Some intriguing properties of extreme geometric quantiles

Gilles STUPFLER (Aix Marseille Université)
Joint work with Stéphane GIRARD (Inria Grenoble Rhône-Alpes)

9th Conference on Extreme Value Analysis
Ann Arbor, MI, USA, June 16$^{\text{th}}$, 2015

Outline

- Geometric quantiles
- Extreme geometric quantiles
  - Under moment conditions
  - In a multivariate regular variation framework
- Numerical illustrations
- Real data illustration
- Discussion

Multivariate quantiles

- The natural order on $\mathbb{R}$ induces a universal definition of quantiles for univariate distribution functions.
- This is not true in $\mathbb{R}^d$, $d \geq 2$, since no natural order exists in this case.

Many definitions of multivariate quantiles have thus been suggested in the literature:

- Generalizations of univariate quantiles:
  - ◊ Depth-based quantiles: Liu *et al.* (1999), Zuo and Serfling (2000) ;
  - ◊ Norm minimisation: Abdous and Theodorescu (1992), Chaudhuri (1996).

  A review is Serfling (2002).
- Recent developments: DOQR paradigm of Serfling (2010), directional quantiles of Kong and Mizera (2012).

Geometric quantiles

If X is a real-valued random variable, its univariate p–th quantile

$$q(p) := \inf\{t \in \mathbb{R} \,|\, \mathbb{P}(X \le t) \ge p\}$$

can be obtained by solving the optimisation problem

$$\arg\min_{q \in \mathbb{R}} \mathbb{E}(|X - q| - |X|) - (2p - 1)q.$$

- When $\mathbb{E}|X| < \infty$, this problem can be simplified as

  $$\arg\min_{q \in \mathbb{R}} \mathbb{E}|X - q| - (2p - 1)q.$$

  In particular, the median $q(1/2)$ of X is obtained by minimising $\mathbb{E}|X - q|$ with respect to q.
- Subtracting $\mathbb{E}|X|$ makes the cost function well-defined even when $\mathbb{E}|X| = \infty$.

In $\mathbb{R}^d$, $d \geq 2$, analogues of the absolute value and product are given by the Euclidean norm $\|\cdot\|$ and Euclidean inner product $\langle\cdot,\cdot\rangle$.

When X is a multivariate random vector, the geometric quantiles of X, introduced by Chaudhuri (1996), are thus obtained by adapting the aforementioned problem in the multivariate context:

Definition 1 (Chaudhuri 1996)

If $u \in \mathbb{R}^d$ is an arbitrary vector, a geometric u–th quantile of X, if it exists, is a solution of the optimisation problem

$$\arg\min_{q\in\mathbb{R}^d} \mathbb{E}(\|X-q\| - \|X\|) - \langle u, q\rangle. \qquad (P_u)$$

Known properties

Central properties:

- For all $u \in \mathbb{R}^d$ such that $\|u\| < 1$, there exists a unique geometric u–th quantile whenever the distribution of X is not concentrated on a single straight line in $\mathbb{R}^d$ (Chaudhuri, 1996).
- Geometric quantiles are equivariant under any orthogonal transformation (Chaudhuri, 1996).
- The geometric quantile function characterises the associated distribution (Koltchinskii, 1997).

These central properties make geometric quantiles reasonable candidates when trying to define multivariate quantiles.

Extreme properties? Our focus here is to define a notion of extreme geometric quantiles and investigate their properties.

A first step

From now on, we assume that the distribution of X is not concentrated on a single straight line in $\mathbb{R}^d$ and non-atomic.

Proposition 1 (Chaudhuri 1996; Koltchinskii 1997; Girard & S. 2015a)

The optimisation problem (P_u) has a solution if and only if $\|u\| < 1$.

- We cannot compute a geometric quantile with unit index vector, unlike in the univariate case when the distribution has a finite (left or right) endpoint.
- We may nevertheless study the asymptotics of a geometric quantile $q(u)$ when $\|u\| \uparrow 1$: such quantiles will be referred to as extreme geometric quantiles.

The phrase "extreme geometric quantiles" had already been used in the pioneering paper of Chaudhuri (1996).

Theorem 1 (Girard & S. 2015a)

Let S^{d-1} be the unit sphere of $\mathbb{R}^d$.

(i) *The magnitude of extreme geometric quantiles diverges to infinity:*

$$\|q(u)\| \to \infty \quad \text{as} \quad \|u\| \uparrow 1.$$

(ii) *The extreme geometric quantile in the direction $u \in S^{d-1}$ has asymptotic direction u:*

$$\frac{q(\lambda u)}{\|q(\lambda u)\|} \to u \quad \text{as} \quad \lambda \uparrow 1.$$

- ◊ Point (i) is rather intriguing: it holds true even if the distribution of X has a compact support;
- ◊ Related point: sample geometric quantiles do not necessarily lie within the convex hull of the sample, see Breckling *et al.* (2001).

The next results specify rates of the convergence in Theorem 1 under further assumptions.

When there are finite moments

Our first result is obtained in the case when $\|X\|$ satisfies certain moment conditions.

Theorem 2 (Girard & S. 2015a)

Let $u \in S^{d-1}$. *Define* $\Pi_u(x) = x - \langle x, u \rangle u$.

(i) *If* $\mathbb{E}\|X\| < \infty$ *then*

$$\|q(\lambda u)\| \left(\frac{q(\lambda u)}{\|q(\lambda u)\|} - u \right) \to \mathbb{E}(\Pi_u(X)) \quad \text{as} \quad \lambda \uparrow 1.$$

(ii) *If* $\mathbb{E}\|X\|^2 < \infty$ *and* Σ *denotes the covariance matrix of* X *then*

$$\|q(\lambda u)\|^2 (1 - \lambda) \to \frac{1}{2} \left(\operatorname{tr} \Sigma - u' \Sigma u \right) > 0 \quad \text{as} \quad \lambda \uparrow 1.$$

Consequences of Theorem 2

If $\|X\|$ has a finite second moment, then the magnitude of an extreme geometric quantile in the direction u is asymptotically determined by u and the covariance matrix Σ.

In particular, the extreme geometric quantiles of two probability distributions with the same finite covariance matrices are asymptotically equivalent.

$\Rightarrow$ When there is a finite second moment, no information can be recovered on the behaviour of the distribution far from the origin basing solely on extreme geometric quantiles.

Further consequences of Theorem 2

If $\|X\|$ has a finite covariance matrix Σ, then:

- The global maximum of the function $u \mapsto \operatorname{tr}\Sigma - u'\Sigma u$ on the unit sphere is reached at a unit eigenvector of Σ associated with its smallest eigenvalue. Thus:
  - ◊ The norm of an extreme geometric quantile is the largest in the direction where the variance is the smallest;
  - ◊ For elliptically contoured distributions, the shapes of extreme geometric quantile contours and iso-density surfaces are in some sense orthogonal.
- Extreme geometric quantiles can be estimated with a plug-in parametric estimator:

$$\widehat{q}_n(\alpha_n u) = (1-\alpha_n)^{-1/2}\left[\frac{1}{2}\left(\operatorname{tr}\widehat{\Sigma}_n - u'\widehat{\Sigma}_n u\right)\right]^{1/2} u$$

where $\widehat{\Sigma}_n$ is the empirical counterpart of the central quantity Σ.

In a multivariate regular variation framework

When the moment conditions are no longer satisfied, the asymptotic properties of extreme geometric quantiles can be studied in a multivariate regular variation framework:

(M_α) The random vector X has a probability density function f which is continuous on a neighborhood of infinity and such that:

- There exist a positive function Q on $\mathbb{R}^d$ and a function V which is regularly varying at infinity with index $-\alpha < 0$, such that

$$\forall y \neq 0, \quad \left| \frac{f(ty)}{t^{-d} V(t)} - Q(y) \right| \quad \to \quad 0 \quad \text{as} \quad t \to \infty$$

$$\text{and} \quad \sup_{w \in S^{d-1}} \left| \frac{f(tw)}{t^{-d} V(t)} - Q(w) \right| \quad \to \quad 0 \quad \text{as} \quad t \to \infty.$$

- The function $y \mapsto \|y\|^d f(y)$ is bounded in any compact neighborhood of 0.

This model is closely related to the one of Cai *et al.* (2011).
If (M_α) holds, then:

- The function Q is homogeneous of degree $-d-\alpha$ on $\mathbb{R}^d \setminus \{0\}$.
- We have that

$$f(x) = \|x\|^{-d} V(\|x\|) Q(x/\|x\|)(1+o(1)) \text{ as } \|x\| \to \infty$$

  and thus $f(x)$ is roughly of order $\|x\|^{-d-\alpha}$ far from the origin.
- The expectation $\mathbb{E}\|X\|^\beta$ is finite if $\beta < \alpha$.

In particular, the case $\alpha > 2$ is covered by Theorem 2.

Theorem 3 (Girard & S. 2015b)

Let $u \in S^{d-1}$.

(i) *If* (M_α) *holds with* $\alpha \in (0,1)$, *then*

$$\frac{1}{V(\|q(\lambda u)\|)}\left(\frac{q(\lambda u)}{\|q(\lambda u)\|} - u\right) \to \int_{\mathbb{R}^d} \frac{\Pi_u(y)}{\|y-u\|} Q(y)dy \quad \text{as} \quad \lambda \uparrow 1.$$

(ii) *If* (M_α) *holds with* $\alpha \in (0,2)$, *then*

$$\frac{1-\lambda}{V(\|q(\lambda u)\|)} \to \int_{\mathbb{R}^d} \left(1 + \frac{\langle y-u,\, u\rangle}{\|y-u\|}\right) Q(y)dy \quad \text{as} \quad \lambda \uparrow 1.$$

Comments on Theorem 3

Since V is regularly varying with index $-\alpha$, it follows that when $\alpha \in (0, 2)$, the magnitude of an extreme geometric quantile roughly behaves like $(1 - \lambda)^{-1/\alpha}$ as $\lambda \uparrow 1$.

$\Rightarrow$ In this case, the magnitude of an extreme geometric quantile features the behaviour of the distribution far from the origin.

However, Theorem 3 excludes the limit cases $\alpha = 1$ for the asymptotic direction and $\alpha = 2$ for the asymptotic magnitude. The corresponding results are obtained in a separate study, for which we introduce:

$$\mathcal{L}(t) = \int_1^t r^{\alpha-1} V(r) dr.$$

The function $\mathcal{L}$ is slowly varying and $\mathcal{L}(t)/L(t) \to \infty$ as $t \to \infty$.

Theorem 4 (Girard & S. 2015b)

Let $u \in S^{d-1}$ and let σ be the standard surface measure on S^{d-1}.

(i) *If (M_1) holds and $\mathcal{L}(t) \to \infty$ as $t \to \infty$ then as $\lambda \uparrow 1$:*

$$\frac{\|q(\lambda u)\|}{\mathcal{L}(\|q(\lambda u)\|)} \left(\frac{q(\lambda u)}{\|q(\lambda u)\|} - u \right) \to \int_{S^{d-1}} \Pi_u(w) Q(w) \sigma(dw).$$

(ii) *If (M_2) holds and $\mathcal{L}(t) \to \infty$ as $t \to \infty$ then as $\lambda \uparrow 1$:*

$$\frac{\|q(\lambda u)\|^2}{\mathcal{L}(\|q(\lambda u)\|)}(1 - \lambda) \to \frac{1}{2} \int_{S^{d-1}} \langle \Pi_u(w),\, w \rangle Q(w) \sigma(dw).$$

In a nutshell

For all $\alpha > 0$, we can write

$$\frac{q(\lambda u)}{\|q(\lambda u)\|} - u \quad \propto \quad R_{1,\alpha}((1-\lambda)^{-1})$$
$$\text{and} \quad \|q(\lambda u)\| \quad \propto \quad R_{2,\alpha}((1-\lambda)^{-1}) \quad \text{as} \quad \lambda \uparrow 1,$$

where $R_{1,\alpha}$ and $R_{2,\alpha}$ are regularly varying functions with respective indices $-\min(1,\alpha)/\min(2,\alpha)$ and $1/\min(2,\alpha)$.

$\Rightarrow$ Extreme geometric quantiles feature the behaviour of a multivariate distribution far from the origin only when its covariance matrix does not exist.

Numerical illustrations: Theorem 2

We choose $d = 2$. The following bivariate distributions are considered:

- the centred Gaussian bivariate distribution $\mathcal{N}(0, v_X, v_Y, v_{XY})$ with covariance matrix

$$\Sigma = \begin{pmatrix} v_X & v_{XY} \\ v_{XY} & v_Y \end{pmatrix}.$$

- a centred double exponential distribution $\mathcal{E}(\lambda_-, \mu_-, \lambda_+, \mu_+)$, with λ_-, μ_-, λ_+, $\mu_+ > 0$, whose probability density function is:

$$f(x,y) = \frac{1}{4} \begin{cases} \lambda_+ \mu_+ e^{-\lambda_+|x| - \mu_+|y|} & \text{if } xy > 0, \\ \lambda_- \mu_- e^{-\lambda_-|x| - \mu_-|y|} & \text{if } xy \leq 0. \end{cases}$$

In this case, X has an explicit finite covariance matrix.

Since both distributions have a finite covariance matrix, Theorem 2 entails that their extreme geometric quantiles are asymptotically equal to:

$$q_{\mathrm{eq}}(\lambda u) := (1-\lambda)^{-1/2}\left[\frac{1}{2}\left(\operatorname{tr}\Sigma - u'\Sigma u\right)\right]^{1/2} u.$$

Goal: to show that

- for these two distributions, equal covariance matrices induce equivalent extreme geometric quantiles;
- and to assess the accuracy of the asymptotic equivalent.

We choose three different sets of parameters, in order that the related covariance matrices coincide:

- $\mathcal{N}(0, 1/2, 1/2, 0)$ and $\mathcal{E}(2, 2, 2, 2)$ with spherical covariance matrices;
- $\mathcal{N}(0, 1/8, 3/4, 0)$ and $\mathcal{E}(4, 2\sqrt{2/3}, 4, 2\sqrt{2/3})$ with diagonal but non-spherical covariance matrices;
- $\mathcal{N}(0, 1/2, 1/2, 1/6)$ and $\mathcal{E}(2\sqrt{3}, 2\sqrt{3}, 2\sqrt{3/5}, 2\sqrt{3/5})$ with full covariance matrices.

Any $u \in S^1$ can be written $u = u_\theta = (\cos\theta, \sin\theta)$, $\theta \in [0, 2\pi)$. We let $\lambda = 0.995$ and in each case, we compute:

- the true iso-quantile curve $\mathcal{C}q(\lambda) = \{q(\lambda u_\theta),\ \theta \in [0, 2\pi)\}$;
- its asymptotic equivalent $\mathcal{C}q_{\mathrm{eq}}(\lambda) = \{q_{\mathrm{eq}}(\lambda u_\theta),\ \theta \in [0, 2\pi)\}$.

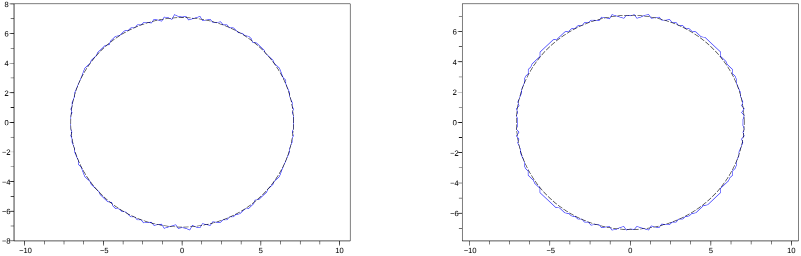


Figure 1: Spherical case: Gaussian (left) and double exponential (right) distributions. Iso-quantile curves $\mathcal{C}q(\lambda)$ (full blue line) and $\mathcal{C}q_{\mathrm{eq}}(\lambda)$ (dashed black line).

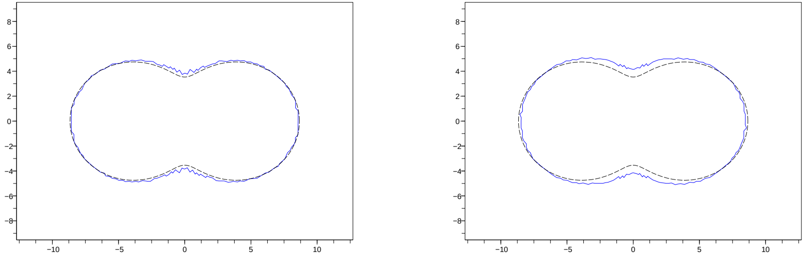


Figure 2: Diagonal case: Gaussian (left) and double exponential (right) distributions. Iso-quantile curves $\mathcal{C}q(\lambda)$ (full blue line) and $\mathcal{C}q_{\mathrm{eq}}(\lambda)$ (dashed black line).

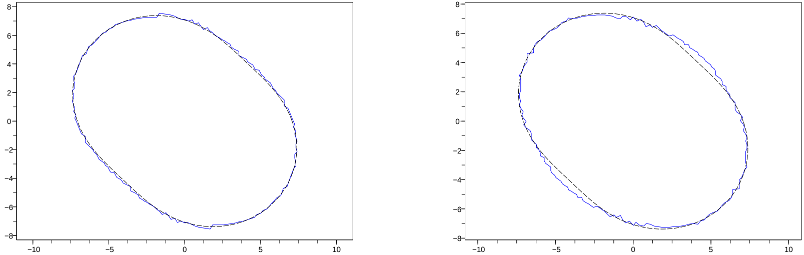


Figure 3: Full case: Gaussian (left) and double exponential (right) distributions. Iso-quantile curves $\mathcal{C}q(\lambda)$ (full blue line) and $\mathcal{C}q_{\rm eq}(\lambda)$ (dashed black line).

Numerical illustrations: Theorem 3

Here, we consider a bivariate Pareto($\alpha, \sigma_1, \sigma_2$) distribution, whose probability density function is:

$$f(x,y) = \frac{\alpha}{2\sigma_1\sigma_2\pi}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}\right)^{(-2-\alpha)/2} \mathbb{1}_{[1,\infty)}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}\right)$$

where α, σ_1^2 and $\sigma_2^2 > 0$. When $\alpha > 2$, this distribution has covariance matrix:

$$M = \frac{1}{2}\cdot\frac{\alpha}{\alpha-2}\Sigma, \quad \text{with} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}.$$

For any $\alpha > 0$, this distribution is part of the class (M_α), with

$$\begin{aligned} Q(x) &= (x'\Sigma^{-1}x)^{(-2-\alpha)/2} \\ \text{and} \quad V(t) &= \frac{\alpha}{2\sigma_1\sigma_2\pi} t^{-\alpha} \mathbb{1}_{[1,\infty)}(t). \end{aligned}$$

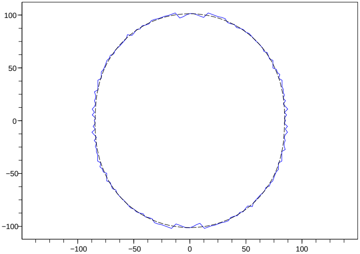
Theorems 2 and 3 thus entail that the extreme geometric quantiles of this distribution are asymptotically equal to:

$$q_{\mathrm{eq}}(\lambda u) := (1-\lambda)^{-1/\alpha} I(\alpha, \sigma_1, \sigma_2) u \quad \text{if} \quad \alpha < 2$$

where $I(\alpha, \sigma_1, \sigma_2)$ is a positive constant, and

$$q_{\mathrm{eq}}(\lambda u) := (1-\lambda)^{-1/2} \left[\frac{1}{2} \left(\operatorname{tr} M - u' M u \right) \right]^{1/2} u \quad \text{if} \quad \alpha > 2.$$

Goal: to examine if both these approximations are satisfactory on this heavy-tailed example.



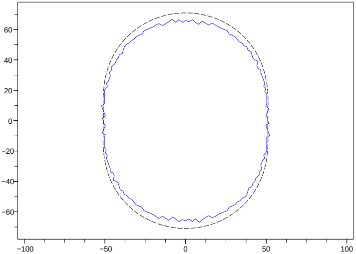


Figure 4: Pareto($\alpha, 2, 1/2$) model, with $\alpha = 1.3$ (left) and $\alpha = 1.5$ (right). Iso-quantile curves $\mathcal{C}q(\lambda)$ (full blue line) and $\mathcal{C}q_{\text{eq}}(\lambda)$ (black dashed line).

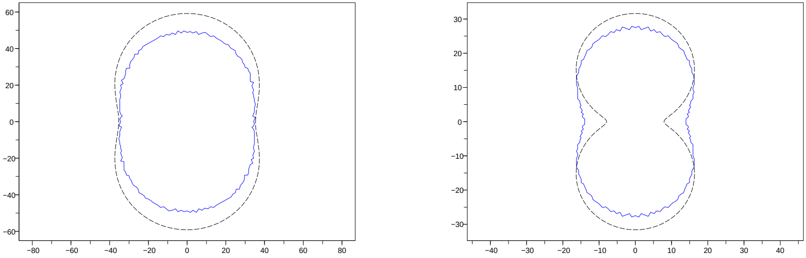


Figure 5: Pareto($\alpha, 2, 1/2$) model, with $\alpha = 1.7$ (left) and $\alpha = 2.5$ (right). Iso-quantile curves $\mathcal{C}q(\lambda)$ (full blue line) and $\mathcal{C}q_{\mathrm{eq}}(\lambda)$ (black dashed line).

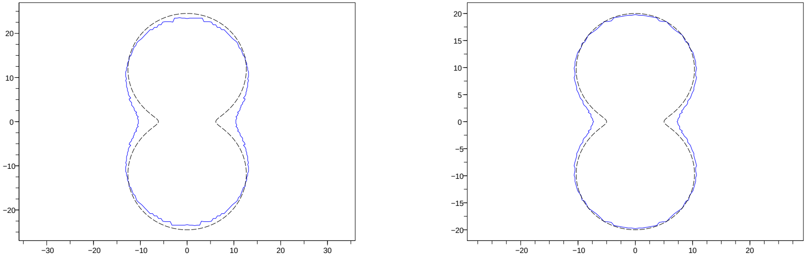


Figure 6: Pareto$(\alpha, 2, 1/2)$ model, with $\alpha = 3$ (left) and $\alpha = 4$ (right). Iso-quantile curves $\mathcal{C}q(\lambda)$ (full blue line) and $\mathcal{C}q_{\text{eq}}(\lambda)$ (black dashed line).

Illustration on the Pima Indians Diabetes dataset

- The sample behaviour of extreme geometric quantiles is illustrated on a two-dimensional dataset extracted from the Pima Indians Diabetes Database.
- The data set consists of $n = 392$ pairs (X_i, Y_i), where X_i is the body mass index of the ith individual and Y_i is its diastolic blood pressure.
- Already considered in Chaouch and Goga (2010) in the context of outlier detection.

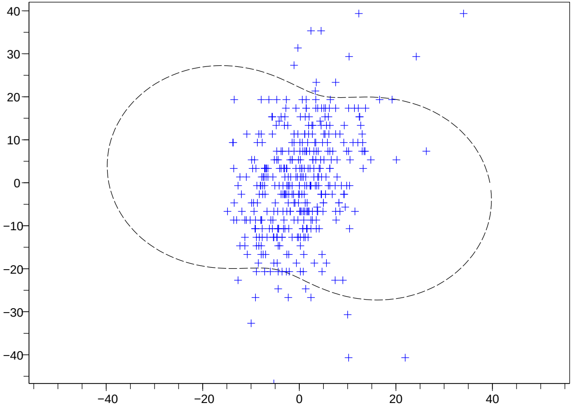


Figure 7: Centred data: estimated geometric iso-quantile curve, $\alpha = 0.95$.

Discussion

- Extreme geometric quantiles in the direction u have asymptotic direction u.
- They are asymptotically equal for two distributions which have the same finite covariance matrix, which is not satisfying from the extreme value perspective.
- The shape of the iso-quantile curves may be totally different from the shape of the density contour plots. Outlier detection should be conducted with great care.
- They do however feature the behaviour of X far from the origin in a multivariate regular variation context when the right tail of $\|X\|$ is sufficiently heavy.

References

Abdous, B., Theodorescu, R. (1992) Note on the spatial quantile of a random vector, *Statistics and Probability Letters* **13**(4): 333–336.

Breckling, J., Kokic, P., Lübke, O. (2001) A note on multivariate $M-$quantiles, *Statistics and Probability Letters* **55**(1): 39–44.

Cai, J.-J., Einmahl, J.H.J., de Haan, L. (2011) Estimation of extreme risk regions under multivariate regular variation, *Annals of Statistics* **39**(3): 1803–1826.

Chaouch, M., Goga, C. (2010) Design-based estimation for geometric quantiles with application to outlier detection, *Computational Statistics and Data Analysis* **54**(10): 2214–2229.

Chaudhuri, P. (1996) On a geometric notion of quantiles for multivariate data, *Journal of the American Statistical Association* **91**(434): 862–872.

Girard, S., Stupfler, G. (2015a) Intriguing properties of extreme geometric quantiles, to appear in *REVSTAT: Statistical Journal*.

Girard, S., Stupfler, G. (2015b) Extreme geometric quantiles in a multivariate regular variation framework, working paper.

Koltchinskii, V.I. (1997) M-estimation, convexity and quantiles, *Annals of Statistics* **25**(2): 435–477.

Kong, L., Mizera, I. (2012) Quantile tomography: using quantiles with multivariate data, *Statistica Sinica* **22**: 1589–1610.

Serfling, R. (2002) Quantile functions for multivariate analysis: approaches and applications, *Statistica Neerlandica* **56**(2): 214–232.

Serfling, R. (2010) Equivariance and invariance properties of multivariate quantile and related functions, and the role of standardisation, *Journal of Nonparametric Statistics* **22**(7): 915–936.

Zuo, Y., Serfling, R. (2000) General notions of statistical depth function, *Annals of Statistics* **28**(2): 461–482.

Thanks for listening!