Distance of observations

Zdeněk Fabián
Ústav informatiky AVČR Praha

ENBIS 2015

full title:

Statistical distance of observations based on the assumed model

Why statistical distance in energy considerations ?

It can be of interest by dividing the data into groups of 'similar' events. We show that the distances depend on the model and are often non-linear

This lecture is food for thought, based on rather non-traditional approach. Apart from preliminary published results, the whole account can be found in Z.Fabián: Score function of distribution and revival of the moment method, accepted 2013 in Communication in Statistics, but yet not appeared

The result

$\mathcal{X} \subseteq \mathbb{R}$ denotes an open interval. Let a continuous random variable X has support (sample space, the space on which is defined) $\mathcal{X}$, distribution function F and density $f(x) = dF(x)/dx$. A 'natural' statistical distance between two observations $x_1, x_2 \in \mathcal{X}$ from F is

$$d_F(x_1, x_2) = \omega |S_F(x_2) - S_F(x_1)|$$

where $S_F(x)$ is the **score function of distribution** of F and ω^2 the **score variance** of F.

Score function

Let 'statistical structure' has a 'center' ξ

$\psi(x)$ ... score function is a function describing the relative influence of observed $x \in \mathcal{X}$ to a construction of ξ

The estimator of ξ is based on the requirement of zero average of oriented distances of observed values to the 'center', measured by their relative influence, that is

$$\sum_{i=1}^{n} \psi(x_i - \xi) = 0$$

The distance of $x_1, x_2 \in \mathcal{X}$ is thus $\hat{d}(x_1, x_2) \sim |\psi(x_2) - \psi(x_1)|$

Score functions of classical statistics

Let F be the parent of parametric family $F_\theta(x), \theta \in \Theta \subseteq \mathbb{R}^m$. Function $u_F = (u_1, ..., u_m)$ where

$$u_j(x;\theta) = \frac{\partial}{\partial\theta_j} \log f(x;\theta)$$

is the likelihood score function (Fisher score) for θ_j

The well-known example: normal distribution

$$F_\theta = \mathcal{N}(\mu, 1) : \qquad u_F(x) = x - \mu$$

Bad news: A vector-valued function cannot be reasonably used for a definition of a distance

Score functions of robust statistics

Bounded $\psi(x)$

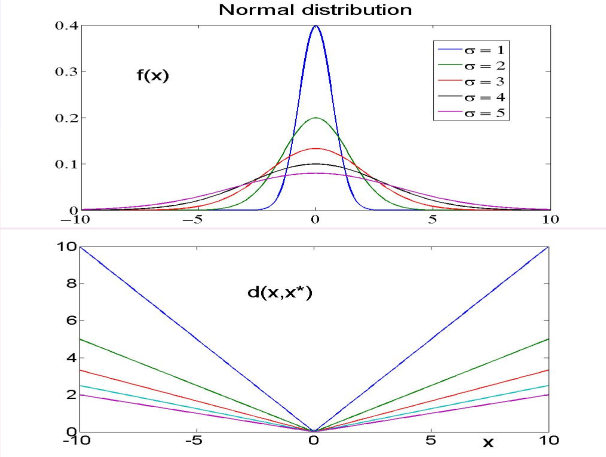
The well-known example: Huber's score function for contaminated normal distribution

$$\psi(x) = \begin{cases} -b & \text{if} \quad x-\xi < -b \\ x-\xi & \text{if} \quad |x-\xi| < b \\ b & \text{if} \quad x-\xi > b \end{cases}$$

Bad news: The assumed model F_θ need not be a location model. A choice of a bounded ψ usually means to resign an assumed model

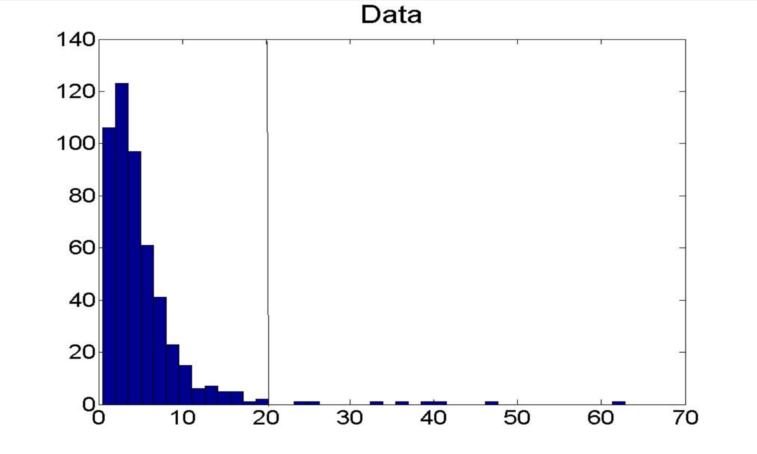
Statistical distance: Normal distribution $N(0,\sigma)$

$$u_N(x) = x; \omega = \sigma, d_F(x,0) = \sigma|S_F(x) - 0| = \sigma|x/\sigma^2| = |x|/\sigma$$



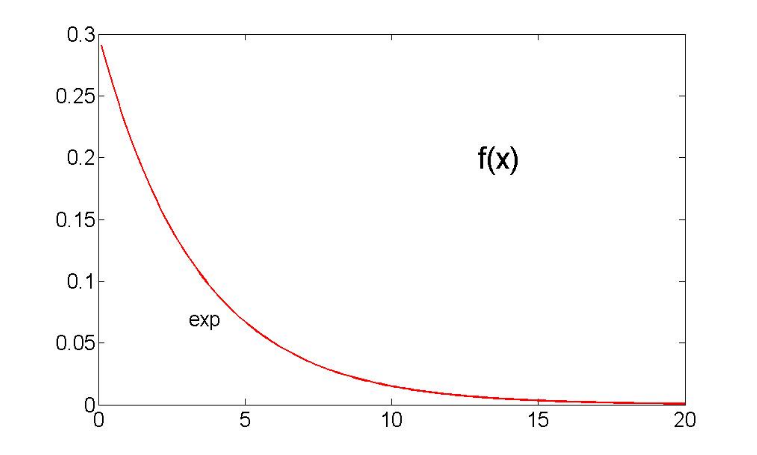
Data

observations $x_1, ..., x_n$

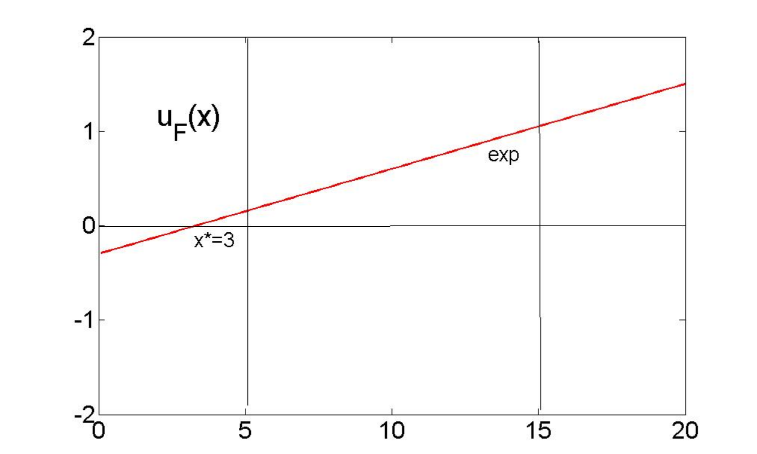


A parametric model: exponential

$$f(x) = \frac{1}{\tau} e^{-x/\tau}, \tau = 3$$



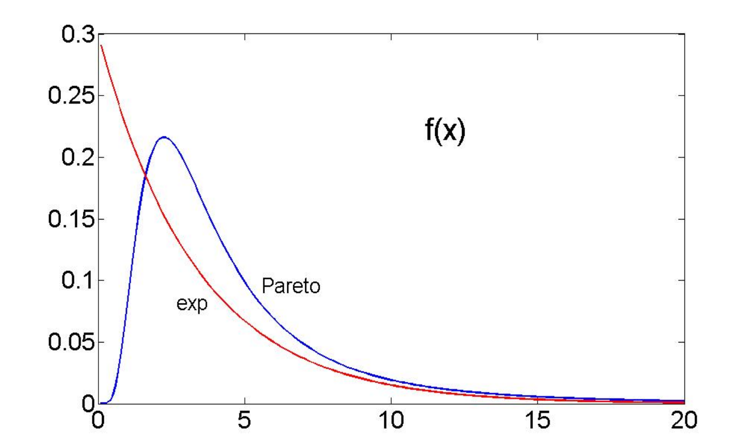
The Fisher score function for τ



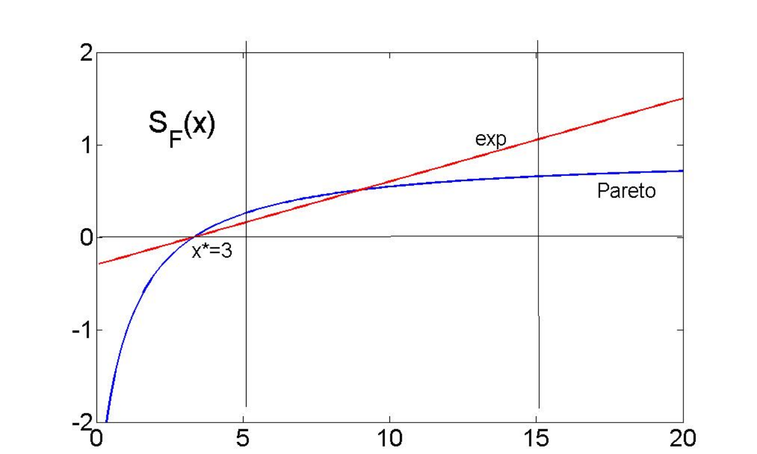
Model gen. Pareto

Distribution F is more probably the Pareto one

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



Score functions of distribution



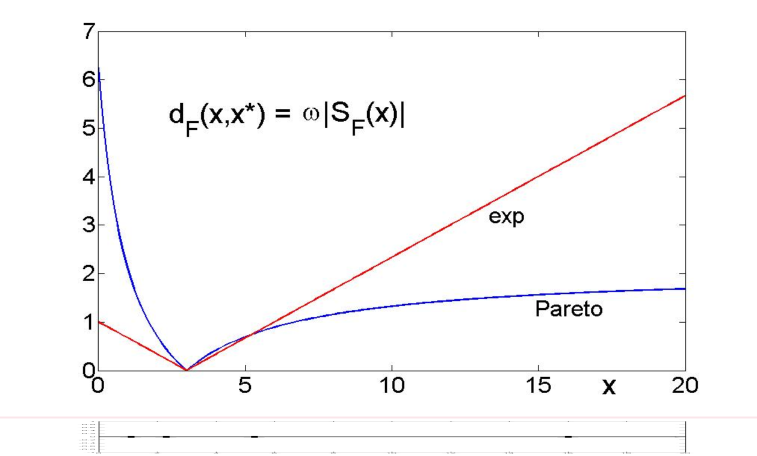
Central point of the distribution

The zero of the sfd, the solution x^* of the equation $S_F(x) = 0$ or, in a parametric case, the solution $x^* = x^*(\theta)$ of equation

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

expresses the typical value of the distribution (the central point in the geometry introduced in $\mathcal{X}$ by S_F), the score mean. It exists even in cases of heavy-tailed distributions with non-existing mean value

Distances from the central point



Score function of distribution (SFD)

My research was stimulated by lectures of P. Kovanic, that showed me, involuntarily, that there must be some scalar-valued score function yet not discovered in classical statistics

SFD, step I: Types of continuous distributions

There are three types distributions:

1) with $\mathcal{X} = \mathbb{R}$ and a 'simple' $f(x)$

2) with arbitrary $\mathcal{X}$ and density which can be decomposed into

$$f(x) = g(\eta(x))\eta'(x),$$

where g is some bell-like function with support $\mathbb{R}$ and $\eta : \mathcal{X} \to \mathbb{R}$ a differentiable strictly increasing function. They can be considered as transformed distributions with Jacobian $\psi'(x)$

3) with $\mathcal{X} \neq \mathbb{R}$ and a 'simple' $f(x)$

Type 2: Examples of transformed distributions

- The gen. Pareto with $\mathcal{X} = (0, \infty)$ has

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

$\eta(x) = \log x$

Type 2: Examples of transformed distributions

- The gen. Pareto with $\mathcal{X} = (0, \infty)$ has

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

$\eta(x) = \log x$

- The Burr V distribution with $\mathcal{X} = (-\pi/2, \pi/2)$ has

$$f(x) = \frac{e^{-\tan x}}{(1+e^{-\tan x})^2} \frac{1}{\cos^2 x}$$

$\eta(x) = \tan x$

Type 2: Examples of transformed distributions

- The gen. Pareto with $\mathcal{X} = (0, \infty)$ has

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

$\eta(x) = \log x$

- The Burr V distribution with $\mathcal{X} = (-\pi/2, \pi/2)$ has

$$f(x) = \frac{e^{-\tan x}}{(1+e^{-\tan x})^2} \frac{1}{\cos^2 x}$$

$\eta(x) = \tan x$

- The log-gamma distribution with $\mathcal{X} = (1, \infty)$ has

$$f(x) = \frac{c^\alpha}{\Gamma(\alpha)} (\log x)^{\alpha-1} \frac{1}{x^{c+1}} = \frac{c^\alpha}{\Gamma(\alpha)} (\log x)^{\alpha} \frac{1}{x^{c}} \frac{1}{x \log x}$$

$\eta(x) = \log \log x$

Type 1: Prototypes

Distributions with support $\mathbb{R}$ and densities in the form

$$f(x) = g(\eta(x))\eta'(x),$$

where $\eta(x) = x, \eta'(x) = 1$. The score function is known to be $S_F(x) = -g'(x)/g(x)$. Example: standard logistic distribution with density $f(x) = e^{-x}/(1 + e^{-x})^2$ and

$$S_F(x) = (e^x - 1)/(e^x + 1).$$

However, a distribution with support $\mathbb{R}$ and density

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{1 + x^2}} e^{-\frac{1}{2}(\sinh^{-1} x)^2}$$

is the standard normal prototype transformed by $\eta : \mathbb{R} \to \mathbb{R}$ in the form $\eta(x) = \sinh^{-1} x$.

Type 3: Problem

The density $f(x) = e^{-x}$ of the exponential distribution with $\mathcal{X} = (0, \infty)$ has no explicitly expressed Jacobian term. Undoubtedly, $\eta(x) = \log x$ and $f(x) = xe^{-x}\frac{1}{x}$

The truncated exponential distribution with $\mathcal{X} = (0, 1)$ and density $f(x) = be^{-\lambda x}$ and an arbitrary function with finite $\mathcal{X}$ integrable to 1. If we write formally

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

to obtain a density in a transformed form, is there a principle according to which can be chosen a 'favorable' $\eta : \mathcal{X} \to \mathbb{R}$?

Type 3: Our solution

Let us call mappings $\mathcal{X} \to \mathbb{R}$ given by

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

with an obvious generalization for a general support (a, b) the Johnson's mappings. The reason for assigning the corresponding Johnson mapping to a distribution with density without an explicitly expressed Jacobian term is the principle of parsimony: They are the simplest mappings, generating in the sample space the simplest distance. (Moreover, most of transformed distributions has Johnson's η)

SFD, step II: Definition

The density of all distributions with arbitrary support can be written in a transformed form

$$f(x) = g(\eta(x))\eta'(x)$$

The **score function of distribution** of F (Fabián, 2007) is

$$S_F(x) = -k\frac{1}{f(x)}\frac{d}{dx}[g(\eta(x))] \quad (1)$$

where k is a constant specified later

To obtain the score function of distribution, it is to differentiate the density without the Jacobian term. The explanation is that after decomposition of $f(x)$ into transformed form (1), the term $\eta'(x)$ does not contain any statistical information

The basic property of SFDs

Recall that x^*, the score mean, is the solution of equation $S_F(x, \theta) = 0$.

If F_θ, $\theta = (\theta_1, \theta_2, ...)$ has some $\theta_j = x^*$, then $S_F(x; \theta)$ with $k = \eta'(x^*)$ equals to the Fisher score for this parameter. **The score function of distribution is thus the (generalized) Fisher score for x^***

An example of distribution without a parameter equal to the score mean is the gen. Pareto or the gamma distribution with $\mathcal{X} = (0, \infty)$,

$$f(x) = \frac{\gamma^\alpha}{x\Gamma(\alpha)} x^\alpha e^{-\gamma x}$$

with $S_F(x) = k(\gamma x - \alpha)$ and $x^* = \alpha/\gamma$

Some other properties

- Score moments $ES_F^k(\theta)$ are finite, θ can be estimated from

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

S_F of heavy-tailed distributions are bounded

Some other properties

- Score moments $ES_F^k(\theta)$ are finite, θ can be estimated from

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

S_F of heavy-tailed distributions are bounded

- $ES_F^2(\theta)$ is the Fisher information for x^*. The characteristic of variability of F is the **score variance**

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

Example

Consider the heavy-tailed loglogistic distribution with $\mathcal{X} = (0, \infty)$ and

$$f(x) = \frac{c}{\tau} \frac{(x/\tau)^{c-1}}{[(x/\tau)^c + 1]^2} = c \frac{(x/\tau)^c}{[(x/\tau)^c + 1]^2} \frac{1}{x}$$

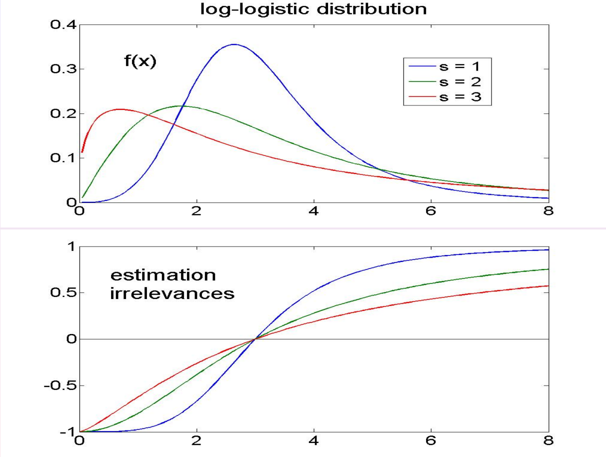
with score mean $x^* = \tau$ and $\omega = t/c$. The SFD is

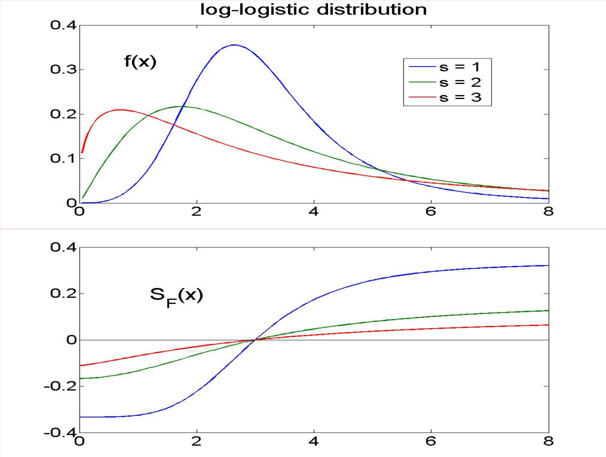
$$S_F(x) = -\frac{1}{\tau} \frac{d}{dx} [xf(x)] = \frac{c}{\tau} \frac{(x/\tau)^c - 1}{(x/\tau)^c + 1} \qquad (2)$$

Multiplying (2) by $(x/\tau)^{-c/2}$ and by setting $c = 4/s$, one obtains

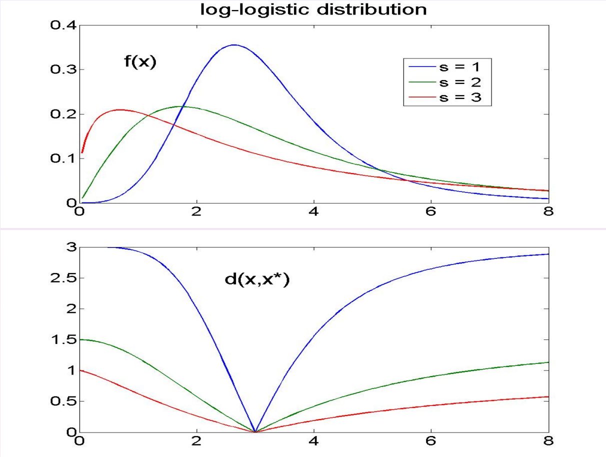
$$S_F(x) \sim \frac{(x/\tau)^{2/s} - (x/\tau)^{-2/s}}{(x/\tau)^{2/s} + (x/\tau)^{-2/s}}$$

which is Kovanic's score function called estimating irrelevance.


log-logistic distribution
f(x)
s = 1
s = 2
s = 3
0.4
0.3
0.2
0.1
0
0
2
4
6
8
estimation
irrelevances
1
0.5
0
-0.5
-1
0
2
4
6
8



log-logistic distribution
f(x)
s = 1
s = 2
s = 3
0.4
0.3
0.2
0.1
0
0
2
4
6
8
$S_F(x)$
0.4
0.2
0
-0.2
-0.4
0
2
4
6
8



Distance of observations

A relevant distance in the sample space

I. A 'small' data sample $(x_1, ..., x_n) \sim F_\theta$ with unknown θ

estimate $\hat{\theta}$ of θ
$\hat{x}^* = x^*(\hat{\theta}), \hat{\omega} = \omega(\hat{\theta})$,

$$d(x, x^*) = \hat{\omega}|S_F(x, \hat{\theta})|$$

II. A large data sample

estimate $\hat{f}(x)$ of $f(x)$ (histogram, kernel estimate),
using a numerical derivative of $\hat{f}(x)$ and computation of $\hat{S}_F(x)$
using the Johnson's $\eta(x)$ for the given support

$$d(x, x^*) \sim |\hat{S}_F(x_2)|$$

Thank you for attention

Thank you for attention