i\}}}(1 - \mathrm{e}^{-\lambda_{\{i\}}\gamma})$ u cenzorovaného. Při daném γ rozeznáme pak hustotu λ a snadno zjistíme i normovací konstanty $C_j(\gamma)$ a $D_j(\gamma)$.

Uveďme konečně tvar bayesovského odhadu funkce spolehlivosti pro dobu života při kvadratické ztrátové funkci, tj. aposteriorní střední hodnotu veličiny $\exp(-\Lambda(t))$.

Tvrzení 3.2. *Za předpokladů předchozího tvrzení máme pro funkci spolehlivosti S bayesovský odhad*

$$\widehat{S}(t) = \frac{\displaystyle\int \Big(\prod_{j<i} C_j^+(\gamma) D_j^+(\gamma)\Big) C_{i1}^+(\gamma) C_{i2}(\gamma) D_i(\gamma) \Big(\prod_{j>i} C_j(\gamma) D_j(\gamma)\Big) \pi(\gamma)\,\mathrm{d}\gamma}{\displaystyle\int \Big(\prod_{j\neq i} C_j(\gamma)\Big) C_{i1}(\gamma) C_{i2}(\gamma) \Big(\prod_{j} D_j(\gamma)\Big) \pi(\gamma)\,\mathrm{d}\gamma}$$

$t \in (Z_{i-1}, Z_i)$, kde $C_{i1}(\gamma)$, resp. $C_{i2}(\gamma)$ jsou jako $C_i(\gamma)$ v (2)*, ale s exponenty $n_0\Lambda_0(Z_{i-1}, t)$, resp. $n_0\Lambda_0(t, Z_i)$ a C^+, resp. D^+ jsou jako C a D, ale s $N_t(1+\gamma)+1$ místo $N_t(1+\gamma)$.*

Důkaz. Použijeme značení jako v důkazu předchozího tvrzení, ale s dělením $0 = Z_0 < Z_1 < \cdots < Z_n$ a s $t \in (Z_{i-1}, Z_i)$. Počítáme

$$\widehat{S}(t) = \mathrm{E}\,\mathrm{e}^{-\Lambda(t)} = \mathrm{E}\Big[\Big(\prod_{j<i} \mathrm{E}(\mathrm{e}^{-\lambda_j} \mid \gamma)\,\mathrm{E}(\mathrm{e}^{-\lambda_{\{j\}}} \mid \gamma)\Big)\,\mathrm{E}(\mathrm{e}^{-\lambda_{i1}} \mid \gamma)\Big],$$

kde všechny střední hodnoty rozumíme navíc jako podmíněné pozorováními (1), tj. při aposteriorním rozdělení. Ve výrazu uvedené střední hodnoty $\mathrm{E}(.|\gamma)$ jsou postupně $C_j^+(\gamma)/C_j(\gamma)$, $D_j^+(\gamma)/\,\mathrm{D}_j(\gamma)$ a $C_{i1}^+(\gamma)/C_{i1}(\gamma)$. □

Omezující předpoklady jsme zavedli jen z důvodu přehlednosti, v jejich uvolnění nám nic nebrání. Pokud např. připustíme skoky v "apriorním odhadu" Λ_0, pak v případě shody některého pozorování s bodem skoku se jako aposteriorní rozdělení skoku v takovém bodě místo rozdělení G_0 objeví rozdělení s hustotou úměrnou rozdílu gama hustot a místo "normovací konstanty" $D(\gamma)$ rozdíl "konstant" $C(\gamma)$. Podobně, lze dospět k aposteriorním rozdělením i v případě shod časů některých pozorování. V případě dvou cenzorovaných pozorování ve stejném čase $t = t_i$ to znamená v příslušném dělícím bodě člen $\mathrm{e}^{-2\lambda_{\{i\}}}(1 - \mathrm{e}^{-\lambda_{\{i\}}\gamma})^2$, když byla obě cenzorovaná, nebo $\mathrm{e}^{-\lambda_{\{i\}}}(1 - \mathrm{e}^{-\lambda_{\{i\}}\gamma})(1 - \mathrm{e}^{-\lambda_{\{i\}}})$, když jedno bylo cenzorované a jedno nikoliv. Tomu pak odpovídají i normovací konstanty, místo $D_i(\gamma)$ dostaneme $\ln \frac{(n_0+2)(n_0+2+2\gamma)}{(n_0+2+\gamma)^2}$, resp. $\ln \frac{(n_0+1)(n_0+2+\gamma)}{(n_0+1+\gamma)(n_0+2)}$. Podobně můžeme uvážit i jiné typy cenzorovaných pozorování jako jsou $X = Y = t$ nebo $X \geqq Y, Y = t$.

Podobný postup jako pro gama proces je možné uplatnit také pro jiná apriorní rozdělení. Např. v případě apriorního Dirichletova procesu $1-\mathrm{e}^{-\Lambda} \sim \mathcal{D}(\alpha)$ v normovacích konstantách vyjdou rozdíly digama funkcí.

Reference

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

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

[3] Ferguson T.S., Phadia E.G. (1979). *Bayesian nonparametric estimation based on censored data*. Ann. Statist. **7**, No. 1, 163 – 186.

[4] Ferguson T.S., Phadia E.G., Tiwari R.C. (1992). *Bayesian nonparametric inference*. In Current issues in statistical inference: essays in honor of D. Basu (Ghosh M., Pathak P.K., eds.), IMS Lecture Notes Monogr. Ser. **17**, Inst. Math. Statist., Hayward, 127 – 150.

[5] Hjort, N.L. (1990). *Nonparametric Bayes estimators based on beta processes in models for life history data*, Ann. Statist. **18**, No. 3, 1259 – 1294.

[6] Kalbfleisch, J.D. (1978). *Non-parametric Bayesian analysis of survival time data*, J. Roy. Statist. Soc. Ser. B **40**, No. 2, 214 – 221.

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

[8] Salinas-Torres V.H., Pereira C.A.B., Tiwari R.C. (2002). *Bayesian nonparametric estimation in a series system or a competing-risks model*. J. Nonparametr. Stat. **14**, No. 4, 449 – 458.

[9] Sethuraman J. (1994). *A constructive definition of Dirichlet priors*. Statist. Sinica **4**, No. 2, 639 – 650.

[10] Susarla V., Van Ryzin J. (1976). *Nonparametric Bayesian estimation of survival curves from incomplete observations*, J. Amer. Statist. Assoc. **71**, No. 356, 897 – 902.

[11] Walker S.G., Damien P., Laud P.W., Smith A.F.M. (1999). *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**, No. 3, 485 – 527.

[12] Walker S., Muliere P. (1997). *Beta-Stacy processes and a generalization of the Pólya-urn scheme*. Ann. Statist. **25**, No. 4, 1762 – 1780.

Poděkování: Tato práce vznikla za podpory výzkumného záměru MSM 235200001.

Adresa: Katedra matematiky, Fakulta aplikovaných věd, Západočeská univerzita v Plzni, Univerzitní 22, 306 14 Plzeň

E-mail: `friesl@kma.zcu.cz`