Contribution à l'estimation non paramétrique des quantiles géométriques et à l'analyse des données fonctionnelles

Mohamed Chaouch

N° d'ordre:

THÈSE

présentée à

L'UNIVERSITÉ DE BOURGOGNE

en vue de l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ DE BOURGOGNE
Mention : Mathématiques Appliquées
Spécialité : Statistique

par

Mohamed CHAOUCH

Laboratoire : Institut de Mathématiques de Bourgogne (IMB)
Équipe d'accueil : Applications des Mathématiques
École Doctorale : Carnot

Titre de la thèse :

Contribution à l'estimation non paramétrique des quantiles géométriques et à l'analyse des données fonctionnelles

Soutenue le 05 Décembre 2008 devant la commission d'examen composée de :

M. :	Hervé Cardot	Prof. Université de Bourgogne	Président
M. :	F. Jay Breidt	Prof. Colorado State University	Rapporteur
Mme. :	Anne Ruiz-Gazen	Prof. Université Toulouse I	Rapporteur
M. :	Keming Yu	Prof. Brunel University	Rapporteur
M. :	Jérôme Saracco	Prof. Université Bordeaux IV	Directeur
M. :	Ali Gannoun	Prof. CNAM Paris	Co-Directeur
Mme. :	Camelia Goga	MCF. Université de Bourgogne	Co-Directeur

Remerciements

Je souhaite tout d'abord remercier vivement Jérôme Saracco pour avoir dirigé ma thèse. Je suis particulièrement reconnaissant de la confiance qu'il m'a su m'accorder durant cette thèse ainsi que de ses precieux conseils.

Je veux également remercier Ali Gannoun et Camelia Goga, qui ont co-dirigé cette thèse, pour leur encadrement et leur encouragement durant toute la période de la réalisation de ce travail.

Par ailleurs, je me considère comme particulièrement chanceux d'avoir pu travailler sous la direction de trois reponsables dont les qualités humaines n'ont cessé de m'encourager. Je les remercie tous les trois de l'attention et de la disponibilité dont ils ont su faire preuve vis à vis de mon travail.

Je remerci chalereusement les trois rapporteurs. Jay F. Breidt, Anne Ruiz-Gazen et Kemin Yu d'avoir accepté de juger mon travail.

Merci à Hervé Cardot pour m'avoir fait profité de son expérience et de ses compétences en analyse de données fonctionnelles ainsi que d'avoir bien voulu présider le jury.

Un grand merci à ma famille pour son soutien constant et chaleureux pendant toute la réalisation de ce travail.

Enfin, j'adresse mes remerciments à tous ceux qui ont participé de prés ou de loin à la bonne réalisation de la soutenence de cette thèse.

Mohamed Chaouch

Table des matières

1 Présentation générale **5**

1.1 Description de la thèse 5

1.2 Description par chapitre des principaux résultats 6

1.2.1 Quantiles géométriques conditionnels ou non 6

1.2.2 Quantiles géométriques et sondage 14

1.2.3 ACP fonctionnelle et sondage 18

1.2.4 Estimation non paramétrique des quantiles géométriques conditionnels sous l'hypothèse de mélange fort 23

1.3 Liste des travaux 29

2 Estimation des quantiles géométriques conditionnels et non conditionnels **31**

2.1 Introduction 32

2.2 Quantile univarié 33

2.2.1 Définition 33

2.2.2 Deux caractérisations du quantile univarié 33

2.2.2.1 Le quantile en tant que racine d'une équation 33

2.2.2.2 Le quantile en tant que solution d'un problème de minimisation 34

2.2.3 Estimation 35

2.3 Quantile géométrique 35

2.3.1 Définition 36

2.3.2 Estimation 36

2.3.3 Existence et unicité de $\mathbf{Q}_n(\mathbf{u})$ 37

2.3.4 Interprétation du vecteur $\mathbf{u}$ 37

2.3.5 Résultats asymptotiques 40

2.3.6 Technique de Transformation-Retransformation (TR) 41

2.3.6.1 Choix de α 41

2.3.7 Un algorithme d'estimation 42

2.4 Quantile géométrique conditionnel 43

2.4.1 Définition 43

2.4.2 Estimation 44

2.4.3 Propriétés asymptotiques de $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ 45

2.4.4 Un algorithme d'estimation du quantile géométrique conditionnel 46
2.5 Implémentation en R des algorithmes de calculs des quantiles (conditionnels) géométriques . . . 47
2.5.1 Cas du quantile géométrique $\mathbf{Q}_n(\mathbf{u})$ . . . 47
2.5.1.1 Description de la fonction "**QuantileNC.est**" . . . 47
2.5.1.2 Procédure de Transformation-Retransformation . . . 48
2.5.2 Cas du quantile géométrique conditionnel $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ . . . 48
2.5.2.1 Choix des paramètres de lissage . . . 49
2.5.2.2 Description des fonctions . . . 49
2.6 Etude par simulation . . . 50
2.6.1 Une première simulation : cas de quantiles géométriques . . . 50
2.6.2 Une seconde simulation : cas de quantiles géométriques conditionnels . . . 51
2.7 Étude sur des données réelles . . . 52

3 Design-Based Estimation for Geometric Quantiles 61
3.1 Introduction . . . 61
3.2 Geometric quantile in finite population setting . . . 63
3.3 Design-based estimator of $Q(u)$ . . . 64
3.3.1 Computation of $\widehat{Q}(u)$ . . . 66
3.4 Main results . . . 67
3.4.1 Asymptotic framework . . . 67
3.4.2 Asymptotic variance . . . 68
3.4.3 Variance estimation . . . 69
3.5 Simulation Study . . . 70
3.6 Conclusion and comments . . . 73
3.7 Appendix . . . 73

4 Functional Principal Components Analysis with Survey Data 79
4.1 Introduction and notations . . . 79
4.2 Survey framework and PCA . . . 81
4.2.1 FPCA in a finite population setting . . . 81
4.2.2 The Horvitz-Thompson Estimator . . . 82
4.2.3 Substitution Estimator for Nonlinear Parameters . . . 83
4.3 Asymptotic Properties . . . 84
4.3.1 ADU-ness and Convergence of Estimators . . . 85
4.3.2 Variance Approximation and Estimation . . . 86
4.4 A simulation study . . . 88

5 On the conditional geometric quantile from dependent observations and its estimation 101
5.1 Introduction . . . 101
5.2 Definitions and notation . . . 102
5.3 Main results . . . 104

5.4 Proofs 105
5.4.1 Preliminary Results 105

Annexes **115**

Table des figures **123**

Chapitre 1

Présentation générale

1.1 Description de la thèse

Le travail développé dans ce mémoire de thèse se situe à l'intersection entre trois thématiques importantes de la Statistique, à savoir l'estimation non paramétrique (avec des méthodes à noyau), la théorie des sondages et l'analyse des données fonctionnelles. Pour la partie "estimation non paramétrique", nous nous sommes intéressés plus particulièrement à l'estimation non paramétrique des quantiles géométriques (qui représentent un cas particulier des quantiles multivariés) conditionnels et non conditionnels. Nous nous sommes placés, dans un premier temps, dans le cas où on dispose de réalisations de variables aléatoires indépendantes et identiquement distribuées (i.i.d.). Ensuite nous nous sommes intéressés au cas d'observations vérifiant une hypothèse de mélange. En ce qui concerne la thématique "théorie des sondages", pour des raisons pratiques et par souci d'optimiser le temps de calcul des estimateurs (qui depend de la taille de l'échantillon), nous avons proposé et étudié des estimateurs des quantiles géométriques construits à partir d'un échantillon sélectionné suivant un plan de sondage. Dans la partie "analyse de données fonctionnelles", nous avons proposé une Analyse en Composantes Principales Fonctionnelle (ACPF) basée sur des estimateurs de type Horvitz-Thompson lorsque les données (les courbes observées) sont issues d'un plan de sondage.

Dans la suite de ce premier chapitre, nous allons décrire brièvement les différents résultats obtenus dans le cadre de cette thèse. Chacun des chapitres suivants est sous la forme d'un article. Un avantage de ce choix est que chaque chapitre peut être lu de manière indépendante, un petit défaut est que certaines notations ou certains arguments pourront paraître redondants lors du parcours de l'ensemble de ce manuscrit.
Le Chapitre 2 s'intéresse à l'estimation non paramétrique des quantiles géométriques conditionnels et non conditionnels. Dans ce but, on construit des estimateurs basés sur un critère de minimisation d'une fonction de perte. Nous donnons également les différents résultats asymptotiques ainsi que les algorithmes de calcul. Des simulations qui ont été faites montrent le bon comportement de nos estimateurs. Une application sur des données réelles montre l'intérêt pratique des quantiles géométriques. Ce chapitre a fait l'objet de l'article intitulé *"Estimation de quantiles géométriques conditionnels et*

non conditionnels". Cet article a été fait en collaboration avec Ali Gannoun et Jérôme Saracco et a été soumis au *Journal de la SFdS/RSA*.
Le troisième chapitre porte sur l'estimation des quantiles géométriques dans le cadre des sondages. Ce chapitre peut être vu comme une solution pratique à un problème qui se pose dans le chapitre précédent à savoir la lourdeur du temps de calcul lorsque la taille de l'échantillon est importante. Nous proposons différents résultats asymptotiques concernant la convergence des estimateurs proposés ainsi que des simulations qui montrent leur bon comportement. Ce chapitre a fait l'objet d'un article intitulé *"Design-Based Estimation for Geometric Quantiles"*, il a été fait en collaboration avec Camelia Goga et il a été soumis à l'*International Statistical Review*.
Le quatrième chapitre consiste à étudier l'ACP de données fonctionnelles issues d'un plan de sondage. Nous démontrons dans ce chapitre le bon comportement asymptotique des estimateurs, de type Horvitz-Thompson, des différents paramètres caractérisant l'ACPF ainsi que l'apport de la technique de linéarisation par la fonction d'influence dans l'estimation de la variance asymptotique des éléments propres de l'ACPF. Ce chapitre a fait l'objet d'un article intitulé *"Functional Principal Components Analysis with Survey Data"* qui a été écrit en collaboration avec Hervé Cardot, Camelia Goga et Catherine Labruère. Cet article a été soumis au *Journal of Statistical Planning and Inference*.
Le Chapitre 5 traite de l'estimation des quantiles géométriques conditionnels dans le cas de mélange fort. Nous définissions, dans un premier lieu, le quantile géométrique conditionnel ainsi que son estimateur non paramétrique. Ensuite nous montrons que, sous certaines hypothèses, cet estimateur est uniformement convergent sur tout ensemble compact. Ce chapitre a fait l'objet d'un article intitulé *"Estimation non paramétrique des quantiles géométriques conditionnels sous une hypothèse d'α-mélange"*. Il a été rédigé par Mohamed Chaouch et soumis aux *Comptes Rendus de l'Académie des Sciences de Paris (C.R.A.S.)*. Les programmes développés avec *R*, permettant de produire les graphiques présentés dans ce mémoire sont détaillés dans l'Annexe.

1.2 Description par chapitre des principaux résultats

Dans cette section nous donnons une brève présentation de chaque chapitre de ce mémoire ainsi que les différents résultats obtenus. Les perspectives de recherche, relatives à chaque chapitre, sont également proposées.

1.2.1 Quantiles géométriques conditionnels ou non

Nous parlerons dans cette section, dans un permier lieu, des quantiles géométriques, ensuite des quantiles géométriques conditionnels.

Quantiles géométriques

Le problème de définition des quantiles pour des variables aléatoires multidimensionnelles n'est pas recent. En effet, au début du vingtième siècle apparaissent les travaux

de Weber (1909) représentant une première tentative de généralisation de la médiane. Considérons le problème de localisation suivant : une entreprise cherche à placer un entrepôt pour servir ses n clients ayant les coordonnées $X_1, X_2, \ldots, X_n$ dans une zone géographique donnée, avec $X_i \in \mathbb{R}^2$ représente les coordonnées du client i. Pour faciliter ce problème, Weber considère que l'entreprise peut placer l'entrepôt dans n'importe quel point de la carte et que le coût de transport des marchandises est proportionnel à la distance Euclidienne entre l'entrepôt et le client.
La solution proposée par Weber à ce problème est de localiser l'entrepôt dans un endroit de telle sorte que le total des coûts de transport soit minimal. En d'autre termes, ceci revient à minimiser la somme des distances entre les clients et l'entrepôt. De cette façon, Weber définit la médiane "spatiale" (ou "multidimensionnelle") comme étant l'argument qui minimise $\sum_{i=1}^{n} \|X_i - \theta\|$.
Cette définition a été utilisée ensuite par Scates (1933) pour rechercher le centre géographique des Etats-Unis. Pour une bibliographie plus large sur les différentes définitions proposées de la médiane multidimensionnel, le lecteur peut se reporter à l'article de Small (1990). Néanmoins, l'absence d'un critère clair permettant d'ordonner les observations multivariées reste un obstacle pour la généralisation des quantiles au cas multidimensionnel.
L'article de Barnett (1976) propose quelques techniques permettant d'ordonner des observations multivariées. Cet article représente le vrai point de départ pour les chercheurs pour définir les quantiles bivariés. Nous citons à ce stade les travaux de Eddy (1982, 1985) et Brown et Hettmansperger (1987, 1989). Par la suite, Babu et Rao (1988), Abdous et Theodorescu (1992) et Kim (1992) ont généralisé la notion de quantile pour un vecteur aléatoire de dimension quelconque. Cependant, cette définition ne tient pas compte de la géométrie des points bien qu'elle soit invariante par rotation, elle n'est pas invariante par transformation affine.

Récemment, deux approches principales ont été développées pour définir des quantiles multivariés qui soient invariants par rotation et/ou 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 (1983) pour la médiane simpliciale et ceux de Donoho et Gasko (1992), Liu et *al.* (1999) et Zuo et Serfling (2000) pour les quantiles multivariés. L'avantage de cette approche est qu'elle est invariante par rotation et par transformation affine.
La seconde approche définit le quantile comme un M-estimateur qui minimise une fonction de perte (ou de coût). Les quantiles multivariés, en tant que M-estimateurs, sont invariant par rotation mais ne sont pas invariants par toutes les transformations affines. Cette approche a été utilisée, en premier lieu, par Haldane (1948), Gower (1974), Brown (1983) et Chaudhuri (1992) pour généraliser la notion de la médiane au cas multivarié. Ensuite, Chaudhuri (1996), Koltchinskii (1997) et Kokic et *al.* (2002) ont proposé différentes généralisations des quantiles multivariés. Pour une description plus détaillée des différentes méthodes ainsi qu'une comparaison entre elles, le lecteur peut se référer à l'article de Serfling (2002).

Dans ce mémoire, nous nous focalisons sur la définition des quantiles, dit *géomé-*

triques, introduite par Chaudhuri (1996). Cette définition est une généralisation d'un résultat connu sur les quantiles univariés (voir Ferguson (1967) p. 51). Soit Y une variable aléatoire réelle. Soient $p \in (0,1)$ et $u = 2p - 1$. Le quantile d'ordre p, noté $Q(p)$ peut être vu, si $\mathbb{E}|Y| < \infty$, comme étant la solution du problème d'optimisation suivant :

$$\arg\min_{\theta \in \mathbb{R}} \mathbb{E}\{|Y - \theta| + u(Y - \theta)\}. \tag{1.1}$$

Dans le cadre de l'estimation des quantiles de régression, Koenker et Basset (1978) ont introduit la fonction de "perte", $\phi(u,t) = |t| + ut$, avec $t \in \mathbb{R}$ et $u \in (-1,1)$. Cette fonction permet de réécrire la relation (1.1) de la façon suivante

$$\arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\{\phi(u, Y - \theta)\} \tag{1.2}$$

Cette défintion basée sur le critère de minimisation, donné par (1.2), permet facilement de généraliser la notion de quantile au cas d'une variable multidimensionnelle. Notons que ce n'est pas le cas de la définition classique où le quantile d'ordre p est défini comme étant la réciproque de la fonction de répartition en p.

En effet, considérons maintenant une variable aléatoire multidimensionnelle $Y \in \mathbb{R}^d$ avec $d \geq 2$, et ϕ une fonction, dite "de perte multivariée", définie sur $B^d \times R^d$ et à valeurs dans $\mathbb{R}$, telle que pour tout $t \in \mathbb{R}^d$ et $u \in B^d = \{u \in \mathbb{R}^d : \|u\| < 1\}$, $\phi(u,t) = \|t\| + \langle u, t\rangle$. Le quantile géométrique, indexé par le vecteur u, est la solution du problème de minimisation suivant

$$\arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\{\phi(u, Y - \theta)\}. \tag{1.3}$$

La fonction $\mathbb{E}\{\phi(u, Y-\theta)\}$ n'est définie que si $\mathbb{E}\|Y\| < \infty$. Utilisant un artitifice de Kemperman (1987), la fonction $\mathbb{E}\{\phi(u, Y - \theta) - \phi(u, Y)\}$ l'est toujours. Ces deux fonctions admettent le même minimum quand celui-ci existe. Ceci permet de définir le quantile géométrique comme suit :

$$Q(u) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\{\phi(u, Y - \theta) - \phi(u, Y)\}. \tag{1.4}$$

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

$$\mathbb{E}(S(\theta - Y)) = u. \tag{1.5}$$

Soit F_n l'estimateur empirique (non paramétrique) de la fonction de répartition F obtenu à partir des observations $Y_1, \ldots Y_n$ de Y. Pour tout $u \in B^d$, on peut définir un estimateur $Q_n(u)$ de $Q(u)$ par :

$$\begin{aligned} Q_n(u) &= \arg\min_{\theta \in \mathbb{R}^d} \int (\phi(u, y - \theta) - \phi(u, y)) F_n(dy) \\ &= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (\phi(u, Y_i - \theta) - \phi(u, Y_i)) \end{aligned} \tag{1.6}$$

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

$$Q_n(u) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \phi(u, Y_i - \theta). \tag{1.7}$$

La démonstration de l'existence et de l'unicité de l'estimateur $Q_n(u)$ est donnée à la Section 3.3 du Chapitre 2. On peut également montrer, à partir de l'équation (1.7), que $Q_n(u)$ est aussi solution de l'équation suivante dont l'inconnue est θ :

$$\frac{1}{n} \sum_{i=1}^{n} S(\theta - Y_i) = u. \tag{1.8}$$

Le vecteur u indexant le quantile géométrique permet de donner des informations sur le quantile géométrique et sur son estimateur. En effet, u étant un vecteur de la boule unitaire B^d, sa norme nous renseigne sur l'"ordre" du quantile : si sa norme est proche de 1 (resp. 0), alors $Q(u)$ est un quantile "extrême" (resp. "central" i.e. proche de la médiane géométrique). La direction de u nous indique la position du quantile par rapport à la médiane. Une étude par simulation ainsi que des exemples, permettant de comprendre l'interprétation du vecteur u, sont détaillés dans la Section 3.4 du Chapitre 2.
Chaudhuri (1996) a démontré, à l'aide d'une représentation linéaire de type Bahadur (1966), que $Q_n(u)$ est un estimateur consistant de $Q(u)$. Ensuite il a déduit, à l'aide de cette représentation, la loi asymptotique de l'estimateur du quantile géométrique. Ces résultats sont détaillés dans la Section 3.5 du Chapitre 2.

Les quantiles géométriques sont invariants par rotation, cependant ils ne le sont pas par transformation affine. La technique dite de *Transformation-Retransformation* (TR) (voir Chaudhuri et Sengupta, 1993) permet d'avoir des quantiles géométriques invariants par rotation et tranformation affine (voir par exemple Chakraborty (2001) et Gannoun et *al.* (2003)). Cette technique ainsi que les différents résultats asymptotiques correspondant sont décrits dans le Chapitre 2, Section 3.6.

Différents auteurs, tels que Bedall et Zimmermann (1979), se sont intéressés au problème d'estimation de la médiane géométrique (ou "spatiale") qui minimise $\sum_{i=1}^{n} \|Y_i - \theta\|$. Chaudhuri (1996) 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 indexé par une direction u fixée. L'avantage de cet algorithme est qu'il converge au bout d'une dizaine d'itérations. En revanche le temps de calcul dépend de la taille de l'échantillon traité. Les différentes étapes de cet algorithme sont données dans la Section 3.7 du Chapitre 2 et son implémentation sous le logiciel R est décrite dans la Section 5 du Chapitre 2.

Quantiles géométriques conditionnels

Dans le cadre d'études industrielles ou biomédicales par exemple, une variable d'intérêt 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 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 Y sachant X. Ceci est l'objet de la Section 2.4 où nous proposons une généralisation, dans le cas conditionnel, de la définition du quantile géométrique ainsi que de son estimateur non paramétrique.

Considérons n observations $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ d'un couple de vecteurs aléatoires (X, Y) à valeurs dans $\mathbb{R}^s \times \mathbb{R}^d$, avec $d > 1$ et $s \geq 1$. Il est d'usage de rechercher la relation qui peut exister entre le vecteur à expliquer Y et la covariable multidimensionnelle X. Les quantiles géométriques conditionnels représentent un moyen pour aborder ce problème. Soit $u \in B^d$, le quantile géométrique conditionnel de Y sachant $X = x$, indexé par u, est défini par :

$$\begin{aligned} Q(u|x) &= \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\Phi(u, Y - \theta) - \Phi(u, Y) \mid X = x\right] \\ &= \arg\min_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} \{\phi(u, y - \theta) - \phi(u, y)\} F(dy|x). \end{aligned} \tag{1.9}$$

où F est la distribution conditionnelle de Y sachant X.
De manière similaire au cas non conditionnel, ce quantile peut être vu comme l'unique solution de l'équation dont l'inconnue est θ :

$$\mathbb{E}\left(S\left(\theta - Y\right) \mid X = x\right) = u. \tag{1.10}$$

Par la suite, on estime la fonction de répartition conditionnelle $F(.|x)$ par un estimateur non paramétrique, noté $F_n(.|x)$. Nous choisissons, par exemple, un estimateur de type Nadaraya-Watson défini pour tout $y \in \mathbb{R}^d$, par

$$F_n(y|x) = \sum_{i=1}^{n} w_{n,i} \mathbb{1}_{\{Y_i \leq y\}},$$

où $\mathbb{1}_{\{Y_i \leq y\}} = \mathbb{1}_{\{Y_i^1 \leq y^1\}} \times \cdots \times \mathbb{1}_{\{Y_i^d \leq y^d\}}$ et $w_{n,i} = K\left(\left(x - X_i\right)/h_n\right) / \sum_{i=1}^{n} K\left(\left(x - X_i\right)/h_n\right)$ représentant le poids associé à 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 noyaux unidimensionnels 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 X est multidimensionnelle, on peut choisir une largeur de fenêtre $h_{n,j}$ spécifique à chaque composante X_j de X ; cependant très souvent on choisit une même fenêtre h_n commune à l'ensemble des composantes. Nous nous sommes placés dans ce cadre afin de simplifier l'écriture de l'estimateur $F_n(y|x)$. Les poids $w_{n,i}$ sont autant plus importants pour les Y_i tels que X_i est proche de x.

On en déduit par la suite un estimateur $Q_n(u|x)$ de $Q(u|x)$ de la forme :

$$\begin{aligned} Q_n(u|x) &= \arg\min_{\theta\in\mathbb{R}^d} \int_{\mathbb{R}^d} \{\phi(u, y-\theta) - \phi(u,y)\} F_n(dy|x) \\ &= \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{n} w_{n,i} \{\phi(u, Y_i-\theta) - \phi(u,Y_i)\} \end{aligned} \tag{1.11}$$

A partir de l'équation (1.11), l'estimateur $Q_n(u|x)$ peut être aussi vu comme la solution de l'équation suivante dont l'inconnue est θ :

$$\sum_{i=1}^{n} w_{n,i} S(\theta - Y_i) = u \tag{1.12}$$

Sous certaines hypothèses détaillées dans la Section 4.3 du Chapitre 2, Cheng et De Gooijer (2007) ont établi une relation de type Bahadur permettant de déduire la loi asymptotique de l'estimateur du quantile géométrique conditionnel.
Le théorème suivant représente une généralisation du théorème 2.1.2 de Chaudhuri (1996) dans le cadre conditionnel.

Théorème 1.2.1 *Soit $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un n-échantillon de couples de vecteurs aléatoires à valeurs dans $\mathbb{R}^s \times \mathbb{R}^d$, avec $n \geq d+s$. Soit $Q_n(u|x)$ l'estimateur de $Q(u|x)$.*

- *Si pour tout $1 \leq i \leq n$, $Q_n(u|x) \neq Y_i$, alors on a :*

$$\sum_{i=1}^{n} S(Y_i - Q_n(u|x)) K_{h_n}(x - X_i) + u \sum_{i=1}^{n} K_{h_n}(x - X_i) = 0 \tag{1.13}$$

- *Si pour un certain i, on a $Q_n(u|x) = Y_i$, alors*

$$\left\| \sum_{\substack{1\leq i\leq n \\ Q_n(u|x)\neq Y_i}} [S(Y_i - Q_n(u|x)) + u] K_{h_n}(x - X_i) \right\| \leq \sum_{\substack{1\leq i\leq n \\ Q_n(u|x)= Y_i}} K_{h_n}(x - X_i)(1 + ||u||) \tag{1.14}$$

A partir de ce théorème nous avons déduit un algorithme de calcul permettant l'estimation des quantiles géométriques conditionnels (voir Section 2.4.4). Une implémentation de cet algorithme sous le logiciel R a été également faite et décrite à la Section 2.5. A la Section 2.6, différentes simulations, dans le cas conditionnel et non conditionnel, ont montré l'importance de l'étape de Transformation-Retransformation quand la distribution s'éloigne du cadre sphérique (en d'autres termes quand les variables qui forment le vecteur aléatoire sont corrélées).
Nous avons proposé dans la Section 2.7 du Chapitre 2 un exemple d'application sur des données issus d'un projet dont l'objectif est de déterminer le taux de pollution autour d'une zone industrielle. Cette application montre que les quantiles géométriques

assurent une meilleure lecture de nuage des points que les quantiles marginaux car ils tiennent compte de la corrélation qui existe entre les variables d'interêt.

Perspectives. L'étape de Transformation-Retransformation a été appliquée pour construire un estimateur invariant par rotation et par transformation affine du quantile géométrique conditionnel. Les différentes simulations faites dans ce chapitre ont montré l'intérêt d'appliquer cette technique dans le cas d'une distribution non sphérique (elliptique par exemple). Il serait intéressant de généraliser les résultats théoriques trouvés par Chakraborty (2001) sur l'estimation des quantiles géométriques avec Transformation-Retransformation, au cas conditionnel.
Récemment Vardi et Zhang (2000) ont proposé un algorithme de calcul de la médiane multivariée défini à la fois à l'aide de la fonction de profondeur et du critére de minimisation de la fonction de perte. Il serait intéressant de généraliser cet algorithme pour estimer le quantile géométrique et de le comparer (au niveau du temps de calcul et de la convergence) à l'algorithme de Chaudhuri (1996).

Généralités sur la théorie des sondages

Les résultats du Chapitre 3, respectivement 4, sont obtenus lorsque les observations sont sélectionnées selon un plan de sondage. La condition i.i.d. supposée dans le cadre de la statistique inférentielle classique n'est plus satisfaite. Nous commonçons par introduire le cadre générale de la théorie des sondages qui sera utilisé ensuite pour trouver les résultats des Chapitre 3 et 4. Le lecteur peut être référé à Särndal et *al.* (1992) et Tillé (2001) pour une étude détaillée sur la théorie des sondages.
Lorsque la taille d'une population est trés élevée ou inconnue, on a souvent recours à un plan de sondage pour évaluer une caractéristique précise de cette population. Le sondage consiste à mesurer le caractère sur une partie de la population (appelée échantillon) et ensuite étendre les tendances observées sur l'échantillon à la population entière.

Soit une population U finie composée de N individus. Chaque individu est représenté par un numéro unique allant de 1 à N, soit $U = \{1, 2, \ldots, N\}$. On souhaite évaluer une fonction d'une caractéristique, appelée aussi variable d'intérêt, Y, sur la population U. On note par Y_k la valeur déterministe du caractère Y mesurée sur l'individu k. En théorie des sondages, l'intérêt porte le plus souvent sur le total de la caractéristique Y noté

$$t_Y = \sum_{k=1}^{N} Y_k.$$

Pour différentes raisons (coûts, temps, ...) on ne peut pas mesurer la caractéristique Y sur tous les individus et par conséquent le total t_Y est inconnu. On sélectionne alors un sous ensemble s de U, constitué de n individus de la population ($n < N$). Ce sous-ensemble est appelé échantillon et il est obtenu par un procédé probabiliste $p(\cdot)$ appelé plan de sondage. En effet, p est une loi de probabilité sur l'ensemble $\mathcal{S}$, des échantillons qu'on peut tirer de U, telle que $p(s) \geq 0$ pour tout $s \in \mathcal{S}$ et $\sum_{s\in\mathcal{S}} p(s) = 1$. On désigne par $Y_1, \ldots, Y_n$ les valeurs de la variable d'intérêt Y observées sur l'échantillon. Ces valeurs sont connues et considérées fixes. En revanche, dans un sondage, le fait qu'un

individu k appartienne ou non à l'échantillon est aléatoire et se traduit par la variable aléatoire $I_k = \mathbb{1}_{\{k \in s\}}$. Notons par $\pi_k = \Pr(k \in s)$, pour tout $k \in U$ la probabilité d'inclusion d'ordre 1 et par $\pi_{kl} = \Pr(k \,\&\, l \in s)$ pour tout $k, l \in U$ la probabilité d'inclusion d'ordre 2. On suppose que $\pi_k > 0$ et $\pi_{kl} > 0$, pour tout $k, l \in U$.
A partir de l'échantillon s, nous pouvons maintenant donner l'estimateur de type Horvitz-Thompson (1952) du total t_Y, soit

$$\widehat{t}_Y = \sum_{k=1}^{n} \frac{Y_k}{\pi_k} = \sum_{k \in U} \frac{Y_k}{\pi_k} I_k. \tag{1.15}$$

Nous voyons clairement que la valeur de $\widehat{t}_Y$ va dépendre des individus présents dans l'échantillon s. C'est en ce sens que nous affirmons que l'estimateur $\widehat{t}_Y$ est une variable aléatoire (il peut prendre différentes valeurs suivant l'échantillon choisi).
La qualité du sondage est mesurée par la qualité de l'estimateur. Si l'estimateur $\widehat{t}_Y$ est sans biais (i.e $\mathbb{E}_p(\widehat{t}_Y) = t_Y$), autrement dit $\widehat{t}_Y$ "tombe" en moyenne sur sa cible t_Y, et si $\widehat{t}_Y$ est de variance minimale, c-à-d l'ensemble des valeurs possibles de $\widehat{t}_Y$ se répartit autour de la cible t_Y, alors l'échantillon s représente bien la population U.
La variance de $\widehat{t}_Y$ calculée par rapport au plan de sondage $p(s)$ est la variance de type Horvitz-Thompson (HT) donnée par la formule suivante :

$$V_p(\widehat{t}_Y) = \sum_{k \in U} \sum_{\ell \in U} \Delta_{k\ell} \frac{Y_k}{\pi_k} \frac{Y_\ell}{\pi_\ell} \tag{1.16}$$

où $\Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell$. Comme nous pouvons le remarquer, la variance reste toujours inconnue vu qu'elle dépend de la population U alors qu'on ne dispose que d'un échantillon s de cette population. L'estimateur de type Horvitz-Thompson de cette variance est donné par l'expression suivante :

$$\widehat{V}_p(\widehat{t}_Y) = \sum_{k \in s} \sum_{\ell \in s} \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{Y_k}{\pi_k} \frac{Y_\ell}{\pi_\ell}. \tag{1.17}$$

L'estimateur de type HT présenté précedemment reste valable lorsqu'on cherche à estimer des fonctions linéaires d'un total. Nous désirons dans la suite étudier le cas où on cherche à estimer des fonctions non linéaires des totaux de la forme $f = f(t_Y, \dots)$ avec f une fonction non linéaire. Pour estimer f, on remplace chaque total par son estimateur de type HT et on obtient l'estimateur par substitution $\widehat{f} = f(\widehat{t}_Y, \dots)$.
Par exemple, lorsque la taille N de la population U est inconnue, la moyenne $\mu = \frac{\sum_{k \in U} Y_k}{N}$ est une fonction non linéaire des totaux.
L'estimation par substitution de la moyenne consiste à remplacer, dans la formule de μ, le total t_Y (resp. la taille de la population N) par son estimateur $\widehat{t}_Y$ donné par la formule (1.15) (resp. $\widehat{N} = \sum_{k \in s} = \frac{1}{\pi_k}$). On obtient alors

$$\widehat{\mu} = \frac{1}{\widehat{N}} \sum_{k \in s} \frac{Y_k}{\pi_k}.$$

La non linéarité de f pose un problème pour le calcul de la variance de l'estimateur $\widehat{f}$. Il existe deux approches pour donner une variance approximative de $\widehat{f}$: l'approche par ré-échantillonnage et l'approche par linéarisation. La première approche consiste à utiliser une des techniques de ré-échantillonnage telles que le jacknife (Rao et *al.*, 1992 et Berger et Skinner, 2005) ou le bootstrap (Gross, 1980, Chauvet, 2007) pour donner une approximation numérique de la variance. Quant à l'approche par linéarisation, nous nous intéressons dans cette thèse à la technique de linéarisation par les équations estimantes (Binder, 1983 et Kovačević et Binder, 1997) (voir plus de détails dans le Chapitre 3) et à la technique de linéarisation par la fonction d'influence (Deville, 1999) (voir les détails de cette méthode dans le Chapitre 4). Les deux méthodes de linéarisation consistent à écrire :

$$\widehat{f} - f = \sum_{k \in s} \frac{u_k}{\pi_k} - \sum_{k \in U} u_k + reste,$$

où u_k est la variable linéarisée de f. Le but est alors de donner le cadre asymptotique dans lequel le reste converge vers zéro ce qui permet d'approximer la variance de $\widehat{f}$ par la variance de $\sum_{k \in s} \frac{u_k}{\pi_k}$. Le cadre asymptotique exige en particulier que la taille de la population et de l'échantillon tendent vers l'infini. Isaki et Fuller (1982) proposent un cadre théorique, que nous décrivons par la suite, qui permet d'avoir une suite de populations croissantes tel que n et N tendent vers l'infini.
Soit $\{r_j\}$ une suite d'éléments. On considère une suite de populations finies $\{U_t\}_{t \geq 1}$ de tailles $\{N_t\}_{t \geq 1}$ tel que $0 < N_1 < N_2 < N_3 < \dots$ créées à partir de la suite $\{r_j\}$. La population U_1 est composée des N_1 premiers termes de $\{r_j\}$, $U_2 \supset U_1$ contient les N_2 premiers individus de $\{r_j\}$ tel que $N_1 < N_2$, et ainsi de suite. Dans chaque population U_t, on sélectionne un échantillon s_t selon un plan de sondage $p(s_t)$ avec des probabiltés d'inclusion π_{kt} et $\pi_{k\ell t}$. Les tailles des échantillons s_t satisfont $n_1 < n_2 < \dots < \dots$ et s_t n'est pas nécessairement inclus dans s_{t+1}. Nous avons ainsi $\lim_{t \to \infty} N_t = \infty$ et que $\lim_{t \to \infty} n_t = \infty$. Pour alléger les notations l'indice t n'est plus utilisé dans les Chapitre 3 et 4.
Les Chapitre 3 et 4 donnent les hypothèses nécessaires pour montrer les propriétés asymptotiques des quantiles géométriques et respectivement les éléments propres d'une ACP fonctionnelle.

1.2.2 Quantiles géométriques et sondage

Etant donné une population U, de taille N, nous souhaitons calculer le quantile géométrique d'ordre u. Dans le Chapitre 2, nous avons parlé du problème de lourdeur des calculs quand N est assez grand. Le Chapitre 3 présente une solution pratique à ce problème. En effet, nous montrons dans la suite que l'estimation du quantile géométrique à l'aide d'un échantillon, noté s de taille $n << N$ sélectionné dans la population U selon un plan de sondage $p(s)$, assure à la fois la résolution du problème de calcul et la convergence des estimateurs.
En se placant dans le cadre d'une population U finie, nous pouvons remarquer que le quantile géométrique, du nuage des points $Y_1, \dots, Y_n$ dans $\mathbb{R}^d$, minimise le total d'une

fonction de perte calculée sur toute la population U,

$$Q_N(u) = \arg\min_{\theta\in\mathbb{R}^d} \sum_{k=1}^{N} \phi(u, Y_k - \theta) \quad \text{avec} \quad u \in B^d = \{z \in \mathbb{R}^d : ||z|| < 1\}. \tag{1.18}$$

Nous pouvons déduire à partir de la relation (1.18) que $Q_N(u)$ est l'unique solution de l'équation suivante dont l'inconnue est θ.

$$\sum_{k=1}^{N} [S(Y_k - \theta) + u] = 0 \tag{1.19}$$

Notons par $h_k(\theta) = S(Y_k - \theta) + u$, pour tout $k \in U$, et par $H_U(\theta) = \sum_{i=1}^{N} h_k(\theta)$ le total de $h_k(\theta)$ calculé sur toute la population U. On peut vérifier que $H_U(Q_N(u)) = 0$. Considérons maintenant un échantillon $s \subset U$, sélectionné suivant un plan de sondage $p(s)$. L'estimateur du quantile géométrique $Q_N(u)$ peut s'écrire de la façon suivante

$$\widehat{Q}(u) = \arg\min_{\theta\in\mathbb{R}^d} \sum_{k\in s} \frac{\phi(u, Y_k - \theta)}{\pi_k}. \tag{1.20}$$

Les deux hypothèses suivante garantissent l'unicité du quantile géométrique ainsi que sont estimateur (voir Kemperman, 1987 et les arguments prouvant l'existence et l'unicité de l'estimateur dans la Section 3.2 du Chapitre 3).

($\star$) $\{Y_1, \ldots, Y_N\}$ ne sont pas alignés dans $\mathbb{R}^d$.

($\star\star$) Supposons que les $Y_k \in \mathbb{R}^d$, pour tout $k \in s$, ne sont pas alignés dans $\mathbb{R}^d$.

Dans la suite nous donnons les propriétés asymptotiques de cet estimateur. Pour cela notons d'abord par $\widehat{H}(\theta) = \sum_{k\in s} \frac{h_k(\theta)}{\pi_k} = \sum_{k\in U} \frac{h_k(\theta)}{\pi_k} I_k$, l'estimateur de type Horvitz-Thompson de $H_U(\theta)$.

Nous pouvons facilement montrer que $\widehat{H}(\theta)$ est un estimateur non biaisé de $H_U(\theta)$, i.e. $\mathbb{E}_p(\widehat{H}(\theta)) = H_U(\theta)$, où $\mathbb{E}_p(\cdot)$ est l'espérance par rapport au plan de sondage. Nous pouvons également écrire la variance de $\widehat{H}(\theta)$ (en utilisant la formule de variance donnée par la relation (1.16)) comme suit

$$\mathbb{V}_p\left(\widehat{H}(\theta)\right) = \sum_U \sum_U \Delta_{k\ell} \frac{h_k(\theta)}{\pi_k} \frac{h_l^T(\theta)}{\pi_\ell}. \tag{1.21}$$

Un estimateur non biaisé de $\mathbb{V}_p\left(\widehat{H}(\theta)\right)$ peut être déduit à partir de la relation (1.17),

$$\widehat{\mathbb{V}}_p\left(\widehat{H}(\theta)\right) = \sum_s \sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}} \frac{h_k(\theta)}{\pi_k} \frac{h_l^T(\theta)}{\pi_\ell}. \tag{1.22}$$

Dans le Chapitre 3, nous démontrons les résultats suivants :

Théorème 1. *Soit $u \in B^d$ et $\widehat{Q}(u)$ l'estimateur de $Q_N(u)$ calculé à partir de l'échantillon s.*

- *Si* $\widehat{Q}(u) \neq Y_k$ *pour tout* $k \in s$, *alors*

$$\widehat{H}(\widehat{Q}(u)) = \sum_{k \in s} \frac{h_k(\widehat{Q}(u))}{\pi_k} = 0, \tag{1.23}$$

- *Si* $\widehat{Q}(u) = Y_k$ *pour un certain* $k \in s$, *alors* $\| \sum_{\substack{k \in s \\ Y_k \neq \widehat{Q}(u)}} \frac{h_k(\widehat{Q}(u))}{\pi_k} \| \leq (1+\|u\|) \sum_{\substack{k \in s \\ Y_k = \widehat{Q}(u)}} \frac{1}{\pi_k}$

La relation (1.23) signifie que $\widehat{Q}(u)$ est la solution unique de l'équation estimante suivante $\sum_{k \in s} \frac{h_k(\theta)}{\pi_k} = 0$ (voir Binder, 1983).

Il est clair que $\widehat{Q}(u)$ est un estimateur non linéaire. Pour cette raison nous utilisons la technique de linéarisation pour obtenir une approximation de la variance. Plaçons nous maintenant dans le cadre asymptotique du sondage qui a été décrit dans la section précédente et introduisons également les hypothèses suivantes, présentées dans la Section 3.4.2 du Chapitre 3, qui serviront pour démontrer les différents résultats asymptotiques.

(A1) $\lim_{N \to \infty} \frac{n}{N} = \pi \in (0,1)$,

(A2) $\min_k \pi_k \geq \lambda_1$, $\min_{k \neq l} \pi_{kl} \geq \lambda_2$ avec λ_1, λ_2 deux constantes positives et $\overline{\lim}_{N \to \infty} n \max_{k \neq l} |\pi_{kl} - \pi_k \pi_l| < \infty$.

(A3) On suppose qu'il existe une constante positive M tel que $||Y_k - \theta|| \geq M$ pour tout $k \in U$ et $\theta \in \mathcal{V}_{Q_N(u)}$.

(A4) $\widehat{Q}(u)$ est un estimateur convergent de $Q_N(u)$, c-à-d pour tout $\varepsilon > 0$ fixé, nous avons $\lim_{N \to \infty} \mathbb{P}\left(||\widehat{Q}(u) - Q_N(u)|| > \varepsilon \right) = 0$.

(A5) $\frac{\sqrt{n}}{N} \left[\widehat{H}(Q_N(u)) - H_N(Q_N(u)) \right] \longrightarrow N(0, \Sigma)$ avec Σ une matrice définie positive.

Lemme 1. *Sous les hypothèses (A1) et (A2), l'estimateur de type HT,* $\widehat{H}(\theta)$, *de* $H_U(\theta)$, *satisfait* $\mathbb{E}_p \left|\left| \frac{1}{N} \left(\widehat{H}(\theta) - H_N(\theta) \right) \right|\right| = O(n^{-1/2})$ *pour tout* $\theta \in \mathcal{V}_{Q_N(u)}$, *où* $\mathcal{V}_{Q_N(u)}$ *est un voisinage de* $Q_N(u)$.

Notons par $J_U(\theta)$ la matrice Jacobienne de $H_U(\theta)$ définie par

$$J_U(\theta) = \sum_U \frac{1}{||Y_k - \theta||} \left[I_d - S(Y_k - \theta) S^T (Y_k - \theta) \right],$$

avec I_d la matrice identité de dimension d. L'estimateur de type HT de $J_U(\theta)$ est

$$\widehat{J}(\theta) = \sum_s \frac{1}{\pi_k ||Y_k - \theta||} \left[I_d - S(Y_k - \theta) S^T (Y_k - \theta) \right].$$

Les deux matrices $J_U(\theta)$ et $\widehat{J}(\theta)$ sont de dimension $d \times d$, symmétriques et définies positives sous les hypothèses $(\star)$ et $(\star\star)$.

Lemme 2. Supposons que les conditions (A1)-(A3) sont vérifiées. *Pour tout* $\theta \in \mathcal{V}_{Q_N(u)}$, *nous avons*

(i) $\frac{1}{N} J_U(\theta) = O(1)$,

(ii) $\mathbb{E}_p \left|\left| \frac{1}{N} \left(\widehat{J}(\theta) - J_U(\theta) \right) \right|\right|_1 = O(n^{-1/2})$ *où* $||\cdot||_1$ *est la norme trace telle que* $||A||_1^2 = tr(A^T A)$ *pour toute matrice* A.

Théorème 2. *Lorsque les hypothèses (A1)-(A5) sont vérifiées, l'estimateur* $\widehat{Q}(u)$ *de* $Q_N(u)$ *basé sur le plan de sondage* $p(s)$ *satisfait la relation suivante :*

$$\begin{aligned} \widehat{Q}(u) - Q_N(u) &= -J_U^{-1}(Q_N(u)) \left(\widehat{H}(Q_N(u)) - H_U(Q_N(u)) \right) + o_p(n^{-1/2}) \\ &= \sum_s \frac{u_k}{\pi_k} + o_p(n^{-1/2}) \end{aligned}$$

où $u_k = -J_U^{-1}(Q_N(u)) h_k(Q_N(u))$ *est la variable linéarisée de* $Q_N(u)$ *avec* $\sum_U u_k = 0$. *Par conséquent la variance asymptotique de* $\widehat{Q}(u)$ *notée* $\mathbb{AV}_p(\widehat{Q}(u))$, *est égale à la variance de l'estimateur de type HT* $\sum_s \frac{u_k}{\pi_k}$ *:*

$$\mathbb{AV}_p(\widehat{Q}(u)) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l^T}{\pi_l}.$$

La variance asymptotique de $\widehat{Q}(u)$ est calculée sur la population entière U, alors qu'on ne dispose que d'un échantillon s de cette population. Nous proposons maintenant l'estimer par l'estimateur de la variance donné par l'expression (1.17) où u_k est remplacé par son estimateur $\widehat{u}_k$. Nous obtenons alors

$$\widehat{\mathbb{V}}_p(\widehat{Q}(u)) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\widehat{u}_k}{\pi_k} \frac{\widehat{u}_l^T}{\pi_l} = \left[\widehat{J}(\widehat{Q}(u)) \right]^{-1} \widehat{V}_p(\hat{H}(\widehat{Q}(u))) \left[\widehat{J}(\widehat{Q}(u))^T \right]^{-1}$$

avec $\widehat{u}_k = -\widehat{J}^{-1}(\widehat{Q}(u)) h_k(\widehat{Q}(u))$. Le résultat suivant montre, sous certaines hypothèses présentées ci dessous, que $\widehat{\mathbb{V}}_p(\widehat{Q}(u))$ converge en probabilité vers $\mathbb{AV}_p(\widehat{Q}(u))$.

Théorème 3. *Supposons que* $(A1)$-$(A5)$ *sont vérifiées et que de plus* $\frac{1}{N^2} \left[\widehat{V}_p(\widehat{H}(Q_N)) - \Sigma \right] = o_p(n^{-1})$ *alors,*

$$\widehat{V}_p(\widehat{Q}(u)) - \mathbb{AV}_p(\widehat{Q}(u)) = o_p(n^{-1}).$$

Implémentation informatique et simulations

Dans le but de faciliter l'interprétation des graphiques, nous nous sommes placés dans la partie simulation (présentée dans la Section 3.5 du Chapitre 3) dans le cas bidimensionnel c-à-d $Y \in \mathbb{R}^2$. Nous avons supposé que Y suit une distribution binormale $\mathcal{N}_2((0,0) I_2)$. Nous avons simulé une population de taille 5000 selon cette loi. Nous avons construit deux strates U_1 et U_2 avec des variances différentes. La première strate

est de taille 1500 et la deuxième est de taille 3500. Aprés avoir adapé l'algorithme de Chaudhuri (1996) au cadre des sondages (voir Section 3.1 du Chapitre 3), nous l'avons implémenté avec le logiciel *R*. Différentes simulations ont montré le bon comportement de l'estimateur du quantile géométrique qu'on a proposé. De plus, on montre que le plan stratifié donne de meilleurs estimations que le Sondage Aléatoire Simple (SAS) notamment pour des échantillons de petites tailles.

Perspectives. Le fait que les quantiles géométriques ne sont pas invariants par transformation affine pose dans certains cas pratiques un problème, par exemple lorsque les variables ne sont pas mesurées sur une même échelle. Dans ce cas on est loin du cadre sphérique et l'application de la technique de Transformation-Retransformation dans le cadre de sondage peut être une solution à ce problème.
Nous pouvons également envisager l'étude de l'estimation des quantiles géométriques lorsque l'on dispose d'une information auxiliaire. Cette idée peut avoir deux aspects : le premier est de tenir compte de l'information auxiliaire pour améliorer la qualité des estimateurs proposés ci-dessus, le second consiste à définir les quantiles géométriques conditionnels dans le cadre des sondages.

1.2.3 ACP fonctionnelle et sondage

L'analyse statistique de courbes, ou analyse de données fonctionnelles, est un thème de recherche en Statistique qui est en pleine expansion et dont les applications concernent de nombreux domaines scientifiques (climatologie, médecine, économie, chimie quantitative, ...). On pourra se reporter à Ramsay et Silverman (2005) pour une revue de différentes méthodes d'analyse illustrées sur des exemple variés. Les outils statistiques mis en œuvre sont issus de l'analyse fonctionnelle et généralisent les procédures classiques de la statistique multivariée.

Le statisticien cherche généralement, dans une première étape, à représenter au mieux ces données fonctionnelles dans un espace de dimension plus petite par l'intermédiaire d'une analyse en composantes principales adaptée au cadre fonctionnel. On peut ainsi déterminer le comportement moyen (courbe moyenne) et les principaux modes de variations autour de la moyenne grâce aux éléments propres de l'opérateur de covariance (Dauxois *et al.* 1982).

La manière dont les données sont obtenues est rarement prise en compte dans ces analyses qui supposent (implicitement) que les observations sont indépendantes et identiquement distribuées. Or cette hypothèse n'est pas systématiquement vérifiée, les courbes observées pouvant provenir d'un plan de sondage élaboré par le statisticien. C'est le cas par exemple pour l'étude de l'évolution de la consommation électrique mesurée à partir d'un échantillon de compteurs tirés selon un plan de sondage complexe (Dessertaine 2006).

Notre objectif dans ce chapitre est de proposer des estimateurs de la courbe moyenne et des éléments propres de l'opérateur de covariance puis d'étudier leurs propriétés dans le cadre des sondages. Les estimateurs sont des fonctions non-linéaires d'estimateurs de

Horvitz-Thompson dont la variance est approchée en utilisant la fonction d'influence (Deville, 1999) et la théorie des perturbations (Kato, 1966). Il existe peu de travaux abordant la question de l'ACP dans la cadre des sondages. Skinner *et al.* (1986) étudient les propriétés de différents plans de sondage sans donner des procédures d'inférence sur les élements propres.

Nous commençons par définir nos estimateurs de la courbe moyenne et de l'acp fonctionnelle dans ce cadre d'échantillonnage. Les techniques de linéarisation basées sur la fonction d'influence permettent ensuite de fournir des estimateurs de la variance dans un cadre asymptotique.
On considère une population U de taille finie N et on s'intéresse à une variable fonctionnelle $\mathcal{Y}$ définie pour chaque individu k de la population U, où $Y_k = (Y_k(t))_{t\in[0,1]}$ appartient à l'espace des fonctions de carré intégrable $L^2[0,1]$ muni de son produit scalaire usuel noté $\langle .,. \rangle$ et de la norme induite $\|.\|$.
On note $\mu \in L^2[0,1]$, la moyenne des Y_k dans la population U

$$\mu(t) = \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,1] \tag{1.24}$$

et Γ, l'opérateur de covariance défini sur $L^2[0,1]$ par

$$\Gamma = \frac{1}{N}\sum_{k\in U} (Y_k - \mu) \otimes (Y_k - \mu) \tag{1.25}$$

où le produit tensoriel de deux éléments a et b de $L^2[0,1]$ est l'opérateur de rang un tel que $a \otimes b(u) = \langle a, u\rangle b$ pour tout u dans $L^2[0,1]$.
L'opérateur Γ est positif et ses éléments propres $(\lambda_j, v_j)_{j=1,\ldots,N}$ vérifient

$$\Gamma v_j(t) \;=\; \lambda_j \; v_j(t), \quad t \in [0,1], \tag{1.26}$$

où les fonctions v_j forment un système orthonormé dans $L^2[0,1]$ et les valeurs propres, qui sont toutes positives ou nulles, sont rangées par ordre décroissant $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_N \geq 0$.
On peut alors montrer que la meilleure approximation linéaire, au sens de l'erreur quadratique, des fonctions Y_k de la population U dans un espace fonctionnel de dimension fixée $q,\ q < N$, est fournie par la projection des Y_k sur les q premières fonctions propres $v_1, \ldots, v_q$ de Γ

$$Y_k^q(t) = \mu(t) + \sum_{j=1}^{q} \langle X_k - \mu, v_j\rangle v_j(t). \tag{1.27}$$

Les fonctions propres v_j indiquent les "principaux modes" de variations de $\mathcal{Y}$ et la variance expliquée par chaque axe v_j est $\lambda_j = \frac{1}{N}\sum_{k\in U}\langle Y_k - \mu, v_j\rangle^2$.
Intéressons nous maintenant à l'estimation des différents paramétres présentés ci-dessus. Soit un échantillon s un échantillon de U, tiré selon un procédé probabiliste $p(s)$.

Les estimateurs par substitution de N, μ et Γ sont définis par

$$\widehat{\mu} = \frac{1}{\widehat{N}} \sum_{k \in s} \frac{Y_k}{\pi_k} \tag{1.28}$$

$$\widehat{\Gamma} = \frac{1}{\widehat{N}} \sum_{k \in s} \frac{Y_k \otimes Y_k}{\pi_k} - \widehat{\mu} \otimes \widehat{\mu} \tag{1.29}$$

où l'estimateur de la taille N de la population est $\widehat{N} = \sum_{k \in s} \frac{1}{\pi_k}$. Les estimateurs des éléments propres $\left(\widehat{\lambda}_j, \widehat{v}_j\right)_{j=1}^{q}$ sont ensuite obtenus par diagonalisation de l'opérateur de covariance $\widehat{\Gamma}$. Les éléments propres sont des fonctions non linéaires de la covariance Γ. La linéarisation par la fonction d'influence (Deville, 1999) combinée à la théorie des perturbations (Kato, 1966) permet cependant d'obtenir des estimateurs de la variance asymptotique des éléments propres. Introduisons la mesure discrète M définie sur $L^2[0,1]$ par $M = \sum_U \delta_{Y_k}$ avec δ_{Y_k} est la fonction de Dirac qui prend la valeur 1 si $\mathcal{Y} = Y_k$ et zéro sinon. Nos paramètres à estimer s'écrivent comme des fonctionnelles non linéaires de totaux par rapport à cette mesure M. Par exemple, $N(M) = \int dM$, $\mu(M) = \int \mathcal{Y} dM / \int dM$ and $\Gamma(M) = \int (\mathcal{Y} - \mu(M)) \otimes (\mathcal{Y} - \mu(M))\, dM / \int dM$. Les éléments propres donnés par by (1.26) sont des fonctionnelles T qui dépendent implicitement de M.
Si on estime la mesure M par la mesure aléatoire $\widehat{M} = \sum_U \frac{\delta_{Y_k}}{\pi_k} I_k$, les estimateurs (1.28) et (1.29) sont obtenus par substitution de M par $\widehat{M}$ et s'écrivent donc comme des fonctionnelles par rapport à la mesure $\widehat{M}$.

Afin de déterminer les propriétés asymptotiques des estimateurs, nous nous plaçons maintenant dans le cadre asymptotique de la théorie des sondages introduit par Isaki et Fuller (1982). On suppose également les hypothèses suivantes

(A1) $\sup_{k \in U} \|Y_k\| \leq C < \infty$,

(A2) $\lim_{N \to \infty} \frac{n}{N} = \pi \in (0,1)$,

(A3) $\min_{k \in U_N} \pi_k \geq \lambda > 0$, $\min_{k \neq l} \pi_{kl} \geq \lambda * > 0$ et $\overline{\lim}_{N \to \infty} n \max_{k \neq l} |\pi_{kl} - \pi_k \pi_l| < \infty$,

avec λ et $\lambda *$ deux constantes positives. Les hypothèses (A1) et (A2) sont vérifiées pour les plans de sondage usuels et (A3) est assez classique en analyse des données fonctionnelles, elle n'implique pas que les courbes $Y_k(t)$ sont uniformément bornées en k et $t \in [0,1]$. On suppose aussi que la fonctionnelle T de chacun de nos paramètres d'intérêt est homogène de degré α, i.e. $T(rM) = r^\alpha T(M)$ et $\lim_{N \to \infty} N^{-\alpha} T(M) < \infty$. Par exemple, μ et Γ sont deux fonctionnelles de degré zéro par rapport à la mesure M. Nous introduisons la norme Hilbert-Schmidt pour les opérateurs définis de $L^2[0,1]$ dans $L^2[0,1]$, notée par $\|\cdot\|_2$.

Dans le résultat suivant, nos montrons que nos estimateurs sont asymptotiquement sans biais, i.e. $\lim_{N \to \infty} \left(E_p(T(\widehat{M})) - T(M) \right) = 0$, en plus ils sont consistants, c-à-d pour tout $\varepsilon > 0$ fixée, on a $\lim_{N \to \infty} P(|T(\widehat{M}) - T(M)| > \varepsilon) = 0$. Notons que $E_p(\cdot)$ est une espérance relative au plan de sondage $p(s)$.

Proposition 1. *Sous les hypothèses (A1), (A2) et (A3),*

$$E_p \left\|\mu - \widehat{\mu}\right\|^2 \;=\; O(n^{-1}), \quad E_p \left\|\Gamma - \widehat{\Gamma}\right\|_2^2 = O(n^{-1}).$$

Si on suppose que les valeurs propres de Γ *sont distinctes et strictement positives, on a :*

$$E_p \left(\sup_j \left|\lambda_j - \widehat{\lambda_j}\right| \right)^2 \;=\; O(n^{-1}), \quad E_p \left\|v_j - \widehat{v_j}\right\|^2 = O(n^{-1}) \quad \textit{pour tout } j \textit{ fixé.}$$

Nous cherchons, dans l'étape suivante, à déterminer une approximation de la variance des différents paramètres d'interêt ainsi que son estimation. Dans ce but nous faisons appel aux techniques de linéarisation par la fonction d'influence. Commençons d'abord par calculer les fonctions d'influence des différents paramètres. Pour cela, notons par $IT(M, \mathcal{Y})$ la fonction d'influence, quand elle existe, d'une fonctionnelle T en un point $\mathcal{Y} \in L^2[0,1]$, définie par

$$IT(M, \mathcal{Y}) \;=\; \lim_{h \to 0} \frac{T(M + h\delta_{\mathcal{Y}}) - T(M)}{h}$$

où $\delta_{\mathcal{Y}}$ est la fonction de Dirac calculée en $\mathcal{Y}$.

Proposition 2. *Supposons que (A1) est vérifiée, les fonctions d'influence de* μ *et* Γ *existent et* $I\mu(M, Y_k) = (Y_k - \mu)/N$ *et* $I\Gamma(M, Y_k) = \frac{1}{N}\left((Y_k - \mu) \otimes (Y_k - \mu) - \Gamma\right)$. *Si les valeurs propres de* Γ *sont distinctes et strictement positives alors*

$$\begin{aligned} I\lambda_j(M, Y_k) &= \frac{1}{N}\left(\langle Y_k - \mu, v_j\rangle^2 - \lambda_j\right), \\ Iv_j(M, Y_k) &= \frac{1}{N}\left(\sum_{\ell \neq j} \frac{\langle Y_k - \mu, v_j\rangle \langle Y_k - \mu, v_\ell\rangle}{\lambda_j - \lambda_\ell} v_\ell\right). \end{aligned}$$

Afin d'obtenir la variance asymptotique de $T(\widehat{M})$ pour toute fonctionnelle T donnée par (1.24), (1.25) et (1.26), nous écrivons le développement de von Mises d'ordre un de notre fonctionnelle T en $\widehat{M}/N$ "proche" de M/N, ensuite, en utilisant le fait que T est homogène de degré 0 et $IT\left(M/N, Y_k\right) = N \cdot IT\left(M, Y_k\right)$,

$$T(\widehat{M}) \;=\; T(M) + \sum_{k \in U} IT\left(M, Y_k\right)\left(\frac{I_k}{\pi_k} - 1\right) + R_T\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right).$$

Proposition 3. *Supposons que les hypothèses (A1), (A2) et (A3) sont vérifiées. On considère les fonctionnelles* T *des différents paramètres donnés par (1.24), (1.25) et*

(1.26). Si on suppose que les valeurs propres sont distinctes et strictement positives, alors $R_T\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right) = o_p(n^{-1/2})$ *et la variance asymptotique de* $T(\widehat{M})$ *est égale à*

$$V_p[\sum_{k\in s} \frac{IT\,(M, Y_k)}{\pi_k}] = \sum_U \sum_U (\pi_{kl} - \pi_k\pi_l) \frac{IT\,(M, Y_k)}{\pi_k} \frac{IT\,(M, Y_l)}{\pi_l}.$$

Comme nous pouvons le remarquer, la variance asymptotique donnée par le résultat ci-dessus n'est pas connue car son expression dépend de la population U entière alors qu'on ne dispose que d'un échantillon de cette population. Pour cela nous proposons de l'estimer par un estimateur de la variance de type Horviz-Thompson et ceci en remplaçant $IT\,(M, Y_k)$ par son estimateur de type HT. Nous obtenons alors les résultats suivants :

$$\begin{aligned}
\widehat{V}_p(\widehat{\mu}) &= \frac{1}{\widehat{N}^2} \sum_{k\in s} \sum_{\ell\in s} \frac{1}{\pi_{k\ell}} \frac{\Delta_{k\ell}}{\pi_k\pi_\ell} (Y_k - \widehat{\mu}) \otimes (Y_\ell - \widehat{\mu}), \\
\widehat{V}_p\left(\widehat{\lambda}_j\right) &= \frac{1}{\hat{N}^2} \sum_{k\in s} \sum_{\ell\in s} \frac{1}{\pi_{k\ell}} \frac{\Delta_{k\ell}}{\pi_k\pi_\ell} \left(\langle Y_k - \widehat{\mu}, \widehat{v}_j\rangle^2 - \widehat{\lambda}_j\right) \left(\langle Y_\ell - \widehat{\mu}, \widehat{v}_j\rangle^2 - \widehat{\lambda}_j\right), \\
\widehat{V}_p\left(\widehat{v}_j\right) &= \sum_{k\in s} \sum_{\ell\in s} \frac{1}{\pi_{k\ell}} \frac{\Delta_{k\ell}}{\pi_k\pi_\ell} \widehat{Iv}_j(M, Y_k) \otimes \widehat{Iv}_j(M, Y_\ell),
\end{aligned}$$

où $\widehat{Iv}_j(M, Y_\ell) = \frac{1}{\hat{N}} \left(\sum_{\ell\neq j} \frac{\langle Y_k - \widehat{\mu}, \widehat{v}_j\rangle\langle Y_k - \widehat{\mu}, \widehat{v}_\ell\rangle}{\widehat{\lambda}_j - \widehat{\lambda}_\ell} \widehat{v}_\ell\right)$.

La proposition suivante montre que ces estimateurs des variances sont asymptotiquement sans biais et convergent en probabilité.

Proposition 4. *Supposons que les hypothèses (A1)-(A3) sont vérifiées, on suppose en plus l'hypothèse A7 dans Breidt et Opsomer (2000). Nous avons alors*

$$\begin{aligned}
E_p \left\| AV(\widehat{\mu}) - \widehat{V}_p(\widehat{\mu}) \right\|_2 &= o\left(\frac{1}{n}\right), \\
E_p \left| AV(\widehat{\lambda}_j) - \widehat{V}_p(\widehat{\lambda}_j) \right| &= o\left(\frac{1}{n}\right).
\end{aligned}$$

Si de plus Γ *est un opérateur de rang fini (indépendant de* N*), alors*

$$\left\| AV(\widehat{v}_j) - \widehat{V}_p(\widehat{v}_j) \right\|_2 = o_p\left(\frac{1}{n}\right)$$

pour $j = 1, \ldots, q$.
Notons que l'opérateur de covariance des vecteurs propres dépend de sommes infinies dont les termes tendent vers zéro et sont tronquées pour la mise en œuvre.

Implémentation informatique et étude sur simulations

Nous avons simulé une population de 10000 trajectoires de mouvements browniens sur l'intervalle [0,1], discrétisées en 100 points équidistants. Nous avons ensuite constitué

deux fois 500 plans de sondage (aléatoire simple et stratifié), avec des échantillons de tailles $n = 100, 500$ et 1000. Nous avons ensuite évalué la qualité des variances approximées par linéarisation des éléments propres de l'opérateur de covariance. Il s'avère que la technique de linéarisation fournit de bonnes approximations de la variance des valeurs propres, même pour des tailles relativement petites d'échantillons. L'approximation des variances des vecteurs propres est aussi satisfaisante pour les premiers mais se révèle moins performante pour les vecteurs propres associés à de petites valeurs propres, ce qui était prévisible. Les programmes permettant de faire ces simulations sont donnés en Annexe.

Perspectives. Différentes perspectives peuvent être proposées au vu de ce chapitre. La première est d'appliquer les résultats trouvés sur un jeu de données réelles. Une autre perspective consiste à améliorer la qualité des estimateurs présentés dans ce chapitre, lorsque nous nous disposons d'une information auxilaire. Pour cela nous pouvons utiliser l'ACP fonctionnelle conditionnelle introduite par Cardot (2007). Enfin il serait également intéressant de trouver des résultats similaires à ceux de ce chapitre lorsque les probabilités d'inclusions dépendent du temps.

1.2.4 Estimation non paramétrique des quantiles géométriques conditionnels sous l'hypothèse de mélange fort

Dans le Chapitre 2, nous nous avons étudié l'estimation des quantiles géométriques conditionnels lorsque les observations sont supposées indépendantes et identiquement distribuées. Cette hypothèse n'est pas toujours valide dans la pratique surtout lorsqu'on on s'intéresse à l'estimation dans le cadre de la prévision.
Pour cette raison, nous proposons d'étudier, dans le Chapitre 5 de ce mémoire, l'estimation des quantiles géométriques lorsque les observations sont dépendantes. Nous nous sommes intéressés à un cas particulier de dépendance à savoir le mélange fort (ou α-mélange).
Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^s \times \mathbb{R}^d$ (avec $s \geq 1$ et $d \geq 2$). Soit $F(.|x)$ la fonction de répartition conditionnelle de Y sachant $X = x$ et $P(.|x)$ une mesure de probabilité définie sur tout borélien V de $\mathbb{R}^d$ par $P(V|x) = \int_V F(dy|x)$. On note par $\varphi(\theta, u, x) = \int_{\mathbb{R}^d} [\phi(u, y - \theta) - \phi(u, y)] F(dy|x)$. Le quantile géométrique conditionnel de Y sachant $X = x$ est définie comme une solution du problème de minimisation suivant :

$$\begin{aligned} Q(u|x) &= \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\phi(u, Y - \theta) - \phi(u, Y)|X = x\right] \\ &= \arg\min_{\theta \in \mathbb{R}^d} \varphi(\theta, u, x). \end{aligned} \tag{1.30}$$

Soient $(X_1, Y_1), \ldots, (X_n, Y_n)$ des observations de (X, Y) vérifiant l'hypothèse d'α-mélange : il existe une suite $(\alpha_n)_{n\in\mathbb{N}}$, avec $\lim_{n\to\infty} \alpha_n = 0$, telle que pour tout $n \in \mathbb{N}$ $\quad |P(A \cap B) - P(A)P(B)| \leq \alpha_n$, avec $A \in \mathcal{F}_1^t$, $B \in \mathcal{F}_{t+n}^\infty$ où $\mathcal{F}_\mu^\nu$ est la tribu engendrée par $\{(X_t, Y_t) : \mu \leq t \leq \nu\}$. Cette définition est celle introduite par Rosenblatt (1956). Il existe d'autres définitions de mélange (voir Doukhan, 1994, pour plus de détails).

Lorsque $\alpha_n = 0$, pour tout n, on retrouve le cas i.i.d.
Pour estimer $Q(u|x)$, il suffit d'introduire $F_n(\cdot|x)$ un estimateur non paramétrique de type Nadaraya-Watson de la distribution conditionnelle de Y sachant X, défini pour tout $y \in \mathbb{R}^d$, par $F_n(y|x) = \sum_{i=1}^n w_{n,i} \mathbb{1}_{\{Y_i \geq y\}}$, où $w_{n,i} = K((x - X_i)/h_n) / \sum_{i=1}^n K((x - X_i)/h_n)$ représente le poids associé à Y_i. La discussion faite, sur le choix du noyau K et de la fenêtre h_n dans le cadre multidimentionnel, dans le Chapitre 2 reste valable dans ce cas. Soit $P_n(\cdot|x)$ un estimateur de la mesure de probabilité $P(\cdot|x)$, définie, pour tout ensemble Borélien $V \in \mathbb{R}^d$, par $P_n(V|x) = \int_V F_n(dy|x)$. Un estimateur du quantile géométrique conditionnel est donné par l'expression suivante :

$$\begin{aligned} Q_n(u|x) &= \arg\min_{\theta \in \mathbb{R}^d} \varphi_n(\theta, u, x) \\ &= \arg\min_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} [\phi(u, y - \theta) - \phi(u, y)] F_n(dy|x). \end{aligned} \tag{1.31}$$

Dans ce qui suit nous montrons que $Q_n(u|x)$ est un estimateur uniformment convergent de $Q(u|x)$ sur tout ensemble compact $\mathcal{C}$ de $\mathbb{R}^s$.

Propriétés asymptotiques. Soit g la densité de X supposée bornée sur tout ensemble compact $\mathcal{C}$ de $\mathbb{R}^s$. Introduisons maintenant les hypothèses suivantes :

(H1) La densité g de X est uniformement continue,
(H2) Le noyau K est continu, borné, positif et satisfait $\int K(v)dv = 1$, $\int v_i K(v)dv = 0$ pour tout $i \in \{1, \ldots, s\}$, and $\|v\|^s K(v) \to 0$ as $\|v\| \to \infty$,
(H3) Il existe $\delta > 0$ tel que $n^{1/4}(h_n^s)^{(1+\delta)/4} / \log n \to \infty$ quand $n \to \infty$,
(H4) Pour tout borélien V de $\mathbb{R}^d$ et $\theta \in \mathbb{R}^d$, les fonctions $P(V|\cdot)$ et $\varphi(\theta, u, \cdot)$ sont continues sur $\mathcal{C}$,
(H5) La fonction $Q(u|\cdot)$ satisfait la propriété suivante sur $\mathcal{C}$, $\forall \epsilon > 0$, $\exists \eta > 0$, $\forall z : \mathcal{C} \to \mathbb{R}^d$,

$$\sup_{x \in \mathcal{C}} \|Q(u|x) - z(x)\| \leq \epsilon \Rightarrow \sup_{x \in \mathcal{C}} |\varphi(Q(u|x), u, x) - \varphi(z(x), u, x)| \leq \eta.$$

Commentaires sur les hypothèses. Les hypothèses (H1) et (H2) sont classiques en estimation non paramétrique. L'hypothèse d'α-mélange est raisonnable car les modéles de séries temporelles satisfont cette condition. L'hypothèse (H3) est utilisée pour montrer que $\sup_{y \in \mathbb{R}^d} |F_n(y|x) - F(y|x)| \to 0$ sous l'hypothèse de mélange. (H4) implique la convergence uniforme de $\varphi_n(\theta, u, \cdot)$ vers $\varphi(\theta, u, \cdot)$. Enfin, l'hypothèse (H5) a déjà été introduite par Collomb et *al.* (1987) pour montrer la convergence de l'estimateur du mode conditionnel.

Théorème 1.2.2 *Si les hypothèses (H1)-(H5) sont satisfaites, alors*
(i) il existe un entier $N > 1$ telle que si $n \leq N$ et $x \in \mathcal{C}$, $Q_n(u|x)$ existe et est unique ;
(ii) la fonction $Q_n(u|\cdot)$ est continue sur $\mathcal{C}$;
(iii) $\sup_{x \in \mathcal{C}} \|Q_n(u|x) - Q(u|x)\| \to 0$ quand $n \to \infty$.

Schéma de la preuve. La démonstration de ce théorème, repose sur les lemmes suivants.

Lemme 1.1 *Sous les hypothèses (H1)-(H4), nous avons*

$$\lim_{\|\theta\|\to\infty} \sup_{n\geq 1} \sup_{x\in\mathcal{C}} \left| \frac{\varphi_n(\theta,u,x) + \langle u,\theta\rangle}{\|\theta\|} - 1 \right| = 0.$$

Lemme 1.2 *Sous les hypothèses (H1)-(H4), et pour tout $A > 0$, nous avons*

$$\sup_{\|\theta\|\leq A} \sup_{x\in\mathcal{C}} |\varphi_n(\theta,u,x) - \varphi(\theta,u,x)| \to 0 \qquad \textit{lorsque} \quad n\to\infty.$$

Les assertions (i) et (ii) du théorème se déduisent à partir du lemme 1.1. L'assertion (iii) se démontre en deux étapes.

Étape 1. En utilisant le Lemme 1.1, nous pouvons démontrer qu'il existe un réel $r > 0$ et $N \geq 1$ tels que

$$\sup_{n\geq N} \sup_{x\in\mathcal{C}} \|Q_n(u|x)\| \leq r \quad \text{et que} \quad \sup_{x\in\mathcal{C}} \|Q(u|x)\| \leq r.$$

Étape 2. En utilisant l'inégalité triangulaire et le résultat de l'étape 1, nous démontrons que

$$\sup_{x\in\mathcal{C}} |\varphi(Q(u|x),u,x) - \varphi(Q_n(u|x),u,x)| \leq 2 \sup_{\|\theta\|\leq r} \sup_{x\in\mathcal{C}} |\varphi_n(\theta,u,x) - \varphi(\theta,u,x)|.$$

De plus sous les hypothèses (H1)-(H4) et le résultat du Lemme 1.2, nous obtenons que

$$\sup_{x\in\mathcal{C}} |\varphi_n(\theta,u,x) - \varphi(\theta,u,x)| \to 0 \quad \text{lorsque} \quad n\to\infty.$$

Ensuite en appliquant l'hypothèse (H5), nous montrons que

$$\sup_{x\in\mathcal{C}} |Q(u|x) - Q_n(u|x)| \to 0 \quad \text{lorsque} \quad n\to\infty.$$

Perspectives. Dans la littérature, plusieurs auteurs présentent les quantiles univariés comme une alternative intéressante au modèle linéaire pour faire de la prévision. L'estimation des quantiles géométriques conditionnels peut servir à la construction de prédicteurs non paramétriques des séries chronologiques. En effet dans la pratique, nous pouvons être amenés à prévoir une série temporelle multivariée : c'est par exemple lorsqu'on cherche à prévoir deux séries de façon simultanée ou à faire de la prévision multihorizon. De Gooijer et *al.* (2006) ont étudié ce problème en utilisant la définition des quantiles multivariés introduite par Abdous et Theodorescu (1992). Comme nous l'avons précisé précédemment, la définition d'Abdous et Theodoresu est invariante par rotation mais elle ne l'est pas par transformation affine. Pour cette raison, l'utilisation de la définition des quantiles géométriques conditionnels présentée dans ce chapitre (qui est invariante par rotation et transformation affine) devrait permettre d'améliorer la qualité de la prévision.

Références

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

Babu, G. J. and Rao, C. R. (1988). Joint asymptotic distribution of marginal quantile functions in samples from a multivariate population. *Journal of Multivariate Analysis*, **27**, 15-23.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, **37**, 577-580.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Ser. A*, **139**, 318-354.

Bedall, F.K. and Zimmermann, H. (1979). Algorithm AS 143, the Mediancenter. *Applied Statistics*, **28**, 325-328.

Berger, Y.G. et Skinner, C.J. (2005). A jackknife variance estimator for unequal probability sampling . *Journal of the Royal Statistical Society Series B*, **67**, 1, 79-89.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.

Brown, B. M. (1983). Statistical use of the spatial median. *Journal of the Royal Statistical Society, Ser. B*, **45**, 25-30.

Brown, B. M. and Hettmansperger, T. P.(1987). Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Ser. B*, **49**, 301-310.

Brown, B. M. and Hettmansperger, T. P.(1989). An affine invariant bivariate version of the sign test. *Journal of the Royal Statistical Society, Ser. B*, **51**, 117-125.

Cardot, H. (2007). Conditional functional principal components analysis. *Scandinavian J. of Statistics* , **34**, 317-335.

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

Chaudhuri, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *The Annals of Statistics*, **20**, 897-916.

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

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

Cheng, Y. and De Gooijer J. (2007). On the uth geometric conditional quantile. *Journal of Statistical Planning and Inference*, **137**, 1914-1930.

Collomb, G., HŁrdle, W. and Hassani, S. (1987). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, **15**, 227-236.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a random vector function : some applications to statistical inference. *J. Multivariate Anal.*, **12**, 136-154.

De Gooijer J. G., Gannoun, A. et Zerom, D. (2006). A multivariate quantile predictor. *Communications in Statistics-Theory and Methods*, 35, 133-147.

Dessertaine A. (2006). Sondage et séries temporelles : une application pour la prévision de la consommation électrique. *38èmes Journées de Statistique*, Clamart, Juin 2006.

Deville, J.C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, **25**, 193-203.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, **20**, 1803-1827.

Doukhan, P. (1994). Mixing : Properties and Examples. *Lecture Notes in Statistics*, **85**, Springer, New York.

Eddy, W.F. (1982). Convex Hull Peeling. *COMPSTAT 1982 for IASC*, Vienna : Pysica-Verlag, 42-47.

Eddy, W.F. (1985). Ordering of Multivariate Data. *Computer Science and Statistics : The Interface*, ed. L. Billard, Amesterdam : North-Holland, 25-30.

Ferguson, T. (1967). *Mathematical Statistics : A Decision Theoric Approach.* Academic Press, New York.

Gannoun, A., Saracco, J., Yan, A. and Bonney, G.E. (2003a). On adaptive transformation-retransformation estimate of conditional spatial median. *Communications in Statistics - Theory and Methods*, **32**, 1981-2011.

Gower, J.C. (1974). Algorithm AS 78 : The Mediancenter. *Applied Statistics*, **23**, 466-470.

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

Isaki, C. I. and Fuller, W. A. (1982). Survey Design under the regression superpopulation model. *Journal of the American Statistical Association*, **77**, 89-96.

Kato, T. (1966). *Perturbation theory for linear operators.* Springer Verlag, Berlin.

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

Kim, S. J. (1992). A metrically trimmed mean as a robust estimator of location. *The annals of Statistics*, **20**, 1534-1547.

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Kokic, P., Breckling, J., Lübke, O. (2002). A new definition of multivariate M-quantiles. Statistical data analysis based on the L_1-norm and related methods. Stat. Ind. Technol. : *Statistical Data Analysis*, 15-24, Birkhäuser Verlag Basel/Switzerland.

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

Kovacevic, M. S. et Binder, D.A. (1997). ÊVariance estimation for measures of income inequality and polarizationÊ. *Journal of Official Statistics*, **13**, 1, 41-58.

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

Oja, H. (1983). Descriptive statistics for multivariate trimming. *Statistics and Probability Letters*, **1**, 327-332.

Ramsay, J. O. and Silverman, B.W. (2002). *Applied Functional Data Analysis : Methods and Case Studies*. Springer-Verlag.

Rao, J. N. K., Wu, C. F. J. et Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, **18**, 209-217.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.*, **42**, 43-47.

Särndal C.E. , Swensson B. and Wretman J. (1992). *Model Assisted Survey Sampling*, Springer, Berlin.

Scates, D. E. (1933). Locating the median of the population in the United States. *Metron*, **11**, 49-66.

Small, C. G. (1990). A survey of multidimensional median. *International Statistical Review*, **58**, 263-277.

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

Skinner, C.J, Holmes, D.J, Smith, T.M.F (1986). The Effect of Sample Design on Principal Components Analysis. *J. Am. Statist. Ass.* **81**, 789-798.

Tillé, Y. (2001). *ThÕorie des sondages : Échantillonnage et estimation en populations finies : cours et exercices*, 284 pages, Paris, Dunod.

Vardi, Y. et Zhang, C. H. (2000). The multivariate L_1-median and associated data depth. *PNAS*, **4**, 1423-1426.

Weber, A. (1909). Über den standard der industrien, Tubingen, English translation by K.J. Freidrich (1929), Alfred Weber's theory of location of industies, Chicago University Press.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, **28**, 461-482.

1.3 Liste des travaux

Publications

1. Chaouch, M., Gannoun, A. et Saracco, J. (2007). Estimation de quantiles géométriques conditionnels et non conditionnels. En révision pour le *JSFdS/RSA*.

2. Chaouch, M. and Goga C. (2008). Design-Based Estimation for Geometric Quantiles. Submitted at *International Statistical Review*.

3. Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2007). Functional Principal Components Analysis with Survey Data. Submitted at *Journal of Statistical Planning and Inference*.

4. Chaouch, M. (2008). Estimation non paramétrique des quantiles géométriques conditionnels sous une hypothèse d'α-mélange. Soumis aux *C. R., Math., Acad. Sci. Paris*.

Chapitres de livres

1. Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2008). Functional Principal Components Analysis with Survey Data. In Functional and Operatorial Statistics, Dabo-Niang, S. and Ferraty, F. (Eds.), Physica-Verlag, Heidelberg, 95-102.

Articles parus dans des Actes de conférences (avec comité de lecture)

1. Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2008). Functional Principal Components Analysis with Survey Data. *International Workshop on Functional an Operatorial Statistics. Toulouse, 19-21 Juin 2008.*
2. Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2007). Sondage et ACP fonctionnelle : une approche par la fonction d'influence. $5^{ème}$ *Colloque francophone sur les sondages. CIRM, Marseille, 5-7 Novembre 2007.*
3. Chaouch, M. (2007). Un exemple de quantile spatial conditionel. $2^{èmes}$ *Journées des jeunes statisticiens. Aussois, 3-7 Septembre 2007.*
4. Cardot, H., Chaouch, M., Goga, C. et Labruère, C. (2007). Echantillonnage et ACP fonctionnelle. $39^{èmes}$ *Journées de Statistique. Angers, 11-15 Juin 2007.*
5. Chaouch, M., Gannoun, A. et Saracco, J. (2007). Quantile spatial conditionnel et non conditionnel. $39^{èmes}$ *Journées de Statistique. Angers, 11-15 Juin 2007.*

Chapitre 2

Estimation des quantiles géométriques conditionnels et non conditionnels*

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 reponsant 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.*

Contents

2.1 Introduction 32
2.2 Quantile univarié 33
2.3 Quantile géométrique 35
2.4 Quantile géométrique conditionnel 43
2.5 Implémentation en R des algorithmes de calculs des quantiles (conditionnels) géométriques 47
2.6 Etude par simulation 50
2.7 Étude sur des données réelles 52

*Article écrit en collaboration avec Ali Gannoun et Jérôme Saracco et il est en révision pour le *JSFdS/RSA*.

2.1 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.* (2002) 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.*, 2002, et Gannoun et *al.*, 2003b). 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 (1976), Plackett (1976), Reiss (1989), Eddy (1982, 1985). Brown et Hettmansperger (1987, 1989) ont introduit la notion de quantile bivarié en se basant sur la définition de la médiane d'Oja (1983). Par la suite, Babu et Rao (1988) et Abdous et Theodorescu (1992) 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 (1983) pour la médiane et ceux de Donoho et Gasko (1992), Liu et *al.* (1999) et Zuo et Serfling (2000) pour les quantiles multivariés. La seconde approche a été utilisée, en premier lieu, par Brown (1983), Gower (1974), Haldane (1948) et Chaudhuri (1992) pour généraliser la notion de la médiane au cas multivarié. Ensuite, Abdous et Theodorescu (1992), Chaudhuri (1996), Koltchinskii (1997) et Kokic et *al.* (2002) 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 (2002).

Dans ce qui suit, nous nous focalisons sur la définition des quantiles, dit *géométriques*, introduite par Chaudhuri (1996). 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 (2001) et Gannoun et *al.* (2003a)). Ces différents points sont décrits à la Section 3. Une utilisation des quantiles géométriques en statistique descriptive multivariée est disponible

dans Serfling (2004).

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 (1996) 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.

2.2 Quantile univarié

2.2.1 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

$$Q_F(p) = F^{-1}(p) = \inf\{y : F(y) \geq p\}, \tag{2.1}$$

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

2.2.2 Deux caractérisations du quantile univarié

2.2.2.1 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 :

$$Q(u) = F^{-1}\left(\frac{1+u}{2}\right) \quad \text{avec} \quad -1 < u < 1. \tag{2.2}$$

Nous remarquons que, contrairement à la définition donnée par (2.1), le quantile est indexé par $u \in (-1,1)$. La définition de la fonction quantile $Q(.)$ donnée par (2.2) 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

$$\mathbb{E}\left(S(\theta - Y)\right) - u = 0, \tag{2.3}$$

où S désigne la fonction "signe" univariée définie par

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

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

Preuve :

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

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 (2.3).

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

Ferguson (1967) et Koenker et Basset (1978), 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

$$Q_M\left(u\right) = \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\{\phi\left(u, Y - \theta\right)\} = \arg\min_{\theta \in \mathbb{R}} \int_{\mathbb{R}} \left(|y - \theta| + u\left(y - \theta\right)\right) F\left(dy\right). \tag{2.4}$$

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

$$\mathbb{E}\left(S\left(Y - \theta\right)\right) + u = 0, \tag{2.5}$$

qui est équivalente à l'équation 2.3.

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

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

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

2.2.3 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

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq y\}}. \tag{2.6}$$

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

$$Q_{F_n}(p) = F_n^{-1}(p) = \inf\{y : F_n(y) \geq p\}. \tag{2.7}$$

Pour $u = 2p - 1$, la caractérisation donnée par l'équation (2.3) permet de définir un estimateur $Q_n(u)$ de $Q(u)$ comme la solution θ de l'équation suivante

$$\frac{1}{n} \sum_{i=1}^{n} S(\theta - Y_i) = u. \tag{2.8}$$

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 (2.4), pour $u = 2p - 1$, $Q_M(p)$ est estimé par

$$Q_{M,n}(u) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} \phi\left(u, Y_i - \theta\right) = \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} |Y_i - \theta| + u\left(Y_i - \theta\right). \tag{2.9}$$

On peut montrer que $Q_{M,n}(u)$ est solution de l'équation (2.8) dont l'inconnue est θ. Ainsi, pour $u = 2p - 1$, ces trois estimateurs sont identiques :

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

2.3 Quantile géométrique

On suppose maintenant que $\mathbf{Y} \in \mathbb{R}^d$. La définition donnée en (2.1), 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 (1996) a proposé une définition du quantile multivarié, dit géométrique, qui généralise la définition du quantile univarié donnée en (2.4).

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.

2.3.1 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

$$\mathbf{Q}(\mathbf{u}) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left\{ \phi(\mathbf{u}, \mathbf{Y} - \theta) \right\}. \tag{2.10}$$

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 (1987), 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 :

$$\mathbf{Q}(\mathbf{u}) = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left\{ \phi(\mathbf{u}, \mathbf{Y} - \theta) - \phi(\mathbf{u}, \mathbf{Y}) \right\}. \tag{2.11}$$

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 :

$$\mathbb{E}\left(\mathbf{S}(\theta - \mathbf{Y}) \right) - \mathbf{u} = 0. \tag{2.12}$$

2.3.2 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 :

$$\begin{aligned} \mathbf{Q}_n(\mathbf{u}) &= \arg\min_{\theta \in \mathbb{R}^d} \int \left(\phi(\mathbf{u}, \mathbf{y} - \theta) - \phi(\mathbf{u}, \mathbf{y}) \right) F_n(d\mathbf{y}) \\ &= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \left(\phi(\mathbf{u}, \mathbf{Y}_i - \theta) - \phi(\mathbf{u}, \mathbf{Y}_i) \right) \end{aligned} \tag{2.13}$$

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

$$\mathbf{Q}_n(\mathbf{u}) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} \phi(\mathbf{u}, \mathbf{Y}_i - \theta). \tag{2.14}$$

On arrive facilement, à partir de l'équation (2.13), à montrer que $\mathbf{Q}_n(\mathbf{u})$ est aussi solution de l'équation suivante dont l'inconnue est θ :

$$\frac{1}{n} \sum_{i=1}^{n} S(\theta - \mathbf{Y}_i) = \mathbf{u}. \tag{2.15}$$

2.3.3 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 alignés, par le théorème 2.17 de Kemperman (1987), 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.

2.3.4 Interprétation du vecteur 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.

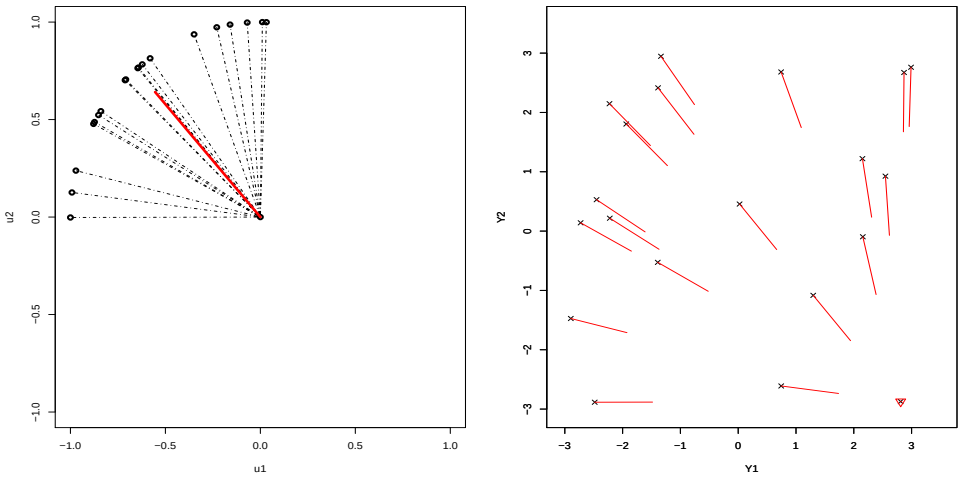


FIG. 2.1 – A droite, le nuage des 20 observations sur lequel chaque segment représente le vecteur normé $S(\theta - \mathbf{Y}_i)$ qui relie une observation i au quantile géométrique $\mathbf{Q}(\mathbf{u})$ (qui est le point représenté par un triangle et situé en bas à droite). A gauche, le vecteur $\mathbf{u}$ (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 (2.15), nous déterminons le vecteur $\mathbf{u}$ correspondant à ce quantile. Ce

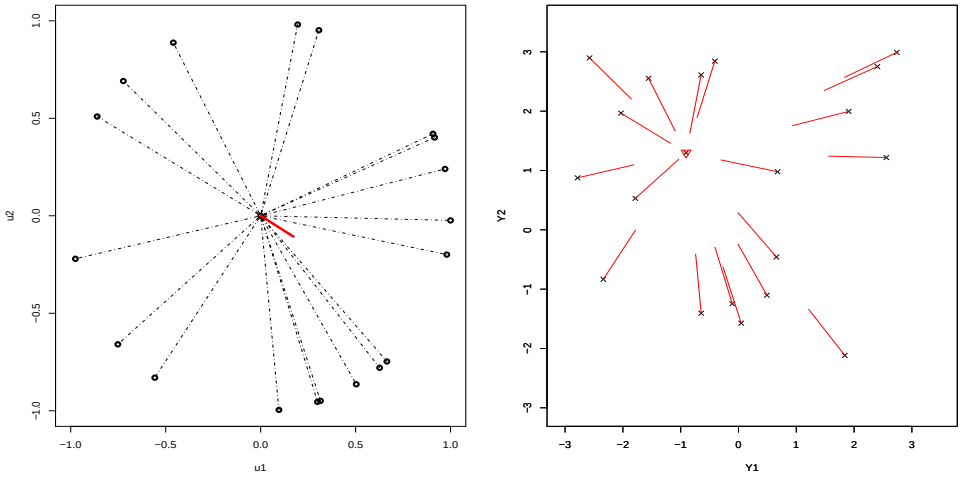


FIG. 2.2 – A droite, le nuage des 20 observations sur lequel chaque segment représente le vecteur normé $S(\theta - \mathbf{Y}_i)$ qui relie une observation i au quantile géométrique $\mathbf{Q}(\mathbf{u})$ (qui est le point représenté par un triangle et situé au centre de nuage des points). A gauche, le vecteur $\mathbf{u}$ (trait continu) est la moyenne des vecteurs unitaires $S(\theta - \mathbf{Y}_i)$ (tracés en pointillés).

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 $\mathbf{u}$ (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.

Remarques. (Serfling, 2002)

(1) 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 $||\mathbf{u}||$.

(2) 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 seconde remarque 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 ainsi que

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

TAB. 2.1 – Pourcentage des simulations où l'ensemble $\mathcal{R} = \{\mathbf{Q}_n(\mathbf{u}) : ||\mathbf{u}|| \leq 0.5\}$ contient 50% des observations.

la Figure 3, ont pour but d'illustrer, graphiquement et numériquement, les deux remarques ci-dessus. L'exemple 1 (resp. l'exemple 2) correspond à la remarque (1) (resp. à la remarque (2)).

Exemple 1 (Serfling, 2002). 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 $Q(-0.1) = F^{-1}(0.45) = -10$.

- 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 2.
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.

2.3.5 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 (1996) a établi une représentation de type Bahadur (Bahadur, 1966) 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 2.3.1 *(Chaudhuri, 1996)*
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 :

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

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 2.3.2 *(Chaudhuri, 1996)*
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 $\sqrt{n}\left(\mathbf{Q}_n(\mathbf{u}_r) - \mathbf{Q}(\mathbf{u}_r)\right)$ et $\sqrt{n}\left(\mathbf{Q}_n(\mathbf{u}_s) - \mathbf{Q}(\mathbf{u}_s)\right)$, où $1 \leq r, s \leq k$, 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, 1996, Koltchinskii, 1997) 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.

2.3.6 Technique de Transformation-Retransformation (TR)

Cette approche a été introduite, dans un premier temps, par Chaudhuri et Sengupta (1993), pour construire des tests de signe multivariés invariants par transformation affine. Ensuite, elle a été utilisée par Chakraborty et Chaudhuri (1996), Gannoun et *al.* (2003a) et De Gooijer et Gannoun (2007) pour donner une version invariante par transformation affine de la médiane spatiale et la médiane spatiale conditionnelle. Chakraborty (2001) 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 $\{1, 2, \ldots, n\}$. 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 la 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 $\mathbf{R}_n^{(\alpha)}(\mathbf{v})$, calculé à partir des observations $\mathbf{Y}_j^{(\alpha)}, j \notin \alpha$, au moyen de la formule (2.14). 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 systéme 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 2.3.3 *(Chakraborty, 2001)*
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 $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$.

2.3.6.1 Choix de α

La qualité de 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 (2001) 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.

2.3.7 Un algorithme d'estimation

Le problème de calcul de la médiane spatiale comme étant la quantité qui minimise $\sum_{i=1}^n ||\mathbf{Y}_i - \mathbf{Q}||$ a été abordé par Bedall et Zimmermann (1979) et Gower (1974). Des algorithmes de minimisation ont été proposés par ces auteurs. Récemment Chaudhuri (1996) 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 2.3.4 *(Chauduri, 1996)*
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* $\boldsymbol{Q}_n(\boldsymbol{u}) \neq \boldsymbol{Y}_i, \quad \forall\ 1 \leq i \leq n$, *alors on a :*

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

- *Si* $\exists\ 1 \leq i \leq n \quad$ *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 :

$$\left\| \sum_{\substack{1 \leq j \leq n \\ j \neq i}} [S(\mathbf{Y}_j - \mathbf{Y}_i)] + (n-1)\mathbf{u} \right\| \leq (1 + ||\mathbf{u}||). \tag{2.16}$$

  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

$$\sum_{i=1}^{n} S(\mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u})) + n\mathbf{u} = 0. \tag{2.17}$$

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. 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 (2001).

2.4 Quantile géométrique conditionnel

2.4.1 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. (2002) pour une rapide description de ces méthodes).

L'intérêt pour les quantiles multivariés conditionnels est tout à fait récent. De Gooijer et al. (2006) ont généralisé au cadre conditionnel la définition du quantile spatial basée sur la minimisation de la semi-norme donnée par Abdous et Theodorescu (1992). Cheng et De Gooijer (2007) se sont aussi intéressés au même problème en généralisant la définition du quantile géométrique, introduite par Chaudhuri (1996). 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 :

$$\begin{aligned} \mathbf{Q}(\mathbf{u}|\mathbf{x}) &= \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\phi(\mathbf{u}, \mathbf{Y} - \theta) - \phi(\mathbf{u}, \mathbf{Y}) \mid \mathbf{X} = \mathbf{x}\right] \\ &= \arg\min_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} \left\{\phi(\mathbf{u}, \mathbf{y} - \theta) - \phi(\mathbf{u}, \mathbf{y})\right\} G(d\mathbf{y}|\mathbf{x}). \end{aligned} \tag{2.18}$$

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 :

$$\mathbb{E}\left(S\left(\theta - \mathbf{Y}\right) \mid \mathbf{X} = \mathbf{x}\right) = \mathbf{u}. \tag{2.19}$$

2.4.2 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 $w_{n,i} = K\left(\left(\mathbf{x} - \mathbf{X}_i\right)/h_n\right) / \sum_{i=1}^{n} K\left(\left(\mathbf{x} - \mathbf{X}_i\right)/h_n\right)$ 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 $\mathbf{Q}(\mathbf{u}|\mathbf{x})$:

$$\begin{aligned} \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) &= \arg\min_{\theta \in \mathbb{R}^d} \int_{\mathbb{R}^d} \left\{\phi(\mathbf{u}, \mathbf{y} - \theta) - \phi(\mathbf{u}, \mathbf{y})\right\} G_n(d\mathbf{y}|\mathbf{x}) \\ &= \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} w_{n,i} \left\{\phi(\mathbf{u}, \mathbf{Y}_i - \theta) - \phi(\mathbf{u}, \mathbf{Y}_i)\right\} \end{aligned} \tag{2.20}$$

A partir de l'équation (2.19), l'estimateur $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ peut être regardé comme la solution de l'équation suivante dont l'inconnue est θ :

$$\int S(\theta - \mathbf{t})\, G_n(d\mathbf{t}|\mathbf{x}) = \sum_{i=1}^{n} w_{n,i} S(\theta - \mathbf{Y}_i) = \mathbf{u}. \tag{2.21}$$

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

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

Cheng et De Gooijer (2007) 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 $\mathbf{t} \in N(\mathbf{x})$,
$D_t = \mathbb{E}\left[(S(\mathbf{Y} - \mathbf{Q}(\mathbf{u}|\mathbf{x})) + \mathbf{u})\,(S(\mathbf{Y} - \mathbf{Q}(\mathbf{u}|\mathbf{x})) + \mathbf{u})^T |\mathbf{X} = \mathbf{t}\right]$ est continue en tout point $\mathbf{t} = \mathbf{x}$. On suppose également que $\gamma > 2/(2+s)$.

Théorème 2.4.1 *(Cheng et De Gooijer, 2007)*
Sous les hypothèses $(H1) - (H7)$, la représentation de type Bahadur de $\boldsymbol{Q}_n(\boldsymbol{u}|\mathbf{x})$ est donnée par la relation suivante

$$\boldsymbol{Q}_n(\boldsymbol{u}|\mathbf{x}) - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x}) = D_{1n}^{-1} \sum_{i=1}^{n} w_{n,i} \left[S(\mathbf{Y}_i - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x})) + \mathbf{u}\right] + R_n, \tag{2.22}$$

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 2.4.2 *(Cheng et De Gooijer, 2007) Sous les hypothèses* $(H1)-(H8)$*, alors on a*

$$\sqrt{\frac{nh_n^s g(\mathbf{x})}{\int K^2(\mathbf{z})d\mathbf{z}}} D_x^{-1/2} D_{1n} \left(\boldsymbol{Q}_n(\boldsymbol{u}|\mathbf{x}) - \boldsymbol{Q}(\boldsymbol{u}|\mathbf{x}) - \frac{h_n^2}{2} D_{1n}^{-1} \xi_s \right) \to N(0, I_d), \tag{2.23}$$

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

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

2.4.4 Un algorithme d'estimation du quantile géométrique conditionnel

Commençons par généraliser le théorème 2.3.4 au cas conditionnel.

Théorème 2.4.3 *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* $\boldsymbol{Q}(\boldsymbol{u}|\boldsymbol{x})$*.*

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

$$\sum_{i=1}^{n} S(\boldsymbol{Y}_i - \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})) K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i) + \boldsymbol{u} \sum_{i=1}^{n} K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i) = 0 \tag{2.24}$$

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

$$\left\| \sum_{\substack{1\leq i\leq n \\ \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})\neq \boldsymbol{Y}_i}} \left[S(\boldsymbol{Y}_i - \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})) + \mathbf{u} \right] K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i) \right\| \leq \sum_{\substack{1\leq i\leq n \\ \boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})= \boldsymbol{Y}_i}} K_{h_n}(\boldsymbol{x} - \boldsymbol{X}_i)(1 + ||\boldsymbol{u}||) \tag{2.25}$$

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 :

$$\left\| \sum_{\substack{1\leq j\leq n \\ j\neq i}} \left[S(\mathbf{Y}_j - \mathbf{Y}_i) + \mathbf{u} \right] K_{h_n}(\mathbf{x} - \mathbf{X}_j) \right\| \leq K_{h_n}(\mathbf{x} - \mathbf{X}_i)(1 + ||\mathbf{u}||) \tag{2.26}$$

  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 (2.24).

- ***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.

2.5 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.

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

Le programme mis en œuvre pour estimer le quantile géométrique est appelé "**QuantileNC.est**".

2.5.1.1 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).
- $\mathbf{u}$: 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 (2.16). 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 (2.16) 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 (2.17).

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 (2.16) était verifiée ou non. Cette variable prend la valeur $TRUE$ si la condition est verifiée et $FALSE$ sinon.

2.5.1.2 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 $MatY$, **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.

2.5.2 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 $X \in \mathbb{R}$.

2.5.2.1 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 (2007) 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 :

$$\tilde{h} = \arg\min_{h>0} \sum_{j=1}^{n} ||\mathbf{Q}_n(\mathbf{u}|x) - \mathbf{Q}_n^{-j}(\mathbf{u}|x)||^2 \tag{2.27}$$

où $\mathbf{Q}_n^{-j}(\mathbf{u}|x)$ désigne 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.

2.5.2.2 Description des fonctions

Pour calculer l'estimateur du quantile géométrique conditionnel on 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 $X = x$
- $\mathbf{u}$ 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 :

- $\mathbf{Q}$: l'estimateur du quantile géométrique de $\mathbf{Y}$ conditionnellement à $X = x$, relatif au vecteur $\mathbf{u}$.
- *test* : une variable logique qui prend la valeur $TRUE$ si la condition (2.26) est verifiée et $FALSE$ sinon.
- $\mathbf{u}$ et $||\mathbf{u}||$: la direction $\mathbf{u}$ 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. La fonction **"fenetre.opt"** nécessite les paramètres suivants : $MatXY$, x, $\mathbf{u}$, 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 $2n^{-1/5}\sigma_n$ oú 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 en (2.27).

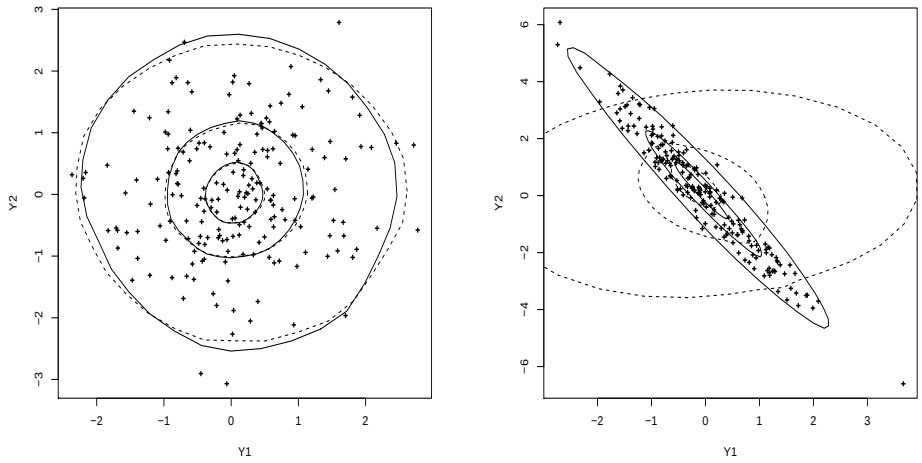
2.6 Etude par simulation

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 $\mathbf{u}$ 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 $\mathbf{u}$ 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 fallait 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 2.3 (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.

2.6.1 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 2.3 (a) représente les trois contours de niveaux 30%, 60% et 90% estimés et le nuage de points correspondant.

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 2.3 (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



(a) (b)

FIG. 2.3 – 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 $\epsilon \sim N(0, 0.5)$.

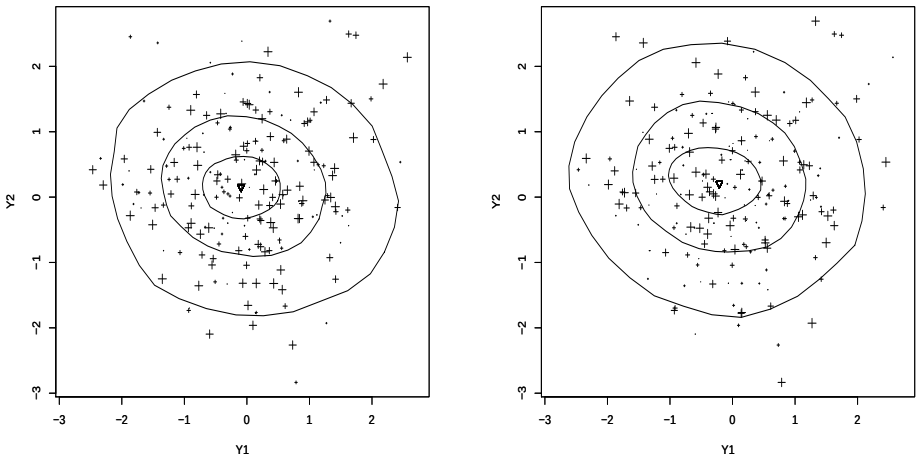
TR (tracés en pointillés) n'ont pas vraiment de sens car ils décrivent mal le support de la distribution dans ce cas. Nous remarquons également sur la Figure 2.3 (a) que dans le cas d'un distribution sphérique l'étape TR n'est pas vraiment nécessaire : en effet les contours calculés avec et sans TR se superposent.

2.6.2 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 2.4 (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 indépedantes. De même, les contours calculés sont très similaires quel que soit le conditionnement.



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

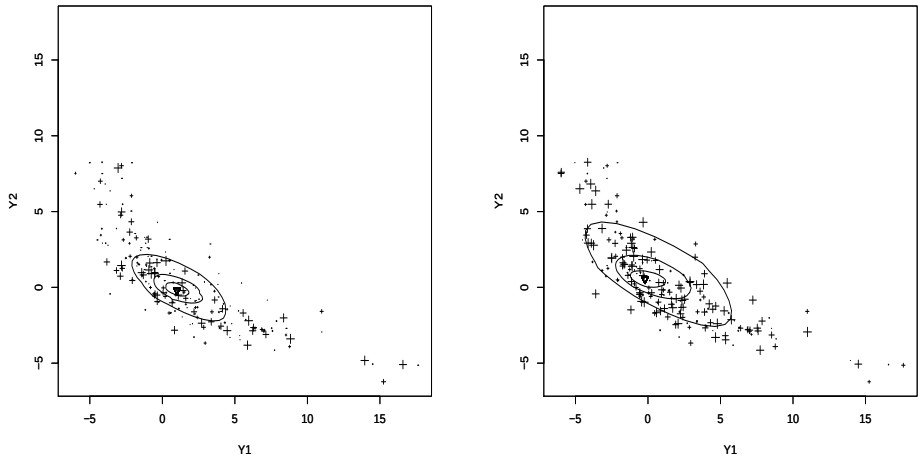
FIG. 2.4 – Tracé de la médiane géométrique conditionnelle (représentée par un triangle) et des contours de niveau 25%, 50% et 75% pour un jeu de données issues du Modèle 1, conditionnellement à $X = -0.5$ et $X = +0.5$

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 2.5 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.

2.7 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



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

FIG. 2.5 – 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 $X = +1$

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 2.6 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).

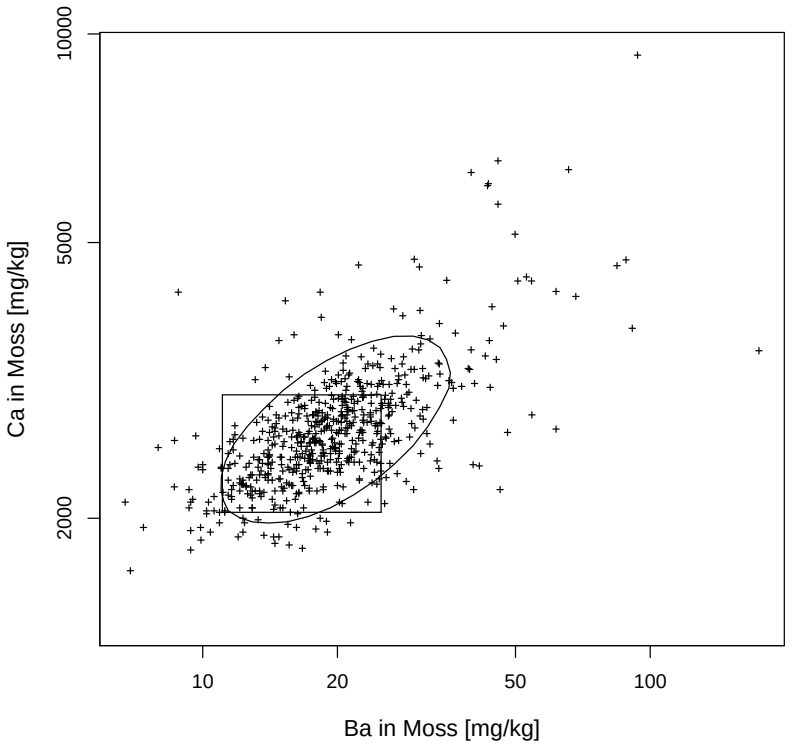


FIG. 2.6 – Comparaison entre les quantiles marginaux et les quantiles géométriques calculés sur les composants chimiques (Ba, Ca) mésurés sur des algues.

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 considérés 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.

Annexe : preuve du Théorème 4.3

- La première partie du théorème se déduit directement de l'équation (2.19). Si les observations $\mathbf{Y}_i$ ne sont pas alignées, le quantile géométrique conditionnel est l'unique solution θ de l'équation (2.19). 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 :

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

$$\sum_{\substack{1 \leq i \leq n \\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) \neq \mathbf{Y}_i}} K_{h_n}(\mathbf{x} - \mathbf{X}_i) < S(\mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u}|\mathbf{x})) + \mathbf{u}, \mathbf{h} > . \tag{2.28}$$

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

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

Ainsi, l'inégalité (2.28) est équivalente à

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

$$\sum_{\substack{1 \leq i \leq n \\ \mathbf{Q}_n(\mathbf{u}|\mathbf{x}) \neq \mathbf{Y}_i}} K_{h_n}(\mathbf{x} - \mathbf{X}_i) < S(\mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u}|\mathbf{x})) + \mathbf{u}, \mathbf{h} > . \tag{2.29}$$

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

$$\mathbf{h} = S(\mathbf{Y}_i - \mathbf{Q}_n(\mathbf{u}|\mathbf{x})) + \mathbf{u}. \tag{2.30}$$

En remplaçant $\mathbf{h}$ par cette valeur dans l'équation (2.29), 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 (2.25).

Références

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

Babu, G. J. and Rao, C. R. (1988). Joint asymptotic distribution of marginal quantile functions in samples from a multivariate population. *Journal of Multivariate Analysis*, **27**, 15-23.

Bahadur, R. R. (1966). A note on quantiles in large samples. *The Annals of Mathematical Statistics*, **37**, 577-580.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Ser. A*, **139**, 318-354.

Bedall, F.K. and Zimmermann, H. (1979). Algorithm AS 143, the Mediancenter. *Applied Statistics*, **28**, 325-328.

Brown, B. M. (1983). Statistical use of the spatial median. *Journal of the Royal Statistical Society, Ser. B*, **45**, 25-30.

Brown, B. M. and Hettmansperger, T. P.(1987). Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Ser. B*, **49**, 301-310.

Brown, B. M. and Hettmansperger, T. P.(1989). An affine invariant bivariate version of the sign test. *Journal of the Royal Statistical Society, Ser. B*, **51**, 117-125.

Chakraborty, B. and Chaudhuri, P. (1996). On a transformation and retransformation technique for constructing an affine equivariant multivariate median. *Proceeding of the American Mathematical Society*, **124**, 2539-2547.

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

Chaudhuri, P. (1992). Multivariate location estimation using extension of R-estimates through U-statistics type approach. *The Annals of Statistics*, **20**, 897-916.

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

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

Cheng, Y. and De Gooijer J. (2007). On the uth geometric conditional quantile. *Journal of Statistical Planning and Inference*, **137**, 1914-1930.

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

De Gooijer, J. G., Gannoun, A. and Zerom, D. (2006). A multivariate quantile predictor. *Communications in Statistics - Theory and Methods*, **35**, 133-147.

Donoho, D. L. and Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, **20**, 1803-1827.

Eddy, W.F. (1982). Convex Hull Peeling. *COMPSTAT 1982 for IASC*, Vienna : Pysica-Verlag, 42-47.

Eddy, W.F. (1985). Ordering of Multivariate Data. *Computer Science and Statistics : The Interface*, ed. L. Billard, Amesterdam : North-Holland, 25-30.

Ferguson, T. (1967). *Mathematical Statistics : A Decision Theoric Approach*. Academic Press, New York.

Gannoun, A., Girard, S., Guinot, C. and Saracco, J. (2002). 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.

Gannoun, A., Saracco, J., Yan, A. and Bonney, G.E. (2003a). On adaptive transformation-retransformation estimate of conditional spatial median. *Communications in Statistics - Theory and Methods*, **32**, 1981-2011.

Gannoun, A., Saracco, J., Yu, K. (2003b). Nonparametric time series prediction by conditional median and quantiles. *Journal of statistical Planning and inference*, **117**, 207-223.

Gower, J.C. (1974). Algorithm AS 78 : The Mediancenter. *Applied Statistics*, **23**, 466-470.

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

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

Koenker, R. and Basset, G. (1978). Regression quantiles. *Econometrica*, **46**, 33-50.

Kokic, P., Breckling, J., Lübke, O. (2002). A new definition of multivariate M-quantiles. Statistical data analysis based on the L_1-norm and related methods. Stat. Ind. Technol. : *Statistical Data Analysis*, 15-24, Birkhäuser Verlag Basel/Switzerland.

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

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

Oja, H. (1983). Descriptive statistics for multivariate trimming. *Statistics and Probability Letters*, **1**, 327-332.

Plackett, R. L. (1976). Comment on the " Ordering of multivariate data", by V. Barnett . *Journal of the Royal Statistical Society, Ser. A*, **139**, 344-346.

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

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

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *J. Statist. Plann. Inference*, **123**, 259-278.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, **28**, 461-482.

Chapitre 3

Design-Based Estimation for Geometric Quantiles*

Abstract : In this paper, we are interested in estimating geometric quantile when data are obtained in a complex survey. The geometric quantile, as it has been introduced by Chaudhuri (1996), may be defined as the unique solution of an implicit estimating equation. This work aims at constructing a design-based estimator of the geometric quantile and computing it by iterative method from survey data. Under broad assumptions, we derive the asymptotic variance of the quantile estimator and propose a consistent estimator of it. Finally, the good behavior of the geometric quantile estimator is verified through a simulation study.

Keywords : Bahadur expansion, consistent estimator, estimating equation, Horvitz-Thompson estimator, variance estimation.

Contents

3.1 Introduction **61**
3.2 Geometric quantile in finite population setting **63**
3.3 Design-based estimator of $Q(u)$ **64**
3.4 Main results **67**
3.5 Simulation Study **70**
3.6 Conclusion and comments **73**
3.7 Appendix **73**

3.1 Introduction

For the last decades, researchers are more and more interested in multivariate location parameters because they play a central role in a wide range of applications. For

*Article écrit en collaboration avec Camelia Goga et soumis à *International Statistical Review*.

instance, Reaven and Miller (1979) examined the relationship between chemical, subclinical and overt nonketotic diabetes in 145 non-obese adult subjects. The variables used in the analysis are glucose intolerance, insuline response to oral glucose and insulin resistance. Chakraborty (2001) used u-th geometric quantiles of two variables at a time, to plot contours which allow to detect some outliers in the data set. De Gooijer and *al.* (2006) give a financial time series analysis application of multivariate quantiles. They apply the multivariate conditional quantiles estimators to predict tails from bivariate time series : Deutsche Mark/US Dollar and Detsche Mark/British Pound.

Unlike the univariate data analysis, the order of observations Y_k lying in $\mathbb{R}^d$ with $d \geq 2$ is not total. As a consequence, several quantile-type multivariate definitions have been formulated. The pioneer papers of Gini and Galvani (1929) and Haldane (1948) considered a multivariate extension of the median defined as an M-estimator. The reader may be referred to Small (1990) and Serfling (2002) for an historical review and comparison. Recently, Chaudhuri (1996) has defined the geometric quantile as an extension of multivariate quantiles based on norm minimization and that uses the geometry of multivariate data clouds. His definition extends to $\mathbb{R}^d$ the well-known characterization of α-univariate quantile q_α, $\alpha \in (0,1)$ (see Ferguson, 1967, p. 51 and Koenker and Basset, 1978)

$$q_\alpha = \arg\min_\theta \mathbb{E}\left(|Y-\theta| + u(Y-\theta)\right)$$

for $u = 2\alpha - 1$ and $\mathbb{E}|Y| < \infty$. Chaudhuri indexed multivariate quantiles by u which belongs to a d-dimensional open unit ball serving for classifying the observations in "extreme" or "central". The vector u corresponds also to a direction in the multivariate data cloud which explains the "geometric aspect" of Chaudhuri's definition. He established asymptotic convergence and distribution theory for the quantile estimator.

In this paper, we consider the more realistic situation when observations are no longer independent and identically distributed. Data are collected then according to a sampling design. Särndal *et al.* (1992, p. 491) affirm that ignoring the sampling design in data analysis may lead to erroneous conclusions. The traditional assumption of independent and identically distributed observations is almost fulfilled for the simple random sampling without replacement with a small sampling fraction. In the contrary case, the traditional statistical procedures can lead to invalid confidence intervals.

Sampling from finite population is characterized by the fact that the statistician may designate in any way he likes a set of samples and the samples may have unequal probability of selection. These two features allow to exist various classes of linear estimators fundamentally different from those dealt with in traditional statistical inference. The Horvitz-Thompson's influential paper (Horvitz and Thompson, 1952) proposes to weight the observations by the inverse of the probability of selection of the corresponding individual.

Univariate quantile estimation with survey data has been studied for example by Kuk, (1988), Francisco and Fuller (1991). These papers consider first the estimation of the finite population distribution function F by the Horvitz-Thomson estimator $\hat{F}$

and next, the estimator of the α-quantile q_α is derived as $\hat{q}_\alpha = \inf_t\{\hat{F}(t) > \alpha\}$ for any $\alpha \in (0,1)$.

Here we focus on geometric quantile estimation with survey data. The geometric quantile proposed by Chaudhuri (1996) is the minimizer of a finite population total function depending on the direction u lying in the unit ball and on the d-dimensional observations. The design-based quantile estimator is obtained by minimizing the Horvitz-Thompson estimator of the objective function. The estimating equation theory (Binder,1983) may be used since our design-based quantile estimator is proved to be the solution of an an implicit sample estimating equation. The resulting estimator is a non-linear function of observations from the sample. In order to calculate and estimate its variance, we linearize the geometric quantile, namely we give a first-order expansion of it. The alternative approach for linearizing the geometric quantile by influence function (Deville, 1999) is also possible.

The paper is structured as follows. Section 3.2 gives the Chaudhuri's geometric quantile in the finite population setting. Section 3.3 presents the derivation of the quantile estimator with survey data as the minimizer of a weighted sample function and gives an iterative method for obtaining it. The estimator obtained in this way is proved to be the unique solution of a sample estimating equation. We introduce in Section 3.4 the asymptotic framework and by similar techniques as in Binder (1983), we give the first-order Bahadur-type expansion of the quantile estimator and derive its asymptotic variance. A variance estimator is proposed and we show that is consistent under broad assumptions. Finally, we present in Section 3.5 a simulation study which confirm the good behavior of the proposed quantile estimator. Some technical details are given in the Appendix.

3.2 Geometric quantile in finite population setting

Let us consider the finite population $U = \{1, \ldots, k, \ldots, N\}$ with size N and a multidimensional vector Y in $\mathbb{R}^d$ with $d \geq 2$. We denote by $Y_1, \ldots, Y_N$ the values taken by Y for each element of U.
We suppose throughout this paper that

($\star$) $\{Y_1, \ldots, Y_N\}$ in $\mathbb{R}^d$ are not all carried by a straight line in $\mathbb{R}^d$.

According to Chaudhuri (1996), the geometric quantile corresponding to a fixed direction u and based on the d-dimensional data $Y_1, \ldots, Y_N$ is defined as follows

$$Q(u) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{k=1}^{N} \phi(u, Y_k - \theta) \quad \text{for} \quad u \in B^d = \{z \in \mathbb{R}^d : ||z|| < 1\}. \tag{3.1}$$

The *multivariate loss function* $\phi : \mathbb{R}^d \times B^d \to \mathbb{R}$ is given by $\phi(u,t) = ||t|| + \langle u, t\rangle$ with $||\cdot||$ the usual Euclidean norm and $<\cdot,\cdot>$ the usual Euclidean inner product. The u-th quantile $Q(u)$ is indexed by a directional "outlyingness" parameter u. The spatial median is obtained for $u = 0$ and $Q(0)$ is also called the center of the data cloud formed

by Y_k's. On the other hand, for $u \neq 0$, Chaudhuri (1996) interprets $\|u\|$ as an "extent of deviation" of $Q(u)$ from the center : the geometric quantile is called *"central"* if $||u||$ closes to 0 and *"extrem"* when $||u||$ closes to 1 (see Chaudhuri, 1996 and Serfling, 2002 for more details).

Chaudhuri (1996) proves also that the minimizer $Q(u)$ exists since the objective function $\sum_{k=1}^{N} \phi(u, Y_k - \theta)$ is continuos with respect of θ and it explodes to infinity when $\|\theta\| \to \infty$. The uniqueness is guaranteed by the fact that the objective function is a strictly convex function of θ (assumption $(\star)$ and theorem 2.17 of Kemperman (1987)).

Result 1 *(Chaudhuri, 1996, theorem 2.1.2) Suppose now that the u-th geometric quantile $Q(u)$ computed from the data set $Y_1, \ldots, Y_N$ is different from all Y_k, $k = 1, \ldots, N$. Then, $Q(u)$ is the unique solution of the following equation*

$$\sum_{k=1}^{N} [S(Y_k - \theta) + u] = 0 \tag{3.2}$$

where S is defined as $S(v) = v/||v||$ for any non null vector $v \in \mathbb{R}^d$.

This result is viewed as a robustness property by Serfling (2002) since the value of $Q(u)$ remains unchanged if the points Y_k are moved outward along the rays joining them with $Q(u)$. It is also of a great practical importance since one can use iterative methods like "Newton-Raphson-type method" (see Bedall and Zimmermann, 1979) in order to obtain $Q(u)$. Chaudhuri (1996) gives such an iterative method and an alternative of it for the case when $Q(u)$ is close to some data points.

Let us denote by $h_k(\theta) = S(Y_k - \theta) + u$, for all $k \in U$ and by $H_U(\theta)$ the finite population total of $h_k(\theta)$, namely $H_U(\theta) = \sum_{k=1}^{N} h_k(\theta)$. Result 1 tells us that the u-th geometric quantile $Q(u)$ verifies then the equation

$$H_U(Q(u)) = \sum_{k=1}^{N} h_k(Q(u)) = 0. \tag{3.3}$$

This means that our parameter of interest is the unique solution of a finite population estimation equation (Binder, 1983).

3.3 Design-based estimator of $Q(u)$

We consider now a sample of individuals s, *i.e.* a subset $s \subset U$, selected according to a sampling design $p(s)$ that assigns probabilities to the elements s of the set of all samples composed from the population U. We denote by $\pi_k = \Pr(k \in s)$ for all $k \in U$ the first order inclusion probabilities and by $\pi_{kl} = \Pr(k \,\&\, l \in s)$ for all $k, l \in U$ with $\pi_{kk} = \pi_k$, the second order inclusion probabilities. We suppose that $\pi_k > 0$ for all $k \in U$ and $\pi_{kl} > 0$ for all $k \neq l \in U$.
In order to obtain a sample estimate $\widehat{Q}(u)$ of the u-th geometric quantile $Q(u)$, we suppose in addition that

($\star\star$) Y_k in $\mathbb{R}^d$ and $k \in s$ are not all carried by a straight line in $\mathbb{R}^d$.

Let us consider, for any fixed $\theta \in \mathbb{R}^d$ and $u \in B^d$, the Horvitz-Thompson (HT) estimator of the objective function from (3.1), namely

$$\sum_{k \in s} \frac{\phi\left(u, Y_k - \theta\right)}{\pi_k} = \sum_{k \in U} \frac{\phi\left(u, Y_k - \theta\right)}{\pi_k} I_k \tag{3.4}$$

where $I_k = \mathbb{1}_{\{k \in s\}}$ is the sample membership indicator of element k (see Särndal *et al.*, 1992). Note that the variables I_k are random with $Pr(I_k = 1) = \pi_k$ and that Y_k are considered fixed with respect to the sampling design $p(\cdot)$.

We propose to minimize the above estimated objective function in order to obtain the u-th geometric quantile estimate $\widehat{Q}(u)$ based on the sample of d-dimensional data Y_k, $k \in s$:

$$\widehat{Q}\left(u\right) = \arg\min_{\theta \in \mathbb{R}^d} \sum_{k \in s} \frac{\phi\left(u, Y_k - \theta\right)}{\pi_k}. \tag{3.5}$$

The existence and uniqueness of $\widehat{Q}(u)$ may be proved with the same arguments given in Section 2 provided that assumption ($\star\star$) is fulfilled.

The derivation of $\widehat{Q}(u)$ in (3.5) gives a *design-based estimator* (Särndal *et al.*, 1992) and may surprise the statistician who is not familiar with survey sampling theory and who would have rather minimized the un-weighted objective function $\sum_{k \in s} \phi\left(u, Y_k - \theta\right)$. A similar situation was met in the case of regression analysis with survey data. The regression coefficient was the subject of discussion and Särndal (1980) and Särndal *et al.* (1992, p. 519) make a comparison between the two possible estimators pleading in favor of the weighted estimator.

Based on such arguments, we consider in the next only the design-based estimator (3.5) and we derive its design-based properties. In order to do that, we need first to construct the HT estimator of $H_U(\theta)$ and give its properties.

For any vector $\theta \in \mathbb{R}^d$, the HT estimator of $H_U(\theta)$ is the vector $\widehat{H}(\theta) \in \mathbb{R}^d$ defined as follows

$$\widehat{H}(\theta) = \sum_{k \in s} \frac{h_k(\theta)}{\pi_k} = \sum_{k \in U} \frac{h_k(\theta)}{\pi_k} I_k. \tag{3.6}$$

The estimator $\widehat{H}(\theta)$ is p-unbiased for $H_U(\theta)$, namely $\mathbb{E}_p(\widehat{H}(\theta)) = H_U(\theta)$ where $\mathbb{E}_p(\cdot)$ is the expectation with respect to the sampling design. One can derive easily the HT variance of $\widehat{H}(\theta)$ (see Särndal *et al.* 1992, p. 170) as being the variance-covariance matrix $\mathbb{V}_p\left(\widehat{H}(\theta)\right)$ given by

$$\mathbb{V}_p\left(\widehat{H}(\theta)\right) = \sum_U \sum_U \Delta_{k\ell} \cdot \frac{h_k(\theta)}{\pi_k} \cdot \frac{h_l^T(\theta)}{\pi_\ell} \quad , \quad \Delta_{k\ell} = \pi_{k\ell} - \pi_k \pi_\ell \tag{3.7}$$

which is estimated unbiasedly by the following matrix

$$\widehat{\mathbb{V}}_p\left(\widehat{H}(\theta)\right) = \sum_s \sum_s \frac{\Delta_{k\ell}}{\pi_{k\ell}} \cdot \frac{h_k(\theta)}{\pi_k} \cdot \frac{h_l^T(\theta)}{\pi_\ell}. \tag{3.8}$$

We may give now a characterization of the design-based geometric quantile in terms of data points from which it is computed.

Result 2 *Let $\widehat{Q}(u)$ be the estimator of $Q(u)$ computed from a sample s.*
- If $\widehat{Q}(u) \neq Y_k$ for all $k \in s$, then

$$\widehat{H}(\widehat{Q}(u)) = \sum_{k\in s} \frac{h_k(\widehat{Q}(u))}{\pi_k} = 0 \quad \textit{for all} \quad u \in B^d. \tag{3.9}$$

- If $\widehat{Q}(u) = Y_k$ for some $k \in s$, then $\| \sum_{\substack{k\in s \\ Y_k \neq \widehat{Q}(u)}} \frac{h_k(\widehat{Q}(u))}{\pi_k} \| \leq (1 + \|u\|) \sum_{\substack{k\in s \\ Y_k = \widehat{Q}(u)}} \frac{1}{\pi_k}$

Proof Because of the convexity of $\phi(u, Y - \theta)$ and that $\sum_{k\in s} \pi_k^{-1} \phi(u, Y_k - \theta)$ is minimum for $\theta = \widehat{Q}(u)$, we get, for all $z \in \mathbb{R}^d$,

$$\lim_{t\to 0} \sum_{k\in s} \frac{\phi(u, Y_k - \widehat{Q}(u) + tz) - \phi(u, Y_k - \widehat{Q}(u))}{\pi_k t} = \sum_{\substack{k\in s \\ Y_k = \widehat{Q}(u)}} \frac{1}{\pi_k} [\|z\| + \langle u, z\rangle] + \sum_{\substack{k\in s \\ Y_k \neq \widehat{Q}(u)}} \frac{1}{\pi_k} \langle h_k(\widehat{Q}(u)), z\rangle \geq$$

0
Then, arguing along the same lines as in the proof of theorem 2.1.2 of Chaudhuri (1996), we conclude the proof of this result.

□

Relation (3.9) from above means that $\widehat{Q}(u)$ is obtained as the unique solution of the sample estimating equation $\sum_{k\in s} \frac{h_k(\theta)}{\pi_k} = 0$ (Binder, 1983). Moreover, suppose that Y is generated from a distribution ξ belonging to a superpopulation model, then Godambe and Thompson (1986) prove that the estimator $\widehat{Q}$ defined in this way is optimal in the sense that its joint mean squared error with respect to the sampling design and the superpopulation model is minimal. Nevertheless, we do not consider here this framework.

3.3.1 Computation of $\widehat{Q}(u)$

In order to compute $\widehat{Q}(u)$, we use result 2 and propose below an algorithm which is in fact the adaption of Chaudhuri's algorithm to the survey sampling framework. This algorithm contains two steps.
Step1. For each element $\ell \in s$ one checks whether or not the following condition is satisfied

$$\| \sum_{\substack{k\in s \\ k\neq \ell}} \frac{h_k(Y_\ell)}{\pi_k} \| \leq \frac{1 + \|u\|}{\pi_l}.$$

If this condition is satistied for some $\ell \in s$, then $\widehat{Q}(u) = Y_\ell$. Otherwise, $\widehat{Q}(u)$ will be the unique solution of (3.9). The second step consists in solving equation (3.9) by iterative procedure.

Step 2. We start with $\widehat{Q}^{(1)}(u)$, the vector of medians of each real-valued components of Y computed on s. Let $\widehat{Q}^{(1)}(u), \ldots, \widehat{Q}^{(m)}(u)$ be a sequence of approximations of $\widehat{Q}(u)$. Then $\widehat{Q}^{(m+1)}(u)$ can be obtained recursively using the Newton-Raphson procedure

$$\widehat{Q}^{(m+1)}(u) = \widehat{Q}^{(m)}(u) + \left[\widehat{J}(\widehat{Q}^{(m)}(u))\right]^{-1} \widehat{H}(\widehat{Q}^{(m)}(u))$$

where $\widehat{J}(\widehat{Q}^{(m)}(u))$ is the HT estimator of the Jacobian of H_U calculated at $\widehat{Q}^{(m)}(u)$ (see relation (3.12) below). By $(\star\star)$, the matrix $\widehat{J}(\widehat{Q}^{(m)}(u))$ is a positive definite matrix. Iteration is continued until two successive approximations of $\widehat{Q}(u)$ happen to be sufficiently close. In practice, this algorithm converge at most after 10 iterations.
The u-th geometric quantile $Q(u)$ is a nonlinear parameter defined implicitly in (3.3). So, the variance of $\widehat{Q}(u)$ as well as its estimator can not be derived by using the HT formulae (3.7) and (3.8). Under broad assumptions, we deduce in the next section an asymptotic variance of $\widehat{Q}(u)$ using the Taylor linearization method.

3.4 Main results

We suppose in the next that the direction u is fixed and we drop it for sake of simplicity from the notations of $Q(u)$ and $\hat{Q}(u)$. Let denote by Θ the parameter space and suppose that the true quantile Q is an inner point of Θ. Let be $\mathcal{V}_Q$ a neighborhood of Q.

3.4.1 Asymptotic framework

In order to obtain the asymptotic properties of $\widehat{Q}$, we consider the conceptual framework of Isaki and Fuller (1982) for having infinite population and sample. Consider also the following assumptions :

(A1) $\lim_{N\to\infty} \dfrac{n}{N} = \pi \in (0,1)$,

(A2) $\min\limits_{k} \pi_k \geq \lambda_1$, $\min\limits_{k\neq l} \pi_{kl} \geq \lambda_2$ for λ_1, λ_2 positive constants and $\overline{\lim}_{N\to\infty} n \max\limits_{k\neq l} |\pi_{kl} - \pi_k\pi_l| < \infty$.

(A3) We suppose that exist a positive constant M such that $||Y_k - \theta|| \geq M$ for any $k \in U$ and $\theta \in \mathcal{V}_Q$.

(A4) The estimator $\widehat{Q}$ is consistent for Q, namely for any fixed $\varepsilon > 0$ we have $\lim_{N\to\infty} \mathbb{P}\left(||\widehat{Q} - Q|| > \varepsilon\right) = 0$.

(A5) $\dfrac{\sqrt{n}}{N}\left[\widehat{H}(Q) - H_N(Q)\right] \longrightarrow N(0,\Sigma)$ for Σ a positive definite matrix.

Assumptions (A1) and (A2) are rather classical in survey theory. The first assumption (Breidt and Opsomer, 2000) means that when n and N go to infinity, we always have $n < N$. Assumption (A2) was also used by Breidt and Opsomer (2000) and is satisfied for several sampling designs. Assumptions similar were used by Robinson and Särndal (1983). The third assumption is needed for technical reasons in results 3 and 4 and it gives in particular the derivability of the objective function. Assumption (A4) assures that the reminder term from the first-order Taylor expansion of $\widehat{H}$ is negligible

(result 3 below). Finally, (A5) is the central limit theorem for the HT estimator $\widehat{H}$ justified by the need of the normal approximation for deriving confidence intervals based on $\widehat{Q}$.
Let us denote by $J_U(\theta)$ the Jacobian matrix of $H_U(\theta)$ and given by

$$J_U(\theta) = \sum_U \frac{1}{||Y_k - \theta||} \left[I_d - S(Y_k - \theta) \cdot S^T(Y_k - \theta) \right].$$

Here, I_d is the $d \times d$ identity matrix. We consider also the Jacobian matrix of $\widehat{H}(\theta)$ denoted by $\hat{J}(\theta)$. One can remark that $\hat{J}(\theta)$ is the HT estimator of $J_U(\theta)$ and is given by

$$\widehat{J}(\theta) = \sum_s \frac{1}{\pi_k ||Y_k - \theta||} \left[I_d - S(Y_k - \theta) \cdot S^T(Y_k - \theta) \right].$$

Both $J_U(\theta)$ and $\widehat{J}(\theta)$ are $d \times d$ symmetric matrix and positively-defined unless Y_k's are carried by a straight line in R^d.

Lemma 3.4.1 *Let the assumptions (A1) and (A2) hold. Then, the HT estimator $\widehat{H}(\theta)$ defined by (3.6) satisfies $\mathbb{E}_p \left|\left| \frac{1}{N} \left(\widehat{H}(\theta) - H_N(\theta) \right) \right|\right| = O(n^{-1/2})$ for any $\theta \in \mathcal{V}_Q$.*

Proof See the Appendix. □

Lemma 3.4.2 *Let the assumptions (A1)-(A3) hold. We have, for any $\theta \in \mathcal{V}_Q$,*
(i) $\frac{1}{N} J_U(\theta) = O(1)$,
(ii) $\mathbb{E}_p \left|\left| \frac{1}{N} \left(\widehat{J}(\theta) - J_U(\theta) \right) \right|\right|_1 = O(n^{-1/2})$ *for* $||\cdot||_1$ *the trace norm given by* $||A||_1^2 = tr(A^T \cdot A)$ *for any matrix* A.

Proof See the Appendix. □

3.4.2 Asymptotic variance

Result 3 *Let the assumptions (A1)-(A5) hold. The first-order Bahadur-type expansion of $\widehat{Q}$ is given as follows :*

$$\begin{aligned} \widehat{Q} - Q &= -J_U^{-1}(Q) \left(\widehat{H}(Q) - H_U(Q) \right) + o_p(n^{-1/2}) \\ &= \sum_s \frac{u_k}{\pi_k} + o_p(n^{-1/2}) \end{aligned}$$

for the linearized variable of Q, $u_k = -J_U^{-1}(Q) \cdot h_k(Q)$ *with* $\sum_U u_k = 0$. *As a consequence, the asymptotic variance of* $\widehat{Q}$ *denoted by* $\mathbb{AV}_p(\widehat{Q})$ *is equal to the variance of the HT estimator* $\sum_s \frac{u_k}{\pi_k}$, *namely*

$$\mathbb{AV}_p(\widehat{Q}) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l^T}{\pi_l}.$$

Proof Consider the first-order Taylor expansion of the θ-differentiable HT estimator $\widehat{H}(\theta)$ at the point $\theta = \widehat{Q}$ and arround the true u-th geometric quantile Q,

$$\widehat{H}(\widehat{Q}) = \widehat{H}(Q) + \widehat{J}(Q)\left(\widehat{Q} - Q\right) + o\left(||\widehat{Q} - Q||\right).$$

Lemma (3.4.2) and assumption (A4) give $\frac{1}{N}\left(\widehat{J}(Q) - J_U(Q)\right)\left(\widehat{Q} - Q\right) = o_p(n^{-1/2})$. We obtain then using $\widehat{H}(\widehat{Q}) = H_U(Q) = 0$ that

$$\frac{\sqrt{n}}{N}\left(H_U(Q) - \widehat{H}(Q)\right) = \frac{\sqrt{n}}{N} J_U(Q) \cdot \left(\widehat{Q} - Q\right) + o_p(1) \tag{3.10}$$

which implies that $\frac{\sqrt{n}}{N} J_U(Q)(\widehat{Q} - Q)$ has the same asymptotic distribution as the first term from the left-side. Moreover, the matrix $\frac{1}{N} J_U(Q)$ is of full rank and bounded in probability by lemma (3.4.2). We obtain from equation (3.10) that

$$\begin{aligned} \widehat{Q} - Q &= -\left(\frac{1}{N} J_U(Q)\right)^{-1} \frac{1}{N}\left(\widehat{H}(Q) - H_U(Q)\right) + o_p(n^{-1/2}) \qquad (3.11) \\ &= -J_U^{-1}(Q)\left(\sum_s \frac{h_k(Q)}{\pi_k} - \sum_U h_k(Q)\right) + o_p(n^{-1/2}) \\ &= \sum_s \frac{u_k}{\pi_k} + o_p(n^{-1/2}) \end{aligned}$$

for $u_k = -J_U^{-1}(Q) h_k(Q)$, $k \in U$ *the linearized variable of* Q with $\sum_U u_k = 0$. □

3.4.3 Variance estimation

As one can notice, the asymptotic variance given by result 3 is unknown since the double sums are considered on the whole population U and the value of Q is unknown. We construct the estimator of $J_U(Q)$ as being the matrix $\widehat{J}(\theta)$ in $\theta = \widehat{Q}$,

$$\widehat{J}(\widehat{Q}) = \sum_{k \in s} \frac{1}{||Y_k - \widehat{Q}||\pi_k}\left[I_d - S(Y_k - \widehat{Q})S^T(Y_k - \widehat{Q})\right] \tag{3.12}$$

and estimate the variance of $\widehat{Q}$ by

$$\widehat{\mathbb{V}}_p(\widehat{Q}) \;=\; \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{\widehat{u}_k}{\pi_k} \frac{\widehat{u}_l^T}{\pi_l} = \left[\widehat{J}(\widehat{Q})\right]^{-1} \widehat{V}_p(\hat{H}(\widehat{Q})) \left[\widehat{J}(\widehat{Q})^T\right]^{-1}$$

for $\widehat{u}_k = -\widehat{J}^{-1}(\widehat{Q}) h_k(\widehat{Q})$ and $\widehat{\mathbb{V}}_p(\widehat{H}(\widehat{Q}))$ obtained from (3.8) for $\theta = \widehat{Q}$.
We intend to show that the estimator $\widehat{\mathbb{V}}_p(\widehat{Q})$ is consistent for $\mathbb{AV}_p(\widehat{Q})$. We suppose that

(A6) $\frac{1}{N^2}\left[\widehat{V}_p(\widehat{H}(Q)) - \Sigma\right] = o_p(n^{-1})$.

Result 4 *Let the assumptions (A1)-(A6) hold. It results then*

$$\widehat{V}_p(\widehat{Q}) - \mathbb{AV}_p(\widehat{Q}) = o_p(n^{-1}).$$

Proof The proof is given in the Appendix. □

3.5 Simulation Study

In this section we verify the good behavior of the geometric quantile estimator for two sampling designs : the Simple Random Sampling Without Replacement (SRSWR) and the stratified sampling with SRSWR within each strata (STRAT). In order to make easier the realization and the interpretation of the graphics, we suppose that $d = 2$ (two-dimensional case).
We consider a variable Y following a bivariate normal distribution $\mathcal{N}_2((0,0), I_2)$ and we take $N = 5000$ replications of it. Let us denote by $Y_i = (Y_i^1, Y_i^2)^T$ for each $i = 1, \ldots, 5000$. We construct two strata U_1 and U_2 of different variances by multiplying the first $N_1 = 1500$ replications of Y by $\sigma_1 = 2$ and the other $N_2 = 3500$ replications by $\sigma_2 = 4$. Our population of study U is the union of these two strata U_1 and U_2 and it will be used to provide the following graphics in this section.

We consider a sequence of vectors $\{u_{ij}\} = \{(r_i \cos(\theta_j), r_i \sin(\theta_j))^T\}$ in the unit ball, with r_i taking the values $r_1 = 0.3$ and $r_2 = 0.9$ and θ taking values in $\{\theta_j = \frac{\pi j}{16}, j = 0, 1, \ldots, 31\}$. For each fixed direction u_{ij}, we compute the population geometric quantile $Q(u_{ij})$ and the estimations of it obtained from samples of size $n = 1000$ selected from U according to SRSWR and STRAT. The stratified sample is built by drawing independently two SRSWR of sizes n_1 in strata U_1, respectively of size $n_2 = n - n_1$ in strata U_2. The sample sizes n_1 and n_2 are chosen according to the optimal allocation.
The set $\mathcal{C}(r_i) = \{Q_n(u_{ij}) : ||u_{ij}|| = r_i\}$, with $0 < r_i < 1$, is named " the quantile contour plot". In general, this set can be viewed as a multivariate analogue of box and whisker plots used for univariate data. It is also useful for detecting the outliers in multivariate data : we compute the quantile contour for some r_i close to 1, and if a particular observation lies outside this contour, then we will call it an outlier. The choice of r_i depends on the problem and the user's preference.
Figure 1. gives the quantile contour plot for $r_1 = 0.3$ and $r_2 = 0.9$:

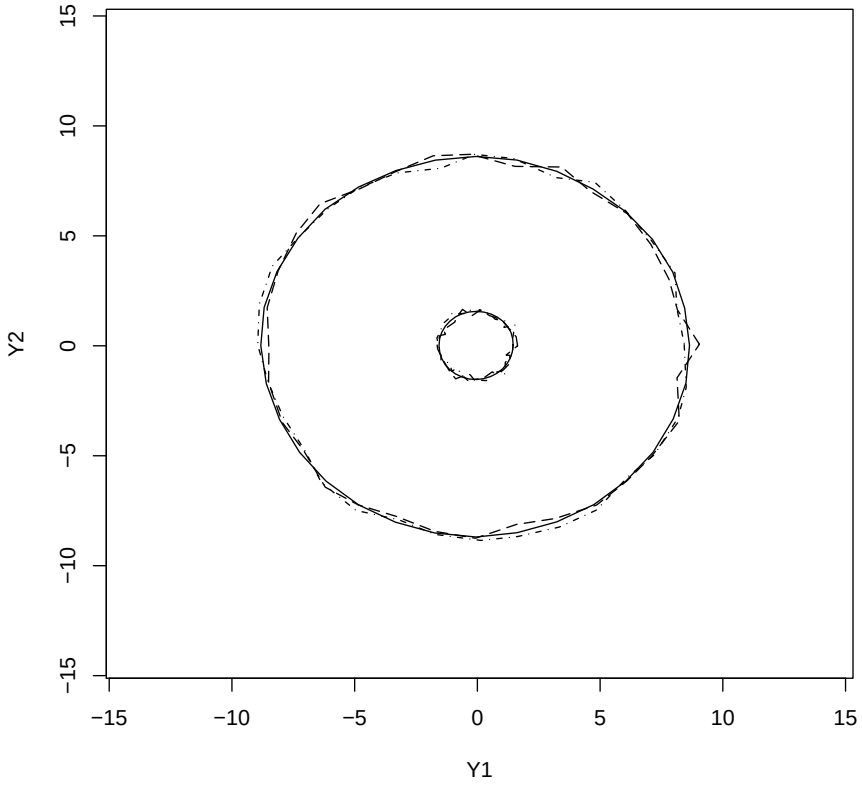
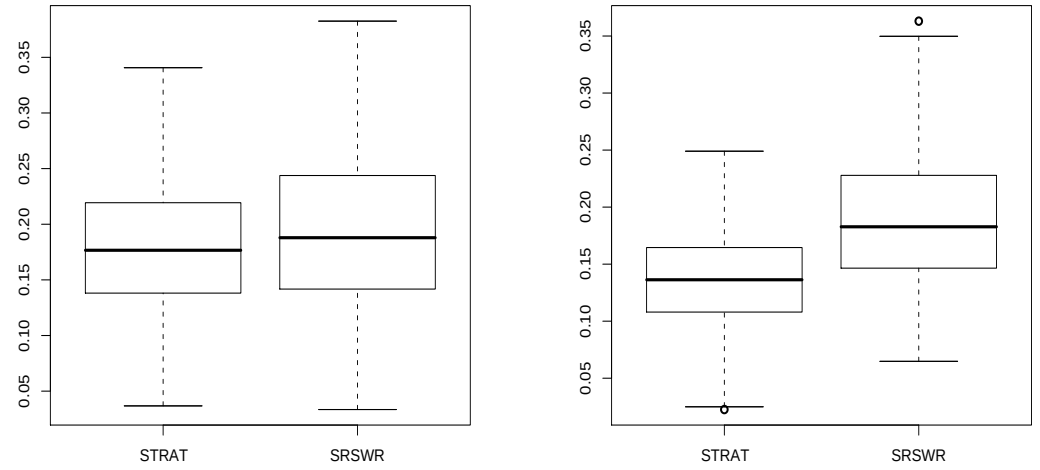


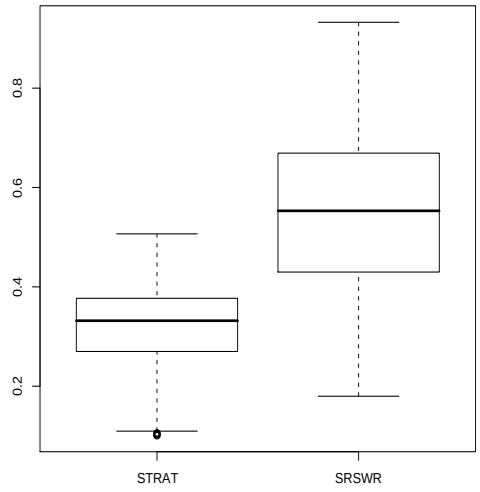
FIG. 3.1 – Quantile contour plots for $r = 0.3$ (inner contours) and $r = 0.9$ (outer contours). The u-th geometric quantile calculated from the population U are represented by solid line and its estimation with SRSWR strategie (resp. stratified strategie) in dotdash line (resp. longdash line).

In Figure 1. we have three inner contours (corresponding to $r_1 = 0.3$) and three outer contours (for $r_2 = 0.9$). For each value of r we represent the geometric quantile calculated from the population U (solid line) and its estimators with SRSWR strategie (dotdash line) and with stratified one (longdash line). We remark that both SRSWR and STRAT seem to give good estimations of the geometric quantile for low and high values of r.

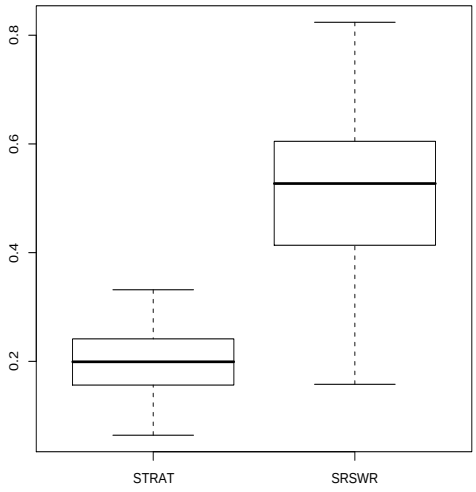


(a)

(b)



(c)



(d)

FIG. 3.2 – Relative estimation errors of geometric quantile for two sampling strategies (SRSWR and stratified sampling). Figure (a) (resp. (b)) is for $u = (0.1, 0.3)^T$ (resp. $u = (0.1, 0.8)^T$) and $n = 1000$. Figure (c) (resp. (d)) is for $u = (0.1, 0.3)^T$ (resp. $u = (0.1, 0.8)^T$) and $n = 500$.

Despite this, we can not conclude from Figure 1. the consistence of the estimators given by SRSWR and stratified stratigies. In order to compare the two procedures, we draw 500 samples of sizes $n = 500$ and $n = 1000$ accordig to each survey sampling strategy, namely SRSWR and stratified design. From each sample $j = 1, \ldots, 500$, we compute the estimator $\widehat{Q}^{(j)}(u)$ of $Q(u)$. Estimation errors for the u-geometric quantile are evaluated by considering the following loss criteria $\frac{||Q(u)-\widehat{Q}^{(j)}(u)||}{||Q(u)||}$ (Euclidean norm) among our 500 replications of the experiments. Figure 2. (a) (resp. (b)) represents the boxplot of relative estimation errors of the u-th geometric quantile for SRSWR and stratified sampling for $u = (0.1, 0.3)^T$ (resp. $u = (0.1, 0.8)^T$) with the sample size $n = 1000$. Figure 2. (c) and (d) represent the same thing but for $n = 500$. It appears that SRSWR and stratified sampling designs give a consistent estimations for $n = 1000$ with slightly better results for stratified one. In fact the median relative error is lower than 18 % (see Figure 2. (a) and (b)). But, from Figure 2. (c) and (d), we deduce that stratified sampling strategy gives much better estimations than SRSWR for n=500. In fact the median relative error is lower than 30 %. This confirm that stratified sampling gives better results for small sizes samples than SRSWR.

3.6 Conclusion and comments

In this paper we have proposed an estimator of geometric quantiles when data are obtained by survey sampling techniques. A Bahadur-type expansion of this estimator has been obtainded in order to get the asymptotic distribution and variance of $\widehat{Q}(u)$. Further, a consistent estimator of the asymptotic variance has been proposed. We give different simulations for a variable with spherical-symmetric distribution and show that stratified strategy gives better estimations for the geometric quantile than SRSWR. The well-known disadvantage of geometric quantiles is that they are not equivariant under arbitrary affine transformations though they are equivariant under rotations of the data cloud. As a consequence, geometric quantiles do not lead to any sensible estimate when the different coordinate variables of the data-vectors are measured in different units or they have different degrees of statistical variations. To overcome this problem, an estimation of geometric quantile by Transformation-Retransformation (TR) technique is required (see Chakraborty, 2001). An extension of the geometric quantile estimation by TR technique which takes into account the survey sampling design is under investigations.

3.7 Appendix

We introduce some new notations.
Let be $S_k(\theta) = S(Y_k - \theta)$ for $k \in U$ and $J_{U,k}(\theta) = \frac{1}{||Y_k - \theta||}\left[I_d - S_k(\theta) \cdot S_k^T(\theta)\right]$. Let us denote by $\beta_k = \frac{I_k}{\pi_k} - 1$.

Proof of lemma 3.4.1.
We note that by assumption $(A2)$, we have $\mathbb{E}_p(\beta_k^2) = (1 - \pi_k)/\pi_k < (1 - \lambda)/\lambda$ and for $k \neq l$,

$|\mathbb{E}_p(\beta_k\beta_\ell)| = |\Delta_{k\ell}/(\pi_k\pi_\ell)| \leq \max_{k\neq l}|\Delta_{k\ell}|/\lambda^2$. Then we get, for any $\theta \in \mathcal{V}_Q$,

$$\mathbb{E}_p\left|\left|\frac{1}{N}\left(\widehat{H}(\theta) - H_U(\theta)\right)\right|\right|^2 = \frac{1}{N^2}\mathbb{E}_p\left|\left|\sum_{k\in U}\beta_k h_k(\theta)\right|\right|^2 \leq \frac{4}{N^2}\sum_{k,\ell\in U}|\mathbb{E}_p(\beta_\ell\beta_k)|$$

since $||h_k(\theta)|| \leq 2$ for all $k \in U$. Let us interest now to $\frac{1}{N^2}\sum_{k,\ell\in U}|\mathbb{E}_p(\beta_\ell\beta_k)|$. We have

$$\begin{aligned}\frac{1}{N^2}\sum_{k,\ell\in U}|\mathbb{E}_p(\beta_\ell\beta_k)| &= \frac{1}{N^2}\sum_{k\in U}\frac{(1-\pi_k)}{\pi_k} + \sum_{k\in U}\sum_{\ell\neq k}\frac{|\Delta_{k\ell}|}{\pi_k\pi_\ell} \leq \frac{1-\lambda}{N\lambda} + \frac{N(N-1)}{N^2 n}\frac{n\max_{k\neq\ell}|\Delta_{k\ell}|}{\lambda^2} \\ &= O\left(\frac{1}{n}\right)\end{aligned}$$

by assumptions $(A1)$ and $(A2)$. This completes the proof of lemma (3.4.1). □

Proof of lemma 3.4.2. *(i)* We have $\frac{1}{N}J_U(\theta) = \frac{1}{N}\sum_{k\in U}J_{U,k}(\theta)$ and

$$\left|\left|\frac{1}{N}J_U(\theta)\right|\right|_1 \leq \frac{1}{N}\sum_{k\in U}||J_{U,k}(\theta)||_1 \leq \frac{1}{N}\sum_{k\in U}\frac{1}{||Y_k-\theta||}(||I_d||_1 + \left|\left|S_k(\theta)\cdot S_k^T(\theta)\right|\right|_1) \leq \frac{\sqrt{d}+1}{M}$$

by assumption (A3) and the fact that $||I_d||_1 = \sqrt{d}$ and $\left|\left|S_k(\theta)\cdot S_k^T(\theta)\right|\right|_1 = 1$.
(ii) We may write $\frac{1}{N}\left(\widehat{J}(\theta) - J_U(\theta)\right) = \frac{1}{N}\sum_{k\in U}\frac{1}{||Y_k-\theta||}J_{U,k}(\theta)\beta_k$ and

$$\left|\left|\frac{1}{N}\left(\widehat{J}(\theta) - J_U(\theta)\right)\right|\right|_1^2 = \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}tr(J_{U,k}(\theta)J_{U,l}(\theta))\beta_k\beta_l \quad \text{with}$$

$$\begin{aligned}& tr(J_{U,k}(\theta)J_{U,l}(\theta)) \\ = \;& \frac{1}{||Y_k-\theta||\cdot||Y_l-\theta||}\left[trI_d - tr(S_k(\theta)S_k^T(\theta)) - tr(S_l(\theta)S_l^T(\theta)) + tr(S_k(\theta)S_k^T(\theta)S_l(\theta)S_l^T(\theta))\right] \\ \leq \;& \frac{1}{M^2}(d-1)\end{aligned}$$

by assumption (A3) and the fact that $tr(S_k(\theta)S_k^T(\theta)) = 1$ and $tr(S_k(\theta)S_k^T(\theta)S_l(\theta)S_l^T(\theta)) \leq 1$. As a consequence,

$$\mathbb{E}_p\left|\left|\frac{1}{N}\left(\widehat{J}(\theta) - J_U(\theta)\right)\right|\right|_1^2 \leq \frac{(d-1)}{N^2M^2}\sum_{k\in U}\sum_{l\in U}\mathbb{E}_p(\beta_k\beta_l) = O\left(\frac{1}{n}\right)$$

by assumptions (A1) and (A2). □

In order to obtain result (4), we give first the following lemma.

Lemma 3.7.1 *1. For all $k \in U$, we have that $h_k(Q) = O(1)$ and $h_k(\widehat{Q}) = O(1)$ hold uniformly in k. Under assumptions (A3) and (A4), we have that $h_k(Q) - h_k(\widehat{Q}) = o_p(1)$ holds uniformly in k.*

2. *Let assumptions (A1)-(A4) hold. It results than* $\frac{1}{N}(\widehat{J}(\widehat{Q}) - \widehat{J}(Q)) = o_p(1)$. *As a consequence,* $\widehat{J}^{-1}(\widehat{Q}) - \widehat{J}^{-1}(Q) = o_p(N^{-1})$.
3. *Let assumptions (A1)-(A4) hold. For all* $k \in U$, *the linearized variable* $u_k = -J_U^{-1}(Q)h_k(Q)$ *of* Q *satisfies* $u_k = O(N^{-1})$ *and* $\hat{u}_k - u_k = o_p(n^{-1})$ *uniformly in* k.

Proof

1. We have obviously that $||h_k(Q)|| \leq 2$ and $||h_k(\widehat{Q})|| \leq 2$. We have

$$h_k(Q) - h_k(\widehat{Q}) = \frac{\widehat{Q} - Q}{||Y_k - Q||} + \frac{(Y_k - \widehat{Q})(||Y_k - \widehat{Q}|| - ||Y_k - Q||)}{||Y_k - Q|| \cdot ||Y_k - \widehat{Q}||}$$

which gives

$$||h_k(Q) - h_k(\widehat{Q})|| \leq \frac{2||\widehat{Q} - Q||}{||Y_k - Q||} = o_p(1) \tag{3.13}$$

since $\Big|||Y_k - \widehat{Q}|| - ||Y_k - Q||\Big| \leq ||\widehat{Q} - Q||$ and using assumptions (A3) and (A4).

2. We have $\frac{1}{N}(\widehat{J}(\widehat{Q}) - \widehat{J}(Q)) = J_1 + J_2$ with

$$J_1 = \frac{1}{N}\sum_s \frac{I_d}{\pi_k}\left[\frac{1}{||Y_k - \widehat{Q}||} - \frac{1}{||Y_k - Q||}\right] \quad \text{and}$$

$$J_2 = \frac{1}{N}\sum_s \frac{1}{\pi_k}\left[\frac{S_k(\widehat{Q}) \cdot S_k^T(\widehat{Q})}{||Y_k - \widehat{Q}||} - \frac{S_k(Q) \cdot S_k^T(Q)}{||Y_k - Q||}\right]$$

for $S_k(Q) = S(Y_k - Q)$ and $S_k(\widehat{Q}) = S(Y_k - \widehat{Q})$. We have

$$||J_1||_1 \leq \frac{1}{N}\sum_s \frac{||I_d||_1}{\pi_k}\frac{\Big|||Y_k - \widehat{Q}|| - ||Y_k - Q||\Big|}{||Y_k - \widehat{Q}|| \cdot ||Y_k - Q||} \leq \sqrt{d}\frac{\hat{N}}{N}\frac{||\widehat{Q} - Q||}{M^2} = O_p(1) \cdot o_p(1) = o_p(1)$$

by assumptions (A1)-(A4). After some manipulations and using the fact that $||S_k(\widehat{Q}) - S_k(Q)|| = ||h_k(Q) - h_k(\widehat{Q})|| = o_p(1)$ by (3.13), we get

$$\begin{aligned} ||J_2||_1 &\leq \frac{1}{NM}\sum_s \frac{1}{\pi_k}\left[||S_k(\widehat{Q}) - S_k(Q)||^2 + 2||S_k(\widehat{Q}) - S_k(Q)|| \cdot ||S_k(Q)|| + ||S_k(Q)||^2 \cdot \frac{||\widehat{Q} - Q||}{M}\right] \\ &= o_p(1). \end{aligned}$$

We also used the facts that $||S_k(Q)|| = 1$ and $\hat{N}/N = O_p(1)$. As a consequence, $\frac{1}{N}(\widehat{J}(\widehat{Q}) - \widehat{J}(Q)) = o_p(1)$.

3. The linearized variable satisfies $||u_k|| \leq N^{-1}||(\frac{1}{N}J_U(Q))^{-1}||_1 \cdot ||h_k(Q)||$ and lemma (3.4.2) and the point 1 of this lemma give that $\|u_k\| = O(N^{-1}$. Next,

$$\begin{aligned} \hat{u}_k - u_k &= J_U^{-1}(Q)h_k(Q) - \widehat{J}^{-1}(\widehat{Q})h_k(\widehat{Q}) \\ &= \left[J_U^{-1}(Q) - \widehat{J}^{-1}(Q)\right]h_k(Q) + \left[\widehat{J}^{-1}(Q) - \widehat{J}^{-1}(\widehat{Q})\right]h_k(\widehat{Q}) + \left[h_k(Q) - h_k(\widehat{Q})\right]\widehat{J}^{-1}(Q) \\ &= o_p\left(\frac{1}{N}\right) \end{aligned}$$

since $J_U^{-1}(Q) - \widehat{J}^{-1}(Q) = O_p(\frac{1}{N\sqrt{n}})$ by lemma 3.4.2, $\widehat{J}^{-1}(Q) - \widehat{J}^{-1}(\widehat{Q}) = o_p(\frac{1}{N})$ by point 2 of this lemma and finally, $h_k(\widehat{Q}) = O(1)$, $h_k(Q) = O(1)$ and $h_k(Q) - h_k(\widehat{Q}) = o_p(1)$ by point 1.

□

Proof of result 4. The asymptotic variance of $\widehat{Q}$ is given by

$$\mathbb{AV}_p(\widehat{Q}) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l^T}{\pi_l} = J_U^{-1}(Q) \cdot V_p(\widehat{H}(Q)) \cdot J_U^{-1}(Q)$$

and is estimated unbiasedly by

$$\widehat{\mathbb{AV}_p}(\widehat{Q}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{u_k}{\pi_k} \frac{u_l^T}{\pi_l} = J_U^{-1}(Q) \cdot \widehat{V}_p(\widehat{H}(Q)) \cdot J_U^{-1}(Q).$$

We may write

$$\widehat{V}_p(\widehat{Q}) - \mathbb{AV}_p(\widehat{Q}) = A_N + B_N$$

with $A_N = \widehat{V}_p(\widehat{Q}) - \widehat{\mathbb{AV}_p}(\widehat{Q})$ and $B_N = \widehat{\mathbb{AV}_p}(\widehat{Q}) - \mathbb{AV}_p(\widehat{Q})$ and we intend to show in the next that both A_N and B_N are $o_p(n^{-1})$.
We have immediately that

$$B_N = J_U^{-1}(Q) \cdot \left(\widehat{V}_p(\widehat{H}(Q)) - V_p(\widehat{H}(Q)) \right) \cdot J_U^{-1}(Q) = o_p(n^{-1})$$

by lemma 3.4.2 and assumption (A5) and (A6).
Let us denote by $c_{kl} = \frac{\Delta_{kl}}{\pi_{kl}} \frac{I_k I_l}{\pi_k \pi_l}$ for all $k, l \in U$. The quantity A_N can be written as follows

$$\begin{aligned} A_N &= \sum_U \sum_U c_{kl}(\hat{u}_k \hat{u}_l^T - u_k u_l^T) \\ &= \sum_U \sum_U c_{kl}(\hat{u}_k - u_k)(\hat{u}_k - u_k)^T + \sum_U \sum_U c_{kl}(\hat{u}_k - u_k) u_l^T + \sum_U \sum_U c_{kl} u_k (\hat{u}_l - u_l)^T \\ &= A_{1N} + A_{2N} + A_{2N}^T. \end{aligned}$$

We have for the first term from the right side that

$$\begin{aligned} ||A_{1N}||_1 &\leq \sum_U |c_k| \cdot ||\hat{u}_k - u_k||^2 + \sum \sum\nolimits_{k \neq l} |c_{kl}| \cdot ||\hat{u}_k - u_k|| \cdot ||\hat{u}_l - u_l|| \\ &\leq \left[\frac{1 - \lambda_1}{\lambda_1^2} + \frac{N}{n} \frac{n \max |\Delta_{kl}|}{\lambda_1^2 \lambda_2} \right] \sum_U ||\hat{u}_k - u_k||^2 = o_p\left(\frac{1}{n}\right) \end{aligned} \tag{3.14}$$

from (A1)-(A4) and lemma 3.7.1. With the same arguments, we obtain

$$||A_{2N}||_1 \leq \frac{1 - \lambda_1}{\lambda_1^2} \sum_U ||\hat{u}_k - u_k|| \cdot ||u_k|| + \frac{\max |\Delta_{kl}|}{\lambda_1^2 \lambda_2} \sum \sum\nolimits_{k \neq l} ||\hat{u}_k - u_k|| \cdot ||u_l|| = o_p\left(\frac{1}{n}\right) \tag{3.15}$$

Relations (3.14) and (3.15) give us that for A_N is also of order $o_p\left(\frac{1}{n}\right)$. This completes the proof.

□

Bibliographie

[1] Bedall, F.K. and Zimmermann, H. (1979), Algorithm AS 143, the Mediancenter, Applied Statistics, **28**, 325-328.

[2] Binder, D.A. (1983), On the variances of asymptotically normal estimators from complex surveys, International Statistical Review, **51**, 279-292.

[3] Breidt, F.J. and Opsomer, J. (2000), Local Polynomial Regression Estimators in Survey Sampling, The Annals of Statistics, **28**, 1026-1053.

[4] Chaudhuri, P. (1996), On a geometric notation of quantiles for multivariate data, Journal of the American Statistical Association **91**, 862-872.

[5] Chakraborty, B. (2001), On affine equivariant multivariate quantiles, The Institute of Statistical Mathematics, **53**, 80-403.

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

[7] Deville, J.C. (1999), Variance estimation for complex statistics and estimators : linearization and residual techniques, Survey Methodology, **25**, 193-203.

[8] Ferguson, T. (1967), Mathematical Statistics : A Decision Theoric Approach, Academic Press, New York.

[9] Francisco, C.A. and Fuller, W.A. (1991), Quantile estimation with a complex survey design, The Annals of Statistics, **19**, 454-469.

[10] Gini, C. and Galvani, L. (1929), Di talune estensioni dei concetti di media ai caratteri qualitativi, Journal of the American Statistical Association, **25**, 448-450.

[11] Godambe, V.P. and Thompson, M. E. (1986), Parameters of Superpopulation and Survey Population : Their Relationships and Estimation, International Statistical Review, **54**, 127-138.

[12] Haldane, J.B.S. (1948), Note on the median of a multivariate distribution, Biometrika **35**, 414-415.

[13] Horvitz, D.G. and Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association, **47**, 663-685.

[14] Koenker, R. and Basset, G. (1978), Regression Qantiles, Econometrica, **46**, 33-50.

[15] Kuk, Estimation of distribution function and medians under sampling with unequal probabilities, Biometrika, **75**, 97-103.

[16] Isaki, C. I. and Fuller, W. A. (1982), Survey Design under the regression superpopulation model, Journal of the American Statistical Association, **77**, 89-96.

[17] Kemperman, J.H.B. (1987), The median of a finite measure on a Banach space, In : Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amesterdam, 217-230.

[18] Robinson, P.M. and Särndal, C.E. (1983), Asymptotic Properties of the Generalized Regression Estimator in Probability Sampling, Sankhya, **45**, 240-248.

[19] Reaven, G. M. and Miller, R. G. (1979), An attempt to define the nature of chemical diabetes using a multidimensional analysis, Diabetologia, **16**, 17-24.

[20] Särndal C.E., (1980), On π-inverse weighting versus best linear unbiased weighting in probability sampling, Biometrika, **67**, 639-650.

[21] Särndal C.E. , Swensson B. and Wretman J. (1992), Model Assisted Survey Sampling, Springer, Berlin.

[22] Serfling, R. (2002), Quantile functions for multivariate analysis : approaches and applications, Statistica Neerlandica, **56**, 214-232.

[23] Small, C.G. (1990), A survey of multidimensional medians, International Statistical Review, **58**, 263-277.

Chapitre 4

Functional Principal Components Analysis with Survey Data*

Abstract : This work aims at performing Functional Principal Components Analysis (FPCA) with Horvitz-Thompson estimators when the observations are curves collected with survey sampling techniques. FPCA relies on estimations of the eigenelements of the covariance operator which can be seen as nonlinear functionals. Adapting to our functional context the linearization technique based on the influence function developed by Deville (1999), we prove that these estimators are asymptotically design unbiased and convergent. Under mild assumptions, asymptotic variances are derived for the FPCA' estimators and convergent estimators of them are proposed. Our approach is illustrated with a simulation study and we check the good properties of the proposed estimators of the eigenelements as well as their variance estimators obtained with the linearization approach.

Keywords : covariance operator, eigenfunctions, Horvitz-Thompson estimator, influence function, perturbation theory, variance estimation, von Mises expansion.

Contents

4.1 Introduction and notations 79
4.2 Survey framework and PCA 81
4.3 Asymptotic Properties 84
4.4 A simulation study 88

4.1 Introduction and notations

Functional Data Analysis whose main purpose is to provide tools for describing and modeling sets of curves is a topic of growing interest in the statistical community. The books by Ramsay and Silverman (2002, 2005) propose an interesting description of the available procedures dealing with functional observations whereas Ferraty and Vieu (2006) present a completely nonparametric point of view. These functional approaches mainly rely on generalizing

*Article écrit en collaboration avec Hervé Cardot, Camelia Goga et Catherine Labruère et soumis au *Journal of Statistical Planning and Inference.*

multivariate statistical procedures in functional spaces and have been proved useful in various domains such as chemometrics (Hastie and Mallows, 1993), economy (Kneip and Utikal, 2001), climatology (Besse *et al.* 2000), biology (Kirkpatrick and Heckman, 1989, Chiou *et al.* 2003) or remote sensing (Cardot *et al.*, 2003).

When dealing with functional data, the statistician generally wants, in a first step, to represent as well as possible the sample of curves in a small dimension space in order to get a description of the functional data that allows interpretation. This objective can be achieved by performing a Functional Principal Components Analysis (FPCA) which provides a small dimension space which is able to capture, in an optimal way according to a variance criterion, the main modes of variability of the data. These modes of variability are given by considering, once the mean function has been subtracted off, projections onto the space generated by the eigenfunctions of the covariance operator associated to the largest eigenvalues. This technique is also known as Karhunen-Loeve expansion in probability or Empirical Orthogonal Functions (EOF) in climatology and numerous works have been published on this topic. From a statistical perspective, the seminal paper by Deville (1974) introduces the functional framework whereas Dauxois *et al.* (1982) give asymptotic distributions. More recent works deal with smoothing or interpolation procedures (Castro *et al.*, 1986, Besse and Ramsay, 1986, Silverman, 1996 or Cardot, 2000) as well as bootstrap properties (Kneip and Utikal, 2001) or sparse data (James *et al.*, 2000).

The way data are collected is seldom taken into account in the literature and one generally supposes the data are independent realizations drawn from a common functional probability distribution. Even if this assumption can be supposed to be satisfied in most situations, there are some cases for which it will lead to estimation procedures that are not adapted to the sampling scheme. Design of experiments approaches have been studied by Cuevas *et al.* (2003) but nothing has been done in the functional framework, as far as we know, from a survey sampling point of view whereas it can have some interest for practical applications. For instance, Dessertaine (2006) considers the estimation with time series procedures of electricity demand at fine time scales with the observation of individual electricity consumption curves. In this study, the individuals are selected according to balancing techniques (Deville and Tillé, 2004) and consequently they do not have the same probability to belong to the sample. More generally, there are now data (data streams) produced automatically by large numbers of distributed sensors which generate huge amounts of data that can be seen as functional. The use of sampling techniques to collect them proposed for instance in Chiky and Hébrail (2007) seems to be a relevant approach in such a framework allowing a trade off between storage capacities and accuracy of the data. In such situations classical estimation procedures will lead to misleading interpretation of the FPCA since the mean and covariance structure of the data will not be estimated properly.

We propose in this work estimators of the FPCA when the curves are collected with survey sampling strategies. Let us note that Skinner *et al.* (1986) have studied some properties of multivariate PCA in such a survey framework. Unfortunately, this work has received little attention in the statistical community. The functional framework is different since the eigenfunctions which exibit the main modes of variability of the data are also functions and can be naturally interpreted as modes of variability varying along time. FPCA is also the background of linear and generalized linear models which make this technique useful if one wants to introduce model-assisted approaches (Särndal *et al.*, 1992) that can take auxiliary information into account. In this new functional framework, we define estimators of the mean function and the covariance operator based on the Horvitz-Thompson approach. In order to calculate and estimate the variance of non-linear estimators, we use the influence function linearization method introduced by Deville (1999). The influence function of an estimator has been introduced in

robust statistics by Hampel (1974) in order to evaluate the sensitivity of the associated functional to infinitesimal contaminations and is also useful for computing the asymptotic variance of the estimator. In survey sampling theory, the influence function indicates the variability due to an infinitesimal variation of the weight associated to an individual. Campbell and Little (1980) proposed, in a pioneer work, to use the influence function for estimating the variance of complex statistics and compared it with a jackknife variance estimator. Deville (1999) gives the theoretical framework for developing a linearization theory for very general nonlinear parameters of interest such as quantiles, measures of income inequality (Gini index or population below the poverty threshold) or principal components analysis in a multivariate setting. In such a context, we can not perform a first-order Taylor expansion of the associated complex statistics but we can make a first-order von Mises (1947) expansion of the functional giving these complex statistics. The influence function appears then in the first-order term of the von Mises expansion which is in fact, under broad assumptions, the asymptotic variance of the complex statistics.

The paper is structured as follows. Section 2 presents the functional principal components analysis in the setting of finite populations and defines the Horvitz-Thompson estimator in the functional framework. The generality of the influence function allows us to extend the estimators proposed by Deville to our functional objects. Section 3 gives the asymptotic properties. We show in section 3.1 that the FPCA' estimators are asymptotically design unbiased and convergent. Section 3.2 provides approximations and convergent estimators of the variances of FPCA' estimators with the help of perturbation theory (Kato, 1966). Section 4 proposes a simulation study which shows the good behavior of our estimators for various sampling schemes as well as the ability of linearization techniques to give good approximations to their theoretical variances. The proofs are gathered in an Appendix.

4.2 Survey framework and PCA

4.2.1 FPCA in a finite population setting

Let us consider a finite population $U = \{1, \ldots, k, \ldots, N\}$ with size N, not necessarily known, and a functional variable $\mathcal{Y}$ defined for each element k of the population $U : Y_k = (Y_k(t))_{t\in[0,1]}$ belongs to the separable Hilbert space $L^2[0,1]$ of square integrable functions defined on the closed interval $[0,1]$ equipped with the usual inner product $\langle\cdot,\cdot\rangle$ and the norm $\|\cdot\|$.

The mean function $\mu \in L^2[0,1]$, is defined by

$$\mu(t) = \frac{1}{N}\sum_{k\in U} Y_k(t), \quad t \in [0,1] \tag{4.1}$$

and the covariance operator Γ by

$$\Gamma = \frac{1}{N}\sum_{k\in U} (Y_k - \mu)\otimes(Y_k - \mu) \tag{4.2}$$

where the tensor product of two elements a and b of $L^2[0,1]$ is the rank one operator such that $a\otimes b(u) = \langle a, u\rangle b$ for all u in $L^2[0,1]$. We have in an equivalent way the following representation of the covariance operator

$$\Gamma u(t) = \int_0^1 \gamma(s,t)u(s)\, ds \tag{4.3}$$

where $\gamma(s,t)$ is the covariance function

$$\gamma(s,t) = \frac{1}{N}\sum_{k\in U} (Y_k(t) - \mu(t))\,(Y_k(s) - \mu(s))\,, \quad (s,t) \in [0,1]\times[0,1]. \tag{4.4}$$

Let us note that the functions Y_k are not random and thus the term covariance should not be understood according to the usual definition but as it is considered in the survey sample terminology.

The operator Γ is symmetric and non negative ($\langle \Gamma u, u\rangle \geq 0$). Its eigenvalues, sorted in decreasing order, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0$, satisfy

$$\Gamma v_j(t) \;=\; \lambda_j \; v_j(t), \quad t \in [0,1], \quad j = 1, \ldots, N, \tag{4.5}$$

where the eigenfunctions v_j form an orthonormal system in $L^2[0,1]$, *i.e* $\langle v_j, v_{j'}\rangle = 1$ if $j = j'$ and zero else.

We can get now an expansion similar to the Karhunen-Loeve expansion or FPCA which allows to get the best approximation in a finite dimension space with dimension q to the curves of the population

$$Y_k(t) \quad = \quad \mu(t) + \sum_{j=1}^{q} \langle Y_k - \mu, v_j\rangle v_j(t) + R_{q,k}(t), \quad t \in [0,1]$$

which is based on the fact that the quadratic loss criterion

$$R_q \;=\; \frac{1}{N}\sum_{i=1}^{N} \left\| Y_k - \left(\phi_0 + \sum_{j=1}^{q} \langle Y_k - \phi_0, \phi_j\rangle \phi_j\right)\right\|^2$$

is clearly minimum for $\phi_0 = \mu$ and $\phi_j = v_j$, $j = 1, \ldots q$. This means that the space generated by the eigenfunctions $v_1, \cdots, v_q$ gives a representation of the main modes of variation along time t of the data around the mean μ and the explained variance of the projection onto each v_j is given by the eigenvalue

$$\lambda_j \quad = \quad \frac{1}{N}\sum_{k\in U} \langle Y_k - \mu, v_j\rangle^2 \; .$$

We aim at estimating the mean function μ and the covariance operator Γ in order to deduce estimators of the eigenelements (λ_j, v_j) when the data are obtained with survey sampling procedures.

4.2.2 The Horvitz-Thompson Estimator

Let us consider a sample s of n individuals, *i.e.* a subset $s \subset U$, selected according to a probabilistic procedure $p(s)$ where p is a probability distribution on the set of 2^N subsets of U. We denote by $\pi_k = \Pr(k \in s)$ for all $k \in U$ the first order inclusion probabilities and by $\pi_{kl} = \Pr(k \;\&\; l \in s)$ for all $k, l \in U$ with $\pi_{kk} = \pi_k$, the second order inclusion probabilities. We suppose that all the individuals and all the pairs of individuals of the population have non null probabilities to be selected in the sample s, namely $\pi_k > 0$ and $\pi_{kl} > 0$. We also suppose that π_k and π_{kl} are not depending on $t \in [0,1]$. This means that once we have selected the sample s of individuals, we observe $Y_k(t)$ for all $t \in [0,1]$ and all $k \in s$. Let us start with the simplest case, the estimation of the finite population total of the Y_k curves denoted by

$$t_Y = \sum_{k\in U} Y_k.$$

The Horvitz-Thompson (HT) estimator $\widehat{t}_{Y\pi}$ of t_Y is a function belonging to $L^2[0,1]$ defined as follows

$$\widehat{t}_{Y\pi} = \sum_{k\in s} \frac{Y_k}{\pi_k} = \sum_{k\in U} \frac{Y_k}{\pi_k} I_k$$

where $I_k = \mathbf{1}_{\{k \in s\}}$ is the sample membership indicator of element k (Särndal *et al.*, 1992). Note that the variables I_k are random with $Pr(I_k = 1) = \pi_k$ whereas the curves Y_k are considered as fixed with respect to the sampling design $p(s)$. So, the HT estimator $\widehat{t}_{Y\pi}$ is p-unbiased, namely

$$E_p(\widehat{t}_{Y\pi}) = t_Y$$

where $E_p(\cdot)$ is the expectation with respect to the sampling design.

The variance operator of $\widehat{t}_{Y\pi}$ calculated with respect to $p(s)$ is the HT variance

$$\mathrm{V}_p(\widehat{t}_{Y\pi}) = \sum_U \sum_U \Delta_{kl} \frac{Y_k}{\pi_k} \otimes \frac{Y_l}{\pi_l}$$

and it is estimated p-unbiasedly by

$$\widehat{\mathrm{V}}_p(\widehat{t}_{Y\pi}) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{Y_k}{\pi_k} \otimes \frac{Y_l}{\pi_l}$$

with the notation $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$. One may obtain equivalent integral representations of $\mathrm{V}_p(\widehat{t}_{Y\pi})$ and $\widehat{\mathrm{V}}_p(\widehat{t}_{Y\pi})$ similar as in equations (4.3) and (4.4). For a fixed-size sampling design, we can give a Yates-Grundy-Sen variance formula,

$$\mathrm{V}_p(\widehat{t}_{Y\pi}) = -\frac{1}{2} \sum_U \sum_U \Delta_{kl} \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right) \otimes \left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l} \right).$$

Example : *Let us select a sample of n curves Y_k according to a simple random sample without replacement (SI) from U. We have $\widehat{t}_{Y\pi} = (N/n) \sum_{k \in s} Y_k$ with variance $V_{SI}(\widehat{t}_{Y\pi}) = N^2 \frac{1-f}{n} S^2_{YU}$ for $f = n/N$ and $S^2_{YU} = \frac{1}{N-1} \sum_U (Y_k - \mu) \otimes (Y_k - \mu)$ the population variance. The variance estimator is given by $\widehat{V}_{SI}(\widehat{t}_{Y\pi}) = N^2 \frac{1-f}{n} S^2_{Ys}$ with $S^2_{Ys} = \frac{1}{n-1} \sum_s (Y_k - \mu_s) \otimes (Y_k - \mu_s)$ and $\mu_s = \frac{1}{n} \sum_s Y_k$.*

4.2.3 Substitution Estimator for Nonlinear Parameters

Consider now the situation when the estimation of a nonlinear function $\Phi = \Phi(t_Y)$ is desired. When the population size is unknown, the mean function (4.1) or the covariance operator (4.2) are particular cases of Φ. Moreover, we would like to calculate and to estimate the variance of the estimator $\widehat{\Phi}$ of Φ. Besides the nonlinearity of $\widehat{\Phi}$, we have to cope with the fact that $\mathcal{Y}$ is a functional variable which makes the variance estimation issue more difficult. In order to overcome this, we adapt the linearization technique based on the influence function introduced by Deville (1999) to the functional framework. This approach is based on the fact that the population parameter of interest can be written as a functional T depending on a certain measure M, namely $\Phi = T(M)$, and the estimator $\widehat{\Phi}$ can be seen as the functional T of a random measure $\widehat{M}$ which is close to M, namely $\widehat{\Phi} = T(\widehat{M})$.
Let us introduce now the discrete measure M defined on $L^2[0,1]$ as follows

$$M = \sum_{k \in U} \delta_{Y_k}$$

where δ_{Y_k} is the Dirac function taking value 1 if $Y = Y_k$ and zero else. The following parameters of interest can be defined as functionals of M :

$$N(M) = \int dM \quad \text{and} \quad \mu(M) = \frac{\int \mathcal{Y} dM}{\int dM}$$

$$\Gamma(M) = \frac{\int (\mathcal{Y} - \mu(M)) \otimes (\mathcal{Y} - \mu(M))\, dM}{\int dM}$$

and the eigenelements given by (4.5) are implicit functionals T of M.
The measure M is estimated by the random measure $\widehat{M}$ associating the weight $1/\pi_k$ for each Y_k with $k \in s$ and zero else,

$$\widehat{M} = \sum_{k \in U} \frac{\delta_{Y_k}}{\pi_k}\, I_k$$

and Φ is then estimated by $\widehat{\Phi} = T(\widehat{M})$ also called the *substitution estimator*. For example, the substitution estimators for μ and Γ are

$$\widehat{\mu} = \frac{1}{\widehat{N}} \sum_{k \in s} \frac{Y_k}{\pi_k} \tag{4.6}$$

$$\widehat{\Gamma} = \frac{1}{\widehat{N}} \sum_{k \in s} \frac{Y_k \otimes Y_k}{\pi_k} - \widehat{\mu} \otimes \widehat{\mu} \tag{4.7}$$

where the size N of the population is estimated by $\widehat{N} = \sum_{k \in s} \frac{1}{\pi_k}$ when it is not known. Then estimators of the eigenfunctions $\{\widehat{v}_j, j = 1, \dots q\}$ associated to the q largest eigenvalues $\{\widehat{\lambda}_j, j = 1, \dots q\}$ are obtained readily by the eigen-analysis of the estimated covariance operator $\widehat{\Gamma}$.

Remark. *In practice we do not observe the whole curves but generally discretized versions at m design points $0 \leq t_1 < t_2 < \cdots < t_m \leq 1$ that we suppose to be the same for all the curves. Quadradure rules are often employed in order to get numerical approximations to integrals and inner product by summations : for each u in $L^2[0,1]$ we get an accurate discrete approximation to the integral*

$$\int_0^1 u(t)dt \approx \sum_{\ell=1}^{m} w_\ell\ u(t_\ell)$$

provided the number of design points p is large enough and the grid is sufficiently fine (Ramsay and Silverman, 2005).

4.3 Asymptotic Properties

We give in this section the asymptotic properties of our estimators considering the super-population asymptotic framework introduced by Isaki and Fuller (1982) which supposes that the population and the sample sizes tend to infinity. Let $U_{\mathbb{N}}$ be a population with infinite (denumerable) number of individuals and consider a sequence of nested sub populations such that $U_1 \subset \cdots \subset U_{\nu-1} \subset U_\nu \subset U_{\nu+1} \subset \cdots \subset U_{\mathbb{N}}$ of sizes $N_1 < N_2 < \ldots < N_\nu < \ldots$. Consider then a

sequence of samples s_ν of size n_ν drawn from U_ν according to the fixed-size sampling designs $p_\nu(s_\nu)$ and denote by $\pi_{k\nu}$ and $\pi_{kl\nu}$ their first and second order inclusion probabilities . Note that the sequence of sub populations is an increasing nested one while the sample sequence is not. For sake of simplicity, we will drop the subscript ν in the following.

We assume that the following assumptions are satisfied :

(A1) $\sup_{k\in U} \|Y_k\| \leq C < \infty$,

(A2) $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0,1)$,

(A3) $\min_{k\in U_N} \pi_k \geq \lambda > 0$, $\min_{k\neq l} \pi_{kl} \geq \lambda* > 0$ and $\overline{\lim}_{N\to\infty} n \max_{k\neq l} |\pi_{kl} - \pi_k\pi_l| < \infty$.

Hypothese (A1) is rather classical in functional data analysis. Note that it does not imply that the curves $Y_k(t)$ are uniformly bounded in k and $t \in [0,1]$. Hypotheses (A2) and (A3) are checked for usual sampling plans (Robinson and Särndal, 1983, Breidt and Opsomer, 2000).
We also suppose that the functional T giving the parameter of interest, $\Phi = T(M)$, is a homogeneous functional of degree α, namely for any positive number r, $T(rM) = r^\alpha T(M)$ and $\lim_{N\to\infty} N^{-\alpha} T(M) < \infty$. For example, if T is the mean function μ or the covariance operator Γ, then $T(M) = T(M/N)$, that is to say μ and Γ are functionals of degree zero with respect to M. Let us note that the eigenelements of Γ are also functionals of degree zero with respect to M.

4.3.1 ADU-ness and Convergence of Estimators

The substitution estimators of μ and Γ defined in (4.6) and (4.7), as well as $\hat{\lambda}_j$ and $\hat{v}_j$, are no longer p-unbiased. Nevertheless, we show in the next that, in large samples, they are *asymptotically design unbiased* (ADU) and *convergent in probability.*
An estimator $\widehat{\Phi}$ of Φ of degree α is said to be *asymptotically design unbiased* (ADU) if

$$\lim_{N\to\infty} N^{-\alpha} \left(E_p(\widehat{\Phi}) - \Phi \right) = 0.$$

We say that $\widehat{\Phi}$ satisfies $N^{-\alpha}\left(\widehat{\Phi} - \Phi\right) = O_p(u_n)$ for a sequence u_n of positive numbers if there is a constant C such that for any $\varepsilon > 0$, $\Pr\left(\left|\widehat{\Phi} - \Phi\right| \geq CN^\alpha u_n\right) \leq \epsilon$. The estimator is *convergent in probability* if one can find a sequence u_n tending to zero as n tends to infinity such as $N^{-\alpha}\left(\widehat{\Phi} - \Phi\right) = O_p(u_n)$.
Let us also introduce the Hilbert-Schmidt norm, denoted by $\|\cdot\|_2$ for operators mapping $L^2[0,1]$ to $L^2[0,1]$. It is induced by the inner product between two operators Γ and Δ defined by $\langle \Gamma, \Delta \rangle_2 = \sum_{\ell=1}^{\infty} \langle \Gamma e_\ell, \Delta e_\ell \rangle$ for any orthonormal basis $(e_\ell)_{\ell\geq 1}$ of $L^2[0,1]$. In particular, we have that $\|\Gamma\|_2^2 = \sum_{\ell=1}^{\infty} \langle \Gamma e_\ell, \Gamma e_\ell \rangle = \sum_{j\geq 1} \lambda_j^2$.

Proposition 4.3.1 *Under hypotheses (A1), (A2) and (A3),*

$$\begin{aligned} E_p\left(\frac{N - \widehat{N}}{N}\right)^2 &= O(n^{-1}), \\ E_p \left\|\mu - \widehat{\mu}\right\|^2 &= O(n^{-1}), \\ E_p \left\|\Gamma - \widehat{\Gamma}\right\|_2^2 &= O(n^{-1}). \end{aligned}$$

If we suppose that the non null eigenvalues are distinct, we also have,

$$E_p\left(\sup_j \left|\lambda_j - \widehat{\lambda_j}\right|\right)^2 = O(n^{-1}),$$

and for each fixed j,

$$E_p \left\| v_j - \widehat{v_j} \right\|^2 \quad = \quad O(n^{-1}).$$

As a consequence, the above estimators are ADU and convergent in probability.
The proof is given in the Appendix.

4.3.2 Variance Approximation and Estimation

Let us now define, when it exists, the influence function of a functional T at point $\mathcal{Y} \in L^2[0,1]$ say $IT(M, \mathcal{Y})$, as follows

$$IT(M, \mathcal{Y}) \quad = \quad \lim_{h \to 0} \frac{T(M + h\delta_{\mathcal{Y}}) - T(M)}{h}$$

where $\delta_{\mathcal{Y}}$ is the Dirac function at $\mathcal{Y}$. Note that this is not exactly the usual definition of the influence function (see *e.g.* Hampel, 1974 or Serfling, 1980) and it has been adapted to the survey sampling framework by Deville (1999). We define the *linearized variables* u_k, $k \in U$ as the influence function of T at M and $\mathcal{Y} = Y_k$, namely

$$u_k = IT(M, Y_k).$$

Note that the linearized variables depend on Y_k for all $k \in U$ and as a consequence, they are all unknown.

We can give a first order von Mises expansion of our functional T in $\dfrac{\widehat{M}}{N}$ close to $\dfrac{M}{N}$,

$$\begin{aligned} T\left(\frac{\widehat{M}}{N}\right) \quad &= \quad T\left(\frac{M}{N}\right) + \int IT\left(\frac{M}{N}, y\right) d\left(\frac{\widehat{M}}{N} - \frac{M}{N}\right) + R_T\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right) \quad \text{or} \\ N^{-\alpha} T(\widehat{M}) \quad &= \quad N^{-\alpha} T(M) + \frac{1}{N} \sum_{k \in U} IT\left(\frac{M}{N}, Y_k\right) \left(\frac{I_k}{\pi_k} - 1\right) + R_T\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right). \end{aligned} \tag{4.8}$$

The above expansion tells us that the approximated variance of the estimator $N^{-\alpha} T(\widehat{M})$ is $N^{-2\alpha}$ times the variance of the HT estimator of the population mean of $IT\left(\frac{M}{N}, Y_k\right)$ provided the remainder term $R_T\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right)$ is negligible.
Before handling the remainder term, let us first calculate the influence function for our parameters of interest.

Proposition 4.3.2 *Under assumption (A1), we get that the influence functions of μ and Γ exist and*

$$\begin{aligned} I\mu(M, Y_k) \quad &= \quad \frac{1}{N}(Y_k - \mu) \\ I\Gamma(M, Y_k) \quad &= \quad \frac{1}{N}\left((Y_k - \mu) \otimes (Y_k - \mu) - \Gamma\right). \end{aligned} \tag{4.9}$$

If moreover, the non null eigenvalues of Γ are distinct, then

$$I\lambda_j(M, Y_k) \quad = \quad \frac{1}{N}\left(\langle Y_k - \mu, v_j \rangle^2 - \lambda_j\right) \tag{4.10}$$

$$Iv_j(M, Y_k) \quad = \quad \frac{1}{N}\left(\sum_{\ell \neq j} \frac{\langle Y_k - \mu, v_j \rangle \langle Y_k - \mu, v_\ell \rangle}{\lambda_j - \lambda_\ell} v_\ell\right). \tag{4.11}$$

If T is one of the above parameter of interest, then $IT\left(\frac{M}{N}, Y_k\right) = N \cdot IT(M, Y_k)$.

Let us remark that the influence functions of the eigenelements are similar to those found in the multivariate framework for classical PCA (Croux and Ruiz-Gazen, 2005).

We are now able to state the main result of the paper which proves that the remainder term R_T defined in equation (4.8) is negligible and that the linearization approach can be used to get the asymptotic variance of our substitution estimators.

Proposition 4.3.3 *Suppose the hypotheses (A1), (A2) and (A3) are true. Consider the functional T giving the parameters of interest defined in (4.1), (4.2) and (4.5). We suppose that the non null eigenvalues are distinct. Then $R_T\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right) = o_p(n^{-1/2})$ and*

$$T(\widehat{M}) - T(M) = \sum_{k \in U} IT\left(M, Y_k\right)\left(\frac{I_k}{\pi_k} - 1\right) + o_p(n^{-1/2}).$$

Thereafter, the asymptotic variance of $\widehat{\mu}$, resp. of $\widehat{v}_j$, is equal to the variance operator of the HT estimator $\sum_s \frac{u_k}{\pi_k}$ with u_k given by (4.9), resp. by (4.11), and its expression is given by

$$AV(T(\widehat{M})) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \otimes \frac{u_l}{\pi_l} \tag{4.12}$$

The asymptotic variance of $\widehat{\lambda}_j$ is

$$AV(\widehat{\lambda}_j) = \sum_U \sum_U \Delta_{kl} \frac{u_k}{\pi_k} \frac{u_l}{\pi_l} \tag{4.13}$$

with u_k given by (4.10).

The proof is given in the Appendix.

As one can notice, the variance approximations given in Proposition 4.3.3 are unknown since the double sums are considered on the whole population U and we have only a subset of it and secondly, the linearized variables u_k are not known. As a consequence, we propose to estimate (4.12) and (4.13) by the HT variance estimators replacing the linearized variables by their estimations. In the case of $\widehat{\mu}$, $\widehat{\lambda}_j$ and $\widehat{v}_j$, we obtain the following variance estimators :

$$\begin{aligned}
\widehat{V}_p(\widehat{\mu}) &= \frac{1}{\widehat{N}^2} \sum_{k \in s} \sum_{\ell \in s} \frac{1}{\pi_{k\ell}} \frac{\Delta_{k\ell}}{\pi_k \pi_\ell} (Y_k - \widehat{\mu}) \otimes (Y_\ell - \widehat{\mu}) \\
\widehat{V}_p\left(\widehat{\lambda}_j\right) &= \frac{1}{\widehat{N}^2} \sum_{k \in s} \sum_{\ell \in s} \frac{1}{\pi_{k\ell}} \frac{\Delta_{k\ell}}{\pi_k \pi_\ell} \left(\langle Y_k - \widehat{\mu}, \widehat{v}_j \rangle^2 - \widehat{\lambda}_j\right)\left(\langle Y_\ell - \widehat{\mu}, \widehat{v}_j \rangle^2 - \widehat{\lambda}_j\right) \\
\widehat{V}_p\left(\widehat{v}_j\right) &= \sum_{k \in s} \sum_{\ell \in s} \frac{1}{\pi_{k\ell}} \frac{\Delta_{k\ell}}{\pi_k \pi_\ell} \widehat{Iv}_j(M, Y_k) \otimes \widehat{Iv}_j(M, Y_\ell),
\end{aligned}$$

with $\widehat{Iv}_j(M, Y_\ell) = \frac{1}{\widehat{N}} \left(\sum_{\ell \neq j} \frac{\langle Y_k - \widehat{\mu}, \widehat{v}_j \rangle \langle Y_k - \widehat{\mu}, \widehat{v}_\ell \rangle}{\widehat{\lambda}_j - \widehat{\lambda}_\ell} \widehat{v}_\ell \right)$.

We prove now that these variance estimators are convergent in probability.

Proposition 4.3.4 *Under assumptions (A1)–(A3), we have that*

$$\begin{aligned} E_p \left\| AV(\widehat{\mu}) - \widehat{V}_p(\widehat{\mu}) \right\|_2 &= O\left(\frac{1}{n}\right) \\ E_p \left| AV(\widehat{\lambda}_j) - \widehat{V}_p(\widehat{\lambda}_j) \right| &= O\left(\frac{1}{n}\right) \\ \left\| AV(\widehat{v}_j) - \widehat{V}_p(\widehat{v}_j) \right\|_2 &= O_p\left(\frac{1}{\sqrt{n}}\right) \end{aligned}$$

for $j = 1, \ldots, q$. Therefore, variance estimators for the mean function and the eigenvalue are asymptotically design unbiased and convergent in probability. Moreover, the variance estimators of the eigenfunctions v_j, $j = 1, \ldots, q$, are convergent.

The proof is given in the Appendix.

4.4 A simulation study

We check now with a simulation study that we get accurate estimations to the eigenelements even for moderate sample sizes as well as good approximation to their variance for simple random sampling without replacement (SRSWR) and stratified sampling. In our simulations all functional variables are discretized in $m = 100$ equispaced points in the interval $[0, 1]$. Riemann approximations to the integrals are employed to deal with the discretization effects.

We consider a random variable Y following a Brownian motion with mean function $\mu(t) = \cos(4\pi t), t \in [0, 1]$ and covariance function $cov(s, t) = \min(s, t)$. We make $N = 10000$ replications of Y. We construct then two strata U_1 and U_2 of different variances by multiplying the $N_1 = 7000$ first replications of Y by $\sigma_1 = 2$ and the $N_2 = 3000$ other replications by $\sigma_2 = 4$. Our population U is the union of these two strata.

To evaluate our estimation procedures we make 500 replications of the following experiment. We draw samples according to two different sampling designs (SRSWR and stratified) and two different sample sizes $n = 100$ and $n = 1000$. Each stratified sample is built by drawing independently two SRSWR of sizes n_1 in stata U_1 and $n_2 = n - n_1$ in strata U_2. The sample sizes are chosen to take into account the different variances in the strata :

$$\frac{n_1}{n} = \frac{N_1}{N} \frac{\sigma_1}{\frac{N_1\sigma_1 + N_2\sigma_2}{N}} \ , \ \frac{n_2}{n} = \frac{N_2}{N} \frac{\sigma_2}{\frac{N_1\sigma_1 + N_2\sigma_2}{N}}$$

in analogy with univariate stratified sampling with optimal allocation (Särndal *et al.*, 1992). A stratified sample s of size $n = 100$ trajectories is drawn in Figure 4.1.

Estimation errors for the first eigenvalue and the first eigenvector are evaluated by considering the following loss criterions $\frac{\lambda_1 - \widehat{\lambda_1}}{\lambda_1}$ and $\frac{\|v_1 - \widehat{v_1}\|}{\|v_1\|}$ (Euclidean norm) among our 500 replications of the experiments. The approximations turn out to be effective as seen in Figure 4.2. For example for both sampling strategies the first eigenvector approximation has a median error lower than 3% for a sample size $n = 1000$. It also appears that the stratified sampling gives better estimations than the SRSWR sampling.

Let us look now at the variance of our estimators. Tables 4.1 and 4.2 give three variance (resp. euclidean norm of variance) approximations to the estimator of respectively the first eigenvalue and the first eigenvector. The first variance approximation to these estimators is their empirical variance and are denoted by $Var(\widehat{\lambda_1})$ and by $Var(\widehat{v_1})$, the second one is the asymptotic variance denoted by $AV(\widehat{\lambda_1})$ and by $AV(\widehat{v_1})$ whereas the third one is a $[25\%, 75\%]$

confidence interval obtained by estimating the asymptotic variance using the HT variance estimator respectively denoted by $\widehat{V}_p\left(\widehat{\lambda}_1\right)$ and $\widehat{V}_p\left(\widehat{v}_1\right)$. Errors (see Figure 4.3) in approximating the variance of the estimators by the linearization approach are evaluated by considering the following criterions : $\left|\frac{Var(\widehat{\lambda_1})-\widehat{V}_p(\widehat{\lambda}_1)}{Var(\widehat{\lambda_1})}\right|$ and $\frac{\|Var(\widehat{v_1})-\widehat{V}_p(\widehat{v}_1)\|}{\|Var(\widehat{v_1})\|}$.

As a conclusion, we first note with this simulation study that HT estimators of the covariance structure of functional observations are accurate enough to derive good estimators of the FPCA. Secondly, linear approximations by the influence function give reasonable estimation of the variance of the eigenelements for small sample sizes and accurate estimations as far as n gets larger (n=1000). We also notice that the variance of the estimators obtained by stratified sampling turns out to be smaller than with SRSWR sampling.

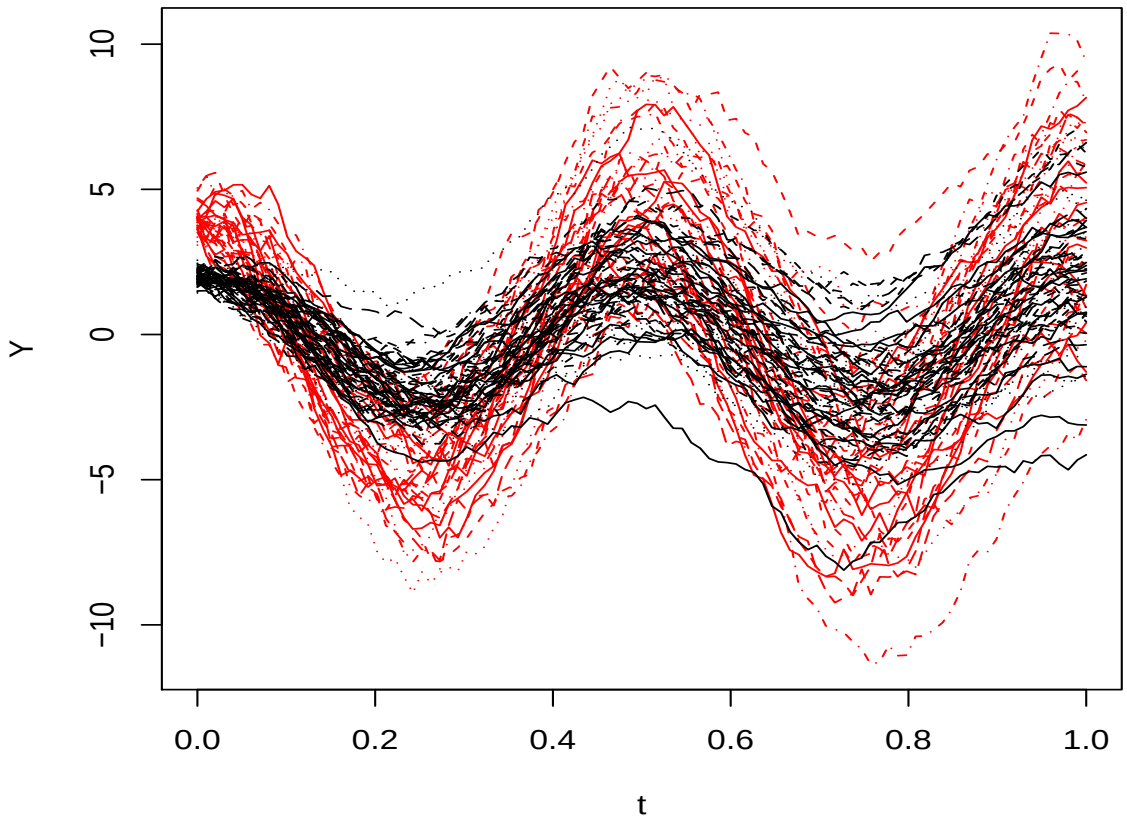


FIG. 4.1 – A stratified sample of $n = 100$ curves

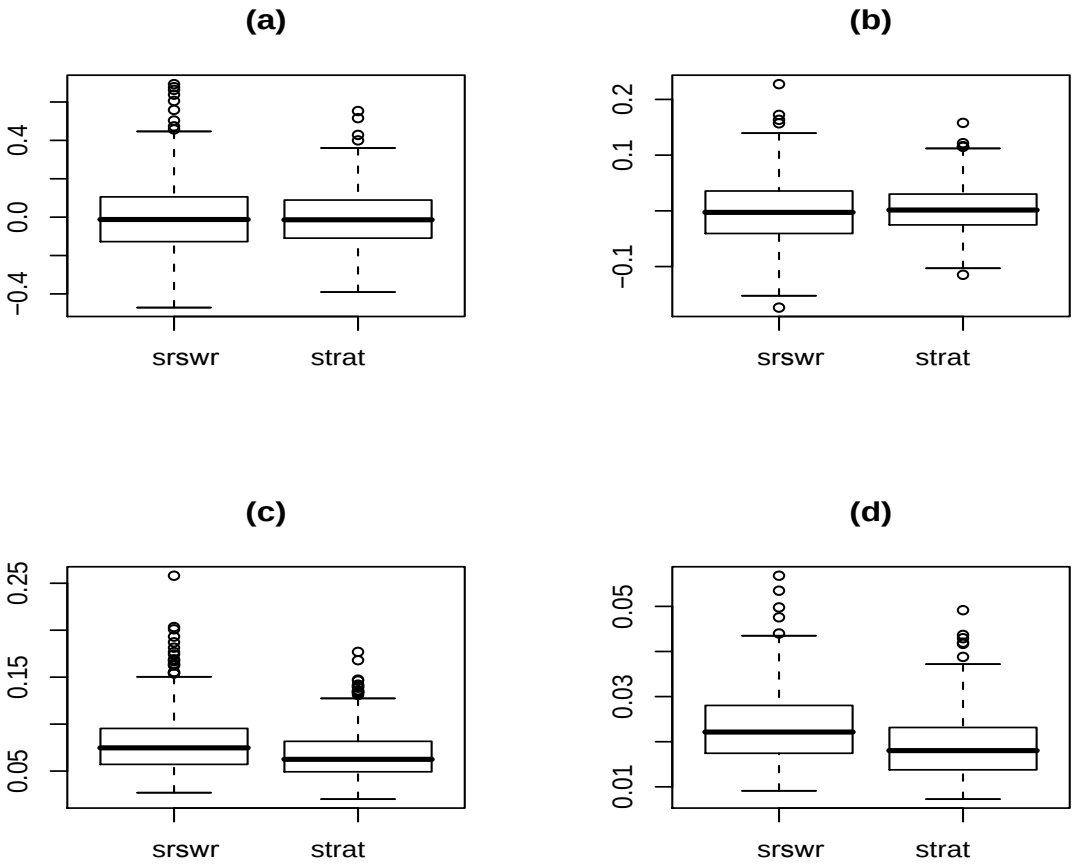


FIG. 4.2 – Estimation errors for two different sampling strategies (SRSWR and stratified sampling). First eigenvalue with $n = 100$. (a) and $n = 1000$ (b). First eigenvector with $n = 100$. (c) and $n = 1000$ (d).

	n=100		n=1000	
	SRSWR	stratified	SRSWR	stratified
$Var(\widehat{\lambda_1})$	0.314	0.223	0.0317	0.0189
$AV(\widehat{\lambda_1})$	0.340	0.209	0.0309	0.0183
$\widehat{V}_p\left(\widehat{\lambda_1}\right)$	[0.208 ;0.430]	[0.155 ;0.257]	[0.027 ;0.034]	[0.0169 ;0.0195]

TAB. 4.1 – Variance approximation of the first eigenvalue estimator.

	n=100		n=1000	
	SRSWR	stratified	SRSWR	stratified
$\|Var(\widehat{v_1})\|$	0.450	0.286	0.0396	0.0265
$\|AV(\widehat{v_1})\|$	0.3997	0.287	0.0386	0.0267
$\|\widehat{V_p}(\widehat{v_1})\|$	[0.335 ;0.491]	[0.252 ;0.354]	[0.0371 ;0.0410]	[0.0256 ;0.0280]

TAB. 4.2 – Norm of the variance approximation of the first eigenvector estimator.

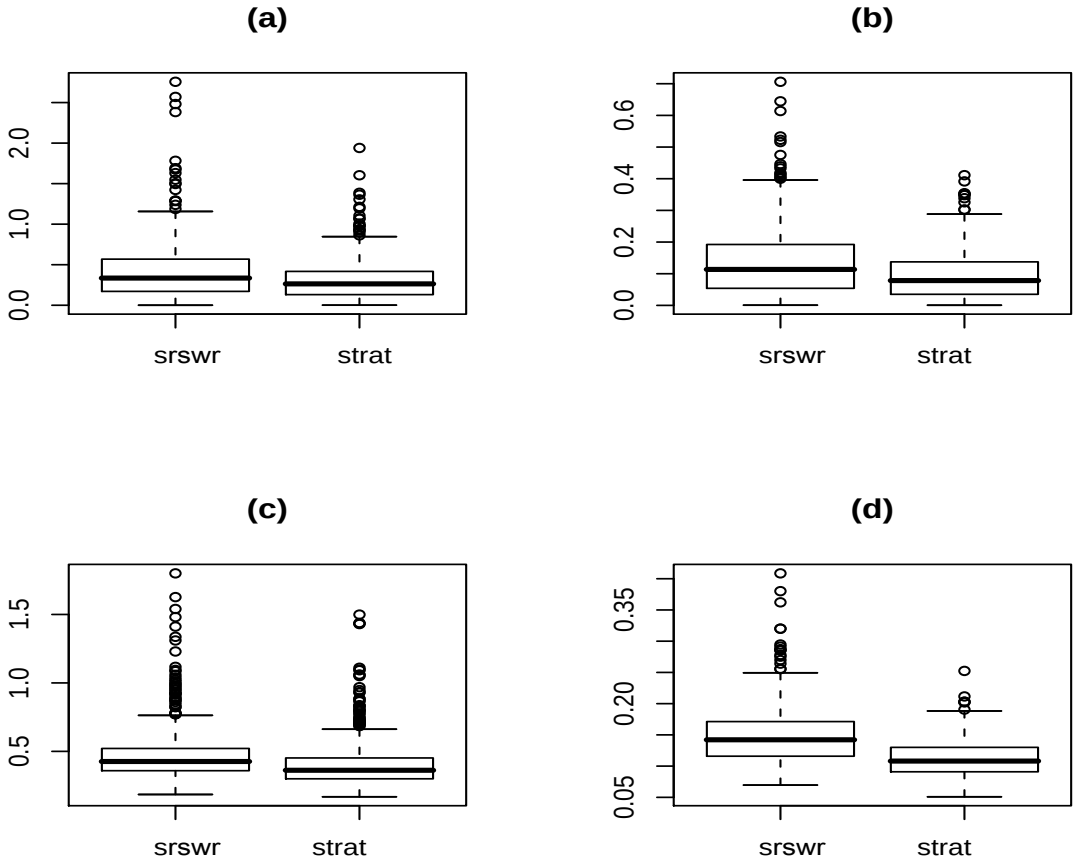


FIG. 4.3 – Estimation errors in the variance approximation for two different sampling strategies (SRSWR and stratified sampling). First eigenvalue with $n = 100$. (a) and $n = 1000$ (b). First eigenvector with $n = 100$. (c) and $n = 1000$ (d).

Appendix : proofs

Proof of proposition 4.3.1.
Let us introduce $\alpha_k = \frac{I_k}{\pi_k} - 1$, we have

$$\frac{\widehat{N} - N}{N} = \frac{1}{N} \sum_{k \in U} \alpha_k.$$

Noting that with assumptions (A2) and (A3), $E(\alpha_k^2) = (1 - \pi_k)/\pi_k < (1 - \pi_k)/\lambda$, $|E(\alpha_k \alpha_\ell)| = |\Delta_{k\ell}/(\pi_k \pi_\ell)| \leq |\Delta_{k\ell}|/\lambda^2$, and taking now the expectation, according to the sampling distribution p, we get

$$\begin{aligned} E_p \left(\frac{\widehat{N} - N}{N} \right)^2 &= \frac{1}{N^2} \sum_{k,\ell \in U} E_p(\alpha_\ell \alpha_k) \\ &= \frac{1}{N^2} \left(\sum_{k \in U} \frac{1 - \pi_k}{\pi_k} + \sum_{k \in U} \sum_{\ell \neq k} \frac{\Delta_{k\ell}}{\pi_k \pi_\ell} \right) \\ &\leq \frac{1}{N^2} \left(\frac{N}{\lambda} + \frac{N(N-1)}{n} \frac{n \max |\Delta_{k\ell}|}{\lambda^2} \right) \\ &= O\left(\frac{1}{n} \right) \end{aligned} \tag{4.14}$$

which is the first result. Looking now at the estimator of the mean function, we have

$$\begin{aligned} \widehat{\mu} - \mu &= \frac{1}{N} \sum_{k \in U} \alpha_k Y_k + \left(\frac{1}{\widehat{N}} - \frac{1}{N} \right) \sum_{k \in s} \frac{1}{\pi_k} Y_k \\ &= \frac{1}{N} \sum_{k \in U} \alpha_k Y_k + \left(\frac{N - \widehat{N}}{N} \right) \widehat{\mu} \end{aligned}$$

By assumptions (A1)-(A3) it is clear that $\|\widehat{\mu}\| = O(1)$ and consequently $E_p \left\| \frac{N - \widehat{N}}{N} \widehat{\mu} \right\|^2 = O(n^{-1})$. The first term of the right side of the inequality is dealt with as in (4.14), noticing that $\|Y_k\| \leq C$ for all k :

$$\begin{aligned} E_p \left\| \frac{1}{N} \sum_{k \in U} \alpha_k Y_k \right\|^2 &= \frac{1}{N^2} \sum_{k,\ell \in U} E_p \left(\alpha_\ell \alpha_k \right) \langle Y_k, Y_\ell \rangle \\ &\leq \frac{1}{N^2} \sum_{k,\ell \in U} |E_p(\alpha_\ell \alpha_k)| \, \|Y_k\| \, \|Y_\ell\| \\ &= O\left(\frac{1}{n} \right). \end{aligned}$$

To complete the proof, let us introduce the operator $Z_k = Y_k \otimes Y_k$ and remark that

$$\widehat{\Gamma} - \Gamma = \frac{1}{N} \sum_{k \in U} \alpha_k Z_k + \left(\frac{1}{\widehat{N}} - \frac{1}{N} \right) \sum_{k \in s} \frac{1}{\pi_k} Z_k + \mu \otimes \mu - \widehat{\mu} \otimes \widehat{\mu}$$

By assumption (A1), we have that $|\langle Z_k, Z_\ell \rangle_2| \leq \|Y_k\|^2 \|Y_\ell\|^2 \leq C^4$, for all k and ℓ and we get with similar arguments as above that

$$E_p \left\| \frac{1}{N} \sum_{k \in U} \alpha_k Z_k \right\|_2^2 = O\left(\frac{1}{n}\right)$$

and $E_p \left\| \left(\frac{1}{\widehat{N}} - \frac{1}{N} \right) \sum_{k \in s} \frac{1}{\pi_k} Z_k \right\|_2^2 = O(n^{-1})$. Remarking now that

$$\|\mu \otimes \mu - \widehat{\mu} \otimes \widehat{\mu}\|_2 \leq \|(\mu - \widehat{\mu}) \otimes \mu\|_2 + \|\widehat{\mu} \otimes (\mu - \widehat{\mu})\|_2$$

the result is proved.

Consistency of the eigenelements is an immediate consequence of classical properties of the eigenelements of covariance operators. The eigenvalues (see *e.g.* Dauxois *et al.*, 1982) satisfy

$$|\widehat{\lambda}_j - \lambda_j| \leq \left\| \widehat{\Gamma} - \Gamma \right\|_2 .$$

On the other hand, Lemma 4.3 by Bosq (2000) tells us that

$$\|\widehat{v}_j - v_j\| \quad \leq \quad C\delta_j \left\| \widehat{\Gamma} - \Gamma \right\|_2$$

where $\delta_1 = 2\sqrt{2}(\lambda_1 - \lambda_2)^{-1}$ and for $j \geq 2$,

$$\delta_j \quad = \quad 2\sqrt{2} \max \left[(\lambda_{j-1} - \lambda_j)^{-1}, (\lambda_j - \lambda_{j+1})^{-1} \right]. \tag{4.15}$$

This concludes the proof. □

Before giving the proof of proposition 4.3.2 let us state the following Lemma
Lemma : Consider the functional $T = \frac{\sum_U Y_k}{N}$. Then $T(M) = T\left(\frac{M}{N}\right)$ and $IT\left(\frac{M}{N}, Y_k\right) = N \cdot IT(M, Y_k)$.
Proof :

$$\begin{aligned} IT\left(\frac{M}{N}, Y_k\right) &= \lim_{h \to 0} \frac{T(M + hN\delta_{Y_k}) - T(M)}{h} = \lim_{h \to 0} \frac{1}{h} \left[\frac{\int \mathcal{Y} d(M + hN\delta_{Y_k})}{\int d(M + hN\delta_{Y_k})} - \frac{\int \mathcal{Y} d(M)}{\int d(M)} \right] \\ &= \lim_{h \to 0} \frac{1}{h+1} \left(y_k - \frac{t_Y}{N} \right) = y_k - \frac{t_Y}{N} = N \cdot IT(M, Y_k) \end{aligned}$$

□

Proof of proposition 4.3.2
Considering first the mean curve μ, we get directly

$$\mu(M + \epsilon\delta y) \quad = \quad \frac{1}{N + \epsilon} \left(\sum_{\ell \in U} Y_\ell + \epsilon y \right) = \mu + \frac{\epsilon}{N} (y - \mu) + o(\epsilon),$$

so that

$$I\mu(M, Y_k) \quad = \quad \frac{1}{N}(Y_k - \mu).$$

Let us first note that perturbation theory (Kato, 1966, Chatelin 1983) allows us to get the influence function of the eigenelements provided the influence function of the covariance operator is known. Indeed, let us consider the following expansion of Γ according to some operator Γ_1,

$$\Gamma(\epsilon) \quad = \quad \Gamma + \epsilon\Gamma_1 + o(\epsilon), \tag{4.16}$$

we get from perturbation theory that the eigenvalues satisfy

$$\lambda_j(\epsilon) \quad = \quad \lambda_j + \epsilon \ \mathrm{tr}\left(\Gamma_1 P_j\right) + o(\epsilon), \tag{4.17}$$

where $P_j = v_j \otimes v_j$ is the projection onto the space spanned by v_j and the trace of an operator Δ defined on $L^2[0,1]$ is defined by $\mathrm{tr}(\Delta) = \sum_j \langle \Delta e_j, e_j \rangle$ for any basis $e_j, j \geq 1$ of $L^2[0,1]$.There exists a similar result for the eigenfunctions which states, provided ϵ is small enough and for simplicity that the non null eigenvalues are distinct, that

$$v_j(\epsilon) \quad = \quad v_j + \epsilon \left(S_j \Gamma_1(v_j)\right) + o(\epsilon), \tag{4.18}$$

where operator S_j is defined on $L^2[0,1]$ as follows

$$S_j \ = \ \sum_{\ell \neq j} \frac{v_\ell \otimes v_\ell}{\lambda_j - \lambda_\ell} \ .$$

So going back to the notion of influence function, if we get an expression for Γ_1 in our case, we will be able to derive the influence function for the eigenelements. The influence function of Γ can be computed directly using the definition,

$$\begin{aligned}
\Gamma(\epsilon) &= \Gamma(M + \epsilon\delta_y) \\
&= \frac{1}{N+\epsilon}\left(\sum_{\ell \in U} (Y_\ell \otimes Y_\ell) + \epsilon(y \otimes y)\right) - \frac{1}{(N+\epsilon)^2}(N\mu + \epsilon y) \otimes (N\mu + \epsilon y) \\
&= \Gamma + \frac{\epsilon}{N}\left(y \otimes y - \mu \otimes \mu - \Gamma\right) - \frac{\epsilon}{N}\left(\mu \otimes (y-\mu) + (y-\mu) \otimes \mu\right) + o(\epsilon) \\
&= \Gamma + \frac{\epsilon}{N}\left((y-\mu) \otimes (y-\mu) - \Gamma\right) + o(\epsilon)
\end{aligned} \tag{4.19}$$

so that

$$I\Gamma(M, Y_k) \quad = \quad \frac{1}{N}\left((Y_k - \mu) \otimes (Y_k - \mu) - \Gamma\right).$$

The combination of (4.17) and (4.19) give us the influence function of the jth eigenvalue

$$I\lambda_j(M, Y_k) \quad = \quad \frac{1}{N}\left(\langle Y_k - \mu, v_j \rangle^2 - \lambda_j\right)$$

as well as the influence function of the jth eigenfunction (since $\langle v_j, v_\ell \rangle = 0$ when $j \neq \ell$)

$$Iv_j(M, Y_k) \quad = \quad \frac{1}{N}\left(\sum_{\ell \neq j} \frac{\langle Y_k - \mu, v_j \rangle \langle Y_k - \mu, v_\ell \rangle}{\lambda_j - \lambda_\ell} v_\ell\right)$$

and the first part of proposition is proved. Now, we have immediately with Lemma above that $IT\left(\frac{M}{N}, Y_k\right) = NIT(M, Y_k)$ for T equal to μ or Γ. As for λ_j, we use (4.17) with $\Gamma_1 = I\Gamma\left(\frac{M}{N}, Y_k\right) = NI\Gamma(M, Y_k)$ which implies that $I\lambda_j\left(\frac{M}{N}, Y_k\right) = NI\lambda_j(M, Y_k)$. We prove in the

same way for v_j using (4.18). □

Proof of proposition 4.3.3
Let us begin with the mean function, we have that $I\mu\left(\frac{M}{N}, Y\right) = NI\mu(M, Y)$ and using (4.8) for $\alpha = 0$, the remainder term is defined as follows

$$R_\mu\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right) = \widehat{\mu} - \mu - \int I\mu(M, Y)d(\widehat{M} - M)$$

and

$$\begin{aligned} R_\mu\left(\tfrac{\widehat{M}}{N}, \tfrac{M}{N}\right) &= \widehat{\mu} - \mu - \frac{1}{N}\sum_{k\in s}\frac{Y_k - \mu}{\pi_k} \\ &= \widehat{\mu}\left(1 - \tfrac{\widehat{N}}{N}\right) + \mu\left(\tfrac{\widehat{N}}{N} - 1\right) \\ &= (\mu - \widehat{\mu})\left(\tfrac{\widehat{N}}{N} - 1\right) \\ &= o_p(n^{-1/2}), \end{aligned}$$

since $\mu - \widehat{\mu} = O_P(n^{-1/2})$ and $(\widehat{N} - N)/N = O_P(n^{-1/2})$ by proposition 4.3.1.

For the covariance operator, we have

$$\begin{aligned} R_\Gamma\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right) &= \widehat{\Gamma} - \Gamma - \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\left((Y_k - \mu)\otimes(Y_k - \mu) - \Gamma\right) \\ &= \Gamma\left(\frac{\widehat{N}}{N} - 1\right) + \widehat{\Gamma} - \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}(Y_k - \mu)\otimes(Y_k - \mu) \\ &= \left(\Gamma - \widehat{\Gamma}\right)\left(\frac{\widehat{N}}{N} - 1\right) - \frac{\widehat{N}}{N}\left((\mu - \widehat{\mu})\otimes(\mu - \widehat{\mu})\right) \\ &= o_p(n^{-1/2}), \end{aligned} \tag{4.20}$$

noticing that

$$\frac{1}{N}\sum_{k\in s}\frac{Y_k \otimes Y_k}{\pi_k} = \frac{\widehat{N}}{N}\left(\widehat{\Gamma} + \widehat{\mu}\otimes\widehat{\mu}\right).$$

To study the remainder terms for the eigenelements, we need to go back to the perturbation theory and equations (4.16), (4.17) and (4.18). According to (4.20), with $\epsilon = n^{-1/2}$, we can write

$$\Gamma_1 = \sqrt{n}\left(\frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\left((Y_k - \mu)\otimes(Y_k - \mu) - \Gamma\right) + R_\Gamma\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right)\right). \tag{4.21}$$

Introducing now (4.21) in equation (4.17), we get noting that $\langle R_\Gamma\left(\frac{\widehat{M}}{N}, \frac{M}{N}\right)v_j, v_j\rangle = o_p(n^{-1/2})$,

$$\begin{aligned} \widehat{\lambda}_j - \lambda_j &= \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\left(\langle Y_k - \mu, v_j\rangle^2 - \langle\Gamma v_j, v_j\rangle\right) + o_p(n^{-1/2}) \\ &= \int I\lambda_j(M, Y)d(\widehat{M} - M) + o_p(n^{-1/2}) \end{aligned}$$

which proves that $R_{\lambda_j}\left(\frac{\widehat{M}}{N},\frac{M}{N}\right)=o_p(n^{-1/2})$. Using now (4.18) and since $S_jR_\Gamma\left(\frac{\widehat{M}}{N},\frac{M}{N}\right)v_j=o_p(n^{-1/2})$, we can check with similar arguments that

$$\begin{aligned}\widehat{v}_j-v_j &= S_j\left(\frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\left(\langle Y_k-\mu,v_j\rangle(Y_k-\mu)-\lambda_jv_j\right)\right)+o_p(n^{-1/2})\\ &= \frac{1}{N}\sum_{k\in s}\frac{1}{\pi_k}\sum_{\ell\neq j}\frac{\langle Y_k-\mu,v_j\rangle\langle Y_k-\mu,v_\ell\rangle}{\lambda_j-\lambda_\ell}v_\ell+o_p(n^{-1/2})\\ &= \int Iv_j(M,Y)d(\widehat{M}-M)+o_p(n^{-1/2})\end{aligned}$$

and the proof is complete. □

Proof of proposition 4.3.4.
We prove the result for functional linearized variables u_k. For real valued linearized variables, for instance for an eigenvalue λ_j, the proof is similar replacing the tensor product with usual product and the norm $||\cdot||_2$ with the absolue value $|\cdot|$. Let us denote by

$$\widehat{AV}(T(\widehat{M}))=\sum_s\sum_s\frac{\Delta_{kl}}{\pi_{kl}}\frac{u_k}{\pi_k}\otimes\frac{u_l}{\pi_l}=\sum_U\sum_U\frac{\Delta_{kl}}{\pi_{kl}}\frac{u_k}{\pi_k}\otimes\frac{u_l}{\pi_l}I_kI_l$$

and by

$$A = \left\|AV(T(\widehat{M}))-\widehat{AV}(T(\widehat{M}))\right\|_2 \quad\text{and}\quad B=\left\|\widehat{AV}(T(\widehat{M}))-\widehat{V}_p(T(\widehat{M}))\right\|_2.$$

It is clear that

$$\left\|AV(T(\widehat{M}))-\widehat{V}_p(T(\widehat{M}))\right\|_2 \leq A+B.$$

We get, under assumptions (A2) and (A3), with similar manipulations as before that it exists a positive constant C_1 such that

$$E_p(A) \leq C_1\sum_{k\in U}\|u_k\|^2.$$

Then under assumption (A1) we can bound

$$\sum_{k\in U}\|u_k\|^2=O(N^{-1})$$

when the u_k are either the linearized variables for the mean function or an eigenfunction, the eigenvalues being distinct. Thus $E_p(A)=O(n^{-1})$.

Let us study now the second term B and examine separately the case of the mean function and the case of the eigenfunctions which can not be dealt with the same way. We can prove, under assumptions (A2) and (A3), with similar manipulations as before that it exists a positive constant C_2 such that

$$E_p(B) \leq C_2E_p\left(\sum_{k\in U}\|u_k\otimes u_k-\widehat{u}_k\otimes\widehat{u}_k\|_2\right).$$

Since the estimated linearized variables satisfy $\|\widehat{u}_k\|^2 \leq C_3/N^2$ for some constant C_3 for the mean function (according to the $L^2[0,1]$ norm) and for the eigenvalues, we easily get that, for some constant C_4,

$$\begin{aligned} E_p(B) &\leq \frac{C_4}{N} E_p\left(\sum_{k\in U} \|u_k - \widehat{u}_k\|\right) \\ &= O(N^{-1}) , \end{aligned}$$

since on can check easily that $E_p\left(\sum_{k\in U} \|u_k - \widehat{u}_k\|\right) = O(1)$.

The technique is different for the eigenfunctions $\widehat{v}_1, \ldots, \widehat{v}_q$ because we can not bound easily terms like $E_p(\widehat{\lambda}_j - \widehat{\lambda}_{j+1})^{-1}$ which appear in the linearized variables. Consider the event

$$E_n = \left\{ \left|\lambda_j - \widehat{\lambda}_j\right| < \delta_j/6, \quad \forall j = 1, \ldots, q \right\}.$$

where the δ_j are defined in (4.15). We have that

$$\begin{aligned} \Pr(E_n^c) &\leq \sum_{\ell=1}^{q} P\left(\left|\lambda_j - \widehat{\lambda}_j\right| > \delta_j/6\right) \\ &\leq 36 \sum_{\ell=1}^{q} E_p \frac{\left|\lambda_j - \widehat{\lambda}_j\right|^2}{\delta_j^2} \\ &\leq 36 E_p \left\|\widehat{\Gamma} - \Gamma\right\|_2^2 \sum_{\ell=1}^{q} \frac{1}{\delta_j^2} \\ &= O\left(\frac{1}{n}\right). \end{aligned}$$

By Cauchy inequality we have

$$B \leq \left(\sum_{k\in U}\sum_{l\in U} \left(\frac{\Delta_{kl} I_k I_l}{\pi_{kl}\pi_k\pi_l}\right)^2\right)^{1/2} \left(\sum_{k\in U}\sum_{l\in U} \|u_k \otimes u_l - \widehat{u}_k \otimes \widehat{u}_l\|_2^2\right)^{1/2} . \tag{4.22}$$

By assumptions (A2) and (A3) we have, for $k \neq l$

$$E_p\left(\frac{\Delta_{kl} I_k I_l}{\pi_{kl}\pi_k\pi_l}\right)^2 = \frac{\Delta_{kl}^2}{\pi_{kl}\pi_k^2\pi_l^2} \leq \frac{C_5}{n^2},$$

for some constant C_5 that does not depend on k and l. When $k = l$, we have $E_p\left(\frac{\Delta_{ll} I_l}{\pi_l^3}\right)^2 \leq C_6$. Thus, by Markov inequality

$$\left(\sum_{k\in U}\sum_{l\in U} \left(\frac{\Delta_{kl} I_k I_l}{\pi_{kl}\pi_k\pi_l}\right)^2\right)^{1/2} = O_p(\sqrt{n}).$$

Considering the second term in (4.22), we have that

$$\begin{aligned} \Pr\left(\sum_{k\in U}\sum_{l\in U} \|u_k \otimes u_l - \widehat{u}_k \otimes \widehat{u}_l\|_2^2 > \epsilon\right) &\leq \Pr\left(\sum_{k\in U}\sum_{l\in U} \|u_k \otimes u_l - \widehat{u}_k \otimes \widehat{u}_l\|_2^2 > \epsilon \mid E_n\right) \\ &\quad + \Pr(E_n^c) \end{aligned}$$

By conditioning on E_n and assumption (A1) we ensure that there exist positive constants C_7 and C_8 such that

$$\begin{aligned} \|\widehat{u}_k\|^2 &\leq C_7 N^{-2} \langle Y_k - \widehat{\mu}, \widehat{v}_j \rangle^2 \left\| \sum_{\ell \neq j} \langle Y_k - \widehat{\mu}, \widehat{v}_\ell \rangle \widehat{v}_\ell \right\|^2 \\ &\leq C_8 N^{-2} \end{aligned}$$

uniformly in k. Consequently, we get under E_n

$$\sum_{k \in U} \sum_{l \in U} \|u_k \otimes u_l - \widehat{u}_k \otimes \widehat{u}_l\|_2^2 \leq C_9 N^{-2}$$

for some positive constant C_9 and $\Pr\left(\sum_{k \in U} \sum_{l \in U} \|u_k \otimes u_l - \widehat{u}_k \otimes \widehat{u}_l\|_2^2 > \epsilon \mid E_n\right) = 0$ for N large enough. This concludes the proof.

□

References

Berger, Y.G, Skinner, C.J (2005). A jacknife variance estimator for unequal probability sampling. *J. R. Statist. Soc B*, **67**, 79-89.

Besse, P.C and Ramsay, J.O. (1986). Principal component analysis of sampled curves. *Psychometrika*, **51**, 285-311.

Besse, P.C., Cardot, H. and Stephenson, D.B. (2000). Autoregressive Forecasting of Some Functional Climatic Variations. *Scand. J. Statist.*, **27**, 673-687.

Bosq, D. (2000). *Linear Processes in Function Spaces.* Lecture Notes in Statistics, 149, Springer.

Breidt, F.J. and Opsomer, J.D. (2000). Local Polynomial Survey Regression Estimators in Survey Sampling. *The Annals of Statistics*, **4**, 1026-1053.

Campbell, C. and Little, A. D. (1980). A Different View of Finite Population Estimation. *Proceeding of the Section on Survey Research Methods*, American Statistical Association. 319-324.

Cardot, H. (2000). Nonparametric estimation of the smoothed principal components analysis of sampled noisy functions. *J. Nonparametr. Stat.*, **12**, 503-538.

Cardot, H., Faivre, R. and Goulard, M. (2003). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *J. of Applied Statistics*, **30**, 1185-1199.

Castro, P., Lawton, W. and Sylvestre, E. (1986). Principal Modes of Variation for Processes with Continuous Sample Curves. *Technometrics*, **28**, 329-337.

Chatelin, F. (1983). *Spectral approximation of linear operators.* Academic Press, New York

Chiky, R., Hébrail, G. (2007). Generic tool for summarizing distributed data streams. *Preprint.*

Chiou, J.M., Müller, H.G., Wang, J.L., Carey, J.R. (2003). A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies. *Statist. Sinica* **13**, 1119-1133.

Croux, C., Ruiz-Gazen, A. (2005). High breakdown estimators for principal components : the projection-pursuit approach revisited. *J. Multivariate Analysis*, **95**, 206-226.

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression : The case of fixed design and functional response. *Canadian Journal of Statistics*, **30**, 285-300.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a random vector function : some applications to statistical inference. *J. Multivariate Anal.*, **12**, 136-154.

Dessertaine A. (2006). Sondage et séries temporelles : une application pour la prévision de la consommation électrique. *38èmes Journées de Statistique*, Clamart, Juin 2006.

Deville, J.C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Ann. Insee*, **15**, 3-104.

Deville, J.C. (1999). Variance estimation for complex statistics and estimators : linearization and residual techniques. *Survey Methodology*, **25**, 193-203.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis, Theory and Applications.* Springer Series in Statistics, Springer, New-York.

Hampel, F. R. (1974). The influence curve and its role in robust statistics. *J. Am. Statist. Ass.*, **69**, 383-393.

Hastie, T. and Mallows, C. (1993). A discussion of "A statistical view of some chemometrics regression tools" by I.E. Frank and J.H. Friedman. *Technometrics*, **35**, 140-143.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Am. Statist. Ass.* **77**, 89-96.

James, G., Hastie, T., and Sugar, C. (2000). Principal Component Models for Sparse Functional Data. *Biometrika*, **87** , 587-602.

Kato, T. (1966). *Perturbation theory for linear operators.* Springer Verlag, Berlin.

Kirkpatrick, M. and Heckman, N. (1989). A quantitative genetic model for growth, shape, reaction norms and other infinite dimensional characters. *J. Math. Biol.*, **27**, 429-450

Kneip, A. and Utikal, K.J. (2001). Inference for Density Families Using Functional Principal Component Analysis. *J. Am. Statist. Ass.*, **96**, 519-542.

Mises, R., v (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statist.*, 18, 309-348.

Ramsay, J. O. and Silverman, B.W. (2002). *Applied Functional Data Analysis : Methods and Case Studies*. Springer-Verlag.

Ramsay, J. O. and Silverman, B.W. (2005). *Functional Data Analysis*. Springer-Verlag, second edition.

Robinson, P.M. and Särndal, C.E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya : The Indian Journal of Statistics*, **45**, 240-248.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons.

Silverman, B.W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, **24**, 1-24.

Skinner, C.J, Holmes, D.J, Smith, T.M.F (1986). The Effect of Sample Design on Principal Components Analysis. *J. Am. Statist. Ass.* **81**, 789-798.

Chapitre 5

On the conditional geometric quantile from dependent observations and its estimation*

Abstract : For fixed $\mathbf{u}$ in $\mathbb{R}^d$, where $\|\mathbf{u}\| \leq 1$, the conditional geometric quantile is defined as the minimiser over $\theta \in \mathbb{R}^d$ of

$$\varphi(\theta, \mathbf{u}, \mathbf{x}) = \int_{\mathbb{R}^d} (\Phi(\mathbf{u}, \mathbf{y} - \theta) - \Phi(\mathbf{u}, \mathbf{y}))F(d\mathbf{y}|\mathbf{x})$$

where $F(.|\mathbf{x})$ is the multivariate conditional distribution function and $\cdot\|$ is the Euclidean norm. From a nonparametric estimate F_n of F, a natural estimate φ_n of φ is defined, which is proved to be uniformly consistent. Under suitable assumptions, the existence and the uniqueness of a minimizer $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ of $\varphi_n(\theta, \mathbf{u}, \mathbf{x})$ are derived. The function $\mathbf{Q}_n(\mathbf{u}|.)$ is proved to be (with probability 1) continuous for n large and the sequence $(\mathbf{Q}_n(\mathbf{u}|.))$ is shown to estimate consistently the function $\mathbf{Q}(\mathbf{u}|.)$, uniformly on compact sets.

Keywords : Conditional geometric quantile, non-parametric estimation, α-mixing process.

Contents

5.1 Introduction 101
5.2 Definitions and notation 102
5.3 Main results 104
5.4 Proofs 105

5.1 Introduction

In the univariate case, quantiles (of a random variable $Y \in \mathbb{R}$) have been received much attention in the literature. On the other hand and due to lack of objective basis for ordering

*Ce chapitre a fait l'objet d'une note écrite par Mohamed Chaouch et soumise aux *C. R. Math. Acad. Sci. Paris.*

multivariate observations, the multivariate quantiles associated to a random vector (variable $\mathbf{Y} \in \mathbb{R}^d$) are less studied than the univariate case.
When the variable of interest Y is concomitant with an explicative variable X, the associated quantiles of Y knowing X are called conditional quantiles. When $\mathbf{Y}$ is a vector, De Gooijer et *al.* (2002) proposed an extension to the conditional case of the definition of the multivariate quantiles proposed by Abdous and Theoderescu (1992). Under certain conditions, they proved the existence and the uniqueness of such quantiles and proposed a nonparametric estimator based on the kernel method. Then, they studied the consistence as well as the asymptotic behavior of this estimator. However, although invariant by the orthogonal transformations, the definition of the multivariate quantiles of De Gooijer et *al.* (2002) is not invariant by affine transformations. To overcome this failure, Gannoun et *al.* (2003) proposed a new approach of the multivariate conditional quantiles. This approach is equivariant by affine transformations and by rotations. It is based on a procedure known as transformation-retransformation inspired from the work of Chakraborty (2001).
In this work, we focus on the definition introduced by Chaudhuri (1996) and investigated by Cheng and De Gooijer (2007) and which relate to the uth geometric conditional quantile (or uth conditional geometric quantile) that will be noted, henceforth, by $\mathbf{Q}(\mathbf{u}|\mathbf{x})$ for $\mathbf{X} = \mathbf{x}$ and $\mathbf{u}$ a unit vector from $B^d = \{\mathbf{u} \in \mathbb{R}^d : ||\mathbf{u}|| < 1\}$. Let $(\mathbf{X}_i, \mathbf{Y}_i),\ i = 1, \ldots, n$ be n independent observations that have the same law than $(\mathbf{X}, \mathbf{Y})$, Cheng and De Gooijer (2007) devoleped a nonparametric estimator of $\mathbf{Q}(\mathbf{u}|\mathbf{x})$. They studied its asymptotic behavior via its Bahadur-type representation. From the asymptotic normality, they developed ellipsoides of confidence around this estimator.
In the following, we are intersting on the estimation of $\mathbf{Q}(\mathbf{u}|\mathbf{x})$ from $\alpha-$mixing observations $(\mathbf{X}_i, \mathbf{Y}_i),\ i = 1, \ldots, n$ which are identically distributed. The paper is organized as follows. Definitions and notations are given in Section 2. The main theoretical result will be exposed in Section 3. Lemmas and proofs are postponed to the final Section 4.

5.2 Definitions and notation

In statistical bibliography we find diffrent attempts to define multidimentional quantile. Eddy (1982, 1983, 1985) proposed an approach for defining quantiles for multivariate data using certain nested sequence of sets, and Brown and Hettmansperger (1987, 1989) introduced a notion of bivariate quantiles based on Oja's criterion function that arises in the definition of Oja's simplex median (see Oja 1983). Two different approaches based on norm minimization, were introduced by Abdous and Theoderescu (1992) and Chaudhuri (1996) to define multivariate quantile as a generalisation of the Ferguson's definition of quantile in the univariate case (see Ferguson, 1967, p.51). Reader can be refered to the perper of Serfling (2002) for a comparision of some approaches defining multivariate quantiles. Throughout this paper, we use the definition of multivariate quantile introduced by Chaudhuri (1996) as

$$\begin{aligned} \mathbf{Q}(\mathbf{u}) &= \arg\min_{\theta \in \mathbb{R}^d} \varphi(\theta, \mathbf{u}) \\ &= \arg\min_{\theta \in \mathbb{R}^d} \int \{\Phi(\mathbf{u}, \mathbf{y} - \theta) - \Phi(\mathbf{u}, \mathbf{y})\} F(d\mathbf{y}) \end{aligned}$$

with $\mathbf{u} = (u_1, u_2, \ldots, u_d)^T$ a vector in the open unit ball $B^d = \{\mathbf{u} \in \mathbb{R}^d : ||\mathbf{u}|| < 1\}$ and $\mathbf{t} \in \mathbb{R}^d$, denote by $\Phi(\mathbf{u}, \mathbf{t}) = ||\mathbf{t}|| + < \mathbf{u}, \mathbf{t} >$, where $||.||$ is the Euclidean norm and $< ., . >$ is the usual Euclidean inner product, and F is the distribution function of $\mathbf{Y}$.

Consider a set of observations $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ in $\mathbb{R}^d$. Let $\mathbf{Q}_n(\mathbf{u})$ be a minimizer (with respect

to θ) of $\sum_{i=1}^{n}(\Phi(\mathbf{u},\mathbf{Y}_i-\theta)-\Phi(\mathbf{u},\mathbf{Y}_i))$. Note that for any fixed $\mathbf{u}\in B^d$, the function $\Phi(\mathbf{u},\mathbf{t})$ tends to infinity as $||\mathbf{t}||$ tends to infinty. Hence $\sum_{i=1}^{n}(\Phi(\mathbf{u},\mathbf{Y}_i-\theta)-\Phi(\mathbf{u},\mathbf{Y}_i))$ must tends to infinity if $||\theta||$ goes to infinity. One must look for a minimizer within a closed and bounded ball around the origin in $\mathbb{R}^d$.
In view of the continuity of $\Phi(\mathbf{u},\mathbf{t})$ as a function of $\mathbf{t}$, which implies the continuity of $\sum_{i=1}^{n}(\Phi(\mathbf{u},\mathbf{Y}_i-\theta)-\Phi(\mathbf{u},\mathbf{Y}_i))$ as a function of θ, there must be a minimizer $\mathbf{Q}_n(\mathbf{u})$ located at a finite distance from the origin in $\mathbb{R}^d$. Next comes the question of uniqueness. Because $\mathbb{R}^d$ equiped with Euclidean norm is a strictly convex Banach space for $d\geq 2$, and $<\mathbf{u},\mathbf{t}>$ is a linear function in $\mathbf{t}$ for every fixed $\mathbf{u}\in B^d$, it follows from theorem 2.17 of Kemperman (1987, p.220) that unless all of the data points $\mathbf{Y}_1,\ldots,\mathbf{Y}_n$ fall on a straight line in $\mathbb{R}^d$, $\sum_{i=1}^{n}(\Phi(\mathbf{u},\mathbf{Y}_i-\theta)-\Phi(\mathbf{u},\mathbf{Y}_i))$ must be strictly convex function of θ. This guarentees the uniqueness of the minimizer $\mathbf{Q}_n(\mathbf{u})$ in $\mathbb{R}^d$ for any $d\geq 2$.

From now on, we suppose that $(\mathbf{X},\mathbf{Y})$, be a random variables defined on a probability space $(\Omega,\mathcal{F},P)$ with values in $\mathbb{R}^s$ x $\mathbb{R}^d$ $(s\geq 1, d\geq 2)$. Let $F(.|\mathbf{x})$ denotes the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$. Then the conditional geometric quantile of $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$ is defined as

$$\begin{aligned}\mathbf{Q}(\mathbf{u}|\mathbf{x}) &= \arg\min_{\theta\in\mathbb{R}^d} E\{\Phi(\mathbf{u},\mathbf{Y}-\theta)-\Phi(\mathbf{u},\mathbf{Y})|\mathbf{X}=\mathbf{x}\}\\ &= \arg\min_{\theta\in\mathbb{R}^d}\int_{\mathbb{R}^d}\{\Phi(\mathbf{u},\mathbf{y}-\theta)-\Phi(\mathbf{u},\mathbf{y})\}F(d\mathbf{y}|\mathbf{x}).\end{aligned}$$

When $||\mathbf{u}||$ is close to one, we may think of $\mathbf{Q}(\mathbf{u}|\mathbf{x})$ as an "extreme" conditional quantile while a "central" conditional quantile corresponds to $||\mathbf{u}||$ being close to zero. Thus $\mathbf{Q}(\mathbf{u}|\mathbf{x})$, as indexed by a directional "outlyingness" parameter $\mathbf{u}$, measures quantitatively through $||\mathbf{u}||$ the extent of deviation from the center of the data cloud formed by $\mathbf{Y}$ given $\mathbf{X}=\mathbf{x}$, see Chaudhuri (1996) and Serfling (2004). If we suppose that

$$\varphi(\theta,\mathbf{u},\mathbf{x})=\int_{\mathbb{R}^d}(\Phi(\mathbf{u},\mathbf{y}-\theta)-\Phi(\mathbf{u},\mathbf{y}))F(d\mathbf{y}|\mathbf{x}),$$

then,

$$\mathbf{Q}(\mathbf{u}|\mathbf{x})=\arg\min_{\theta\in\mathbb{R}^d}\varphi(\theta,\mathbf{u},\mathbf{x}). \tag{5.1}$$

An estimate of the conditional geometric quantile $\mathbf{Q}(\mathbf{u}|\mathbf{x})$ is defined through an estimate of the conditional distribution function. We can choose for instance the nonparametric estimate of the conditional distribution function $F(\cdot|\mathbf{x})$ given, for $\mathbf{y}\in\mathbb{R}^d$, by the Nadaraya-Watson estimate as $F_n(\mathbf{y}|\mathbf{x})=\sum_{i=1}^{n}\omega_{n,i}(\mathbf{x})\mathbb{1}_{\{\mathbf{Y}_i\leq\mathbf{y}\}}$, where the weight functions $\omega_{n,i}(\mathbf{x})$ $(i=1,2,\ldots,n)$ are given by $\omega_{n,i}(\mathbf{x})=\frac{K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}$, where $K(\cdot)$ denotes a probability density function on $\mathbb{R}^s$ (the kernel), $(h_n)_{n\geq 1}$ denotes a sequence of real positive numbers tending to 0 as n increases to infinity, and

$$\mathbb{1}_{\{\mathbf{Y}_i\leq\mathbf{y}\}}=\mathbb{1}_{\{\mathbf{Y}_i^1\leq\mathbf{y}^1\}}\times\cdots\times\mathbb{1}_{\{\mathbf{Y}_i^d\leq\mathbf{y}^d\}},$$

if $\mathbf{y}=\left(y^1,\ldots,y^d\right)\in\mathbb{R}^d$ and $\mathbf{Y}_i=\left(Y_i^1,\ldots,Y_i^d\right)$ for $i\geq 1$.

For $\theta\in\mathbb{R}^d$, let $\varphi_n(\theta,\mathbf{u},\mathbf{x})$ be the estimate of $\varphi(\theta,\mathbf{u},\mathbf{x})$ defined by

$$\begin{aligned}\varphi_n(\theta,\mathbf{u},\mathbf{x}) &= \int_{\mathbb{R}^d}(\Phi(\mathbf{u},\mathbf{y}-\theta)-\Phi(\mathbf{u},\mathbf{y}))F_n(d\mathbf{y}|\mathbf{x}) \\ &= \frac{\sum_{i=1}^{n}(||\mathbf{y}-\theta||-||\mathbf{y}||-<\mathbf{u},\theta>)\,K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^{n}K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}.\end{aligned}$$

From the definition of $\mathbf{Q}(\mathbf{u}|\cdot)$, it is natural to estimate it by minimizing the estimate of $\varphi(\cdot,\mathbf{u},\mathbf{x})$. Denoting by $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ any minimizer of $\varphi_n(\cdot,\mathbf{u},\mathbf{x})$, we will prove, under mild assumptions, that the sequence $(\mathbf{Q}_n(\mathbf{u}|\cdot))_{n\geq 1}$ estimates consistently the conditional geometric quantile $\mathbf{Q}(\mathbf{u}|\cdot)$. Indeed, for n large enough, $\varphi_n(.,\mathbf{u},\mathbf{x})$ has a unique minimum

$$\mathbf{Q}_n(\mathbf{u}|\mathbf{x}) = \arg\min_{\theta\in\mathbb{R}^d}\varphi_n(\theta,\mathbf{u},\mathbf{x}) = \arg\min_{\theta\in\mathbb{R}^d}\sum_{i=1}^{n}(||\mathbf{y}-\theta||-<\mathbf{u},\theta>)\,K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right), \qquad (5.2)$$

where the last equality is obtained by removing terms independent of θ in the expression of $\varphi_n(\theta,\mathbf{u},\mathbf{x})$.
Let $P_n(.|\mathbf{x})$ be the estimate of the probability measure $P(.|\mathbf{x})$, defined for any Borel set V in $\mathbb{R}^d$ by

$$P_n(V|\mathbf{x}) = \int_V F_n(d\mathbf{y}|\mathbf{x}).$$

5.3 Main results

The sudy of consistence of the estimator $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ developed by Cheng and De Gooijer (2007) was made under the hypothesis of independence of the observations $(\mathbf{X}_i,\mathbf{Y}_i)$, $i=1,\ldots,n$. This assumption is rather restrictive in particular in the forecast of the processes. In the following, we introduce a realistic assumption of strong dependence (α-mixing) between the observations.

Definition. The process $\{(\mathbf{X}_i,\mathbf{Y}_i), i\geq 1\}$ is strongly mixing or α-mixing if there exists a sequence of numbers $(\alpha_n)_{n\in\mathbb{N}}$ with $\lim_{n\to\infty}\alpha_n = 0$, such that :

$$\forall n\in\mathbb{N} \quad |P(A\cap B)-P(A)P(B)|\leq\alpha_n$$

where $A\in\mathcal{F}_1^t$ and $B\in\mathcal{F}_{t+n}^{\infty}$ and $\mathcal{F}_\mu^\nu$ is the σ-field generated by the random vectors $\{(\mathbf{X}_t,\mathbf{Y}_t) : \mu\leq t\leq\nu\}$.

Remarks.

- This definition is introduced by Rosenblatt (1956). There are other kind of definitions of mixing, see for example Doukhan (1994).
- if $\alpha_n = 0$ for all n, it is the independent case.

Throughout this paper, $\mathcal{C}$ will denote a fixed compact subset of $\mathbb{R}^d$ on which the density g of $\mathbf{X}$ is lower bounded by some positive constant.
We introduce now the following hypotheses :

(H1) The density g of $\mathbf{X}$ is uniformly continuous.

(H2) The kernel $K:\mathbb{R}^s\to\mathbb{R}$ is continuous, bounded, non negative and satisfying

$$\int K(v)dv = 1,\ \int v_i K(v)dv = 0 \text{ for } i\in\{1,\ldots,s\}, \text{ and } ||v||^s K(v)\to 0 \text{ as } ||v||\to\infty$$

(H3) The process $\{(\mathbf{X}_i, \mathbf{Y}_i), i \geq 1\}$ is α-mixing as defined previously.
(H4) There exists $\delta > 0$ such that

$$n^{1/4}(h_n^s)^{(1+\delta)/4}/logn \to \infty \text{ as } n \to \infty.$$

(H5) For any Borel set $V \subset \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$, the functions $P(V|.)$ and $\varphi(\theta, \mathbf{u}, .)$ are continuous on $\mathcal{C}$.
(H6) The function $\mathbf{Q}(\mathbf{u}|.)$, satisfies a uniform uniqueness property over $\mathcal{C}$

$$\forall\ \varepsilon > 0, \quad \exists \eta > 0, \quad \forall\ t : \mathcal{C} \to \mathbb{R}^d,$$

$$\sup_{\mathbf{x}\in\mathcal{C}} ||\mathbf{Q}(\mathbf{u}|\mathbf{x}) - t(\mathbf{x})|| \geq \varepsilon \Rightarrow \sup_{\mathbf{x}\in\mathcal{C}} |\varphi\left(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}\right) - \varphi\left(t(\mathbf{x}), \mathbf{u}, \mathbf{x}\right)| \geq \eta.$$

Remarks. Assumption (**H1**) and (**H2**) are classical in nonparametric estimation. In (**H3**), the α-mixing assumption is reasonable because many time series models fulfill this condition. Asumption (**H4**) is needed to prove $\sup_{\mathbf{y}\in\mathbb{R}^d} |F_n(\mathbf{y}|\mathbf{x}) - F(\mathbf{y}|\mathbf{x})| \to 0$ under the mixing condition. From Berlinet et *al.* (2001a), in the i.i.d. case, assumption (**H4**) is replaced by $(nh_n^s/logn) \to 0$. Assumption (**H5**) implies the uniform convergence of $\varphi_n(\theta, \mathbf{u}, .)$ to $\varphi(\theta, \mathbf{u}, .)$. Finally, the uniform uniqueness property (**H6**) was introduced by Collomb et *al.* (1987) in order to get consistency of an estimate of the conditional mode.

Theorem 5.3.1 *Assume that Assumptions (**H1**) - (**H6**) are satisfied. Then*
(i) with probability 1, one can find an integer $N > 1$ such that if $n \geq N$ and $\boldsymbol{x} \in \mathcal{C}$, $\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})$ exists and is unique ;
(ii) the function $\boldsymbol{Q}(\boldsymbol{u}|\cdot)$ is continuous on $\mathcal{C}$;
(iii) with probability 1, we have

$$\sup_{\boldsymbol{x}\in\mathcal{C}} ||\,\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x}) - \boldsymbol{Q}(\boldsymbol{u}|\boldsymbol{x})|| \to 0, \quad \textit{if} \quad n \to \infty.$$

5.4 Proofs

5.4.1 Preliminary Results

In order to obtain Theorem 3.1, we have to prove some lemmas which also are of independent interest.

Lemma 5.4.1 *Assume assumptions (**H1**) - (**H5**) holds, then we have*

$$\lim_{||\theta||\to\infty} \sup_{n\geq 1} \sup_{\boldsymbol{x}\in\mathcal{C}} \left| \frac{\varphi_n(\theta, \boldsymbol{u}, \boldsymbol{x}) + < \boldsymbol{u}, \theta >}{||\theta||} - 1 \right| = 0. \tag{5.3}$$

Proof
Let us recall that

$$\begin{aligned} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) &= \int_{\mathbb{R}^d} \left(\Phi(\mathbf{u}, \mathbf{y} - \theta) - \Phi(\mathbf{u}, \mathbf{y})\right) F_n(d\mathbf{y}|\mathbf{x}) \\ &= \int_{\mathbb{R}^d} \left(||\mathbf{y} - \theta|| - ||\mathbf{y}|| - < \mathbf{u}, \theta >\right) F_n(d\mathbf{y}|\mathbf{x}). \end{aligned}$$

Then we have

$$\varphi_n(\theta,\mathbf{u},\mathbf{x})+<\mathbf{u},\theta>=\int_{\mathbb{R}^d}\left(||\mathbf{y}-\theta||-||\mathbf{y}||\right)F_n(d\mathbf{y}|\mathbf{x}),$$

and equation (5.3) leads to

$$\lim_{||\theta||\to\infty}\sup_{n\geq 1}\left|\frac{\int_{\mathbb{R}^d}\left(||\mathbf{y}-\theta||-||\mathbf{y}||\right)F_n(d\mathbf{y}|\mathbf{x})}{||\theta||}-1\right|=0. \tag{5.4}$$

So to proof equation (5.3) we have to proof (5.4).
For all $\mathbf{x}\in\mathcal{C}$ and $p\geq 1$, $\mathbf{x}\longmapsto P(||\mathbf{Y}||>p\mid\mathbf{X}=\mathbf{x})$ is a continuous function (by (**H5**)), one can find $\mathbf{x}_p\in\mathcal{C}$ such that

$$\sup_{\mathbf{x}\in\mathcal{C}}P(||\mathbf{Y}||>p|\mathbf{X}=\mathbf{x})=P(||\mathbf{Y}||>p|\mathbf{X}=\mathbf{x}_p). \tag{5.5}$$

The sequence $(\mathbf{x}_p)_{p\geq 1}$ being in the compact $\mathcal{C}$, one can extract a subsequence $(p_k)_{k\geq 1}$ such that

$$\mathbf{x}_{p_k}\to\mathbf{x}_\infty \quad \text{if} \quad k\to\infty. \tag{5.6}$$

Then, with probability 1 (w.p. 1), for any $\theta\in\mathbb{R}^d-\{0\}$, $\mathbf{x}\in\mathcal{C}$, $n\geq 1$ and $k\geq 1$, we have

$$\begin{aligned}\left|\frac{\varphi_n(\theta,\mathbf{u},\mathbf{x})+<\mathbf{u},\theta>}{||\theta||}-1\right| &\leq \int_{\mathbb{R}^d}\left|\frac{||\mathbf{y}-\theta||-||\mathbf{y}||-||\theta||}{||\theta||}\right|F_n(d\mathbf{y}|\mathbf{x})\\ &\leq \int_{\{||\mathbf{y}||\leq p_k\}}\left|\frac{||\mathbf{y}-\theta||-||\mathbf{y}||-||\theta||}{||\theta||}\right|F_n(d\mathbf{y}|\mathbf{x})\\ &+ \int_{\{||\mathbf{y}||>p_k\}}\left|\frac{||\mathbf{y}-\theta||-||\mathbf{y}||-||\theta||}{||\theta||}\right|F_n(d\mathbf{y}|\mathbf{x}).\end{aligned}$$

Now for all $\theta\in\mathbb{R}^d-\{0\}$ and $\mathbf{y}\in\mathbb{R}^d$:

$$\left|\tfrac{||\mathbf{y}-\theta||-||\mathbf{y}||-||\theta||}{||\theta||}\right|\leq 2\tfrac{||\mathbf{y}||}{||\theta||} \quad \text{and} \quad \left|\tfrac{||\mathbf{y}-\theta||-||\mathbf{y}||-||\theta||}{||\theta||}\right|\leq 2$$

Thus , we get the inequality

$$\begin{aligned}\left|\frac{\varphi_n(\theta,\mathbf{u},\mathbf{x})+<\mathbf{u},\theta>}{||\theta||}-1\right| &\leq \int_{\{||\mathbf{y}||\leq p_k\}}2\,\frac{||\mathbf{y}||}{||\theta||}\,F_n(d\mathbf{y}|\mathbf{x})+\int_{\{||\mathbf{y}||>p_k\}}2\,F_n(d\mathbf{y}|\mathbf{x})\\ &\leq 2\frac{p_k}{||\theta||}+2\int_{\{||\mathbf{y}||>p_k\}}F_n(d\mathbf{y}|\mathbf{x}).\end{aligned}$$

When $||\theta||$ tends to infinity, we obtain that, w.p.1, and for all $k\geq 1$

$$\lim_{||\theta||\to\infty}\sup_{n\geq 1}\sup_{\mathbf{x}\in\mathcal{C}}\left|\frac{\varphi_n(\theta,\mathbf{u},\mathbf{x})+<\mathbf{u},\theta>}{||\theta||}-1\right|\leq 2\sup_{n\geq 1}\sup_{\mathbf{x}\in\mathcal{C}}\int_{\{||\mathbf{y}||>p_k\}}F_n(d\mathbf{y}|\mathbf{x}). \tag{5.7}$$

The last upper bound is now proved to tend to zero as k tends to infinity. If $k,n\geq 1$ and $\mathbf{x}\in\mathcal{C}$, let us denote

$$q_n^{\mathbf{x}}(k) = \int_{\{||\mathbf{y}||>p_k\}} F_n(d\mathbf{y}|\mathbf{x}) = \frac{\sum_{i=1}^n \mathbb{1}_{\{||\mathbf{y}||>p_k\}} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h_n}\right)}.$$

Assuming **(H1)** - **(H5)**, we have, w.p.1, for any $k \geq 1$:

$$\sup_{\mathbf{x}\in\mathcal{C}} |q_n^{\mathbf{x}}(k) - P(||\mathbf{Y}|| > p_k|\mathbf{X} = \mathbf{x})| \to 0, \quad \text{as} \quad n \to \infty, \tag{5.8}$$

(Bosq and Lecoutre, 1987).
Of course, the P-null set where the convergence in (5.8) is not true may be chosen to be independent of k. Moreover, let us note that

$$\left(\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k)\right)_{k\geq 1} \quad \text{and} \quad \left(\sup_{\mathbf{x}\in\mathcal{C}} P(||\mathbf{Y}|| > p_k|\mathbf{X} = \mathbf{x})\right)_{k\geq 1}$$

are decreasing sequences of positive numbers (because the sequence $(p_k)_{k\geq 1}$ is increasing), hence both convergence as $k \to \infty$. Consequently, w.p.1, the convergence in (5.8) is uniform in $k \geq 1$, i.e.

$$\sup_{k\geq 1}\sup_{\mathbf{x}\in\mathcal{C}} |q_n^{\mathbf{x}}(k) - P(||\mathbf{Y}|| > p_k|\mathbf{X} = \mathbf{x})| \to 0, \quad \text{if} \quad n \to \infty.$$

Let $\epsilon > 0$. By the above property, one can find, w.p.1, an integer $N \geq 1$ such that if $n > N$, $k \geq 1$ and $\mathbf{x} \in \mathcal{C}$

$$q_n^{\mathbf{x}}(k) \leq \epsilon + P(||\mathbf{Y}|| > p_k|\mathbf{X} = \mathbf{x}). \tag{5.9}$$

Now, w.p.1, if $k \geq 1$

$$\sup_{n\geq 1}\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k) \leq \sup_{n=1,\ldots,N}\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k) + \sup_{n>N}\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k).$$

On the one hand, by the very definition of $q_n^{\mathbf{x}}(k)$:

$$\sup_{n=1,\ldots,N}\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k) \leq \sup_{n=1,\ldots,N} \mathbb{1}_{\{||\mathbf{Y}_i||>p_k\}}.$$

On the other hand, according to (5.9) :

$$\sup_{n>N}\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k) \leq \epsilon + \sup_{\mathbf{x}\in\mathcal{C}} P(||\mathbf{Y}|| > p_k|\mathbf{X} = \mathbf{x}).$$

But, the Y_i's are P-a.s. finite random variables, so that, with probability 1

$$\limsup_{k\to\infty}\sup_{n\geq 1}\sup_{\mathbf{x}\in\mathcal{C}} q_n^{\mathbf{x}}(k)$$

is upper bounded by

$$\epsilon + \limsup_{k\to\infty} \sup_{\mathbf{x}\in\mathcal{C}} P(||\mathbf{Y}|| > p_k|\mathbf{X}=\mathbf{x}). \tag{5.10}$$

As ϵ is arbitrary, the proof will be completed by showing that the last term in (5.10) is equal to 0.

Let $k \geq 1$. According to (5.5) :

$$\sup_{\mathbf{x}\in\mathcal{C}} P(||\mathbf{Y}|| > p_k|\mathbf{X}=\mathbf{x}) = P(||\mathbf{Y}|| > p_k|\mathbf{X}=\mathbf{x}_{p_k}).$$

Moreover, $\mathbf{x}_{p_k} \to \mathbf{x}_\infty$ if $k \to \infty$, according to (5.6). Now, if $k \geq 1$ and $p' \geq p_k$:

$$P(||\mathbf{Y}|| > p_k|\mathbf{X}=\mathbf{x}_{p_k}) \leq P(||\mathbf{Y}|| > p'|\mathbf{X}=\mathbf{x}_{p_k}),$$

So that if $p' \geq 1$:

$$\limsup_{k\to\infty} P(||\mathbf{Y}|| > p'|\mathbf{X}=\mathbf{x}_{p_k}) = P(||\mathbf{Y}|| > p'|\mathbf{X}=\mathbf{x}_\infty),$$

because $P(||\mathbf{Y}|| > p'|\mathbf{X}=.)$ is a continous function on $\mathcal{C}$ by **(H5)**. Letting $p' \to \infty$, we get

$$\lim_{k\to\infty} \sup_{\mathbf{x}\in\mathcal{C}} P(||\mathbf{Y}|| > p_k|\mathbf{X}=\mathbf{x}) = 0,$$

Finally, according to (5.10), (5.7) and the previous result, we have proved the result of lemma 1.

□

Lemma 5.4.2 *Assume K is a positive probability density. W.p.1, One can find out an integer $N \geq 1$ such that if $n \geq N$ and $\boldsymbol{x} \in \mathcal{C}$, $\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})$ exists and is unique.*

Proof

For a fixed $\mathbf{x} \in \mathcal{C}$, $\varphi_n(\theta, \mathbf{u}, \mathbf{x})$ is bounded by $2||\theta||$.

In fact we have

$$\begin{aligned} |\varphi_n(\theta,\mathbf{u},\mathbf{x})| &= \left| \int_{\mathbb{R}^d} (||\mathbf{y}-\theta|| - ||\mathbf{y}|| - < \mathbf{u}, \theta >) F_n(d\mathbf{y}|\mathbf{x}) \right| \\ &\leq \int_{\mathbb{R}^d} |(||\mathbf{y}-\theta|| - ||\mathbf{y}||)|\, F_n(d\mathbf{y}|\mathbf{x}) + | < \mathbf{u}, \theta > |, \end{aligned}$$

and because

$$|(||\mathbf{y}-\theta|| - ||\mathbf{y}||)| \leq ||\mathbf{y}-\theta-\mathbf{y}|| = ||\theta|| \quad \text{and} \quad | < \mathbf{u}, \theta > | \leq ||\mathbf{u}||\, ||\theta|| \leq ||\theta||$$

then we have

$$|\varphi_n(\theta, \mathbf{u}, \mathbf{x})| \leq 2\ ||\theta||$$

Also due to $||\mathbf{u}|| < \mathbf{1}$, we obtain

$|\varphi_n(\theta_1, \mathbf{u}, \mathbf{x}) - \varphi_n(\theta_2, \mathbf{u}, \mathbf{x})|$

$$\begin{aligned}
&= \left| \int_{\mathbb{R}^d} \{(||\mathbf{y} - \theta_1|| - ||\mathbf{y}|| - < \mathbf{u}, \theta_1 >) - (||\mathbf{y} - \theta_2|| - ||\mathbf{y}|| - < \mathbf{u}, \theta_2 >)\} F_n(d\mathbf{y}|\mathbf{x}) \right| \\
&= \left| \int_{\mathbb{R}^d} \{(||\mathbf{y} - \theta_1|| - ||\mathbf{y} - \theta_2||) - < \mathbf{u}, \theta_1 - \theta_2 >\} F_n(d\mathbf{y}|\mathbf{x}) \right| \\
&= \left| \int_{\mathbb{R}^d} \{(||\mathbf{y} - \theta_2 - (\theta_1 - \theta_2)|| - ||\mathbf{y} - \theta_2||) - < \mathbf{u}, \theta_1 - \theta_2 >\} F_n(d\mathbf{y}|\mathbf{x}) \right| \\
&= \left| \int_{\mathbb{R}^d} \varphi_n(\theta_1 - \theta_2, \mathbf{u}, \mathbf{x}) F_n(d\mathbf{y}|\mathbf{x}) \right| \\
&< 2||\theta_1 - \theta_2||.
\end{aligned}$$

In one hand, from the last equation, we deduce that $\varphi_n(\theta, \mathbf{u}, \mathbf{x})$ is Lipschizian, so it is continuous. Also it is convex because it is the integrand (with respect to $P_n(.|\mathbf{x})$) of a convex function. In the other hand, from Berlinet et *al.* (2001b), $P_n(.|\mathbf{x})$ is not carried by a straight line in $\mathbb{R}^d$. Thus, from Kemperman (1987, p. 220), we can prove that $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ exists and is unique.

□

Lemma 5.4.3 *Assume K is a continuous positive kernel, then the function $\boldsymbol{Q}_n(\boldsymbol{u}|\boldsymbol{x})$ is continuous on $\mathcal{C}$.*

Proof According to Lemme 5.4.2, w.p.1, we can find $N \geq 1$ such that if $n \geq N$ and $\mathbf{x} \in \mathcal{C}$, $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ exists and is unique. Let $\mathbf{x} \in \mathcal{C}$ and $(\mathbf{x}_k)_{k\geq 1}$ be a sequence such that $\mathbf{x}_k \to \mathbf{x}$, if $k \to \infty$.

By the continuity of K, we can get that the sequence of probability measures $(P_n(.|\mathbf{x}_k))_{k\geq 1}$ converge weakly to the measure $P_n(.|\mathbf{x})$. Because we already know the measure $P_n(.|\mathbf{x})$ is not supported by any straight line, according to corollary 2.26 of Kemperman (1987), we obtain $\mathbf{Q}_n(\mathbf{u}|\mathbf{x}_k) \to \mathbf{Q}_n(\mathbf{u}|\mathbf{x})$, if $k \to \infty$. Hence w.p.1, if $n \geq N$, $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$ is continuous on $\mathcal{C}$. □

Lemma 5.4.4 *Let us assume that **(H1)** - **(H5)** holds. For any $A > 0$, we have w.p.1*

$$\sup_{||\theta|| \leq A} \sup_{\boldsymbol{x} \in \mathcal{C}} |\varphi_n(\theta, \boldsymbol{u}, \boldsymbol{x}) - \varphi(\theta, \boldsymbol{u}, \boldsymbol{x})| \to 0 \quad \textit{if} \quad n \to \infty$$

Proof We have proved previously that because $||\mathbf{u}|| < 1$:

$$|||\mathbf{y} - \theta|| - ||\mathbf{y}|| - < \mathbf{u}, \theta >| < 2\ ||\theta||.$$

and we know that

$$\begin{aligned}
\varphi_n(\theta, \mathbf{u}, \mathbf{x}) &= \textstyle\int_{\mathbb{R}^d} \{||\mathbf{y} - \theta|| - ||\mathbf{y}|| - < \mathbf{u}, \theta >\} F_n(d\mathbf{y}|\mathbf{x}) \\
\varphi(\theta, \mathbf{u}, \mathbf{x}) &= \textstyle\int_{\mathbb{R}^d} \{||\mathbf{y} - \theta|| - ||\mathbf{y}|| - < \mathbf{u}, \theta >\} F(d\mathbf{y}|\mathbf{x}).
\end{aligned}$$

Consequently, assuming **(H1)** - **(H5)**, Berlinet et *al.* (2001c) proved that

$$\sup_{\mathbf{x} \in \mathcal{C}} \sup_{\mathbf{y} \in \mathbb{R}^d} |F_n(\mathbf{y}|\mathbf{x}) - F(\mathbf{y}|\mathbf{x})| \to 0, \quad \text{if} \quad n \to \infty,$$

then according to Bosq and Lecoutre (1987), we obtain that for all $\theta \in \mathbb{R}^d$ and w.p.1,

$$\sup_{\mathbf{x}\in\mathcal{C}} |\varphi_n(\theta, \mathbf{u}, \mathbf{x}) - \varphi(\theta, \mathbf{u}, \mathbf{x})| \to 0, \quad \text{if} \quad n \to \infty \tag{5.11}$$

But, w.p.1, if $n \geq 1,\ \mathbf{x} \in \mathcal{C}$ and $\theta_1, \theta_2 \in \mathbb{R}^d$:

$$|\varphi(\theta_1, \mathbf{u}, \mathbf{x}) - \varphi(\theta_2, \mathbf{u}, \mathbf{x})| < 2\ ||\theta_1 - \theta_2||$$

and

$$|\varphi_n(\theta_1, \mathbf{u}, \mathbf{x}) - \varphi_n(\theta_2, \mathbf{u}, \mathbf{x})| < 2\ ||\theta_1 - \theta_2||.$$

Moreover, according to the last inequality and w.p.1, the sequence of functions $(\varphi_n(., \mathbf{u}, \mathbf{x}), n \geq 1)$ is equicontinuous, and this property is independent of $\mathbf{x} \in \mathcal{C}$. Thus according to (5.11) and the Ascoli's theorem, we get that, w.p.1, if $A > 0$:

$$\sup_{||\theta||\leq A} \sup_{\mathbf{x}\in\mathcal{C}} |\varphi_n(\theta, \mathbf{u}, \mathbf{x}) - \varphi(\theta, \mathbf{u}, \mathbf{x})| \to 0, \quad \text{if} \quad n \to \infty.$$

This achived the proof of Lemma 5.4.4.

□

Proof of Theorem
Assertion (i) follows from Lemmas 5.4.2 and 5.4.3. Moreover, (ii) is a straightforward consequence of (i) and (iii). One only needs to prove (iii). The proof is divided into two steps.
step1 : We want to prove that one can find $r > 0,\ N \geq 1$ such that

$$\sup_{n\geq N} \sup_{\mathbf{x}\in\mathcal{C}} ||\mathbf{Q}_n(\mathbf{u}|\mathbf{x})|| \leq r, \quad \text{and} \quad \sup_{\mathbf{x}\in\mathcal{C}} ||\mathbf{Q}(\mathbf{u}|\mathbf{x})|| \leq r.$$

From Lemma 5.4.1, we know $\forall\ \epsilon > 0,\ \exists\ r_1$, such that if $||\mathbf{Q}|| > r_1,\ \forall\ n \leq 1$ and $\forall\ \mathbf{x} \in \mathcal{C}$:

$$\frac{\varphi_n(\theta, \mathbf{u}, \mathbf{x}) + < \mathbf{u}, \theta >}{||\theta||} > 1 - \epsilon.$$

Then,

$$\begin{aligned} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) &> (1 - \epsilon)\ ||\theta|| - < \mathbf{u}, \theta > \\ &> (1 - \epsilon)||\theta|| - ||\mathbf{u}||\ ||\theta|| \\ &> (1 - \epsilon - ||\mathbf{u}||)\ ||\theta||. \end{aligned}$$

Because $||\mathbf{u}|| < 1$, always $\exists\ \delta > 0$, such that when $0 < \epsilon < \delta$, we have

$$\varphi_n(\theta, \mathbf{u}, \mathbf{x}) > 0.$$

Assuming that $||\mathbf{Q}_n(\mathbf{u}|\mathbf{x})|| > r_1$, then we have

$$\varphi_n(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) > 0. \tag{5.12}$$

However, by the definition of $\mathbf{Q}_n(\mathbf{u}|\mathbf{x})$,

$$\begin{aligned} \varphi_n(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) &= \inf_{\theta\in\mathbb{R}^d} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) \\ &\leq \varphi_n(0, \mathbf{u}, \mathbf{x}) = 0. \end{aligned} \tag{5.13}$$

From (5.12) and (5.13), we have a contradiction. Hence, w.p.1, for all $n \geq N$,

$$||\mathbf{Q}_n(\mathbf{u}|\mathbf{x})|| \leq r_1.$$

Similar arguments leads to

$$||\mathbf{Q}(\mathbf{u}|\mathbf{x})|| \leq r_2.$$

Now, the desired result is obtained with $r = \max(r_1, r_2)$.

step2 : According to the triangle inequality, w.p.1, if $n \geq N$ (N was fixed in step 1), we will have

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} |\varphi(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) - \varphi(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x})| &\leq \sup_{\mathbf{x}\in\mathcal{C}} |\varphi(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) - \varphi_n(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x})| \\ &+ \sup_{\mathbf{x}\in\mathcal{C}} |\varphi_n(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) - \varphi(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x})|.\end{aligned}$$

From step1, we obtain $\sup_{\mathbf{x}\in\mathcal{C}} ||\mathbf{Q}_n(\mathbf{u}|\mathbf{x})|| \leq r, \sup_{\mathbf{x}\in\mathcal{C}} ||\mathbf{Q}(\mathbf{u}|\mathbf{x})|| \leq r$.

Then

$$\varphi(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) = \inf_{\theta\in\mathbb{R}^d} \varphi(\theta, \mathbf{u}, \mathbf{x}) = \inf_{||\theta||\leq r} \varphi(\theta, \mathbf{u}, \mathbf{x}),$$

$$\varphi_n(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) = \inf_{\theta\in\mathbb{R}^d} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) = \inf_{||\theta||\leq r} \varphi_n(\theta, \mathbf{u}, \mathbf{x})$$

Thus w.p.1, if $n \geq N$, we have

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} |\varphi(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) - \varphi(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x})| &\leq \sup_{\mathbf{x}\in\mathcal{C}} \left| \inf_{||\theta||\leq r} \varphi(\theta, \mathbf{u}, \mathbf{x}) - \inf_{||\theta||\leq r} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) \right| \\ &+ \sup_{||\theta||\leq r} \sup_{\mathbf{x}\in\mathcal{C}} |\varphi_n(\theta, \mathbf{u}, \mathbf{x}) - \varphi(\theta, \mathbf{u}, \mathbf{x})|.\end{aligned} \tag{5.14}$$

Because

$$\begin{aligned}\sup_{\mathbf{x}\in\mathcal{C}} \left| \inf_{||\theta||\leq r} \varphi(\theta, \mathbf{u}, \mathbf{x}) - \inf_{||\theta||\leq r} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) \right| &= \sup_{\mathbf{x}\in\mathcal{C}} \left| \sup_{||\theta||\leq r} (-\varphi(\theta, \mathbf{u}, \mathbf{x})) - \sup_{||\theta||\leq r} (-\varphi_n(\theta, \mathbf{u}, \mathbf{x})) \right| \\ &= \sup_{\mathbf{x}\in\mathcal{C}} \left| \sup_{||\theta||\leq r} \varphi_n(\theta, \mathbf{u}, \mathbf{x}) - \sup_{||\theta||\leq r} \varphi(\theta, \mathbf{u}, \mathbf{x}) \right| \\ &\leq \sup_{||\theta||\leq r} \sup_{\mathbf{x}\in\mathcal{C}} |\varphi_n(\theta, \mathbf{u}, \mathbf{x}) - \varphi(\theta, \mathbf{u}, \mathbf{x})|,\end{aligned} \tag{5.15}$$

from (5.14) and (5.15), we obtain :

$$\sup_{\mathbf{x}\in\mathcal{C}} |\varphi(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) - \varphi(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x})| \leq 2 \sup_{||\theta||\leq r} \sup_{\mathbf{x}\in\mathcal{C}} |\varphi_n(\theta, \mathbf{u}, \mathbf{x}) - \varphi(\theta, \mathbf{u}, \mathbf{x})|.$$

Then based on the assumptions **(H1)** - **(H5)** and the result of Lemma 5.4.4, we obtain

$$\sup_{\mathbf{x}\in\mathcal{C}} |\varphi(\mathbf{Q}(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x}) - \varphi(\mathbf{Q}_n(\mathbf{u}|\mathbf{x}), \mathbf{u}, \mathbf{x})| \to 0, \quad \text{if} \quad n \to \infty.$$

We applt now assumption **(H6)**, to get

$$\sup_{\mathbf{x}\in\mathcal{C}} |\mathbf{Q}(\mathbf{u}|\mathbf{x}) - \mathbf{Q}_n(\mathbf{u}|\mathbf{x})| \to 0, \quad \text{if} \quad n \to \infty.$$

□

References

Abdous, B. and Theodorescu, R. (1992). Note on the geometric quantile of a random vector. *Statistics and Probability Letter*, **13**, 333-336.

Berlinet, A., Cadre, B. and Gannoun, A. (2001a). On the conditional L_1-median and its estimation. *Journal of Nonparametric Statistics*, **13**, 631-645.

Berlinet, A., Cadre, B. and Gannoun, A. (2001b). Estimation of conditional L_1-median from dependent observations. *Statistics & Probability Letters*, **55**, 353-358.

Berlinet, A., Gannoun, A., Matzner-L£ber, E. (2001c). Asymptotic normality of convergent estimates of conditional quantiles. *Statistics*, 35, 139-169.

Bosq, D. and Lecoutre, J.P. (1987). Théorie de l'estimation fonctionnelle. Economica.

Brown, B. M., and Hettmansperger, T. P. (1987). Affine Invariant Rank Methods in the Bivariate Location Model. *Journal of the Royal Statistical Society, Ser. B*, **49**, 301-310.

Brown, B. M., and Hettmansperger, T. P. (1989). An Affine Invariant Bivariate Versions of the Sign Test. *Journal of the Royal Statistical Society, Ser. B*, **51**, 117-125.

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

Chaouch, M., Gannoun, A. and Saracco, J. (2007). Quantile geometric conditionnel et non conditionnel : une méthode d'estimation et son implementation en R. *submited*.

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

Cheng, Y. and De Gooijer, J. (2007). On the uth geometric conditional quantile. *Journal of statistical planning and inference*, **137**, 1914-1930.

Collomb, G., HŁrdle, W. and Hassani, S. (1987). A note on prediction via estimation of the conditional mode function. *Journal of Statistical Planning and Inference*, **15**, 227-236.

De Gooijer, J., Gannoun, A. and Zerom, D. (2002). A multivariate quantile predictor. *Communications in Statistics - Theory and Methods*, **35**, 133-147.

Doukhan, P. (1994). Mixing : Properties and Examples. *Lecture Notes in Statistics*, **85**, Springer, New York.

Eddy, W.F. (1982). Convex Hull Peeling. *COMPSTAT 1982 for IASC*, Vienna : Pysica-Verlag, 42-47.

Eddy, W.F. (1983). Set Valued Ordering of Bivariate Data. *Stochastics Geometry, Geometric Statistics and Stereology*, eds. R. V. Ambartsumian and W. Weil, Leipzig : Tuebner, 79-90.

Eddy, W.F. (1985). Ordering of Multivariate Data. *Computer Science and Statistics : The Interface*, ed. L. Billard, Amesterdam : North-Holland, 25-30.

Ferguson, T. (1967). *Mathematical Statistics : A Decision Theoric Approach*. Academic Press, New York.

Gannoun, A., Saracco, J., Yan, A. and Bonney, G.E. (2003). On adaptive transformation-retransformation estimate of conditional geometric median. *Communications in Statistics - Theory and Methods*, **32**, 1981-2011.

Kemperman, J. H. B. (1987). The median of a finite measure on a Banach space. In *Statistical Data Analysis based on the L_1-norm and related methods,* Y. Dodge (ed), North-Holland, Amesterdam, 217-230.

Oja, H. (1983). Descriptive Statistics for Multivariate Trimming. *Statistics and Probability Letters*, **1**, 327-332.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.*, **42**, 43-47.

Serfling, R. (2002). Quantile Functions for Multivariate Analysis : Approaches and applications. *Statistica Neerlandica*, **56**, 214-232.

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on geometric quantiles. *Journal of statistical Planning and Inference*, **123**, 259-278.

Annexes

Ce chapitre contient les programmes qui ont été décrits dans le chapitre 2 ainsi que ceux de l'ACP fonctionnelle. Ces programmes ont été réalisés avec le logiciel R et ils sont présentés en deux parties. Dans la première partie, nous donnons les programmes permettant d'estimer les quantiles géométriques et les quantiles géométriques conditionnels. La partie 1 est composée de deux sections. Dans la première section, on se place dans le cas d'une distribution sphérique et on donne deux programmes, le premier permet d'estimer le quantile géométrique indexé par un vecteur u fixé et le deuxième permet de calculer le quantile géométrique conditionnel indexé par un vecteur u pour un x fixé. Dans la deuxième section, on s'intéresse au cas d'une distribution non sphérique et on donne le programme permettant d'effectuer l'étape de Transformation-Retransformation pour estimer les quantiles géométriques conditionnels et non conditionnels. La partie 2 est destinée à décrire les programmes relatifs à l'ACPF et sondage.

Partie I : Programmes des quantiles géométriques ou non

Cas de l'estimation sans une étape TR

Les deux programmes suivant permettent de calculer le quantile géométrique conditionnel ou non conditionnel sans passer par une étape de Transformation-Retransformation.

```
#====================================================================
##################  Quantile NON conditionnel ########################
#====================================================================
   QuantileNC.est = function(MatY, u, m)
{
   d=ncol(MatY)
   n=nrow(MatY)
   Q = matrix(NA,nrow=d,ncol=1)
   test = FALSE
   Qm = matrix(apply(MatY,2,median),nrow=d,ncol=1)
   Id = diag(1,d)
   matQ = matrix(NA,nrow=d,ncol=m)
   u=matrix(u,ncol=1)

 for (i in 1:n){
 somme = matrix (0,nrow=d,ncol=1)
 for (j in 1:n){
  if (i!=j && norme((MatY[j,] - MatY[i,]))!=0)
```

```
   {somme = somme + ((MatY[j,] - MatY[i,])/norme((MatY[j,] - MatY[i,])))}}
   normexpr = norme((somme + (n-1)*u))
    if (normexpr <= (1+norme(u)))
    {Q = MatY[i,]
     test = TRUE
     break}}

   for(j in 1:m){
                  phi = matrix(0,nrow=d,ncol=d)
                  delta = matrix(0,nrow=d,ncol=1)
                  Sdelta = matrix(0,nrow=d,ncol=1)
         for(i in 1:n){
          if (norme((MatY[i,]-Qm)) != 0){
          Sdelta = Sdelta +((MatY[i,]-Qm)/norme((MatY[i,]-Qm)))
          phi=phi+(Id-((MatY[i,]-Qm)%*%t(MatY[i,]-Qm)/(norme(MatY[i,]-Qm)^2)))/norme(MatY[i,]-Qm)

   }}

   delta = Sdelta + n*u
   Qm = Qm + (solve(phi))%*%delta
   matQ[,j] = Qm
}

 if (test == TRUE) { list(Q = Q, u = u, test= test, matQ=matQ)}
 else { list(Q = matQ[,m], u = u, test= test,matQ=matQ)}

}
}
```

Le programme suivant permet de calculer le quantile géométrique conditionnel $Q_n(u|x)$, pour u et x fixés.

```
#===============================================================
############## Quantiles conditionnels ######################
#===============================================================



QuantileC.est = function (MatXY, x, u, m,hn = nrow(MatXY)^(-0.2))   {

    Z = matrix((x - MatXY[,3])/hn,nrow = nrow(MatXY), ncol = 1)
    MatY = MatXY[,1:2]
    k = exp(-0.5*(Z^2))/sqrt(2*pi)        #  noyau Gaussien
    Q = matrix(NA, nrow = ncol(MatY), ncol = 1)
    test = FALSE
     Qm = matrix(apply(MatY, 2, median), nrow = ncol(MatY), ncol = 1)
     Id = diag(1, ncol(MatY))
     matQ = matrix(NA, nrow = ncol(MatY), ncol = m)
     for( i in 1:nrow(MatY)){
```

```
sommeg = matrix(0, nrow = ncol(MatY), ncol = 1)
for( j in 1:nrow(MatY)){
if((i != j) && (norme((MatY[j,] - MatY[i,])) != 0) ){
sommeg = sommeg + ((MatY[j,]-MatY[i,])/norme((MatY[j,]-MatY[i,])) + u) *k[j,]  }  }
nsommeg = norme(sommeg)
 if(nsommeg <= (k[i,]*(1+norme(u)))){test = TRUE
 Q = MatY[i,]
 break}}
 if(test == FALSE){
 for(j in 1:m){
 phi = matrix(0, nrow = ncol(MatY), ncol = ncol(MatY))
 Sdelta = matrix(0, nrow = ncol(MatY), ncol = 1)
 for( i in 1:nrow(MatY)){
 if ((norme(MatY[i,] - Qm)) != 0 )  {
 Sdelta = Sdelta + (((MatY[i,] - Qm)/norme(MatY[i,] - Qm)) + u)*k[i,]
 phi = phi + 1/(norme(MatY[i,]- Qm)) * (Id - (MatY[i,] - Qm)%*%
 t((MatY[i,] - Qm))/((norme(MatY[i,] - Qm))^2))*k[i,]}}
 s = svd(phi)
 D = round(diag(s$d),3)
 invD = D
 for(i in 1:nrow(D)){if (D[i,i]== 0.000){invD[i,i] = 0} else {invD[i,i] = 1/D[i,i]} }
 U = s$u
 V = s$v
 invphi = V%*%invD%*%t(U)
 Qm = Qm + invphi%*%Sdelta
 matQ[,j] = Qm
 }}
 if(test == TRUE) {list(Q = Q,matQ=matQ)}
 else {Q = matQ[,m]}
 list(Q = Q, test=test, matQ=matQ,k=k)

}
```

Cas de l'estimation avec une étape TR

Ces programmes permettent de calculer le quantile géométrique conditionnel et non conditionnel aprés une étape de Transformation-Retransformation lorsque la distribution est loin du cadre sphérique.

```
#=====================================================================
######## TR version pour les quantiles conditionnels ou non conditionnels
#=====================================================================
ChoixIndice=function(matY,epsilon){
d=ncol(matY)
n=nrow(matY)
Sigma=var(matY)
continue=T
while (continue==T){
indice=sample(1:n,size=(d+1))
```

```
matYalpha=matY[indice,]
Yalpha=cbind(matYalpha[2,]-matYalpha[1,],matYalpha[3,]-matYalpha[2,])
M=t(Yalpha)%*%solve(Sigma)%*%Yalpha
moy.arith=sum(diag(M))/d
moy.geo=sqrt(det(M))
ratio=moy.arith/moy.geo
if (ratio<1+epsilon){continue=F}
}
list(indice=indice,ratio=ratio)
}


TRversion.QuantileNC.est = function(MatY, u, m,indice){
d=ncol(MatY)
n=nrow(MatY)
matY0=MatY[indice,]
Y0=cbind(matY0[2,]-matY0[1,],matY0[3,]-matY0[2,])
matZ=solve(Y0)%*%t(MatY[-indice,])
MatZ=t(matZ)
res=QuantileNC.est(MatZ,u,m)
Qn.Z=res$Q
Qn.Y=Y0%*%matrix(Qn.Z,ncol=1)
list(Q=Qn.Y,matQ=res$matQ)
}

TRversion.QuantileC.est = function(MatXY,x,u,m,h,indice){
matY0=MatXY[indice,-1]
Y0=cbind(matY0[2,]-matY0[1,],matY0[3,]-matY0[2,])
matZ=solve(Y0)%*%t(MatXY[-indice,-1])
MatZ=t(matZ)
res=QuantileC.est(cbind(MatXY[-indice,1],MatZ),x,u,m,h)
Qn.Z=res$Q
Qn.Y=Y0%*%matrix(Qn.Z,ncol=1)
list(Q=Qn.Y,matQ=res$matQ,k=res$k)
}
```

Partie II : Programmes de l'ACPF et sondage

Le programme suivante donne les différents paramètres de l'ACPF, à savoir la fonction moyenne, les valeurs propres, les fonctions propres et les composantes principales.

```
### Sondages et ACP#######
acpf <-  function(X)
{
############################
# principal components analysis
### ARGUMENTS
# X: matrix (size nxp) of functional  variables (p=number of design points)
```

```
### VALUES
# valp: valeurs propres de l'operateur de covariance
# vecp: vecteurs propres de l'operateur de covariance
# compP = composantes principales
############################

     N <- nrow(X)
     p <- ncol(X)
     mu = apply(X,2,mean)
     X.cent <- sweep(X,2,mu)
     Gamma <- crossprod(X.cent)/(N*p)
     Vk <- eigen(Gamma, s = T)
     valp <- Vk$values
     vecp <- sqrt(p)*Vk$vectors         ## on les normalise a 1 selon la norme "discrete"
     CPk <- X.cent %*% as.matrix(vecp)/p
     list(mu=mu,Gamma=Gamma,valp = valp, vecp=vecp, compP = CPk, X.cent = X.cent, N.est = N)
}
```

Ce programme permet d'estimer les paramètres de l'ACPF lorsqu'on dispose des poids de sondage.

```
acp.survey = function(Xsample,pik){
### ACP avec des poids de sondage
## ENTREES
## Xsample matrice de taille n*p ou n =taille de l'echantillon
## pik = vecteur des probas d'inclusion des lignes de X
### SORTIES
## N.est = taille estimee de la pop.
## mu.est = courbe moyenne estimee
## Gamma.est = matrice de variance estimee
## vecp = matrice des vecteurs (fonctions) propres
## valp = vecteur des valeurs propres

  p = ncol(Xsample)
pi.inv = 1/pik
N.est = sum(pi.inv)
mu.est = t(pi.inv)%*%Xsample/(N.est)
X.cent = sweep(Xsample,2,mu.est)
Gamma.est = t(X.cent)%*%diag(pi.inv)%*%X.cent/(N.est*p)
eig.gamma = eigen(Gamma.est,symmetric=TRUE)
vecp=sqrt(p)*eig.gamma$vectors
valp=eig.gamma$values
list(N.est=N.est,Gamma=Gamma.est,mu=mu.est, vecp=vecp,valp=valp, X.cent = X.cent)
}
```

Le programme suivant calcule les fonctions d'influence de chacun des paramètres d'intérêt.

```
#Calcul des fonctions d'influence des estimateurs de la moyenne,
#valeurs propre, vecteur propre

influence.acp = function(res.acp,nbv=10){
### calcul des fonctions d'influence de la moyenne et des elements propres
### ENTREE
### une sortie de acp.survey
### SORTIE
### les variables lineaisees : Imu, IGamma,
### Ilambda (premiere et deuxieme valeur propre)
### et Ivecp (premier et second vecteur propre)
### Variables linearisees

 N.est = res.acp$N.est
 X.cent = res.acp$X.cent
 p = ncol(X.cent)
 n = nrow(X.cent)
 Imu = X.cent/N.est

 Gamma = res.acp$Gamma
 IGamma = array(NA,c(p,p,n))
 for ( i in 1:n){
 IGamma[,,i] = (X.cent[i,]%*%t(X.cent[i,]) - Gamma)/N.est
}

 vecp = res.acp$vecp
 vecp = vecp[,1:nbv]
 CompPrinc = X.cent%*%vecp/p
 valp = res.acp$valp
 valp = valp[1:nbv]

 Ilambda1.est = ((CompPrinc[,1])^2 - valp[1])/N.est
 Ilambda2.est = ((CompPrinc[,2])^2 - valp[2])/N.est

 diff.valp1 = valp[1] - valp[-1]
 diff.valp2 = valp[2] - valp[-2]
 Ivecp1 = (CompPrinc[,-1]*matrix(rep(CompPrinc[,1],nbv-1),ncol=(nbv-1)))%*%
 t((vecp[,-1]%*%diag(1/diff.valp1)))

list(Imu = Imu , IGamma = IGamma, Ilambda1=Ilambda1.est, Ilambda2=Ilambda2.est,Ivecp1=Ivecp1/N.est)
}
```

Ceci est le programme permattant d'estimer la variance asymptotique d'un paramètre donné de l'ACPF.

```
varest.sas = function(res.influence,pik){
##calcul de la variance d'estimation ds un SAS
### ENTRE
### une sortie de la fonction influence.acp
### SORTIE
```

```
### la variance estimée
  n = length(pik)
  pi.inv = 1/pik
  N.est = sum(pi.inv)
  f=n/N.est
  res.influence.cent = sweep(res.influence,2,apply(res.influence,2,mean))
s2 = t(res.influence.cent) %*% res.influence.cent/(n-1)
  var.est = N.est^2*(1-f)*s2/n
  return(var.est)
  }
```

Le programme suivant détermine la taille optimale de chaque échantillon à tirer de chacune des deux strates.

```
tailleOpt = function(n,Npop,sigma1,sigma2){
### SORTIE
### la taille optimale de chaque echantillon ds chaque starte
  f = n/Npop
  #N2 = Npop-N1
  sigmabar = ((N1*sigma1)+ (N2*sigma2))/Npop
  n1 = round(N1*f*sigma1/sigmabar)
  n2 = round(N2*f*sigma2/sigmabar)
  list(n1 = n1,n2 = n2)
  }
```

Le programme suivant calcule la norme d'une matrice

```
matnorm = function(x)
 {
### calcul de la norme d'une matrice

 normX = sum(diag(crossprod(x)))
 list(normX = normX)
 }
```

Ceci est le programme de simulation d'un mouvement Brownien.

```
SimulBrownien = function(N,p)
{
### Simulation Mouvement Browien
### n trajectoires sur [0,1]
### p points de discretisation equidistants
  X = rnorm(N*p)/sqrt(p)
  Xmat = matrix(X,nrow=N)
  B = matrix(0,nrow=N,ncol=p)
    for (i in 1:N)
```

```
    {
      B[i,] = cumsum(Xmat[i,])
    }
  B
}
```

Programme permettant d'estimer les différent paramétres de l'ACPF à partir d'un plan de Sondage Aléatoire Simple (SAS).

```
######################### PLAN SAS ############################

SAS = function(Xpop,Npop,n,p,n.sim){
### ENTREE
### Xpop une matrice Npop ligne p colonnes: les lignes sont les indivudus (courbe) et les colonne
### sont les point de dÕscritisation de mes courbes
### n la taille de l'Õchantillon Ĺ tirer de la popuation
### n.sim est le nombre de simulation
### SORTIE
### mu.est l'estimateur de la fonction moyenne
### vecp1.varest l'estimateur de la variance du premier vecteur propre
### valp.est les estimateurs des n valeurs propres
### valp.varest l'estimateur de la variance des valeurs propres
### vecp1.est l'estimateur du premier vecteur propre
### vecp2.est l'estimateur du second vecteur propre

pop.acp = acpf(Xpop)
ind.pop = c(1:Npop)
Mat.indech = matrix(NA,ncol=n.sim,nrow=n)
for (i in 1:n.sim){ Mat.indech[,i] = sort(sample(ind.pop,n))}

### Sauvegarde des resultats des simulations
mu.est  = matrix(NA,ncol=p,nrow=n.sim)
Gamma.est = array(NA,c(p,p,n.sim))
vecp1.varest = Gamma.est
valp.est = matrix(NA,ncol=2,nrow=n.sim)
valp.varest = matrix(NA,ncol=2,nrow=n.sim)

vecp1.est = matrix(NA,ncol=p,nrow=n.sim)
vecp2.est = matrix(NA,ncol=p,nrow=n.sim)

for (i in 1:n.sim){
   Xsample = Xpop[Mat.indech[,i],]
   res.acp = acp.survey(Xsample,rep(n/Npop,n))
   mu.est[i,] = res.acp$mu
   Gamma.est[,,i] = res.acp$Gamma
   valp.est[i,] = res.acp$valp[1:2]
   vecp1.est[i,] = res.acp$vecp[,1]
   vecp2.est[i,] = res.acp$vecp[,2]
```

```
     res.influence = influence.acp(res.acp)
     Ilambda1 = matrix(res.influence$Ilambda1,ncol=1)
     Ilambda2 = matrix(res.influence$Ilambda2,ncol=1)
     valp.varest[i,1] = varest.sas(Ilambda1,rep(n/Npop,n))
     valp.varest[i,2] = varest.sas(Ilambda2,rep(n/Npop,n))
     vecp1.varest[,,i] = varest.sas(res.influence$Ivecp1,rep(n/Npop,n))

   cat("sim:",i,"\n")
}
list( mu.est = mu.est,vecp1.varest = vecp1.varest,valp.est = valp.est,
valp.varest = valp.varest,vecp1.est = vecp1.est,vecp2.est =vecp2.est, Gamma.est = Gamma.est
}
```

Table des figures

2.1 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). 37

2.2 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). 38

2.3 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 $\epsilon \sim N(0,0.5)$. 51

2.4 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 $X=+0.5$ 52

2.5 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 $X=+1$ 53

2.6 Comparaison entre les quantiles marginaux et les quantiles géométriques calculés sur les composants chimiques (Ba, Ca) mésurés sur des algues. 54

3.1 Quantile contour plots for $r=0.3$ (inner contours) and $r=0.9$ (outer contours). The u-th geometric quantile calculated from the population U are represented by solid line and its estimation with SRSWR strategie (resp. stratified strategie) in dotdash line (resp. longdash line). 71

3.2 Relative estimation errors of geometric quantile for two sampling strategies (SRSWR and stratified sampling). Figure (a) (resp. (b)) is for $u=(0.1,0.3)^T$ (resp. $u=(0.1,0.8)^T$) and $n=1000$. Figure (c) (resp. (d)) is for $u=(0.1,0.3)^T$ (resp. $u=(0.1,0.8)^T$) and $n=500$. 72

4.1 A stratified sample of $n=100$ curves 89

4.2 Estimation errors for two different sampling strategies (SRSWR and stratified sampling). First eigenvalue with $n=100$. (a) and $n=1000$ (b). First eigenvector with $n=100$. (c) and $n=1000$ (d). 90

4.3 Estimation errors in the variance approximation for two different sampling strategies (SRSWR and stratified sampling). First eigenvalue with $n = 100$. (a) and $n = 1000$ (b). First eigenvector with $n = 100$. (c) and $n = 1000$ (d). 91

Contribution à l'estimation non paramétrique des quantiles géométriques et à l'analyse des données fonctionnelles

Résumé : Ce mémoire de thèse est consacré à l'estimation non paramétrique des quantiles géométriques conditionnels ou non et à l'analyse des données fonctionnelles. Nous nous sommes intéressés, dans un premier temps, à l'étude des quantiles géométriques. Nous avons montré, avec plusieurs simulations, qu'une étape de Transformation-Retransformation est nécessaire, pour estimer le quantile géométrique, lorsqu'on s'éloigne du cadre d'une distribution sphérique. Une étude sur des données réelles a confirmée que la modélisation des données est mieux adaptée lorsqu'on utilise les quantiles géométriques à la place des quantiles mariginaux, notamment lorsque les variables qui constituent le vecteur aléatoire sont corrélées. Ensuite nous avons étudié l'estimation des quantiles géométriques lorsque les observations sont issues d'un plan de sondage. Nous avons proposé un estimateur sans biais du quantile géométrique et à l'aide des techniques de linéarisation par les équations estimantes, nous avons déterminé la variance asymptotique de l'estimateur. Nous avons ensuite montré que l'estimateur de type Horvitz-Thompson de la variance converge en probabilité. Nous nous sommes placés par la suite dans le cadre de l'estimation des quantiles géométriques conditionnels lorsque les observations sont dépendantes. Nous avons démontré que l'estimateur du quantile géométrique conditionnel converge uniformement sur tout ensemble compact. La deuxième partie de ce mémoire est consacrée à l'étude des différents paramètres caractérisant l'ACP fonctionnelle lorsque les observations sont tirées selon un plan de sondage. Les techniques de linéarisation basées sur la fonction d'influence permettent de fournir des estimateurs de la variance dans le cadre asymptotique. Sous certaines hypothèses, nous avons démontré que ces estimateurs convergent en probabilité.

Mots-clés : Quantiles géométriques, quantiles géométriques conditionnels, ACP fonctionnelle, sondage, fonction d'influence, linéarisation, α-mélange, Transformation-Retransformation.

Contribution to the nonparametric geometric quantiles estimation and functional data analysis

Abstract : In this dissertation we study the nonparametric geometric quantile estimation, conditional geometric quantiles estimation and functional data analysis. First, we are interested to the definition of geometric quantiles. Different simulations show that Transformation-Retransformation technique should be used to estimate geometric quantiles when the distribution is not spheric. A real study shows that, data are better modelized by geometric quantiles than by marginal one's, especially when variables that make up the random vector are correlated. Then we estimate geometric quantiles when data are obtained by survey sampling techniques. First, we propose an unbaised estimator, then using linearization techniques we give its asymptotic variance. Further, we proove the consistensy of the Horvitz-Thompson estimator of the variance. Conditional geometric quantile estimation is also studied when data are dependent realisations. We prove that the proposed estimator converge uniformly on every compact sets. The second part of this thesis is devoded to the study of the Functional Principal Components Analysis parameters when data are curves selected with survey sampling techniques. Linearization techniques using influence functions allows us to give estimators of asymptotic variances. Under suitable conditions, we prove that the proposed estimators are consistent.

Key-words : Geometric quantiles, conditional geometric quantiles, Functional PCA, survey sampling, influence function, linerization, α-mixing, Transformation-Retransformation.