Journal de la Société Française de Statistique
Volume 150, numéro 2, 2009

Estimation de quantiles géométriques conditionnels et non conditionnels

Mohamed Chaouch[1], Ali Gannoun[2] et Jérôme Saracco[3]

Title

Nonparametric estimation of non-conditional or conditional geometric quantiles

Résumé

L'absence d'un critère pour ordonner les observations représente un obstacle pour étendre la définition classique des quantiles univariés au cas multidimensionnel. Dans le cadre d'études biomédicales ou industrielles, par exemple, on cherche souvent à déterminer le quantile d'un vecteur aléatoire conditionnellement à un autre. Plusieurs définitions des quantiles (conditionnels) multivariés, ne reposant pas sur une relation d'ordre, ont été proposées dans la littérature statistique. Dans cet article, nous nous focalisons sur la notion de quantile géométrique et de quantile géométrique conditionnel, fondée sur la minimisation d'une fonction de perte.

Mots-clés : *algorithmes de calcul, estimateur à noyau, contours, quantile géométrique, quantile géométrique conditionnel, Transformation-Retransformation*

Abstract

Lack of objective basis for ordering multivariate observations is a major problem in extending the notion of quantiles in a multidimensional setting. Conditional quantiles are required in various biomedical or industrial problems. Numerous alternative definitions of (conditional) quantile for multidimensional variables, have been proposed in statistical literature. In this article, we focus on the notion of geometric quantile and conditional geometric quantile, based on the minimization of a loss function.

Keywords : *algorithm, geometric quantile,conditional geometric quantile, kernel estimator, contour plot, tranformation-retransformation estimate*

Mathematics Subject Classification: (62G05, 62H11, 62G20)

[1]Institut de Mathématiques de Bourgogne (UMR CNRS 5584), Université de Bourgogne, Mohamed.Chaouch@u-bourgogne.fr

[2]CEDRIC, CNAM, Paris, ali.gannoun@cnam.fr

[3]GREThA (UMR CNRS 5113) & Institut de Mathématiques de Bordeaux (UMR CNRS 5251), Université de Bordeaux, Jerome.Saracco@math.u-bordeaux1.fr

1 Introduction

Les quantiles univariés, conditionels ou non conditionnels, sont fréquemment utilisés en Statistique. Par exemple, la médiane est un indicateur robuste de la tendance centrale d'une population, l'intervalle interquartile est un bon indicateur de sa dispersion. Dans la pratique, les domaines d'utilisation des quantiles sont assez variés. En biologie, Gannoun *et al* [22] utilisent les quantiles conditionnels pour estimer des courbes de référence permettant d'analyser certaines propriétés biophysiques de la peau. Les quantiles représentent également un moyen robuste de prévision (voir par exemple De Gooijer *et al* [16] et Gannoun *et al* [24]). En pratique, ces quantiles sont calculés suivant un critère d'ordre sur les observations. Un rappel sur les caractérisations des quantiles univariés sera présenté à la Section 2. L'ordre n'étant pas total sur $\mathbb{R}^d$, une extension de la définition classique des quantiles au cas où les observations sont à valeurs dans $\mathbb{R}^d$, avec $d \geq 2$, ne peut être que partielle. Il s'agit dans ce cas du vecteur quantile (dit *"arithmétique"*) dont les composantes sont les quantiles marginaux. Cette définition souffre de plusieurs faiblesses. Notamment, elle n'est pas invariante par rotation et elle ne tient pas compte de l'existence possible de corrélations entre les différentes composantes des vecteurs des observations. Le problème d'ordre des données multivariées est assez ancien. Plusieurs auteurs se sont attelés à le résoudre. Nous citons par exemple les travaux de Barnett [4], Plackett [33], Reiss [34], Eddy [19], [20]. Brown et Hettmansperger [7], [8] ont introduit la notion de quantile bivarié en se basant sur la définition de la médiane d'Oja [32]. Par la suite, Babu et Rao [2] et Abdous et Theodorescu [1] ont généralisé la notion de quantile pour un vecteur aléatoire. Cependant, cette définition ne tient pas compte de la géométrie des points, de plus elle n'est pas invariante par rotation.

Récemment, deux approches principales ont été développées pour définir des quantiles multivariés qui soient invariants par transformation affine. La première approche est basée sur la fonction de profondeur (en anglais "depth function") ; nous citons à ce propos les travaux d'Oja [32] pour la médiane et ceux de Donoho et Gasko [18], Liu *et al* [31] et Zuo et Serfling [37] pour les quantiles multivariés. La seconde approche a été utilisée, en premier lieu, par Brown [6], Gower [25], Haldane [26] et Chaudhuri [11] pour généraliser la notion de la médiane au cas multivarié. Ensuite, Abdous et Theodorescu [1], Chaudhuri [12], Koltchinskii [30] et Kokic *et al* [29] ont proposé différentes généralisations des quantiles multivariés. Cette approche définit le quantile comme un M-estimateur qui minimise une fonction de perte (ou de coût). Pour une description plus détaillée des différentes méthodes ainsi qu'une comparaison entre elles, le lecteur peut se référer à l'article de Serfling [35].

Dans ce qui suit, nous nous focalisons sur la définition des quantiles, dit *géométriques*, introduite par Chaudhuri [12]. Les quantiles géométriques sont invariants par rotation, cependant ils ne le sont pas par transformation affine. La technique dite de *Transformation-Retransformation* (TR) permet d'avoir des quantiles géométriques invariants par rotation et tranformation affine (voir par exemple Chakraborty [9] et Gannoun *et al* [23]). Ces différents points sont décrits à la Section 3. Une utilisation des quantiles géométriques en statistique descriptive multivariée est disponible dans Serfling [36].

Dans le cadre d'études industrielles ou biomédicales par exemple, une variable d'intérêt $\mathbf{Y}$ à valeurs dans $\mathbb{R}^d$ (par exemple la pression artérielle avec ses deux composantes : la pression systolique et la pression diastolique) peut être concomitante à une variable explicative $\mathbf{X}$ à

valeurs dans $\mathbb{R}^s$ (par exemple l'âge et le poids du patient). Dans ce cas, il est question de définir et d'étudier les quantiles géométriques conditionnels multivariés de $\mathbf{Y}$ sachant $\mathbf{X}$. Ceci est l'objet de la Section 4 où nous proposons une généralisation, dans le cas conditionnel, du théorème 2.1.2 de Chaudhuri [12] et de l'algorithme d'estimation correspondant. Dans la Section 5, nous décrivons l'implémentation des différents algorithmes sous le logiciel **R**. Des exemples sur des données simulées sont présentés afin d'illustrer les notions présentées dans les sections précédentes à la Section 6. Enfin nous donnons, dans la Section 7, un exemple d'application sur des données environnementales.

2 Quantile univarié

2.1 Définition

Pour une variable $Y \in \mathbb{R}$, la fonction quantile se définit à partir de l'inverse de sa fonction de répartition. Quand cette fonction de répartition est strictement croissante, son inverse est défini sans ambiguïté. Mais une fonction de répartition reste constante sur tout intervalle dans lequel la variable aléatoire ne peut pas prendre de valeurs. De manière générale, si $F(.)$ désigne la fonction de répartion de la variable Y, on appelle fonction quantile de Y la fonction qui, à $p \in (0,1)$, associe

$$Q_F(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\}, \tag{1}$$

où $F^{-1}(.)$ est souvent appelée l'inverse généralisée de $F(.)$.

2.2 Deux caractérisations du quantile univarié

2.2.1 Le quantile en tant que racine d'une équation

Soit $p \in]0,1[$, posons $u = 2p-1$. On introduit une nouvelle fonction quantile notée $Q(.)$ définie sur l'intervalle $(-1,1)$ par :

$$Q(u) = F^{-1}\left(\frac{1+u}{2}\right) \quad \text{avec} \quad -1 < u < 1. \tag{2}$$

Nous remarquons que, contrairement à la définition donnée par (1), le quantile est indexé par $u \in (-1,1)$. La définition de la fonction quantile $Q(.)$ donnée par (2) nous donne, à l'aide du signe (resp. la valeur absolue) de u, une idée sur l'orientation (resp. l'ordre) du quantile par rapport à la médiane. En effet :

- pour $u = 0$, $Q(0)$ est la médiane (le quantile d'ordre $p = 1/2$),
- si u est négatif (resp. positif), le quantile d'ordre u est à gauche (resp. à droite) de la médiane.
- si $|u|$ est proche de 0, le quantile correspondant est proche de la médiane (quantile d'ordre $p = 1/2$) ; si $|u|$ est proche de 1, le quantile correspondant est un quantile "extrême" (quantile d'ordre p proche de 0 ou de 1).

Il est facile de montrer que $Q_F(p) = Q(u)$, pour $u = 2p-1$, est solution de l'équation suivante dont l'inconnue est θ

$$\mathbb{E}\left(S(\theta - Y)\right) - u = 0, \tag{3}$$

où S désigne la fonction "signe" univariée définie par

$$S\left(\theta - Y\right) = \left\{ \begin{array}{rl} 1 & \text{si } \theta - Y > 0, \\ -1 & \text{si } \theta - Y < 0. \end{array} \right.$$

Par convention, on pose $S(0) = 0$.

Démonstration.

$$\begin{array}{rcl} F\left(F^{-1}\left(p\right)\right) - p & = & \mathbb{P}\left(Y \leq F^{-1}\left(p\right)\right) - p \\ & = & \mathbb{E}(\mathbb{1}_{\{Y \leq F^{-1}(p)\}}) - p \\ & = & \mathbb{E}(\mathbb{1}_{\{Y \leq Q(u)\}}) - p \\ & = & \mathbb{E}(\mathbb{1}_{\{Q(u) - Y \geq 0\}} - \frac{1+u}{2}) \\ & = & \frac{1}{2}\mathbb{E}([2\mathbb{1}_{\{Q(u) - Y \geq 0\}} - 1] - u) \\ & = & \frac{1}{2}\mathbb{E}(S\left(Q\left(u\right) - Y\right) - u) \end{array}$$

Comme $F\left(F^{-1}\left(p\right)\right) - p = 0$, on en déduit que, pour un u fixé, le quantile $Q_F(p) = Q\left(u\right)$ est bien la solution de l'équation (3). □

2.2.2 Le quantile en tant que solution d'un problème de minimisation

Ferguson [21] et Koenker et Basset [28], dans le cadre des quantiles de régression, définissent le quantile comme la solution du problème de minimisation suivant. Soient $p \in]0,1[$ une probabilité fixée et $u = 2p-1$. On note $\phi(u,t) = |t| + ut$, pour tout couple $(u,t) \in]-1,1[\times\mathbb{R}$, la fonction dite *de coût* ou *de perte*. La fonction quantile de Y, notée $Q_M(.)$, est alors définie comme suit

$$\begin{array}{rcl} Q_M\left(u\right) & = & \arg\min\limits_{\theta \in \mathbb{R}} \mathbb{E}\{\phi\left(u, Y - \theta\right)\} \\ & = & \arg\min\limits_{\theta \in \mathbb{R}} \int_{\mathbb{R}} \left(|y - \theta| + u\left(y - \theta\right)\right) F\left(dy\right). \end{array} \tag{4}$$

A partir de cette équation, on peut montrer que $Q_M(u)$ est aussi la solution de l'équation suivante dont l'inconnue est θ

$$\mathbb{E}\left(S\left(Y - \theta\right)\right) + u = 0,$$

qui est équivalente à l'équation (3).

Par conséquent, pour $u = 2p-1$, les trois caractérisations sont équivalentes :

$$Q_M(p) = Q(u) = Q_F(p).$$

Remarque 2.1. Contrairement à la définition du quantile donnée par l'équation (1), l'approche de minimisation permet facilement de généraliser la notion de quantile dans le cadre multidimensionnel.

2.3 Estimation

On considère un échantillon $Y_1, \ldots, Y_n$, de n observations de Y dans $\mathbb{R}$. Un estimateur non paramétrique de la fonction de répartition F de Y est donné par

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}}.$$

Ainsi, pour $p \in]0,1[$, un estimateur de $Q_F(p)$ est

$$Q_{F_n}(p) = F_n^{-1}(p) = \inf\{y : F_n(y) \geq p\}.$$

Pour $u = 2p - 1$, la caractérisation donnée par l'équation (3) permet de définir un estimateur $Q_n(u)$ de $Q(u)$ comme la solution θ de l'équation suivante

$$\frac{1}{n}\sum_{i=1}^{n} S(\theta - Y_i) = u. \tag{5}$$

Il est facile de voir que $Q_n(u) = Q_{F_n}(\frac{1+u}{2}) = Q_{F_n}(p)$. En effet, nous avons

$$F_n\left(F_n^{-1}(p)\right) - p = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{Y_i \leq F_n^{-1}(p)\}} - p\right) = \frac{1}{2n}\sum_{i=1}^{n}\left[S(Q_n(u) - Y_i) - u\right].$$

En utilisant l'équation (4), pour $u = 2p - 1$, $Q_M(p)$ est estimé par

$$Q_{M,n}(u) = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} \phi\left(u, Y_i - \theta\right) = \arg\min_{\theta\in\mathbb{R}} \sum_{i=1}^{n} |Y_i - \theta| + u\left(Y_i - \theta\right).$$

On peut montrer que $Q_{M,n}(u)$ est solution de l'équation (5) dont l'inconnue est θ. Ainsi, pour $u = 2p - 1$, ces trois estimateurs sont identiques :

$$Q_{M,n}(u) = Q_n(u) = Q_{F_n}(p).$$

3 Quantile géométrique

On suppose maintenant que $\mathbf{Y} \in \mathbb{R}^d$. La définition donnée en (1), reposant sur la notion de relation d'ordre total dans $\mathbb{R}$, ne peut pas être étendue à $\mathbb{R}^d$ du fait que l'ordre n'est pas total sur cet espace. Dans ce cadre, Chaudhuri [12] a proposé une définition du quantile multivarié, dit *géométrique*, qui généralise la définition du quantile univarié donnée en (4).

Dans ce qui suit, les symboles $||.||$ et $\langle ., . \rangle$ désignent la norme et le produit scalaire Euclidiens. Les vecteurs sont considérés comme étant des matrices colonnes et le symbole "T" désignera la transposée d'une matrice.

3.1 Définition

Considérons donc la fonction de perte *multivariée* définie par

$$\phi(\mathbf{u}, \mathbf{t}) = ||\mathbf{t}|| + \langle \mathbf{u}, \mathbf{t} \rangle ,$$

avec $\mathbf{t} \in \mathbb{R}^d$ et $\mathbf{u} \in B^d = \left\{ \mathbf{u} \in \mathbb{R}^d \ : ||\mathbf{u}|| < 1 \right\}$. Le quantile géométrique, indexé par le vecteur $\mathbf{u}$, est défini par la relation suivante

$$\mathbf{Q}(\mathbf{u}) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left\{ \phi(\mathbf{u}, \mathbf{Y} - \theta) \right\}.$$

La fonction $\mathbb{E}\left\{ \phi(\mathbf{u}, \mathbf{Y} - \theta) \right\}$ n'est définie que si $\mathbb{E}\|\mathbf{Y}\| < \infty$. Utilisant un artifice de Kemperman [27], la fonction $\mathbb{E}\left\{ \phi(\mathbf{u}, \mathbf{Y} - \theta) - \phi(\mathbf{u}, \mathbf{Y}) \right\}$ l'est toujours. Ces deux fonctions admettent le même minimum quand celui-ci existe. Ceci permet de définir le quantile comme suit :

$$\mathbf{Q}(\mathbf{u}) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left\{ \phi(\mathbf{u}, \mathbf{Y} - \theta) - \phi(\mathbf{u}, \mathbf{Y}) \right\}.$$

Soit maintenant $\mathbf{S}(.)$ la fonction définie de $\mathbb{R}^d$ dans $\mathbb{R}^d$ par $\mathbf{S}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|$ si $\mathbf{v} \neq 0$, avec par convention $\mathbf{S}(0) = 0$. De manière analogue au cas univarié, on peut montrer que le quantile géométrique est solution de l'équation suivante dont l'inconnue est θ :

$$\mathbb{E}\left(\mathbf{S}(\theta - \mathbf{Y}) \right) - \mathbf{u} = 0.$$

3.2 Estimation

Soit F_n l'estimateur empirique (non paramétrique) de F obtenu à partir des observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ de $\mathbf{Y} \in \mathbb{R}^d$. Pour tout $\mathbf{u} \in B^d$, on peut définir un estimateur $\mathbf{Q}_n(\mathbf{u})$ de $\mathbf{Q}(\mathbf{u})$ par :

$$\begin{aligned} \mathbf{Q}_n(\mathbf{u}) &= \arg\min_{\theta \in \mathbb{R}^d} \int \left(\phi(\mathbf{u}, \mathbf{y} - \theta) - \phi(\mathbf{u}, \mathbf{y}) \right) F_n(d\mathbf{y}) \\ &= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\phi(\mathbf{u}, \mathbf{Y}_i - \theta) - \phi(\mathbf{u}, \mathbf{Y}_i) \right) \end{aligned} \tag{6}$$

De plus, si $\mathbb{E}\|\mathbf{Y}\| < \infty$, on a

$$\mathbf{Q}_n(\mathbf{u}) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta). \tag{7}$$

On arrive facilement, à partir de l'équation (6), à montrer que $\mathbf{Q}_n(\mathbf{u})$ est aussi solution de l'équation suivante dont l'inconnue est θ :

$$\frac{1}{n} \sum_{i=1}^{n} S(\theta - \mathbf{Y}_i) = \mathbf{u}. \tag{8}$$

3.3 Existence et unicité de $\mathbf{Q}_n(\mathbf{u})$

Puisque $\sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta)$ tend vers l'infini quand $\|\theta\| \to \infty$ et $\sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta)$, en tant que fonction de θ, est continue, alors la fonction $\sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta)$ possède au moins un minimum. Ensuite, sous l'hypothèse que les observations $\{\mathbf{Y}_i, i = 1, \ldots, n\}$ ne se sont pas alignées, par le théorème 2.17 de Kemperman [27], la fonction $\sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta)$ est strictement convexe en θ. Ceci prouve que l'estimateur du quantile géométrique $\mathbf{Q}_n(\mathbf{u})$ existe et est unique.

3.4 Interprétation du vecteur $\mathbf{u}$

Le vecteur $\mathbf{u}$ permet de donner des informations sur le quantile $\mathbf{Q}(\mathbf{u})$ et son estimateur $\mathbf{Q}_n(\mathbf{u})$. En effet, $\mathbf{u}$ étant un vecteur de B^d,

- sa norme nous renseigne sur l'"ordre" du quantile : si $||\mathbf{u}|| \approx 1$ (resp. 0), alors $\mathbf{Q}(\mathbf{u})$ (ou $\mathbf{Q}_n(\mathbf{u})$) est un quantile "extrême" (resp. "central", i.e. proche de la médiane géométrique).
- sa direction nous indique la position du quantile par rapport à la médiane.

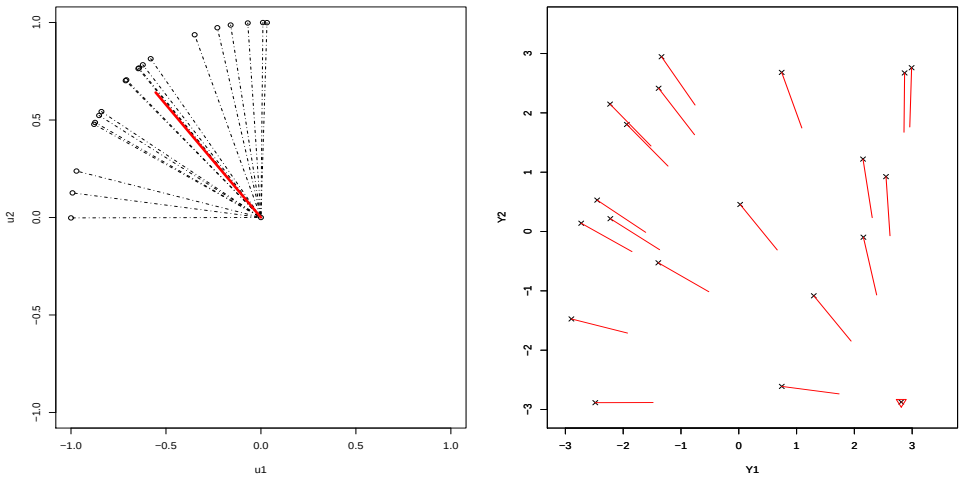


FIGURE 1 – *A droite, le nuage des* 20 *observations sur lequel chaque segment représente le vecteur normé* $S(\theta - \mathbf{Y}_i)$ *qui relie une observation* i *au quantile géométrique* $\mathbf{Q}(\mathbf{u})$ *(qui est le point représenté par un triangle et situé en bas à droite). A gauche, le vecteur* $\mathbf{u}$ *(trait continu) est la moyenne des vecteurs unitaires* $S(\theta - \mathbf{Y}_i)$ *(tracés en pointillés).*

Les figures 1 et 2 donnent une illustration graphique de l'interprétation du vecteur $\mathbf{u}$. Pour chacune des deux figures, nous avons simulé 20 observations $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{20}\}$ de $\mathbb{R}^2$ indépendantes telles que chaque composante est générée selon la loi uniforme $U_{[-3,3]}$, les deux composantes étant indépendantes l'une de l'autre. Nous avons fixé un point de ce nuage qui sera considéré comme un quantile géométrique. Ensuite, à l'aide de l'équation (8), nous déterminons le vecteur $\mathbf{u}$ correspondant à ce quantile. Ce vecteur $\mathbf{u}$ peut être vu comme la moyenne de tous les vecteurs unitaires $S(\mathbf{Q}(\mathbf{u}) - \mathbf{Y}_i)$, pour $i = 1, \ldots, 20$, qui partent d'une observation i (de coordonnées $\mathbf{Y}_i$) vers le quantile géométrique $\mathbf{Q}(\mathbf{u})$. Nous remarquons que si $\mathbf{Q}(\mathbf{u})$ est un point "hors norme" (voir la figure 1 à droite avec le point de coordonnées (2.8, -2.8)), son vecteur $\mathbf{u}$

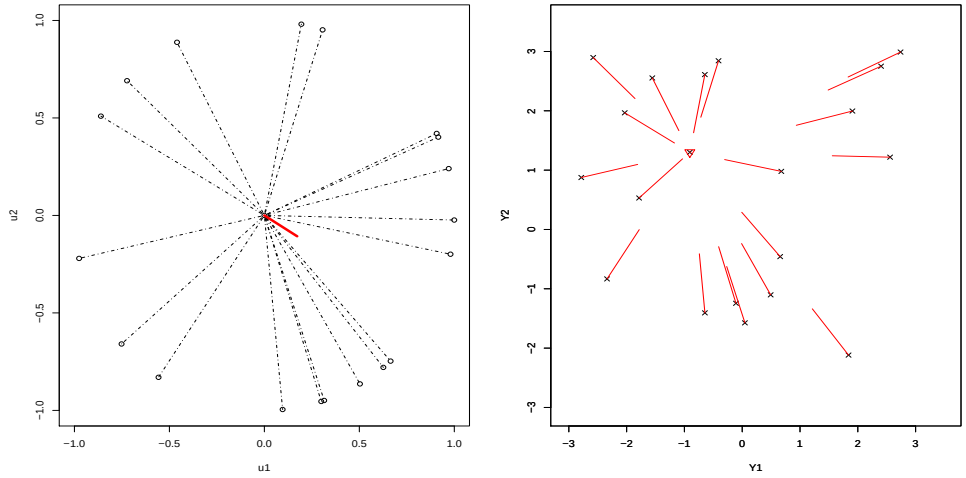


FIGURE 2 – *A droite, le nuage des* 20 *observations sur lequel chaque segment représente le vecteur normé* $S(\theta - \mathbf{Y}_i)$ *qui relie une observation* i *au quantile géométrique* $\mathbf{Q}(\mathbf{u})$ *(qui est le point représenté par un triangle et situé au centre de nuage des points). A gauche, le vecteur* $\mathbf{u}$ *(trait continu) est la moyenne des vecteurs unitaires* $S(\theta - \mathbf{Y}_i)$ *(tracés en pointillés).*

(voir le vecteur en trait continu sur la figure 1 à gauche) a une norme proche de 1. Si $\mathbf{Q}(\mathbf{u})$ est un point plus central, (voir la figure 2 à droite avec le point de coordonnées (-0.9, 1.3)), il lui correspondra un vecteur $\mathbf{u}$ (voir le vecteur en trait continu sur la figure 2 à gauche) de norme proche de 0.

Remarque 3.1 (Serfling [35]). Le terme $||\mathbf{u}||$ (appelé en anglais "extent of deviation") ne doit pas être pris comme la distance Euclidienne entre $\mathbf{Q}(\mathbf{u})$ et la médiane spatiale $\mathbf{M} = \mathbf{Q}(0)$. De plus, la distance entre $\mathbf{Q}(\mathbf{u})$ et $\mathbf{M}$ ne croît pas forcément en fonction de $||\mathbf{u}||$.

Remarque 3.2 ([35]). Contrairement au cas univarié où $u = 2p - 1$, la "grandeur" $||\mathbf{u}||$ ne porte aucune interprétation probabiliste lorsque $d \geq 2$. En particulier, considérons la région $\mathcal{R} = \{\mathbf{Q}_n(\mathbf{u}) \ : ||\mathbf{u}|| \leq 0.5\}$. Dans le cas univarié, elle correspond à la région interquartile avec $\frac{1}{4} \leq p \leq \frac{3}{4}$. Par contre dans le cas multivarié, cet ensemble ne contient pas forcément 50% des observations.

La remarque 3.1 signifie que la région $\mathcal{R}$ représente mal la forme du support de la distribution en particulier lorsque celle-ci est (trés) allongée. A ce niveau, ceci est l'inconvénient majeur des quantiles géométriques. Les deux exemples suivants ont pour but d'illustrer numériquement les deux remarques ci-dessus. L'exemple 3.2 (resp. l'exemple 3.3) correspond à la remarque (3.1) (resp. à la remarque (3.2)).

Exemple 3.3 (Serfling [35]). Soit $F = \frac{1}{2}F_1 + \frac{1}{2}F_2$, avec F_1 et F_2 deux distributions uniformes univariées respectivement sur $[-100, 0]$ et $[0, 1]$.

On calcule les quantiles suivants :

$M = 0$, $Q\left(\frac{1}{2}\right) = F^{-1}\left(\frac{3}{4}\right) = \frac{1}{2}$, $Q\left(-\frac{1}{2}\right) = F^{-1}\left(\frac{1}{4}\right) = -50$ et $Q(-0.1) = F^{-1}(0.45) = -10$.

Modèle	Pourcentage
$(M1) : \begin{pmatrix} Y1 \sim \mathcal{N}(0,1) \\ Y2 \sim \mathcal{N}(0,1) \end{pmatrix}$	$\begin{cases} 0\% & \text{pour } n^* = 100 \\ 3\% & \text{pour } 70 \leq n^* < 100 \\ 97\% & \text{pour } n^* < 70 \end{cases}$
$(M2) : \begin{pmatrix} Y1 \sim \mathcal{N}(200, \sigma = 1) \\ Y2 \sim -2 * Y1 + \mathcal{N}(200, \sigma = 3) \end{pmatrix}$	$\begin{cases} 0\% & \text{pour } n^* = 100 \\ 24\% & \text{pour } 85 \leq n^* < 100 \\ 76\% & \text{pour } n^* < 85 \end{cases}$
$(M3) : \begin{pmatrix} Y1 \sim \mathcal{N}(200, \sigma = 1) \\ Y2 \sim -2 * Y1 + \mathcal{N}(200, \sigma = 0.01) \end{pmatrix}$	$\begin{cases} 99\% & \text{pour } n^* = 100 \\ 1\% & \text{pour } 85 \leq n^* < 100 \\ 0\% & \text{pour } n^* < 85 \end{cases}$

TABLEAU 1 – *Pourcentage des simulations où l'ensemble* $\mathcal{R} = \{\boldsymbol{Q}_n(\boldsymbol{u}) : ||\boldsymbol{u}|| \leq 0.5\}$ *contient* 50% *des observations.*

- Pour $u = \pm\frac{1}{2}$, on a $|u| = \frac{1}{2}$ mais les quantiles correspondants $Q(\frac{1}{2}) = \frac{1}{2}$ et $Q(-\frac{1}{2}) = -50$ ne sont pas équidistants par rapport à la médiane.
- Pour $u_1 = -0.1$ et $u_2 = \frac{1}{2}$ on a $|u_1| < |u_2|$ mais $|Q(-0.1)| > |Q\left(\frac{1}{2}\right)|$. On observe donc ici que la distance Euclidienne entre le quantile et la médiane ne croît pas en fonction de $|u|$.

Exemple 3.4. A travers cet exemple, nous illustrons à l'aide de plusieurs simulations la seconde remarque. Considérons un vecteur aléatoire bidimensionnel $\mathbf{Y} = (Y_1, Y_2)$. Nous avons simulé des échantillons de taille $n = 200$ en considérant différentes lois . Ensuite, pour chacun des échantillons simulés, nous avons déterminé le nombre d'observations (noté n^*) qui appartiennent à l'ensemble $\mathcal{R}$. Les différentes lois considérées pour $\mathbf{Y}$ et les résultats obtenus sont résumés dans la Table 1. Sur les 100 simulations qui ont été réalisées avec les modèles $(M1)$ et $(M2)$, l'ensemble $\mathcal{R}$ ne contient jamais la moitié des observations. En revanche, pour le modèle $(M3)$, on voit que pour 99 % des simulations, $\mathcal{R}$ contient 50% des observations. Ceci vient tout simplement du fait que ce modèle se réduit au cas univarié et l'ensemble $\mathcal{R}$ n'est autre que l'intervalle interquartile qui contient la moitié des observations.

3.5 Résultats asymptotiques

Soient $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, n-observations indépendantes et identiquement distribuées suivant la même loi que $\mathbf{Y}$. Soit $\mathbf{Q_n}(\mathbf{u})$ l'estimateur de $\mathbf{Q}(\mathbf{u})$ calculé à partir de ces observations. Chaudhuri [12] a établi une représentation de type Bahadur (voir Bahadur [3]) de cet estimateur. Cette représentation a été utilisée pour le comportement asymptotique de $\mathbf{Q_n}(\mathbf{u})$. Ces résultats sont l'objet des deux théorèmes qui vont suivre. L'énoncé de ces théorèmes nécessitent l'introduction des notations suivantes.

Soit $P(\mathbf{v}) = \frac{1}{\|\mathbf{v}\|}\left(I_d - S(\mathbf{v})S^T(\mathbf{v})\right)$ pour $\mathbf{v} \neq 0$, où I_d est la matrice identité d'ordre d. Soit $D_1(\theta) = \mathbb{E}\{P(\mathbf{Y} - \theta)\}$, pour tout $\theta \in \mathbb{R}^d$, une matrice de dimension $d \times d$, symétrique.

L'espérance qui définit $D_1(\theta)$ est finie, pour $d \geq 2$, quand la densité de $\mathbf{Y}$ est bornée sur tout compact de $\mathbb{R}^d$. Ceci est une conséquence immédiate du fait que $\int_{\mathbb{R}^d} \frac{1}{\|\mathbf{x}-\mathbf{y}\|} f(\mathbf{x})d\mathbf{x} < \infty$ pour toute densité f bornée sur tout compact de $\mathbb{R}^d$. Pour tous vecteurs $\theta_1, \theta_2 \in \mathbb{R}^d$ et $\mathbf{u}, \mathbf{v} \in B^d$, on note par $D_2(\theta_1, \theta_2, \mathbf{u}, \mathbf{v})$ la matrice de dimension $d \times d$, définie par $D_2(\theta_1, \theta_2, \mathbf{u}, \mathbf{v}) = \mathbb{E}\left\{[S(\mathbf{Y}-\theta_1)+\mathbf{u}]\,[S(\mathbf{Y}-\theta_2)+\mathbf{v}]^T\right\}$.

Théorème 3.5 (Chaudhuri [12]). *Soient $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ une suite de vecteurs aléatoires, de $\mathbb{R}^d$, indépendants et identiquement distribués suivant une densité bornée sur tout ensemble compact de $\mathbb{R}^d$. Pour tout vecteur $\mathbf{u} \in B^d$ fixé, nous avons la représentation suivante de type Bahadur :*

$$\mathbf{Q}_n(\mathbf{u}) - \mathbf{Q}(\mathbf{u}) = n^{-1}\left[D_1(\mathbf{Q}(\mathbf{u}))\right]^{-1} \times \sum_{i=1}^{n} [S(\mathbf{Y_i} - \mathbf{Q}(\mathbf{u})) + \mathbf{u}] + R_n(\mathbf{u}),$$

avec $R_n(\mathbf{u}) = O(\log n/n)$ quand $d \geq 3$ et $R_n(\mathbf{u}) = o(n^{-\beta})$ quand $d = 2$, où $0 < \beta < 1$.

Théorème 3.6 (Chaudhuri [12]). *Soient $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k$, des vecteurs de la boule unitaire ouverte B^d. Sous les hypothèses du théorème 3.1, la loi jointe asymptotique du k-uplet,*

$$\sqrt{n}\left(\mathbf{Q}_n(\mathbf{u}_1) - \mathbf{Q}(\mathbf{u}_1)\right), \ldots, \sqrt{n}\left(\mathbf{Q}_n(\mathbf{u}_k) - \mathbf{Q}(\mathbf{u}_k)\right),$$

est une loi multinormale centrée dont les termes de covariances, entre les vecteurs

$$\sqrt{n}\left(\mathbf{Q}_n(\mathbf{u}_r) - \mathbf{Q}(\mathbf{u}_r)\right) \quad \textit{et} \quad \sqrt{n}\left(\mathbf{Q}_n(\mathbf{u}_s) - \mathbf{Q}(\mathbf{u}_s)\right) \quad \textit{où} \quad 1 \leq r, s \leq k,$$

sont donnés par : $\left[D_1(\mathbf{Q}_n(\mathbf{u}_r))\right]^{-1}\left[D_2(\mathbf{Q}(\mathbf{u}_r), \mathbf{Q}(\mathbf{u}_s), \mathbf{u}_r, \mathbf{u}_s)\right]\left[D_1(\mathbf{Q}_n(\mathbf{u}_s))\right]^{-1}$.

Ces théorèmes permettent de conclure que $\mathbf{Q}_n(\mathbf{u})$ est un estimateur consistant du quantile géométrique $\mathbf{Q}(\mathbf{u})$ avec une vitesse de convergence de l'ordre de $1/\sqrt{n}$. De plus, cet estimateur est asymptotiquement normal.

Cependant, l'inconvénient majeur du quantile géométrique (Chaudhuri [12], Koltchinskii [30]) est qu'il n'est pas invariant par transformation affine. Aussi si les composantes qui forment le vecteur $\mathbf{Y}$ n'ont pas les mêmes unités de mesure ou qu'elles ont des variations assez différentes les unes des autres, l'estimation du quantile géométrique ne donne pas des résultats convenables. Pour remédier à cette défaillance (lorsque l'on s'éloigne du cadre d'une distribution sphérique des données), une technique d'estimation, dite de Transformation-Retransformation (TR), est proposée dans la littérature. Nous allons la présenter dans la paragraphe suivant.

3.6 Technique de Transformation-Retransformation (TR)

Cette approche a été introduite, dans un premier temps, par Chaudhuri et Sengupta [13], pour construire des tests de signe multivariés invariants par transformation affine. Ensuite, elle a été utilisée par Chakraborty et Chaudhuri [10], Gannoun *et al* [23] et De Gooijer et Gannoun [15] pour donner une version invariante par transformation affine de la médiane spatiale et la médiane spatiale conditionnelle. Chakraborty [9] a généralisé cette technique dans le cadre de l'estimation du quantile géométrique.

Soient $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$, n-observations de $\mathbb{R}^d$, avec $n > d+1$. On note $\alpha = \{i_0, i_1, \ldots, i_d\}$ un sous-ensemble de $(d+1)$ indices inclus dans $\{1, 2, \ldots, n\}$. On définit la matrice suivante $\mathbf{Y}(\alpha) = [\mathbf{Y}_{i_1} - \mathbf{Y}_{i_0}, \ldots, \mathbf{Y}_{i_d} - \mathbf{Y}_{i_0}]$. Cette matrice de dimension $d \times d$ sert à transformer le reste des points $\mathbf{Y}_j, j \notin \alpha$, en les exprimant dans un nouveau système de coordonnées de façon suivante : $\mathbf{Y}_j^{(\alpha)} = [\mathbf{Y}(\alpha)]^{-1} \mathbf{Y}_j$, c'est l'étape de la transformation (T). Notons que, si la distribution de probabilité des $\mathbf{Y}_j$ est absolument continue par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, la matrice $\mathbf{Y}(\alpha)$ est inversible. Cette étape de transformation des données nécessite également une transformation du vecteur $\mathbf{u}$ qui indexe le quantile géométrique $\mathbf{Q}(\mathbf{u})$:

$$\mathbf{v}(\alpha) = \begin{cases} \frac{\|\mathbf{u}\|}{\|[\mathbf{Y}(\alpha)]^{-1}\mathbf{u}\|} [\mathbf{Y}(\alpha)]^{-1} \mathbf{u} & \text{si } \mathbf{u} \neq 0 \\ 0 & \text{si } \mathbf{u} = 0. \end{cases}$$

Cette transformation du vecteur $\mathbf{u}$ dans le nouveau système de coordonnée devrait être de la forme $[\mathbf{Y}(\alpha)]^{-1} \mathbf{u}$, comme cela a été fait pour la matrice des données. Cependant rien ne garantit que le vecteur $\mathbf{u}$ ainsi transformé appartiendra à la boule unitaire ouverte B^d. Pour cette raison, la pondération $\|\mathbf{u}\| / \| [\mathbf{Y}(\alpha)]^{-1} \mathbf{u}\|$ a été ajoutée dans la formule de $\mathbf{v}(\alpha)$. Une fois ces transformations faites, l'estimateur du quantile géométrique d'ordre $\mathbf{v}(\alpha)$, noté $\mathbf{R}_n^{(\alpha)}(\mathbf{v})$ est calculé à partir des observations $\mathbf{Y}_j^{(\alpha)}, j \notin \alpha$, au moyen de la formule (7). Ensuite, par une étape de Retransformation (R), l'estimateur du quantile géométrique, d'ordre $\mathbf{u}$ est donné par $\mathbf{Q}_n^{(\alpha)}(\mathbf{u}) = \mathbf{Y}(\alpha)\mathbf{R}_n^{(\alpha)}(\mathbf{v})$ dans le système de coordonnées d'origine. Le théorème ci-aprés garantit le bon fonctionnement théorique de cette approche. La validité pratique de la méthode TR sera illustrée par des simulations à la Section 6.

Théorème 3.7 (Chakraborty [9])**.** *Soit n-vecteurs aléatoires $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n \in \mathbb{R}^d$ transformés en $A\mathbf{Y}_1 + \mathbf{b}, \ldots, A\mathbf{Y}_n + \mathbf{b}$, où A est une matrice non singulière de dimension $d \times d$ et $\mathbf{b} \in \mathbb{R}^d$. Soit $\mathbf{w} = (\|\mathbf{u}\| / \|A\mathbf{u}\|) A\mathbf{u}$. Le TR quantile géométrique d'ordre $\mathbf{w}$, calculé à partir des observations $A\mathbf{Y}_1 + \mathbf{b}, \ldots, A\mathbf{Y}_n + \mathbf{b}$, est égal à $A\mathbf{Q}_n^{(\alpha)}(\mathbf{u}) + \mathbf{b}$, où $\mathbf{Q}_n^{(\alpha)}(\mathbf{u})$ est le TR quantile géométrique d'ordre $\mathbf{u}$, calculé à partir des observations $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$.*

3.6.1 Choix de α

L'estimateur $\mathbf{Q}_n^{(\alpha)}(\mathbf{u})$ dépend du choix de la matrice de transformation $\mathbf{Y}(\alpha)$, et par conséquent du choix de α. Chakraborty [9] propose de le choisir tel que la matrice $[\mathbf{Y}(\alpha)]^T \Sigma^{-1} \mathbf{Y}(\alpha)$ soit proche d'une matrice diagonale de la forme λI_d, où $\lambda > 0$, la matrice Σ étant la matrice de variances covariances de $\mathbf{Y}$.

La matrice $\mathbf{Y}(\alpha)$ est choisie telle qu'elle minimise le ratio entre la moyenne arithmétique et la moyenne géométrique des valeurs propres de la matrice définie positive $[\mathbf{Y}(\alpha)]^T \widehat{\Sigma}^{-1} \mathbf{Y}(\alpha)$, où $\widehat{\Sigma}$ est un estimateur convergent de Σ.

Notons que la moyenne arithmétique (resp. la moyenne géométrique) des valeurs propres d'une matrice symétrique est égale à sa trace (resp. son déterminant). En pratique, nous n'avons pas besoin de balayer tous les sous-ensembles α de $\{1, \ldots, n\}$. Nous nous arrêterons au premier sous-ensemble qui donne une valeur du ratio qui soit inférieur à $1 + \epsilon$, où ϵ est un réel assez petit fixé par l'utilisateur. Des simulations ont montré que cette procédure n'affecte pas la qualité des estimateurs.

3.7 Un algorithme d'estimation

Le problème de calcul de la médiane spatiale comme étant la quantité qui minimise l'expression $\sum_{i=1}^{n} ||\mathbf{Y}_i - \mathbf{Q}||$ a été abordé par Bedall et Zimmermann [5] et Gower [25]. Des algorithmes de minimisation ont été proposés par ces auteurs. Récemment Chaudhuri [12] a proposé, en modifiant légèrement l'algorithme de Newton-Raphson pour déterminer les racines d'une équation multivariée, un algorithme itératif permettant de calculer l'estimateur du quantile géométrique correspondant à une direction $\mathbf{u}$ fixée. Cet algorithme est basé sur le résultat suivant.

Théorème 3.8 (Chaudhuri [12])**.** *Considérons un n-échantillon* $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ *avec* $\boldsymbol{Y}_i \in \mathbb{R}^d$ *et* $\boldsymbol{Q}_n(\boldsymbol{u})$ *un estimateur du quantile géométrique* $\boldsymbol{Q}(\boldsymbol{u})$.

- Si $\boldsymbol{Q}_n(\boldsymbol{u}) \neq \boldsymbol{Y}_i, \ \forall\, 1 \leq i \leq n$, *alors on a :*

$$\sum_{i=1}^{n} S(\boldsymbol{Y}_i - \boldsymbol{Q}_n(\boldsymbol{u})) + n\boldsymbol{u} = 0.$$

- Si $\exists\, 1 \leq i \leq n$ *tel que* $\boldsymbol{Q}_n(\boldsymbol{u}) = \boldsymbol{Y}_i$, *alors on a :*

$$\left\| \sum_{\substack{1 \leq i \leq n \\ \boldsymbol{Y}_i \neq \boldsymbol{Q}_n(\boldsymbol{u})}} [S(\boldsymbol{Y}_i - \boldsymbol{Q}_n(\boldsymbol{u})) + \boldsymbol{u}] \right\| \leq \sum_{\substack{1 \leq i \leq n \\ \boldsymbol{Y}_i = \boldsymbol{Q}_n(\boldsymbol{u})}} (1 + ||\boldsymbol{u}||).$$

Cet algorithme comporte deux étapes :

– *Etape 1.* Pour chaque $1 \leq i \leq n$, on teste la condition suivante :

$$\left\| \sum_{\substack{1 \leq j \leq n \\ j \neq i}} [S(\mathbf{Y}_j - \mathbf{Y}_i)] + (n-1)\mathbf{u} \right\| \leq (1 + ||\mathbf{u}||). \tag{9}$$

Si cette condition est vérifiée pour un certain i, alors $\mathbf{Q}_n(\mathbf{u}) = Y_i$.
Sinon il faut aller à l'étape 2.

– *Etape 2.* Cette étape consiste à résoudre, par une méthode itérative, l'équation

$$\sum_{i=1}^{n} S(\mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u})) + n\mathbf{u} = 0. \tag{10}$$

Notons par $\mathbf{Q}_n^{(1)}(\mathbf{u})$ une approximation initiale de $\mathbf{Q}_n(\mathbf{u})$. En pratique nous pouvons, par exemple, prendre pour $\mathbf{Q}_n^{(1)}(\mathbf{u})$ le vecteur des médianes (marginales) empiriques des d composantes de $\mathbf{Y}$, calculées à partir des observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$.
Soient $\mathbf{Q}_n^{(1)}(\mathbf{u}), \ldots, \mathbf{Q}_n^{(m)}(\mathbf{u})$ les approximations successives de $\mathbf{Q}_n(\mathbf{u})$ obtenues après les m premières itérations. La $(m+1)^{\text{ème}}$ approximation est calculée de la manière suivante.
Soient

$$\Delta = \sum_{i=1}^{n} S(\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u})) + n\mathbf{u},$$

et

$$\Phi = \sum_{i=1}^{n} P(\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u})),$$

Dans le cas où les observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ ne sont pas toutes alignées, la matrice Φ est définie positive, et dans ce cas, on pose :

$$\mathbf{Q}_n^{(m+1)}(\mathbf{u}) = \mathbf{Q}_n^{(m)}(\mathbf{u}) + \Phi^{-1}\Delta.$$

On arrête les itérations quand on obtient deux approximations successives très proches. En général, l'algorithme converge au bout d'une dizaine d'itérations.

Remarque 3.9. Une généralisation des théorèmes 3.3 et 3.4 ainsi que l'algorithme de calcul, dans le cadre de la procédure d'estimation par Transformation-Retransformation, est détaillée dans l'article de Chakraborty [9].

4 Quantile géométrique conditionnel

4.1 Définition

Considérons n observations $\{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}$ d'un couple de vecteurs aléatoires ($\mathbf{X}$,$\mathbf{Y}$) à valeurs dans $\mathbb{R}^s \times \mathbb{R}^d$. Il est d'usage de rechercher la relation qui peut exister entre le vecteur à expliquer $\mathbf{Y}$ et la covariable multidimensionnelle $\mathbf{X}$. Les quantiles géométriques conditionnels représentent un moyen pour aborder ce problème.

Dans le cas univarié, c'est à dire lorsque Y à valeurs dans $\mathbb{R}$, il existe une grande variété d'approches paramétriques ou non paramétriques permettant d'estimer les quantiles conditionnels univariés. Citons par exemple, parmi les méthodes non paramétriques, celle du noyau, celle de la constante locale et celle du noyau produit (voir Gannoun *et al* [22]) pour une rapide description de ces méthodes).

L'intérêt pour les quantiles multivariés conditionnels est tout à fait récent. De Gooijer *et al* [17] ont généralisé au cadre conditionnel la définition du quantile spatial basée sur la minimisation de la semi-norme donnée par Abdous et Theodorescu [1]. Cheng et De Gooijer [14] se sont aussi intéressés au même problème en généralisant la définition du quantile géométrique, introduite par Chaudhuri [12]. Dans ce qui suit, nous nous focalisons sur cette dernière défintion.

Soit $\mathbf{u} \in B^d$, le quantile géométrique conditionnel de $\mathbf{Y}$ sachant $\mathbf{X} = \mathbf{x}$, indexé par $\mathbf{u}$, est défini par :

$$\begin{aligned} \mathbf{Q}(\mathbf{u}|\mathbf{x}) &= \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\phi(\mathbf{u}, \mathbf{Y} - \theta) - \phi(\mathbf{u}, \mathbf{Y}) \mid \mathbf{X} = \mathbf{x}\right] \\ &= \arg\min_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} \{\phi(\mathbf{u}, \mathbf{y} - \theta) - \phi(\mathbf{u}, \mathbf{y})\}\, G(d\mathbf{y}|\mathbf{x}). \end{aligned}$$

où G est la distribution conditionnelle de $\mathbf{Y}$ sachant $\mathbf{X}$. De la même manière qu'à la section précédente, ce quantile peut être vu comme l'unique solution de l'équation dont l'inconnue est θ :

$$\mathbb{E}\left(S\left(\theta - \mathbf{Y}\right) \mid \mathbf{X} = \mathbf{x}\right) = \mathbf{u}. \tag{11}$$

4.2 Estimation

Soit $G_n(.|\mathbf{x})$ un estimateur non paramétrique de type Nadaraya-Watson de la distribution conditionnelle de $\mathbf{Y}$ sachant $\mathbf{X} = \mathbf{x}$, défini pour tout $\mathbf{y} \in \mathbb{R}^d$, par

$$G_n(\mathbf{y}|\mathbf{x}) = \sum_{i=1}^{n} w_{n,i} \mathbb{1}_{\{\mathbf{Y}_i \leq \mathbf{y}\}},$$

avec $w_{n,i} = K\left((\mathbf{x} - \mathbf{X}_i)/h_n\right) / \sum_{i=1}^{n} K\left((\mathbf{x} - \mathbf{X}_i)/h_n\right)$ représentant le poids associé à $\mathbf{Y}_i$, où le noyau K est une application de $\mathbb{R}^s$ dans $\mathbb{R}$, bornée, intégrable par rapport à la mesure de Lebesgue et d'intégrale 1(on choisit souvent pour K un noyau produit, i.e. un produit de noyau unidimensionnel qui sont généralement des densités de probabilité) et le paramètre h_n est la fenêtre de lissage. Lorsque X est unidimensionnelle, (h_n) est une suite de réels positifs tendant vers zéro pour n tendant vers l'infini. Quand $\mathbf{X}$ est multidimensionnelle, on peut choisir une largeur de fenêtre $h_{n,j}$ spécifique à chaque composante X_j de $\mathbf{X}$; cependant très souvent on choisit une même fenêtre h_n commune à l'ensemble des composantes. Nous nous sommes plaçés dans ce cadre afin de simplifier l'écriture de l'estimateur $G_n(\mathbf{y}|\mathbf{x})$. Les poids $w_{n,i}$ sont autant plus importants pour les $\mathbf{Y}_i$ tels que $\mathbf{X}_i$ est proche de $\mathbf{x}$.

A partir de l'estimateur $G_n(.|\mathbf{x})$, on en déduit un estimateur $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ de $\mathbf{Q}(\mathbf{u}|\mathbf{x})$:

$$\begin{aligned}
\mathbf{Q}_n(\mathbf{u}|\mathbf{x}) &= \arg\min_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} \{\phi(\mathbf{u}, \mathbf{y} - \theta) - \phi(\mathbf{u}, \mathbf{y})\} G_n(d\mathbf{y}|\mathbf{x}) \\
&= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} w_{n,i} \{\phi(\mathbf{u}, \mathbf{Y}_i - \theta) - \phi(\mathbf{u}, \mathbf{Y}_i)\}
\end{aligned}$$

A partir de l'équation (11), l'estimateur $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ peut être regardé comme la solution de l'équation suivante dont l'inconnue est θ :

$$\int S(\theta - \mathbf{t}) G_n(d\mathbf{t}|\mathbf{x}) = \sum_{i=1}^{n} w_{n,i} S(\theta - \mathbf{Y}_i) = \mathbf{u}.$$

Nous nous donnons maintenant deux propriétés asymptotiques de l'estimateur $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$.

4.3 Propriétés asymptotiques de $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$

Cheng et De Gooijer [14] ont établi, sous certaines hypothèses, une représentation de type Bahadur de $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$. Ils en ont déduit la loi asymptotique de cet estimateur. Dans un premier temps, nous reportons les hypothèses sous lesquelles les différents résultats ont été établis.

(H1) Pour tout vecteur $\mathbf{t}$ appartenant à un voisinage de $\mathbf{x}$ noté $N(\mathbf{x})$, la densité conditionnelle $f(.|\mathbf{x})$ est bornée sur tout ensemble de $\mathbb{R}^d$.

(H2) La loi marginale de $\mathbf{X}$ admet une densité $g(.)$ continue et strictement positive au point $\mathbf{x}$.

(H3) Il existe trois réels positifs c_1, c_2, c_3 tels que $c_1 \mathbb{1}_{\{\|\mathbf{z}\| \leq c_3\}} \leq K(\mathbf{z}) \leq c_2 \mathbb{1}_{\{\|\mathbf{z}\| \leq c_3\}}$, pour $\mathbf{z} \in \mathbb{R}^s$, et $\int K(\mathbf{z}) d\mathbf{z} = 1$ et $\int \mathbf{z} K(\mathbf{z}) d\mathbf{z} = 0$.

(H4) $nh_n^s \sim Cn^\gamma$ pour tout $C > 0$ et $0 < \gamma < 1$ et $\limsup_{n\to\infty} nh_n^{s+4} < \infty$.
(H5) Pour tout $\mathbf{t} \in N(\mathbf{x})$ et $\theta \in \mathbb{R}^d$, soit

$$r(\theta, \mathbf{t}) = (r_1(\theta, \mathbf{t}), \ldots, r_d(\theta, \mathbf{t}))^T = \mathbb{E}\left[S(\mathbf{Y} - \mathbf{Q}(\mathbf{u}|\mathbf{x}) - \theta) + \mathbf{u}|\mathbf{X} = \mathbf{t}\right].$$

Pour tout $M > 0$, $\sup_{\|\theta\|\leq M, \mathbf{t}\in N(\mathbf{x})} \|\partial^2 r(\theta, \mathbf{t})/\partial\mathbf{t}\partial\mathbf{t}^T\| < \infty$.
(H6) Pour un $M > 0$ assez petit et ω un paramètre positif,

$$\sup_{\mathbf{t}\in N(\mathbf{x})} \sup_{\theta:\|\theta\|\leq M} \int \frac{f(\mathbf{y}|\mathbf{t})}{\|\mathbf{y} - \theta - \mathbf{Q}(\mathbf{u}|\mathbf{x})\|^{1+\omega}} < \infty.$$

Cette relation est vérifiée pour $\omega = 1$ quand $d \geq 3$, et elle l'est aussi pour $0 < \omega < 1$ quand $d = 2$.
(H7) La fonction $\mathbf{t} \longrightarrow \mathbb{E}\left[P(\mathbf{Y} - \mathbf{Q}(\mathbf{u}|\mathbf{x}))|\mathbf{X} = \mathbf{t}\right]$ est continue au point $\mathbf{t} = \mathbf{x}$.
(H8) On suppose que la dérivée seconde de la fonction $g(.)$ est bornée sur $N(\mathbf{t})$ et que pour tout $\mathbf{t} \in N(\mathbf{x})$,
$D_t = \mathbb{E}\left[\left(S(\mathbf{Y} - \mathbf{Q}(\mathbf{u}|\mathbf{x})) + \mathbf{u}\right)\left(S(\mathbf{Y} - \mathbf{Q}(\mathbf{u}|\mathbf{x})) + \mathbf{u}\right)^T |\mathbf{X} = \mathbf{t}\right]$ est continue en tout point $\mathbf{t} = \mathbf{x}$. On suppose également que $\gamma > 2/(2 + s)$.

Théorème 4.1 (Cheng et De Gooijer [14])**.** *Sous les hypothèses* $(H1) - (H7)$*, la représentation de type Bahadur de* $\boldsymbol{Q}_n(\boldsymbol{u}|\mathbf{x})$ *est donnée par la relation suivante*

$$\boldsymbol{Q}_n(\boldsymbol{u}|\mathbf{x}) - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x}) = D_{1n}^{-1} \sum_{i=1}^{n} w_{n,i}\left[S(\mathbf{Y}_i - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x})) + \mathbf{u}\right] + R_n,$$

avec $D_{1n} = \mathbb{E}\left[K_{h_n}(\mathbf{x} - \mathbf{X})P(\mathbf{Y} - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x}))\right]/\mathbb{E}K_{h_n}(\mathbf{x}-\mathbf{X})$ *et* $K_{h_n}(\mathbf{x}-\mathbf{X}) = K((\mathbf{x}-\mathbf{X})/h_n)$. *Quand* $d \geq 3$, $R_n = O\left(\log n/nh_n^s\right)$, *et lorsque* $d = 2$, $R_n = o\left((\log n/(nh_n^s))^\omega\right)$ *pour tout* $0 < \omega < 1$.

Théorème 4.2 (Cheng et De Gooijer [14])**.** *Sous les hypothèses* $(H1) - (H8)$*, alors on a*

$$\sqrt{\frac{nh_n^s g(\mathbf{x})}{\int K^2(\mathbf{z})d\mathbf{z}}} D_x^{-1/2} D_{1n}\left(\boldsymbol{Q}_n(\boldsymbol{u}|\mathbf{x}) - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x}) - \frac{h_n^2}{2} D_{1n}^{-1}\xi_s\right) \to N(0, I_d), \text{ avec}$$

$$\xi_s = (\xi_{s1}, \ldots, \xi_{sd})^T \text{ où } \xi_{sk} = \sum_{1\leq i,j\leq s}\left(\frac{\partial^2 r_k(0,\mathbf{t})}{\partial t_i \partial t_j} + 2\frac{\partial \log g(\mathbf{t})}{\partial t_i}\frac{\partial r_k(0,\mathbf{t})}{\partial t_j}\right)\Bigg|_{\mathbf{t}=\mathbf{x}} \int z_i z_j K(\mathbf{z})d\mathbf{z}.$$

Dans le paragraphe suivant nous proposons un algorithme pour calculer un estimateur de $\mathbf{Q}(\mathbf{u}|\mathbf{x})$.

4.4 Un algorithme d'estimation du quantile géométrique conditionnel

Commençons par généraliser le théorème 3.8 au cas conditionnel.

Théorème 4.3. *Soit* $\{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$ *un* n*-échantillon de couples de vecteurs aléatoires à valeurs dans* $\mathbb{R}^s \times \mathbb{R}^d$*, avec* $n \geq d + s$*. Soit* $\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})$ *l'estimateur de* $\boldsymbol{Q}(\boldsymbol{u}|\boldsymbol{x})$.

1. *Si pour tout* $1 \leq i \leq n$, $\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x}) \neq \boldsymbol{Y}_i$, *alors on a :*

$$\sum_{i=1}^{n} S(\boldsymbol{Y}_i - \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x}))K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i) + \boldsymbol{u}\sum_{i=1}^{n} K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i) = 0 \tag{12}$$

2. *Si pour un certain* i, *on a* $\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x}) = \boldsymbol{Y}_i$, *alors*

$$\left\| \sum_{\substack{1\leq i\leq n \\ \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})\neq \boldsymbol{Y}_i}} \left[S(\boldsymbol{Y}_i - \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})) + \mathbf{u}\right] K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i) \right\| \leq \sum_{\substack{1\leq i\leq n \\ \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})= \boldsymbol{Y}_i}} K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i)(1 + ||\boldsymbol{u}||)$$

Utilisant ce théorème, l'algorithme pour le calcul de l'estimateur du quantile géométrique conditionnel se décompose en deux étapes.

– *Etape 1.* Pour chaque $1 \leq i \leq n$, on teste l'inégalité suivante :

$$\left\| \sum_{\substack{1\leq j\leq n \\ j\neq i}} \left[S(\mathbf{Y}_j - \mathbf{Y}_i) + \mathbf{u}\right] K_{h_n}(\mathbf{x} - \mathbf{X}_j) \right\| \leq K_{h_n}(\mathbf{x} - \mathbf{X}_i)(1 + ||\mathbf{u}||) \tag{13}$$

Si cette condition est satisfaite pour un certain i, alors $\mathbf{Q}_n(\mathbf{u}|\mathbf{x}) = \mathbf{Y}_i$.
Sinon on passe à l'étape suivante qui consiste à résoudre numériquement l'équation (12).

– *Etape 2.* Notons par $\mathbf{Q}_n^{(1)}(\mathbf{u}|\mathbf{x}), ..., \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x})$ des approximations successives de $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ avec comme initialisation pour $\mathbf{Q}_n^{(1)}(\mathbf{u}|\mathbf{x})$ $(\in \mathbb{R}^d)$ le vecteur des médianes (marginales) empiriques conditionnelles des d composantes de $\mathbf{Y}$, calculé à partir des observations $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$. La $(m+1)^{\text{ème}}$ approximation $\mathbf{Q}_n^{(m+1)}(\mathbf{u}|\mathbf{x})$ est calculée comme suit.
Soient

$$\Delta = \sum_{i=1}^{n} \frac{\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x})}{||\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x})||} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right) + \mathbf{u}\sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right)$$

et

$$\Phi = \sum_{i=1}^{n} \frac{1}{||\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x})||} \left[I_d - \frac{(\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x}))\,(\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x}))^T}{||\mathbf{Y}_i - \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x})||^2}\right] K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h_n}\right).$$

Si les observations $\mathbf{Y}_i$ ne sont pas alignées, alors la matrice Φ est définie positive et on pose :

$$\mathbf{Q}_n^{(m+1)}(\mathbf{u}|\mathbf{x}) = \mathbf{Q}_n^{(m)}(\mathbf{u}|\mathbf{x}) + \Phi^{-1}\,\Delta.$$

On arrête les itérations quand on obtient deux approximations successives très proches. En général, l'algorithme converge au bout d'une dizaine d'itérations.

5 Implémentation en R des algorithmes de calculs des quantiles (conditionnels) géométriques

Dans cette section, nous donnons une implémentation en **R** des algorithmes présentés dans les paragraphes précédents.

5.1 Cas du quantile géométrique $\mathbf{Q}_n(\mathbf{u})$

Le programme estimant le quantile géométrique est appelé "**QuantileNC.est**".

5.1.1 Description de la fonction "QuantileNC.est"

Paramètres de la fonction Trois paramètres doivent être donnés afin d'exécuter cette fonction :

- $MatY$: une matrice des données constituée de n lignes (nombre d'observations) et d colonnes (nombre de variables).
- **u** : un vecteur de $\mathbb{R}^d$ de norme inférieure ou égale à 1. Sa direction nous indique la position du quantile $\mathbf{Q}(\mathbf{u})$ par rapport à la médiane géométrique et sa norme indique l'ordre du quantile correspondant.
- m : un entier naturel qui représente le nombre maximal d'itérations à faire.

Rapide description de l'implémentation Le programme teste en premier lieu l'inégalité (9). Si cette inégalité est vérifiée pour une certaine observation i, on arrête l'exécution du programme et le quantile est le vecteur composé des éléments de la $i^{\text{ème}}$ ligne de $MatY$. Si l'inégalité (9) n'est pas vérifiée pour tous les $i = 1, ..., n$, alors on passe à la deuxième étape du programme qui consiste à résoudre à l'aide d'un algorithme itératif l'équation (10).

Sorties de la fonction Les sorties de cette fonctions sont :

- Le quantile géométrique, noté **Q**.
- La direction **u** pour laquelle on a calculé le quantile géométrique.
- La norme Euclidienne du vecteur **u** qui permet d'avoir une idée sur le caractère extrême ou non du quantile.
- La variable logique $test$ qui nous indique si la condition (9) était verifiée ou non. Cette variable prend la valeur $TRUE$ si la condition est verifiée et $FALSE$ sinon.

5.1.2 Procédure de Transformation-Retransformation

Cette procédure se déroule en deux étapes.

La première consiste à sélectionner le sous-ensemble optimal α de $\{1, \ldots, n\}$ selon le critère qui a été détaillé dans le paragraphe 3.6.1. La fonction permettant d'effectuer cette étape, appelée "**ChoixIndice**", dépend de deux paramètres : $MatY$ et ϵ (fixé par défaut à 0.01). Les

sorties de cette fonction sont le vecteur des indices, de dimension $d+1$, noté $indice$, ainsi que la valeur du ratio correspondant.

La deuxième étape consiste à estimer, à l'aide de la fonction "**TRversion.QuantileNC.est**", le quantile géométrique. Cette fonction dépend des paramètres $MatY$, $\mathbf{u}$, m, $indice$. Dans cette fonction, nous faisons appel à la fonction "**QuantileNC.est**" appliquée à la matrice transformée de $MatY$, c'est-à-dire $[\mathbf{Y}(\alpha)]^{-1}MatY$. En sortie, on obtient le vecteur Q correspondant quantile géométrique estimé.

5.2 Cas du quantile géométrique conditionnel $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$

Dans ce paragraphe nous présentons deux fonctions nécessaires pour calculer un estimateur du quantile géométrique conditionnel. La première fonction notée **"QuantileC.est"**, basée sur l'algorithme présenté dans le paragraphe 4.3, calcule un estimateur du quantile géométrique conditionnel. La seconde fonction **"fenetre.opt"** calcule la fenêtre optimale de lissage. Le critère de choix de cette fenêtre est décrit dans le paragraphe suivant. Dans ce qui suit, nous traitons le cas où le vecteur $\mathbf{Y} \in \mathbb{R}^d$ et la variable $X \in \mathbb{R}$.

5.2.1 Choix des paramètres de lissage

La qualité des estimateurs n'étant pas trés affectée par le choix du noyau, la densité gaussienne est utilisée comme noyau K :

$$K(z) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2}\right) \quad \text{où } z \in \mathbb{R}.$$

Ce choix est suffisant pour les résultats théoriques de convergence de Cheng et De Gooijer [14] et donne de bons résultats en simulations. Le choix de la fenêtre est crucial. La qualité des estimateurs non paramétriques basés sur les noyaux y est étroitement liée. Une importante littérature est consacrée à ce sujet et, en particulier, aux méthodes de sélection automatique par minimisation d'un critère. La méthode de validation croisée entre dans ce cadre. Pour estimer $\mathbf{Q}(\mathbf{u}|x)$, une approche dérivée du critère de validation croisée est utilisée :

$$\tilde{h} = \arg\min_{h>0} \sum_{j=1}^{n} ||\mathbf{Q}_n(\mathbf{u}|x) - \mathbf{Q}_n^{-j}(\mathbf{u}|x)||^2$$

où $\mathbf{Q}_n^{-j}(\mathbf{u}|x)$ désigne classiquement l'estimateur de $\mathbf{Q}(\mathbf{u}|x)$ calculé à partir de l'échantillon $\{(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)\}$ privé de la $j^{\text{ème}}$ observation.

5.2.2 Description des fonctions

Le calcul de l'estimatio du quantile géométrique conditionnel utilise la fonction **"QuantileC.est"**. Cette fonction nécessite cinq paramètres d'entrée qui sont :

- $MatXY$: le tableau des données, representé sous la forme d'une matrice à n lignes (individus) et $(d+1)$ colonnes (variables) telles la première colonne est celle qui correspond à la variable unidimensionnelle X et les d dernières colonnes aux d composantes de $\mathbf{Y}$.

– x : la valeur affectée à la variable réelle X qui définit la condition $X = x$
– **u** et m : mêmes paramètres que ceux de la fonction **"QuantileNC.est"**.
– h_n : la fenêtre de lissage. Cette fenêtre peut être calculée à l'aide de la fonction **"fenetre.opt"** décrite à la remarque suivante.

Les sorties de la fonction **"QuantileC.est"** sont :

– **Q** : l'estimateur du quantile géométrique de $\mathbf{Y}$ conditionnellement à $X = x$, relatif au vecteur **u**.
– $test$: une variable logique qui prend la valeur $TRUE$ si la condition (13) est verifiée et $FALSE$ sinon.
– **u** et $||\mathbf{u}||$: la direction **u** pour laquelle on a calculé le quantile géométrique conditionnel, ainsi que sa norme.

L'étape de Transformation-Retransformation décrite dans le paragraphe 5.1.2 reste valable dans le cas de l'estimation des quantiles géométriques conditionnels.

Remarque 5.1. La fonction **"fenetre.opt"** nécessite les paramètres suivants : $MatXY$, x, **u**, m et $seqhn$, une séquence de valeurs de la fenêtre. En pratique, nous préconisons pour $seqhn$ une séquence d'une dizaine de valeurs équidistribuées entre $n^{-1/5}\sigma_n$ et $2n^{-1/5}\sigma_n$ où σ_n désigne l'écart-type empirique de la variable X. En sortie, cette fonction fournit la valeur $hnopt$ correspondant à la fenêtre optimale calculée en utilisant la méthode de validation croisée donnée dans le paragraphe 5.2.1.

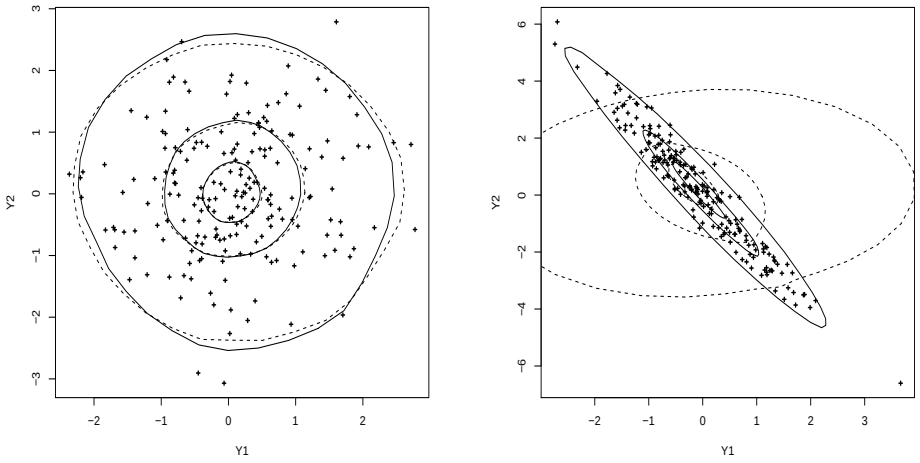
6 Etude par simulation

Dans la suite nous nous plaçons dans le cadre où $d = 2$ afin de faciliter l'interprétation et la réalisation des graphiques. L'identification des observations "extrêmes" dans un échantillon est une étape importante dans une étude statistique. Dans le cas univarié, il est possible, à l'aide du boxplot, de déterminer ces observations. Nous donnons dans cette section un graphique (fondé sur des contours) qui peut jouer un rôle équivalent à celui du boxplot dans un cadre multivarié.

Etant donné une séquence de vecteurs $\mathbf{u} \in B^2$, de même norme r et de directions différentes, nous calculons pour chaque **u** le quantile géométrique correspondant. L'ensemble $C(r) = \{\mathbf{Q}_n(\mathbf{u}) : ||\mathbf{u}|| = r\}$, avec $0 < r < 1$, est appelé contour (en anglais "quantile contour plot") de niveau r. Dans le cas où la distribution est sphérique l'ensemble $C(r)$ peut être l'équivalent du boxplot, en tant qu'un outil pour détecter les données "hors norme", dans un cadre multivarié. En effet, lorsque la norme r de **u** est proche de 1, les observations qui se situent à l'extérieur de ce contour peuvent alors être considérées comme "hors normes". En revanche, si on est loin du cadre sphérique, il faut passer par la procédure TR pour estimer les quantiles géométriques. Les contours estimés, par le biais des quantiles géométriques transformés retransformés, permettent à la fois l'identification des individus "hors normes" et la bonne description du support de la distribution (voir par exemple la figure 3 (b)). Pour les applications, le choix de r dépend du cadre de l'étude. Généralement, c'est le spécialiste qui le fixe selon ses objectifs.

6.1 Une première simulation : cas de quantiles géométriques

Pour illustrer la construction des contours, nous avons généré $n = 200$ réalisations de $\mathbf{Y} = (Y_1, Y_2)$ suivant la loi binormale centrée réduite $N_2(0, I_2)$. Soit $\mathbf{u} = (r\cos\theta, r\sin\theta)^T$ la direction suivant laquelle nous allons calculer le quantile géométrique. Fixons des valeurs de $r = 0.3, 0.6, 0.9$ et prenons une séquence de valeurs pour l'angle, $\theta = k\pi/16$ où $k = 0, 4, 8, 12, \ldots, 28$. Pour calculer les quantiles formant le contour de niveau r fixé, on fait varier l'angle θ et on obtient ainsi tous les quantiles qui forment le contour, quantiles que l'on relie ensuite par des segments de droite. La figure 3 (a) représente les trois contours de niveaux 30%, 60% et 90% estimés et le nuage de points correspondant.



(a) (b)

FIGURE 3 – *Tracé des contours de niveaux 30%, 60% et 90% estimés avec une étape de TR (resp. sans TR), en ligne continue (resp. en pointillés) pour des données provenant* (a) *d'une binormale centrée réduite,* (b) $Y_1 \sim N(0,1)$ *et* $Y_2 = -2Y_1 + \epsilon$ *avec* $\epsilon \sim N(0, 0.5)$.

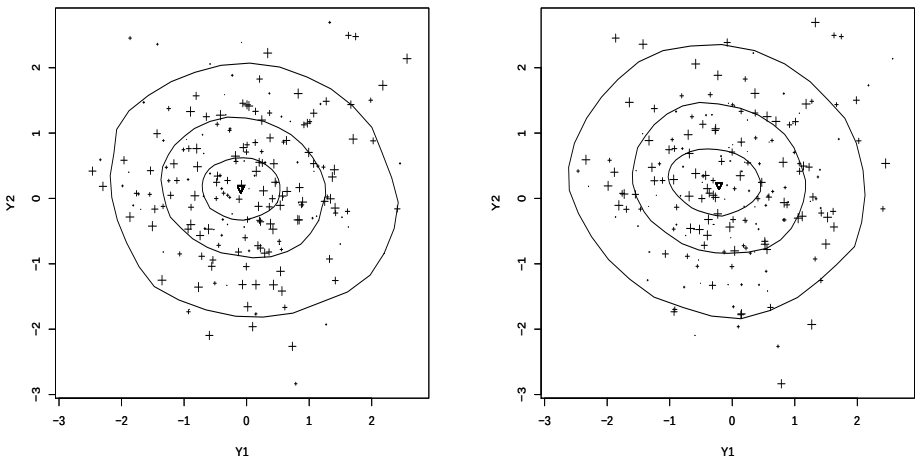
Pour illustrer l'apport de l'approche TR lors de l'estimation des quantiles géométriques dans le cas d'une distribution non sphérique, nous avons simulé $n = 200$ observations selon le modèle suivant : $Y_1 \sim N(0,1)$ et $Y_2 = -2Y_1 + \epsilon$ avec $\epsilon \sim N(0, 0.5)$. La figure 3 (b) montre bien que les contours estimés aprés l'étape TR (tracés en ligne continue) tiennent compte de la corrélation qui existe entre Y_1 et Y_2 et par conséquent ils décrivent bien le support de la distribution. Par contre les contours calculés sans passer par l'étape TR (tracés en pointillés) n'ont pas vraiment de sens car ils décrivent mal le support de la distribution dans ce cas. Nous remarquons également sur la figure 3 (a) que dans le cas d'un distribution sphérique l'étape TR n'est pas vraiment nécessaire : en effet les contours calculés avec et sans TR se superposent.

6.2 Une seconde simulation : cas de quantiles géométriques conditionnels

Dans cette section nous estimons, pour différentes directions $\mathbf{u}$, le quantile géométrique du vecteur $\mathbf{Y} = (Y_1, Y_2)$ conditionnellement à une variable unidimensionnelle X. Afin de

tracer les contours de niveaux 25%, 50% et 75%, on considère des vecteurs $\mathbf{u}$ de la forme $\mathbf{u} = (rcos\theta, rsin\theta)^T$ avec $r = 0, 0.25, 0.5, 0.75$ et $\theta = k\pi/16$ où $k = 0, 4, 8, 12, \ldots, 28$. Nous avons considéré deux modèles.

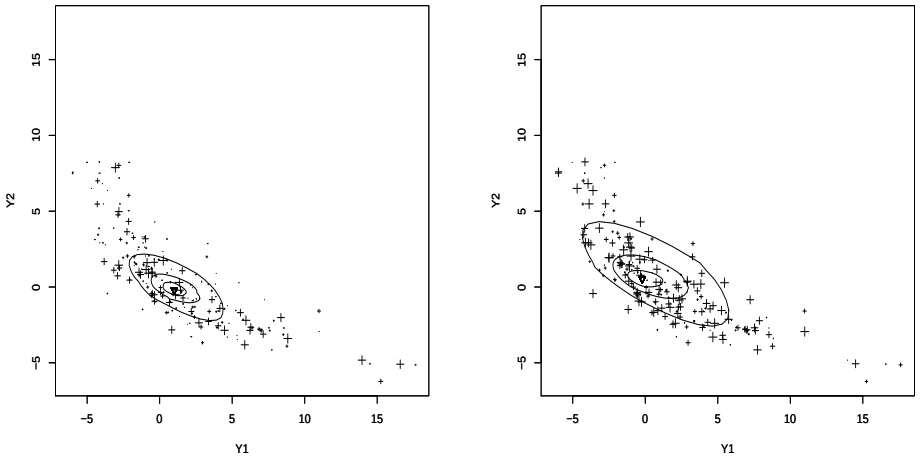
Dans le premier modèle (noté Modèle 1), on suppose que les variables Y_1, Y_2 et X suivent la loi $N(0, 1)$ et sont indépendantes. La figure 4 (a) (resp. (b)) représente l'estimation des contours de niveaux 25%, 50% et 75% de $\mathbf{Y}$ conditionnellement à $X = -0.5$ (resp. $X = 0.5$). La médiane géométrique, qui correspond au contour de niveau 0% (i.e. pour $\mathbf{u} = (0, 0)$), est représentée par un triangle. Les observations qui ont le plus de poids dans l'estimation des quantiles $\mathbf{Q}_n(\mathbf{u}|x)$, c'est à dire celles dont la valeur de X_i est proche du x fixé, sont représentées par des croix de taille d'autant plus grandes que les poids sont importants. Nous remarquons dans les deux cas que les médianes géométriques conditionnelles estimées sont voisines de la vraie médiane conditionnelle qui n'est autre que l'espérance de $\mathbf{Y}$ (égale à $(0, 0)$), les variables $\mathbf{Y}$ et X étant indépendantes. De même, les contours calculés sont très similaires quel que soit le conditionnement.



(a) $X = -0.5$ (b) $X = +0.5$

FIGURE 4 – *Tracé de la médiane géométrique conditionnelle (représentée par un triangle) et des contours de niveau 25%, 50% et 75% pour un jeu de données issues du Modèle 1, conditionnellement à* $X = -0.5$ *et* $X = +0.5$

Considérons maintenant un deuxième modèle (noté Modèle 2) défini de la façon suivante : $X \sim N(0, 1)$, $Y_1 = X^2 + 4X + \epsilon_1$ et $Y_2 = |X| - 3X + \epsilon_2$, avec $\epsilon_1 \sim N(0, 1)$ et $\epsilon_2 \sim N(0, 1)$, les variables X, ϵ_1 et ϵ_2 étant indépendantes. Les quantiles géométriques conditionnels sont estimés en utilisant la technique TR pour $X = -1$ et $X = +1$. Il apparaît clairement sur la figure 5 que le conditionnement a un effet sur les contours estimés de niveau 25%, 50% et 75%. De même, l'estimation de la médiane conditionnelle diffère pour $X = -1$ et pour $X = +1$. Cela n'a rien de surprenant au vu du Modèle 2 dans lequel le vecteur $\mathbf{Y}$ dépend de la covariable X.



(a) $X = -1$ (b) $X = +1$

FIGURE 5 – *Tracé de la médiane conditionnelle (représentée par un triangle) et des contours de niveau 25%, 50% et 75% pour des données issues du Modèle 2, conditionnellement à* $X = -1$ *et* $X = +1$

7 Étude sur des données réelles

Dans ce paragraphe nous mettons en exergue, sur un jeu de données réelles, l'avantage des quantiles géométriques par rapport aux quantiles marginaux dans la détermination des valeurs "hors normes" dans le cas d'un vecteur $\mathbf{Y}$ de dimension $d = 2$.

Présentation des données et de l'étude. Les données traitées sont celles du projet "Kola Ecogeochemistry" (pour plus de détails sur ce projet le lecteur pourra se reporter à l'adresse suivante : www.ngu.no/Kola). L'objectif de ce projet est de déterminer le taux de pollution autour d'une zone industrielle et ceci à partir des mesures de différents composants chimiques faites sur des plantes situées autour de cette zone industrielle. L'algue est une plante qui se nourrit principalement de composants chimiques qui se trouvent dans l'atmosphère. Par conséquent elle représente un bon indicateur biologique du degré de pollution. Des mesures de taux de Calcium (Ca) et de Barium (Ba) ont été faites sur un échantillon de $n = 594$ plantes. Le Barium est un composant chimique trés utilisé dans l'industrie et sa dissolution dans l'eau peut provoquer des problèmes respiratoires et cardiaques. La première étape pour les chercheurs consiste à identifier les observations "hors normes" pour le couple de variables aléatoires (taux de Ba, taux de Ca).

Travail effectué Pour répondre à cette problématique, une idée naturelle consiste à déterminer, de façon indépendante, les quantiles d'ordres 25% et 75% correspondant à chacune des variables. Ces quantiles sont notés $q^{Ca}_{0.25}$, $q^{Ca}_{0.75}$, $q^{Ba}_{0.25}$ et $q^{Ba}_{0.75}$. Les sommets du rectangle représenté à la figure 6 sont les points de coordonnées $(q^{Ba}_{0.25}, q^{Ca}_{0.25})$, $(q^{Ba}_{0.75}, q^{Ca}_{0.25})$, $(q^{Ba}_{0.25}, q^{Ca}_{0.75})$ et $(q^{Ba}_{0.75}, q^{Ca}_{0.75})$.

Les observations qui se trouvent en dehors de ce rectangle peuvent être considérées comme “hors normes”. D’autre part nous avons tracé le contour de niveau 75% (basé sur les quantiles géométriques du vecteur (Ba, Ca) calculés pour différentes directions $\mathbf{u}$ de norme égale à $r = 0.75$ en utilisant la technique TR).

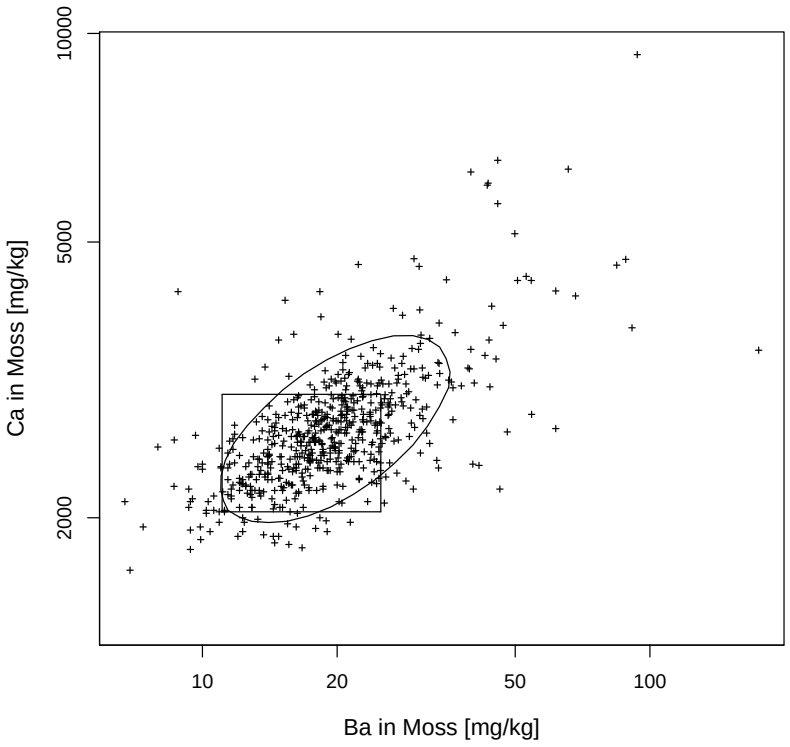


FIGURE 6 – *Comparaison entre les quantiles marginaux et les quantiles géométriques calculés sur les composants chimiques (Ba, Ca) mesurés sur des algues.*

Commentaires et conclusion Nous remarquons que les quantiles géométriques assurent une meilleure lecture le nuage des points que les quantiles marginaux car ils tiennent compte de la corrélation qui existe entre les deux variables. De plus des observations sont considérées par les quantiles marginaux comme “hors normes” (i.e. se situant à l’extérieur du rectangle) alors qu’elles ne le sont pas en se basant sur le contour de niveau 75% (car elles se situent à l’intérieur de la zone délimitée par le contour). De même des observations sont considérées “dans la norme” (i.e. se trouvant dans le rectangle) alors qu’elles apparaissent plutôt “hors normes” puisqu’elles sont en dehors du contour. Les quantiles géométriques donnent donc ici des résultats plus intéressants que ceux obtenus au moyen des quantiles marginaux.

Remerciements. Qu’il nous soit permis de remercier l’Éditeur en Chef du Journal de la SFdS-RSA, ainsi que les deux relecteurs anonymes : leurs commentaires, leurs critiques et leurs suggestions constructives nous ont permis d’améliorer substantiellement la qualité de cet article.

Références

[1] B. Abdous and R. Theodorescu. Note on the geometric quantile of a random vector. *Statistics and Probability Letters*, 13 :333–336, 1992.

[2] G. J. Babu and C. R. Rao. Joint asymptotic distribution of marginal quantile functions in samples from a multivariate population. *Journal of Multivariate Analysis*, 27 :15–23, 1988.

[3] R. R. Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37 :577–580, 1966.

[4] V. Barnett. The ordering of multivariate data. *Journal of the Royal Statistical Society, Ser. A*, 139 :318–354, 1976.

[5] F.K. Bedall and H. Zimmermann. Algorithm as 143, the mediancenter. *Applied Statistics*, 28 :325–328, 1979.

[6] B. M. Brown. Statistical use of the spatial median. *Journal of the Royal Statistical Society, Ser. B*, 45 :25–30, 1983.

[7] B. M. Brown and T. P. Hettmansperger. Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Ser. B*, 49 :301–310, 1987.

[8] B. M. Brown and T. P. Hettmansperger. An affine invariant bivariate version of the sign test. *Journal of the Royal Statistical Society, Ser. B*, 51 :117–125, 1989.

[9] B. Chakraborty. On affine equivariant multivariate quantiles. *The Institute of Statistical Mathematics*, 53 :380–403, 2001.

[10] B. Chakraborty and P. Chaudhuri. On a transformation and retransformation technique for constructing an affine equivariant multivariate median. *Proceeding of the American Mathematical Society*, 124 :2539–2547, 1996.

[11] P. Chaudhuri. Multivariate location estimation using extension of r-estimates through u-statistics type approach. *The Annals of Statistics*, 20 :897–916, 1992.

[12] P. Chaudhuri. On a geometric notation of quantiles for multivariate data. *Journal of the American Statistical Association*, 91 :862–872, 1996.

[13] P. Chaudhuri and D. Sengupta. Sign tests in multidimension : inference based on the geometry of the data cloud. *Journal of the American Statistical Association*, 88 :1363–1370, 1993.

[14] Y. Cheng and De Gooijer J. G. On the uth geometric conditional quantile. *Journal of Statistical Planning and Inference*, 137 :1914–1930, 2007.

[15] J. G. De Gooijer and A. Gannoun. Tr multivariate conditional median estimation. *Communications in Statistics - Simulation and Computation*, 36 :165–176, 2007.

[16] J. G. De Gooijer, A. Gannoun, and D. Zerom. Mean squared error properties of kernel-based multi-stage conditional median predictor for time series. *Statistics and Probablility Letters*, 56 :51–56, 2002.

[17] J. G. De Gooijer, A. Gannoun, and D. Zerom. A multivariate quantile predictor. *Communications in Statistics - Theory and Methods*, 35 :133–147, 2006.

[18] D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, 20 :1803–1827, 1992.

[19] W.F. Eddy. Convex hull peeling. *COMPSTAT 1982 for IASC, Vienna : Pysica-Verlag*, pages 42–47, 1982.

[20] W.F. Eddy. Ordering of multivariate data. *Computer Science and Statistics : The Interface. L. Billard, Amesterdam : North-Holland*, pages 25–30, 1985.

[21] T. Ferguson. *Mathematical Statistics : A Decision Theoric Approach*. Academic Press, New York, 1967.

[22] A. Gannoun, S. Girard, C. Guinot, and J. Saracco. Trois méthodes non paramétriques pour l'estimation de courbes de référence. application à l'analyse des propriétés biophysiques de la peau. *Revue de Statistique Appliquée*, 1 :65–89, 2002.

[23] A. Gannoun, J. Saracco, A. Yan, and G.E. Bonney. On adaptive transformation-retransformation estimate of conditional spatial median. *Communications in Statistics - Theory and Methods*, 32 :1981–2011, 2003.

[24] A. Gannoun, J. Saracco, and K. Yu. Nonparametric time series prediction by conditional median and quantiles. *Journal of Statistical Planning and Inference*, 117 :207–223, 2003.

[25] J.C. Gower. Algorithm as 78 : The mediancenter. *Applied Statistics*, 23 :466–470, 1974.

[26] J. B. S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35 :414–415, 1948.

[27] J. H. B. Kemperman. The median of a finite measure on a banach space. *Statistical Data Analysis based on the L_1-norm and related methods, Y. Dodge (ed.), North-Holland, Amsterdam*, pages 217–230, 1987.

[28] R. Koenker and G. Basset. Regression quantiles. *Econometrica*, 46 :33–50, 1978.

[29] P. Kokic, J. Breckling, and O. Lübke. A new definition of multivariate m-quantiles. *Statistical data analysis based on the L_1-norm and related methods, Birkhäuser Verlag, Basel*, pages 15–24, 2002.

[30] V. Koltchinskii. M-estimation, convexity and quantiles. *The Annals of Statistics*, 25 :435–477, 1997.

[31] R. Y. Liu, J. M. Parelius, and K. Singh. Multivariate analysis by data depth : descriptive statistics, graphics and inference (with discussion). *The Annals of Statistics*, 27 :783–858, 1999.

[32] H. Oja. Descriptive statistics for multivariate trimming. *Statistics and Probability Letters*, 1 :327–332, 1983.

[33] R. L. Plackett. Comment on the " ordering of multivariate data", by v. barnett. *Journal of the Royal Statistical Society, Ser. A*, 139 :344–346, 1976.

[34] R. D. Reiss. *Approximate distributions of order statistics with applications to nonparametric statistics*. New York : Springer, 1989.

[35] R. Serfling. Quantile functions for multivariate analysis : approaches and applications. *Statistica Neerlandica*, 56 :214–232, 2002.

[36] R. Serfling. Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference*, 123 :259–278, 2004.

[37] Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28 :461–482, 2000.

Annexe : preuve du Théorème 4.3

- La première partie du théorème se déduit directement de l'équation (11). Si les observations $\mathbf{Y}_i$ ne sont pas alignées, le quantile géométrique conditionnel est l'unique solution θ de l'équation (11). On en déduit que $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ satisfait l'équation suivante :

$$\sum_{i=1}^{n} S(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}) - \theta) K_{h_n}(\mathbf{x} - \mathbf{X}_i) = \mathbf{u} \sum_{i=1}^{n} K_{h_n}(\mathbf{x} - \mathbf{X}_i).$$

- Montrons maintenant la deuxième partie du théorème. La fonction $\phi(\mathbf{u}, \mathbf{y})$ est convexe sur $\mathbb{R}^d$. On en déduit que

$$\mathbf{Q}_n(\mathbf{u}|\mathbf{x}) = \arg\min_{\theta} \sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta) K_{h_n}(\mathbf{x} - \mathbf{X}_i)$$

si et seulement si, pour tout $\mathbf{h} \in \mathbb{R}^d$, on a

$$\lim_{t \to 0^+} \left[\sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) + t\mathbf{h}) K_{h_n}(\mathbf{x} - \mathbf{X}_i) - \sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u}|\mathbf{x})) K_{h_n}(\mathbf{x} - \mathbf{X}_i) \right] \geq 0.$$

Cependant, pour tout $\mathbf{y}, \mathbf{h} \in \mathbb{R}^d$ tel que $\mathbf{y} \neq 0$, on a :

$$\lim_{t \to 0^+} \frac{\phi(\mathbf{u}, \mathbf{y} + t\mathbf{h}) - \phi(\mathbf{u}, \mathbf{y})}{t} = \lim_{t \to 0^+} \frac{||\mathbf{y} + t\mathbf{h}|| - ||\mathbf{y}|| + < \mathbf{u}, t\mathbf{h} >}{t} = < \frac{\mathbf{y}}{||\mathbf{y}||} + \mathbf{u}, \mathbf{h} > .$$

De plus, pour tout $\mathbf{h} \in \mathbb{R}^d$ et $\mathbf{y} = 0$, on a

$$\lim_{t \to 0^+} \frac{\phi(\mathbf{u}, t\mathbf{h}) - \phi(\mathbf{u}, 0)}{t} = || \mathbf{h} || + < \mathbf{u}, \mathbf{h} > .$$

Ensuite, en utilisant les résultats précédents, on obtient :

$$\sum_{\substack{1 \leq i \leq n \\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) \neq \mathbf{Y}_i}} K_{h_n}(\mathbf{x} - \mathbf{X}_i) < S(\mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u}|\mathbf{x})) + \mathbf{u}, \mathbf{h} > + \sum_{\substack{1 \leq i \leq n \\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) = \mathbf{Y}_i}} K_{h_n}(\mathbf{x} - \mathbf{X}_i)(|| \mathbf{h} || + < \mathbf{u}, \mathbf{h} >) \geq 0.$$

Puisque cette inégalité est vraie pour tout $\mathbf{h} \in \mathbb{R}^d$, elle reste aussi vraie pour $-\mathbf{h}$. En remplaçant $\mathbf{h}$ par $-\mathbf{h}$ dans l'inégalité précédente, on obtient :

$$\sum_{\substack{1 \leq i \leq n \\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) = \mathbf{Y}_i}} K_{h_n}(\mathbf{x} - \mathbf{X}_i)(|| \mathbf{h} || - < \mathbf{u}, \mathbf{h} >) \quad \geq$$

$$\sum_{\substack{1\leq i\leq n\\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x})\neq \mathbf{Y}_i}} K_{h_n}(\mathbf{x}-\mathbf{X}_i) < S(\mathbf{Y}_i-\mathbf{Q}_n(\mathbf{u}|\mathbf{x}))+\mathbf{u},\mathbf{h}> . \tag{14}$$

D'autre part, en utilisant l'inégalité de Schwartz, on a :

$$\mid \ ||\,\mathbf{h}\,|| \pm <\mathbf{u},\mathbf{h}>| \leq ||\,\mathbf{h}\,|| + |<\mathbf{u},\mathbf{h}>| \leq (1+||\,\mathbf{u}\,||\,)||\,\mathbf{h}\,||.$$

Ainsi, l'inégalité (14) est équivalente à

$$\sum_{\substack{1\leq i\leq n\\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x})=\mathbf{Y}_i}} K_{h_n}(\mathbf{x}-\mathbf{X}_i)(1+||\,\mathbf{u}\,||)||\,\mathbf{h}\,|| \ \ \geq$$

$$\sum_{\substack{1\leq i\leq n\\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x})\neq \mathbf{Y}_i}} K_{h_n}(\mathbf{x}-\mathbf{X}_i) < S(\mathbf{Y}_i-\mathbf{Q}_n(\mathbf{u}|\mathbf{x}))+\mathbf{u},\mathbf{h}> . \tag{15}$$

Puisque cette inégalitée est vraie pour tout $\mathbf{h}\in\mathbb{R}^d$, donc on peut choisir en particulier

$$\mathbf{h} = S(\mathbf{Y}_i-\mathbf{Q}_n(\mathbf{u}|\mathbf{x}))+\mathbf{u}.$$

En remplaçant $\mathbf{h}$ par cette valeur dans l'équation (15), on a :

$$\sum_{\substack{1\leq i\leq n\\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x})=\mathbf{Y}_i}} K_{h_n}(\mathbf{x}-\mathbf{X}_i)(1+||\,\mathbf{u}\,||) \geq \left\| \sum_{\substack{1\leq i\leq n\\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x})\neq \mathbf{Y}_i}} K_{h_n}(\mathbf{x}-\mathbf{X}_i)[S(\mathbf{Y}_i-\mathbf{Q}_n(\mathbf{u}|\mathbf{x}))+\mathbf{u}] \right\| .$$

On en déduit ainsi l'inégalité du deuxième point du théorème.