ROBUST 2010
*
Míry a váhy

Zdeněk Fabián
Ústav informatiky AVČR

February 3, 2010

Pravděpodobnostní míra a metrika ve výběrovém prostoru

- $\mathcal{X} \subseteq R$

Pravděpodobnostní míra a metrika ve výběrovém prostoru

- $\mathcal{X} \subseteq R$
- F, f

Pravděpodobnostní míra a metrika ve výběrovém prostoru

- $\mathcal{X} \subseteq R$
- F, f
- skalární skórová funkce S

Pravděpodobnostní míra a metrika ve výběrovém prostoru

- $\mathcal{X} \subseteq R$
- F, f
- skalární skórová funkce S
- diference v $\mathcal{X}$:

$$d(x_2, x_1) = S(x_2) - S(x_1) = \int_{x_1}^{x_2} dw$$

Pravděpodobnostní míra a metrika ve výběrovém prostoru

- $\mathcal{X} \subseteq R$
- F, f
- skalární skórová funkce S
- diference v $\mathcal{X}$:

$$d(x_2, x_1) = S(x_2) - S(x_1) = \int_{x_1}^{x_2} dw$$

- x_k v modelu $F(x; \theta)$: $\quad f(x_k; \theta), S(x_k; \theta), w(x_k; \theta)$

funkce $T_G(y) = -\frac{g'(y)}{g(y)}$

- popis rozdělení na $\mathbb{R}$

funkce $T_G(y) = -\frac{g'(y)}{g(y)}$

- popis rozdělení na $\mathbb{R}$
- $$g(y-\mu): \qquad T_G(y-\mu) = \frac{\partial}{\partial \mu} \log g(y-\mu)$$

funkce $T_G(y) = -\frac{g'(y)}{g(y)}$

- popis rozdělení na $\mathbb{R}$
- $$g(y-\mu): \qquad T_G(y-\mu) = \frac{\partial}{\partial \mu} \log g(y-\mu)$$
- $$T_G(y) = 0$$

funkce $T_G(y) = -\frac{g'(y)}{g(y)}$

- popis rozdělení na $\mathbb{R}$
- $$g(y-\mu): \qquad T_G(y-\mu) = \frac{\partial}{\partial \mu} \log g(y-\mu)$$
- $$T_G(y) = 0$$
- T_G rozdělení s těžkými chvosty omezená, $ET_G^k < \infty$

funkce $T_G(y) = -\frac{g'(y)}{g(y)}$

- popis rozdělení na $\mathbb{R}$
- $$g(y-\mu): \qquad T_G(y-\mu) = \frac{\partial}{\partial \mu} \log g(y-\mu)$$
- $$T_G(y) = 0$$
- T_G rozdělení s těžkými chvosty omezená, $ET_G^k < \infty$
- ale jen pro rozdělení na $\mathbb{R}$

inferenční funkce na $\mathcal{X} \neq \mathbb{R}$

- $f(x)$ na $\mathcal{X}$

inferenční funkce na $\mathcal{X} \neq \mathbb{R}$

- $f(x)$ na $\mathcal{X}$
- jaký má 'střed' ?

inferenční funkce na $\mathcal{X} \neq \mathbb{R}$

- $f(x)$ na $\mathcal{X}$
- jaký má 'střed' ?
- těžiště x^*

inferenční funkce na $\mathcal{X} \neq \mathbb{R}$

- $f(x)$ na $\mathcal{X}$
- jaký má 'střed' ?
- těžiště x^*
- $\eta : \mathcal{X} \to \mathbb{R}$, $Y = \eta(X)$ $\qquad T_F(x) = T_G(\eta(x))$

inferenční funkce na $\mathcal{X} \neq \mathbb{R}$

- $f(x)$ na $\mathcal{X}$
- jaký má 'střed' ?
- těžiště x^*
- $\eta : \mathcal{X} \to \mathbb{R}$, $Y = \eta(X)$ $T_F(x) = T_G(\eta(x))$
- $$T_F(x) = \frac{1}{f(x)} \frac{d}{dx} \left(-\frac{1}{\eta'(x)} f(x) \right)$$

  transformation-based score (t-skór)

inferenční funkce na $\mathcal{X} \neq \mathbb{R}$

- $f(x)$ na $\mathcal{X}$
- jaký má 'střed' ?
- těžiště x^*
- $\eta : \mathcal{X} \to \mathbb{R}$, $Y = \eta(X)$ $\qquad T_F(x) = T_G(\eta(x))$
- $$T_F(x) = \frac{1}{f(x)} \frac{d}{dx} \left(-\frac{1}{\eta'(x)} f(x) \right)$$

  transformation-based score (t-skór)

- $$x^* : T_F(x) = 0$$

Jaké η ?

- $\eta : \mathcal{X} \to \mathbb{R}$

Jaké η ?

- $\eta : \mathcal{X} \to \mathbb{R}$
-

$$\eta(x) = \begin{cases} x & \text{if} \quad \mathcal{X} = \mathbb{R} \\ \log x & \text{if} \quad \mathcal{X} = (0, \infty) \\ \log \frac{x}{1-x} & \text{if} \quad \mathcal{X} = (0, 1) \end{cases}$$

Jaké η ?

- $\eta : \mathcal{X} \to \mathbb{R}$
- $$\eta(x) = \begin{cases} x & \text{if} \quad \mathcal{X} = \mathbb{R} \\ \log x & \text{if} \quad \mathcal{X} = (0, \infty) \\ \log \frac{x}{1-x} & \text{if} \quad \mathcal{X} = (0, 1) \end{cases}$$
- $$T_F(x) = \begin{cases} -\frac{f'(x)}{f(x)} & \mathcal{X} = \mathbb{R} \\ -1 - x\frac{f'(x)}{f(x)} & \mathcal{X} = (0, \infty) \\ -1 + 2x - x(1-x)\frac{f'(x)}{f(x)} & \mathcal{X} = (0, 1) \end{cases}$$

Shoda s klasickou statistikou

- rozdělení na $\mathbb{R}$: př. standard normal $T_G(y) = y$

Shoda s klasickou statistikou

- rozdělení na $\mathbb{R}$: př. standard normal $T_G(y) = y$
- Pro rozdělení třídy τ na $(0, \infty)$: $\tau = \exp(\mu)$

Shoda s klasickou statistikou

- rozdělení na $\mathbb{R}$: př. standard normal $T_G(y) = y$
- Pro rozdělení třídy τ na $(0, \infty)$: $\tau = \exp(\mu)$
- $f(x, \tau) = g(\log \frac{x}{\tau})\frac{1}{x}$ př. Weibull $T_F(x) = c[(\frac{x}{\tau})^c - 1]$

Shoda s klasickou statistikou

- rozdělení na $\mathbb{R}$: př. standard normal $T_G(y) = y$
- Pro rozdělení třídy τ na $(0, \infty)$: $\tau = \exp(\mu)$
- $f(x, \tau) = g(\log \frac{x}{\tau})\frac{1}{x}$ př. Weibull $T_F(x) = c[(\frac{x}{\tau})^c - 1]$
- platí věta

$$\eta'(\tau)\, T_F(x;\tau) = \frac{\partial}{\partial \tau} \log f(x;\tau)$$

Shoda s klasickou statistikou

- rozdělení na $\mathbb{R}$: př. standard normal $T_G(y) = y$
- Pro rozdělení třídy τ na $(0, \infty)$: $\tau = \exp(\mu)$
- $f(x, \tau) = g(\log \frac{x}{\tau})\frac{1}{x}$ př. Weibull $T_F(x) = c[(\frac{x}{\tau})^c - 1]$
- platí věta

$$\eta'(\tau) T_F(x; \tau) = \frac{\partial}{\partial \tau} \log f(x; \tau)$$

- $S_F(x; \tau, \theta_2, ...\theta_m) = \eta'(\tau) T_F(x; \tau, \theta_2, ..., \theta_m)$

  ES_F^2 Fisherova informace pro τ

'Blbá rozdělení'

- beta-prime: $\mathcal{X} = (0, \infty)$

$$f(x) = \frac{1}{B(p,q)} \frac{x^{p-1}}{(x+1)^{p+q}} \qquad T_F(x) = \frac{qx - p}{x + 1}$$

'Blbá rozdělení'

- beta-prime: $\mathcal{X} = (0, \infty)$

$$f(x) = \frac{1}{B(p,q)} \frac{x^{p-1}}{(x+1)^{p+q}} \qquad T_F(x) = \frac{qx - p}{x + 1}$$

-

$$x^* = \frac{p}{q}$$

Obecná skalární inferenční funkce

- parametr τ

Obecná skalární inferenční funkce

- parametr τ
- $S_F(x;\theta) = \eta'(x^*)T_F(x;\theta)$

Obecná skalární inferenční funkce

- parametr τ
- $S_F(x;\theta) = \eta'(x^*)T_F(x;\theta)$
- ES_F^2 Fisherova informace pro těžiště

Obecná skalární inferenční funkce

- parametr τ
- $S_F(x;\theta) = \eta'(x^*)T_F(x;\theta)$
- ES_F^2 Fisherova informace pro těžiště
- pro beta-prime $S_F(x) = \frac{q}{p}\frac{qx-p}{x+1}$

$$ES_F^2 = \frac{q^2}{p^2}\frac{pq}{p+q+1}$$

Charakteristiky dat

- Míra variability: t-variance

$$\omega^2 = \frac{1}{ES_F^2}$$

Charakteristiky dat

- Míra variability: t-variance

$$\omega^2 = \frac{1}{ES_F^2}$$

- 'Střed a poloměr' rozdělení x^*, ω
  'Střed a poloměr' dat $\hat{x}^* = x^*(\hat{\theta}), \hat{\omega} = \omega(\hat{\theta})$
  pro porovnání odhadu pro modely libovolnými parametry

Charakteristiky dat

- Míra variability: t-variance

$$\omega^2 = \frac{1}{ES_F^2}$$

- 'Střed a poloměr' rozdělení x^*, ω
  'Střed a poloměr' dat $\hat{x}^* = x^*(\hat{\theta}), \hat{\omega} = \omega(\hat{\theta})$
  pro porovnání odhadu pro modely libovolnými parametry
- Odhad θ zobecněnou momentovou metodou

$$\frac{1}{n}\sum_{i=1}^{n} T_F^k(x_i;\theta) = \mathrm{E}\, T_F^k(\theta) \qquad k = 1, ..., m$$

Charakteristiky dat

- Míra variability: t-variance

$$\omega^2 = \frac{1}{ES_F^2}$$

- 'Střed a poloměr' rozdělení x^*, ω
  'Střed a poloměr' dat $\hat{x}^* = x^*(\hat{\theta}), \hat{\omega} = \omega(\hat{\theta})$
  pro porovnání odhadu pro modely libovolnými parametry
- Odhad θ zobecněnou momentovou metodou

$$\frac{1}{n}\sum_{i=1}^{n} T_F^k(x_i;\theta) = \mathrm{E}T_F^k(\theta) \qquad k = 1, ..., m$$

- Korelační koeficient $Corr_T(X, Y) = \frac{ET_X T_Y}{\sqrt{ET_X^2 ET_Y^2}}$

Momenty vyššího řádu

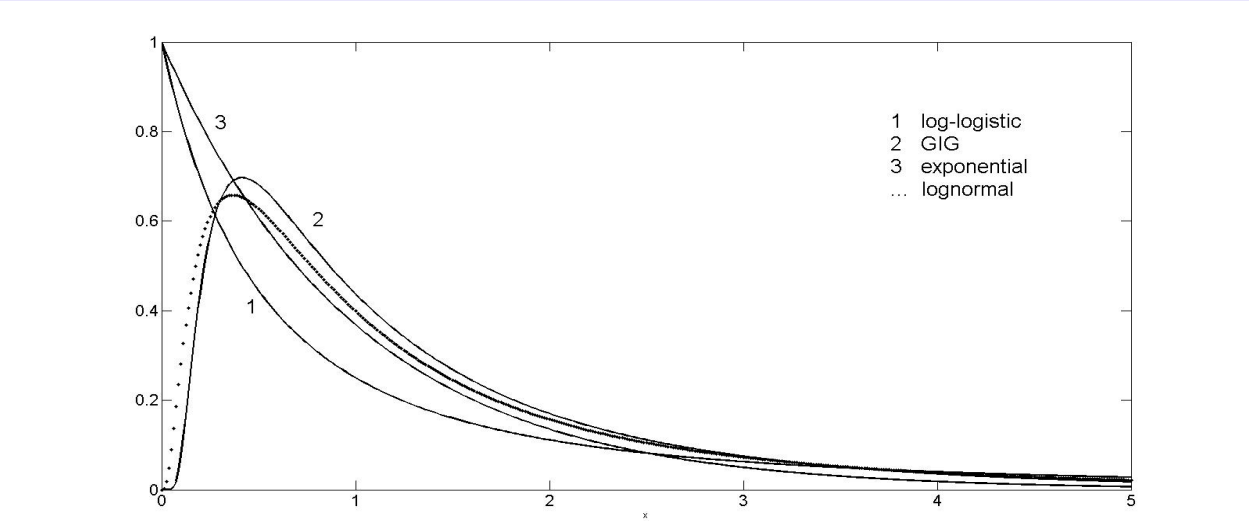
- M_3 šikmost

Momenty vyššího řádu

- M_3 šikmost
- M_4 'plochost' "Pearson's" $\Gamma_2 = \frac{M_4}{(M_2^2)^2}$

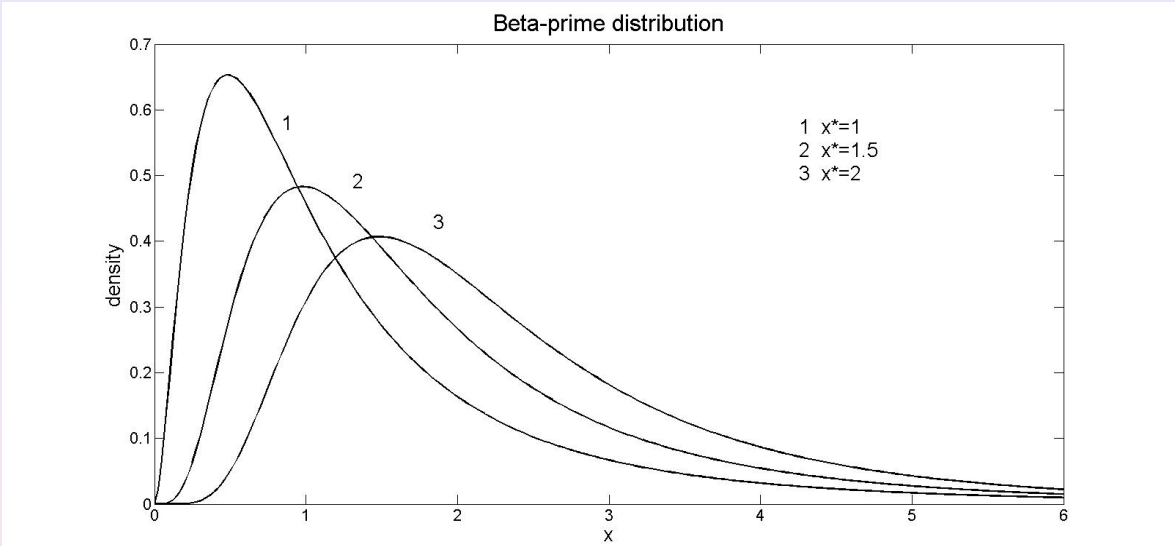
Model	$f(x)$	$S(x)$	M_2	M_4	Γ_2	γ_2
Laplace	$\frac{1}{2}e^{-\lvert x\rvert}$	$\operatorname{sgn} x$	1	1	1	6
Cauchy	$\frac{1}{\pi}\frac{1}{1+x^2}$	$\frac{2x}{1+x^2}$	1/2	3/8	1.5	?
logistic	$\frac{e^x}{(1+e^x)^2}$	$\frac{e^x-1}{e^x+1}$	1/3	0.2	1.8	4.2
normal	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$	x	1	3	3	3
hyperb.	$\frac{1}{K}e^{-\cosh x}$	$\frac{e^x-e^{-x}}{2}$	1.43	11.6	5.66	?
$*$	$\frac{1}{c}e^{-\frac{1}{4}x^4}$	x^3	2.03	45	10.9	1.71

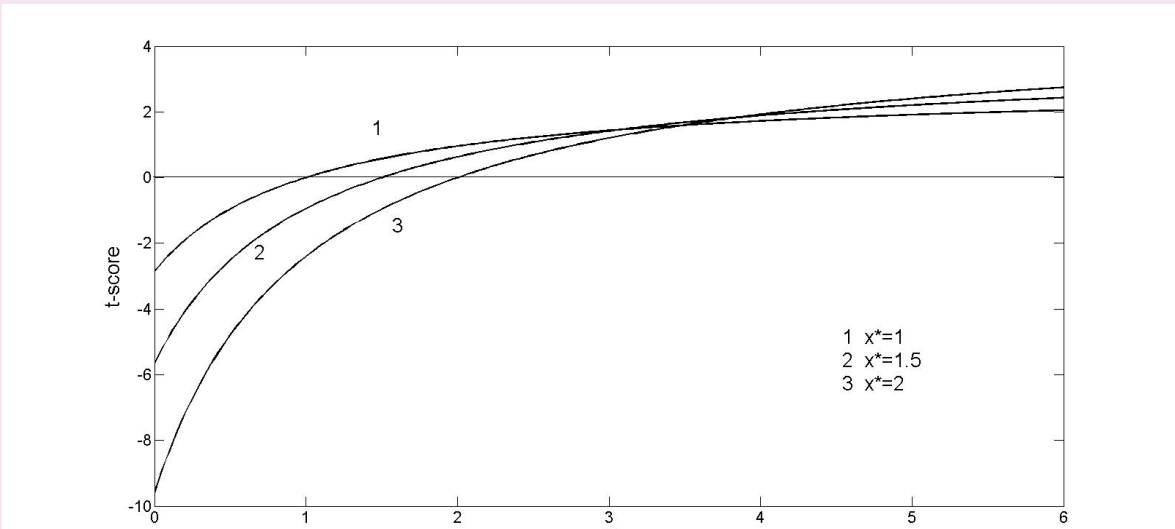
leptokurtic $\Gamma_2 < 3$, platykurtic $\Gamma_2 > 3$



Model	$f(x)$	$S(x)$	M_2	M_3	M_4	Γ_1	Γ_2
log-logistic	$\frac{1}{(1+x)^2}$	$\frac{x-1}{x+1}$	1/3	0	0.2	0	1.8
lognormal	$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\log^2 x}$	$\log x$	1	0	3	0	3
GIG	$\frac{e^{-\frac{1}{2}(x+1/x)}}{2K_0(1)}$	$\frac{x-1/x}{2}$	1.43	0	11.58	0	5.66
exponential	e^{-x}	$x-1$	1	2	9	2	9

Hustoty a t-skóry beta-prime rozdělení’





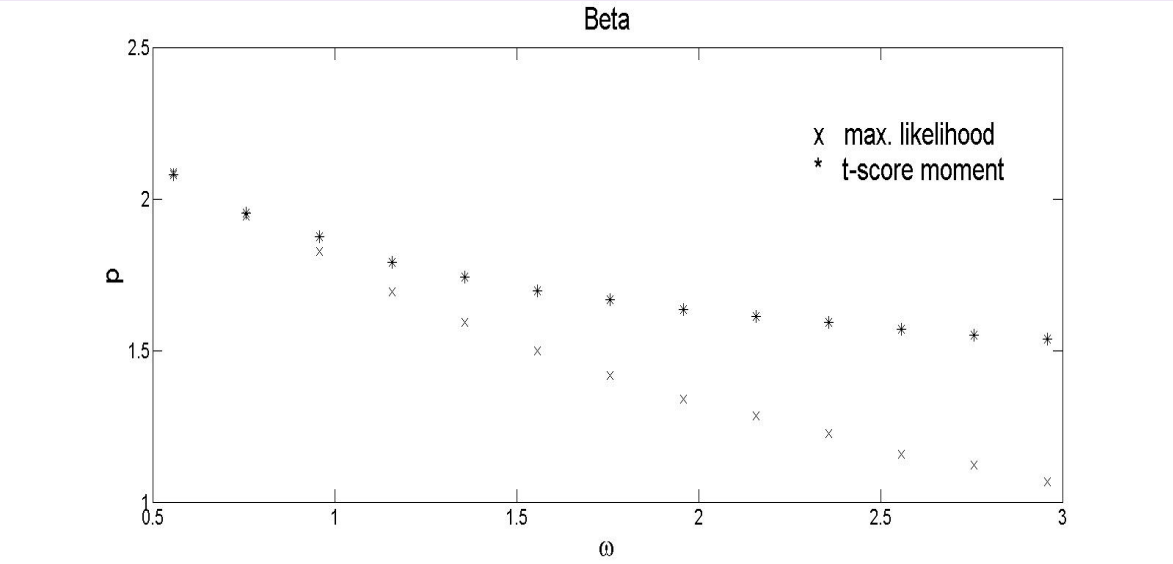
Odhady pro beta rozdělení

- $f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1} \quad T_F(x) = (p+q)x - p$

Odhady pro beta rozdělení

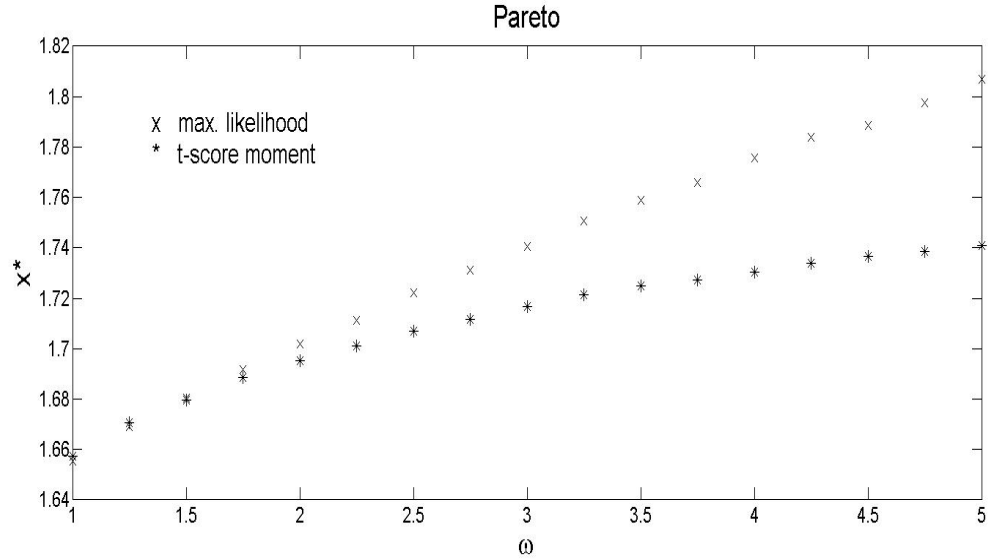
- $f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1} \quad T_F(x) = (p+q)x - p$
- $x^* = \frac{p}{p+q} \quad S_F(x) = p(x - x^*) \quad \hat{x}^* = \bar{x}$

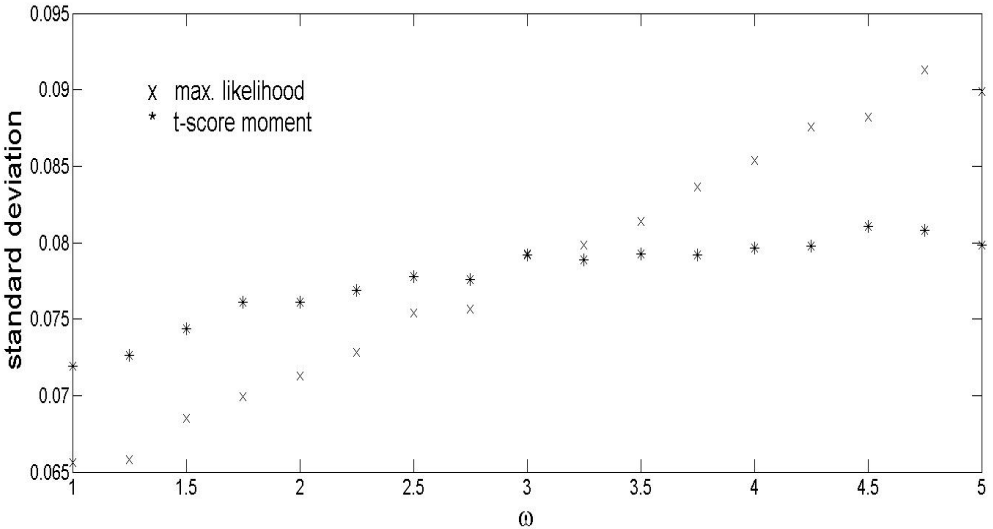
Odhady pro beta rozdělení

- $f(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1} \quad T_F(x) = (p+q)x - p$
- $x^* = \frac{p}{p+q} \quad S_F(x) = p(x - x^*) \quad \hat{x}^* = \bar{x}$
- 

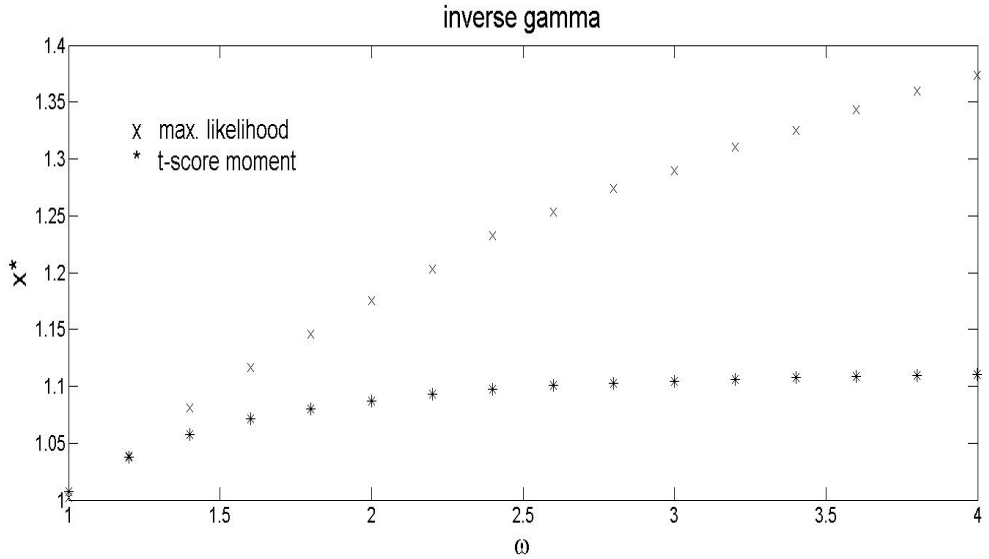
$$F_{kont.} = 0.9 * Beta(2,2) + 0.1 * Beta(\omega)$$

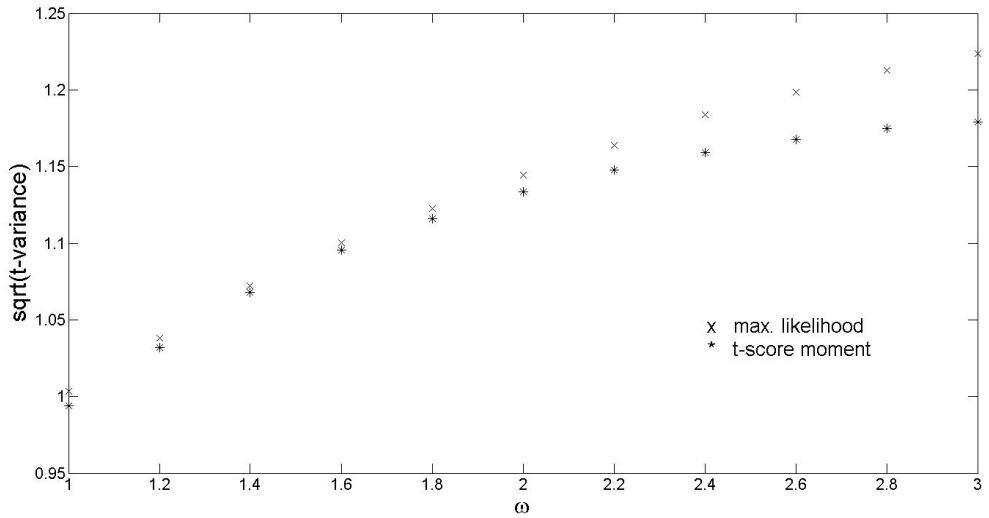
$\hat{x}^*$ a $std(\hat{x}^*)$ Pareto





$\hat{x}^*$ a $\hat{\omega}$ inversní gamma





Exponenciální na (a, ∞)

- $f(x) = \frac{1}{\tau} e^{-\frac{x-a}{\tau}}$ $\qquad\qquad$ $T_F(x) = \frac{(x-a)}{\tau} - 1$

Exponenciální na (a, ∞)

- $f(x) = \frac{1}{\tau} e^{-\frac{x-a}{\tau}}$ $\qquad$ $T_F(x) = \frac{(x-a)}{\tau} - 1$
-

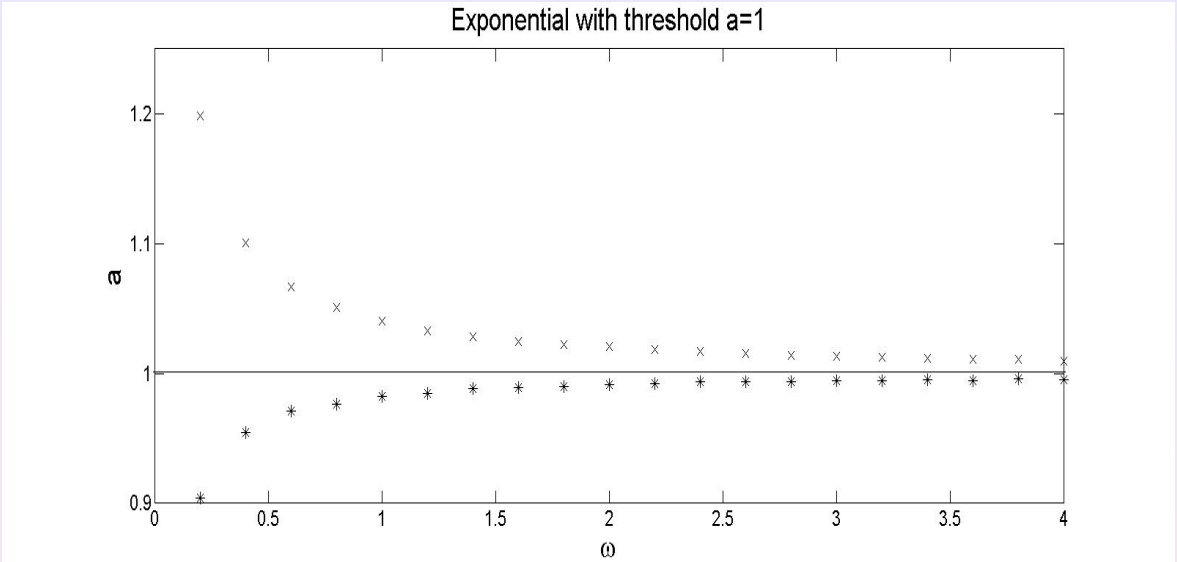
$$\sum_{i=1}^{n} \left(\frac{x_i - a}{\tau} - 1 \right) = 0$$

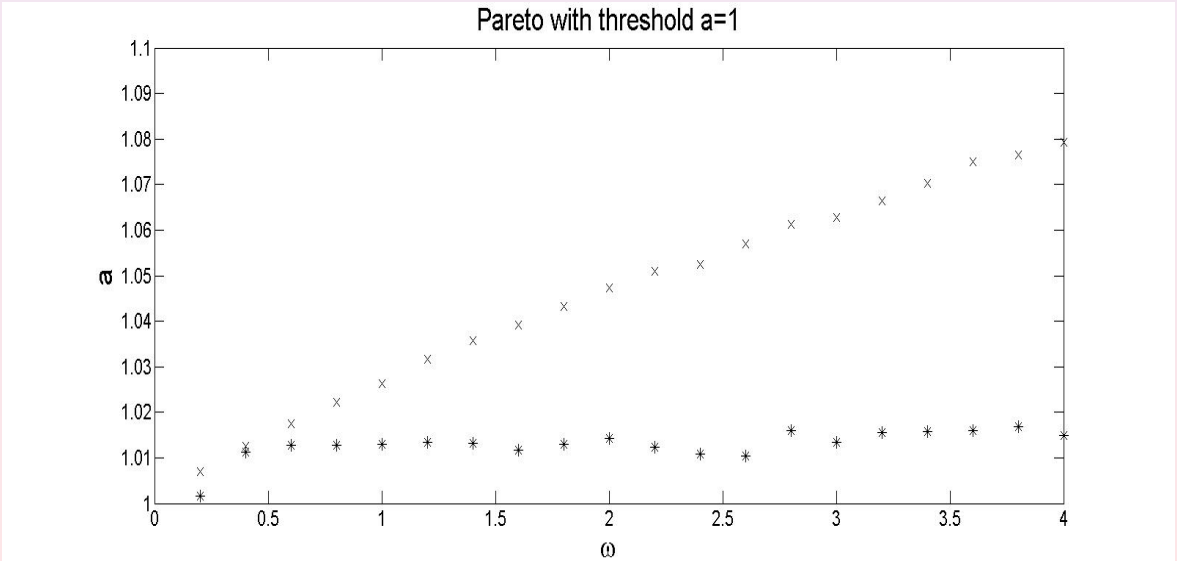
$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - a}{\tau} - 1 \right)^2 = 1$$

$\hat{\tau} = (\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2)^{1/2}$ $\qquad$ $\hat{a} = \bar{x} - \hat{\tau}$

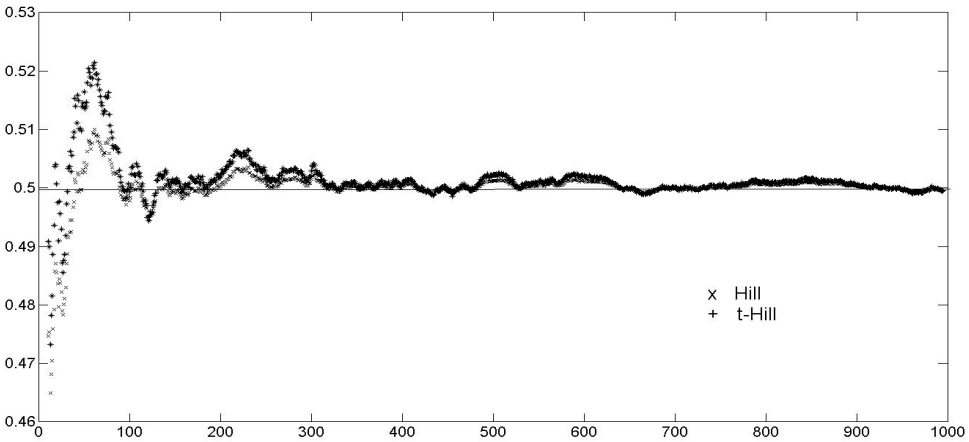
$\hat{a}_M = \min(\hat{a}, x_{(1)})$

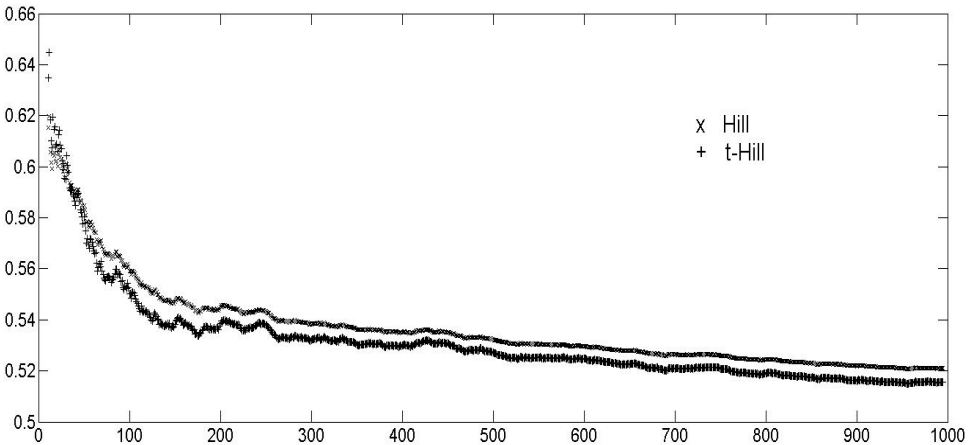
Exponential and Pareto with threshold





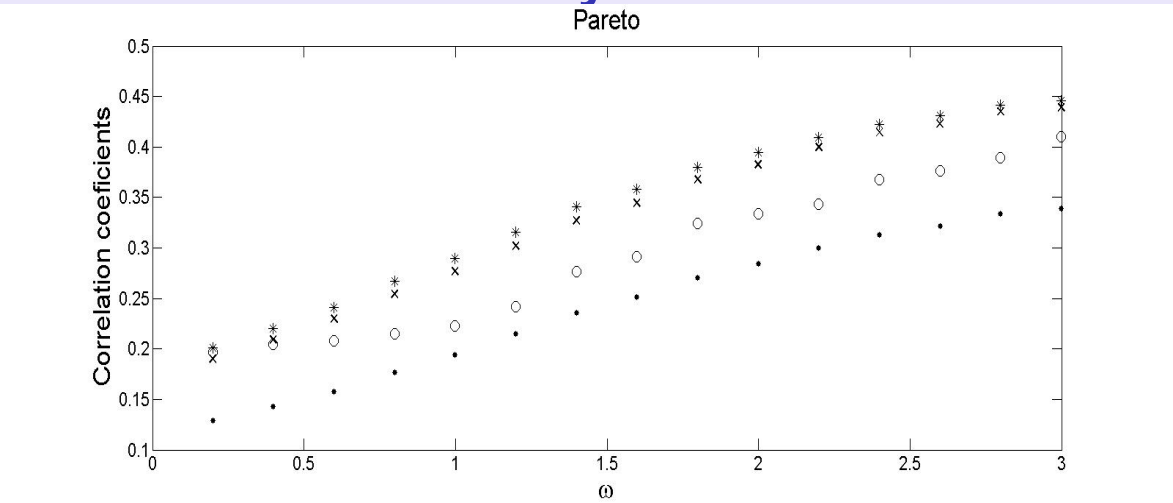
Hill plots

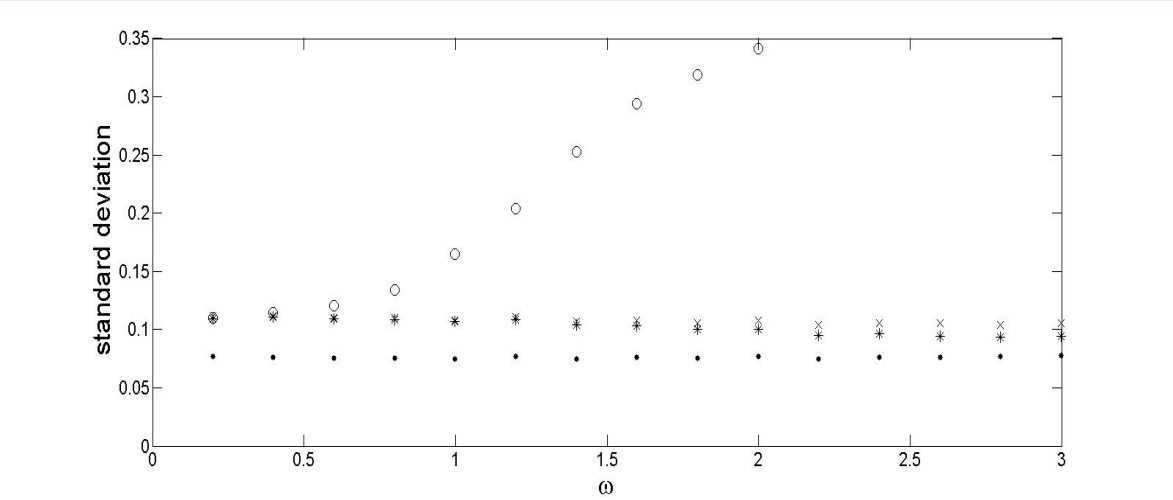




$$F_1 = P(\omega = 0.71), F_2 = 0.9 * P(0.71) + 0.1 * P(\omega = 3)$$

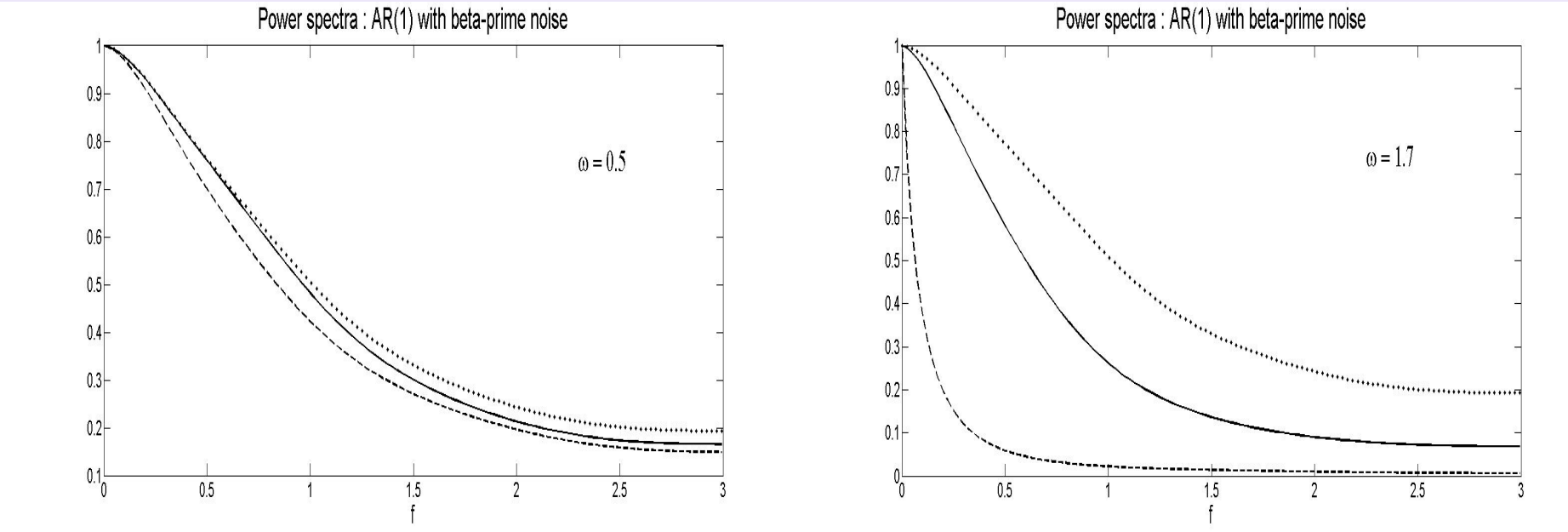
Korelační koeficienty





o Pearson, . Kendall, x Spearman, * t-score

Odhady spektrální hustoty



$X_t = 0.4X_{t-1} + Z_t$ with beta-prime noise with different ω
Full line: $T_F(X_t)$, dashed line: $\log X_t$, dotted line: normal Z_t

Závěr

Fabián, Z. (2001). Induced cores and their use in robust parametric estimation,. Comm. in Statist., Theory Meth. 30, 537-556.
Fabián, Z. (2008). New measures of central tendency and variability of continuous distributions. Comm. in Statist., Theory Meth. 37 159-174.
Fabián, Z., Stehlík, M. (2008). A note on favorable estimation when data is contaminated. Comm. Dep. and Quality Management 11, 36-43.
Fabián, Z. (2009). Confidence intervals for a new characteristic of central tendency of distributions. Comm. Statist. Theory Methods 38, 1804-1814.