



Institute of Computer Science
Academy of Sciences of the Czech Republic

Geometry in the data space

Zdeněk Fabián

Technical report No. 998

May 2007



Institute of Computer Science
Academy of Sciences of the Czech Republic

Geometry in the data space ¹

Zdeněk Fabián

Technical report No. 998

May 2007

Abstract:

Geometry of probability distributions [1] continues by an introduction of new data characteristics, estimated in accordance with geometry of the assumed model, and by a short study of their properties.

Keywords:

Basic characteristics of distributions, basic characteristics of the data, statistics.

¹This work was supported by GA ASCR under grant No. 1ET 400300513 and the Institutional Research Plan AV0Z10300504. Prepared for conference ITAT2007.

1 Introduction

It was shown in [1] that any continuous probability distribution with arbitrary interval support $\mathcal{X} \in \mathbb{R}$ can be characterized, besides the distribution function $F(x)$ and density $f(x)$, by its Johnson score $S(x)$ ([1], Definition 1), information function $S^2(x)$ and weight function $S'(x) = dS(x)/dx$. Instead of the usual moments, we have the Johnson score moments

$$ES^k = \int_{\mathcal{X}} S^k(x)f(x) dx, \quad k = 1, 2, \dots \quad (1.1)$$

It was shown that $ES = 0$ and it can be shown that $ES^k < \infty$ if $ES^2 < \infty$. The last condition is equivalent to the usual regularity requirements. As a 'center' of the distribution can be taken the Johnson mean $x^* : S(x) = 0$ and as a measure of variability of values around x^* the Johnson variance $\omega^2 = (I^*)^{-1}$ where $I^* = ES^2$ is the mean information. If we introduce for $x_1, x_2 \in \mathcal{X}$ a Johnson difference

$$\tilde{d}(x_1, x_2) = \omega[S(x_2) - S(x_1)], \quad (1.2)$$

we obtain in the sample space \mathcal{X} a non-Euclidean Johnson distance $d(x_1, x_2) = |\tilde{d}(x_1, x_2)|$, which can be used for the testing of hypotheses and determination of confidence intervals.

Densities, Johnson scores, Johnson means and Johnson variances of distributions discussed in this paper are given for reference in Table 1. Apart from the normal distribution, the support of all other distributions is $\mathcal{X} = (0, \infty)$.

Table 1. Some distributions and their characteristics.

Distribution	$f(x)$	$S(x)$	x^*	ω^2
normal	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	$\frac{x-\mu}{\sigma^2}$	μ	σ^2
lognormal	$\frac{\beta}{\sqrt{2\pi x}} e^{-\frac{1}{2}\log^2\left(\frac{x}{t}\right)^\beta}$	$\frac{\beta}{t} \log(x/t)^\beta$	t	t^2/β^2
Weibull	$\frac{\beta}{x} \left(\frac{x}{t}\right)^\beta e^{-\left(\frac{x}{t}\right)^\beta}$	$\frac{\beta}{t} [(x/t)^\beta - 1]$	t	t^2/β^2
gamma	$\frac{\gamma^\alpha}{x\Gamma(\alpha)} x^\alpha e^{-\gamma x}$	$\gamma\left(\frac{x}{\alpha/\gamma} - 1\right)$	α/γ	α/γ^2
inv. gamma	$\frac{\gamma^\alpha}{x\Gamma(\alpha)} x^{-\alpha} e^{-\gamma/x}$	$\alpha\left(1 - \frac{\gamma/\alpha}{x}\right)$	γ/α	γ^2/α^3
beta-prime	$\frac{1}{xB(p,q)} \frac{x^p}{(x+1)^{p+q}}$	$\frac{q}{p} \frac{qx-p}{x+1}$	p/q	$\frac{p(p+q+1)}{q^3}$

Fig.1 shows densities and Johnson scores of Weibull distributions with $\beta = 1$ (exponential distribution), $\beta = 2$ (Rayleigh distribution) and $\beta = 3$ (Maxwell distribution). The densities are quickly decreasing to zero showing low probability of large observed values. Johnson scores are sensitive to large values; this sensitivity increases with increasing β . Johnson mean of all these distributions is $x^* = 1$. The usual means (denoted by stars) are near to x^* , and in the case $\beta = 1$ equal to x^* .

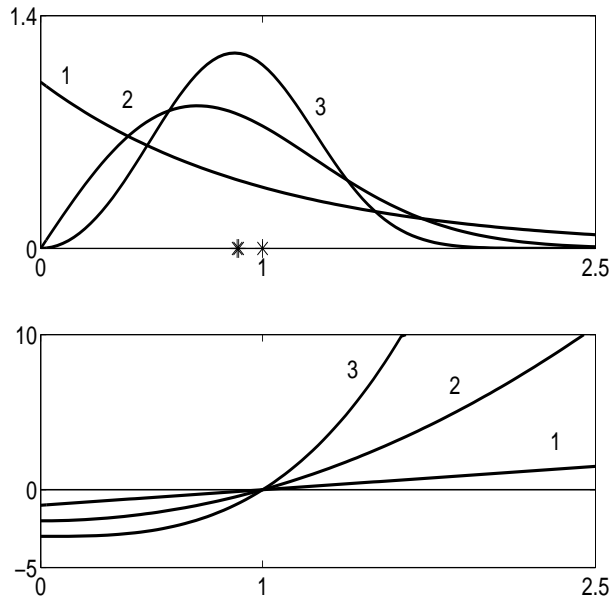


Figure 1. Densities and Johnson scores of Weibull distributions.

Fig.2 shows densities and Johnson scores of inverse gamma distributions with $\gamma = \alpha = 0.6(1), 1$ and $1.5(3)$. The densities decrease slowly to zero showing that in this case extremely large values can be observed. Johnson scores of this 'heavy-tailed' distributions are bounded in infinity, so that averages $\frac{1}{n} \sum S(x_i)$ containing large observed values are robust (the averages can be heavily influenced, however, by observed values near zero, which occur with low probability). The means of distributions denoted by 1 and 2 do not exist, the mean of distribution 3 is plotted by the star. The usual description of distributions by the mean and variance in this quite regular case fails. However, all three distributions have the same Johnson mean $x^* = 1$, which seems to give a reasonable description of the position of a distribution on the x -axis.

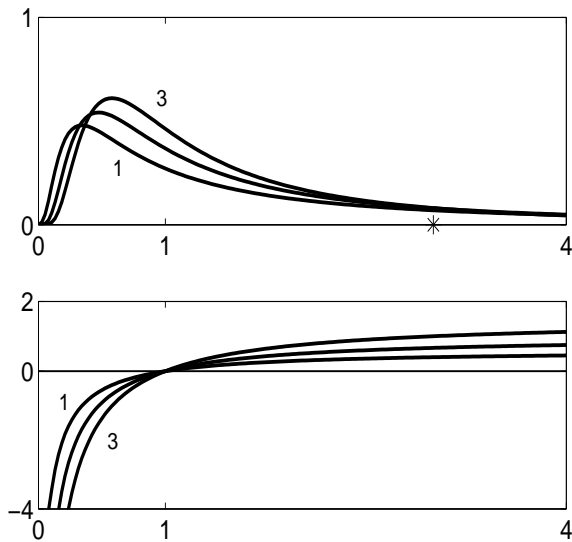


Figure 2. Densities and Johnson scores of inverse gamma distributions.

Further reasons for our conviction that they are the Johnson characteristics, which are to be used for the description of distributions instead of the usual mean and variance, can be found in [2].

2 Treatment of the data

Consider parametric family $\{F_\theta, \theta \in \Theta\}$ supported by $\mathcal{X} \subseteq \mathbb{R}$ with parameter $\theta = (\theta_1, \dots, \theta_m), \Theta \subseteq \mathbb{R}^m$. Let $\mathbf{X}_n = [x_1, \dots, x_n]$ be a random sample from F_{θ_0} (a realization of independent identically distributed according to F_{θ_0} random variables X_1, \dots, X_n) with unknown θ_0 . What can be said about θ_0 and how to characterize the data by a small number (two) values ?

A solution of this basic statistical problem consists of three steps:

i/ choosing an 'inference function' Q and treat the data \mathbf{X}_n as

$$\mathbf{Q}_n = [Q(x_1), \dots, Q(x_n)],$$

ii/ making some averages based on \mathbf{Q}_n ,

iii/ study the properties of the estimates.

The still used inference function of the approach which we will call the 'naive statistics' is the identity function $Q(x) = x$. The distance among data is thus Euclidean, the 'center' of the data is the sample mean and the dispersion of values around it the sample variance. However, their theoretical counterparts, the mean EX and variance $EX^2 - (EX)^2$, may not exist for some (heavy-tailed) distributions. In such cases (in Table 1: the inverse gamma and the beta-prime distribution) this approach does not offer any reasonable characteristics of the data.

The inference function of classical statistics is the vector of partial scores for parameters,

$$\mathbf{Q}(x) = \left(\frac{\partial}{\partial \theta_1} \log f(x; \theta), \dots, \frac{\partial}{\partial \theta_m} \log f(x; \theta) \right).$$

The maximum likelihood method uses \mathbf{Q} in a system of m equations for components of θ , giving the 'best' estimate of F_{θ_0} . However, the problem of simple characteristics of the 'center' and variability of the data still remains.

The inference function of the robust statistics is $Q(x) = \psi(x)$ where ψ (so called 'psi-function') is a suitable bounded function, suppressing the influence of large observations. The ψ -function prescribes a finite distance in the sample space, $d_R(x_1, x_2) = c|\psi(x_2) - \psi(x_1)|$ where $c = (E\psi')^{-1}$, and offers simple characteristics of the 'center' and variability of the data. The drawback of this approach is the lack of the connection of the ψ -function with properties of the assumed distribution F .

On the base of the account given in [1] we suggest to use as the inference function the Johnson score. Our 'treated' data are thus $[S(x_1), \dots, S(x_n)]$.

3 Basic Johnson characteristics of the data

Unlike the usual moments, the sample versions of Johnson score moments cannot be determined without an assumption about the underlying distribution family. On the other hand, by substituting the empirical distribution function into (1.1), the resulting system of equations

$$\frac{1}{n} \sum_{i=1}^n S^k(x_i; \theta) = ES^k(\theta), \quad k = 1, \dots, m \quad (3.1)$$

appears to be an alternative to the system of the maximum likelihood equations. The estimates $\hat{\theta}_n$ from (3.1) are shown (Fabián, 2001) to be asymptotically normally distributed with mean θ_0 and a certain variance σ^2 , i.e., $AN(\theta_0, \sigma^2)$. They can have slightly large variances than the maximum likelihood estimates, but they are robust 'if the situation demands it' (the heavy tailed distribution have bounded Johnson scores).

The first or the first two equations of system (3.1) give for particular distributions simple estimates of the Johnson mean or of both Johnson characteristics. Let us call the estimate \hat{x}_n^* and $\hat{\omega}_n^2$ of x^* and ω^2 based on observations x_1, \dots, x_n the *sample Johnson mean* and *sample Johnson variance*, respectively

For some particular distributions, the first equation of the system (3.1) can be written in the form

$$\sum_{i=1}^n S(x_i; \hat{x}_n^*) = 0. \quad (3.2)$$

Proposition 1 *Sample \hat{x}_n^* determined from (3.2) is $AN(x^*, \omega^2)$.*

Proof. Random variables $S(X)$ have zero mean $ES = 0$ and finite variance ES^2 so that \hat{x}_n^* is $AN(x^*, 1/ES^2)$ according to the Lindeberg-Lévy central limit theorem. $\omega^2 = 1/ES^2$ from the definition.

This is a nice result saying that the sample Johnson mean has normal distribution for any considered distribution, including distributions for which the Central limit theorem cannot be applied, and that its variance attains the Cramér-Rao lower bound.

Proposition 2 *$\sqrt{n}\tilde{d}(x^*, \hat{x}_n^*)$ is $AN(0, 1)$.*

Outline of the proof. The delta method theorem says that if q is $AN(q_0, \omega^2)$ then $\varphi(q)$ is $AN(\varphi(q_0), [\varphi'(q_0)]^2\omega^2)$. By this theorem and Proposition 1, $S(\hat{x}_n^*) - S(x^*)$ is $AN(0, [S'(x^*)]^2\omega^2)$. It can be shown that $ES' = ES^2$ so that $\sqrt{n}\tilde{d}(\hat{x}_n^*, x^*)$ is $AN(0, \omega^2(ES^2)^2\omega^2) = AN(0, 1)$.

This is another nice result saying that the approximate $(1 - \alpha)\%$ confidence intervals for the sample Johnson mean can be determined from a simple condition

$$\sqrt{n}|\tilde{d}(x^*, \hat{x}_n^*)| \leq u_{\alpha/2}, \quad (3.3)$$

where \tilde{d} is given by (1.2) and $u_{\alpha/2}$ is the $(\alpha/2)$ -th quantile of the normal distribution ($u_{\alpha/2} = 1.96$ for $\alpha = 5$).

Definition 1 *Let X, Y be random variables supported by \mathcal{X} and \mathcal{Y} , respectively, with joint distribution F , marginal distributions with Johnson scores S_X, S_Y and Johnson information I_X^*, I_Y^* . Let f be the joint density of (X, Y) . Value*

$$i_{XY}^* = \frac{1}{\sqrt{I_X^* I_Y^*}} \int_{\mathcal{X}} \int_{\mathcal{Y}} S_X(x) S_Y(y) f(x, y) dx dy$$

will be called a Johnson mutual information of X and Y .

Obviously, $|i_{XY}^*| \leq 1$ according the Cauchy-Schwartz inequality. Having sample $(x_i, y_i), i = 1, \dots, n$, taken from (X, Y) , the *sample Johnson mutual information* is

$$\hat{i}_{XY}^* = \frac{\sum_{i=1}^n S_X(x_i) S_Y(y_i)}{\left(\sum_{i=1}^n S_X^2(x_i) \sum_{i=1}^n S_Y^2(y_i) \right)^{1/2}}. \quad (3.4)$$

(3.4) can serve as an empirical measure of the association between X and Y .

4 Examples

In this section we show examples of statistical procedures which take into account the particular geometry in the sample space of the assumed distribution.

Normal distribution. Johnson score of the normal distribution is $S(x) = \frac{x-\mu}{\sigma^2}, x^* = \mu$ and $I^* = ES^2 = 1/\sigma^2$. The sample Johnson mean and sample Johnson variance are the usual mean and variance; other statistics are the usual statistics. The Johnson mutual information is the usual correlation coefficient.

On the other hand, from our point of view, the use of the data without 'treatment' is equivalent to the implicit assumption of the normal distribution.

In the rest of this section we denote $\lambda_n = 1.96/\sqrt{n}$.

Lognormal distribution. The first two equations (3.1) are

$$\begin{aligned}\beta \sum_{i=1}^n \log \left(\frac{x_i}{t} \right) &= 0 \\ \beta^2 \sum_{i=1}^n \log^2 \left(\frac{x_i}{t} \right) &= 1\end{aligned}$$

from which $\hat{x}_n^* = \hat{t}_n = \frac{1}{n} \sum_{i=1}^n \log x_i$, $\hat{\beta}_n^2 = n / \sum_{i=1}^n \log^2(x_i/\hat{t}_n)$ and $(\hat{\omega}_n)^2 = \hat{t}_n^2 / \hat{\beta}_n^2$. Since by (1.2) $\tilde{d}(x^*, \hat{x}_n^*) = \beta \log(x^*/\hat{x}_n^*)$, the 95% confidence interval for the Johnson mean is, according to (3.3),

$$\hat{x}_n^* e^{-\lambda_n / \hat{\beta}_n} \leq x^* \leq \hat{x}_n^* e^{\lambda_n / \hat{\beta}_n}.$$

Weibull distribution. The first two equations (3.1) are

$$\begin{aligned}\sum_{i=1}^n [(x_i/t)^\beta - 1] &= 0 \\ \sum_{i=1}^n [(x_i/t)^\beta - 1]^2 &= 1\end{aligned}$$

which are to be solved iteratively. For a constant β , the sample Johnson mean $\hat{x}_n^* = \hat{t}_n = (n^{-1} \sum_{i=1}^n x_i^\beta)^{1/\beta}$ is the β -th mean. By (1.2), $\tilde{d}(x^*, \hat{x}_n^*) = (x^*/\hat{x}_n^*)^{\hat{\beta}_n} - 1$ so that the 95% confidence interval for x^* is

$$\hat{x}_n^* (1 - \lambda_n)^{1/\hat{\beta}_n} \leq x^* \leq \hat{x}_n^* (1 + \lambda_n)^{1/\hat{\beta}_n}.$$

Gamma distribution. The first two equations (3.1) are

$$\begin{aligned}\sum_{i=1}^n (\gamma x_i - \alpha) &= 0 \\ \sum_{i=1}^n (\gamma x_i - \alpha)^2 &= n\alpha\end{aligned}$$

from which $\hat{x}_n^* = \alpha/\gamma = n^{-1} \sum_{i=1}^n x_i = \bar{x}$ and $\hat{\omega}_n^2 = \alpha/\gamma^2 = n^{-1} \sum_{i=1}^n x_i^2 - \bar{x}^2$. Johnson mean and Johnson variance are thus equal to the normal mean and normal variance. Since $\tilde{d}(x^*, \hat{x}_n^*) = \sqrt{\alpha}(x^*/\hat{x}_n^*) - 1$ and $\sqrt{\alpha} = \bar{x}/\hat{\omega}_n$, the 95% confidence interval for x^* is

$$\bar{x} - \lambda_n \omega_n \leq x^* \leq \bar{x} + \lambda_n \omega_n.$$

For distribution with linear Johnson score we obtained the usual symmetrical confidence interval.

Inverse gamma distribution. The first two equations (3.1) are

$$\begin{aligned}\sum_{i=1}^n (\alpha - \gamma/x_i) &= 0 \\ \sum_{i=1}^n (\alpha - \gamma/x_i)^2 &= n\alpha\end{aligned}$$

from which $\hat{x}_n^* = \gamma/\alpha = n/\sum_{i=1}^n 1/x_i = \bar{x}_H$, which is the harmonic mean, and

$$\hat{\omega}_n^2 = \bar{x}_H^2 \frac{\bar{x}_H^2 - \bar{x}_{2H}}{\bar{x}_{2H}}$$

where $\bar{x}_{2H} = n/\sum_{i=1}^n 1/x_i^2$. Since $\tilde{d}(x^*, \hat{x}_n^*) = \sqrt{\alpha}(1 - \bar{x}_H/x^*)$ and $\sqrt{\alpha} = \bar{x}_H \hat{\omega}_n$, the 95% confidence interval for x^* is

$$\frac{\bar{x}_H}{1 + \lambda_n/\bar{x}_H \hat{\omega}_n} \leq x^* \leq \frac{\bar{x}_H}{1 - \lambda_n/\bar{x}_H \hat{\omega}_n}.$$

Beta-prime distribution. The first two equations (3.1) are

$$\begin{aligned} \sum_{i=1}^n \frac{qx_i - p}{x_i + 1} &= 0 \\ \sum_{i=1}^n \left(\frac{qx_i - p}{x_i + 1} \right)^2 &= \frac{pq}{p + q + 1}. \end{aligned} \quad (4.1)$$

As $x^* = p/q$, from the first equation we obtain

$$\hat{x}_n^* = \frac{\sum_{i=1}^n \frac{x_i}{1 + x_i}}{\sum_{i=1}^n \frac{1}{1 + x_i}}. \quad (4.2)$$

Multiplying (4.1) by $1/pq$, substituting $p = \hat{x}_n q$ and using formula for ω^2 from Table 1, we have

$$\hat{\omega}_n^2 = \frac{\hat{\rho}_n \hat{x}_n^* (1 + \hat{x}_n^*)^2}{(\hat{\rho}_n - 1)^2}$$

where \hat{x}_n^* is given by (4.2) and

$$\frac{n}{\hat{\rho}_n} = \frac{1}{\hat{x}_n^*} \sum_{i=1}^n \frac{x_i^2}{(x_i + 1)^2} - 2 \sum_{i=1}^n \frac{x_i}{(x_i + 1)^2} + \hat{x}_n^* \sum_{i=1}^n \frac{1}{(x_i + 1)^2}.$$

Condition (3.3) is

$$q\hat{\omega}_n \left| \frac{x^* - \hat{x}_n^*}{x^* + 1} \right| \leq \lambda_n$$

so that the 95% confidence interval for x^* is

$$\frac{\hat{x}_n - \hat{\tau}_n}{1 + \hat{\tau}_n} \leq x^* \leq \frac{\hat{x}_n - \hat{\tau}_n}{1 - \hat{\tau}_n}$$

where $\hat{\tau}_n = \lambda_n/\hat{q}\hat{\omega}_n$ and \hat{q} is to be determined from the system $\hat{x}_n^* = p/q, \hat{\omega}_n^2 = p(p + q + 1)/q^3$. For example, if $\hat{p} = \hat{q} = 1$, the 95% confidence interval for $\hat{x}_n^* = 1$ and $n = 50$ is (0.72, 1.38).

5 Simulations

Example 1. The sample Johnson mean and sample Johnson deviance (the square root of variance) of samples of length 50 generated from distributions listed in the first column of Table 1 and with parameters determined by values $x^* = 1$ and $\omega = 1.118$ were determined for various assumed families listed in the first row of Table 2. The presented values are the average values after 5000 experiments.

Table 2. Comparison of the estimated Johnson characteristics.

\hat{i}_{5000}^*	gamma	Weibull	lognorm.	beta-pr.	inv.gam
gamma	1.000	0.94	0.60	0.49	0.12
Weibull	1.06	1.005	0.64	0.53	0.15
lognormal	1.66	1.66	1.010	1.01	0.63
beta-prime	2.00	1.77	1.01	1.008	0.54
inv.gamma	84.4	4.71	1.70	2.13	1.022
$\hat{\omega}_{5000}$					
gamma	1.094	1.06	0.81	0.72	0.31
Weibull	1.17	1.108	0.83	0.75	0.39
lognormal	2.04	1.62	1.082	1.09	0.74
beta-prime	3.52	2.00	1.11	1.113	0.82
inv.gamma	187.	8.52	2.32	3.23	1.117

It is apparent from Table 2 that erroneous assumptions often lead to unacceptable estimates (note, however, the similar results obtained under assumptions of the lognormal and beta-prime distributions). Estimating the Johnson mean and Johnson variance, it is easy to compare mean characteristics of the data from distributions parametrized by arbitrary ways.

Example 2. In the left part of Fig.3 we plot samples $(x_i, y_i)_{i=1, \dots, 12}$ from random vector (X, Y) , where $Y = 0.35X + 0.65Z$ and where X and Z are independent random variables with inverse gamma distribution. In the right part are the corresponding samples $[S_X(x_i; \hat{\theta}_X), S_Y(y_i; \hat{\theta}_Y)]$ computed under the right assumption. $\hat{\theta}_X, \hat{\theta}_Y$ are the estimated values of the parameters.

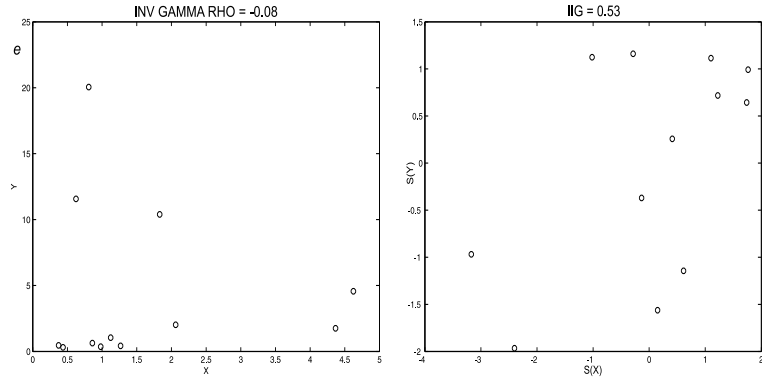


Figure 3.

Making different assumptions on underlying marginal distributions we obtained the following values of \hat{i}_{XY}^* :

f_X, f_Y	gamma	Weibull	lognormal	beta-prime	inv.gamma
\hat{i}_{XY}^*	-0.08	-0.01	0.29	0.40	0.53

It is apparent that for the estimation of the degree of the association of random variables, the assumption on the underlying distribution is substantial.

Acknowledgements. The work was supported by GA ASCR under grant No.1ET 400300513.

References

- Pawitan Y. (2001): In all likelihood. Oxford, Clarendon press.
- Fabián Z. (2006): Geometry of probabilistic models. Sb. ITAT 2006, 35-40.
- Fabián Z. (2008): New measures of central tendency and variability of continuous distributions. To appear in *Commun. in Statist.-Theory Meth.*, 2.