Journal of Classification 21:139-154 (2004)
DOI: 10.1007/s00357-004-0009-9

Asymptotic Properties of the Multivariate L_p-mean

Yadolah Dodge

Statistics Group, University of Neuchâtel, Switzerland

Pirmin Lemburger

Statistics Group, University of Neuchâtel, Switzerland

Abstract: For a multivariate random variable $\mathbf{X}$ and probability density f, we define a central parameter $\boldsymbol{\mu}_*$ as the unique vector which minimizes $\mathbb{E}[||\mathbf{X}-\boldsymbol{\mu}||_p^q]$. When $q=2$, we call $\boldsymbol{\mu}_*$ the L_p-mean of the distribution f. Taking a sample $\mathbf{X}_1,\ldots,\mathbf{X}_n$ of size n from this distribution, we define the corresponding empirical central parameter $\widehat{\boldsymbol{\mu}}_n$ as a specific the random variable minimizing $\sum_{i=1}^n ||\mathbf{X}_i-\boldsymbol{\mu}||_p^q$. In this article we prove that $\widehat{\boldsymbol{\mu}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu}_*$ as $n\to\infty$ and that $\sqrt{n}\,(\widehat{\boldsymbol{\mu}}_n-\boldsymbol{\mu}_*) \xrightarrow{\mathcal{L}} Z \sim \mathcal{N}(0,\mathbf{V})$ as $n\to\infty$. We give the covariance matrix $\mathbf{V}=\mathbf{V}[f]$ of the multivariate normal distribution as an explicit functional of the p.d.f. f. A simulation study is provided in which a relative performances of L_p-mean and L_p-median are compared.

Keywords : Classification; Multivariate; Data depth; L_p-norms; Measure of Centrality.

We wish to thank the four unknown referees for their excellent remarks on the paper that improved substantially the quality of the paper. We also thank Giuseppe Melfi for his valuable technical help. Our special thanks go to the editor for his constructive criticism and comments.

Authors' Address: Statistics Group, University of Neuchâtel,P.O. Box 805, 2002 Neuchâtel, Switzerland; e-mail: yadolah.dodge@unine.ch.

1. Introduction

Suppose we are given a vector-valued random variable $\mathbf{X} = (X_1, \ldots, X_k)' : (\Omega, \mathfrak{A}, P) \to \mathbb{R}^k$ where Ω is the fundamental space, $\mathfrak{A}$ the σ-algebra and P the probability measure. Assume that $\mathbf{X}$ has a p.d.f. f. How can we define a center $\mathbf{c}_*$ for the p.d.f. f? Clearly there is no unique answer to this question. Perhaps the most natural way is to take $\mathbf{c}_*$ as the center of gravity of f, that is,

$$\mathbf{c}_* = \mathbb{E}[\mathbf{X}] = \int \mathbf{x} f(\mathbf{x}) d\mathbf{x}.$$

This center of gravity $\mathbf{c}_*$ can also be characterized as the unique vector which minimizes the function

$$\mathbf{c} \to \mathbb{E}[\|\mathbf{X} - \mathbf{c}\|_2^2] = \sum_{j=1}^{k} \mathbb{E}[|X_j - c_j|^2], \tag{1}$$

where $\|\mathbf{x}\|_p = \left(\sum_{j=1}^k |x_j|^p\right)^{1/p}$ is the usual L_p-norm on $\mathbb{R}^k$. This follows simply from the fact that each term $\mathbb{E}[|X_j - c_j|^2]$, $j = 1, \ldots, k$ in (1) is minimized separately by $c_{*j} = \mathbb{E}[X_j] = \int x_j f(\mathbf{x}) d\mathbf{x}$. Now, generalizing this, we can define, for each $p \geq 1$ and $q \geq 1$ a central parameter as one of the minimizers of the function

$$\mathbf{c} \to \Phi_p^q(\mathbf{c}) = \mathbb{E}[\|\mathbf{X} - \mathbf{c}\|_p^q]. \tag{2}$$

The question of uniqueness will be discussed in detail in section 2. Clearly p defines how we decide to measure distances in $\mathbb{R}^k$, for instance $p = 2$ corresponds to the standard Euclidean distance. When $k = 1$, p obviously plays no role, in that case when $q = 2$, $\mathbf{c}_* = c_*$ is just the usual expectation $\mathbb{E}[X]$ of X, whereas when $q = 1$, c_* is a median M of X (not unique when n is even). In the multivariate case $k \geq 2$, we define by analogy $\mathbf{c}_*$ to be the L_p-mean of $\mathbf{X}$ if it minimizes $\Phi_p^2(\mathbf{c})$ and the L_p-median of $\mathbf{X}$ if it minimizes $\Phi_p^1(\mathbf{c})$.

Although quite surprisingly the general asymptotic properties of the central parameter $\mathbf{c}$ for arbitrary p and q have never been considered before, various definitions of central location parameters have been proposed earlier. The $(p, q) = (1, 1)$ case was first introduced by Hayford (1902). It is also studied in Mood (1941) where it is called the *arithmetic median.* For $(p, q) = (2, 1)$ the parameter $\mathbf{c}$ is known as the L_2-median. It was first introduced by Gini and Galvani (1929) and by Eells (1930) and was called the *spatial median* by Brown (1997) although Haldane (1948) introduced the term *geometrical median* and Gower (1974) the term *mediancenter*. In his article, Small (1990) unfortunately introduces the term L_1-*median* to denote the spatial median. Dodge and Rousson (1998) introduced the L_1-*mean* that is the $(p, q) = (1, 2)$ case with our notations. They provided an efficient algorithm to compute it numerically. See also references Chaudhuri (1992), Gentleman (1965), Oja and Niinimaa (1985) and Pollard (1984). For the convenience of the reader we list below special cases of our L_p-norm central parameter that have been considered in the literature:

- $(p, q) = (2, 2)$: As we mentioned, the L_2-mean $\mathbf{c}_*$ is just the center of gravity $\mathbb{E}[\mathbf{X}]$ of the p.d.f. f.

- $(p, q) = (1, 1)$: Here $\mathbf{c}_*$ is the L_1-median. Note that $E[\|\mathbf{X} - \mathbf{c}\|_1] = \sum_{j=1}^{k} |X_j - c_j|$ is minimized by taking each c_{*j} to be a median M_j of X_j for $j = 1, \ldots, k$. The problem is therefore one-dimensional. For any finite sample size n, the p.d.f. of the corresponding random variable $\widehat{c}_{*j}$ can be found explicitly using the p.d.f. of the order statistics of the components X_j .

- $(p, q) = (2, 1)$: The L_2-median has been considered by Eells (1930) and by Gini and Galvani (1929). It is just a special case of our treatment.

- $(p, q) = (1, 2)$: The L_1-mean is in some sense the simplest non-euclidean mean, it will also be a special case of our results. This case was introduced recently in Dodge and Rousson (1998).

The paper is organized as follows: in section 2. we examine the question of the uniqueness of the minimizers of Φ and of its empirical counterpart $\widehat{\Phi}_n$. When needed, we shall single out one of these minimizers, and denote it respectively by $\boldsymbol{\mu}_*$ and $\widehat{\boldsymbol{\mu}}_n$. Then we state a law of large numbers and a central limit theorem which describe the asymptotic behavior of $\widehat{\boldsymbol{\mu}}_n$ as $n \to \infty$. In section 3. we give a detailed proof of both theorems and finally we leave the reader with an open problem.

2. Statement of the Results

Before we state our results, we define the central parameter $\boldsymbol{\mu}_*$ of the distribution f and its estimator $\widehat{\boldsymbol{\mu}}_n$ as an unbiased statistic on the sample. We then establish in which case they are unique. From here on we shall drop the indices on Φ_p^q to avoid too heavy annotation. We must essentially investigate the convexity properties of the following two functions of μ:

$$\begin{aligned} \Phi(\mu) &= \int f(\mathbf{x})\|\mathbf{x}-\mu\|_p^q d\mathbf{x} \quad \text{and} \\ \Phi_n(\mu) &= \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i-\mu\|_p^q, \end{aligned} \tag{3}$$

depending on the values of p and q. The points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are here thought of as being realizations of the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$, i.e. $\mathbf{x}_i = \mathbf{X}_i(\omega)$ for some occurrence $\omega \in \Omega^{\mathbb{N}}$. We write $\widehat{\Phi}_n(\mu)$ for the random variable defined by replacing $\mathbf{x}_i$ in (3) by $\mathbf{X}_i$.

Let $0 < \lambda < 1$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$ and $p \geq 1, q \geq 1$. Convexity of Φ and Φ_n is a consequence of the following two inequalities

$$\|\lambda\mathbf{u}+(1-\lambda)\mathbf{v}\|_p^q \leq (\lambda\|\mathbf{u}\|_p+(1-\lambda)\|\mathbf{v}\|_p)^q \tag{4}$$

$$\leq \lambda\|\mathbf{u}\|_p^q+(1-\lambda)\|\mathbf{v}\|_p^q \tag{5}$$

where (4) is the triangle inequality for $\|.\|_p$, $p \geq 1$ and (5) is the convexity of $x \to x^q$, $q \geq 1$. Obviously any convex linear combination of minimizers is still a minimizer. This allows us to define μ_* and μ_n as the centers of gravity of the respective sets of minimizers of Φ and Φ_n. Again we shall denote by $\widehat{\mu}_n$ the random variable corresponding to μ_n.

Now equality in (4) holds

- for $p > 1$ if and only if $\mathbf{u} = \alpha\mathbf{v}$, for some $\alpha > 0$ and
- for $p = 1$ if and only if $\operatorname{sgn}(u_j) = \operatorname{sgn}(v_j)$, for all $j = 1, \ldots, k$.

whereas (5) is saturated

- for $q > 1$ if and only if $\|\mathbf{u}\|_p = \|\mathbf{v}\|_p$ and
- for $q = 1$ always.

Uniqueness of μ_*

Notice first that all minimizers of Φ, and in particular μ_*, lie in the convex hull $\mathcal{H}(\operatorname{supp}(f))$ of the support $\operatorname{supp}(f)$ of p.d.f. f. We shall assume that $\operatorname{supp}(f)$ is connected. See Figure 1.

- Let $p > 1$. Take $\mathbf{u} = \mathbf{x} - \mu_*$ and $\mathbf{v} = \mathbf{x} - \nu_*$ in (4) and (5), multiply both members by the positive density f and integrate over $\mathbf{x}$. We obtain

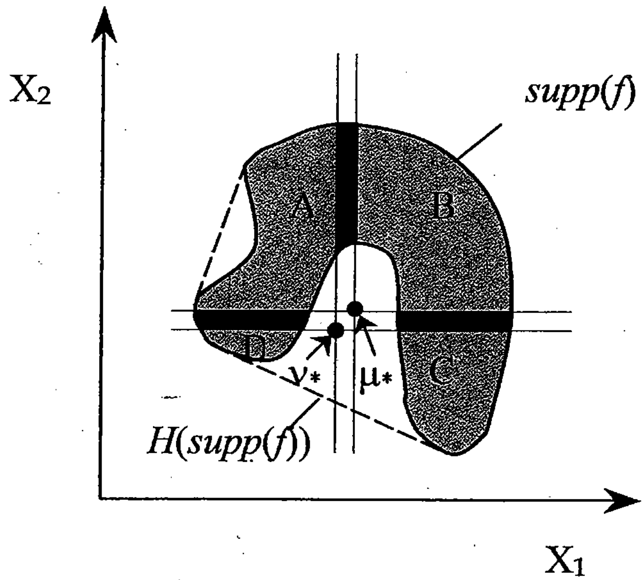

Figure 1. The support supp(f), its convex hull $\mathcal{H}$(supp(f)) and two candidates μ_* and ν_* for the minimizer of Φ.

$\Phi(\lambda\mu_* + (1-\lambda)\nu_*) \leq \lambda\Phi(\mu_*) + (1-\lambda)\Phi(\nu_*)$ for $0 < \lambda < 1$. Now suppose equality holds. This implies that $\mathbf{x}-\mu_* = \alpha(\mathbf{x})(\mathbf{x}-\nu_*)$, $\alpha(\mathbf{x}) > 0$ for almost all $\mathbf{x} \in \text{supp}(f)$. But if $\mu_* \neq \nu_*$ this requires that supp(f) be contained in a line, which is clearly impossible in $\mathbb{R}^k$, $k \geq 2$ for a p.d.f. f of a continuous random variable. Therefore we conclude that $\mu_* = \nu_*$ for $p > 1$.

- Let $p = 1$. Perform the same steps as above to deduce the convexity inequality for Φ. Now, this time equality implies that $\text{sgn}(x_j - \mu_j) = \text{sgn}(x_j - \nu_j)$ for all $j = 1, \ldots, k$ and for almost all $\mathbf{x} \in \text{supp}(f)$. But if $\mu_* \neq \nu_*$ this inequality will not hold for any point $\mathbf{x} \in \text{supp}(f)$ for which $x_l \in [\min(\mu_l, \nu_l), \max(\mu_l, \nu_l)]$ for some index l (see the dark-gray zones in Figure 1). The fact that supp(f) is connected and that the candidates for the minimizer have to lie inside $\mathcal{H}$(supp(f)) imply that those points form a set of non-zero measure and thus the inequality can not hold. We conclude that $\mu_* = \nu_*$ also for $p = 1$.

Uniqueness of $\boldsymbol{\mu}_n$

Let $\boldsymbol{\mu}_n = (\mu_{n1}, \ldots, \mu_{nk})^{\mathrm{T}}$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})^{\mathrm{T}},\ i = 1, \ldots, n.$

- Let $p > 1$. Define the event $B_n \in \mathfrak{A}$ as the set of occurences $\omega \in \Omega^{\mathrm{N}}$ for which there are at least two different minimizers $\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_n$ of $\Phi_n := \widehat{\Phi}_n(\omega)$. To saturate the convexity inequality for Φ_n we must require that the points $\mathbf{x}_i = \mathbf{X}_i(\omega)$ satisfy $\mathbf{x}_i - \boldsymbol{\mu}_n = \alpha_i(\mathbf{x}_i - \boldsymbol{\nu}_n),\ \alpha_i > 0,\ i = 1, \ldots, n$ which means that all $\mathbf{x}_i$ should lie on a line. But this is clearly an event of probability zero, for a continuous multivariate random variable with p.d.f. f. Thus for $p > 1$ and for any n the minimizer $\boldsymbol{\mu}_n$ is almost surely unique.

- Let $p = 1$ and $q = 1$. Define the event $B_l(\varepsilon) \in \mathfrak{A}$ as the set of occurrences $\omega \in \Omega^{\mathrm{N}}$ for which there are at least two different minimizers $\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_n$ which satisfy $\mu_{nl} - \nu_{nl} > \varepsilon$ for some $\varepsilon > 0$ and *for all* $n > 1$. To saturate the convexity inequalities for all Φ_n we must require that all points $\mathbf{x}_i = \mathbf{X}_i(\omega)$ satisfy the equalities $\mathrm{sgn}(x_{ij} - \mu_{nj}) = \mathrm{sgn}(x_{ij} - \nu_{nj})$ for all $j = 1, \ldots, k$ and all $n > 1$. But points $\mathbf{x}_i \in \mathrm{supp}(f)$ for which $x_{il} \in [\inf_{n>1} \nu_{nl}, \sup_{n>1} \mu_{nl}]$ will violate these equalities. As $\varepsilon > 0$, the set of these points clearly has non-zero measure with respect to our p.d.f. f and therefore $B_l(\varepsilon)$ is an event with zero probability. Thus for the case $p = q = 1$ (the L_1-median) almost surely we can find $N(\varepsilon)$ such that $n > N(\varepsilon)$ implies $|\nu_{nj} - \mu_{nj}| < \varepsilon$ for all $j = 1, \ldots, k$.

- Let $p = 1$ and $q > 1$. Define $B = \bigcap_{n>1} B_n$. In words, B is the event that the $\widehat{\Phi}(\omega)$ has no unique minimizer *for any* $n > 1$. Let $\boldsymbol{\mu}_n$ and $\boldsymbol{\nu}_n$ be two different minimizers of Φ_n. Because $q > 1$ we must now impose an additional condition on the points $\mathbf{x}_i$ to saturate the second inequality (5), namely that $\|\mathbf{x}_i - \boldsymbol{\mu}\|_1 = \|\mathbf{x}_i - \boldsymbol{\nu}\|_1$ for all $i = 1, \ldots, n$ and all $n > 1$. Figure 2 illustrates the situation in $k = 2$ dimensions for $n = 2$ and $n = 3$.

  The requirement of having non-uniqueness for the minimizer *for all* $n > 1$ is the same as asking that all points $\mathbf{x}_i$ fall only in two opposite sectors, as for instance A and C in Figure 1. This is clearly an event of probability zero. Therefore when $p = 1$ and $q > 1$ the function $\widehat{\Phi}_n(\omega)$ will almost surely have a unique minimizer if n is large enough.

Finally, let us note that in contradistinction to the usual L_2-mean, for which $\boldsymbol{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, there is in general *no explicit formula* for $\boldsymbol{\mu}_n$ in terms of the coordinates x_{ij} of the n points $\mathbf{x}_i$.

We are now ready to state our two main results:

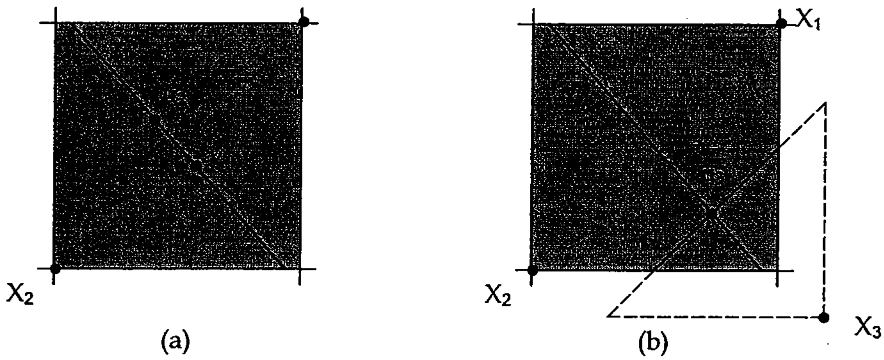

Figure 2. Example of non-uniqueness for $\widehat{\boldsymbol{\mu}}_n$ for $p = 1$ and $q > 1$. In (a) $\boldsymbol{\mu}_2$ is not unique whereas in (b) $\boldsymbol{\mu}_3$ is unique.

Theorem 1 *Let* $\mathbf{X}$ *be a continuous random variable with values in* $\mathbb{R}^k$, $k \geq 2$ *and p.d.f.* f *with connected support. For any* $p \geq 1$ *and any* $q \geq 1$, *define the central parameter* $\boldsymbol{\mu}_*$ *of* $\mathbf{X}$ *as the unique minimizer of the function* $\boldsymbol{\mu} \to \mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|_p^q]$. *For each sample* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *of random variable i.i.d. as* $\mathbf{X}$, *define an empirical central parameter* $\widehat{\boldsymbol{\mu}}_n$ *as a (not necessarely unique) minimizer of the function* $\boldsymbol{\mu} \to \frac{1}{n} \sum_{i=1}^{\infty} \|\mathbf{X}_i - \boldsymbol{\mu}\|_p^q$. *Then if* $\mathbb{E}[\|\mathbf{X}\|_p^q] < \infty$ *the following holds:*

$$\widehat{\boldsymbol{\mu}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu}_* \text{ as } n \to \infty,$$

meaning precisely almost sure componentwise convergence of the vectors $\widehat{\boldsymbol{\mu}}_n$ *towards* $\boldsymbol{\mu}_*$.

Theorem 2 *Assume* $(p > 1$ *and* $q \geq 1)$ *or* $(p = 1$ *and* $q > 1)$, *then the following holds:*

$$\sqrt{n}\,(\widehat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}_*) \xrightarrow{\mathcal{L}} \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}) \text{ as } n \to \infty,$$

where $\mathbf{Z}$ *is a normal multivariate random variable with covariance matrix* $\mathbf{V}$. *If we denote* $\frac{\partial}{\partial \mu_k} \equiv \partial_k$, *we have:*

$$\begin{aligned} \mathbf{V} &= \mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-1} \text{ where} \\ \mathbf{M} &= (M_{kl}) \text{ with } M_{kl} = \partial_k \partial_l \mathbb{E}\left[\|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q\right] \text{ and} \\ \mathbf{A} &= (A_{kl}) \text{ with } A_{kl} = \mathbb{E}\left[\left(\partial_k \|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q\right) \cdot \left(\partial_l \|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q\right)\right]. \end{aligned} \tag{6}$$

The matrix elements of A_{kl} and M_{kl} can be written out more explicitly if we use the relations:

$$\begin{aligned} \partial_k : |X_l - \mu_l| &= -\delta_{kl} : \operatorname{sgn}(X_l - \mu_l) \\ \partial_k : \operatorname{sgn}(X_l - \mu_l) &= -2\delta_{kl} : \delta(X_l - \mu_l), \end{aligned} \tag{7}$$

where $\delta(x)$ is the Dirac distribution at 0. However the expressions are not particularly illuminating.

For the special case $(p,q)=(2,2)$, $\boldsymbol{\mu}_*=\mathbb{E}[\mathbf{X}]$ and $\partial_k\|\mathbf{X}-\boldsymbol{\mu}_*\|_2^2 = -2\,(X_k-\mu_{*k})$, $\partial_k\partial_l\|\mathbf{X}-\boldsymbol{\mu}_*\|_2^2 = 2\delta_{kl}$ which translates into $\mathbf{M}=2\mathbb{I}$ and $A_{kl}=4\mathrm{Cov}(X_k,X_l)$. We thus simply recover the standard result $V_{kl}=\mathrm{Cov}(X_k,X_l)$.

3. Proofs

The proof of Theorem 1 uses standard convexity arguments and the strong law of large numbers.

Proof of Theorem 1: From the assumption $\mathbb{E}[\|\mathbf{X}\|_p^q]<\infty$ and the triangle inequality for $\|.\|_p$ we get:

$$\mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|_p^q] \le \mathbb{E}[(\|\mathbf{X}\|_p+\|\boldsymbol{\mu}\|_p)^q] \le \sum_{k=0}^{q}\binom{q}{k}\,\mathbb{E}[\|\mathbf{X}\|_p^k]\cdot\|\boldsymbol{\mu}\|_p^{q-k}<\infty. \quad (8)$$

Each $\mathbf{X}_i$, $i=1,\dots,n$ being distributed as $\mathbf{X}$, (8) establishes that the random variable $Y_i=\|\mathbf{X}_i-\boldsymbol{\mu}\|_p^q$, $i=1,\dots,n$ are all in $L_1(\Omega^{\mathbb{N}},P^{\mathbb{N}})$. We can therefore apply the strong law of large numbers to the random variable Y_i and conclude that:

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_i-\boldsymbol{\mu}\|_p^q \xrightarrow{\text{a.s.}} \mathbb{E}[\|\mathbf{X}-\boldsymbol{\mu}\|_p^q] \text{ as } n\to\infty$$

for all $\boldsymbol{\mu}\in\mathbb{R}^k$. Using the notation defined in section 2, this means there is a set $A\subset\Omega^{\mathbb{N}}$, negligible with respect to $P^{\mathbb{N}}$, such that:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_i(\omega)-\boldsymbol{\mu}\|_p^q=\lim_{n\to\infty}\Phi_n(\boldsymbol{\mu})=\Phi(\boldsymbol{\mu})$$

for all $\omega\in\Omega^{\mathbb{N}}\setminus A$. From now on we fix such an ω and choose an arbitrary $\varepsilon>0$. We will show that for large enough n, any minimizer $\boldsymbol{\mu}_n$ of Φ_n lies inside the closed disc $D_\varepsilon(\boldsymbol{\mu}_*)$ of radius ε around the unique minimizer $\boldsymbol{\mu}_*$ of Φ, thus proving $\widehat{\boldsymbol{\mu}}_n \xrightarrow{\text{a.s.}} \boldsymbol{\mu}_*$.

Simple convergence of Φ_n towards Φ on the compact domain $D_\varepsilon(\boldsymbol{\mu}_*)$ implies uniform convergence on $D_\varepsilon(\boldsymbol{\mu}_*)$. Therefore, $\forall\delta>0$, we can find $N(\varepsilon,\delta)$ such that $n>N(\varepsilon,\delta)$ implies $|\Phi_n(\boldsymbol{\mu})-\Phi(\boldsymbol{\mu})|<\delta$ for all $\boldsymbol{\mu}\in D_\varepsilon(\boldsymbol{\mu}_*)$. Let $\partial D_\varepsilon(\boldsymbol{\mu}_*)$ be the boundary circle of $D_\varepsilon(\boldsymbol{\mu}_*)$. In section 2 we established that

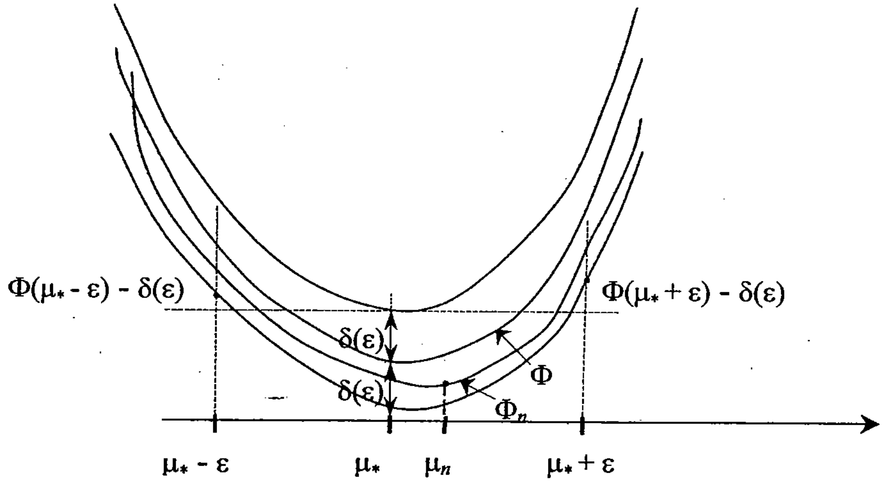

Figure 3. Convergence of the functions Φ_n towards the strictly convex function Φ.

Φ is strictly convex for any $p \geq 1$ and any $q \geq 1$. It is thus possible to take $\delta = \delta(\varepsilon)$ so small that

$$\Phi(\boldsymbol{\mu}_*) + \delta(\varepsilon) < \min_{\boldsymbol{\mu} \in \partial D_\varepsilon(\boldsymbol{\mu}_*)} \Phi(\boldsymbol{\mu}) - \delta(\varepsilon).$$

This implies, for $n > N(\varepsilon, \delta(\varepsilon))$, that

$$\Phi_n(\boldsymbol{\mu}_*) < \min_{\boldsymbol{\mu} \in \partial D_\varepsilon(\boldsymbol{\mu}_*)} \Phi_n(\boldsymbol{\mu})$$

and thus convexity of Φ_n implies that any of its minimizers $\boldsymbol{\mu}_n$ lies inside $D_\varepsilon(\boldsymbol{\mu}_*)$. Figure 3 illustrates the situation. ■

The general idea of the proof of the second theorem is the following: because $\boldsymbol{\mu}_*$ and $\boldsymbol{\mu}_n = \widehat{\boldsymbol{\mu}}_n(\omega)$ solve respectively the equations $\nabla\Phi(\boldsymbol{\mu}) = \mathbf{0}$ and $\nabla\Phi_n(\boldsymbol{\mu}) = \mathbf{0}$, we use a first order expansion to compare the two equations. To get the asymptotic distribution of the corresponding random variable $\widehat{\boldsymbol{\mu}}_n$ we apply the central limit theorem to the random variable $\partial_j \widehat{\Phi}_n(\boldsymbol{\mu})$.

Proof of Theorem 2: Recall the two definitions:

$$\begin{aligned} \Phi(\boldsymbol{\mu}) &= \mathbb{E}[\|\mathbf{X} - \boldsymbol{\mu}\|_p^q], \\ \widehat{\Phi}_n(\boldsymbol{\mu}) &= \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_i - \boldsymbol{\mu}\|_p^q. \end{aligned} \tag{9}$$

We first assume that $p > 1$, because this will allow us to differentiate Φ_n and Φ with respect to μ_j. Denoting $\partial_j \equiv \frac{\partial}{\partial \mu_j}$ and $\nabla \equiv (\partial_1, \ldots, \partial_k)$ we can define the random variable $\nabla\widehat{\Phi}_n(\boldsymbol{\mu})$ in an obvious way. Let us also define $\delta\boldsymbol{\mu} = \boldsymbol{\mu} - \boldsymbol{\mu}_*$. The differentiability of Φ allows us to define $\boldsymbol{\mu}_*$ as the unique solution to the equation $\nabla\Phi(\boldsymbol{\mu}) = 0$, therefore:

$$\begin{aligned}
\nabla\widehat{\Phi}_n(\boldsymbol{\mu}) &= \nabla\widehat{\Phi}_n(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu}) \\
&= \nabla\widehat{\Phi}_n(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu}) - \nabla\Phi(\boldsymbol{\mu}_*) \\
&= \left[\nabla\widehat{\Phi}_n(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu}) - \nabla\Phi(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu})\right] \\
&\quad + \left[\nabla\Phi(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu}) - \nabla\Phi(\boldsymbol{\mu}_*)\right]. \qquad (10)
\end{aligned}$$

For $p > 1$, Φ is infinitely differentiable and therefore we can make a first order expansion of $\nabla\Phi$ to express the second bracket in (10) as (here, $\mathbf{e}_k$ denotes the unit vector in the k^{th} direction)

$$\begin{aligned}
\nabla\Phi(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu}) - \nabla\Phi(\boldsymbol{\mu}_*) &= \sum_{k,l} \mathbf{e}_k \delta\mu_l \partial_k \partial_l \Phi(\boldsymbol{\mu}_*) + O(\delta\boldsymbol{\mu}^2) \\
&\equiv \mathbf{M}\delta\boldsymbol{\mu} + O(\delta\boldsymbol{\mu}^2) \qquad (11)
\end{aligned}$$

The first bracket in (10) is an arithmetic mean of n i.i.d random variable and we shall apply the central limit theorem to get its asymptotic distribution when $n \to \infty$. To get the covariance matrix of the k random variable defined by the first bracket we need to compute, for arbitrary α_j, the following variance:

$$\begin{aligned}
\mathrm{Var}\left[\sum_{j=1}^{k} \alpha_j \partial_j \widehat{\Phi}_n(\boldsymbol{\mu})\right] &= \mathrm{Var}\left[\sum_{j=1}^{k} \alpha_j \frac{1}{n} \sum_{i=1}^{n} \partial_j \|\mathbf{X}_i - \boldsymbol{\mu}\|_p^q\right] \\
&= \frac{1}{n}\mathrm{Var}\left[\sum_{j=1}^{k} \alpha_j \partial_j \|\mathbf{X} - \boldsymbol{\mu}\|_p^q\right] \\
&= \frac{1}{n}\sum_{i,j} \alpha_i\alpha_j \mathrm{Cov}\left[\partial_i\|\mathbf{X} - \boldsymbol{\mu}\|_p^q, \partial_j\|\mathbf{X} - \boldsymbol{\mu}\|_p^q\right] \\
&= \frac{1}{n}\sum_{i,j} \alpha_i\alpha_j (\mathrm{Cov}\left[\partial_i\|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q, \partial_j\|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q\right] \\
&\quad + O(\delta\boldsymbol{\mu})) \\
&= \frac{1}{n}\sum_{i,j} \alpha_i\alpha_j (\mathbb{E}\left[(\partial_i\|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q) \cdot (\partial_j\|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q)\right] \\
&\quad + O(\delta\boldsymbol{\mu}))
\end{aligned}$$

$$\equiv \frac{1}{n}\sum_{i,j}\left(\alpha_i\alpha_j A_{ij} + O(\delta\boldsymbol{\mu})\right), \tag{12}$$

where we used once more that $\partial_j\Phi(\boldsymbol{\mu}_*) = \mathbb{E}[\partial_j \|\mathbf{X} - \boldsymbol{\mu}_*\|_p^q] = 0$. From the central limit theorem we conclude that:

$$\begin{aligned} W_n &= \sqrt{n}\sum_{j=1}^{k}\alpha_j\left(\partial_j\widehat{\Phi}_n(\boldsymbol{\mu}) - \partial_j\Phi(\boldsymbol{\mu})\right) \\ &\xrightarrow{\mathcal{L}} W \sim \mathcal{N}\left(0, \sum_{i,j}\alpha_i\alpha_j A_{ij} + O(\delta\boldsymbol{\mu})\right) \end{aligned}$$

as $n \to \infty$, for all α_j and therefore the asymptotic joint distribution of the $k$random variable is given by:

$$\mathbf{W}_n = \sqrt{n}\left(\nabla\widehat{\Phi}_n(\boldsymbol{\mu}) - \nabla\Phi(\boldsymbol{\mu})\right) \xrightarrow{\mathcal{L}} \mathbf{W} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{A} + O(\delta\boldsymbol{\mu})\right) \text{ as } n \to \infty. \tag{13}$$

Putting everything together we get:

$$\nabla\widehat{\Phi}_n(\boldsymbol{\mu}_* + \delta\boldsymbol{\mu}) = \mathbf{M}\delta\boldsymbol{\mu} + O(\delta\boldsymbol{\mu}^2) + \frac{\mathbf{W}_n}{\sqrt{n}}. \tag{14}$$

Let us reparametrize the problem introducing $\boldsymbol{\gamma} = \sqrt{n}\delta\boldsymbol{\mu}$ and the following sequence of random variable

$$\mathbf{Z}_n = \sqrt{n}\left[\nabla\widehat{\Phi}_n\left(\boldsymbol{\mu}_* + \frac{\boldsymbol{\gamma}}{\sqrt{n}}\right) - \nabla\Phi\left(\boldsymbol{\mu}_* + \frac{\boldsymbol{\gamma}}{\sqrt{n}}\right)\right].$$

From (13) we deduce that $\mathbf{Z}_n \xrightarrow{\mathcal{L}} \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{A})$as $n \to \infty$and (14) becomes:

$$\sqrt{n}\nabla\widehat{\Phi}_n\left(\boldsymbol{\mu}_* + \frac{\boldsymbol{\gamma}}{\sqrt{n}}\right) = \mathbf{M}\boldsymbol{\gamma} + O\left(\frac{\boldsymbol{\gamma}^2}{\sqrt{n}}\right) + \mathbf{Z}_n.$$

First assume $p > 1$. In that case $\nabla\widehat{\Phi}_n$is differentiable, so we can define a new random variable which we denote by $\widehat{\boldsymbol{\gamma}}_n$, as the function (on Ω) defined implicitly by the equation:

$$\nabla\widehat{\Phi}_n\left(\boldsymbol{\mu}_* + \frac{\widehat{\boldsymbol{\gamma}}_n}{\sqrt{n}}\right) = \mathbf{0}.$$

Letting $n \to \infty$, we obtain an equation for the limit $\widehat{\boldsymbol{\gamma}}$of $\widehat{\boldsymbol{\gamma}}_n$as $n \to \infty$:

$$\mathbf{M}\widehat{\boldsymbol{\gamma}} + \mathbf{Z} = \mathbf{0},$$

which immediately implies

$$\widehat{\gamma} \sim \mathcal{N}(\mathbf{0}, \mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-1}). \tag{15}$$

Thus the theorem is proved for $p > 1$. To conclude for $p = 1$ and $q > 1$, we use a simple continuity argument. In section 2 we saw that Φ has in that case a unique minimizer $\boldsymbol{\mu}_*^{(1,q)}$ and that $\widehat{\Phi}_n(\omega)$ a.s. has a unique minimizer $\boldsymbol{\mu}_n^{(1,q)}$ if n is large enough. These results followed from strict convexity of Φ and of $\widehat{\Phi}_n(\omega)$ for large enough n. Using exactly the same convergence argument as in the proof of Theorem 1 we can show that $\lim_{p \xrightarrow{>} 1} \boldsymbol{\mu}_*^{(p,q)} = \boldsymbol{\mu}_*^{(1,q)}$ and (for large n) $\lim_{p \xrightarrow{>} 1} \boldsymbol{\mu}_n^{(p,q)} = \boldsymbol{\mu}_n^{(1,q)}$. The theorem thus also holds in the case $p = 1$ and $q > 1$.

4. Relative Performances of L_p-means and L_p-medians

If μ_1 and μ_2 are respectively the median and the mean of a univariate data set $\mathbf{x} = (x_1, ..., x_n)'$, Stavig and Gibbons (1977) introduced the following coefficients

$$\mu_{1(2)adv} = \frac{\sum_{i=1}^{n} |x - \mu_2| - \sum_{i=1}^{n} |x - \mu_1|}{\sum_{i=1}^{n} |x - \mu_2|} \tag{16}$$

$$\mu_{2(1)adv} = \frac{\sum_{i=1}^{n} (x - \mu_1)^2 - \sum_{i=1}^{n} (x - \mu_2)^2}{\sum_{i=1}^{n} (x - \mu_1)^2} \tag{17}$$

to measure the relative advantage of the median over the mean, respectively of the mean over the median. The first coefficient is the proportional reduction in the sum of the absolute values obtained when the median is used instead of the mean, while the second one is the proportional reduction in the sum of square deviations obtained when the mean is used instead of the median. The net advantage coefficient is then defined to be $\delta_{1(2)} = \mu_{1(2)adv} - \mu_{2(1)adv}$ which can be used as a quantitative criterion for selecting between the median and the mean to describe the center of a univariate data set. The rule is to choose the median if $\delta_{1(2)} > 0$ and to choose the mean if $\delta_{1(2)} < 0$. Stavig and Gibbons (1977) found that $\delta_{1(2)}$ is positively correlated with the classical coefficients of skewness and kurtosis of the data set $\mathbf{x}$ and hence that large skewness or large kurtosis leads to choose the median as a measure of centrality rather than the mean. Note that this result is consistent with the well known "robustness" of the median since high coefficients of skewness and kurtosis may indicate the presence of outliers in the data set, cases where the median is generally preferred to the mean.

Using the notation of Section 1, (16) and (17) may be rewritten as follows:

$$\mu_{1(2)adv} = \frac{\|\mathbf{y}(\mu_2)\|_1 - \|\mathbf{y}(\mu_1)\|_1}{\|\mathbf{y}(\mu_2)\|_1}$$
$$\mu_{2(1)adv} = \frac{\|\mathbf{y}(\mu_1)\|_2^2 - \|\mathbf{y}(\mu_2)\|_2^2}{\|\mathbf{y}(\mu_1)\|_2^2}.$$

Now, if $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)'$ is a multivariate data set with L_1-median $\boldsymbol{\mu}_{11}$, L_1-mean $\boldsymbol{\mu}_{12}$, L_2-median $\boldsymbol{\mu}_{21}$ and L_2-mean $\boldsymbol{\mu}_{22}$, it seems natural to define the coefficients

$$\mu_{11(12)adv} = \frac{\|\mathbf{y}_1(\boldsymbol{\mu}_{12})\|_1 - \|\mathbf{y}_1(\boldsymbol{\mu}_{11})\|_1}{\|\mathbf{y}_1(\boldsymbol{\mu}_{12})\|_1}$$
$$\mu_{11(21)adv} = \frac{\|\mathbf{y}_1(\boldsymbol{\mu}_{21})\|_1 - \|\mathbf{y}_1(\boldsymbol{\mu}_{11})\|_1}{\|\mathbf{y}_1(\boldsymbol{\mu}_{21})\|_1}$$
$$\mu_{11(22)adv} = \frac{\|\mathbf{y}_1(\boldsymbol{\mu}_{22})\|_1 - \|\mathbf{y}_1(\boldsymbol{\mu}_{11})\|_1}{\|\mathbf{y}_1(\boldsymbol{\mu}_{22})\|_1}$$

to measure the relative advantages of the L_1-median over respectively the L_1-mean, the L_2-median and the L_2-mean,

$$\mu_{12(11)adv} = \frac{\|\mathbf{y}_1(\boldsymbol{\mu}_{11})\|_2^2 - \|\mathbf{y}_1(\boldsymbol{\mu}_{12})\|_2^2}{\|\mathbf{y}_1(\boldsymbol{\mu}_{11})\|_2^2}$$
$$\mu_{12(21)adv} = \frac{\|\mathbf{y}_1(\boldsymbol{\mu}_{21})\|_2^2 - \|\mathbf{y}_1(\boldsymbol{\mu}_{12})\|_2^2}{\|\mathbf{y}_1(\boldsymbol{\mu}_{21})\|_2^2}$$
$$\mu_{12(22)adv} = \frac{\|\mathbf{y}_1(\boldsymbol{\mu}_{22})\|_2^2 - \|\mathbf{y}_1(\boldsymbol{\mu}_{12})\|_2^2}{\|\mathbf{y}_1(\boldsymbol{\mu}_{22})\|_2^2}$$

to measure the relative advantages of the L_1-mean over respectively the L_1-median, the L_2-median and the L_2-mean,

$$\mu_{21(11)adv} = \frac{\|\mathbf{y}_2(\boldsymbol{\mu}_{11})\|_1 - \|\mathbf{y}_2(\boldsymbol{\mu}_{21})\|_1}{\|\mathbf{y}_2(\boldsymbol{\mu}_{11})\|_1}$$
$$\mu_{21(12)adv} = \frac{\|\mathbf{y}_2(\boldsymbol{\mu}_{12})\|_1 - \|\mathbf{y}_2(\boldsymbol{\mu}_{21})\|_1}{\|\mathbf{y}_2(\boldsymbol{\mu}_{12})\|_1}$$
$$\mu_{21(22)adv} = \frac{\|\mathbf{y}_2(\boldsymbol{\mu}_{22})\|_1 - \|\mathbf{y}_2(\boldsymbol{\mu}_{21})\|_1}{\|\mathbf{y}_2(\boldsymbol{\mu}_{22})\|_1}$$

to measure the relative advantages of the L_2-median over respectively the L_1-median, the L_1-mean and the L_2-mean, and

$$\begin{aligned}\mu_{22(11)adv} &= \frac{\|\mathbf{y}_2(\mu_{11})\|_2^2 - \|\mathbf{y}_2(\mu_{22})\|_2^2}{\|\mathbf{y}_2(\mu_{11})\|_2^2}\\ \mu_{22(12)adv} &= \frac{\|\mathbf{y}_2(\mu_{12})\|_2^2 - \|\mathbf{y}_2(\mu_{22})\|_2^2}{\|\mathbf{y}_2(\mu_{12})\|_2^2}\\ \mu_{22(21)adv} &= \frac{\|\mathbf{y}_2(\mu_{21})\|_2^2 - \|\mathbf{y}_2(\mu_{22})\|_2^2}{\|\mathbf{y}_2(\mu_{21})\|_2^2}\end{aligned}$$

to measure the relative advantages of the L_2-mean over respectively the L_1-median, the L_1-mean and the L_2-median. In order to select one of these four criteria for describing the center of $\mathbf{X}$ we can then calculate the net advantage coefficients

$$\delta_{11(12)} = \mu_{11(12)adv} - \mu_{12(11)adv} \quad (18)$$
$$\delta_{11(21)} = \mu_{11(21)adv} - \mu_{21(11)adv} \quad (19)$$
$$\delta_{11(22)} = \mu_{11(22)adv} - \mu_{22(11)adv} \quad (20)$$
$$\delta_{12(21)} = \mu_{12(21)adv} - \mu_{21(12)adv} \quad (21)$$
$$\delta_{12(22)} = \mu_{12(22)adv} - \mu_{22(12)adv} \quad (22)$$
$$\delta_{21(22)} = \mu_{21(22)adv} - \mu_{22(21)adv} \quad (23)$$

and apply the following rule: choose the L_1-median if (18), (19) and (20) are positive, choose the L_1-mean if (18) is negative and (21) and (22) positive, choose the L_2-median if (19) and (21) are negative and (23) positive and choose the L_2-mean if (20), (22) and (23) are negative. For the other cases, the choice is not straightforward. For example, if (18), (22) and (23) are negative and (19), (20) and (21) positive, we have the following conclusion: the L_1-mean is preferable to the L_1-median which is preferable to the L_2-mean which is itself preferable to the L_1-mean, all of these three criteria beeing preferable to the L_2-median. In such a case, a solution is to compute the following coefficients

$$\begin{aligned}\delta_{11} &= +(18) + (19) + (20)\\ \delta_{12} &= -(18) + (21) + (22)\\ \delta_{21} &= -(19) - (21) + (23)\\ \delta_{22} &= -(20) - (22) - (23)\end{aligned}$$

and to choose the L_1-median, the L_1-mean, the L_2-median or the L_2-mean if respectively $\delta_{11}, \delta_{12}, \delta_{21}$ or δ_{22} is the largest coefficient of the four.

This rule was used to select one of these four criteria for describing the center of bivariate data sets of size $n = 11$, $n = 31$ and $n = 51$ drawn from

Table 1 : Number of times that each criteria was chosen over 1000 Monte Carlo bivariate samples for different sample sizes and different populations.

Population	n	L_1-median	L_1-mean	L_2-median	L_2-mean
Bivariate Uniform	11	120	287	16	577
	31	54	324	2	620
	51	30	317	4	649
Bivariate Normal	11	198	352	21	429
	31	118	374	13	495
	51	70	469	6	455
Bivariate Logistic	11	219	376	26	379
	31	146	430	14	410
	51	126	481	8	385
Bivariate Laplace	11	290	397	35	278
	31	251	497	25	227
	51	248	501	22	229

populations distributed successively as independent bivariate uniform, normal, logistic and Laplace using the S-Plus random generator. Table 1 gives the number of times that each of the criteria was chosen over 1000 Monte Carlo samples. Results show that in uniform cases, the L_2-mean is the most preferable criterion of the four. In normal and in logistic cases, the L_2-mean is found to be preferable when n is small while the L_1-mean is the most chosen criteria when n is larger. In Laplace cases, except for very small data sets for which the L_1-median can be considered, the L_1-mean should generally be chosen to describe the center of the data set.

Open problem: In the univariate case it is well known that there is no général inequality between the variance of the mean $\widehat{\mu}_n$ and the median $\widehat{M_n}$ of a sample $X_1, \ldots, X_n$. For instance if the distribution f is normal $\mathcal{N}(0,1)$, one has $\text{Var}(\widehat{\mu}_n) < \text{Var}(\widehat{M_n})$ whereas for a Cauchy law $C(0,1)$ the mean does not even exist. So, in the multivariate case we would like to ask: "How can we optimize the choice of the integers (p,q), in terms of smaller variance, for a given p.d.f. f, or how can we compare the variances $\mathbf{V}$ for various choices of indices (p,q) ?"

References

BROWN, B.M. (1983), "Statistical uses of spatial median", *Journal of the Royal Statistical Society Series B, 45*, 25-30.

BROWN, B.M., HALLAND, P. and YOUNG, A.(1997), "On the effects of Inliers on the Spatial Median", *Journal of Multivariate Analysis 63*, 88-104

CHAKRABORTY, B. (2001), "On affine equivariant multivariate quantiles", *Annals of Institute of Statistical Mathematics, 53*, 380-403.

CHAUDHURI, P. (1992), "Multivariate location estimation using extension of R-estimates through U-statistics type approach", *Annals of Statististics 20*, 1803-1827.

DODGE, Y. and ROUSSON, V. (1998), "Multivariate L_1-mean", *Metrika, 49*, 127-134.

EELLS, W.C. (1930), "A mistaken conception of the center of the population", *Journal of the American Statistical Association 25*, 33-40.

GENTLEMAN, W.M. (1965), *Robust Estimation of Multivariate Location by Minimizing pth Power Deviations*. Ph. D. thesis, Princeton University.

GINI, C. and GALVANI, L. (1929), "Di talune estensioni dei concetti di media ai caratteri qualitativi", *Metron, 8*, 3-209.

GOWER, J.G. (1974)., "The mediancentre", *Applied Statistics 23*, 466-470.

HABERMAN, S.J. (1989), "Concavity and estimation", *Annals of Statistics, 17*, 1631-1661.

HALDANE, J.B.S. (1948), "Note on the median of a multivariate distribution", *Biometrika, 35*, 414-415.

HAYFORD, J.F. (1902), "What is the center of an area, or the center of the population?", *Journal of the American Statististical Association 8*, 47-58.

MOOD, A.M. (1941), "On the joint distribution of the median in the samples from a multivariate population", *Annals of Mathematical Statistics 12*, 268-278.

MOSLER, K. (2002), *Multivariate Dispersion, Central Regions and Depth. The Lift Zoniod Approach*, Springer-Verlag, New-York.

OJA, G. and NIINIMAA, A. (1985)., "Asymptotic properties of the generalized median in the case of multivariate normality", *Journal of the Royal Statistical Society Series B, 47*, 372-377.

POLLARD, D. (1984), *Convergence of Stochastic Processes*. Springer-Verlag, New-York.

SMALL, C.G. (1990), "A survey of multidimentional medians", *International Statistical Review, 58*, 263-277.

STAVIG, G.R. and GIBBONS, J.D. (1977)., "Comparing mean and median as measures of centrality", *International Statistical Review 45*, 63-70.