Resuscitace momentové metody

Zdeněk Fabián
Ústav informatiky AV ČR Praha

ROBUST 2012

$(x_1, ..., x_n)$: X s hustotou f na $\mathcal{X} \subseteq \mathbb{R}$

- Parametrický model $f(x; \theta), \theta = (\theta_1, ..., \theta_m)$

$(x_1, ..., x_n)$: X s hustotou f na $\mathcal{X} \subseteq \mathbb{R}$

- Parametrický model $f(x; \theta), \theta = (\theta_1, ..., \theta_m)$
- Momentová metoda

  $S(X)$: $\quad ES^k(\theta) = \int_{\mathcal{X}} S^k(x; \theta) f(x; \theta)\, dx$

$(x_1, ..., x_n)$: X s hustotou f na $\mathcal{X} \subseteq \mathbb{R}$

- Parametrický model $f(x; \theta), \theta = (\theta_1, ..., \theta_m)$
- Momentová metoda

  $S(X)$: $\quad ES^k(\theta) = \int_{\mathcal{X}} S^k(x; \theta) f(x; \theta)\, dx$
- $$\frac{1}{n} \sum_{i=1}^{n} S^k(x_i; \theta) = ES^k(\theta) \quad k = 1, ..., m$$

$(x_1, ..., x_n)$: X s hustotou f na $\mathcal{X} \subseteq \mathbb{R}$

- Parametrický model $f(x; \theta), \theta = (\theta_1, ..., \theta_m)$

Momentová metoda

$S(X)$: $\quad ES^k(\theta) = \int_{\mathcal{X}} S^k(x; \theta) f(x; \theta)\, dx$

$$\frac{1}{n} \sum_{i=1}^{n} S^k(x_i; \theta) = ES^k(\theta) \qquad S(x; \theta) = x$$

Metoda maximum likelihood $\quad L(\theta) = f(x; \theta)$

Skórová funkce $\quad U(\theta) = \frac{\partial}{\partial \theta} \log L(\theta)$

$$\frac{1}{n} \sum_{i=1}^{n} U_{\theta_k}(x_i; \theta) = 0 \qquad k = 1, ..., m$$

Skórová funkce rozdělení na *R*

- odhady momentovou metodou:
  nejsou optimální ale jsou to názorné charakteristiky datového souboru

Skórová funkce rozdělení na R

- odhady momentovou metodou:
  nejsou optimální ale jsou to názorné charakteristiky datového souboru
- Pro některá rozdělení obě metody dávají identické odhady

  $\mathbb{R}, \theta = \mu, f(x-\mu)$:

$$U(x-\mu) = \frac{\partial}{\partial \mu} \log f(x-\mu) = -\frac{f'(x-\mu)}{f(x-\mu)} \equiv S(x-\mu)$$

Skórová funkce rozdělení na R

- odhady momentovou metodou:
  nejsou optimální ale jsou to názorné charakteristiky datového souboru
- Pro některá rozdělení obě metody dávají identické odhady

  $\mathbb{R}, \theta = \mu, f(x-\mu)$:

$$U(x-\mu) = \frac{\partial}{\partial \mu} \log f(x-\mu) = -\frac{f'(x-\mu)}{f(x-\mu)} \equiv S(x-\mu)$$

- Na $\mathbb{R}$ máme skórovou funkci rozdělení $S(x) = -\frac{f'(x)}{f(x)}$

Obecně to nejde

- exponenciální $f(x) = \frac{1}{\tau} e^{-x/\tau}$ na $\mathcal{X} = (0, \infty)$

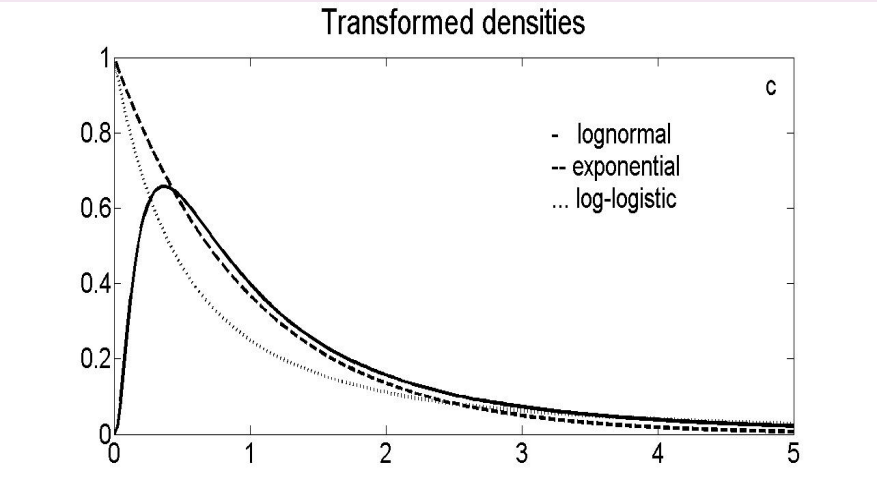
$$-\frac{f'(x)}{f(x)} = 1$$

Obecně to nejde

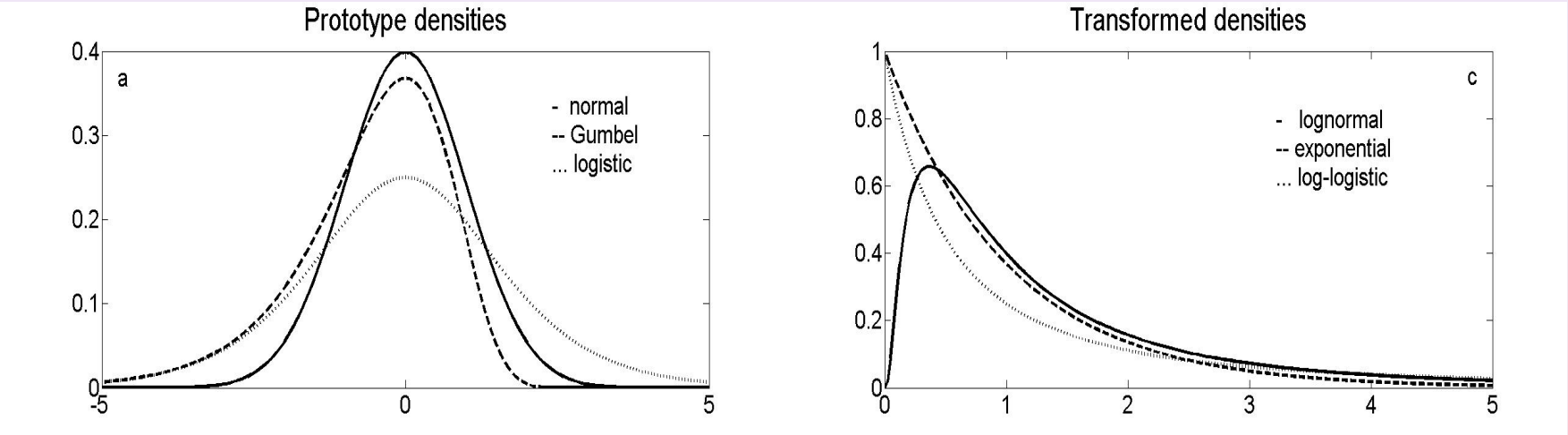
- exponenciální $f(x) = \frac{1}{\tau} e^{-x/\tau}$ na $\mathcal{X} = (0, \infty)$

$$-\frac{f'(x)}{f(x)} = 1$$

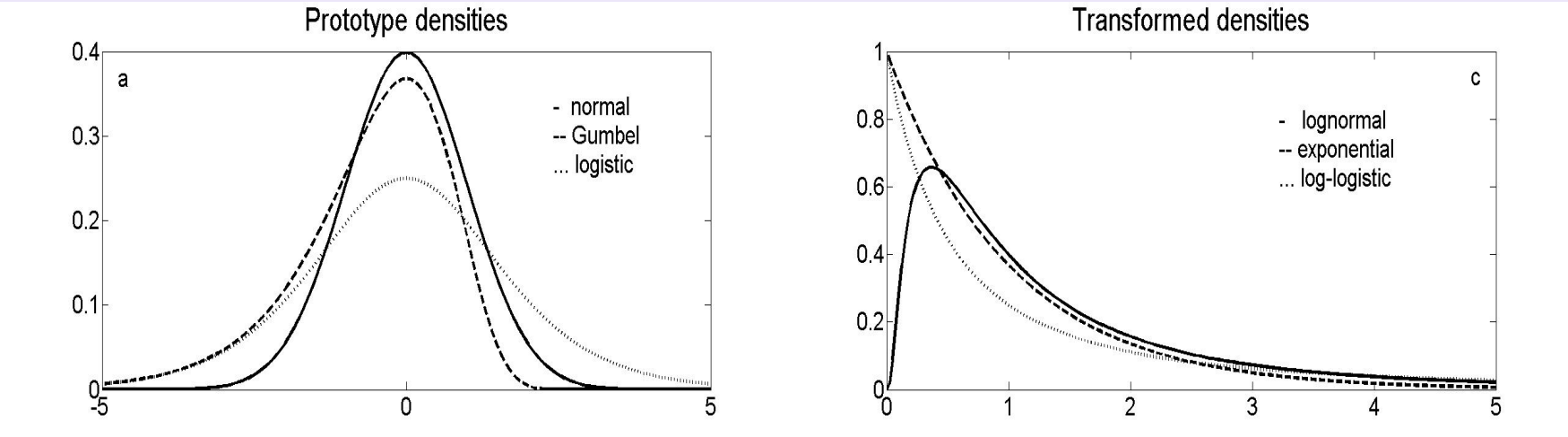
- Rozdělení na R_+



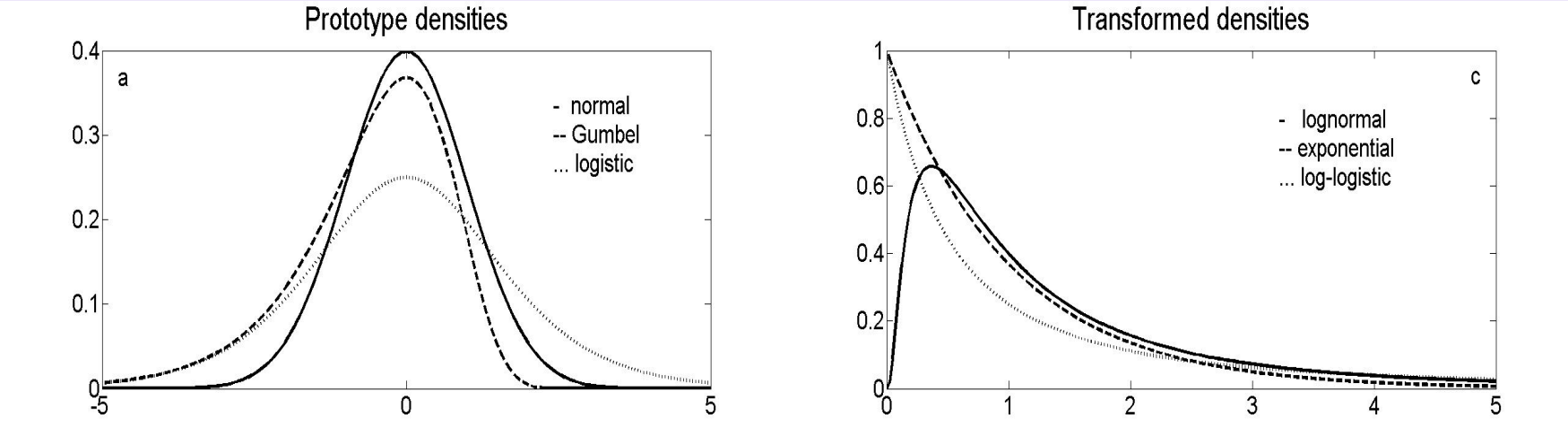
Transformovaná rozdělení



Transformovaná rozdělení



Skalární skór (pro libovolné rozdělení)



Definice

F na intervalu $\mathcal{X}$ s hustotou $f(x)$, $\eta : \mathcal{X} \to \mathbb{R}$:

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

$$T(x) = -\frac{1}{f(x)} \frac{d}{dx} \left(\frac{1}{\eta'(x)} f(x) \right)$$

$$x^* : \qquad T(x) = 0$$

Pak $S(x) = \eta'(x^*) T(x)$ je skórová funkce rozdělení F

na $\mathcal{X} = (0, \infty)$ je $S(x) = \frac{1}{x^*}[-1 - x \frac{f'(x)}{f(x)}]$

Nový pohled na pravd. rozdělení

- $\mathcal{X}, \theta \in \Theta \subseteq \mathbb{R}$

  $F(x), f(x), S(x)$ ... skórová funkce rozdělení

  $F(x;\theta), f(x;\theta), S(x;\theta)$ ... skalární skór

Nový pohled na pravd. rozdělení

- $\mathcal{X}, \theta \in \Theta \subseteq \mathbb{R}$

  $F(x), f(x), S(x)$ ... skórová funkce rozdělení

  $F(x;\theta), f(x;\theta), S(x;\theta)$ ... skalární skór

- Typická hodnota: těžiště $x^*(\theta)$: ES=0

$$S(x;\theta) = 0$$

Nový pohled na pravd. rozdělení

- $\mathcal{X}, \theta \in \Theta \subseteq \mathbb{R}$

  $F(x), f(x), S(x)$ ... skórová funkce rozdělení

  $F(x;\theta), f(x;\theta), S(x;\theta)$ ... skalární skór

- Typická hodnota: těžiště $x^*(\theta)$: ES=0

$$S(x;\theta) = 0$$

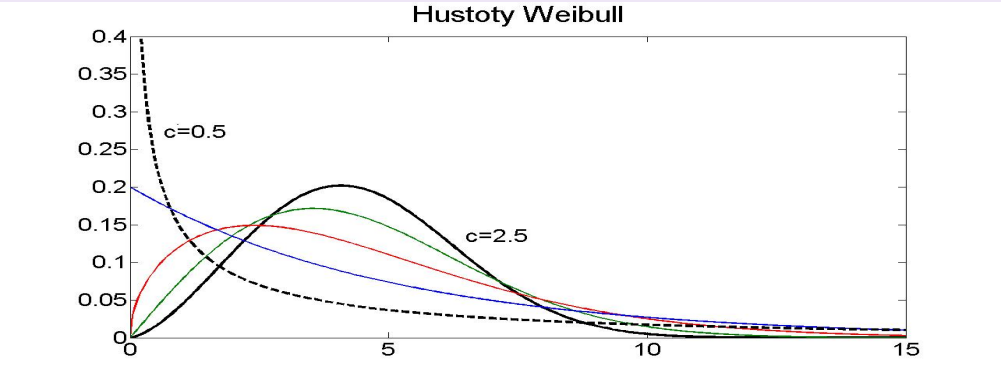
- Shoda s klasickou statistikou:

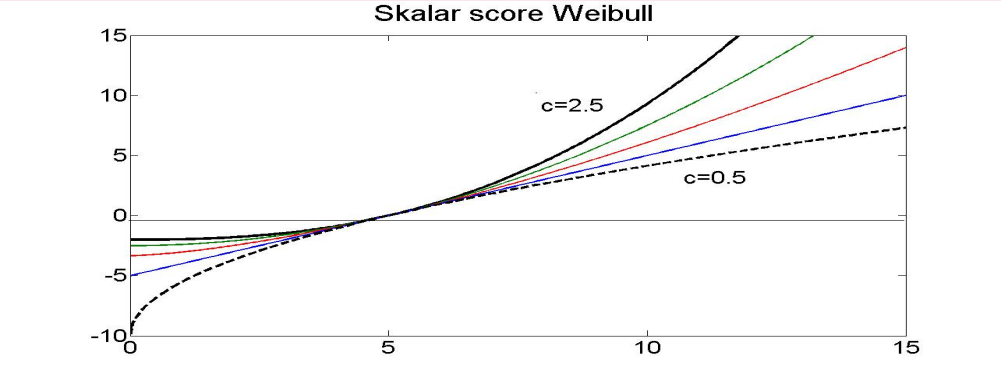
  Věta. Když $\theta = \tau = \eta^{-1}(\mu)$

$$S(x;\tau) = \frac{\partial}{\partial \tau}(-\log f(x;\tau))$$

$\mathcal{X} = (0, \infty)$, Weibull

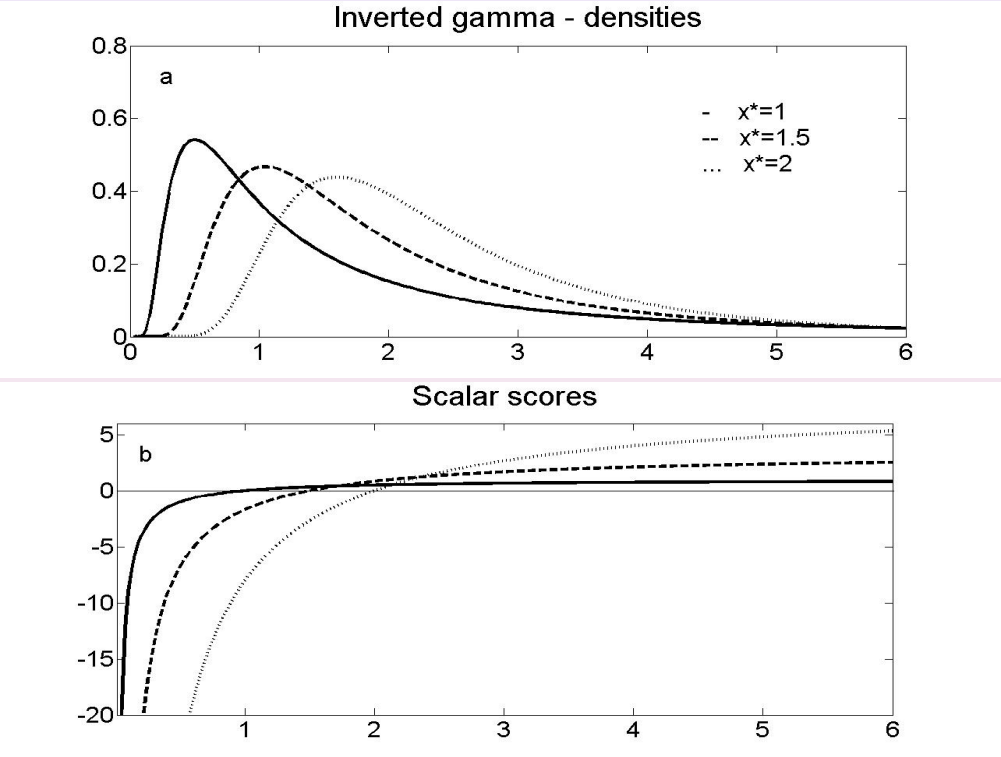
$$f(x; \tau, c) = \frac{c}{\tau}\left(\frac{x}{\tau}\right)^{c-1} e^{-\left(\frac{x}{\tau}\right)^{c}} \qquad S(x; \tau, c) = \frac{c}{\tau}\left(\left(\frac{x}{\tau}\right)^{c} - 1\right)$$





$\mathcal{X} = (0, \infty)$, inverted gamma

$$f(x; \alpha, \gamma) = \frac{\gamma^\alpha}{x\Gamma(\alpha)} x^{-\alpha} e^{-\gamma/x} \qquad S(x; \alpha, \gamma) = \frac{\gamma^2}{\alpha}(1 - x^*/x)$$



Co přináší S nového pro popis modelu

- A. Systematika rozdělení: podle chování S na koncích intervalu $\mathcal{X}$

Co přináší S nového pro popis modelu

- A. Systematika rozdělení: podle chování S na koncích intervalu $\mathcal{X}$
- B. Nové numerické charakteristiky: x^*, ω^2

Co přináší S nového pro popis modelu

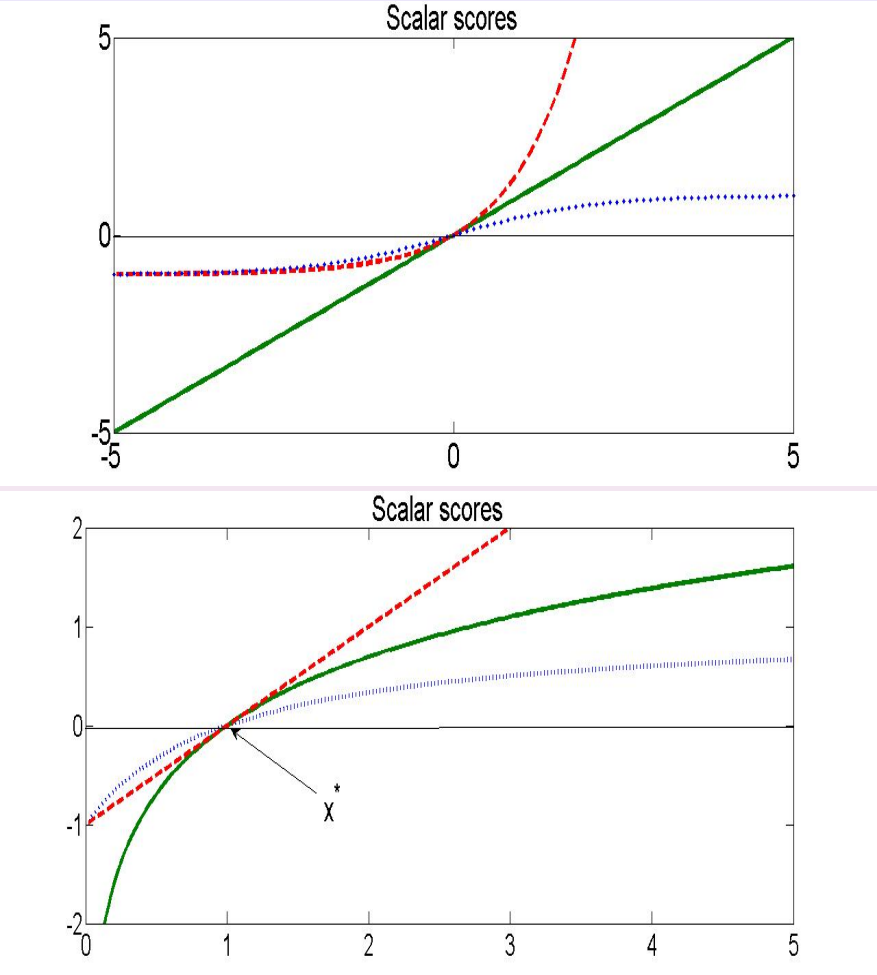
- A. Systematika rozdělení: podle chování S na koncích intervalu $\mathcal{X}$
- B. Nové numerické charakteristiky: x^*, ω^2
- C. Nové funkce charakterizující rozdělení:
  $S(x)$ vlivová funkce
  $S^2(x)$ informační funkce
  $S'(x) = \frac{dS(x)}{dx}$ vahová funkce

Co přináší S nového pro popis modelu

- A. Systematika rozdělení: podle chování S na koncích intervalu $\mathcal{X}$
- B. Nové numerické charakteristiky: x^*, ω^2
- C. Nové funkce charakterizující rozdělení:
  $S(x)$ vlivová funkce
  $S^2(x)$ informační funkce
  $S'(x) = \frac{dS(x)}{dx}$ vahová funkce
- D. Vzdálenosti mezi daty a modely

$$d(x_1, x_2) = |S(x_1) - S(x_2)| \qquad D(f, g) = E_f(S_f - S_g)^2$$

A. Systematika rozdělení



Beta triplet

$\mathcal{X}$	distribution	density
R	'prototype beta'	$\frac{1}{B(p,q)} \frac{e^{py}}{(e^y+1)^{p+q}}$
$(0,\infty)$	beta-prime	$\frac{1}{B(p,q)} \frac{x^{p-1}}{(x+1)^{p+q}}$
$(0,1)$	beta	$\frac{z^{p-1}(1-z)^{q-1}}{B(p,q)}$

B. Numerické charakteristiky: momenty ES^k

- $ES = 0$. Poloha na ose x: x^*: $S(x) = 0$

  ES^2 Fisherova informace rozdělení

B. Numerické charakteristiky: momenty ES^k

- $ES = 0$. Poloha na ose x: x^*: $S(x) = 0$

  ES^2 Fisherova informace rozdělení

- Míra variability rozdělení: score variance

$$\omega^2(\theta) = \frac{1}{ES^2(\theta)}$$

B. Numerické charakteristiky: momenty ES^k

- $ES = 0$. Poloha na ose x: x^*: $S(x) = 0$

  ES^2 Fisherova informace rozdělení

- Míra variability rozdělení: score variance

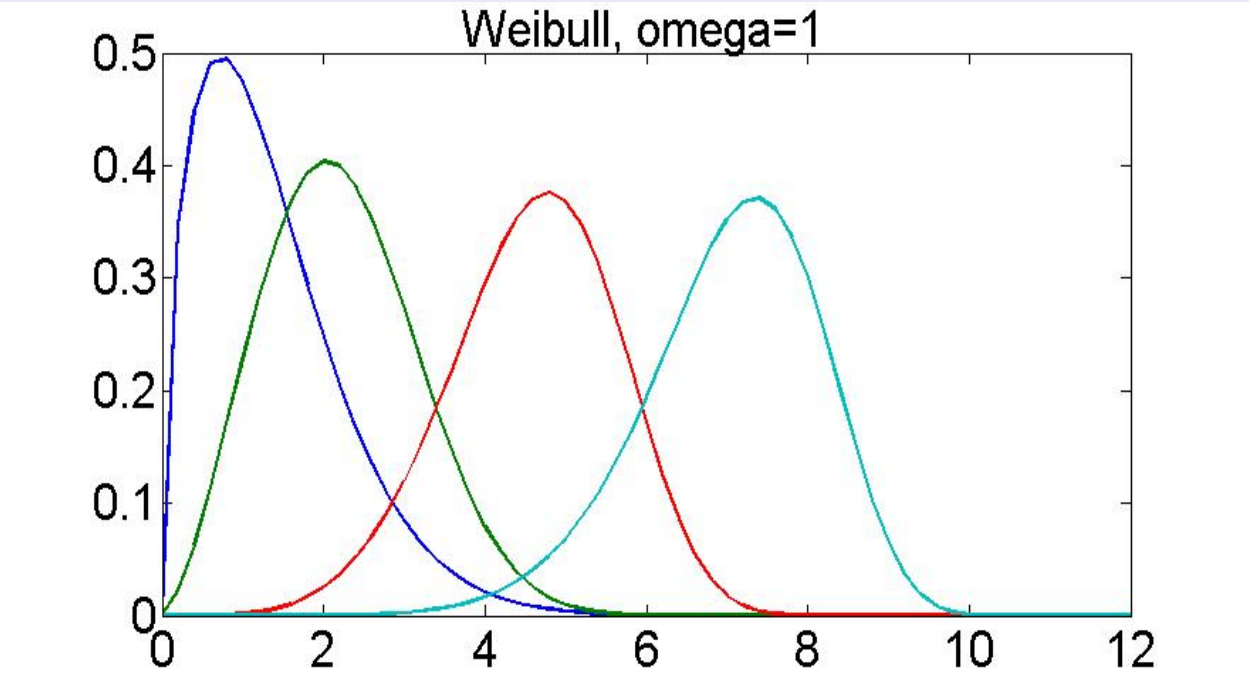
$$\omega^2(\theta) = \frac{1}{ES^2(\theta)}$$

- pozn. normální rozdělení $S(x; \mu, \sigma) = \frac{x-\mu}{\sigma^2}$,
  $x^* = \mu, \omega = \sigma$

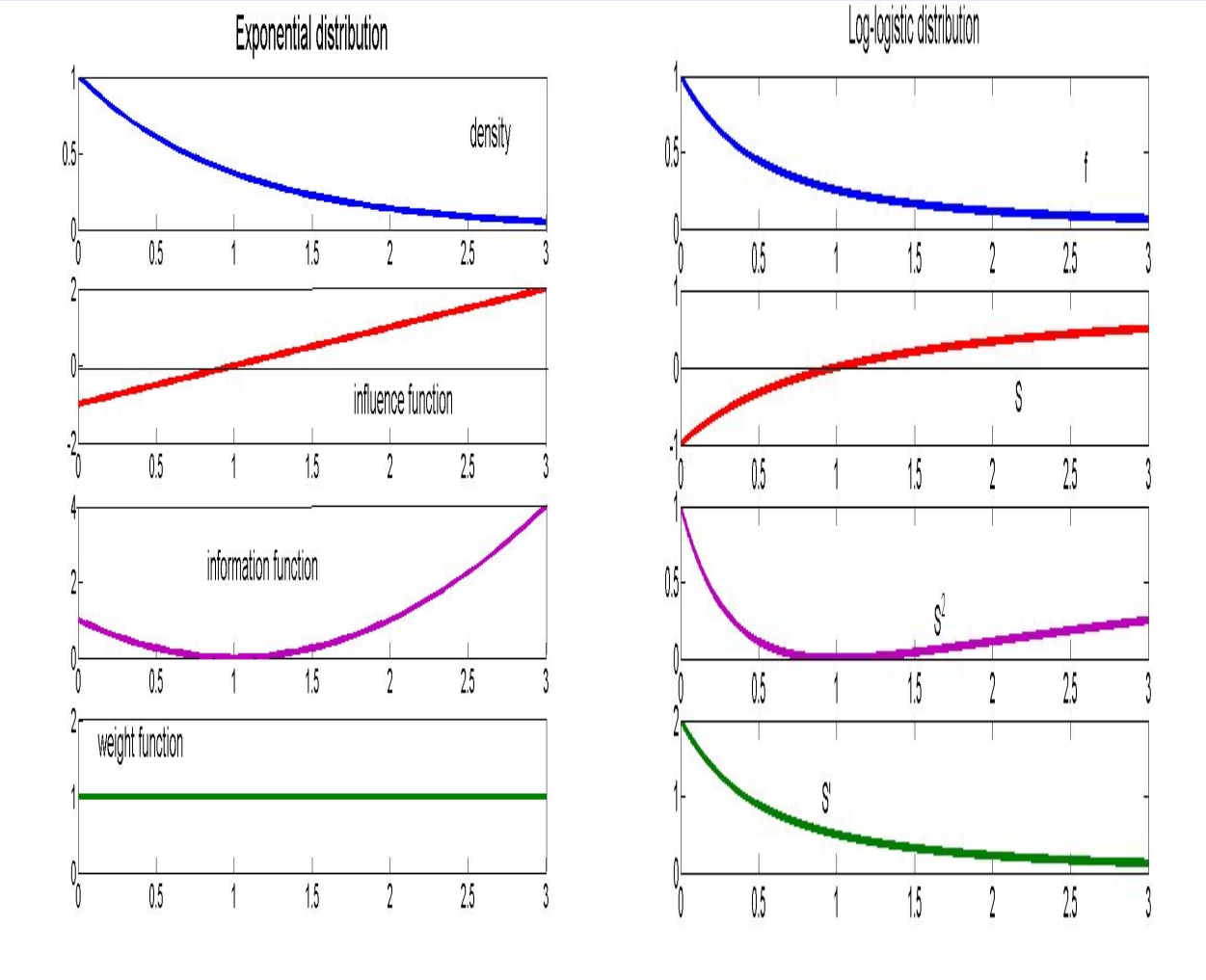
Těžiště a score variance

F	$f(x)$	$S(x)$	m	x^*	ω^2
expon.	$\frac{1}{\tau}e^{-x/\tau}$	$\frac{1}{\tau}(\frac{x}{\tau}-1)$	τ	τ	τ^2
lognor.	$\frac{c}{\sqrt{2\pi}x}e^{-\frac{1}{2}\ln^2(\frac{x}{\tau})^c}$	$\frac{c}{\tau}\ln(\frac{x}{\tau})^c$	$\tau(e^{\frac{1}{c^2}})^{1/2}$	τ	τ^2/c^2
Weibull	$\frac{c}{x}(\frac{x}{\tau})^c e^{-(\frac{x}{\tau})^c}$	$\frac{c}{\tau}((\frac{x}{\tau})^c-1)$	$\tau\Gamma(\frac{1}{c}+1)$	τ	τ^2/c^2
log-log.	$\frac{c}{x}\frac{(x/\tau)^c}{((x/\tau)^c+1)^2}$	$\frac{c}{\tau}\frac{(x/\tau)^c-1}{(x/\tau)^c+1}$		τ	$\frac{3\tau^2}{c^2}$
gamma	$\frac{\gamma^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\gamma x}$	$\frac{\gamma^2}{\alpha}(x-x^*)$	$\frac{\alpha}{\gamma}$	$\frac{\alpha}{\gamma}$	$\frac{\alpha}{\gamma^2}$
Pareto	c/x^{c+1}	$\frac{c^2}{c+1}(1-\frac{x^*}{x})$	$\frac{c}{c-1}$	$\frac{c+1}{c}$	$\frac{c+2}{c^3}$
beta-pr.	$\frac{1}{B(p,q)}\frac{x^{p-1}}{(x+1)^{p+q}}$	$\frac{q^2}{p}\frac{x-x^*}{x+1}$	$\frac{p}{q-1}$	$\frac{p}{q}$	$\frac{p(p+q+1)}{q^3}$
Fréchet	$\frac{c}{x}(\frac{\tau}{x})^c e^{-(\frac{\tau}{x})^c}$	$\frac{c}{\tau}[1-(\frac{\tau}{x})^c]$	$\tau\Gamma(1-\frac{1}{c})$	τ	τ^2/c^2

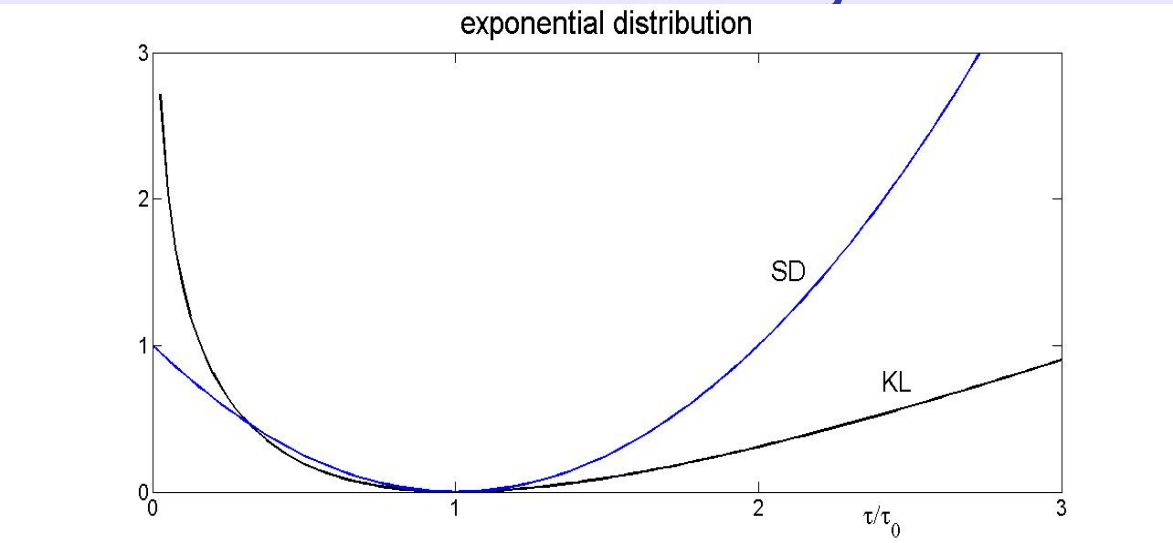
Variabilita pomocí ω^2

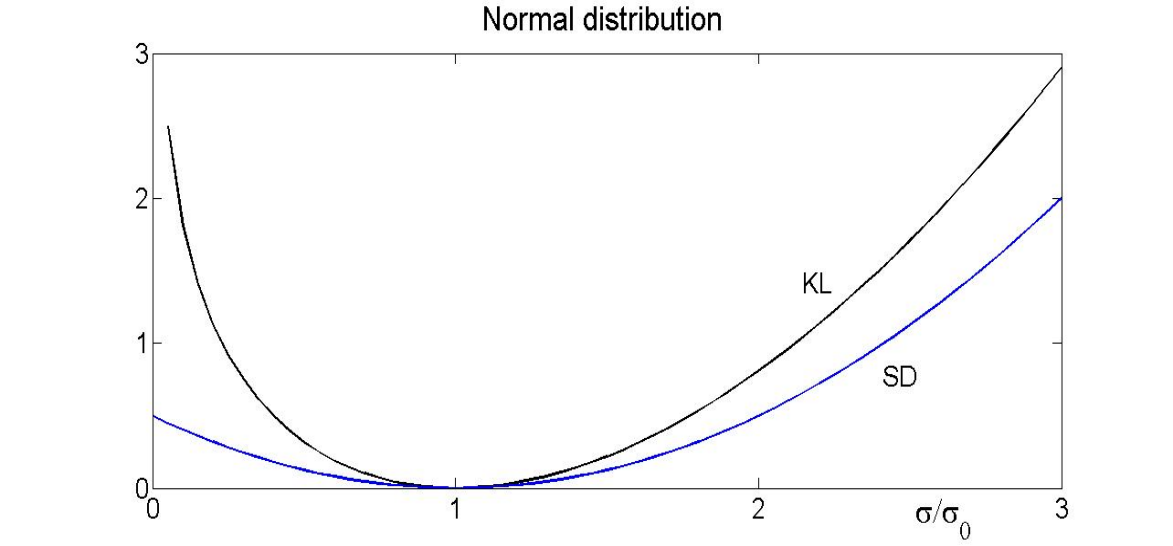


C. Funkce charakterizující rozdělení

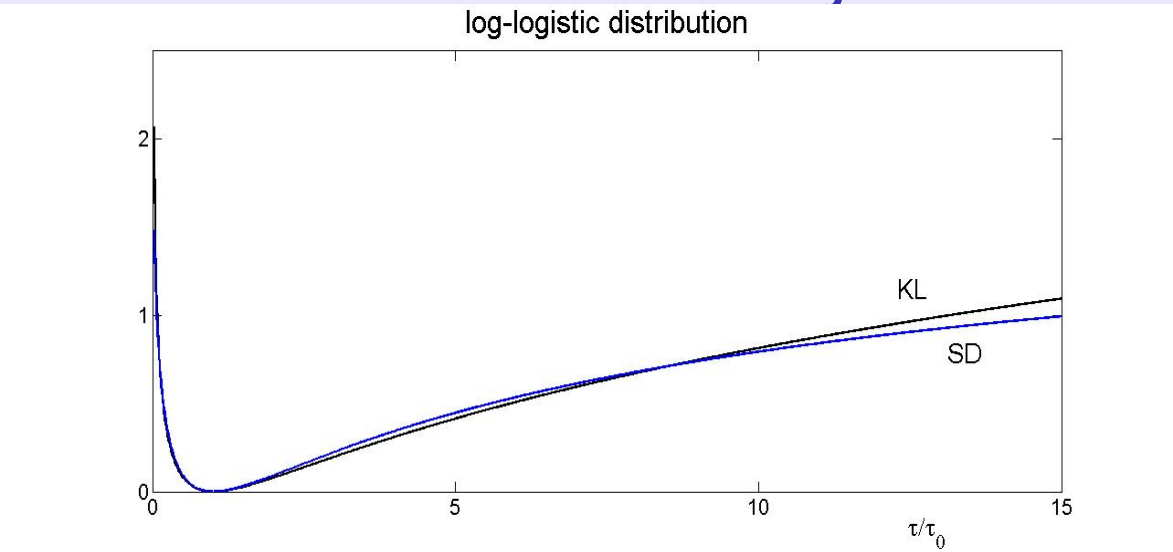


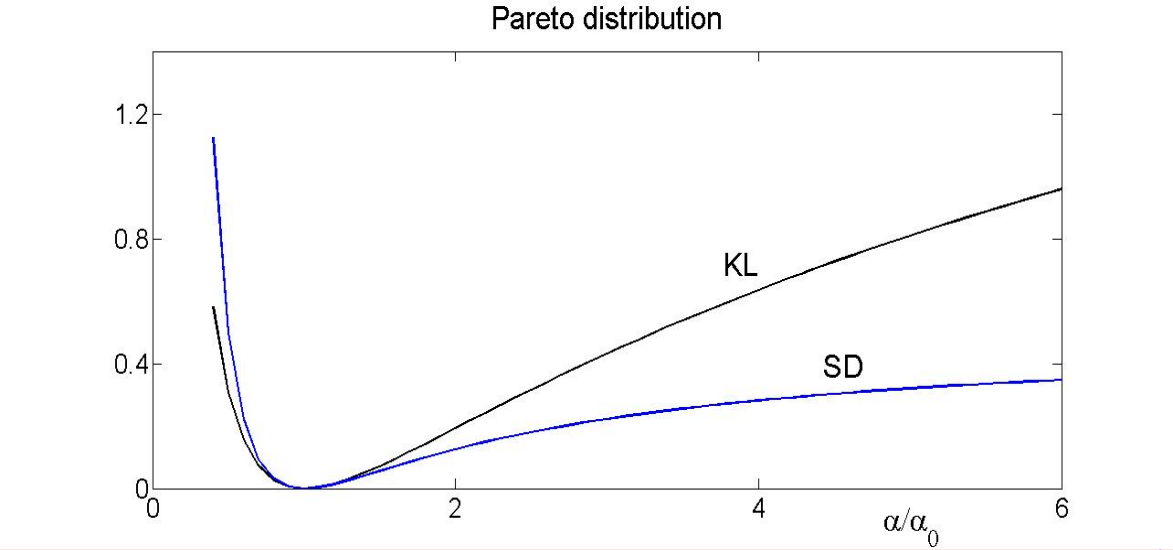
D. Vzdálenosti mezi modely





D. Vzdálenosti mezi modely





Co přináší *S* nového pro statistiku

Málo dat:

- Neznáme model, ’čistá’ data: neparametrické metody

Co přináší *S* nového pro statistiku

Málo dat:

- Neznáme model, ’čistá’ data: neparametrické metody
- Neznáme model, kontaminovaná data: metody robustní statistiky

Co přináší *S* nového pro statistiku

Málo dat:

- Neznáme model, ’čistá’ data: neparametrické metody
- Neznáme model, kontaminovaná data: metody robustní statistiky
- Známe model, ’čistá’ data: parametrické metody klasické statistiky

Co přináší *S* nového pro statistiku

Málo dat:

- Neznáme model, ’čistá’ data: neparametrické metody
- Neznáme model, kontaminovaná data: metody robustní statistiky
- Známe model, ’čistá’ data: parametrické metody klasické statistiky
- Známe model, kontaminovaná data: parametrické metody založené na skalárním skóru

S jako inferenční funkce pro bodové odhady

- 1. Rovnice pro odhady

$$\hat{\theta}_{SM}: \qquad \frac{1}{n}\sum_{i=1}^{n} S^k(x_i;\theta) = ES^k(\theta) \qquad k = 1, ..., m$$

M-estimátor konsistentní, asympt. normální
a robustní pro **všechny** parametry když S je omezená

S jako inferenční funkce pro bodové odhady

- 1. Rovnice pro odhady

$$\hat{\theta}_{SM}: \quad \frac{1}{n}\sum_{i=1}^{n} S^k(x_i;\theta) = ES^k(\theta) \quad k = 1, ..., m$$

M-estimátor konsistentní, asympt. normální
a robustní pro **všechny** parametry když S je omezená

- 2. Neomezená skalární funkce jde snadno 'huberizovat'

S jako inferenční funkce pro bodové odhady

- 1. Rovnice pro odhady

$$\hat{\theta}_{SM}: \quad \frac{1}{n}\sum_{i=1}^{n} S^k(x_i;\theta) = ES^k(\theta) \quad k = 1, ..., m$$

M-estimátor konsistentní, asympt. normální
a robustní pro **všechny** parametry když S je omezená

- 2. Neomezená skalární funkce jde snadno 'huberizovat'
- 3. Výsledky pro rozdílné modely jdou porovnávat pomocí vzdáleností $d(x_0^*, \hat{x}^*)$ kde

$$\hat{x}^* = x^*(\hat{\theta}), \quad \hat{\omega}^* = \omega(\hat{\theta})$$

a $D(F_{\theta_0}, F_{\hat{\theta}})$

Těžiště: $\sum_i S(x_i;\theta) = 0$

exponential	$\sum_i (\frac{x_i}{\tau} - 1) = 0$	$\hat{\tau} = \bar{x}$
lognormal	$\sum_i \log(x_i/\tau) = 0$	$\hat{\tau} = \bar{x}_g$
Weibull	$\sum_i \left[(\frac{x_i}{\tau})^c - 1 \right] = 0$	$\hat{\tau} = (\frac{1}{n} \sum_i x_i^c)^{1/c}$
Pareto	$\sum_i (1 - x^*/x_i) = 0$	$\hat{x}^* = \frac{n}{\sum_i \frac{1}{x_i}}$

Konstrukce rodiny s těžkým chvostem

$$S(x) = \frac{x-1}{x+1}$$

Integrací rovnice

$$S(x) = \frac{1}{x^*}\left(-1 - x\frac{f'(x)}{f(x)}\right)$$

dostaneme hustotu $f(x) = 1/(1+x)^2$. Zobecníme $S(x)$ na

$$S(x;\tau,c,\alpha,\nu) = \frac{1}{\tau}\alpha c\frac{(x/\tau)^c - 1}{(x/\tau)^c + 1/\nu}$$

Konstrukce rodiny s těžkým chvostem

$$S(x) = \frac{x-1}{x+1}$$

Integrací rovnice

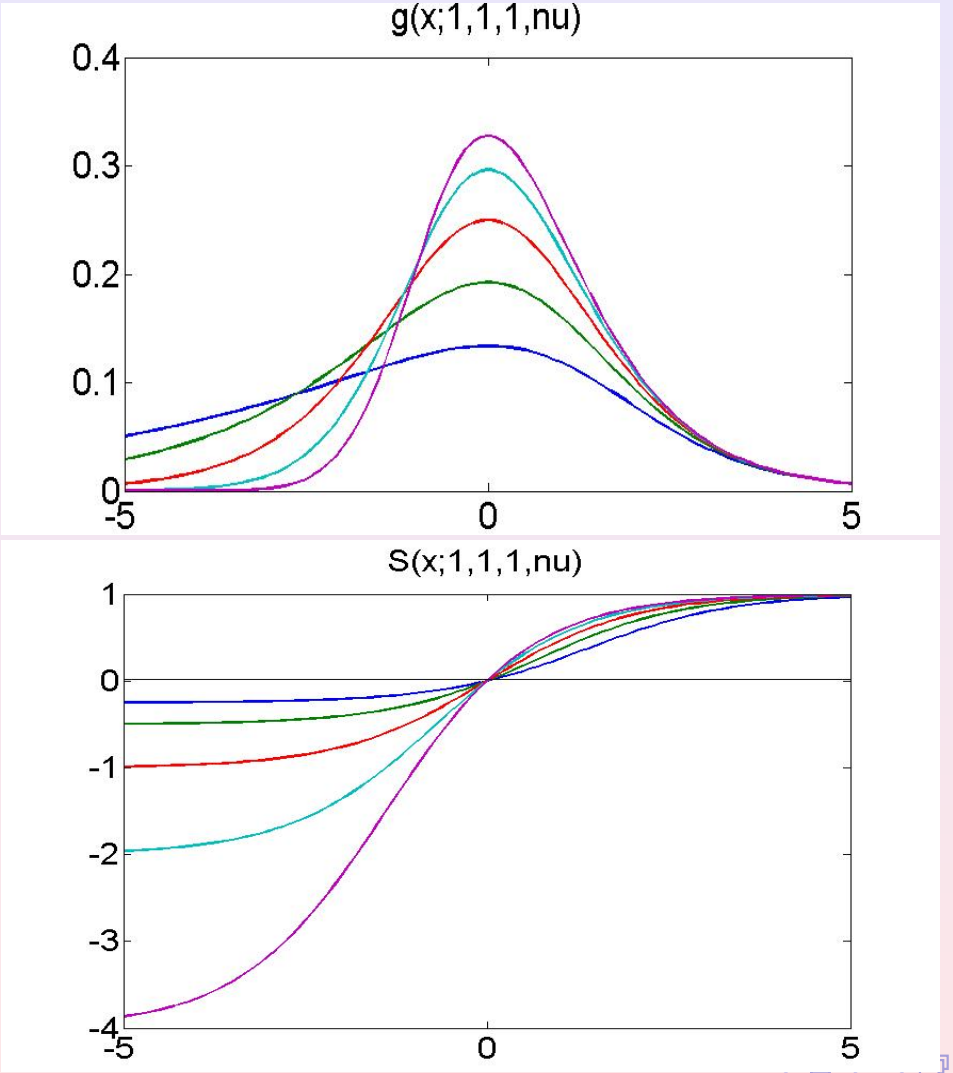
$$S(x) = \frac{1}{x^*}\left(-1 - x\frac{f'(x)}{f(x)}\right)$$

dostaneme hustotu $f(x) = 1/(1+x)^2$. Zobecníme $S(x)$ na

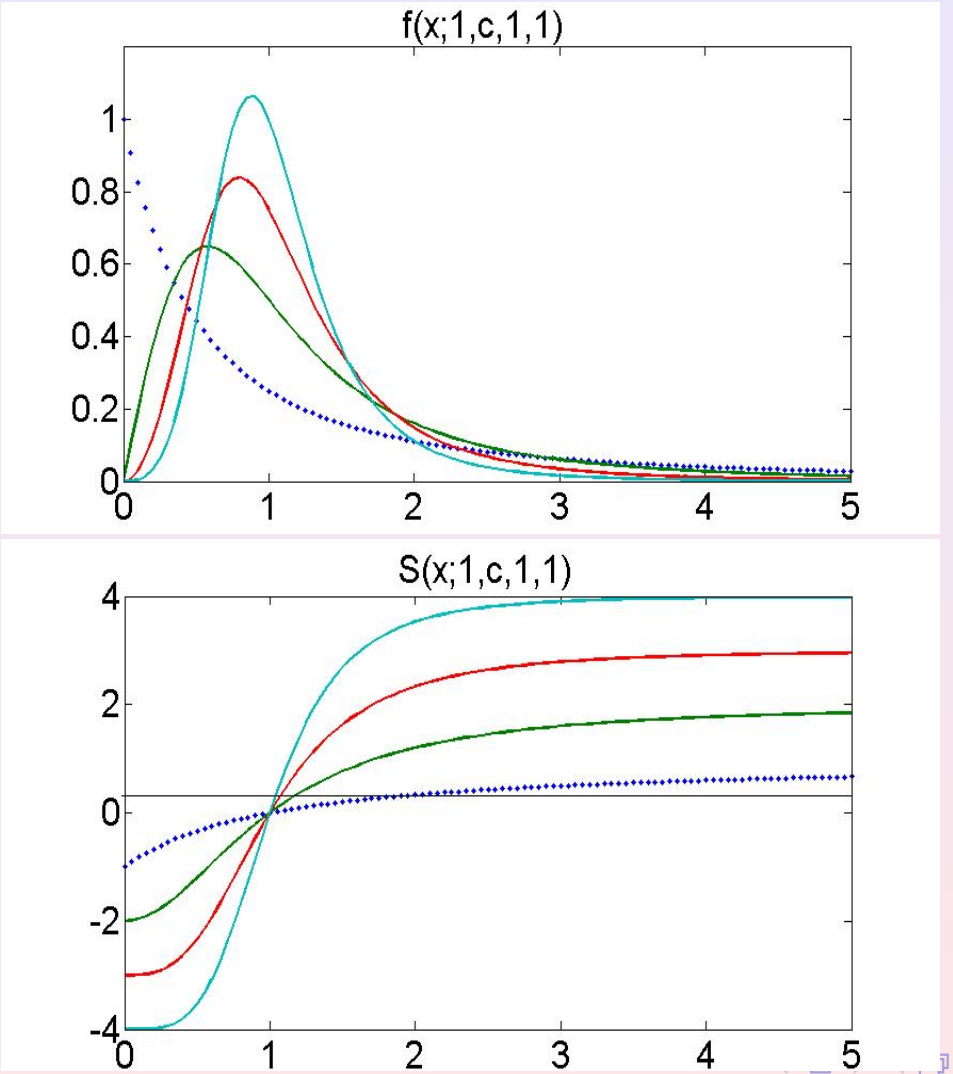
$$S(x;\tau,c,\alpha,\nu) = \frac{1}{\tau}\alpha c\frac{(x/\tau)^c - 1}{(x/\tau)^c + 1/\nu}$$

$$f(x;\tau,c,\alpha,\nu) = \frac{c}{\nu^\alpha B(\nu\alpha,\alpha)x}\frac{(x/\tau)^{c\nu\alpha}}{[(x/\tau)^c + 1/\nu]^{(1+\nu)\alpha}}$$

Prototype of $f(x; 1, 1, 1, \nu)$



$f(x; 1, c, 1, 1)$



Rozdělení beta-prime

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

$$x^* = p/q, ES^2 = \frac{q^3}{p(p+q+1)}.$$

Score moment rovnice

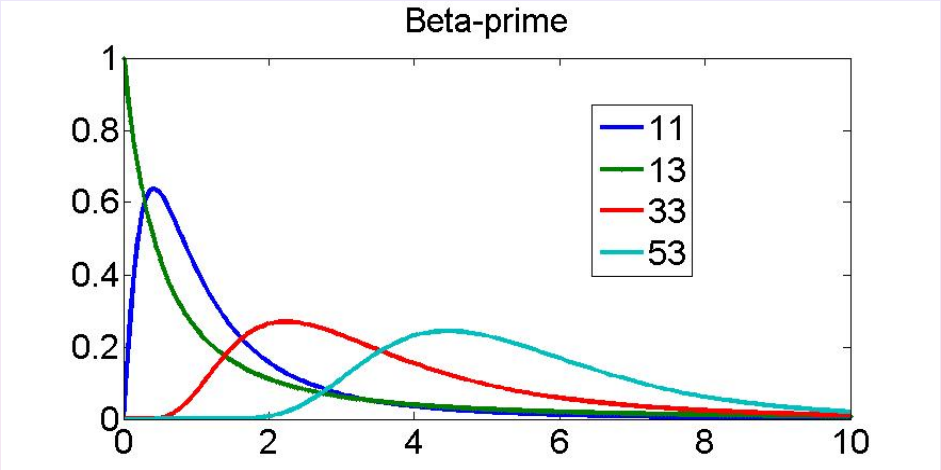
$$\sum_{i=1}^{n} \frac{x_i - x^*}{x_i + 1} = 0$$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - x^*}{x_i + 1} \right)^2 = \frac{p}{q(p+q+1)}$$

Z první rovnice

$$\hat{x}^* = \frac{\sum_{i=1}^{n} \frac{x_i}{1+x_i}}{\sum_{i=1}^{n} \frac{1}{1+x_i}}$$

Efficiencies: beta-prime(x^*, ω)



x^*	ω	x^*_{SM}	x^*_{ML}	ω_{SM}	ω_{ML}	eff(x^*_{SM})	eff(ω^2_{SM})
1	1	1.007	1.007	0.998	0.993	0.98	0.94
1	3	1.019	1.017	3.079	3.008	0.81	0.66
3	3	3.029	3.031	2.993	2.988	0.99	0.92
5	3	5.031	5.031	2.967	2.969	1	0.99

Rozdělení s neomezeným skalárním skórem

Aplikace postupu robustní statistiky:

$$\Psi(x;\theta) = \begin{cases} -b & \text{if} \quad x < u \\ S(x;\theta) & \text{if} \quad u \le x \le v \\ b & \text{if} \quad x > v \end{cases}$$

a odhad $\hat{\theta}$ parametru θ z rovnic

$$\sum_{i=1}^{n} \psi(x_i;\theta) = 0 \qquad \frac{1}{n}\sum_{i=1}^{n} \psi^2(x_i;\theta) = E\psi^2(\theta)$$

kde $\psi = \Psi - E\Psi$

Simulace

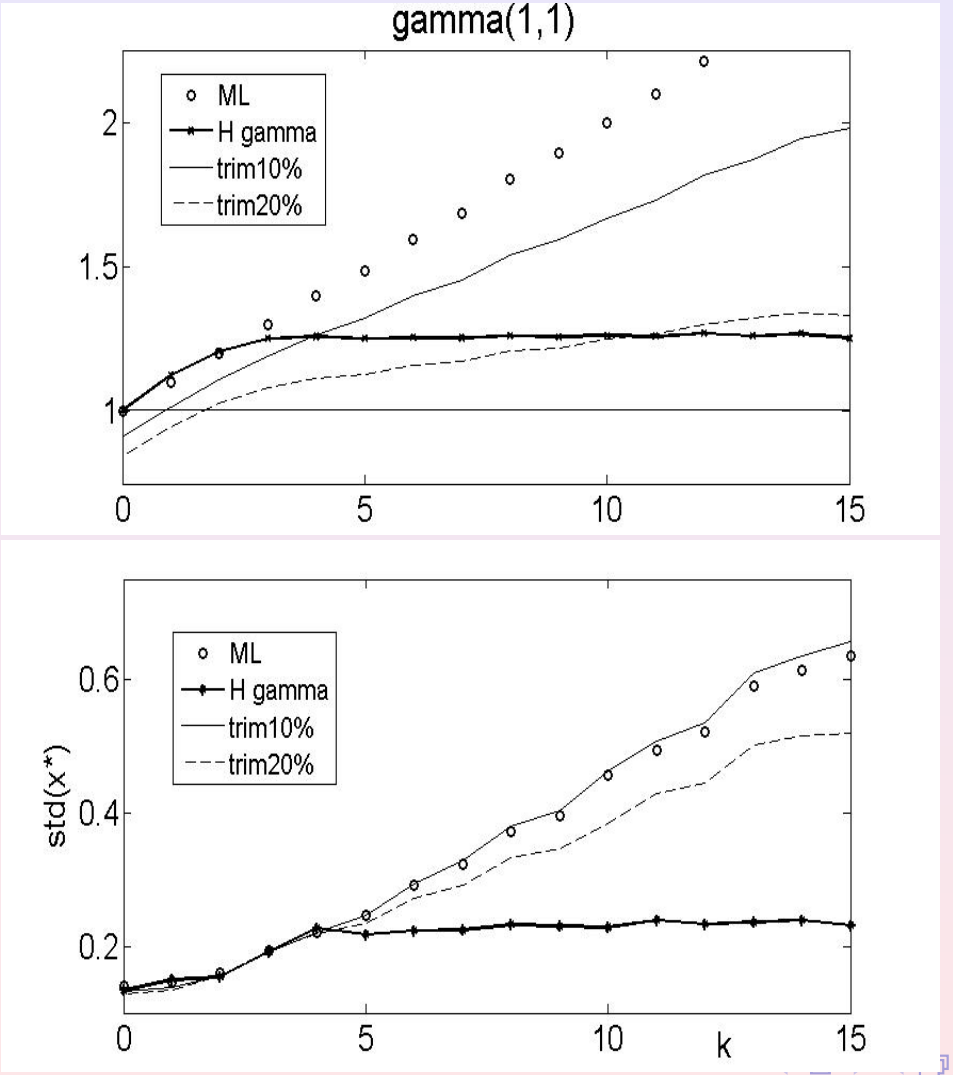
$f(x^*, \omega)$ na $(0, \infty)$

Useknutí v bodě $v = x_0 + k\omega_0$

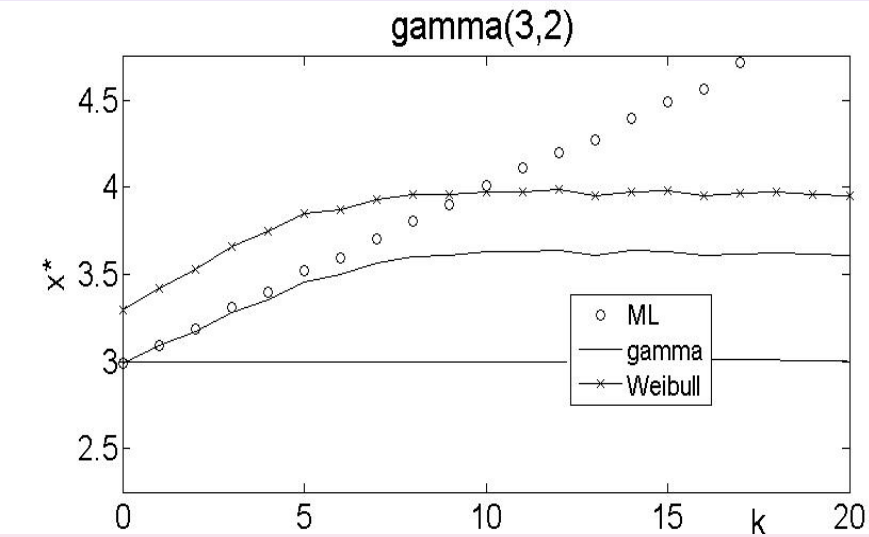
kde x_0 je median a $\omega_0 = MAD$

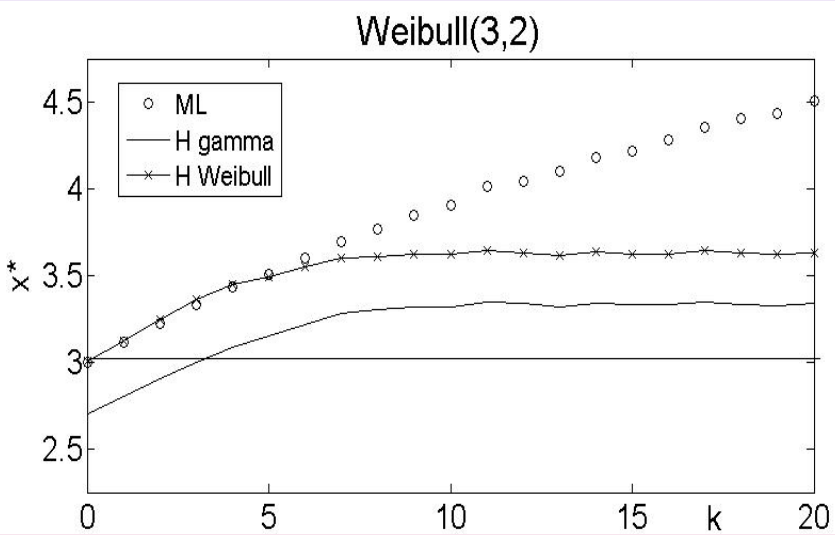
Kontaminace $f_{cont} = 0.9f(1,1) + 0.1f(1 + \mathbf{k}, 1)$

Porovnání s useknutým průměrem



Huberized gamma a Weibull





Huberized Weibull