Intervalově cenzorovaná data

Iveta Selingerová

Ústav matematiky a statistiky
Přírodovědecká fakulta
Masarykova univerzita

Finanční matematika v praxi II
11.-13. září 2012

Obsah

Úvod

Typy cenzorování:

- Cenzorování zprava
- Cenzorování zleva
- Intervalové cenzorování

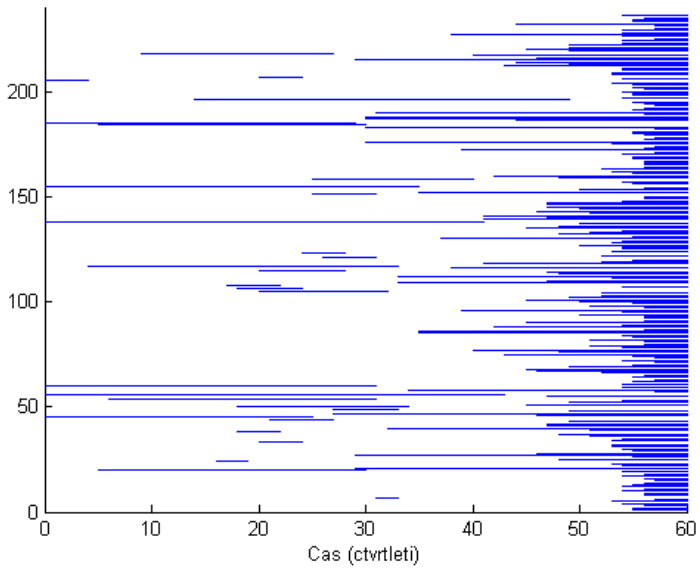
Intervalově cenzorovaná data

$T \in (L, R]$

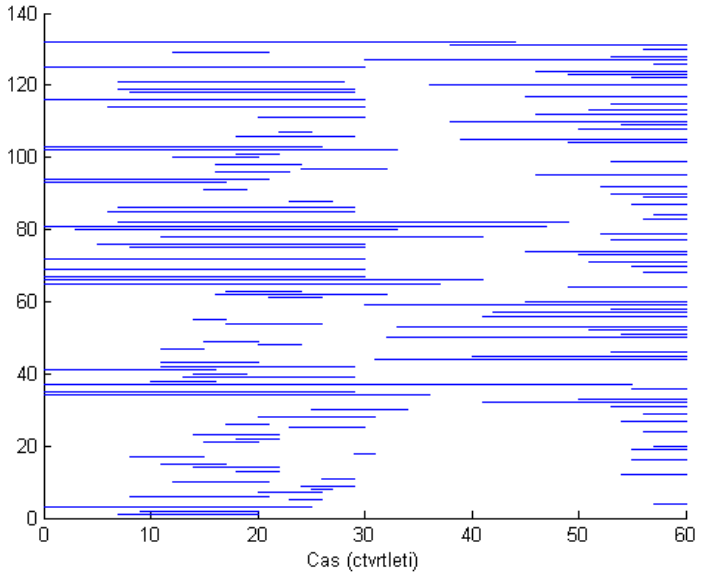
- Current status data, $\{C, \delta\}$, kde $\delta = I(T \leq C)$
- Obecně intervalově cenzorovaná data, $\{U, V, \delta_1 = I(T \leq U), \delta_2 = I(U < T \leq V), \delta_3 = 1 - \delta_1 - \delta_2\}$
- Dvojitě cenzorovaná data

HIV-1 infekce

Studie 368 pacientů s hemofilií z 80. let



Obrázek: Bez faktoru VIII, 236 pacientů



Obrázek: S nízkou dávkou faktoru VIII, 132 pacientů

NPMLE funkce přežití

Čas přežití ... T_i (funkce přežití $S(t)$)
Pozorujeme

$$O = \{(L_i, R_i] \; ; \; i = 1, \ldots, n\}$$

Označení:
$s_j, \; j = 0, \ldots, m$ ... různé seřazené prvky množiny
$\{0, L_i, R_i; i = 1, \ldots, n\}$
$\alpha_{ij} = I(s_j \in (L_i, R_i]), \;\; i = 1, \ldots, n; j = 1, \ldots, m$
$p_j = S(s_{j-1}) - S(s_j), \;\; j = 1, \ldots, m$
Věrohodnostní funkce

$$L_S(\mathbf{p}) = \prod_{i=1}^{n} [S(L_i) - S(R_i)] = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} p_j,$$

kde $\mathbf{p} = (p_1, \ldots, p_m)'$.

NPMLE funkce přežití

NPMLE funkce přežití S určíme maximalizací $L_S(\mathbf{p})$ vzhledem k $\mathbf{p}$ s omezením

$$\sum_{j=1}^{m} p_j = 1$$

$$p_j \geq 0 \quad (j = 1, \ldots, m)$$

Definujme

$$d_j(\mathbf{p}) = \sum_{i=1}^{n} \frac{\alpha_{ij}}{\sum_{l=1}^{m} \alpha_{il} p_l}$$

Z Lagrangeova multiplikativního kritéria získáme, že $\hat{\mathbf{p}}$ je NPMLE, jestliže $d_j(\hat{\mathbf{p}}) = n$ pro všechna $j = 1, \ldots, m$.

Turnbullovy intervaly

Množina disjunktních intervalů, jejichž levé koncové body jsou v $L = \{L_1, L_2, \ldots, L_n\}$ a pravé koncové body jsou v $R = \{R_1, R_2, \ldots, R_n\}$, ale neobsahují jiné členy L nebo R kromě koncových bodů.

Turnbullovy intervaly

	s_0	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
1	0		7										
2	0			8									
3		6			10								
4			7				16						
5			7			14							
6								17					∞
7									37	44			
8											45		∞
9												46	∞
10												46	∞
		(6	7]	(7	8]				(37	44]		(46	∞]

Algoritmy pro NPMLE

- Turnbullův algoritmus

  E-krok Výpočet střední hodnoty logaritmické věrohodnostní funkce podmíněné pozorovanými daty a odhadem $\hat{\mathbf{p}}$ z předchozí iterace

  M-krok Maximalizace Lagrangeovou metodou přes

$$\mathbf{C_p} = \left\{ \mathbf{p} \in [0,1]^m; \sum_{j=1}^{m} p_j = 1, p_j \geq 0 \right\}$$

Algoritmy pro NPMLE

- Turnbullův algoritmus

  1.krok Zvolíme počáteční odhad $\hat{\mathbf{p}}^0$.

  2.krok V l-té iteraci definujeme odhad $\hat{\mathbf{p}}$ jako

$$\hat{p}_j^{(l)} = \frac{d_j(\hat{\mathbf{p}})^{(l-1)}\hat{p}_j^{(l-1)}}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_{ij}\hat{p}_j^{(l-1)}}{\sum_{k=1}^{m}\alpha_{ik}\hat{p}_k^{(l-1)}}$$

  3.krok Opakujeme krok 2 dokud nedosáhneme požadované přesnosti.

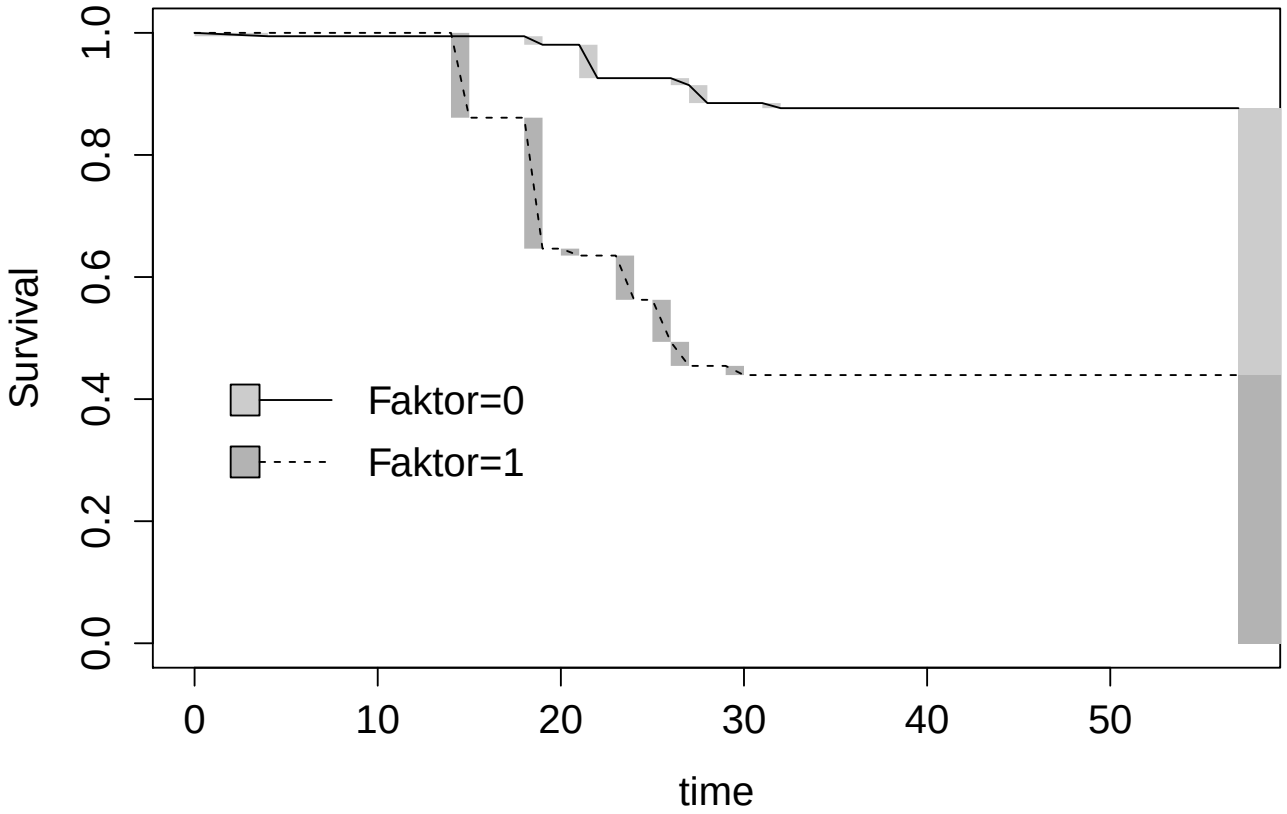
$$\sum_{j=1}^{m-1}|\hat{p}_j^{(l)} - \hat{p}_j^{(l-1)}| < \varepsilon$$

Algoritmy pro NPMLE

- ICM algoritmus

  Transformace maximalizace věrohodnostní funkce na maximalizaci kvadratické funkce.
- EM-ICM algoritmus

Odhad funkce přežití pro HIV-1 infekci



Srovnání algoritmů (přesnost 10^{-8})

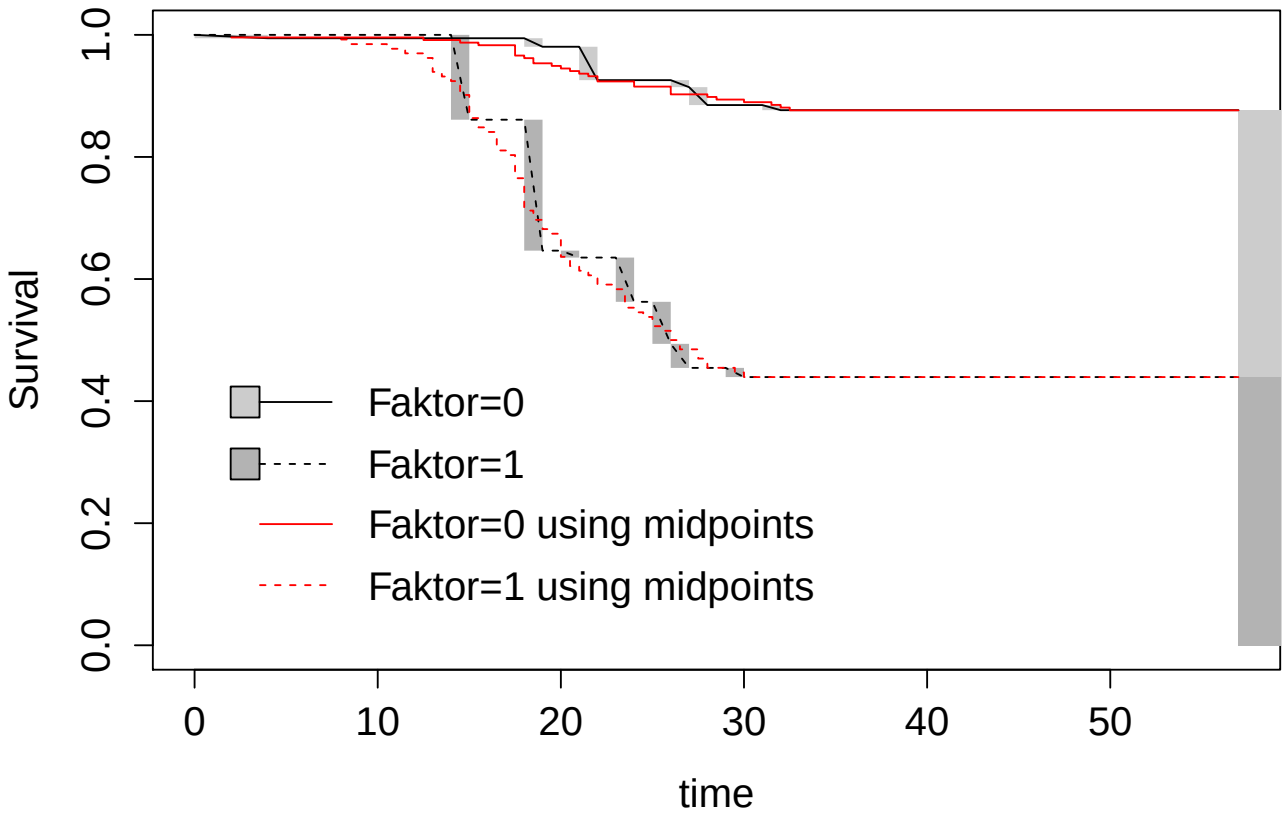
	Turnbullův alg.	ICM alg.	EM-ICM
počet iterací	339	227	11
čas (s)	0,27	0,56	0,25

Tabulka: Bez faktoru VIII

	Turnbullův alg.	ICM alg.	EM-ICM
počet iterací	2602	308	14
čas (s)	1,02	0,79	0,28

Tabulka: S nízkou dávkou faktoru VIII

Odhad funkce přežití pro HIV-1 infekci



Porovnávání funkcí přežití

Nulová hypotéza: $S_1(x) = S_2(x)\ \forall x$
Alternativní hypotéza: $S_1(x) \neq S_2(x)\ \forall x$

- zobecněný log-rank test
- zobecněný Wilcoxonův test

Test pro data HIV-1 infekce

- zobecněný log-rank test: p-hodnota $2,2 \cdot 10^{-16}$
- zobecněný Wilcoxonův test: p-hodnota $2,2 \cdot 10^{-16}$

Jádrový odhad rizikové funkce

Riziková funkce $\lambda(t)$ - pravděpodobnost, že událost nastane v čase t, za předpokladu, že do času t nenastala

Odhad rizikové funkce v čase s_j:

Pro pravostranně cenzorovaná data $\hat{\lambda}(s_j) = \frac{d_j}{r_j}$

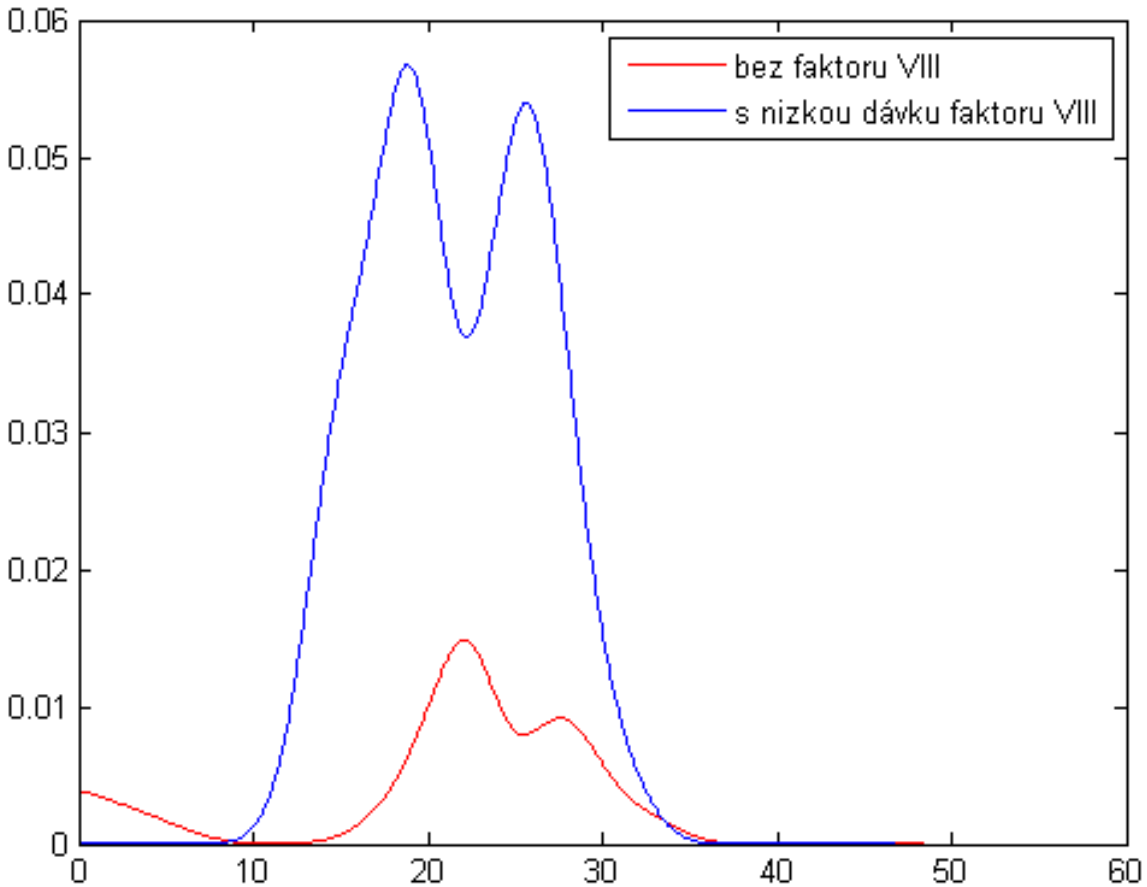
Pro intervalově cenzorovaná data

$$\hat{\lambda}(s_j) = \hat{\lambda}_j = \frac{d_j(\hat{\mathbf{p}})\hat{p}_j}{\sum_{u=j}^{m} d_u(\hat{\mathbf{p}})\hat{p}_u}$$

Jádrový odhad:

$$\hat{\lambda}(t) = \sum_{j=1}^{m} \frac{\frac{1}{h}K\left(\frac{t-s_j}{h}\right)}{\sum_{u=1}^{m}\frac{1}{h}K\left(\frac{t-s_u}{h}\right)}\hat{\lambda}_j$$

Jádrový odhad rizikové funkce pro HIV-1 infekci



Literatura

- J. Sun, *The Statistical Analysis of Interval-censored Failure Time Data*. Springer, 2006.
- S. R. Giolo, *Turnbull's Nonparametric Estimator for Interval-Censored Data*. Federal University of Paraná, 2004.
- M. P. Fay, *Interval Censored Data Analysis*, `http://user2010.org/tutorials/Fay_1.pdf`
- J. P. Klein, M. L. Moeschberger, *SURVIVAL ANALYSIS Techniques for Censored and Truncated Data*. Springer, 2003.