A new concept of quantiles for directional data and the angular Mahalanobis depth

C. Ley, C. Sabbah and T. Verdebout

University of Lille

May 21, 2014

Spherical data

Directional data or spherical data are multivariate data for which we are interested in their direction.

A spherical random variable X in $\mathbb{R}^k$ is a vector

$$X = (X_1, \cdots, X_k)', \text{ such that } \sum_{i=1}^{k} X_i^2 = 1.$$

A spherical random variable takes its values on the unit sphere

$$\mathcal{S}^{k-1} = \left\{ v \in \mathbb{R}^k, \; v'v = 1 \right\}, \; k \geq 2.$$

Spherical data application for $k = 2$ and $k = 3$ (resp. circle in $\mathbb{R}^2$ and sphere in $\mathbb{R}^3$) in e.g. meteorology, biology, neurosciences, oceanography.

Quantiles and depth on spheres

Despite an extensive literature in recent years (see Kong and Mizera (2012) and Hallin, Paindaveine and Siman (2010)) , no conclusive definition of quantiles in multivariate and directional data has emerged.

Koenker (2005) : *The search for a satisfactory notion of multivariate quantiles has become something of a quest for the statistical holy grail in recent years ... It is fair to say that no general agreement has emerged.*

However a related to the quantiles notion is the *depth*. The notion of depth gives an ordering for multivariate data w.r.t. a center point. One famous depth function is the angular Tukey depth that can be summarized as follows

$$ATD_f(\theta) = \inf_{S:\theta\in S} P_f(S),$$

where the infimum is taken over all closed hemispheres containing θ and P denotes a distribution on $\mathcal{S}^{k-1}$.

Quantiles and depth on spheres

Unfortunately, depth functions suffers for some drawbacks

1. Computationally heavy
2. Hard to base inference as no asymptotic representations exists
3. Whatever f is considered, there always exists a hemisphere with constant minimal depth

Outline

1. The new concept of quantiles

2. Applications of the new quantiles

3. Quantile-driven analysis of a cosmic rays data set

Outline

1. The new concept of quantiles

2. Applications of the new quantiles

3. Quantile-driven analysis of a cosmic rays data set

Assumptions

Before defining our new concept of quantiles, we introduce two Assumptions on the distributions we hereafter consider.

Assumption $\mathcal{F}$

The distribution of X belongs to the class $\mathcal{F}$ of probability laws on $\mathcal{S}^{k-1}$ with bounded density and which admit a unique median direction θ_m.

Assumption $\mathcal{R}$

The distribution of X belongs to the class $\mathcal{R}$ of rotationally symmetric distributions on $\mathcal{S}^{k-1}$ with bounded density and which admit a unique median direction θ_m.

New concept of quantiles on spheres

Consider X, a spherical random variable with density f in $\mathcal{F}$ and mode θ_m.
Let c_τ (the projection quantile) be the τ-quantile of the univariate r.v. $X'\theta_m$ and define

$$\begin{aligned} \mathcal{C}_\tau^+ &= \left\{x \in \mathcal{S}^{k-1} \ / \ x'\theta_m \geq c_\tau\right\}, \\ \mathcal{C}_\tau^- &= \left\{x \in \mathcal{S}^{k-1} \ / \ x'\theta_m < c_\tau\right\}. \end{aligned}$$

Observe that

$$\int_{\mathcal{C}_\tau^-} f(x)dx = \tau,$$

formula that parallels

$$\int_{y<q_\tau} f_Y(y)dy = \tau,$$

where Y is some univariate random variable with density f_Y and τ-quantile q_τ.

Angular Mahanalobis Depth

Consider the depth function

$$D_f(x) = \arg \min_{\tau \in [0,1]} \left\{ c_\tau \geq x' \theta_m \right\},$$

which fulfill the following requirements (see Zuo and Serfling (2000))

1. rotationally invariant
2. maximal at $x = \theta_m$
3. decreasing along each great semi-circle form θ_m to $-\theta_m$
4. $D_f(-\theta_m) = 0$

We then define the angular Mahanalobis depth ($AMHD_f$) as

$$AMHD_f(x) = \frac{D_f(x)}{1 + D_f(x)} = \frac{1}{1 + \frac{1}{D_f(x)}}, \quad x \in \mathcal{S}^{k-1}.$$

The $AMHD_f$ satisfies the four properties aforementioned together with $AMHD_f(\theta_m) = 1/2$ (as in the classical depth definitions).

Empirical aspects

Let $X_1, \cdots, X_n$, be i.i.d. observations on $\mathcal{S}^{k-1}$. Our estimators of the quantile caps $\mathcal{C}_\tau^+$ and $\mathcal{C}_\tau^-$ are respectively

$$\begin{aligned}\widehat{\mathcal{C}}_\tau^+ &= \left\{x \in \mathcal{S}^{k-1} \ / \ x'\widehat{\theta}_m \geq \widehat{c}_\tau\right\},\\ \widehat{\mathcal{C}}_\tau^- &= \left\{x \in \mathcal{S}^{k-1} \ / \ x'\widehat{\theta}_m < \widehat{c}_\tau\right\}.\end{aligned}$$

where $\widehat{\theta}_m$ is any root-n consistent estimator of θ_m and $\widehat{c}_\tau$ is the empirical quantile of the (non i.i.d.) sequence

$$X_1'\widehat{\theta}_m, \cdots, X_n'\widehat{\theta}_m.$$

More precisely let

$$\widehat{c}_\tau = \arg \min_{c \in [-1,1]} \sum_{i=1}^{n} \ell_\tau\left(X_i'\widehat{\theta}_m - c\right)$$

the estimator of the projection quantile

$$c_\tau = \arg \min_{c \in [-1,1]} \mathbb{E}\left[\ell_\tau\left(X'\theta_m - c\right)\right]$$

where $\ell_\tau(t) = t(\tau - \mathbb{I}(t \leq 0))$ is the classical quantile *check function*.

Empirical Results

Let $f_{proj}(\cdot)$ be the density of the univariate r.v. $X'\theta_m$

Proposition 1: Bahadur representation

Under Assumption $\mathcal{F}$, there exists a k-vector Γ_{θ_m,c_τ} such that

$$n^{1/2}\left(\widehat{c}_\tau - c_\tau\right) = \frac{n^{-1/2}}{f_{proj}(c_\tau)} \sum_{i=1}^{n} \left(\tau - \mathbb{I}\left[X_i'\theta_m \leq c_\tau\right]\right) - \frac{\Gamma'_{\theta_m,c_\tau}}{f_{proj}(c_\tau)} n^{1/2} \left(\widehat{\theta}_m - \theta_m\right) + o_P(1)$$

as $n \to \infty$.

Proposition 2: Rotationally symmetric case

Under Assumption $\mathcal{R}$

$$n^{1/2}\left(\widehat{c}_\tau - c_\tau\right) = \frac{n^{-1/2}}{f_{proj}(c_\tau)} \sum_{i=1}^{n} \left(\tau - \mathbb{I}\left[X_i'\theta_m \leq c_\tau\right]\right) + o_P(1)$$

as $n \to \infty$.

sketch of proofs : Bahadur representation for non-smooth objective functions (Proposition 1)

We want to estimate $\theta = \arg\min_t \mathbb{E}[m(X,t)]$ given a i.i.d. sample $X_1, \ldots, X_n$.

Consider

- $M_n(t) = \sum_{i=1}^n m(X_i, t)$
- $\widehat{\theta} = \arg\min M_n\left(t\right)$
- The FOC $M_n^{(1)}(\widehat{\theta}) = 0$
- suppose that $M_n^{(2)}$ does not exist. Instead $\mu\left(t\right) = \mathbb{E}[M_n^{(1)}\left(t\right)/n]$ is differentiable.

Then

$$\begin{aligned} 0 &= \frac{M_n^{(1)}\left(\widehat{\theta}\right)}{\sqrt{n}} = \frac{M_n^{(1)}(\theta)}{\sqrt{n}} + \frac{M_n^{(1)}\left(\widehat{\theta}\right) - M_n^{(1)}(\theta)}{\sqrt{n}} \\ &= \frac{M_n^{(1)}(\theta)}{\sqrt{n}} + \sqrt{n}\left(\mu\left(\widehat{\theta}\right) - \mu(\theta)\right) \\ &\quad + \frac{M_n^{(1)}\left(\widehat{\theta}\right) - M_n^{(1)}(\theta) - n\left(\mu\left(\widehat{\theta}\right) - \mu(\theta)\right)}{\sqrt{n}}. \end{aligned}$$

In a preliminary step suppose that for some $v \in (0, 1]$

$$\sup_{\Delta}\left\{\frac{1}{|\Delta|^{\nu}} \frac{\left|M_n^{(1)}(\theta+\Delta) - M_n^{(1)}(\theta) - n(\mu(\theta+\Delta) - \mu(\theta))\right|}{\sqrt{n}}\right\} = O_{\mathbb{P}}(1).$$

Then, it can be shown that $\widehat{\theta} - \theta = O_{\mathbb{P}}\left(n^{-1/2}\right)$, so that

$$0 = \frac{M_n^{(1)}(\theta)}{\sqrt{n}} + \sqrt{n}\mu^{(1)}(\theta)\left(\widehat{\theta} - \theta\right) + O_{\mathbb{P}}(n^{-1/2}) + O_{\mathbb{P}}\left(\left(\widehat{\theta} - \theta\right)^{\nu}\right).$$

This gives

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) = -\frac{1}{\mu^{(1)}(\theta)}\frac{M_n^{(1)}(\theta)}{\sqrt{n}} + \underbrace{O_{\mathbb{P}}\left(n^{-\nu/2} + n^{-1/2}\right)}_{\text{Bahadur term}}.$$

For quantile estimation typically $\nu = 1/2$ which gives a $O_{\mathbb{P}}\left(n^{-1/4}\right)$ Bahadur term. This is done in the context of non-i.i.d. variables and where we use the Kreiss approximation for discretized estimators in order to deal with the estimator $\widehat{\theta}_m$.

sketch of proof of Proposition 2

Proof of Proposition 2 builds on two facts

1. $\widehat{\theta}_m - \theta_m = (I_k - \theta_m \theta_m') \left(\widehat{\theta}_m - \theta_m \right) + o_P(n^{-1/2})$.
2. there exists a constant $\gamma_{\theta_m, c_\tau}$ such that $\Gamma_{\theta_m, c_\tau} = \gamma_{\theta_m, c_\tau} \theta_m$ (Watson's (1983) decomposition).

   The combination of these two facts yields that

$$\Gamma'_{\theta_m, c_\tau} n^{1/2} \left(\widehat{\theta}_m - \theta_m \right) = o_P(1),$$

   which is the desired result.

Outline

1 The new concept of quantiles

2 Applications of the new quantiles

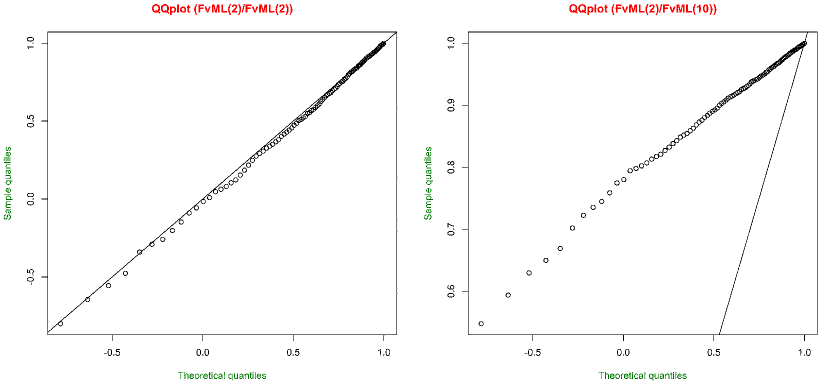
3 Quantile-driven analysis of a cosmic rays data set

Descriptive statistics

We give the values of the deciles $c_{j/10}$, $j = 1, \ldots, 9$ for some classical distributions with different concentration parameter. This gives an alternative to the single value of the resultant length $R = (\mathbb{E}[X]'\mathbb{E}[X])^{1/2}$.

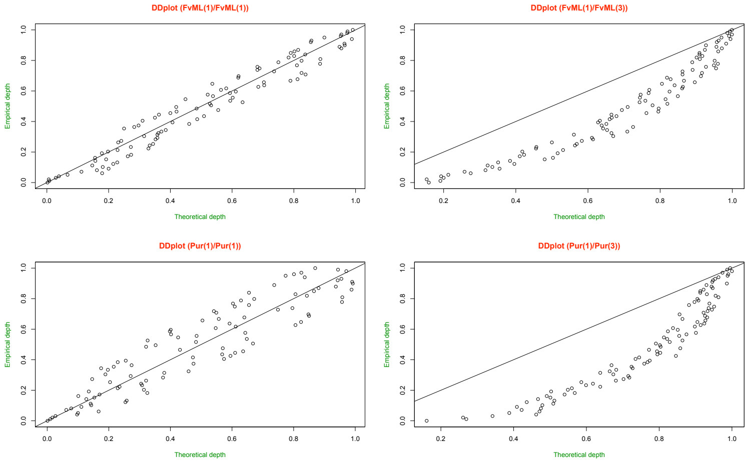
Density	$c_{1/10}$	$c_{2/10}$	$c_{3/10}$	$c_{4/10}$	$c_{5/10}$	$c_{6/10}$	$c_{7/10}$	$c_{8/10}$	$c_{9/10}$
$f_{\mathrm{FvML}(1)}$	-.5059	-.1767	.0705	.2686	.4338	.5756	.6998	.8102	.9096
$f_{\mathrm{FvML}(2)}$	-.0750	.2307	.4190	.5555	.6626	.7507	.8256	.8908	.9484
$f_{\mathrm{FvML}(5)}$	.5396	.6782	.7593	.8168	.8614	.8979	.9287	.9554	.9790
$f_{\mathrm{FvML}(10)}$	.7698	.8391	.8797	.9084	.9307	.9490	.9644	.9777	.9895
$f_{\mathrm{lin}(2)}$	-.6583	-.3875	-.1560	.0494	.2361	.4084	.5691	.7203	.8636
$f_{\mathrm{lin}(5)}$	-.7573	-.5278	-.3095	-.1010	.0991	.2916	.4773	.6569	.8310
$f_{\mathrm{lin}(10)}$	-.7804	-.5660	-.3563	-.1511	.0499	.2470	.4404	.6302	.8167
$f_{\mathrm{Pur}(1)}$	-.4373	-.1078	.1386	.3358	.4986	.6356	.7519	.8507	.9337
$f_{\mathrm{Pur}(5)}$	.7359	.8386	.8911	.9243	.9474	.9644	.9772	.9870	.9946

QQ-Plots



QQ-plots (theoretical quantiles versus sample quantiles) using theoretical $FvML(2)$ quantiles in the plots. In each case, we generated a sample of 1000 observations from two distributions: $FvML(2)$ and $FvML(10)$.

DD-Plots



DDplot (FvML(1)/FvML(1))

DDplot (FvML(1)/FvML(3))

DDplot (Pur(1)/Pur(1))

DDplot (Pur(1)/Pur(3))

DD-Plots defined as in Liu et al. (1999). The left DD-Plots should look like a homoskedastic white noise around the same 45-degree line while the right DD-Plots should show a clear departure from the homoskedastic white noise situation.

Goodness-of-fit test

Our quantiles are well-suited to Goodness-of-fit test. We want to test $\mathcal{H}_0 : f = f_0$ against $\mathcal{H}_1 : f \neq f_0$ for some rotationally symmetric f_0. Proposition 2 yields that

$$T_{\tau}^{(n)} := n^{1/2} \begin{pmatrix} \widehat{c}_{\tau_1} - c^0_{\tau_1} \\ \vdots \\ \widehat{c}_{\tau_m} - c^0_{\tau_m} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma\right)$$

with

$$\Sigma_{i,j} = \frac{\min(\tau_i, \tau_j) - \tau_i, \tau_j}{f_{proj}(c_{\tau_i}) f_{proj}(c_{\tau_j})}.$$

A Goodness-of-fit test can be based on the statistic

$$Q_{\tau}^{(n)} := \left(T_{\tau}^{(n)}\right)' \Sigma^{-1} T_{\tau}^{(n)}$$

which converges in distribution to a chi-square distribution with m degrees of freedom.

Outline

1. The new concept of quantiles

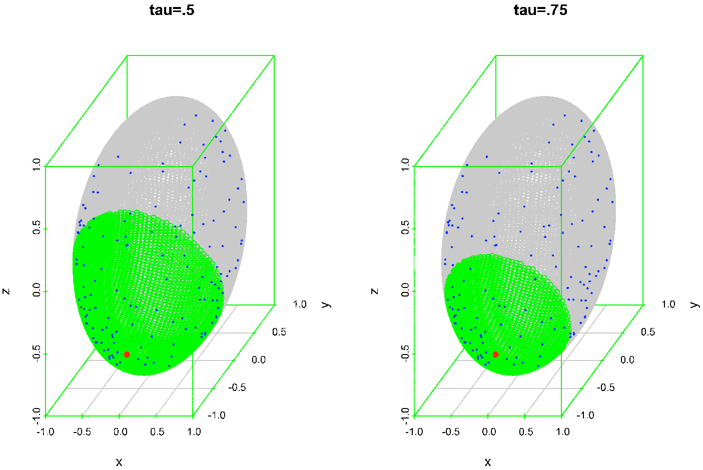
2. Applications of the new quantiles

3. Quantile-driven analysis of a cosmic rays data set

Analysis of a comic ray data

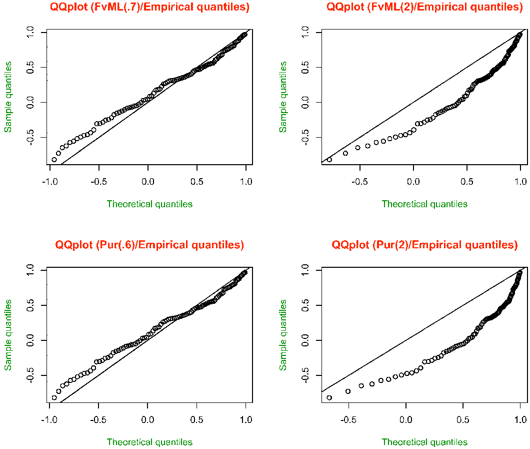
The data was first used in Toyoda et al. (1965) in order to study primary cosmic rays in certain energy regions. The sample size is 148.
We will also apply our Goodness -of-fit test to that sample in order to say which distribution described at best the data.

Visual inspection



The zone in grey are the τ (equal to .5 and .75) empirical upper quantile caps. The red point is the Fisher (1985) empirical median.

QQ-Plot exploration



QQ-Plots of the quantiles of various FvML and Purkayastha distributions versus the sample quantiles of the cosmic rays data.

Formal goodness-of-fit tests

Density	*p*-value	Density	*p*-value
$f_{\mathrm{FvML}(.3)}$	.00135	$f_{\mathrm{Pur}(.3)}$	.00261
$f_{\mathrm{FvML}(.4)}$	.00839	$f_{\mathrm{Pur}(.4)}$	.01316
$f_{\mathrm{FvML}(.5)}$	.02780	$f_{\mathrm{Pur}(.5)}$	.02897
$f_{\mathrm{FvML}(.6)}$	.05274	$f_{\mathrm{Pur}(.6)}$	.02905
$f_{\mathrm{FvML}(.7)}$	.06001	$f_{\mathrm{Pur}(.7)}$	.01295
$f_{\mathrm{FvML}(.8)}$	.04163	$f_{\mathrm{Pur}(.8)}$	.00236
$f_{\mathrm{FvML}(.9)}$	.01737	$f_{\mathrm{Pur}(.9)}$	.00015
$f_{\mathrm{FvML}(1)}$	.00422	$f_{\mathrm{Pur}(1)}$	.00001

Table : *p*-values, for the cosmic rays data, of the goodness-of-fit tests based on the quartiles $(\hat{c}_{.25}, \hat{c}_{.5}, \hat{c}_{.75})'$ for various FvML and Purkayastha distributions.

As a conclusion, the FvML distribution with concentration $\kappa = 0.7$ (or 0.6) provides a reasonable fit to this data set, and we can moreover well describe the data in terms of quantile values.

Thanks for your attention!