# Profondeur de Monge-Kantorovich

**Auteur :** Tixhon, Stéphanie
**Promoteur(s) :** Haesbroeck, Gentiane
**Faculté :** Faculté des Sciences
**Diplôme :** Master en sciences mathématiques, à finalité didactique
**Année académique :** 2017-2018
**URI/URL :** http://hdl.handle.net/2268.2/4964

Université de Liège


Faculté des Sciences - Département de Mathématiques

# Profondeur de Monge-Kantorovich

Travail de fin d'études présenté par Stéphanie Tixhon en vue de l'obtention du grade de Master en Sciences Mathématiques à finalité didactique

Année académique 2017-2018

Université de Liège


Faculté des Sciences - Département de Mathématiques

# Profondeur de Monge-Kantorovich

Travail de fin d'études présenté par Stéphanie Tixhon en vue de l'obtention du grade de Master en Sciences Mathématiques à finalité didactique

Année académique 2017-2018

# Remerciements

*Je souhaite adresser mes remerciements les plus sincères aux personnes qui m'ont apporté leur aide et leur soutien durant la rédaction de ce mémoire, mais aussi tout au long de mes études. Je commencerai par remercier* Gentiane Haesbroeck *pour m'avoir permis d'effectuer ce travail de fin d'études sous sa direction. Son aide, ses conseils, sa disponibilité malgré un emploi du temps chargé m'ont permis de mener à bien ce travail.*

*Je tiens aussi à remercier Yvik Swan qui m'a accordé de son temps pour répondre à mes questions sur le sujet et pour me guider dans l'application de la théorie à la pratique.*

*Merci également à Madame C.Timmermans, et Messieurs Y.Swan et P.Mathonet d'avoir accepté de faire partie du comité de lecture.*

*Je remercie également mes parents qui m'ont soutenue durant l'entièreté de mon parcours universitaire et qui m'ont encouragée à persévérer. Merci à ceux qui ont cru en mes capacités et plus particulièrement à Christine qui me soutient depuis le début.*

*Je finirai en remerciant mes amis et amies, rencontrés lors de mon parcours universitaire, qui ont rendu ces études plus belles et agréables.*

*Stéphanie Tixhon*

# Table des matières

**Introduction** **1**

**1 Fonctions de profondeur** **7**
1.1 Introduction . . . 7
1.2 Fonctions de profondeur . . . 7
1.3 Fonction de profondeur de Tukey pour des distributions particulières . . . 9
1.3.1 Famille $\mathcal{P}^1$ . . . 10
1.3.2 Famille $\mathcal{P}^p_{ell}$ . . . 12
1.4 Exemples empiriques . . . 17
1.5 Contours de profondeur . . . 20
1.5.1 Contours de profondeur pour la famille $\mathcal{P}^1$ . . . 21
1.5.2 Contours de profondeur pour la famille $\mathcal{P}^p_{ell}$ . . . 22
1.5.3 Contexte empirique . . . 22
1.6 Notion de quantiles . . . 23
1.7 Notion de rangs . . . 24
1.8 Convexité des contours de profondeur . . . 26

**2 Profondeur de Monge-Kantorovich** **33**
2.1 Introduction . . . 33
2.2 Le transport . . . 34
2.2.1 Exemple de transport en dimension 1 . . . 34
2.2.2 Transport en dimension 1 pour une application croissante . . . 35
2.3 Définitions et propriétés relatives à l'application du transport . . . 38
2.4 Définitions des vecteurs quantiles et rangs de Monge-Kantorovich . . . 40

**3 Propriétés de la profondeur de MK** **44**
3.1 Introduction . . . 44
3.2 Caractérisation des contours et régions de profondeur . . . 44
3.3 Profondeur de Monge-Kantorovich pour les familles $\mathcal{P}^1$ et $\mathcal{P}^p_{ell}$ . . . 45
3.4 Application . . . 49
3.5 Implémentation des contours de Monge-Kantorovich . . . 52

**4 Application réelle : ranking des universités** **56**
4.1 Introduction . . . 56
4.2 Analyse en dimension $p = 1, 2$ . . . 57

4.3 Universités belges . . . . . 61
4.4 Analyse en dimension 5 . . . . . 62
4.5 Analyse en composantes principales . . . . . 68

**Conclusion** **70**

**A Annexe** **72**

**Bibliographie** **80**

# Introduction

La statistique multivariée est la branche de la statistique qui s'intéresse à l'étude de plusieurs variables simultanément.
Le développement effectué tout au long de ce mémoire portera sur la théorie relative à l'analyse exploratoire réalisée dans le contexte quantitatif.
Tout au long de ce travail, nous illustrerons les différents concepts théoriques grâce à une base de données reprenant le classement 2018 des 200 meilleures universités du monde et publiée par le *"Times Higher Education"*[1]. Ce classement a été effectué sur base des indicateurs suivants : la recherche, l'enseignement, le nombre de citations dans des articles scientifiques, le financement par les industries et les perspectives internationales.

Le classement correspond à l'attribution d'un rang univarié à chaque université déterminé sur base d'un score global donné par la variable "*OverallScore*". Chacun des indicateurs se voit attribuer une importance différente dans le classement final. En effet, pour chacun des indicateurs, chaque université se voit attribuer un score sur 100. Par exemple, pour le nombre de citations dans des articles scientifiques, l'université comptabilisant le plus grand nombre de citations obtient le score de 100 et les autres universités se voient attribuer un score qui est un ratio du nombre de citations que celles-ci comptabilisent par rapport à l'université en comptabilisant le plus. Ensuite, un score global est calculé en fonction de l'importance accordée à chaque indicateur. Parmi ces scores globaux, il y a des ex-aequo entre universités. Dans ce cas, le *Times Higher Education* attribue le même rang à ces universités ex-aequo. Cette façon de procéder fait que certains rangs ne sont pas occupés. Prenons un exemple afin d'expliquer ce qu'il se passe lorsque plusieurs universités ont le même score global. Le *California Institute of Technology* et l'*université de Stanford* occupent tous les deux la 3ème place du classement avec un score de 93. L'université suivante dans le classement se verra alors attribuer le rang 5 et non 4. Généralisons au cas de $t$ ex-aequo, $t \in \mathbb{N}_0$. Si $t$ universités occupent le rang $i > 0$ alors, l'université qui vient après ces $t$ universités dans le classement se verra attribuer le rang $i + t$.

Dans ce mémoire, nous nous intéresserons à la définition de rangs multivariés et nous déterminerons ces rangs pour la base de données étudiée. Nous pourrons alors comparer les rangs multivariés obtenus avec les rangs univariés de la base de données.

Détaillons ce que représente chaque indicateur. Tout d'abord, l'indicateur de l'enseignement englobe notamment le prestige relatif à l'enseignement dans l'université concernée mais aussi la proportion de l'équipe éducative par rapport au nombre d'étudiants et le revenu institutionnel qui donne une vue des installations dont les étudiants et le personnel bénéficient. Celui-ci représente 30% du score global. Ensuite, l'indicateur de recherche comptant pour 30% du score se base entre autre sur la quantité d'articles publiés dans des périodiques académiques et revus par les pairs, ainsi que sur la réputation (relative à

1. Classement des universités : https ://www.timeshighereducation.com/world-university-rankings/2018/

la recherche) d'une université parmi ses pairs. Passons au financement par les industries. Ce dernier indicateur concerne les contrats passés par les industries aux universités et représente 2.5% du score. Les industries font appel à des membres de l'université avec pour objectif de recevoir des conseils ou encore de demander une aide en matière d'innovation. Le nombre de citations dans les articles scientifiques compte pour 30% du score. Enfin, les perspectives internationales concernent la capacité des universités à attirer des étudiants d'universités étrangères ainsi que des post-doc. Cela prend également en compte les collaborations entre universités sur l'écriture d'articles scientifiques et contribue au score pour un ratio de 7.5%.

Ce mémoire se veut théorique. Toutefois, par souci de compréhension, nous définirons certains concepts de façon empirique dans cette introduction. Pour cela, nous utiliserons uniquement les 60 premières universités du classement et nous n'interpréterons pas les résultats dans cette introduction. Dans le dernier chapitre, vous trouverez une analyse plus approfondie et commentée de ces données et cette analyse sera effectuée à partir des 200 meilleures universités.

Commençons par introduire certains outils.

En statistique univariée, les quantiles sont constamment utilisés par les statisticiens, notamment afin d'étudier les queues de distributions. Dans son livre *Exploratory Data Analysis* [27], John Tukey incite à l'utilisation du résumé à cinq valeurs, afin de compléter l'analyse descriptive d'une série quantitative en dimension un :

| Min | $Q_1$ | Médiane | $Q_3$ | Max |
|---|---|---|---|---|
| 49.5 | 60.68 | 68.15 | 80.32 | 90.3 |

où $Q_1, Q_3$ et Médiane sont les trois quartiles.

Ce résumé statistique effectué sur les 60 premières valeurs prises par la variable *"Teaching"* de la base de données décrit le comportement d'une série univariée ordonnée et peut être visualisé à l'aide d'une boîte à moustaches telle qu'illustrée à la Figure 1.

Cette représentation a été obtenue via la commande *boxplot* du logiciel *R* et nous permet d'étudier la distribution d'une variable statistique (notamment sa symétrie, ainsi que sa dispersion au centre mais aussi parmi les plus petites, respectivement les plus grandes, valeurs). Elle nous permet également de déterminer s'il existe des valeurs atypiques. Cependant, en dimension supérieure, il n'y a plus de notion d'ordre prédéfinie. Il en va de même pour les notions de quantiles et de rangs. Dès lors, comment s'y prendre pour repérer les observations "centrales" (équivalentes à celles situées dans la "boîte" de la boîte à moustaches en dimension un) et les observations "extérieures" qui comptent parmi elles les valeurs atypiques ?

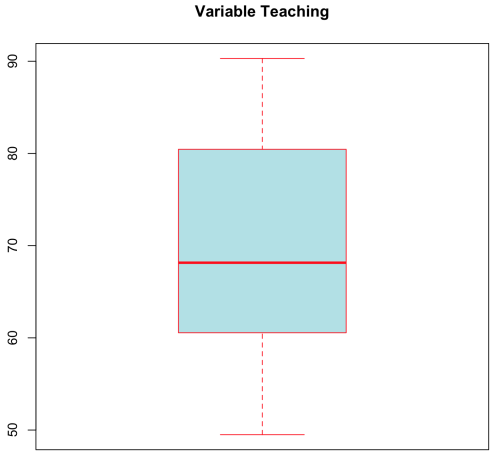


FIGURE 1 – Boîte à moustaches de la variable "*Teaching*"

Dans la base de données, la variable nous donnant le score global des universités ("*OverallScore*") définit le classement des universités. Ce classement consiste à organiser les observations pour cette variable de la plus grande à la plus petite (en dimension un) i.e de la meilleure université (l'*université d'Oxford* avec un score de 94.3) à la moins bonne (l'*université de Purdue* avec un score de 68.2 si l'on considère uniquement les 60 premières universités du classement). Toutefois, nous pourrions définir un autre type de classement. Pour cela, nous considérons un centre (nous décidons de prendre la médiane en dimension un) et au lieu de démarrer de la plus petite (respectivement de la plus grande) observation comme nous en avons l'habitude et de progresser vers l'observation la plus grande (respectivement la plus petite), nous pourrions plutôt partir de ce centre pour ensuite s'en écarter progressivement aussi bien vers la "gauche" que vers la "droite" comme décrit à la FIGURE 2 et à la FIGURE 3 :

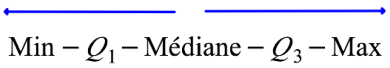


FIGURE 2 – Classement orienté à partir du centre en dimension 1

On aurait alors le classement des observations suivant :

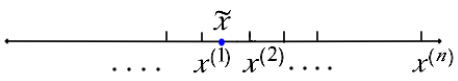


FIGURE 3 – Classement orienté d'observations à partir du centre en dimension 1

où, pour tout $i \in \{1, ..., n\}$, $x^{(i)}$ est la $i$ème observation dans le classement orienté à partir du centre en dimension 1. Cette notation est à ne pas confondre avec la notation $x_{(1)}, ..., x_{(n)}$ qui désigne en général le classement des observations de manière croissante (si $i \leqslant j$, alors $x_{(i)} \leqslant x_{(j)}$).

Afin d'illustrer cette idée de classement orienté à partir du centre, nous avons considéré les 60 premières observations pour la variable *OverallScore* et nous avons déterminé la médiane de ces observations. Celle-ci est de 78.85. Nous avons ensuite répertorié les écarts entre la médiane et chacune des 60 valeurs prises par *OverallScore*. Les universités occupant la première place du classement orienté à partir du centre sont l'université de *ETH Zurich Swiss Federal Institute Technology* et l'université de *Pennsylvanie* aux USA avec un écart de 0.15 par rapport à la médiane. L'université étant dernière de ce classement est l'université *d'Oxford* avec un écart de 15.45 par rapport à la médiane.

Cette idée de nouveau classement sera celle utilisée en dimension $p$, avec $p > 1$. Il faudra cependant commencer par définir la notion de "centre".

Dans le cas particulier de la dimension 2, nous disposons également d'un outil nous permettant de visualiser la dispersion des observations bivariées, le "*sac de points*" plus connu sous son nom anglais *bagplot*. Celui-ci est proposé par Rousseuw, Ruts et Tukey [23] et est construit sur le principe de classement des observations représenté Figures 2 et 3. Il est souvent considéré comme une généralisation de la boîte à moustaches. La construction d'un bagplot sera expliquée un peu plus tard, mais en voici déjà une représentation illustrée à la Figure 4 et obtenue à partir de la fonction *bagplot* de la librairie *aplack* du logiciel *R* appliquée aux variables *"Research"* et *"Teaching"* de la base de données :

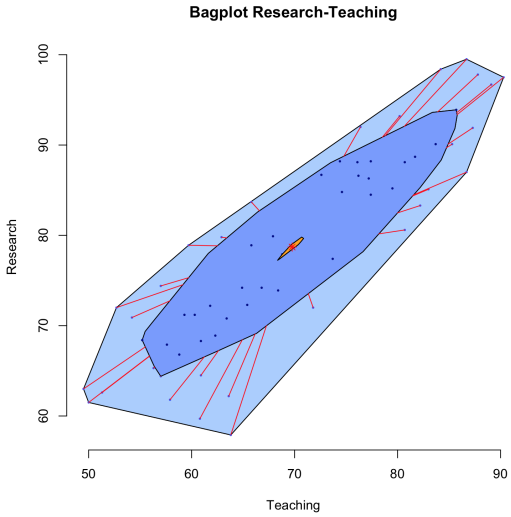


Figure 4 – Bagplot pour les variables *"Teaching"* et *"Research"*

- L'étoile rouge représente le point le plus "au centre".
- La partie bleu foncé représente le "*le sac*" ("*bag*" en anglais) et contient 50% des observations.
- La partie bleu clair représente la "*boucle*" ("*loop*" en anglais). Cette partie est l'enveloppe convexe des observations se situant à l'intérieur d'une partie du *bagplot* appelée la "*frontière*" ("*fence*" en anglais) et extérieures à la partie centrale. La *fence* n'est pas représentée ici. Elle est définie par *le sac* gonflé d'un facteur 3.

Ce *bagplot* capture la masse d'observations centrales dans le *sac* tout comme un *boxplot* capture les observations centrales dans sa *boîte*. La région centrale, i.e le *sac*, est assimilée à l'intervalle interquartile en dimension 1. Le triangle rouge et les segments rouges peuvent être respectivement assimilés à la médiane et aux moustaches de la boîte à moustaches.

Revenons à l'idée de classement des observations. Les observations bivariées seront classées selon leur position par rapport au centre du *bagplot*. Ce classement sera réalisé grâce à la notion de fonction de profondeur qui sera introduite dans le chapitre suivant. Il en va de même lorsque l'on travaille dans une dimension supérieure quelconque, même si la visualisation des observations devient difficile.

De cette nouvelle définition d'ordre établie en dimension $p > 1$ découleront les définitions de quantiles multivariés et de rangs, ce qui permettra de repérer les observations "centrales" comme le permettait la boîte à moustaches en dimension un. Au final, un résumé statistique en dimension $p > 1$ équivalent à celui effectué en univarié pourra être réalisé.

Venons-en à la description des différents chapitres qui seront développés dans ce mémoire.
Pour commencer, dans le Chapitre 1 nous introduirons la notion de fonction de profondeur et en particulier celle de *profondeur de demi-espace*. Cette dernière notion sera définie à la fois théoriquement et empiriquement. Nous terminerons ce chapitre en constatant que la convexité des contours de la fonction de profondeur de demi-espace pose problème pour certaines distributions.
Ensuite, dans le Chapitre 2, il sera question d'une autre fonction de profondeur appelée *fonction de profondeur de Monge-Kantorovich*. Il sera nécessaire de faire appel à la théorie du transport pour la définir. Nous verrons sur quelle intuition s'est basée la construction de cette profondeur ainsi que la construction des quantiles et des rangs dits de Monge-Kantorovich.
Dans le Chapitre 3, il sera question des propriétés de la fonction de profondeur de Monge-Kantorovich. Encore une fois, il sera d'abord question du contexte théorique puis nous enchaînerons avec le contexte empirique afin que le lecteur puisse s'approprier au mieux la théorie définie. Nous nous placerons dans le cas de deux distributions particulières afin d'illustrer ces concepts.

Enfin dans le Chapitre 4, il sera question d'une analyse plus approfondie des données des universités du monde sur base des indicateurs précités. Nous comparerons les rangs

univariés définis dans la base de données avec les rangs multivariés qui seront définis dans ce mémoire. Nous nous demanderons si la pondération effectuée par le *Times Higher Education* semble être une pondération correcte ou si nous ne pourrions pas en trouver une autre qui serait plus représentative. Pour cela, nous effectuerons une analyse en composantes principales et comparerons la pondération de chaque indicateur obtenue dans cette analyse avec la pondération définie par le *Times Higher Education*. Les analyses se baseront sur les 200 meilleures universités du monde uniquement.

# Chapitre 1

# Fonctions de profondeur

## 1.1 Introduction

Ce chapitre est consacré à la définition et aux propriétés des fonctions de profondeur ainsi qu'à la notion de contours de profondeur, pour finir avec l'application de ces concepts à deux familles de distributions particulières. Nous allons tout d'abord définir les notions dans un cadre théorique puis nous tenterons d'illustrer celles-ci d'un point de vue empirique afin d'expliciter la théorie. Tout au long de ce chapitre, nous considérerons des distributions continues.

Tout d'abord, dans la suite, nous serons amenés à parler de symétrie. Or en multivarié, il existe plusieurs types de symétries. Les plus connues sont : les symétries du type central, sphérique, elliptique et angulaire. Celles qui seront utiles dans ce travail sont les symétries du type sphérique et elliptique.

Si les autres symétries intéressent le lecteur, il est invité à se référer au mémoire de Luc Thoma : "Quantiles en statistique multivariée" [25].

## 1.2 Fonctions de profondeur

Ce sont Liu [15], Zuo et Serfling [30] qui définissent ce qu'est une fonction de profondeur.

**Définition 1.1.** Soit $\mathcal{F}$ l'ensemble des distributions $p$-variées, continues, ayant une propriété de symétrie de type quelconque par rapport à $\theta \in \mathbb{R}^p$.

Une fonction de profondeur est une fonction bornée

$$D : \mathbb{R}^p \times \mathcal{F} \to \mathbb{R}^+$$

qui devrait satisfaire les axiomes suivants :

1) Invariance affine : Si $A$ est une matrice $p$-carrée inversible et $b$ un $p$-vecteur, alors

$$D(Ax + b, F_{AX+b}) = D(x, F), \forall x \in \mathbb{R}^p .$$

2) Maximalité au centre : $D(\theta, F) = \sup_{x \in \mathbb{R}^p} D(x, F)$.

3) Décroissance par rapport au centre : $D(x, F) \leqslant D(\theta + t(x - \theta), F)$, $\forall t \in [0, 1]$ et $\forall x \in \mathbb{R}^p$.

4) Annulation à l'infini : $D(x, F)$ tend vers 0 si $||x||$ tend vers l'infini.

Ainsi, la profondeur d'un élément $x$ de $\mathbb{R}^p$ sera d'autant plus grande que $x$ se situe proche du centre de la distribution.

Il existe différentes fonctions de profondeur. Par exemple, la fonction de profondeur de Mahalanobis [16], celle de Liu [15], celle de Tukey [26] (aussi connue sous le nom de profondeur de demi-espace) et bien d'autres. Toutefois, seule la fonction de profondeur de Tukey sera évoquée dans ce travail. Le lecteur est renvoyé aux articles donnés en référence pour plus de détails sur ces deux autres profondeurs.

Revenons à la notion de symétrie et introduisons un autre type de symétrie que nous n'avons pas encore évoqué : la symétrie de demi-espace.

**Définition 1.2.** Un vecteur aléatoire $X \in \mathbb{R}^p$ suit une distribution ayant la propriété de symétrie de demi-espace, aussi appelée $H$-symétrie, par rapport à $\theta \in \mathbb{R}^p$ si $\mathbb{P}(X \in H) \geqslant \frac{1}{2}$ pour tout demi-espace fermé $H = \{x \in \mathbb{R}^p : \alpha^T x \leqslant \beta\}$ avec $\alpha \in \mathcal{S}^{p-1} := \{x \in \mathbb{R}^p : \|x\| = 1\}$ et $\beta \in \mathbb{R}$, tel que $\theta \in \partial H$ avec $\partial H$ la frontière de $H$.

En particulier, en dimension un, toute distribution continue est $H$-symétrique par rapport à $\mu$ (la médiane) car alors, $\mathbb{P}(X \in H) = \frac{1}{2}$, pour tout demi-espace fermé $H$ contenant $\mu$ sur sa frontière.

Cette dernière notion de symétrie est moins forte que celles de symétrie elliptique, sphérique, centrale et angulaire car toute distribution à symétrie elliptique, sphérique, centrale ou angulaire est une distribution ayant la propriété de $H$-symétrie. Pour montrer la Proposition 1.1, nous aurons besoin des définitions suivantes, dans lesquelles "$\stackrel{\mathcal{L}}{=}$" définit l'égalité en loi.

**Définition 1.3.** La distribution d'un vecteur aléatoire $X \in \mathbb{R}^p$ est dite de symétrie sphérique si $AX \stackrel{\mathcal{L}}{=} X$, $\forall A \in \mathcal{O}(p)$, $\mathcal{O}(p)$ étant l'ensemble des matrices de dimension $p$ orthogonales.

**Définition 1.4.** Un vecteur aléatoire $X \in \mathbb{R}^p$ est issu d'une distribution à symétrie elliptique de paramètre de localisation $\mu \in \mathbb{R}^p$ et de paramètre de dispersion $\Sigma$ (matrice carrée, symétrique, réelle et définie positive de dimension $p$) si $X \stackrel{\mathcal{L}}{=} \mu + AY$ où $Y$ est issu d'une distribution à symétrie sphérique et $AA^T = \Sigma$.

**Proposition 1.1.** *Toute distribution ayant la propriété de symétrie elliptique possède la propriété de symétrie de demi-espace.*

*Démonstration.* Si $X$ est un vecteur aléatoire de distribution à symétrie elliptique par rapport à $\mu$ (en particulier sphérique), alors

$$X - \mu \overset{\mathcal{L}}{=} \mu - X$$

et dans ce cas,

$$\frac{X - \mu}{\|X - \mu\|} \overset{\mathcal{L}}{=} \frac{\mu - X}{\|\mu - X\|}$$

par définition de la norme.

Si cette dernière égalité est vérifiée, alors la propriété de $H$-symétrie est satisfaite car dans ce cas, tout hyperplan contenant $\mu$ divise $\mathbb{R}^p$ en deux demi-espaces ouverts de même masse de probabilité (cette masse étant égale à $1/2$ dans le cas d'une distribution continue). Mais tout hyperplan contenant $\mu$ définit également deux demi-espaces fermés de masse de probabilité supérieure ou égale à $1/2$ et donc, $\mathbb{P}(X \in H) \geqslant 1/2$ pour tout demi-espace fermé $H$ contenant $\mu$ sur sa frontière ce qui nous rend la définition de H-symétrie.

□

Définissons maintenant la profondeur avec laquelle nous allons travailler par la suite. Nous considérerons toujours le cas de distributions continues.

**Définition 1.5.** La profondeur de demi-espace (*half-space depth*) d'une valeur $x \in \mathbb{R}^p$ est donnée par la plus petite masse de probabilité contenue dans un demi-espace fermé $H \subset \mathbb{R}^p$ contenant $x$ :

$$D(x, F) = \inf\{\mathbb{P}(X \in H) : H \in \mathcal{H}, x \in H\} \tag{1.1}$$

où $\mathcal{H}$ est l'ensemble des demi-espaces fermés de $\mathbb{R}^p$.

Cette fonction vérifie les quatre axiomes de Liu-Zuo-Serfling [15] et [30] pour toute distribution $F$ H-symétrique sur $\mathbb{R}^p$ [25].

Introduisons cette fonction de profondeur sur deux exemples théoriques.

## 1.3 Fonction de profondeur de Tukey pour des distributions particulières

Dans cette partie, nous allons nous attarder sur deux familles de distributions particulières notées $\mathcal{P}^1$ et $\mathcal{P}^p_{ell}$. L'objectif est de déterminer la forme de la fonction de profondeur de Tukey de ces distributions.

### 1.3.1 Famille $\mathcal{P}^1$

Notons $\mathcal{P}^1$ la famille des distributions de densité non nulle sur des ensembles à support convexe, par rapport à la mesure de Lebesgue.

Nous supposons, dans cet exemple, que les distributions considérées sont continues et que la fonction de répartition $F$ est une fonction strictement croissante, telle que la fonction $F^{-1}$ est bien définie.

Nous savons que, en dimension un, toute distribution continue est $H$-symétrique par rapport à la médiane notée $\mu$. Nous considérons donc une distribution continue issue de la famille $\mathcal{P}^1$ en dimension un et nous allons établir la forme de la fonction de profondeur de Tukey pour une telle distribution en partant de la définition de la fonction de profondeur de Tukey donnée en (1.1). Voici une illustration à la FIGURE 1.1 d'une distribution issue de la famille $\mathcal{P}^1$.

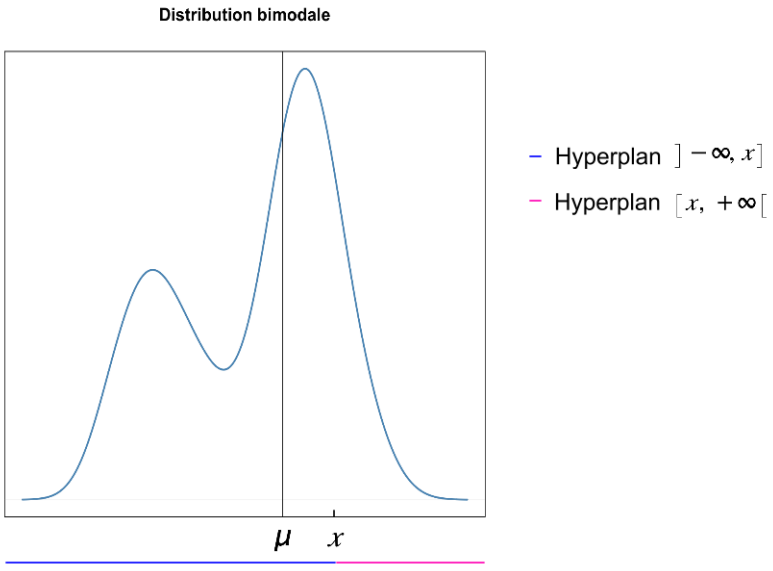


FIGURE 1.1 – Distribution issue de la famille $\mathcal{P}^1$

Par définition de $\mu$, il y a 50% de la masse de probabilité totale à gauche de $\mu$ et 50% à droite. En d'autres termes, $\mu = F^{-1}\left(\frac{1}{2}\right)$.

$\forall x \in \mathbb{R}$, deux hyperplans (pour $p = 1$, ce sont des demi-droites) peuvent être définis : $H_1 = ]-\infty, x]$ et $H_2 = [x, +\infty[$ comme illustrés à la FIGURE 1.1.

- Si $x \leqslant \mu$, nous avons la configuration illustrée à la FIGURE 1.2.

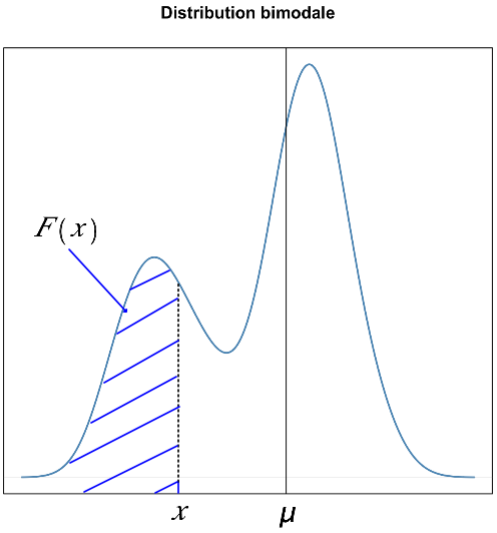


FIGURE 1.2 – Distribution issue de la famille $\mathcal{P}^1$ avec $x \leqslant \mu$

Dans ce cas, au vu de la Définition 1.5, il est clair que la fonction de profondeur de Tukey de n'importe quelle valeur $x \in \mathbb{R}$ coïncide avec la fonction de répartition de cette valeur. Autrement dit, $D(x, F) = F(x)$.

- Si $x > \mu$, nous sommes dans la situation illustrée à la FIGURE 1.3.

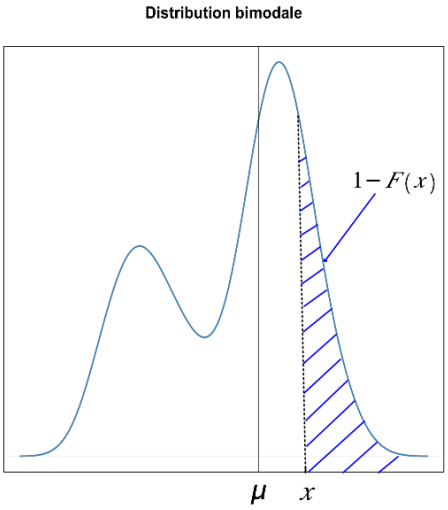


FIGURE 1.3 – Distribution issue de la famille $\mathcal{P}^1$ avec $x > \mu$

Dans cette situation, il est clair que la fonction de profondeur de Tukey pour toute valeur réelle $x$ est donnée par $D(x, F) = 1 - F(x)$.

Ainsi, dans le cas d'une distribution continue issue de la famille $\mathcal{P}^1$, la fonction de profondeur de demi-espace s'écrit :

$$D(x, F) = \begin{cases} F(x) \text{ si } x \leqslant F^{-1}\left(\frac{1}{2}\right) \\ 1 - F(x) \text{ si } x > F^{-1}\left(\frac{1}{2}\right) \end{cases}$$

Ce qui peut se réécrire comme suit :

$$D(x, F) = \min\{F(x), 1 - F(x)\}$$

Cette fonction vérifie bien toutes les propriétés attendues d'une fonction de profondeur grâce aux propriétés de la fonction de répartition $F(x)$.

Le cas discret se traite de façon analogue si ce n'est qu'il faut prendre en compte la probabilité ponctuelle $\mathbb{P}(X = x)$. Pour plus d'informations sur le traitement du cas discret, le lecteur est renvoyé au mémoire de Luc Thoma [25].

### 1.3.2 Famille $\mathcal{P}^p_{ell}$

Notons $\mathcal{P}^p_{ell}$ la famille des distributions à symétrie elliptique sur $\mathbb{R}^p$ $(p > 1)$.

La famille $\mathcal{P}^p_{ell}$ des distributions à symétrie elliptique est une famille de distributions paramétriques indexées par un paramètre de localisation $\mu \in \mathbb{R}^p$ ainsi qu'un paramètre de dispersion $\Sigma$, ce dernier paramètre étant une matrice symétrique, réelle et définie positive. Si $g$ est la fonction de densité radiale, que nous supposerons décroissante afin d'avoir l'unimodalité de la fonction de densité $f$, et $G$ la fonction de répartition associée, alors la distribution à symétrie elliptique associée se note : $P_{\mu,\Sigma,g}$. De plus, la fonction de densité d'un vecteur aléatoire $X$ issu d'une distribution à symétrie elliptique $P_{\mu,\Sigma,g}$ se note :

$$f(x) = \frac{c_p}{\sqrt{\det \Sigma}} g((x - \mu)^T \Sigma^{-1} (x - \mu)) \tag{1.2}$$

$c_p$ étant une constante de normalisation dépendant de $p$ i.e une constante telle que $\int_{\mathbb{R}^p} f(x)dx = 1$.

De plus, par les Définitions 1.3 et 1.4, un vecteur aléatoire $X$ est de distribution à symétrie elliptique $P_{\mu,\Sigma,g}$ si et seulement si $Y := \Sigma^{-1/2}(X - \mu)$ est issu d'une distribution à symétrie sphérique $P_{0,I,g}$.

Donnons un exemple de distribution à symétrie elliptique. Tout vecteur aléatoire $X \in \mathbb{R}^p$ suivant une distribution multinormale $(X \sim \mathcal{N}_p(\mu, \Sigma))$, est distribué selon une loi à symétrie elliptique. En effet, si $X \sim \mathcal{N}_p(\mu, \Sigma)$, alors la fonction de densité de $X$ est donnée par :

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sqrt{\det \Sigma}} \exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

C'est une conséquence du théorème de standardisation d'un vecteur multinormal [8]. Nous supposons à nouveau être en présence de distributions continues.

Soit un vecteur aléatoire $X \in \mathbb{R}^p$ issu de la famille $\mathcal{P}^p_{ell}$. Tout demi-espace fermé peut s'écrire sous la forme $H = \{x \in \mathbb{R}^p : \alpha^T x \leqslant \beta\}$, avec $\alpha \in \mathcal{S}^{p-1}$ et $\beta \in \mathbb{R}$. De plus, par la Proposition 1.1, toute distribution issue de la famille $\mathcal{P}^p_{ell}$ est $H$-symétrique par rapport à $\mu$.

Pour une valeur $x^*$, l'objectif est de chercher le demi-espace contenant $x^*$ et dont la masse de probabilité est la plus faible. Par définition d'une distribution $H$-symétrique nous savons déjà que $\mathbb{P}(X \in H) \geqslant \frac{1}{2}$, avec $H$ un demi-espace fermé contenant $\mu$ sur sa frontière et contenant $x^*$. Cependant, ce demi-espace n'est pas celui dont la masse de probabilité est la plus faible. En effet, en considérant la situation représentée à la FIGURE 1.4, si l'on translate la frontière de $H$ vers la droite, on peut se rendre compte que le demi-espace $H_{translate}$ contenant $x^*$ sur sa frontière est de masse de probabilité plus faible.

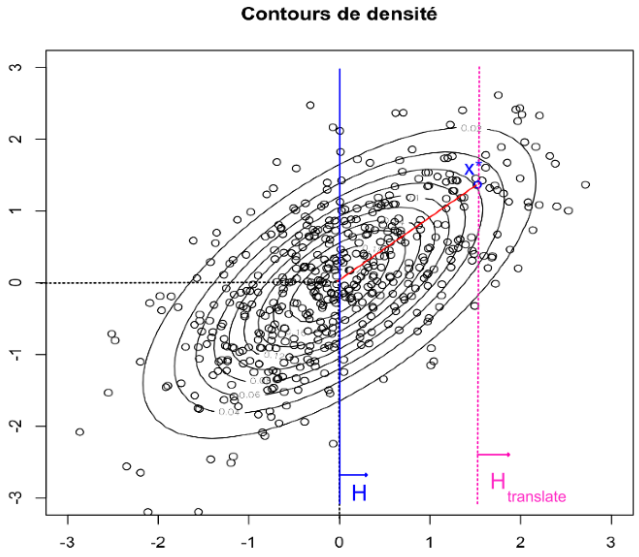


FIGURE 1.4 – Demi-espaces et contours de densité

Afin que l'exemple théorique soit plus parlant, nous avons représenté le nuage de points constitué de 500 réalisations d'un 2-vecteur aléatoire $X = (X_1, X_2)$ issu d'une distribution multinormale de moyenne nulle et de matrice de variance-covariance $\Sigma = \begin{pmatrix} 1 & 0.65 \\ 0.65 & 1 \end{pmatrix}$, ainsi que les contours de densité.

Ainsi pour une valeur $x^*$, la borne inférieure de (1.1) est réalisée par un demi-espace contenant cette valeur sur sa frontière. Nous pouvons constater sur la FIGURE 1.5 que le demi-espace réalisant la borne inférieure de (1.1) est celui dont la frontière est perpendiculaire à la direction définie par $\mu$ et $x^*$ (cette intuition est confirmée plus loin d'un point de vue théorique) et ne contenant pas $\mu$ . En effet, si on considère le demi-espace contenant $\mu$ et dont la frontière est perpendiculaire à la direction définie par $\mu$ et $x^*$ alors, étant donné que celui-ci contient un demi-espace $H$ contenant $\mu$ sur sa frontière, la masse de probabilité de ce demi-espace est supérieure ou égale à la masse de probabilité de $H$ or nous cherchons un demi-espace qui minimise la masse de probabilité.

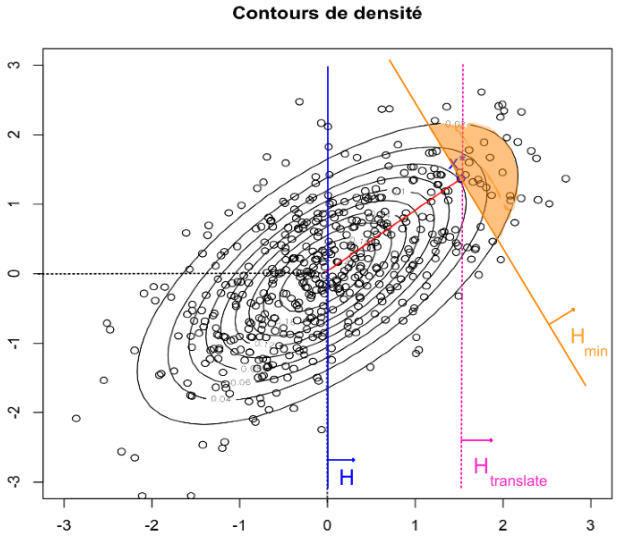


FIGURE 1.5 – Demi-espace minimisant (1.1)

Pour rappel, nous cherchons à déterminer la forme que prend la fonction de profondeur de Tukey pour une distribution issue de la famille $\mathcal{P}^p_{ell}$.

Chaque fois que l'on cherche à définir un demi-espace, on en définit en fait deux comme illustré à la FIGURE 1.6. Ainsi, pour tout $\alpha \in \mathcal{S}^{p-1}$ et $\beta \in \mathbb{R}$ définissant l'hyperplan $H_1 = \{x \in \mathbb{R}^p : \alpha^T x \geqslant \beta\}$ contenant $x^*$ sur sa frontière, l'hyperplan $H_2 = \{x \in \mathbb{R}^p : \alpha^T x \leqslant \beta\}$ contient également $x^*$ sur sa frontière.

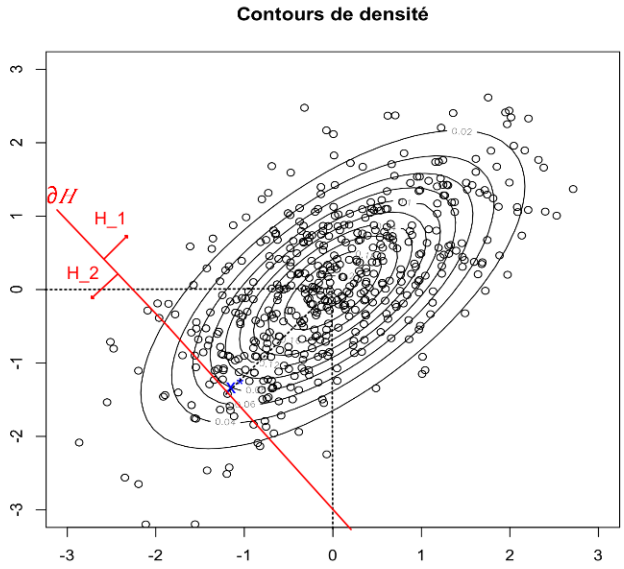


FIGURE 1.6 – Exemple de demi-espaces contenant $x^*$ sur leur frontière

Passons à la recherche d'une forme générale pour la profondeur de demi-espace d'une valeur $x^*$ dans le cas d'une distribution issue de la famille $\mathcal{P}^p_{ell}$. Pour cela, nous devons trouver le demi-espace qui minimise (1.1) et nous savons que ce demi-espace contient $x^*$ sur sa frontière, est tel que sa frontière est perpendiculaire à la direction définie par $x^*$ et $\mu$ et que ce demi-espace ne contient pas $\mu$. Reprenons les hyperplans $H_1$ et $H_2$ définis ci-dessus et supposons que leur frontière soit perpendiculaire à la direction définie par $x^*$ et $\mu$. Dans ce cas, on obtient

- $\alpha^T(x - x^*) \leqslant 0, \forall x \in H_2$
- $\alpha^T(x - x^*) \geqslant 0, \forall x \in H_1$

Supposons que ce soit l'hyperplan $H_2$ qui ne contient pas $\mu$. La profondeur de demi-espace est alors donnée par $\mathbb{P}(X \in H_2)$. Avant de développer cette probabilité, nous allons remarquer que nous pouvons ne traiter que le cas particulier d'une distribution à symétrie sphérique puisque par la propriété d'invariance affine de la profondeur de Tukey,

$$D(\Sigma^{1/2}y + \mu, F_{\Sigma^{1/2}Y+\mu}) = D(y, F_Y), \forall y \in \mathbb{R}^p$$

et, par la Définition 1.4, $X$ étant un vecteur aléatoire de distribution à symétrie elliptique, il peut s'écrire comme $X = \Sigma^{1/2}Y + \mu$ avec $Y$ un vecteur aléatoire issu d'une distribution à symétrie sphérique.

Ainsi, si l'on définit $H_{2modif}$ comme étant le demi-espace espace $H_2$ transformé grâce à la relation $X = \Sigma^{1/2}Y + \mu$ alors,

$\mathbb{P}(Y \in H_{2modif}) = \mathbb{P}(\alpha^T(Y - y^*) \leqslant 0) = \mathbb{P}(\alpha^T Y \leqslant \alpha^T y^*)$

Or, $Y = \Sigma^{-1/2}(X - \mu) \sim P_{0,I,g}$.

En posant $\gamma = \alpha^T y^*$,

$$\mathbb{P}(Y \in H_{2modif}) = \frac{c_p}{\sqrt{\det \Sigma}} \int_{\{y \in \mathbb{R}^p : \alpha^T y \leqslant \gamma\}} g(y^T y) dy$$

Il est difficile de trouver une forme plus précise de la profondeur de demi-espace dans le cas de distributions à symétrie sphérique car celle-ci va dépendre de la fonction de densité radiale $g(.)$. Toutefois, nous pouvons dire que la fonction de profondeur de Tukey dans le cas d'une distribution à symétrie sphérique est une fonction qui dépend d'un rayon et qui décroit en fonction de ce rayon. En effet, cette dernière affirmation provient du fait que l'hyperplan minimisant (1.1) et contenant $y$ sur sa frontière est celui perpendiculaire à la direction définie par les points 0 et $y$. Montrons que c'est bien cet hyperplan qui minimise (1.1).

Commençons par représenter la situation à la FIGURE 1.7. Par la propriété de symétrie sphérique, déterminer la profondeur de Tukey d'une valeur $y^*$ revient à déterminer la profondeur de Tukey de la valeur $y$, qui est l'image de $y^*$ sur l'axe horizontal par la rotation de centre $(0, 0)$ et de rayon $\|y^*\|$. Toutefois, il y a une multitude de demi-espaces contenant $y$ sur leur frontière (voir FIGURE 1.7). Pour montrer que c'est l'hyperplan perpendiculaire à l'axe horizontal qu'il faut choisir, nous montrons que la probabilité d'être dans la partie hachurée en mauve est plus petite que la probabilité d'être dans la partie hachurée en rouge. Autrement dit que

$$\int_{\text{cône rouge}} g(y^T y) dy \geqslant \int_{\text{cône mauve}} g(y^T y) dy$$

Par symétrie centrale de centre $y$, à tout point $(y_1, y_2)$ du cône mauve correspond un point $(y_1', -y_2)$. De plus, nous avons supposé à la Sous-Section 1.3.2 que la fonction de densité radiale $g$ est décroissante. Ainsi, comme pour $y_2$ fixé, les abscisses des points sont plus grandes dans le cône mauve que dans le cône rouge, en appliquant la décroissance de $g$ et un changement de variable, on obtient

$$\int_{\text{cône rouge}} g(y_1^2 + y_2^2) dy_1 dy_2 \geqslant \int_{\text{cône mauve}} g(y_1^2 + y_2^2) dy_1 dy_2$$

Ce qui donne le résultat souhaité. De là, on en tire également que la profondeur de Tukey de $y^*$ est une fonction $j$ de $\|y\|^2 = y^T y$ et que celle-ci décroît lorsque $\|y\|^2$ croît. Ce qui est bien en adéquation avec les observations réalisées auparavant.

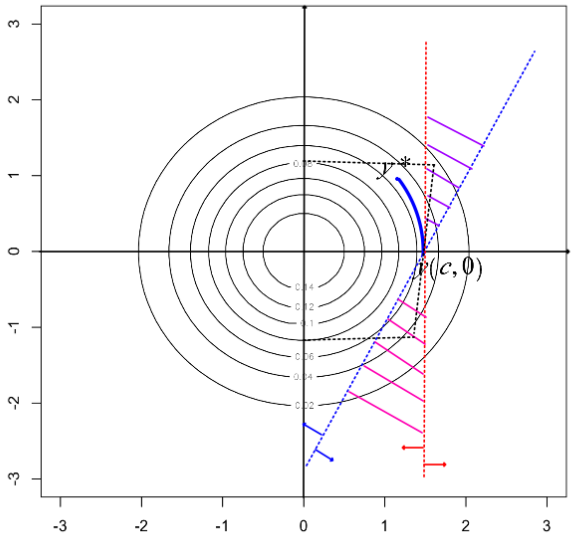


FIGURE 1.7 – Distribution à symétrie sphérique et hyperplan qui minimise (1.1)

Dans le cas des données centrées, le point de profondeur maximale sera l'origine dans $\mathbb{R}^p$.

## 1.4 Exemples empiriques

Illustrons ce concept de profondeur de demi-espace de façon empirique en dimension un. Nous considérons une variable aléatoire $X$ et $n \in \mathbb{N}_0$ réalisations de cette variable que l'on ordonne par ordre croissant $x_{(1)}, ..., x_{(n)}$. Nous souhaitons déterminer la profondeur de l'observation $x_{(i)}$, $D(x_{(i)}, F_n)$, avec $F_n$ la fonction de répartition empirique associée aux $n$ réalisations. Nous devons pour cela trouver un demi-espace $H$ contenant $x_{(i)}$ et tel que la fréquence cumulée des observations contenues dans ce demi-espace soit minimale. Une représentation des données ordonnées est illustrée FIGURE 1.8.

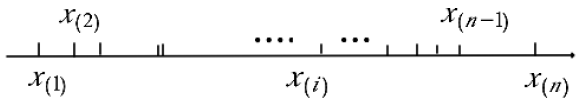


FIGURE 1.8 – Données ordonnées par ordre croissant

Nous pouvons constater que le demi-espace répondant aux conditions précitées est celui

contenant $x_{(i)}$ sur sa frontière et contenant le moins d'observations car il faut en minimiser la masse. Les deux demi-espaces contenant $x_{(i)}$ sur leur frontière sont représentés à la FIGURE 1.9.

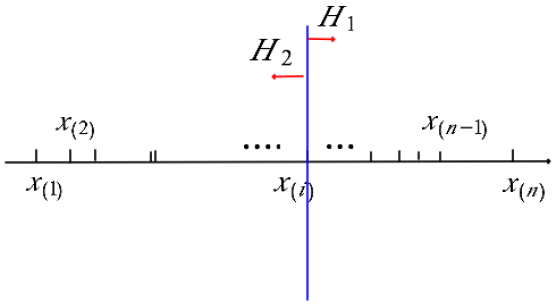


FIGURE 1.9 – Demi-espaces contenant $x_{(i)}$ sur leur frontière

Ainsi, pour déterminer la profondeur de l'observation $x_{(i)}$, il suffit de compter le nombre d'observations se situant dans le demi-espace en contenant le moins et de diviser le résultat obtenu par l'effectif total. La profondeur de cette observation dépend de la position de celle-ci par rapport à la médiane.
Autrement dit, la profondeur de l'observation $x_{(i)}$ est donnée par

$$D(x_{(i)}, F_n) = \begin{cases} \dfrac{n-i+1}{n} & \text{si } x_{(i)} \geqslant \tilde{x} \\ \dfrac{i}{n} & \text{si } x_{(i)} \leqslant \tilde{x} \end{cases} \tag{1.3}$$

où $\tilde{x}$ est l'estimation de la médiane.

Illustrons maintenant la fonction de profondeur de Tukey en dimension 2, en exploitant la base de données. Considérons les variables *"Research"* et *"Teaching"* et représentons la profondeur de Tukey pour l'échantillon constitué par les 60 premières observations de ces indicateurs. Nous obtenons deux représentations illustrées par les FIGURES 1.10 et 1.11.

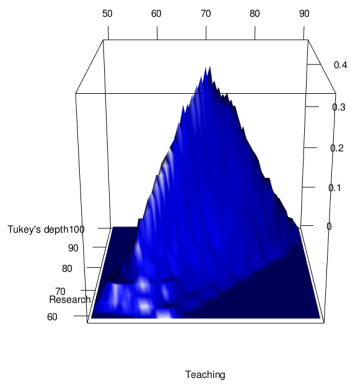


FIGURE 1.10 – Profondeur de Tukey en perspective

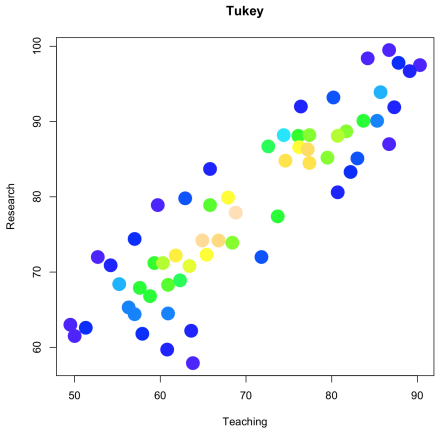


FIGURE 1.11 – Profondeur de Tukey

Celle de gauche est obtenue par la commande *perspdepth* de la librairie *depth* du logiciel *R*. L'axe vertical nous donne la profondeur de Tukey des observations bivariées. Celle de droite représente les observations colorées en fonction de leur profondeur de Tukey. Autrement dit, les observations bivariées ayant la même profondeur de Tukey sont de même couleur. Cette représentation est obtenue à partir de la fonction de profondeur de Tukey *depth* de la librairie *depth* du logiciel *R*. Ces deux représentations nous montrent la même chose. On constate qu'on a bien la propriété de maximalité au centre et que plus on s'écarte du centre de la distribution, plus la profondeur diminue. Sur la FIGURE 1.11, la couleur jaune est attribuée aux observations de plus grande profondeur et la couleur bleu foncé est attribuée aux observations de plus petite profondeur.

Pour comprendre comment est déterminée la profondeur de demi-espace d'une observation, repartons de la FIGURE 1.12 illustrant le nuage de points pour les variables *Teaching* et *Research* et sur lequel une observation est colorée en rouge. Nous allons déterminer la profondeur de cette observation.

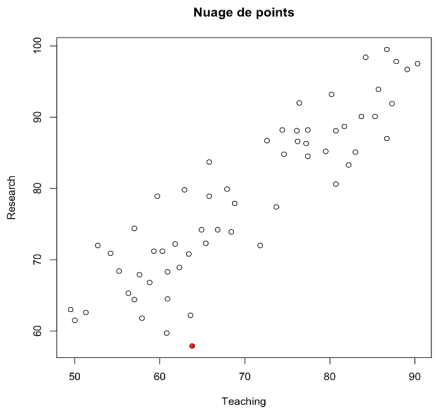


FIGURE 1.12 – Observations bivariées pour les variables *Teaching* et *Research*

Pour déterminer la profondeur de cette observation, il faut considérer tous les demi-espaces dont la frontière passe par cette observation (certains sont illustrés à la FIGURE 1.13) et prendre celui recouvrant la plus petite masse d'observations.

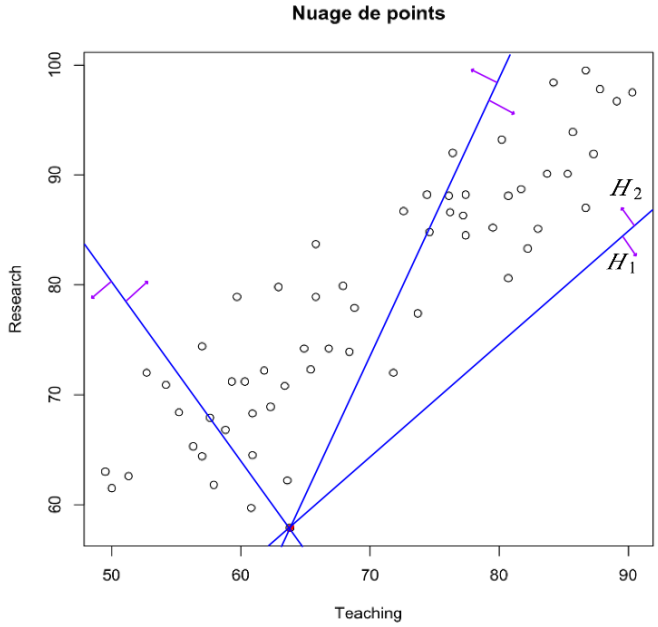


FIGURE 1.13 – Représentation de quelques hyperplans dont la frontière passe par l'observation d'intérêt

On constate clairement sur la FIGURE 1.13 que le demi-espace cherché est $H_1$. En effet, celui-ci contient une seule observation (l'observation d'intérêt). Il minimise donc bien (1.1) et la profondeur de demi-espace de cette observation est donnée par le nombre d'observations contenues dans cet hyperplan divisé par l'effectif total i.e que la profondeur de Tukey de cette observation est égale à $1/60$.

## 1.5 Contours de profondeur

Les contours de profondeur caractérisent la concentration au centre de la distribution et sont une généralisation des quantiles univariés. Un contour d'ordre $\alpha$ délimite une partie de $\mathbb{R}^p$ contenant l'ensemble des valeurs de $\mathbb{R}^p$ de profondeur supérieure ou égale à $\alpha$ pour une distribution $F$. Autrement dit, un contour de profondeur est formé par l'ensemble des valeurs de $\mathbb{R}^p$ de profondeur égale à une constante donnée pour cette distribution. La région de profondeur d'ordre $\alpha$, appelée aussi région tronquée au seuil $\alpha$, est définie par

$$D^{\alpha}(F) = \{x \in \mathbb{R}^p : D(x, F) \geqslant \alpha\} \tag{1.4}$$

La frontière de $D^\alpha(F)$ est le contour d'ordre $\alpha$ de $F$.
Dans le cas où la fonction de profondeur est continue, le contour d'ordre $\alpha$ est défini pour tout $\alpha > 0$. Par contre, dans le cas contraire, certains contours pourraient ne pas être définis. C'est notamment le cas dans le contexte empirique comme nous le verrons à la FIGURE 1.14.

Ces régions de profondeur vérifient certaines propriétés qui découlent des axiomes de LIU [15] et ZUO-SERFLING [30] donnés par la Définition 1.1 et ce pour n'importe quelle fonction de profondeur sauf dans le cas où l'on précise celle-ci.
Tout d'abord, si la fonction de profondeur vérifie l'axiome 1), alors les régions tronquées au seuil $\alpha$ respectent la propriété d'équivariance affine, à savoir : $D^\alpha(F_{AX+b}) = AD^\alpha(F_X)+b$.
Ensuite, par l'axiome 2), ces régions s'emboîtent de façon croissante.
Par ailleurs, si $D(., F)$ vérifie l'axiome 3), alors $D^\alpha(F)$ est connexe.
Enfin, si l'on considère la fonction de profondeur de demi-espace alors, les régions de profondeur d'ordre $\alpha$ sont compactes. Ce résultat est valable pour d'autres fonctions de profondeur, mais seule celle de Tukey nous intéresse dans ce mémoire. Pour les démonstrations relatives à ces propriétés, veuillez vous référer au mémoire de LUC THOMA [25].

On peut également définir les régions de profondeur de probabilité $\beta \in [0, 1]$. Celles-ci correspondent aux régions tronquées au seuil $t_\beta := \inf\{t : \mathbb{P}(D^t(F)) \geqslant \beta\}$. Le contour associé à ce type de régions est la frontière de ces régions.

Reprenons les familles particulières $\mathcal{P}^1$ et $\mathcal{P}^p_{ell}$ des Sous-Sections 1.3.1 et 1.3.2 et déterminons les contours d'ordre $\alpha$ relatifs à la profondeur de Tukey de ces distributions.

### 1.5.1 Contours de profondeur pour la famille $\mathcal{P}^1$

Nous avons obtenu, à la Sous-Section 1.3.1, la forme suivante pour la profondeur de Tukey :

$$D(x, F) = \min\{F(x), 1 - F(x)\}, \forall x \in \mathbb{R}^p$$

avec $F$ une distribution continue issue de la famille $\mathcal{P}^1$.
Ainsi, par la définition de la région tronquée au seuil $\alpha$ donnée en (1.4), on obtient que le contour d'ordre $\alpha$ de la distribution $F$ est donné par :

- $\{x \in \mathbb{R}^p : F(x) = \alpha\} = \{F^{-1}(\alpha)\}$ si $x \leqslant \mu$
- $\{x \in \mathbb{R}^p : 1 - F(x) = \alpha\} = \{F^{-1}(1 - \alpha)\}$ si $x > \mu$

Ce qui correspond, comme nous le verrons plus tard, à la définition des quantiles univariés donnée au début de la Section 1.6.

Passons au cas de la famille $\mathcal{P}^p_{ell}$

### 1.5.2 Contours de profondeur pour la famille $\mathcal{P}^p_{ell}$

Considérons le cas d'une distribution $F$ à symétrie sphérique. Nous généraliserons au cas d'une distribution à symétrie elliptique plus tard. Nous avons remarqué à la Sous-Section 1.3.2 que la fonction de profondeur de Tukey d'une telle distribution $F$ est une fonction qui dépend d'un rayon. Appelons $j(r)$ cette fonction.

Nous obtenons alors le résultat général suivant.

**Proposition 1.2.** *Les contours de profondeur de demi-espace d'une distribution à symétrie elliptique coïncident avec les contours elliptiques.*

*Démonstration.* Nous avons montré que la fonction de profondeur de demi-espace, dans le cas d'une distribution à symétrie sphérique, s'exprime comme une fonction $j$ d'un rayon et que cette fonction $j$ est décroissante. Autrement dit, les contours de profondeur de Tukey sont donnés par $\{y \in \mathbb{R}^p : j(\|y\|^2) = k\}$, avec $k$ une constante. Cela signifie que les contours de profondeur de Tukey pour une distribution à symétrie sphérique sont sphériques.

Généralisons ce résultat au cas d'une distribution à symétrie elliptique. Soit $Z \sim P_{0,I,g}$. Les contours de profondeur d'une telle distribution sont sphériques. Posons $X = \Sigma^{1/2}Z+\mu$. Le vecteur aléatoire $X$ est de distribution à symétrie elliptique $P_{\mu,\Sigma,g}$. Les contours de profondeur d'ordre $\alpha > 0$, pour une distribution à symétrie sphérique, sont donnés par $\{z \in \mathbb{R}^p : D(z, P_{0,I,g}) = \alpha\}$ et par la propriété d'invariance affine de cette fonction de profondeur, on a $\{z \in \mathbb{R}^p : D(\Sigma^{1/2}z + \mu, P_{\mu,\Sigma,g}) = \alpha\} = \{z \in \mathbb{R}^p : D(z, P_{0,I,g}) = \alpha\}$. Ce qui nous donne des contours de profondeur elliptiques pour la distribution $P_{\mu,\Sigma,g}$ car $D(\Sigma^{1/2}z + \mu, P_{\mu,\Sigma,g}) = j(\|x\|^2)$ avec $X \sim P_{\mu,\Sigma,g}$.

□

Cette propriété nous satisfait car lorsque nous ne connaissons pas la distribution mais que nous avons la possibilité de croire qu'elle est à symétrie elliptique, nous pouvons utiliser la méthode non paramétrique qui consiste à utiliser les contours de profondeur pour déterminer cette distribution.

### 1.5.3 Contexte empirique

Plaçons-nous dans le cas empirique afin de faire le lien entre le *bagplot* et la profondeur de demi-espace. C'est à partir de cette profondeur que le *bagplot* d'une distribution est construit. La partie bleu foncé contient les 50% des observations de plus grande profondeur et constitue la région de profondeur de probabilité $\beta = 50\%$.

Une représentation de certains contours de profondeur de Tukey sur les données *"Research"* et *"Teaching"* est donnée à la FIGURE 1.14. Cette illustration est obtenue par la fonction *isodepth* de la librairie *depth* du logiciel *R*. Le contour le plus à l'intérieur est le contour d'ordre $\alpha = 0.45$. Il reprend les observations de profondeur égale à 0.45. Il n'y a qu'une seule observation répondant à ce critère. Le contour le plus à l'extérieur est le

contour d'ordre $\alpha = 1/60$ et il reprend les observations de profondeur égale à $1/60$. Pour une analyse plus précise, veuillez vous référer au dernier chapitre du mémoire. Dans le cas empirique, certains contours n'existent pas. Par exemple, pour ces données, le contour d'ordre $\alpha = 0.38$ n'existe pas. Il n'y a donc pas d'observation dont la profondeur est de 0.38.

Nous pouvons observer à la FIGURE 1.14 que les contours de profondeur représentés ont une structure elliptique.

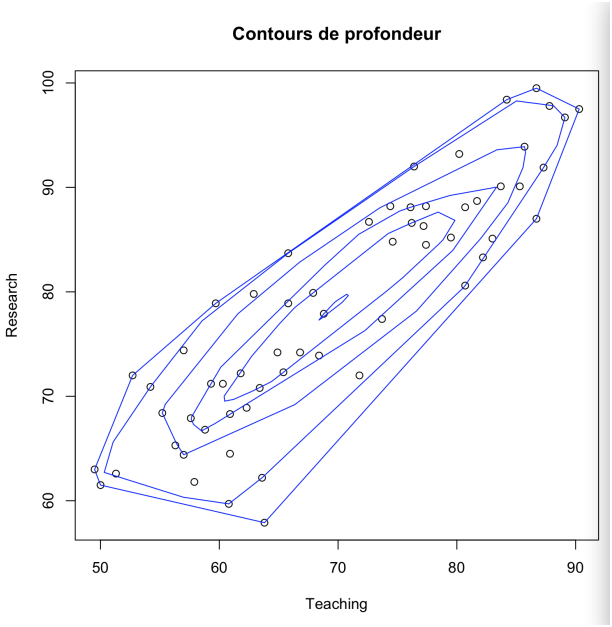


FIGURE 1.14 – Contours de profondeur de Tukey

Passons maintenant à la définition des quantiles et des rangs multivariés. Ces notions nous seront utiles plus tard.

## 1.6 Notion de quantiles

En dimension un, les quantiles d'une distribution $F$ sont définis comme suit :

$$Q(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geqslant \alpha\}$$

avec $\alpha \in [0, 1]$. On dira que $Q(\alpha)$ est le quantile associé à la masse de probabilité $\alpha$ et on notera $Q(\alpha) = F^{-1}(\alpha)$ de façon abusive.

Cette façon de définir les quantiles de la distribution $F$ se base sur le classement des valeurs par ordre croissant mais nous avons vu que nous pouvions aussi classer les valeurs à partir du centre de la distribution. De même, nous pouvons définir les quantiles univariés à partir du centre de la distribution grâce à une fonction $Q(u, \alpha)$ où $u$ représente la direction dans laquelle on va à partir de la médiane $\mu$ i.e $u = \pm 1$ et $\alpha$ est une probabilité.

Les quantiles définis par cette fonction sont appelés *quantiles orientés.*
La fonction $Q(u, \alpha)$ est définie comme suit :

$$\begin{aligned} Q(u,0) &= \mu \\ Q(-1,\alpha) &= F^{-1}\left(\frac{1-\alpha}{2}\right) \\ Q(1,\alpha) &= F^{-1}\left(1-\frac{1-\alpha}{2}\right) \end{aligned}$$

Or, pour tout $\alpha \in [0,1]$, on peut construire l'intervalle $I_\alpha$ de probabilité $\alpha$ autour de la médiane comme suit :

$$I_\alpha = \left[F^{-1}\left(\frac{1-\alpha}{2}\right), F^{-1}\left(1-\frac{1-\alpha}{2}\right)\right]$$

Les intervalles $I_\alpha$ s'emboîtent de façon croissante selon $\alpha$. Si $\alpha = 1/2$ alors $I_\alpha$ représente l'intervalle interquartile et si $\alpha = 0$ alors $I_\alpha$ se réduit à $\{F^{-1}(1/2)\}$.

Ainsi, à partir de la définition donnée de la fonction $Q(u, \alpha)$ et de l'intervalle $I_\alpha$ construit, on peut définir $Q(u, \alpha)$ comme étant le point frontière de l'intervalle $I_\alpha$ dans la direction $u$ choisie à partir de la médiane.

Cette fonction $Q(u, \alpha)$ peut être généralisée au cas d'une dimension quelconque $p$. Etant donné que nous avons déjà donné une définition de la fonction de profondeur ainsi qu'une définition des contours de profondeur, nous allons nous en servir.
Notons $\mu$ le point central de la distribution $F$ i.e le point de plus grande profondeur et caractérisons les directions issues de $\mu$ par les points de la sphère unité $\mathcal{S}^{p-1}$. Le quantile orienté $Q(u, \alpha)$ est défini par le point d'intersection entre le contour de profondeur associé à la région de profondeur de probabilité $\alpha$ et le prolongement de la direction $u$ partant de $\mu$.

## 1.7 Notion de rangs

En dimension un, considérons $n \in \mathbb{N}_0$ variables aléatoires $X_1, ..., X_n$ indépendantes et identiquement distribuées. Un rang peut être attribué à ces variables aléatoires selon le classement habituel i.e le classement croissant $X_{(1)}, ...., X_{(n)}$.
Ce rang est noté $rg(.)$ et est défini par :

$$rg(X_i) = j \text{ si } X_i = X_{(j)} \text{ pour } i, j \in \{1, ..., n\}$$

Si l'on se place dans le cadre empirique, on obtient

$$rg^{(n)}(X_i) = \sum_{j=1}^{n} \mathbb{I}_{[X_j \leqslant X_i]},\ \forall i \in \{1, ..., n\}$$

Or la fonction de répartition empirique, basée sur cet échantillon, est définie par

$$F^{(n)}(x) = \frac{1}{n+1} \sum_{j=1}^{n} \mathbb{I}_{[X_j \leqslant x]}, \forall x \in \mathbb{R}$$

Nous divisons par $n+1$ et non par $n$ afin que la fonction de répartition empirique n'atteigne pas la valeur 1.

Ainsi, en dimension 1, $\frac{1}{n+1} rg^{(n)}(X_i) = F^{(n)}(X_i)$. On dira que la fonction de répartition empirique coïncide avec les rangs normalisés.

Cependant, cette définition de rang se base sur une notion d'ordre, ce qui n'existe plus en dimension supérieure comme nous l'avons déjà évoqué. Pourquoi ne pas définir un rang basé sur un classement orienté à partir de la médiane ?

Cette autre définition du rang est donnée par

$$R^{(n)}(X_i) = \begin{cases} \left| rg^{(n)}(X_i) - \frac{n+1}{2} \right| + \frac{1}{2} \text{ si } n \text{ est pair} \\ \left| rg^{(n)}(X_i) - \frac{n+1}{2} \right| \text{ si } n \text{ est impair} \end{cases}$$

En effet, cela revient à centrer $X_{(1)}, ..., X_{(n)}$ en 0 puis à les parcourir à partir de 0 en leur attribuant un numéro en fonction de leur ordre de rencontre à gauche et à droite. Les $X_i$ à gauche de $\mu$ (respectivement à droite) se verront attribuer des rangs entre 1 et $\frac{n}{2}$ si $n$ est pair et entre 0 et $\frac{n-1}{2}$ si $n$ est impair. Autrement dit, si $i \in \{1, ..., n\}$ est tel que $rg^{(n)}(X_i) = 1$, alors $X_i$ deviendra la dernière valeur rencontrée sur l'axe et sera donc de rang maximal à savoir $\frac{n}{2}$ si $n$ est pair ou $\frac{n-1}{2}$ si $n$ est impair. Il en va de même si $rg^{(n)}(X_i) = n$

En dimension supérieure, un rang $R(.)$ pourra être attribué en fonction de la profondeur. En effet, les valeurs prises par un vecteur aléatoire peuvent être classées en fonction de leur profondeur. Le rang 1 est attribué à l'observation de $\mathbb{R}^p$ dont la profondeur est la plus grande et le rang augmente au fur et à mesure que la profondeur diminue, i.e que l'on s'écarte du centre de la distribution (point de profondeur maximale).

Autrement dit, pour deux éléments $x_1$ et $x_2$ de $\mathbb{R}^p$,

$$R(x_1) \geqslant R(x_2) \Leftrightarrow D(x_1, F) \leqslant D(x_2, F).$$

Regardons ce que l'on obtient dans le cas empirique en dimension 1, grâce à la définition des rangs à partir du centre et de l'égalité 1.3.
Supposons que l'on part des observations ordonnées et que $x_{(i)}$, $x_{(j)} \leqslant \mu$ pour $i, j \in \{1, ..., n\}$ alors

$$R^{(n)}(x_{(i)}) \geqslant R^{(n)}(x_{(j)}) \Leftrightarrow \left| i - \frac{n+1}{2} \right| \geqslant \left| j - \frac{n+1}{2} \right| \Leftrightarrow i \leqslant j \Leftrightarrow D(x_{(i)}, F) \leqslant D(x_{(j)}, F)$$

Le raisonnement est identique dans le cas où $x_{(i)}$, $x_{(j)} \geqslant \mu$.

## 1.8 Convexité des contours de profondeur

Commençons par montrer que les régions de profondeur de Tukey sont convexes.

**Proposition 1.3.** *Pour tout $\alpha > 0$, les régions tronquées au seuil $\alpha$ pour la profondeur de demi-espace sont convexes, ces ensembles pouvant être vides.*

*Démonstration.* Considérons $x_1$ et $x_2$ des éléments de $D^\alpha(F) = \{x \in \mathbb{R}^p : D(x, F) \geqslant \alpha\}$ et $\lambda \in [0, 1]$. L'inégalité suivante est alors vérifiée :

$$D((1-\lambda)x_1 + \lambda x_2, F) \geqslant \min\{D(x_1, F), D(x_2, F)\} \geqslant \alpha$$

par décroissance par rapport au centre de la fonction de profondeur de demi-espace. Ainsi, $(1-\lambda)x_1 + \lambda x_2 \in D^\alpha(F)$ et par définition d'un ensemble convexe, le résultat est démontré. □

La propriété précédente nous fait dire que, dans le cas d'une distribution un peu plus atypique telle que la distribution de *Cauchy bivariée*, la convexité des contours de profondeur de demi-espace fait que ces contours n'arrivent pas à capturer le caractère non convexe de la distribution. En d'autres termes, les contours de profondeur n'épouseront pas le nuage de points aussi bien que souhaité. Il en va de même pour les autres fonctions de profondeur telles que celles de Liu [15] ou de Mahalanobis [16] pour lesquelles la propriété ci-avant est encore vraie.

Considérons l'exemple de la distribution de Cauchy bivariée (qui est une distribution continue) [22]. Avant de déterminer les contours de cette distribution, attardons-nous sur certaines propriétés.

**Définition 1.6.** La fonction de densité d'une distribution de Cauchy univariée de paramètres de localisation $\mu \in \mathbb{R}$ et de dispersion $\sigma \in \mathbb{R}$ est donnée par

$$f(x) = \frac{1}{\pi\sigma\left(1 + \left(\frac{x-\mu}{\sigma}\right)^2\right)}.$$

**Définition 1.7.** Un vecteur aléatoire $X$ suit une distribution de Cauchy multivariée caractérisée par un paramètre de localisation $\mu \in \mathbb{R}^p$ et un paramètre de dispersion $\Sigma$ (une matrice carrée définie positive de dimension $p$) et notée $Cauchy_p(\mu, \Sigma)$ si et seulement si pour tout élément $a$ de $\mathbb{R}^p$, $a^T X$ est de distribution de Cauchy univariée [13].

Pour démontrer la Proposition 1.4, nous aurons besoin de définir ce qu'est la fonction caractéristique d'une variable aléatoire.

**Définition 1.8.** La fonction caractéristique d'une variable aléatoire $X$ est la fonction $\varphi_X : \mathbb{R} \to \mathbb{C} : t \mapsto E[e^{itX}]$.

Elle détermine de façon unique la loi de probabilité de $X$.

**Propriété 1.1.** *On peut vérifier que la fonction caractéristique*

*a) d'un vecteur aléatoire $X$ de $\mathbb{R}^p$ distribué selon la loi de Cauchy multivariée $Cauchy_p(\mu, \Sigma)$ est définie par $\varphi_X(t) = e^{it^T\mu - \|\Sigma^{1/2}t\|}$, $\forall t \in \mathbb{R}^p$ avec $\|t\| = \sqrt{t^T t}$.*

*b) d'une variable aléatoire $Z$ distribuée selon la loi de Cauchy univariée $Cauchy_1(\nu, \sigma^2)$ est donnée par $\varphi_Z(t) = e^{i\nu t - \sigma|t|}$, $\forall t \in \mathbb{R}$.*

**Proposition 1.4.** *Un vecteur aléatoire $X$ de $\mathbb{R}^p$ est distribué selon la loi $Cauchy_p(\mu, \Sigma)$ si et seulement si $a^T X$ est distribué selon la loi $Cauchy_1(a^T\mu, a^T\Sigma a)$.*

*Démonstration.* Au vu de la Définition 1.7, il ne nous reste plus qu'à donner la forme des paramètres de localisation et de dispersion. Pour cela nous allons utiliser les définitions des fonctions caractéristiques d'une distribution de Cauchy univariée et d'une distribution de Cauchy multivariée données par la Définition 1.1.

Supposons d'abord que $X \sim Cauchy_p(\mu, \Sigma)$. Dans ce cas,

$$\varphi_X(at) = e^{ita^T\mu - \|\Sigma^{1/2}at\|}, \forall a \in \mathbb{R}^p \text{ et } \forall t \in \mathbb{R}$$

Or $t$ étant un réel et $a$ un vecteur, on peut encore écrire

$$\varphi_X(at) = e^{ita^T\mu - \sqrt{a^T\Sigma a}|t|} = \varphi_{a^TX}(t), \forall a \in \mathbb{R}^p \text{ et } \forall t \in \mathbb{R}.$$

Supposons maintenant que $a^T X \sim Cauchy_1(a^T\mu, a^T\Sigma a)$. Dans ce cas,

$$\varphi_{a^TX}(t) = e^{ia^T\mu t - \sqrt{a^T\Sigma a}|t|}, \forall a \in \mathbb{R}^p \text{ et } \forall t \in \mathbb{R}.$$

En particulier, pour $t = 1$,

$$\varphi_{a^TX}(1) = e^{ia^T\mu - \|\Sigma^{1/2}a\|} = \varphi_X(a), \forall a \in \mathbb{R}^p.$$

La conclusion en découle.

□

**Lemme 1.1.** *Si $X$ et $Y$ sont deux variables indépendantes distribuées selon la loi de Cauchy univariée $Cauchy_1(0, 1)$, alors les distributions de $aX+bY$ et $(|a|+|b|)X$ coïncident pour tous $a, b$ réels positifs.*

*Démonstration.* Nous allons montrer que les fonctions caractéristiques de $aX + bY$ et $(|a| + |b|)X$ coïncident.

Soient $a, b \in \mathbb{R}^+$. Selon la Définition 1.1, $\varphi_X(t) = e^{-|t|}, \forall t \in \mathbb{R}$. De plus, par indépendance des variables aléatoires $X$ et $Y$,

$$\varphi_{aX+bY}(t) = \varphi_{aX}(t)\varphi_{bY}(t), \forall t \in \mathbb{R}$$

De plus, $\varphi_{aX}(t) = \varphi_X(at)$.

En effet, $E[aX] = aE[X]$ et $Var(aX) = a^2 Var(X)$ donc $\varphi_{aX}(t) = e^{-a|t|} = e^{-|at|} = \varphi_X(at)$, $\forall t \in \mathbb{R}$.

Ainsi,

$$\varphi_{aX+bY}(t) = e^{-|at|}.e^{-|bt|} = e^{-(|a|+|b|)|t|} = \varphi_{(|a|+|b|)X}(t)$$

□

**Proposition 1.5.** *La fonction de profondeur de demi-espace pour une distribution de Cauchy bivariée est constante sur des carrés autour de l'origine.*

*Démonstration.* Soit $(X, Y) \sim Cauchy_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$ avec $X$ et $Y$ des variables indépendantes. Par la Définition 1.7 avec $a = \begin{pmatrix}1\\0\end{pmatrix}$, la variable $X$ suit la loi de Cauchy univariée $Cauchy_1(0, 1)$. Il en va de même pour la variable $Y$ avec $a = \begin{pmatrix}0\\1\end{pmatrix}$.

Les variables $X$ et $Y$ étant indépendantes et au vu de la Définition 1.6, la fonction de densité jointe $f(x, y)$ est alors donnée par

$$f(x, y) = f(x).f(y) = \frac{1}{\pi^2}\left(\frac{1}{1+x^2}\right)\left(\frac{1}{1+y^2}\right).$$

Déterminons la forme prise par la fonction de profondeur de Tukey pour la distribution bivariée de Cauchy de $(X, Y)$.

Nous repartons de la Définition 1.5. Soit $(u, v) \in \mathbb{R}^2$ avec $0 \leqslant u < v$, cherchons à déterminer la profondeur de demi-espace de ce couple sous la distribution
$Cauchy_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$.
Le demi-espace qui réalise (1.1) pour $(u, v)$ est un des demi-espaces qui contiennent ce point sur leur frontière.

Définissons le demi-espace $H = \{(x, y) \in \mathbb{R}^2 : \alpha_1 x + \alpha_2 y \geqslant \alpha_1 u + \alpha_2 v\}$ pour $\alpha_1, \alpha_2 \in \mathbb{R}^+$ non nuls simultanément. Ce demi-espace contient le point $(u, v)$ sur sa frontière et ne contient pas le point $(0, 0)$.

On obtient,

$$\begin{aligned}\mathbb{P}((X,Y)\in H) &= \mathbb{P}(\alpha_1 X+\alpha_2 Y \geqslant \alpha_1 u+\alpha_2 v)\\ &\overset{(*)}{=} \mathbb{P}((\alpha_1+\alpha_2)X \geqslant \alpha_1 u+\alpha_2 v)\\ &= \mathbb{P}\left(X \geqslant \frac{\alpha_1 u+\alpha_2 v}{\alpha_1+\alpha_2}\right)\\ &= \mathbb{P}(X \geqslant \lambda u+(1-\lambda)v),\ \ \lambda=\frac{\alpha_1}{\alpha_1+\alpha_2}\in[0,1]\end{aligned}$$

où $(*)$ se justifie par le Lemme 1.1.

Cette probabilité est à minimiser. Or,

$$\mathbb{P}(X \geqslant \lambda u+(1-\lambda)v) = 1-F(\lambda u+(1-\lambda)v)$$

Il faut donc que $F(\lambda u+(1-\lambda)v)$ soit maximal et donc que $\lambda u+(1-\lambda)v$ soit maximal par la propriété de croissance de la fonction de répartition. Or, comme $u<v$, $\lambda u+(1-\lambda)v$ est maximal en $\lambda=0$.

Ainsi, la profondeur de demi-espace de $(u,v)$ pour la distribution bivariée de Cauchy est donnée par

$$D((u,v),Cauchy_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)) = \min(\mathbb{P}(X \geqslant \lambda u+(1-\lambda)v)) = \mathbb{P}(X \geqslant v).$$

Le cas $u>v\geqslant 0$ se traite de la même manière.
On obtient

$$\mathbb{P}((X,Y)\in H) = \mathbb{P}\left(X \geqslant \frac{\alpha_1 u+\alpha_2 v}{\alpha_1+\alpha_2}\right) = \mathbb{P}(X \geqslant (1-\gamma)u+\gamma v)$$

avec $\gamma=\dfrac{\alpha_2}{\alpha_1+\alpha_2}$. Cette probabilité est minimale en $\gamma=0$.

On a alors $D((u,v),Cauchy_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)) = \mathbb{P}(X \geqslant u)$.

Le cas des autres quadrants du plan se traite de façon analogue.

Nous obtenons, $\forall (x,y)\in\mathbb{R}^2$

$$D((x,y),Cauchy_2\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}1&0\\0&1\end{pmatrix}\right)) = \mathbb{P}(X \geqslant \max\{|x|,|y|\}) = 1-F(\max\{|x|,|y|\}).$$

En utilisant la Définition 1.6, on a

$$F(\max\{|x|,|y|\}) = \int_{-\infty}^{\max\{|x|,|y|\}} \frac{1}{\pi}\left(\frac{1}{1+t^2}\right)dt = \frac{1}{\pi}\arctan(\max\{|x|,|y|\}) - \frac{1}{2}$$

En conclusion, $D((x,y), Cauchy_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)) = \frac{1}{2} - \frac{1}{\pi}\arctan(\max\{|x|,|y|\})$

Ce qui conclut la démonstration puisque cette fonction est bien constante sur des carrés autour de zéro.

□

Les contours de profondeur de Tukey sont donc des carrés autour de zéro pour la distribution de Cauchy bivariée. Or les contours de densité d'une telle distribution ne sont pas convexes. Une représentation des contours de profondeur de Tukey et des contours de densité est donnée FIGURE 1.15.

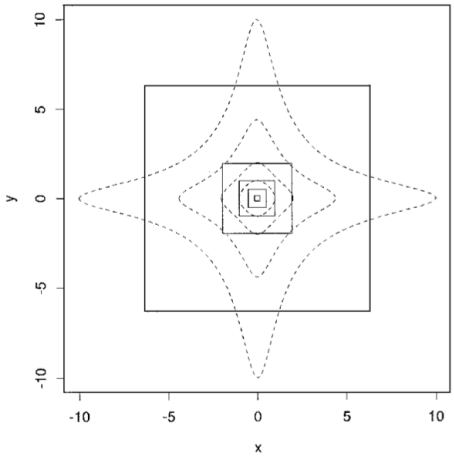


FIGURE 1.15 – Contours de profondeur pour $\alpha = 0.05, 0.15, 0.25, 0.35$ et $0.45$ (en lignes pleines) et les contours de densité de la distribution de Cauchy bivariée (en pointillés) [22].

Ce cas de figure peut également être visualisé empiriquement. Pour cela, nous avons généré dans le logiciel $R$ 400 observations bivariées telles que leur distribution soit en forme de banane. Pour cela, voici la démarche suivie dans le logiciel $R$. On peut d'abord générer une variable aléatoire $X$ uniforme sur l'intervalle $[-1,1]$ via la fonction *runif* du logiciel $R$. Ensuite, nous générons deux autres variables uniformes sur les intervalles $[0,2\pi]$ et $[0,1]$ que l'on nomme respectivement $\varphi$ et $Z$.
Nous construisons ensuite la matrice à deux colonnes et à autant de lignes que de valeurs générées pour $X$ telle que les éléments de la première colonne sont donnés par

$$X + R\cos(\varphi)$$

et les éléments de la deuxième colonne sont définis par

$$X^2 + R\sin(\varphi)$$

où $R = 0,2Z\left(1 + \frac{1-|X|}{2}\right)$.

Nous obtenons alors une représentation comme celle qui est donnée à la FIGURE 1.16.

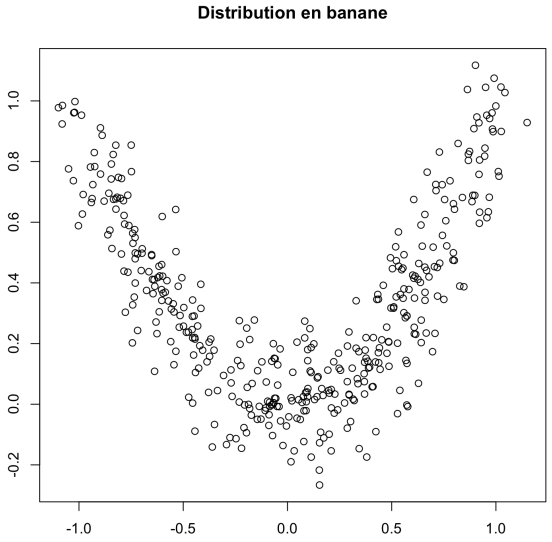


FIGURE 1.16 – Génération d'une distribution sous forme de banane

Nous pouvons ensuite représenter les contours de profondeur de Tukey pour ces observations. Ceux-ci sont construits à l'aide de la commande *isodepth* de la librairie *depth* du logiciel, appliquée à ces 400 observations. Nous pouvons constater sur la FIGURE 1.17 que les contours de profondeur de Tukey (représentés en bleu) pour une telle distribution ne caractérisent pas cette distribution.

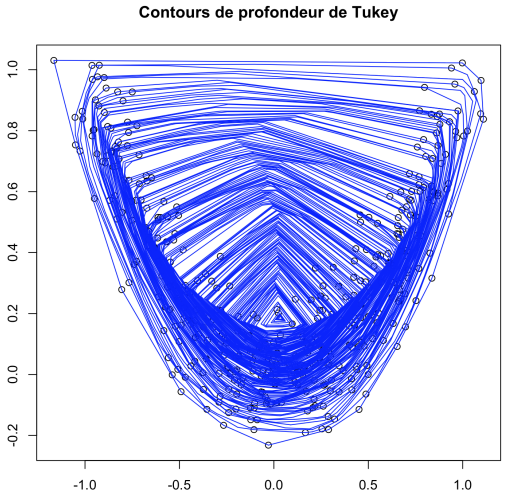


FIGURE 1.17 – Contours de profondeur de Tukey pour une distribution sous forme de banane, produits à partir d'un échantillon de 400 observations.

Nous pouvons faire de même dans le cas d'une distribution de Cauchy bivariée. Pour cela, nous générons deux échantillons de 400 observations issus tous les deux d'une distribution de Cauchy univariée de paramètre de localisation de valeur nulle et un paramètre de dispersion de valeur 1. Pour ce faire, nous utilisons la commande *rcauchy*. A partir de la FIGURE 1.18, les mêmes constatations que pour la distribution en forme de banane peuvent être faites. Seuls quelques contours ont été représentés afin d'assurer une meilleure visibilité.

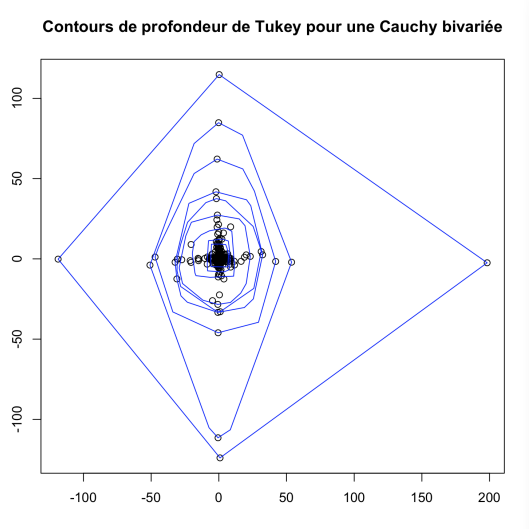


FIGURE 1.18 – Contours de profondeur de Tukey produits à partir d'un échantillon de 400 observations issues d'une distribution de Cauchy bivariée.

L'article *Monge-Kantorovich Depth, Quantiles, Ranks and Signs* de VICTOR CHERNOZHUKOV, ALFRED GALICHON, MARC HALLIN et MARC HENRY [7] présente une nouvelle fonction de profondeur qui s'adapte mieux que la profondeur de demi-espace au cas des distributions de contours de densité non-convexes. Ainsi, l'objectif de ce mémoire est de présenter une nouvelle fonction de profondeur dont les contours ne seront pas convexes, ainsi que les notions liées à cette nouvelle fonction : quantiles, rangs et signes. Cette nouvelle profondeur portera le nom de *profondeur de Monge-Kantorovich*.

# Chapitre 2

# Profondeur de Monge-Kantorovich

## 2.1 Introduction

L'idée suggérée par les auteurs de l'article *Monge-Kantorovich Depth, Quantiles, Ranks and Signs* [7] est de définir une nouvelle fonction de profondeur qui "collera" mieux aux distributions à contours non convexes telles que la distribution en forme de banane. C'est ce qu'ils ont fait en définissant une nouvelle fonction de profondeur dont les contours ne remplissent plus les propriétés d'équivariance affine et de convexité contrairement à la profondeur de demi-espace. Dans ce chapitre, nous considérons toujours des distributions continues.

Pour définir les notions de quantiles multivariés et de rangs de Monge-Kantorovich d'une distribution quelconque donnée, nous utiliserons la notion de *transport*. En effet, le principe sera de déterminer les régions et contours de profondeur d'une distribution d'intérêt sur $\mathbb{R}^p$ à partir d'une distribution de référence pour laquelle la forme des régions et des contours est connue. On dira que l'on *transporte* les régions et contours de profondeur d'une distribution de référence vers la distribution d'intérêt.

L'appellation *"Monge-Kantorovich depth"* vient d'ailleurs du fait que Monge et Kantorovich sont les deux mathématiciens ayant élaboré la *théorie du transport*. Pour plus d'informations sur cette théorie fort utilisée dans le monde économique, le lecteur est renvoyé à l'article *"Théorie générale du transport et applications"* de Cécile Carrère, Didier Lesesvre et Paul Pegon [6].

## 2.2 Le transport

Villani [28] définit le transport d'une mesure de probabilité $\mu$ par une application $T : \mathbb{R}^p \to \mathbb{R}^p$ par $T\#\mu(A) = \mu(T^{-1}(A))$, pour tout ensemble borélien $A$ [19] et où $T^{-1}(A) = \{x \in \mathbb{R}^p : T(x) \in A\}$. Ici, si $X$ et $Y$ sont deux vecteurs aléatoires, nous aimerions transporter une distribution de référence $P^X$ sur une distribution d'intérêt $P^Y$ via une application $T$ bien choisie. Donc, si $P^Y = T\#P^X$, alors $P^Y(A) = P^X(T^{-1}(A))$ par la définition donnée par Villani.
Ainsi,

$$\begin{aligned} P^Y(A) &= \mathbb{P}(X \in T^{-1}(A)) \\ &= \mathbb{P}(T(X) \in A) \end{aligned}$$

donc, si la fonction de répartition de $X$ est notée $F_X$ et la fonction de répartition de $Y$ est notée $F_Y$ alors

$$F_Y(.) = F_X(T^{-1}(.)) \tag{2.1}$$

Illustrons cela via un exemple.

### 2.2.1 Exemple de transport en dimension 1

Considérons $X \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{X}_1^2$ (loi $\mathcal{X}^2$ à 1 degré de liberté). Le transport de la distribution normale sur la distribution $\mathcal{X}^2$ consiste en la recherche d'une application $T$ telle que $T\#\mathcal{N}(0,1) = \mathcal{X}_1^2$.

Autrement dit, au vu de ce qui précède, nous cherchons une application $T$ telle que pour $x \geqslant 0$,

$$\mathbb{P}(T(X) \leqslant x) = \frac{1}{2^{1/2}\sqrt{\pi}} \int_0^x u^{-1/2} e^{-u/2} du \text{ alors que } X \sim \mathcal{N}(0,1)$$

Or, $\mathbb{P}(T(X) \leqslant x) = \mathbb{P}(X \in T^{-1}(]-\infty, x])) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{T^{-1}(]-\infty,x])} e^{-u^2/2} du.$

L'application de transport qui convient est $T(.) = (.)^2$. En effet, pour $x \geqslant 0$,

$$\begin{aligned} \mathbb{P}(X^2 \leqslant x) &= \mathbb{P}(-\sqrt{x} \leqslant X \leqslant \sqrt{x}) \\ &= \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{x}} e^{-u^2/2} du \end{aligned}$$

En effectuant un changement de variable,

$$\mathbb{P}(X^2 \leqslant x) = \frac{1}{\sqrt{2\pi}} \int_0^x \frac{1}{\sqrt{v}} e^{-v/2} dv = \mathbb{P}(Y \leqslant x).$$

Ce résultat est en accord avec la théorie puisque si une variable aléatoire est distribuée selon la loi $\mathcal{N}(0,1)$ alors le carré de cette variable aléatoire est distribué selon une loi $\mathcal{X}_1^2$.

### 2.2.2 Transport en dimension 1 pour une application croissante

Considérons $X$ et $Y$ deux variables aléatoires continues de fonctions de répartition $F_X$ et $F_Y$ respectivement et de fonctions de densité associées $f_X$ et $f_Y$.

Supposons que l'application de transport $T$ qui envoie la loi de $X$ sur la loi de $Y$ soit croissante strictement et dérivable et déterminons une forme générale pour cette application $T$.
En dérivant l'égalité (2.1), on obtient

$$f_X(T^{-1}(x))\frac{1}{T'(T^{-1}(x))} = f_Y(x),\ \forall x \in \mathbb{R}$$

Posons $T^{-1} = S$. Nous cherchons l'application $S$ qui vérifie

$$f_X(S(x)) = (S^{-1})'(S(x))f_Y(x)$$

L'application $S$ est donc solution d'une équation différentielle.

Considérons maintenant le transport d'une variable aléatoire $X$ continue vers la variable aléatoire $F_X(X)$.

Tout d'abord, $F_X(X) \sim U_{[0,1]}$. En effet,

$$\mathbb{P}(F_X(X) \leqslant y) = \mathbb{P}(X \leqslant F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y$$

Ainsi, nous pouvons écrire que $X \overset{\mathcal{L}}{=} F_X^{-1}(U)$, avec $U \sim U_{[0,1]}$.

Nous obtenons alors

$$S\#P^U = P^X, \text{ avec } S = F_X^{-1}$$

L'application $F_X^{-1} : ]0,1[ \to \mathbb{R}$ est donc une application qui transporte la loi $U_{[0,1]}$ sur n'importe quelle loi.

De plus, en dimension 1, cette application $F_X^{-1}$ n'est rien d'autre que la fonction quantile habituelle. En effet, $F_X^{-1}(p)$ est un réel tel que $\mathbb{P}(X \leqslant F_X^{-1}(p)) = p$.

Ainsi, nous pouvons donner une définition de la fonction quantile en termes de transport.

**Définition 2.1.** La fonction quantile habituelle est la fonction qui transporte la loi uniforme sur $[0,1]$ sur la loi d'une variable aléatoire d'intérêt.

Nous pouvons aussi déterminer l'application qui transporte une loi quelconque sur la loi uniforme sur $[0,1]$. En effet, $F_X(X) \overset{\mathcal{L}}{=} U$ donc,

$$F_X\#P^X = P^U$$

Or nous avons vu à la Section 1.7 du Chapitre 1 que la fonction de répartition empirique coïncide avec les rangs normalisés. De plus, la fonction de répartition empirique transporte la loi de $X_1, ..., X_n$ variables aléatoires, indépendantes et identiquement distribuées, avec $n \in \mathbb{N}_0$, sur la loi uniforme discrète sur $\left\{\frac{1}{n+1}, ..., \frac{n}{n+1}\right\}$. En effet, $\frac{1}{n+1} rg^{(n)}(X_i) \in \left\{\frac{1}{n+1}, ...., \frac{n}{n+1}\right\}$ et $\mathbb{P}\left(\frac{1}{n+1} rg^{(n)}(X_i) = \frac{j}{n+1}\right) = \frac{1}{n}$.

Nous pouvons donc donner une définition des rangs empiriques normalisés comme une application de transport.

**Définition 2.2.** La fonction des rangs empiriques normalisés est l'application qui transporte la loi de $n \in \mathbb{N}_0$ variables aléatoires indépendantes et identiquement distribuées $X_1, ..., X_n$, sur la loi uniforme discrète sur $\left\{\frac{1}{n+1}, ..., \frac{n}{n+1}\right\}$.

Par le théorème de Glivenko-Cantelli [10], $F_X^{(n)}(x) \to F_X(x)$ lorsque $n \to +\infty$. On en déduit donc la définition suivante :

**Définition 2.3.** Le rang est défini comme l'application de transport de la loi d'une variable aléatoire quelconque $X$ sur la loi uniforme sur $[0, 1]$.

Cependant, chacune des définitions précédentes est basée sur les notions de quantiles et rangs habituels, i.e elles se basent sur une notion d'ordre comme illustré à la Figure 2.1. Donc, ces concepts ne peuvent pas être généralisés sous cette forme en dimension supérieure.

L'idée est alors de définir une fonction qui remplacera la fonction de répartition $F$ qui se basait dans le cas empirique sur l'arrangement des valeurs par ordre croissant. Pour cela, nous regardons la répartition à partir du centre comme illustré à la Figure 2.2.


Figure 2.1 – Fonction de répartition classique

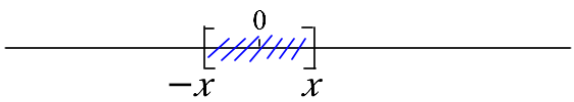


Figure 2.2 – Fonction de répartition définie à partir du centre (pris en zéro) pour $x \geqslant 0$

Supposons que la médiane est nulle, quitte à translater la distribution. Cette nouvelle fonction de répartition notée $F_\pm : \mathbb{R} \to [-1, 1]$ est définie, pour $x \geqslant 0$ par

$$F_\pm(x) := F(x) - F(-x)$$

Ce qui correspond à l'aire sous la courbe de densité entre les points $-x$ et $x$.

En particulier,

$$F_\pm := 2F - 1, \text{ si } F \text{ est symétrique par rapport à la médiane.}$$

Pour l'exemple ci-après, nous ne considérerons que le cas symétrique.
Si $X \sim F$ alors, comme $F(X) \sim U_{[0,1]}$, $F_\pm(X) \sim U_{[-1,1]}$ et par croissante de $F$, la fonction $F_\pm$ est croissante.

Cette fonction de répartition ainsi définie est la fonction qui transporte une distribution d'intérêt quelconque $P^X$ sur la distribution uniforme sur $[-1, 1]$.

De plus, une version empirique de cette nouvelle fonction de répartition pour des variables aléatoires $X_1, ..., X_n$ indépendantes et identiquement distribuées est définie comme suit,

$$F_\pm^{(n)}(X_i) := S_i \frac{R^{(n)}(X_i)}{n+1} \text{ pour } i \in \{1, ..., n\}$$

où $S_i$ est le signe du rang $R^{(n)}(X_i)$, à savoir

$$S_i = \begin{cases} 1 \text{ si } rg^{(n)}(X_i) > \dfrac{n+1}{2} \\ \\ -1 \text{ si } rg^{(n)}(X_i) < \dfrac{n+1}{2} \end{cases}$$

Ainsi, la fonction de répartition $F_\pm^{(n)}$ transporte une loi quelconque sur la loi uniforme discrète sur $\left\{ \dfrac{-n}{2(n+1)}, ..., \dfrac{-1}{n+1}, \dfrac{1}{n+1}, ..., \dfrac{n}{2(n+1)} \right\}$ si $n$ est pair et sur $\left\{ \dfrac{-(n-1)}{2(n+1)}, ..., 0, ..., \dfrac{n-1}{2(n+1)} \right\}$ si $n$ est impair.

Donc, la fonction de transport d'une distribution d'intérêt sur la distribution uniforme sur $[-1, 1]$ est appelée la fonction rang (par analogie au cas où l'on considérait la fonction de répartition $F$) lorsqu'on applique le théorème de Glivenko-Cantelli à $F_\pm^{(n)}$.

Quant à la fonction de transport inverse, il s'agit de la fonction quantile définie à partir de la fonction de répartition $F_\pm$. Autrement dit, le transport de la distribution uniforme sur $[-1, 1]$ sur une distribution d'intérêt $P^X$ se fera via la fonction quantile définie par $Q_\pm = F_\pm^{-1}$, par analogie du transport via la fonction $F^{-1}$.

C'est sur cette intuition que Chernozhukov et al. [7] se basent pour définir les quantiles et rangs de Monge-Kantorovich en dimension $p$ quelconque.

## 2.3 Définitions et propriétés relatives à l'application du transport

Afin d'effectuer le transport d'une distribution de référence bien connue $F$ vers une distribution d'intérêt $P$, VICTOR CHERNOZHUKOV et al. [7] proposent d'utiliser le gradient d'une fonction convexe $\psi$ comme application de transport et ils notent celui-ci "$\bigtriangledown\psi$" tel que $\bigtriangledown\psi : X \mapsto Y$ avec $X$ et $Y$ des vecteurs aléatoires de distribution $F$ et $P$ respectivement.
L'existence d'une telle fonction convexe $\psi$ est assurée par le théorème [7] de BRENIER [3] et MCCANN [18] :

**Théorème 2.1.** *Soient $F$ et $P$ deux distributions sur $\mathbb{R}^p$. Si la distribution $F$ est absolument continue sur $\mathbb{R}^p$ pour la mesure de Lebesgue et à support inclus dans un ensemble convexe $\mathcal{X}$ alors il existe une fonction convexe $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ telle que le gradient de cette fonction existe et est unique presque partout et telle que ce gradient est la fonction qui transporte $F$ sur $P$. De plus, si $P$ vérifie les mêmes hypothèses que $F$ avec son support inclus dans un ensemble convexe $\mathcal{Y}$ alors il existe une fonction $\psi^* : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ dont le gradient transporte $P$ sur $F$ et telle que $\bigtriangledown\psi^* = \bigtriangledown\psi^{-1}$ presque partout.*

Ce théorème est accepté sans démonstration.

En dimension 1, la fonction $F_{\pm}^{-1}$ définie à la Section 2.2 de ce chapitre étant croissante, elle est la dérivée d'une fonction convexe $\psi$ par la Proposition 2.1.

**Proposition 2.1.** *Si $f$ est dérivable sur un intervalle $I$ de $\mathbb{R}$, $f$ est convexe si et seulement si sa dérivée est croissante.*

L'existence de la fonction convexe est assurée par le Théorème 2.1.

Quant à l'existence du gradient d'une fonction convexe choisie (en dimension quelconque), elle est assurée par les deux propriétés suivantes, pour lesquelles nous avons tout d'abord besoin d'introduire les définitions d'une application *localement lipschitzienne* et d'une fonction *différentiable*.

**Définition 2.4.** Soient $\Omega$ une partie de $\mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ et $\psi : \Omega \mapsto \mathbb{R}^p$ une application. Pour tout $L > 0$, l'application $\psi$ est dite *L-lipschitzienne* par rapport à $x_2$ sur $\Omega$ si

$$|\psi(x_1, y_2) - \psi(x_1, x_2)| \leqslant L|y_2 - x_2| \tag{2.2}$$

$\forall (x_1, x_2), (x_1, y_2) \in \Omega$.

Si la condition (2.2) est satisfaite pour une valeur de $L$ au moins, alors $\psi(x_1, x_2)$ sera dite *lipschitzienne* par rapport à $x_2$ dans $\Omega$.

Une telle application sera qualifiée de *localement lipschitzienne* par rapport à $x_2$ dans $\Omega$ si tout point de $\Omega$ possède un voisinage sur lequel $\psi(x_1, x_2)$ est *lipschitzienne* par rapport à $x_2$ sur $\Omega$.

**Définition 2.5.** Soit $f$ une fonction définie sur un ouvert $\mathcal{U}$ de $\mathbb{R}^p$ et à valeurs dans $\mathbb{R}^n$ et soit $a \in \mathcal{U}$. La fonction $f$ est différentiable en $a$ s'il existe une application linéaire $l : \mathbb{R}^p \mapsto \mathbb{R}^n$ telle que

$$f(a+h) = f(a) + l(h) + o(\|h\|)$$

pour $h \to 0$ dans $\mathcal{U}$.
Si une telle application $l$ existe, alors elle est unique et est appelée *application linéaire tangente* de $f$ en $a$ ou encore *différentielle* de $f$ en $a$. Dans ce cas,

$$l(h) = \lim_{t \to 0} \frac{f(a+th) - f(a)}{t}$$

**Proposition 2.2.** *Si $\psi$ est une fonction convexe dont le domaine de définition est donné par $dom\psi = \{x \in \mathcal{X} : \psi(x) < +\infty\}$ où $\mathcal{X}$ est un ensemble convexe de $\mathbb{R}^p$, et si $\psi$ est borné sur un voisinage $V$ de $x$, $\forall x \in dom\psi$ alors $\psi$ est localement lipschitzienne sur l'intérieur de ce voisinage.*

*Démonstration.* Commençons par définir quelques notations. Tout d'abord, $A^\circ$ désigne *l'intérieur* de l'ensemble $A$. Ensuite, nous noterons $\overline{A}$ *l'adhérence* de l'ensemble $A$. Pour plus d'informations concernant ces deux derniers concepts, le lecteur est renvoyé à la lecture du document [17].

Par hypothèse, $\psi$ étant bornée sur un voisinage de $x$, il existe une constante $C > 0$ ainsi qu'un voisinage $V$ de $x$ tels que $|\psi(z)| \leqslant C$ pour tout $z \in V$. Soit $z \in V^\circ$, par définition de l'intérieur, il existe $r > 0$ tels que $B(z,r) \subset V$ et, quitte à prendre un rayon $r$ plus petit, $\overline{B(z,r)} \subset V$.

Considérons $y, y' \in \overline{B(z, \frac{r}{2})}$ avec $y \neq y'$ et posons $R := \|y - y'\|$. Tout d'abord, $R \in ]0, r]$ car $\|y - y'\| = \|y - z + z - y'\| \leqslant \|y - z\| + \|z - y'\| \leqslant \frac{r}{2} + \frac{r}{2} \leqslant r$.

Ensuite, définissons $y''$ par :

$$y'' = y' + \frac{r}{2R}(y' - y).$$

On a alors $y'' \in \overline{B(z,r)}$. En effet, $\|y'' - z\| \leqslant \|y' - z\| + \left\|\frac{r}{2R}\right\| \|y' - y\| \leqslant \frac{r}{2} + \frac{r}{2} \leqslant r$.

De plus, $\left(1 + \frac{r}{2R}\right) y' = y'' + \frac{r}{2R} y \Leftrightarrow y' = \frac{2R}{2R+r} y'' + \frac{r}{2R+r} y$.

En utilisant la propriété de convexité de $\psi$ ainsi que le caractère borné de cette fonction, nous pouvons écrire

$$\begin{aligned}
\psi(y') &\leqslant \frac{2R}{2R+r}\psi(y'') + \frac{r}{2R+r}\psi(y) \\
&= \psi(y) + \frac{2R}{2R+r}(\psi(y'') - \psi(y)) \\
&\leqslant \psi(y) + \frac{4C}{r}\|y - y'\|
\end{aligned}$$

On obtient alors $\psi(y') - \psi(y) \leqslant \frac{4C}{r}\|y' - y\|$ et le résultat est ainsi démontré car il suffit d'intervertir les rôles de $y$ et $y'$ afin d'obtenir l'inégalité $\frac{4C}{r}\|y' - y\| \leqslant \psi(y') - \psi(y)$ et ainsi retomber sur la définition d'une fonction localement lipschitzienne. □

**Proposition 2.3.** *La fonction convexe $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ est différentiable presque partout sur le domaine de définition de $\psi$.*

*Démonstration.* Le résultat est démontré grâce à la proposition précédente ainsi que par le théorème de Rademacher selon lequel toute fonction localement lipschitzienne est différentiable presque partout. □

Le théorème de Rademacher n'étant utilisé que pour montrer que la fonction $\psi$ est bien différentiable et donc que tout est bien défini, ne sera pas démontré ici. Si le lecteur est intéressé, il est renvoyé à l'article [12].

NB : Dans la suite, nous noterons toujours $F$ la distribution de référence et $P$ la distribution d'intérêt si celles-ci ne sont pas spécifiées à l'avance.

## 2.4 Définitions des vecteurs quantiles et rangs de Monge-Kantorovich

Tout d'abord, dans cette partie nous allons faire appel à la notion de distribution uniforme sphérique. Nous commençons donc par définir une telle distribution.

**Définition 2.6.** La distribution uniforme sphérique est la distribution d'un vecteur aléatoire $r\phi$, où $r$ est distribué uniformément sur l'intervalle $[0, 1]$ et $\phi$ est distribué uniformément sur la sphère $\mathcal{S}^{p-1} := \{x \in \mathbb{R}^p : \|x\| = 1\}$, $r$ et $\phi$ étant indépendants [7].

Les définitions qui suivent font appel à l'application $\psi$ mais aussi à son application conjuguée, notée $\psi^*$ par la suite, et dont voici une définition.

**Définition 2.7.** L'application conjuguée de $\psi : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ est définie par l'application $\psi^* : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ telle que

$$\forall y \in \mathcal{Y}, \psi^*(y) := \sup_{z \in \mathcal{X}} [y^T z - \psi(z)]$$

**Définition 2.8.** Considérons une distribution de référence $F$ absolument continue dont le domaine de définition est inclus dans un ensemble convexe $\mathcal{X}$ de $\mathbb{R}^p$ et $P$ une distribution quelconque dont le domaine est inclus dans un ensemble convexe $\mathcal{Y}$ de $\mathbb{R}^p$. On définit le vecteur des quantiles de Monge-Kantorovich par :

$$Q(x, P) \in \arg\sup_{y \in \mathcal{Y}} [y^T x - \psi^*(y)] \text{ pour } x \in \mathcal{X}$$

**Définition 2.9.** Sous les mêmes conditions que la définition du vecteur des quantiles, nous pouvons définir le vecteur des rangs de Monge-Kantorovich par :

$$R(y,P) \in \arg\sup_{x\in\mathcal{X}}[y^Tx - \psi(x)] \text{ pour } y \in \mathbb{R}^p$$

Nous allons pouvoir donner une forme générale à ces vecteurs quantiles et rangs de Monge-Kantorovich, mais pour cela nous avons besoin du théorème de l'enveloppe que nous allons énoncer ci-après. Ce théorème s'intéresse à la façon dont la valeur d'une fonction évaluée en un de ses extrema varie en fonction de l'évolution d'un paramètre de cette fonction.

**Théorème 2.2.** ***(de l'enveloppe)***
*Soient $f : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}$ une fonction de classe $C^1$ et $x^* : \mathcal{U} \to \mathbb{R}^{p_1}$ une fonction de classe $C^1$ où $\mathcal{U}$ est un ouvert de $\mathbb{R}^{p_2}$ telle que, $\forall\alpha \in \mathcal{U}$, $x^*(\alpha)$ est un extremum de la fonction $f$. On définit alors une fonction $V : \mathbb{R}^{p_2} \to \mathbb{R}$ par $V : \alpha \mapsto f(x^*(\alpha), \alpha)$. Cette fonction ainsi définie est telle que,*

$$\forall\alpha \in \mathcal{U} \text{ et } \forall i \in \{1,...,p_2\},\ \frac{\partial}{\partial\alpha_i}V(\alpha) = \frac{\partial f}{\partial\alpha_i}(x^*(\alpha),\alpha).$$

*Démonstration.* La démonstration de ce théorème découle du théorème de dérivation des fonctions composées. En effet, si la fonction $f$ est définie par $f(x_1, x_2, ...., x_{p_1}, \alpha)$ alors

$$\frac{\partial}{\partial\alpha_i}V(\alpha) = \sum_{j=1}^{p_1}\frac{\partial f}{\partial x_j}(x^*(\alpha),\alpha)\frac{\partial x_j}{\partial\alpha_i}(\alpha) + \frac{\partial f}{\partial\alpha_i}(x^*(\alpha),\alpha), \text{ pour } i \in \{1,...,p_2\}.$$

Or, $\frac{\partial f}{\partial x_j}(x^*(\alpha),\alpha) = 0\ \forall j \in \{1,...p_1\}$ étant donné que $x^*(\alpha)$ a été défini comme une fonction du paramètre $\alpha$ qui maximise $f$. □

Nous sommes maintenant en mesure de définir le vecteur rang et le vecteur quantile de Monge-Kantorovich dans le contexte multivarié.

**Proposition 2.4.** *Le vecteur des quantiles de Monge-Kantorovich est tel que $Q(.,P) = \bigtriangledown\psi(.)$ presque partout sur $\mathcal{X}$ par rapport à la mesure de Lebesgue.*

*De plus, le vecteur des rangs de Monge-Kantorovich est tel que $R(.,P) = \bigtriangledown\psi^*(.)$ presque partout sur $\mathcal{Y}$ par rapport à la mesure de Lebesgue.*

*Démonstration.* Ce résultat est direct puisque $Q(x,P) \in \arg\sup_{y\in\mathcal{Y}}[y^Tx - \psi^*(y)]$ pour $x \in \mathcal{X}$. Or,

$$\bigtriangledown_y(y^Tx - \psi^*(y)) = 0 \Leftrightarrow x = \bigtriangledown_y\psi^*(y)$$

et donc par le Théorème 2.1 $\bigtriangledown\psi(x) = y$ presque partout, avec $y = Q(x,P)$.

De même pour $R(y,P) \in \arg\sup_{x\in\mathcal{X}}[y^Tx - \psi(x)]$ pour $y \in \mathbb{R}^p$,

$$\bigtriangledown_x(y^T x - \psi(x)) = 0 \Leftrightarrow y = \bigtriangledown_x \psi(x)$$

Par le Théorème 2.1, $\bigtriangledown \psi^*(y) = x$ presque partout, avec $x = R(y, P)$.

□

Les quantiles de Monge-Kantorovich correspondent à une application qui transporte la distribution uniforme sphérique sur la boule unité en dimension $p$ notée $U_p$ sur une distribution d'intérêt $P$ de $\mathbb{R}^p$. Si cette distribution $P$ possède des moments finis d'ordre 2, alors cette application sera appelée *transport optimal* dans le sens qu'elle minimise le coût quadratique moyen :

$$\min_T E(T(U) - U)^2$$

avec $U \sim U_p$. Le lemme suivant dû à CAFFARELLI [5] nous permettra d'étendre les égalités annoncées dans la propriété précédente.

**Lemme 2.1.** *Soient $\mathcal{X}$ et $\mathcal{Y}$ deux ouverts convexes de $\mathbb{R}^p$ et soient $f$ et $g$ deux fonctions de densité bornées, de domaine de définition $\mathcal{X}$ et $\mathcal{Y}$ respectivement. Si $f$ et $g$ sont de classe $C^\alpha$ avec $\alpha > 0$ sur $\overline{\mathcal{X}}$ et $\overline{\mathcal{Y}}$ respectivement, alors $\psi$ est de classe $C^{\alpha+2}$ sur $\mathcal{X}$.*

Grâce à ce lemme, nous obtenons le résultat suivant.

**Proposition 2.5.** *Si $\mathcal{X}_0$ est un ouvert de $\overline{\mathcal{X}}$ et si $\mathcal{Y}_0$ est un ouvert de $\overline{\mathcal{Y}}$ alors $\bigtriangledown\psi|_{\mathcal{X}_0} : \mathcal{X}_0 \to \mathcal{Y}_0$ et $\bigtriangledown\psi^*|_{\mathcal{Y}_0} : \mathcal{Y}_0 \to \mathcal{X}_0$ sont des homéomorphismes et l'égalité suivante est vérifiée $\bigtriangledown\psi|_{\mathcal{X}_0} = (\bigtriangledown\psi^*|_{\mathcal{Y}_0})^{-1}$.*

Rappelons qu'une application est un homéomorphisme si cette application est continue, bijective et si son application réciproque est continue.

Cette propriété est en effet vérifiée car par le Théorème de Brenier-McCann 2.1, la fonction convexe $\psi$ possède un gradient qui est unique et par le lemme précédent, le gradient de cette fonction convexe est continu.
De cette proposition, on obtient que $Q(., P) = \bigtriangledown\psi$ sur $\mathcal{X}_0$ et $R(., P) = \bigtriangledown\psi^*$ sur $\mathcal{Y}_0$.

**Proposition 2.6.** *La distribution uniforme sphérique est une distribution à symétrie sphérique.*

*Démonstration.* Prenons les mêmes notations que celles utilisées dans la Définition 2.6. Notons $r\phi$ un vecteur aléatoire distribué selon une loi uniforme sphérique.

Soit $A \in \mathcal{O}(p)$. Le vecteur aléatoire $r\phi$ suit une distribution à symétrie elliptique si $A(r\phi) \overset{\mathcal{L}}{=} r\phi$.

Or, l'égalité $A(r\phi) = r(A\phi)$ est satisfaite et comme $\phi$ est distribué uniformément sur la sphère $\mathcal{S}^{p-1}$, $A\phi$ est également distribué uniformément sur cette sphère. De plus, $A\phi$ et $r$ sont indépendants puisque $\phi$ et $r$ sont indépendants par définition.

Ainsi, $r(A\phi)$ est de distribution uniforme sphérique.

□

L'idée de CHERNOZHUKOV et al. [7] est de transformer les contours et régions de profondeur de la distribution uniforme sphérique $U_p$ relatifs à la profondeur de demi-espace (qui sont connus et sphériques au vu de la Propriété 1.2) en les contours et régions de profondeur d'une distribution d'intérêt $P$. Ils définissent alors la fonction de profondeur, les quantiles, les rangs et les signes de Monge-Kantorovich comme suit.

**Définition 2.10.** Soient $X$ un vecteur aléatoire issu de la distribution uniforme sphérique notée $U_p$ sur la boule unité $\mathcal{X} = \mathbb{S}^p := \{x \in \mathbb{R}^p : \|x\| \leqslant 1\}$ et $P$ la fonction de répartition issue d'une distribution quelconque dont le support est inclus dans un ensemble convexe $\mathcal{Y}$ de $\mathbb{R}^p$. Les quantiles, rangs et signes de Monge-Kantorovich sont définis comme suit

- Le rang d'une valeur $y$ de $\mathbb{R}^p$ est donné par $\|R(y,P)\|$ et le signe est défini par $\dfrac{R(y,P)}{\|R(y,P)\|}$.
- Le contour de profondeur d'ordre $\alpha$ de Monge-Kantorovich est donné par $Q(B(0,\alpha)^\bullet, P)$. Quant à la région de profondeur de probabilité $\alpha$, elle est donnée par $Q(B(0,\alpha), P)$. La notation $B(0,\alpha)$ désigne la boule de centre 0 et de rayon $\alpha$ dans $\mathbb{R}^p$ et $B(0,\alpha)^\bullet := \{x \in \mathbb{R}^p : \|x\| = \alpha\}$ .
- La profondeur de Monge-Kantorovich d'une valeur $y$ de $\mathbb{R}^p$ est définie comme étant la profondeur de demi-espace du vecteur $R(y,P)$ sous la distribution uniforme sphérique ce que l'on note : $MK(y,P) := D(R(y,P), U_p)$.

La profondeur de Monge-Kantorovich est définie comme la profondeur de demi-espace de l'application qui transporte une loi d'intérêt quelconque $P$ sur la loi uniforme sphérique $U_p$. La distribution de référence est une distribution à symétrie sphérique. Pour une telle distribution, nous connaissons la forme des contours de profondeur de Tukey et nous souhaitons transporter ces contours afin d'obtenir les contours d'une distribution d'intérêt quelconque.

Il est possible de généraliser cette définition au cas d'une distribution de référence absolument continue et quelconque. Toutefois, ayant travaillé exclusivement avec la profondeur de demi-espace dans ce mémoire, la définition précédente nous suffit pour les raisons évoquées ci-avant. Si le lecteur est intéressé par la généralisation de la définition de la profondeur de Monge-Kantorovich, il est renvoyé à la Définition 2.3 de l'article *Monge-Kantorovich Depth, Quantiles, Ranks and Signs* de CHERNOZHUKOV et al. [7].

# Chapitre 3

# Propriétés de la profondeur de MK

## 3.1 Introduction

Dans ce chapitre, nous nous attarderons sur les propriétés de la nouvelle profondeur définie dans l'article de CHERNOZHUKOV et al. [7]. Nous reviendrons également sur les familles de distributions particulières évoquées aux Sous-Sections 1.3.1 et 1.3.2 du Chapitre 1 pour lesquelles nous étudierons la fonction de profondeur de Monge-Kantorovich. Nous constaterons que cette nouvelle fonction de profondeur coïncide avec la profondeur de Tukey dans le cas des distributions issues des familles $\mathcal{P}^1$ et des distributions à symétrie sphérique.

Nous commencerons par l'étude théorique de cette fonction de profondeur pour les familles de distributions précitées, puis nous nous placerons dans le cadre empirique afin d'appliquer cette théorie. Nous considérons toujours le cas de distributions continues.

A la fin de ce chapitre, nous donnerons une procédure d'implémentation des contours de profondeur de Monge-Kantorovich dans le logiciel *R*.

## 3.2 Caractérisation des contours et régions de profondeur

**Propriété 3.1.** *Les contours de profondeur de Monge-Kantorovich définis ci-avant peuvent être déformés en sphères grâce au transport.*

En effet, les contours de profondeur de Monge-Kantorovich étant donnés par $Q(B(0,\alpha)^\bullet, P)$ et les applications $R(.,P)$ et $Q(.,P)$ étant réciproques, on obtient que $R(Q(B(0,\alpha)^\bullet, P), P) = B(0,\alpha)^\bullet$. On transporte donc les contours de profondeur de Monge-Kantorovich grâce à l'application de transport $R(.,P)$ et nous obtenons des contours sphériques. C'est dans ce sens que nous parlons de déformations en sphères.

**Propriété 3.2.** *Les régions de profondeur de Monge-Kantorovich sont emboîtées.*

En effet, si on considère $\alpha, \alpha' > 0$ tels que $\alpha > \alpha'$ alors $B(0, \alpha') \subset B(0, \alpha)$ et donc $Q(B(0, \alpha'), P) \subset Q(B(0, \alpha), P)$.

## 3.3 Profondeur de Monge-Kantorovich pour les familles $\mathcal{P}^1$ et $\mathcal{P}^p_{ell}$

Revenons aux familles particulières $\mathcal{P}^1$ et $\mathcal{P}^p_{ell}$ que nous avons considérées dans le premier chapitre. Pour ces familles, nous avons deux propriétés de la profondeur de Monge-Kantorovich qui permettent d'affirmer que la définition établie pour cette profondeur est idéale.

Mais avant d'établir ces propriétés nous avons besoin d'introduire les notations suivantes. Pour commencer, pour tout vecteur aléatoire $X \sim P$ où $P$ est une distribution quelconque de paramètre central $\mu$ et de paramètre de dispersion $\Sigma$ (une matrice carrée définie positive), nous noterons $\|X\|_{\mu,\Sigma}$ l'expression $\sqrt{(X-\mu)^T\Sigma^{-1}(X-\mu)}$. Cette expression est appelée *"Pseudo Mahalanobis distance"* [11] où le mot "pseudo" est utilisé pour signaler que l'on travaille avec une matrice $\Sigma$ qui n'est pas nécessairement la matrice de variance-covariance associée à la distribution (donc $\|X\|_{\mu,\Sigma}$ n'est pas la vraie distance de Mahalanobis).

**Propriété 3.3.** *Dans le cas de la famille $\mathcal{P}^1$ avec une distribution $P$ symétrique par rapport à la médiane, la fonction de profondeur $MK(., P)$ de Monge-Kantorovich coïncide avec la fonction de profondeur de Tukey $D(., P)$.*

*Démonstration.* Repartons de la Définition 2.10 avec $p = 1$ et $P$ la fonction de répartition relative à une distribution symétrique issue de la famille $\mathcal{P}^1$. Nous voulons montrer que

$$MK(x, P) = D(x, P),\ \forall x \in \mathbb{R}$$

Or, nous avons obtenu à la Sous-Section 1.3.1 que $D(x, P) = \min\{P(x), 1-P(x)\}, \forall x \in \mathbb{R}$.

De plus, par la Définition 2.10, $MK(x, P) = D(R(x, P), U_1),\ \forall x \in \mathbb{R}$.

Définissons le vecteur des rangs $R(x, P)$. Tout d'abord, $R(., P)$ transporte par définition la distribution d'intérêt vers la distribution de référence $U_1$ qui n'est rien d'autre que la distribution uniforme sur l'intervalle $[-1, 1]$.

Considérons donc $Y$ une variable aléatoire issue de la distribution $U_1$. La fonction de répartition d'une telle variable aléatoire est donnée par $F(y) = \dfrac{y+1}{2}$. Le vecteur des rangs de Monge-Kantorovich est défini par

$$R(x, P) = 2P(x) - 1,\ \forall x \in \mathbb{R}$$

puisque à la Sous-Section 2.2.1, nous avons remarqué que la fonction $F_{\pm}$ transporte une loi d'intérêt sur la loi uniforme sur $[-1, 1]$. Autrement dit, $F_{\pm}(.) = R(., P)$

Cette application est bien la dérivée d'une fonction convexe et elle est unique par le Théorème 2.1.
Ensuite, nous pouvons écrire que

$$D(R(x, P), U_1) = \inf\{\mathbb{P}_{U_1}(2P(X) - 1 \in H) : H \in \mathcal{H}\}, \forall x \in \mathbb{R}$$

avec $\mathcal{H}$ l'ensemble des demi-droites $H_1 = \{z \in \mathbb{R} : z \leqslant \beta\}$ ou $H_2 = \{z \in \mathbb{R} : z \geqslant \beta\}$, $\beta \in \mathbb{R}$ comme cela a été traité dans la Sous-Section 1.3.1.

On obtient alors, dans le cas où on considère le demi-espace $H_1$,

$$\mathbb{P}_{U_1}(2P(X) - 1 \in H_1) \overset{(*)}{=} \mathbb{P}_P(X \in H_1)$$

où $(*)$ se justifie par le fait que, comme $2P(X) - 1$ est distribué selon $U_1$ (par transport), on a $F(2P(x)-1) = P(x), \forall x \in \mathbb{R}$ et donc $\mathbb{P}_F(2P(X)-1 \leqslant \beta) = \mathbb{P}_P(X \leqslant \beta)$, avec $\beta \in \mathbb{R}$.

Le cas du demi-espace $H_2$ se traite exactement de la même façon.

Ainsi, $\inf\{\mathbb{P}_{U_1}(2P(X) - 1 \in H) : H \in \mathcal{H}\} = \inf\{\mathbb{P}_P(X \in H) : H \in \mathcal{H}\}$.

Nous pouvons conclure que $D(R(x, P), U_1) = D(x, P), \forall x \in \mathbb{R}$ et donc que $MK(x, P) = D(x, P), \forall x \in \mathbb{R}$.

□

Passons maintenant au cas des distributions issues de la famille $\mathcal{P}^p_{ell}$. Nous allons tout de suite considérer un vecteur aléatoire issu d'une distribution à symétrie sphérique car nous savons que tout vecteur aléatoire de distribution à symétrie elliptique peut être transformé en vecteur aléatoire de distribution à symétrie sphérique.

La distribution d'intérêt sera donc une distribution à symétrie sphérique et la distribution de référence sera la distribution uniforme sphérique $U_p$ pour laquelle nous connaissons la forme des contours de profondeur. Nous pouvons dès lors appliquer la Définition 2.10.

Pour démontrer la Propriété 3.4, nous avons besoin de plusieurs résultats que nous allons d'abord énumérer et pour certains démontrer.

**Proposition 3.1.** *Si un vecteur aléatoire $X$ est de distribution à symétrie sphérique, de paramètre de localisation $\mu = 0_p$ et de paramètre de dispersion $\Sigma = I_p$, alors $X \overset{\mathcal{L}}{=} RU$ avec $R \geqslant 0$, $R \overset{\mathcal{L}}{=} \|X\|_{0_p, I_p}$, $U$ un vecteur aléatoire distribué selon une loi uniforme sur $\mathcal{S}^{p-1}$ et $R$ et $U$ indépendants.*

Si le lecteur est intéressé par la démonstration de ce résultat, il est renvoyé à l'article [4].

**Proposition 3.2.** *Si $Y$ est un vecteur aléatoire de dimension $p$ issu d'une distribution à symétrie sphérique $P_{0,I,g}$ et si $G$ est la fonction de répartition associée à la fonction de densité radiale $g$ alors, le vecteur aléatoire $\frac{Y}{\|Y\|_{0_p,I_p}} G(\|Y\|_{0_p,I_p})$ est distribué selon $U_p$.*

*Démonstration.* Montrons que le vecteur aléatoire $\frac{Y}{\|Y\|_{0_p,I_p}}$ est distribué uniformément sur la sphère $\mathcal{S}^{p-1}$, que $G(\|Y\|_{0_p,I_p})$ est distribué selon une loi uniforme sur $[0,1]$ et que $\frac{Y}{\|Y\|_{0_p,I_p}}$ et $G(\|Y\|_{0_p,I_p})$ sont indépendants.
Le résultat sera alors démontré par la Définition 2.6 d'une distribution uniforme sphérique.

Tout d'abord, pour la variable aléatoire $G(\|Y\|_{0_p,I_p})$, on obtient

$$F(y) = \mathbb{P}(G(\|Y\|_{0_p,I_p}) \leqslant y) = \mathbb{P}(\|Y\|_{0_p,I_p} \leqslant G^{-1}(y)) = G(G^{-1}(y)) = y,\ \forall y \in \mathbb{R}.$$

où $F$ est la fonction de répartition de $G(\|Y\|_{0_p,I_p})$.

Ensuite, pour le vecteur aléatoire $\frac{Y}{\|Y\|_{0_p,I_p}}$, le vecteur aléatoire $Y$ étant issu d'une distribution à symétrie elliptique, la Proposition 3.1 nous donne

$Y \overset{\mathcal{L}}{=} RU$, avec $R \overset{\mathcal{L}}{=} \|Y\|_{0_p,I_p}$, $U$ distribué uniformément sur $\mathcal{S}^{p-1}$ et $R$ et $U$ indépendants.
De plus, comme $Y = \|Y\|_{0_p,I_p} \frac{Y}{\|Y\|_{0_p,I_p}}$, nous pouvons écrire $\frac{Y}{\|Y\|_{0_p,I_p}} \overset{\mathcal{L}}{=} U$.

Le résultat est démontré puisque $\|Y\|_{0_p,I_p}$ et $\frac{Y}{\|Y\|_{0_p,I_p}}$ sont alors indépendants et par conséquent $G(\|Y\|_{0_p,I_p})$ et $\frac{Y}{\|Y\|_{0_p,I_p}}$ le sont également.

□

**Proposition 3.3.** *Soient $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe et $g : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe et croissante. La fonction définie par la composition des fonctions $g$ et $f$ est une fonction convexe sur $\mathbb{R}^p$.*

*Démonstration.* Soient $x, y \in \mathbb{R}^p$ et $\lambda \in [0,1]$. Comme $f$ est une fonction convexe sur $\mathbb{R}^p$, nous pouvons écrire

$$f((1-\lambda)x + \lambda y) \leqslant (1-\lambda)f(x) + \lambda f(y).$$

La fonction $g$ étant croissante, nous pouvons appliquer celle-ci de part et d'autre de l'inégalité. On obtient

$$g(f((1-\lambda)x + \lambda y)) \leqslant g((1-\lambda)f(x) + \lambda f(y))$$

Or $g$ est également convexe donc en appliquant cette propriété de convexité au membre de droite de l'inégalité, celle-ci devient

$$g(f((1-\lambda)x + \lambda y)) \leqslant (1-\lambda)g(f(x)) + \lambda g(f(y))$$

La conclusion en découle. □

Passons à une autre propriété qui nous servira juste après.

**Proposition 3.4.** *La fonction* $\psi^* : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\} : z \mapsto \int_{-\infty}^{\|z\|} G(r)dr$ *est une fonction convexe.*

*Démonstration.* Cette fonction est la composée de la fonction $\varphi : \mathbb{R} \to \mathbb{R} : t \mapsto \int_{-\infty}^{t} G(r)dr$ avec la fonction $\|.\| : \mathbb{R}^p \to \mathbb{R}$. De plus, ces deux dernières fonctions sont toutes les deux convexes. Il est clair que $\varphi$ est une fonction convexe puisque une fonction réelle et dérivable sur un ouvert $\mathcal{U}$ de $\mathbb{R}$ est convexe si et seulement si sa dérivée est croissante sur cet ouvert. Or, $D\varphi(t) = G(t)$, $\forall t \in \mathbb{R}$ et $G(.)$ est une fonction croissante car c'est une fonction de répartition.
Ensuite, la fonction $\|.\|$ est également convexe car par *l'inégalité de Minkowski* [24] on obtient, $\forall x, y \in \mathbb{R}^p$ et $\forall \lambda \in [0, 1]$,

$$\|(1-\lambda)x + \lambda y\| \leqslant \|(1-\lambda)x\| + \|\lambda y\| \leqslant (1-\lambda)\|x\| + \lambda\|y\|$$

et comme $\varphi(.)$ est une fonction croissante, on conclut que $\psi^*(.)$ est une fonction convexe par la Proposition 3.3.

□

**Propriété 3.4.** *Dans le cas de la famille* $\mathcal{P}^p_{ell}$*, la fonction de profondeur* $MK(., P)$ *de Monge-Kantorovich coïncide avec la fonction de profondeur de Tukey* $D(., P)$*.*

*Démonstration.* Pour démontrer cela, nous allons montrer que les régions de profondeur de Monge-Kantorovich sont également les régions de profondeur de Tukey pour le cas particulier des distributions à symétrie sphérique. La distribution d'intérêt est donc une distribution à symétrie sphérique et la distribution de référence est la distribution uniforme sphérique $U_p$.

L'idée émise par Chernozhukov et al. [7] est de poser

$$\varphi(t) = \int_{-\infty}^{t} G(r)dr, \ \forall t \in \mathbb{R} \text{ et } \psi^*(z) = \varphi(\|z\|_{0_p, I_p}), \ \forall z \in \mathcal{Y}.$$

où $\mathcal{Y}$ est le domaine de définition de la fonction de $\psi^*$.

Le vecteur des rangs de Monge-Kantorovich est donné par le gradient de cette fonction convexe $\psi^*$

$$R(Y, P) := \frac{Y}{\|Y\|_{0_p, I_p}} G(\|Y\|_{0_p, I_p})$$

Au vu de la Proposition 3.2, ce vecteur des rangs est de distribution uniforme sphérique (la distribution de référence).

Par la Définition 2.10, la région de profondeur d'ordre $\alpha$ est définie par $Q(B(0,\alpha),P)$ et, étant donné que $R(.,P)$ et $Q(.,P)$ sont des applications réciproques l'une de l'autre, l'égalité suivante est satisfaite : $R(Q(B(0,\alpha),P),P) = B(0,\alpha)$.

Or, $R(Y,P) = \{x \in \mathbb{R}^p : \exists y \text{ tel que } R(y,P) = x\}$.

Ainsi, on obtient $A := \{x \in \mathbb{R}^p : \exists y \in Q(B(0,\alpha),P) \text{ tel que } R(y,P) = x\} = \{x \in \mathbb{R}^p : \|x\| \leqslant \alpha\} = B(0,\alpha)$.

Soit $x \in A$. Il existe $y \in Q(B(0,\alpha),P)$ tel que $R(y,P) = x$. Or $x \in B(0,\alpha)$ donc $\|x\| \leqslant \alpha$ et donc $\|R(y,P)\| \leqslant \alpha$. Au vu de la définition de $R(y,P)$, on obtient que $\|y\| \leqslant G^{-1}(\alpha)$ pour $y \in Q(B(0,\alpha),P)$.

Ceci achève la démonstration car les régions de profondeur de Monge-Kantorovich sont alors données par

$$\{y \in \mathbb{R}^p : \|y\| \leqslant G^{-1}(\alpha)\},\ \forall \alpha > 0$$

et correspondent donc aux régions de profondeur de Tukey.

□

Ainsi, la profondeur de Monge-Kantorovich apparaît comme une extension de la profondeur de Tukey.

## 3.4 Application

Pour cette partie, nous considérons les variables *Teaching* et *Research* de la base de données [1] et nous ne considérons que les soixante premières observations pour ces variables. Ce choix de nous limiter aux soixante premières observations a été fait dans le but de visualiser les choses.

La normalité bivariée de la distribution bivariée est vérifiée. En effet, en utilisant la procédure *mardiaTest* de la librairie *MVN* sur les 60 premières observations des variables *Teaching* et *Research* dans le logiciel *R*, nous obtenons le non-rejet de la multinormalité.

Commençons par le cas d'une distribution en dimension un. Considérons la variable *Teaching*. Notons $P_n$ la fonction de répartition empirique de cette variable. Nous travaillons dans un contexte non paramétrique.

Nous nous servons de la définition du vecteur des rangs de Monge-Kantorovich donnée à la Section 3.3, de la Définition 2.10 ainsi que du résultat obtenu à la Sous-Section 1.3.1, afin de les appliquer dans un contexte empirique sur la variable *Teaching*. Dans le

contexte empirique, le vecteur des rangs de Monge-Kantorovich est donné par $R(x, P_n) = 2P_n(x) - 1$, pour n'importe quelle observation $x$ de la variable *Teaching*.

Ainsi, dans l'outil statistique $R$, nous déterminons la fonction de répartition empirique de $R(., P_n)$ qui est la fonction de répartition d'une variable aléatoire de taille $n$ suivant une distribution uniforme sur $[-1, 1]$ et en appliquant le résultat de la Sous-Section 1.3.1 aux données, nous déterminons à la fois la fonction de profondeur de Tukey des données et la fonction de profondeur de Tukey de $R(., P_n)$. Nous constatons que celles-ci coïncident comme illustré à la FIGURE 3.1 et donc que la profondeur de Monge-Kantorovich est la profondeur de demi-espace de ces mêmes données.

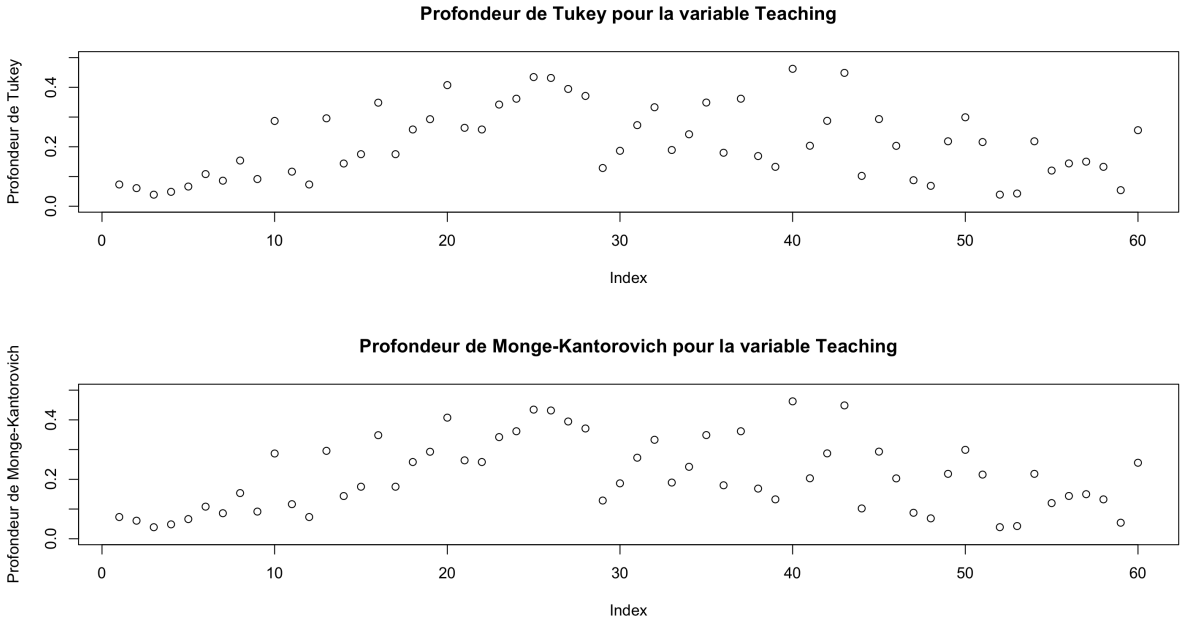


FIGURE 3.1 – Comparaison entre la profondeur de Tukey pour les données relatives à la variable *Teaching* avec la profondeur de Monge-Kantorovich de ces mêmes données

Passons maintenant au cas particulier d'une distribution à symétrie elliptique.

La FIGURE 3.2 représente les données pour les deux variables considérées et la FIGURE 3.3 représente les contours de profondeur de Tukey associés. Ces derniers ont été obtenus par la fonction *isodepth* dans le logiciel *R*. Comme nous n'avons pas rejeté l'hypothèse de multinormalité, nous avons une distribution bivariée qui est à symétrie elliptique, ce qui est bien confirmé par la forme des contours de profondeur empiriques.

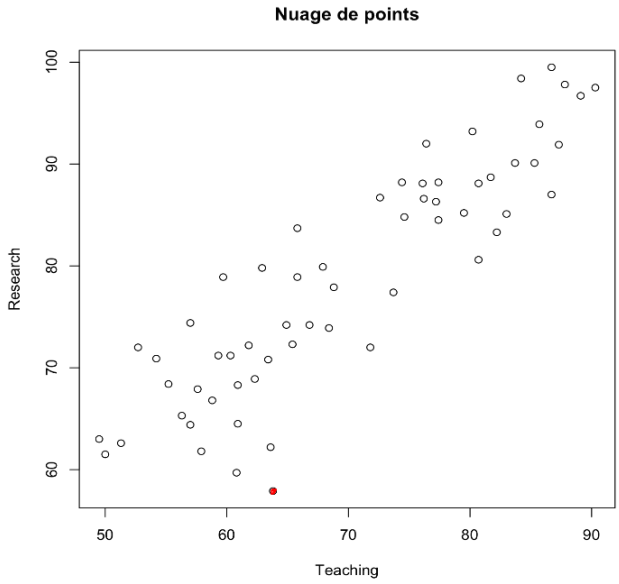


FIGURE 3.2 – Représentation des variables *Teaching* et *Research*

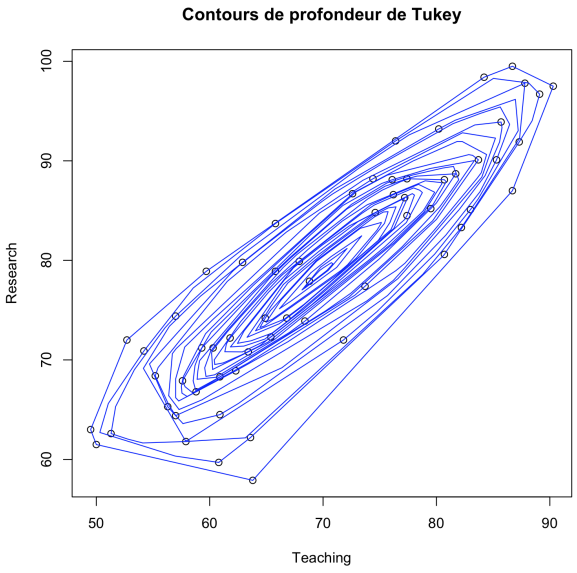


FIGURE 3.3 – Contours de profondeur de Tukey pour les variables *Teaching* et *Research*

Afin d'obtenir la profondeur de Monge-Kantorovich pour ce cas particulier dans le logiciel $R$, nous implémentons le vecteur des rangs comme défini à la Propriété 3.4 avec $G_n(.)$ la fonction de répartition radiale empirique associée à la distribution empirique bivariée des variables *Teaching* et *Research* et $Y$ le vecteur des observations bivariées défini par $Y := S^{-1/2}(X - \overline{x})$, où $S$ est la matrice de variance-covariance pour les indicateurs *Teaching* et *Research* et $\overline{x}$ est le vecteur des moyennes de ces indicateurs. Ensuite, via la fonction *depth* de la librairie *depth*, nous déterminons la profondeur de Tukey de ce vecteur

des rangs. Nous comparons alors les résultats obtenus grâce à cette dernière fonction appliquée sur le vecteur des rangs avec les résultats obtenus en appliquant cette même fonction sur le vecteur des données $Y$. Nous obtenons les représentations données à la FIGURE 3.4.

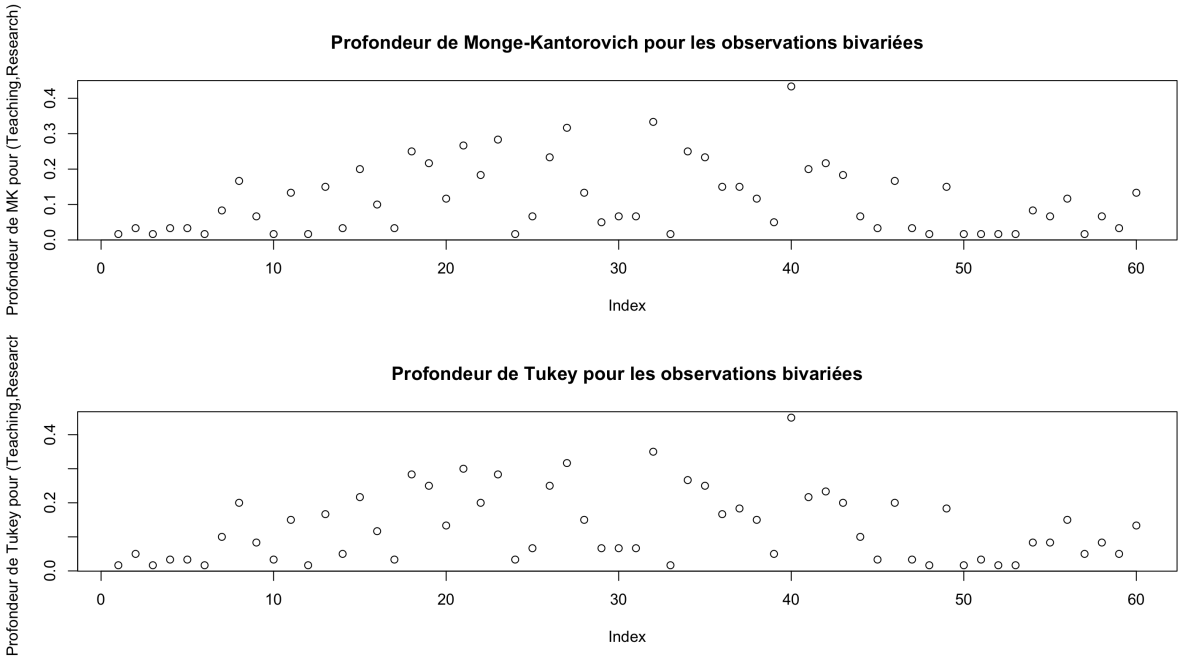


FIGURE 3.4 – Comparaison entre la profondeur de Tukey pour les données relatives aux indicateurs *Teaching* et *Research* avec la profondeur de Monge-Kantorovich de ces mêmes données

Les légères différences que l'on peut voir entre les deux représentations sont dues à des arrondis effectués par le logiciel *R*.

## 3.5 Implémentation des contours de Monge-Kantorovich

Il est possible d'implémenter les contours de profondeur de Monge-Kantorovich dans le logiciel *R*. Prenons l'exemple d'une distribution en forme de banane pour laquelle nous aimerions déterminer les contours de profondeur de Monge-Kantorovich, puisque nous avons constaté que les contours de profondeur de Tukey n'étaient pas adéquats pour cette distribution. En effet, ces derniers ne caractérisent pas cette distribution.

Pour obtenir les contours de profondeur de la distribution en forme de banane à partir d'une distribution uniforme sphérique sur la boule unité, la Définition 2.10 nous dit qu'il faut transporter les contours sphériques de la distribution de référence, grâce à la fonction des quantiles de Monge-Kantorovich. De plus, la fonction des quantiles minimise le coût quadratique moyen $\min_T E(T(U) - U)$ où $U \sim U_p$ ($p = 2$ ici), pour autant que la distribution d'intérêt possède des moments finis d'ordre 2.

Ainsi, la marche à suivre va être d'envoyer les points des différentes sphères définies pour la distribution de référence sur les points de la banane, mais pas de n'importe quelle manière. En effet, il faut envoyer ces points de la distribution de référence vers celle d'intérêt de manière à minimiser le coût quadratique moyen. La dernière étape consistera alors à relier les observations de la banane correspondant à une même sphère dans la distribution de départ, ce qui nous donnera les contours recherchés puisque la fonction des quantiles a pour caractéristique de transporter les contours d'une distribution de référence vers les nouveaux contours.

Générons une distribution de référence (une distribution uniforme sphérique sur la boule unité). Pour cela, générons tout d'abord une variable uniforme via la fonction *runif*. Notons cette variable $r$. Elle représente un rayon et nous allons en générer un certain nombre $n_r$ (que nous avons pris égal à 20 pour l'exemple). Nous générons $r$ comme suit :

$$r \sim \frac{U\{0, ..., n_r\}}{n_r}$$

Toujours via la fonction *runif*, nous générons ensuite une variable $\theta$ qui représentera un angle. Nous en prenons un certain nombre $n_\theta$ (que nous avons également pris égal à 20 pour l'exemple). La variable $\theta$ est définie comme suit :

$$\theta \sim 2\pi \frac{U\{0, ..., n_\theta\}}{n_\theta}$$

Le vecteur aléatoire fournissant les observations de la distribution d'intérêt est alors défini par

$$(X, Y) = (r \cos\theta, r \sin\theta)$$

Une telle distribution est illustrée à la FIGURE 3.5.

Comme dit précédemment, nous devons envoyer les observations de la distribution uniforme sphérique sur la boule unité sur les observations de la banane de manière à ce que le coût quadratique moyen soit minimal. Nous allons donc construire la matrice des distances euclidiennes au carré entre toutes les observations de la distribution d'intérêt et celles de la distribution de référence.

Une fois cette matrice obtenue, nous lui appliquons la commande *assignment* de la librairie *adagio*. Cette fonction nous donne la permutation des observations de la distribution en forme de banane qui minimise le coût quadratique moyen. Autrement dit, cette commande nous dit sur quelle observation de la banane doit être envoyée chaque observation de la distribution de référence afin de minimiser le coût.

Il ne nous reste plus qu'à appliquer la fonction *ahull* de la librairie *alphahull* à la permutation de nos observations, avec *alpha*= 0.5 représentant l'inverse du rayon d'un disque généralisé [1].

1. http ://cgm.cs.mcgill.ca/∼godfried/teaching/projects97/belair/alpha.html

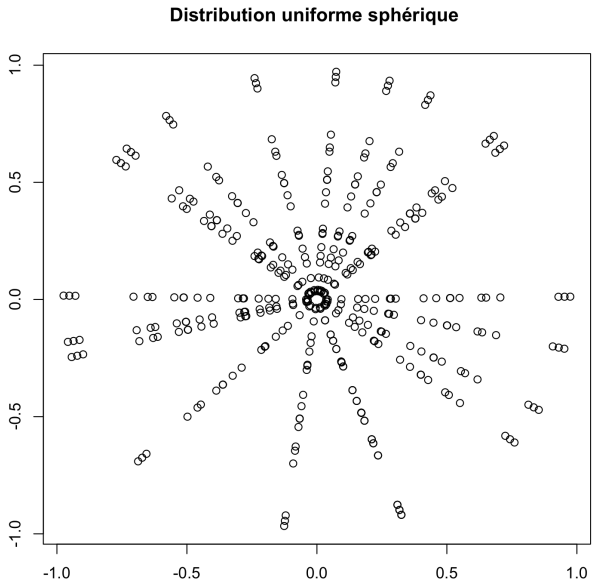


FIGURE 3.5 – Génération d'une distribution uniforme sphérique sur la boule unité

Quand on applique cette fonction à toutes les observations permutées, on obtient le contour de profondeur le plus extérieur. Pour représenter d'autres contours, il faut appliquer cette fonction *ahull* à une partie plus petite des observations permutées.

Une représentation des contours de Monge-Kantorovich pour la distribution en forme de banane est donnée à la FIGURE 3.6.

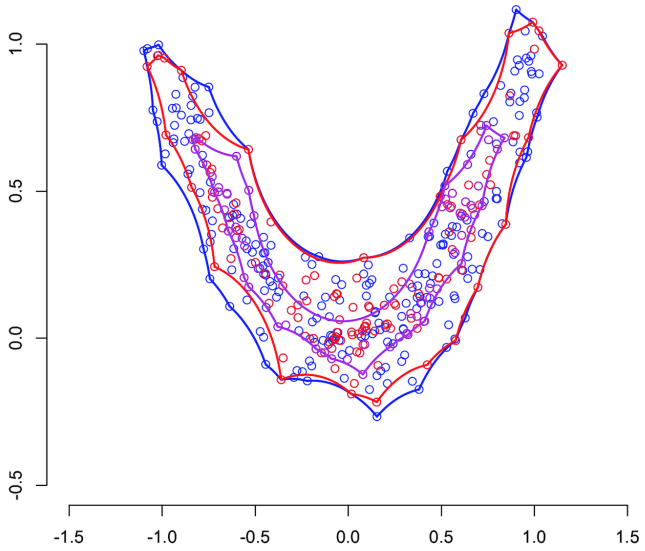


FIGURE 3.6 – Représentation des contours de profondeur de Monge-Kantorovich

Nous pouvons constater que ces contours de profondeur collent mieux à la distribution que les contours de profondeur de Tukey. De plus, plus le nombre d'observations sera élevé et plus les contours obtenus via le logiciel $R$ seront lisses.

Nous pouvons aussi déterminer la profondeur de Monge-Kantorovich pour n'importe quelle observation de la banane. En effet, la définition de la fonction de profondeur de Monge-Kantorovich 2.10 nous dit que la profondeur d'une observation $(x_b, y_b)$ de la banane est donnée par la profondeur de Tukey appliquée à l'observation $(x_u, y_u)$ de la distribution uniforme sphérique sur laquelle est envoyée $(x_b, y_b)$ par la fonction des rangs.

Autrement dit, pour pouvoir déterminer la profondeur de Monge-Kantorovich d'une observation quelconque $(x_b, y_b)$, il suffit d'utiliser la fonction *assignment* dans le logiciel $R$ qui nous donnera la permutation des observations de sorte que l'écart quadratique moyen soit minimisé et puis, une fois que nous aurons déterminé l'observation de la distribution uniforme sphérique correspondant à l'observation $(x_b, y_b)$, nous pourrons lui appliquer la fonction *depth* de la librairie *depth*.

Aux FIGURES 3.7 et 3.8 sont respectivement représentés les contours de Monge-Kantorovich pour une distribution de Cauchy bivariée de paramètre de localisation $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ et de paramètre de dispersion $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et les contours de Monge-Kantorovich pour les indicateurs *Teaching* et *Research* pour lesquels nous avons considéré les 60 premières valeurs uniquement. Les légères différences entre la FIGURE 3.8 et la FIGURE 1.14 sont dues à des approximations faites dans le logiciel afin de pouvoir appliquer la fonction *assignment* à la matrice des distances. Seuls certains contours sont représentés afin d'avoir une meilleure visibilité.

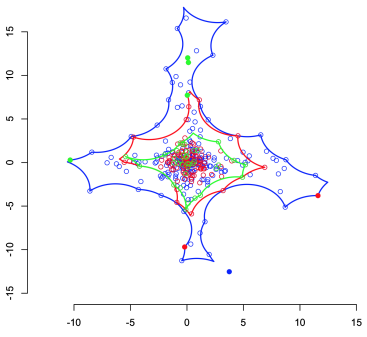


FIGURE 3.7 – Représentation des contours de MK pour une distribution de Cauchy bivariée

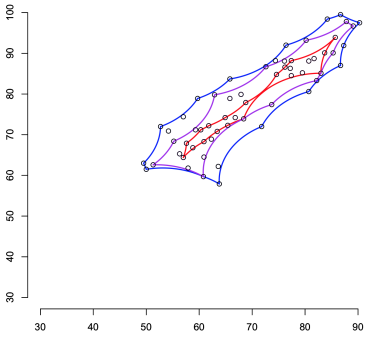


FIGURE 3.8 – Représentation des contours de MK sur les données

# Chapitre 4

# Application réelle : ranking des universités

## 4.1 Introduction

Comme annoncé dans l'introduction de ce mémoire, ce chapitre portera sur une analyse exploratoire de la base de données reprenant le classement des 200 meilleures universités du monde publié par le *"Times Higher Education"*. Ce classement est basé sur les critères énoncés et décrits dans l'introduction (la recherche, l'enseignement, le nombre de citations dans des articles scientifiques, le financement par les industries et les perspectives internationales) et consiste en l'attribution de rangs sur base du résultat obtenu par les universités pour la variable *OverallScore*.

Ce rang est donc un rang obtenu en univarié sur base du classement de la plus grande à la plus petite valeur prise par la variable des scores. Or, la notion de profondeur donne une notion de rang différente. En effet, dans le cas des profondeurs, nous attribuons un rang aux observations à partir du centre (point de profondeur maximale), que ce soit en univarié ou en multivarié. Nous allons donc comparer les rangs donnés dans la base de données et les rangs que nous calculerons en multivarié. De plus, les valeurs prises par la variable *OverallScore* (et donc les rangs basés sur cette variable) sont déterminées en attribuant une importance différente aux indicateurs (comme décrit dans l'introduction), ce qui ne sera plus le cas avec les profondeurs (tous les indicateurs auront la même importance).

Nous nous intéresserons ensuite à la pondération du *Times Higher Education* et nous déterminerons si il n'existe pas une pondération plus adéquate.

La table des données étudiées est reprise en annexe.

## 4.2 Analyse en dimension $p = 1, 2$

Nous pourrions commencer cette analyse par un bref aperçu de ce qui se passe en dimension 1 et 2 pour les données, afin de pouvoir visualiser les choses. Commençons donc avec un résumé à 5 valeurs effectué pour chaque indicateur. Ces résumés sont donnés ci-après (de la TABLE 4.1 à la TABLE 4.5). Une visualisation de ce résumé est alors illustrée à la FIGURE 4.1 par la construction des boîtes à moustaches.

| Min | $Q_1$ | Médiane | $Q_3$ | Max |
|---|---|---|---|---|
| 26.3 | 40.48 | 47.15 | 59.85 | 90.3 |

TABLE 4.1 – Résumé à 5 valeurs pour *Teaching*

| Min | $Q_1$ | Médiane | $Q_3$ | Max |
|---|---|---|---|---|
| 24.4 | 41.82 | 51.6 | 68.32 | 99.5 |

TABLE 4.2 – Résumé à 5 valeurs pour *Research*

| Min | $Q_1$ | Médiane | $Q_3$ | Max |
|---|---|---|---|---|
| 15.8 | 80.4 | 87.3 | 94.22 | 100 |

TABLE 4.3 – Résumé à 5 valeurs pour *Citations*

| Min | $Q_1$ | Médiane | $Q_3$ | Max |
|---|---|---|---|---|
| 31.8 | 39.18 | 51.5 | 76.35 | 100 |

TABLE 4.4 – Résumé à 5 valeurs pour *Industry Income*

| Min | $Q_1$ | Médiane | $Q_3$ | Max |
|---|---|---|---|---|
| 25.2 | 55.65 | 69.4 | 85.85 | 99.8 |

TABLE 4.5 – Résumé à 5 valeurs pour *International Outlook*

Nous constatons dans ces résumés à 5 valeurs que pour les indicateurs *Teaching*, *Research* et *International Outlook*, les valeurs maximales ne sont pas 100. Cela signifie simplement que les universités ayant acquis la meilleure note pour chacun de ces indicateurs ne se trouvent pas dans les 200 meilleures universités. Nous pouvons aussi dire que les indicateurs *Teaching*, *Research, Industry Income* et *International Outlook* ne possèdent pas vraiment de zone de forte concentration contrairement à la variable *Citations* pour laquelle on constate qu'environ 75% des universités obtiennent une note supérieure à 80%.

Passons à la représentation en boîtes à moustaches donnée à la FIGURE 4.1.

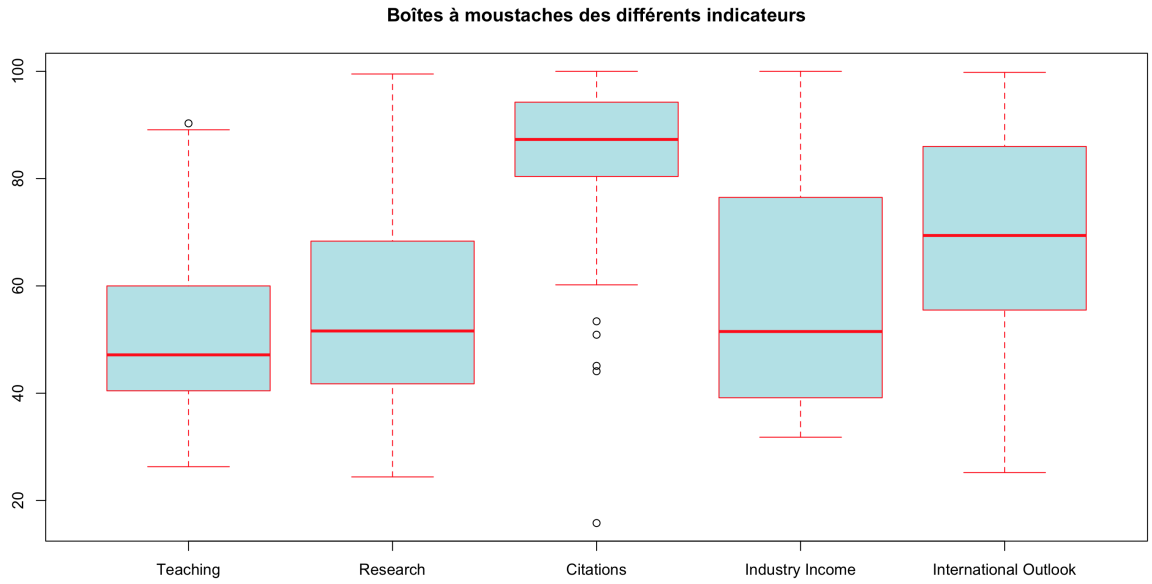


FIGURE 4.1 – Boîtes à moustaches pour les cinq indicateurs

Nous constatons que dans le cas de la variable *Teaching*, une valeur se trouve à l'extérieur de la boîte à moustaches. Il s'agit de l'observation relative à l'université *California Institute of Technology*. Cette université est troisième du classement général avec un score global de 93.0%. On constate également que 50% des universités obtiennent une valeur comprise entre 40% et 60% pour cette variable. Si on analyse les valeurs prises par l'indicateur *Citations*, on constate qu'il y a quatre valeurs extérieures à la boîte à moustaches dont une qui est vraiment très éloignée des autres. Cette observation en question est relative à l'université *Lomonosov Moscow State University* de Russie. Nous serions en droit de nous demander si cette observation n'est pas aberrante, si elle n'est pas le résultat d'une erreur d'encodage par exemple. Si on analyse la dispersion au centre de la variable *Citations*, on se rend compte que cette dispersion est faible. En effet, il y a une forte concentration entre 80% et 95%, ce qui illustre ce qui a été dit pour le résumé à 5 valeurs de cet indicateur.

Passons en dimension 2 avec l'analyse des bagplots des indicateurs pris deux à deux représentés aux FIGURES 4.2, 4.3 et 4.4.

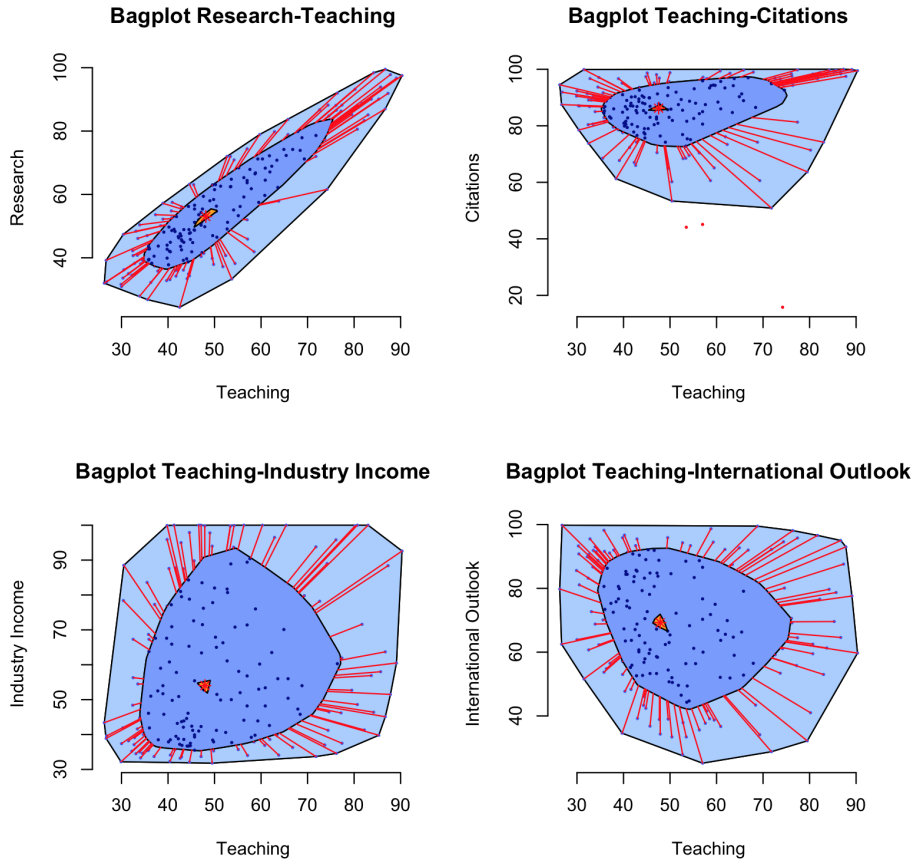


FIGURE 4.2 – Bagplot pour les cinq indicateurs, 1ère partie

Nous pouvons observer que 3 observations bivariées se trouvent à l'extérieur de la *boucle* du *bagplot* des variables *Teaching* et *Citations*. Si ces observations sont extérieures à la *fence* non représentée ici, alors ces observations seront atypiques par rapport aux autres observations. Il semble clair que l'observation la plus basse selon l'axe "*Citations*" est atypique car extérieure à la *fence*. Cette observation correspond à l'université *Lomonosov Moscow State University*. Nous constatons des choses identiques pour les indicateurs *Research* et *Citations*. Une observation semble très éloignée du *bag* et cette observation est à nouveau celle relative à l'université *Lomonosov Moscow State University*. Nous pouvons donc nous demander comment cette université s'est retrouvée parmi les 200 meilleures universités du monde. Cela peut s'expliquer par le fait que ce faible résultat obtenu par cette université pour l'indicateur *Citations* est contre-balancé par le résultat élevé qu'elle obtient pour l'indicateur *Teaching*. En effet, pour ce dernier indicateur, elle obtient un résultat de 74.2, ce qui est le 16ème meilleur résultat obtenu pour cet indicateur. De plus, les indicateurs *Teaching* et *Citations* obtiennent la même pondération (de 30%). L'indicateur *Teaching* compense donc l'indicateur *Citations*, ce qui permet à cette université de se maintenir dans les 200 premières universités du monde, en étant toutefois dans le bas de classement.

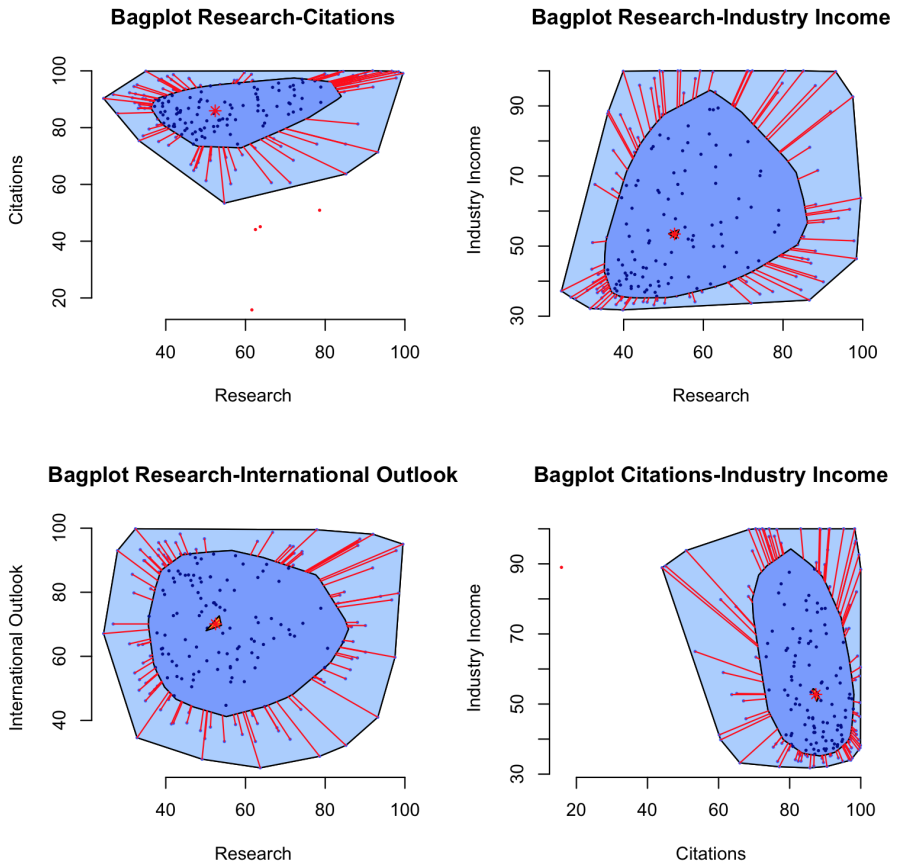


FIGURE 4.3 – Bagplot pour les cinq indicateurs, 2ème partie

Si l'on considère le *bagplot* pour les indicateurs *Teaching* et *Research*, l'étoile rouge i.e l'observation de profondeur maximale est l'observation relative à l'université de Warwick au Royaume-Uni, classée 91ème dans le classement du *Times Higher Education*. Autrement dit, si on considère uniquement ces deux indicateurs et s'ils possèdent tous les deux la même importance, alors cette observation sera classée première du classement bivarié réalisé à l'aide de la fonction de profondeur de Tukey. Nous pourrions effectuer une analyse plus poussée de chaque *boxplot* et *bagplot* mais l'objet principal de notre analyse est de travailler en dimension 5 de sorte de prendre en compte les 5 indicateurs simultanément. Nous en resterons là pour l'analyse univariée et bivariée. Ajoutons toutefois qu'au vu des bagplots, l'introduction de la nouvelle profondeur n'était pas nécessaire dans ce contexte du classement des universités. En effet, les bagplots ne font pas apparaître de formes naturelles non convexes. Nous pouvions donc nous contenter de la profondeur de Tukey dans ce contexte.

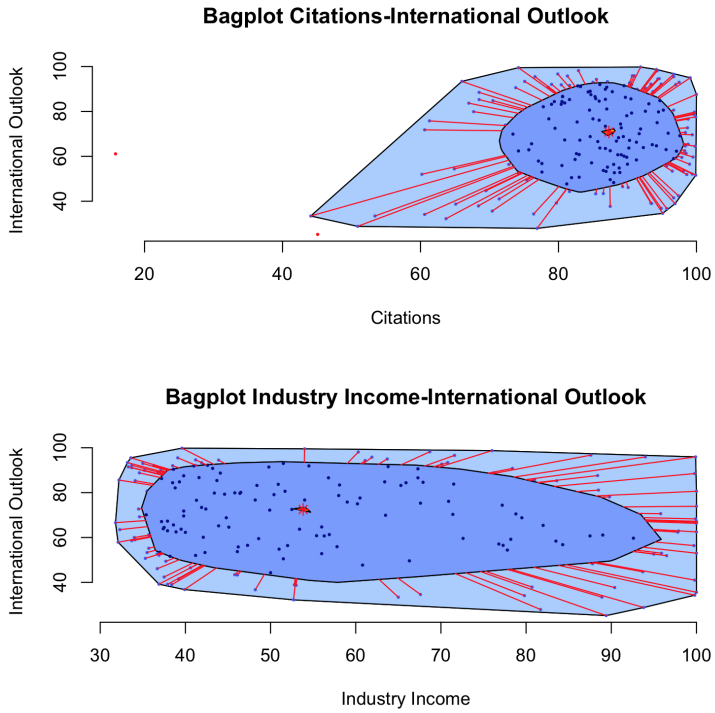


FIGURE 4.4 – Bagplot pour les cinq indicateurs, 3ème partie

## 4.3 Universités belges

Mais où se trouvent les universités belges ?

Quatre universités belges se trouvent parmi les 200 meilleures universités. Tout d'abord la *KUL* avec un score global de 71.8% occupe la 47ème place du classement. Ensuite, il y a l'*Université de Gand* (*Gent-University*) qui arrive à la 107ème place avec un score global de 59.8%. L'*Université Catholique de Louvain* (*UCL*) ne se place pas très loin derrière, à la 129ème place avec un score de 58% et enfin, l'*Université Libre de Bruxelles* (*ULB*) atteint le score de 54% et se classe 175ème. L'*Université de Liège* est quant à elle classée entre la 301 et 350ème place du classement avec un score compris entre 42.4% et 45.1%.

Les TABLES 4.6 et 4.7 donnent les résultats obtenus pour les 5 indicateurs par les universités belges se retrouvant parmi les 200 meilleures universités du monde.

| Rang | Université | Score global | Enseignement | Recherche | Citations |
|---|---|---|---|---|---|
| 47 | KUL | 71.8 | 54.2 | 70.9 | 88.7 |
| 107 | Gent-University | 59.8 | 44.4 | 56.2 | 77.6 |
| 129 | UCL | 58 | 38.9 | 52.2 | 78.6 |
| 175 | ULB | 54 | 32.4 | 42.2 | 80.4 |

TABLE 4.6 – Résultats des universités belges pour les indicateurs *Enseignement, Recherche* et *Citations*

| Rang | Université | Financement par les entreprises | Perspectives internationales |
|---|---|---|---|
| 47 | KUL | 99.9 | 68.3 |
| 107 | Gent-University | 84.3 | 56.8 |
| 129 | UCL | 54 | 76.7 |
| 175 | ULB | 48.4 | 83.2 |

TABLE 4.7 – Résultats des universités belges pour les indicateurs *Financement des entreprises* et *Perspectives internationales*

Nous pouvons constater que la *KUL* est fortement financée par les entreprises. Elle reçoit le deuxième score le plus élevé du classement pour l'indicateur *Industry Income* et celui-ci prend 162 valeurs différentes dans ce classement. De même pour l'*Université de Gand* qui reçoit le 25ème plus haut résultat du classement pour cet indicateur. Nous pouvons également constater que les universités belges, de manière générale, obtiennent des résultats élevés pour l'indicateur *Citations*. Cela signifie que les universités belges sont très citées dans des articles scientifiques et ont donc une certaine réputation dans ce domaine. Par contre, si l'on considère l'indicateur *Teaching*, on se rend compte que les scores des universités belges sont relativement faibles, surtout pour l'*Université Libre de Bruxelles* qui obtient le 11ème moins bon score du classement.

## 4.4 Analyse en dimension 5

Pour cette partie, nous prenons en compte les 5 indicateurs simultanément. Déterminons les rangs multivariés grâce à la profondeur de Tukey, puis grâce à la profondeur de Monge-Kantorovich via la technique évoquée lors de la construction des contours de Monge-Kantorovich dans le logiciel *R*.

Nous utilisons la fonction *depth* que nous appliquons à la matrice dont les 5 colonnes représentent les valeurs prises par les 5 indicateurs, afin d'obtenir la profondeur de Tukey. Sachez toutefois que, dans le logiciel *R*, la profondeur de Tukey ne peut être qu'approchée pour une dimension supérieure à 3. Il en va de même pour la profondeur de Monge-Kantorovich puisque celle-ci est définie à partir de la profondeur de Tukey. Pour la profondeur de Monge-Kantorovich, nous générons d'abord une distribution uniforme sphérique en dimension 5 dans *R*, nous construisons la matrice des distances au carré entre les observations de la distribution uniforme sphérique et les observations de la base de

données et nous lui appliquons la commande *assignment* de la librairie *adagio*. Nous utilisons ensuite la commande *invPerm* de la librairie *Matrix* qui détermine la permutation inverse de la permutation obtenue par la commande *assignment*. Enfin, nous appliquons la fonction *depth* à l' échantillon issu de la distribution uniforme sphérique, permuté selon la permutation obtenue par *invPerm*.

Les résultats obtenus sont retranscrits dans la TABLE A.1 donnée en annexe. Les 3 dernières colonnes de ce tableau reprennent les rangs calculés grâce à différentes méthodes. La colonne *Rg HD* nous donne les rangs déterminés à partir de la profondeur de demi-espace (le rang le plus grand est attribué à l'observation de plus petite profondeur et le plus petit rang est attribué à l'observation de profondeur la plus grande). Ensuite, la colonne *Rg MK* donne les rangs de Monge-Kantorovich définis par $\|R(., P)\|$. Toutefois, ces rangs prennent leurs valeurs dans l'intervalle $[0, 1]$ ainsi, afin de pouvoir les comparer aux rangs de Tukey, nous avons appliqué la commande *rank* à ces rangs obtenus, de manière à avoir des rangs compris entre 1 et 200. Enfin, la dernière colonne nous donne les rangs attribués aux universités sur base du calcul de la profondeur de Monge-Kantorovich, mais cette fois-ci en utilisant la même définition des rangs que pour la profondeur de Tukey.

Regardons ce que deviennent les rangs des 5 meilleures universités du classement. La TABLE 4.8 reprend ces résultats.

| Rg | Université | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|
| [1] | UniversityofOxford | 161 | 103 | 116 |
| [2] | UniversityofCambridge | 161 | 181 | 167 |
| [3] | CaliforniaInstituteofTechnology | 161 | 120 | 116 |
| [3] | StanfordUniversity | 111 | 151 | 167 |
| [5] | MassachusettsInstituteTechnology | 161 | 174 | 167 |

TABLE 4.8 – Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich, pour les 5 meilleures universités

Rappelons que les différents rangs multivariés déterminés sur ces données sont attribués à partir d'un centre (l'observation de profondeur maximale). De plus, le traitement des universités qui ont la même profondeur et donc qui se voient attribuer le même rang (MK, HD ou MK-prof) est différent du traitement des ex-aequo effectué par le *Times Higher Education*. En effet, dans le logiciel *R*, lorsque l'on cherche à attribuer un rang à un ensemble d'observations, les valeurs vont être organisées de la plus petite à la plus grande. Lorsque des valeurs égales sont rencontrées, leur position dans la suite ordonnée va être prise en compte et, pour leur attribuer un rang, le logiciel va effectuer une moyenne des positions quelles occupent dans la suite.

Prenons un exemple afin d'expliquer ce que fait le logiciel. Supposons avoir $n$ observations univariées $x_1, ..., x_n$. Nous pouvons organiser ces observations de la plus petite à la plus grande. On obtient alors la suite $x_{(1)}, ...., x_{(n)}$ où $x_{(i)} \leqslant x_{(j)}, \forall i, j \in \{1, ..., n\}$

avec $i \leqslant j$. Supposons qu'il existe $k \in \{1, ..., n\}$ et $l \in \{0, ..., n\}$, avec $k + l \leqslant n$ tels que $x_{(k)} = x_{(k+1)} = ... = x_{(k+l)}$. Dans ce cas, le rang de ces dernières observations sera donné par $\frac{1}{l+1}\sum_{i=k}^{k+l} i$. L'observation $x_{(k+l+1)}$ aura quant à elle le rang $k + l + 1$ si celle-ci est différente de toutes les observations qui suivent, sinon on recommence le processus explicité ci-dessus. Rappelons que dans le classement univarié effectué par le *Times Higher Education*, les rangs attribués aux universités possédant le même score global sont définis comme suit : si $t > 0$ universités occupent le rang $i > 0$ alors, l'université qui vient après ces $t$ universités dans le classement se verra attribuer le rang $i + t$. Autrement dit, dans une situation telle que décrite ci-avant, le principe est d'attribuer le rang $k$ aux observations $x_{(k)}, x_{(k+1)}, ..., x_{(k+l)}$ puis le rang $k + l + 1$ à l'observation $x_{(k+l+1)}$.

Ensuite, si on s'intéresse aux rangs attribués aux 5 moins bonnes universités, on obtient la TABLE 4.9.

| Rg | Université | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|
| [196] | ParisSorbonneUniversityParis4 | 161 | 135 | 167 |
| [197] | RoyalHollowayUniversityLondon | 161 | 91 | 104 |
| [198] | UniversityofCaliforniaRiverside | 161 | 184 | 167 |
| [198] | UniversityofGothenburg | 161 | 164 | 116 |
| [198] | NationalTaiwanUniversity | 161 | 178 | 167 |

TABLE 4.9 – Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich pour les 5 moins bonnes universités du classement

Allons maintenant chercher les 5 universités du milieu du classement. La TABLE 4.10 reprend les différents types de rangs établis pour ces universités.

| Rg | Université | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|
| [98] | EmoryUniversity | 161 | 137 | 167 |
| [99] | UniversityCaliforniaIrvine | 8 | 2 | 2 |
| [100] | UniversityofBonn | 90 | 96 | 86 |
| [100] | UniversityofColoradoBoulder | 111 | 199 | 167 |
| [100] | UniversityofPittsburgh | 161 | 200 | 167 |

TABLE 4.10 – Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich pour les 5 universités du milieu de classement

Globalement, l'attribution des rangs reste cohérente d'une profondeur à l'autre. Des universités qui occupent un rang de Tukey élevé occupent également un rang de MK élevé en général. Plus loin, nous allons approfondir l'analyse en comparant l'ensemble des rangs.

Enfin, nous nous demandons quelles seraient les places occupées par les universités belges dans le cas de classements multivariés. Pour répondre à cette question, voici la

TABLE 4.11 reprenant les différents rangs pour les universités belges se trouvant dans les 200 meilleures universités du monde.

| Rg | Université | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|
| [47] | KULeuven | 161 | 64 | 70 |
| [107] | GhentUniversity | 53 | 67 | 64 |
| [129] | UniversiteCatholiqueLouvain | 44 | 1 | 1 |
| [175] | UniversiteLibredeBruxelles | 44 | 7 | 8 |

TABLE 4.11 – Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich pour les universités belges

Nous pouvons constater que selon les rangs de Monge-Kantorovich, aussi bien déterminés par le classement des profondeurs de Monge-Kantorovich par ordre décroissant que par la Définition 2.10, c'est une université belge qui occupe la 1ère place du classement. En effet, l'*Université Catholique de Louvain* est l'université de plus grande profondeur de Monge-Kantorovich lorsque l'on considère les 5 indicateurs simultanément. C'est l'université à partir de laquelle tout le classement s'effectue.

De manière générale, nous pouvons constater dans les TABLES 4.8, 4.9, 4.10, 4.11 que des universités qui ont obtenu le même rang en univarié ne sont plus nécessairement de même rang en multivarié. Réciproquement, des universités qui se voyaient octroyer des rangs différents en univarié, obtiennent des rangs égaux selon une méthode multivariée particulière.

Intéressons-nous maintenant au classement complet et aux différents rangs définis.

Comparons tout d'abord les rangs multivariés obtenus à l'aide de la profondeur de demi-espace à ceux obtenus grâce à la profondeur de Monge-Kantorovich (que ce soit ceux obtenus par la Définition 2.10 ou ceux définis de la même manière que les rangs de Tukey i.e le rang le plus grand à la plus petite profondeur et le rang le plus petit à la plus grande profondeur). Des illustrations de ces rangs sont données aux FIGURES 4.5, 4.6 et 4.7. Les classements obtenus sont en apparence fort différents l'un de l'autre mais analysons cela via la corrélation entre ces rangs de Tukey et de Monge-Kantorovich. Pour cela, nous utilisons la commande *cor* du logiciel $R$ avec pour argument supplémentaire *method="spearman"*. Cette dernière méthode est utilisée dans $R$ dans le cadre du calcul de la corrélation entre des rangs. Le coefficient de corrélation de SPEARMAN est plus robuste que le coefficient habituellement utilisé et est conseillé dans l'aide du logiciel pour des données qui ne sont pas nécessairement issues d'une distribution normale bivariée. Nous obtenons une corrélation d'environ 70% (que l'on compare les rangs de Tukey avec la 1ère ou la 2ème définition des rangs de Monge-Kantorovich), ce qui représente une forte corrélation entre ces rangs. On a donc une relation forte entre les rangs de Tukey et ceux de Monge-Kantorovich sur cette base de données.

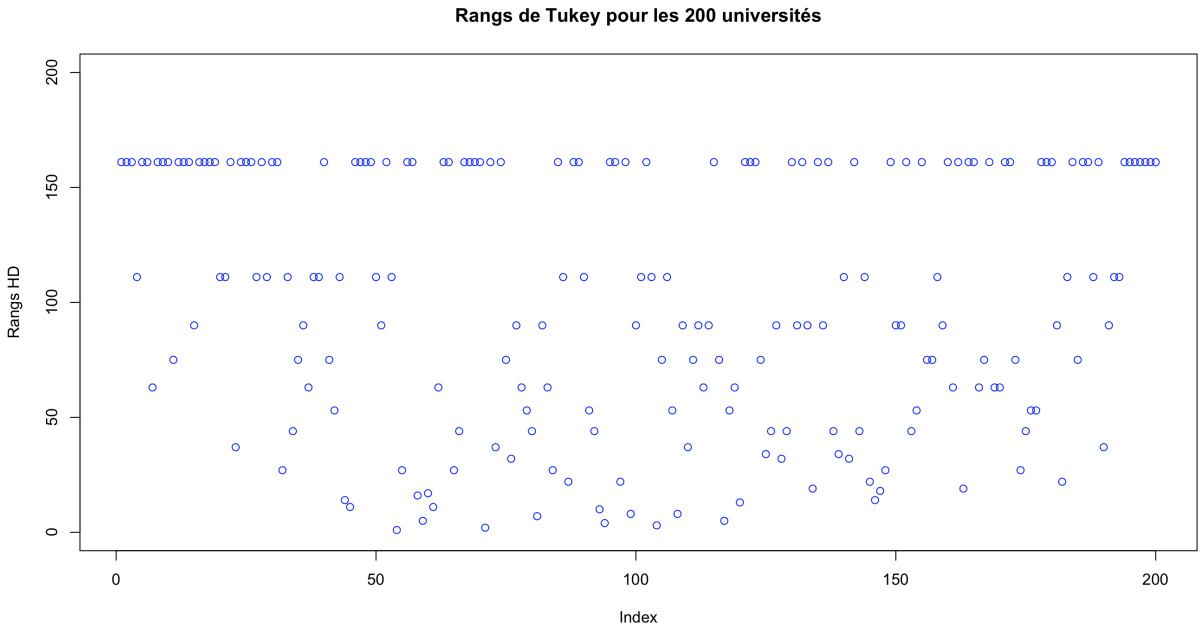


FIGURE 4.5 – Rangs de Tukey

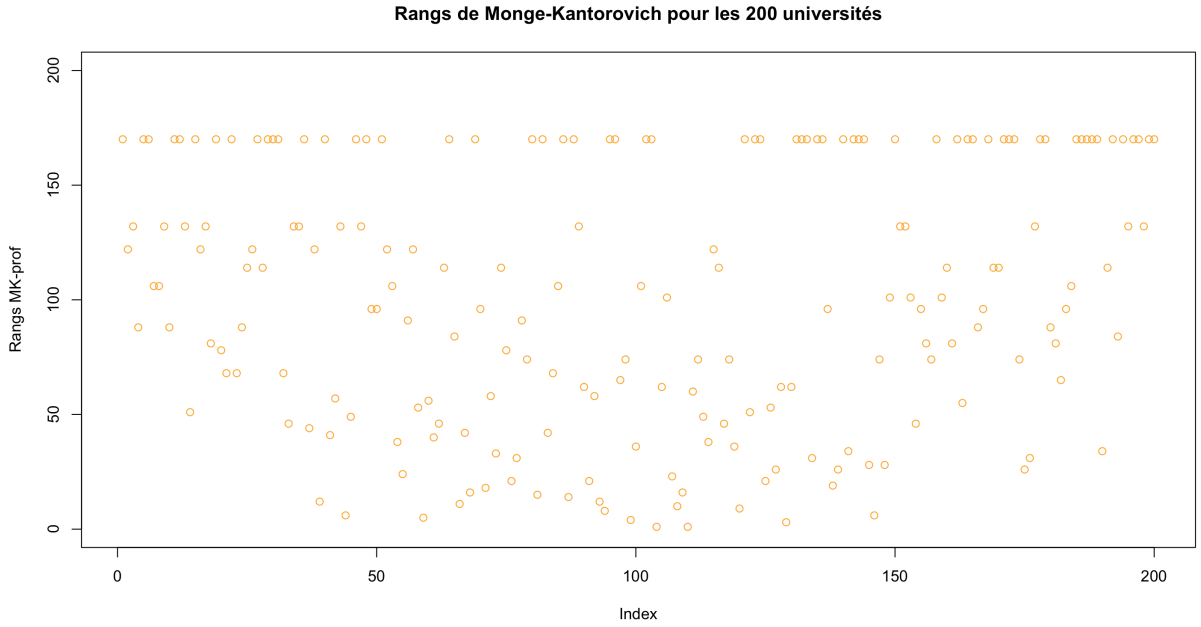


FIGURE 4.6 – Rangs de Monge-Kantorovich (Rg MK-prof)

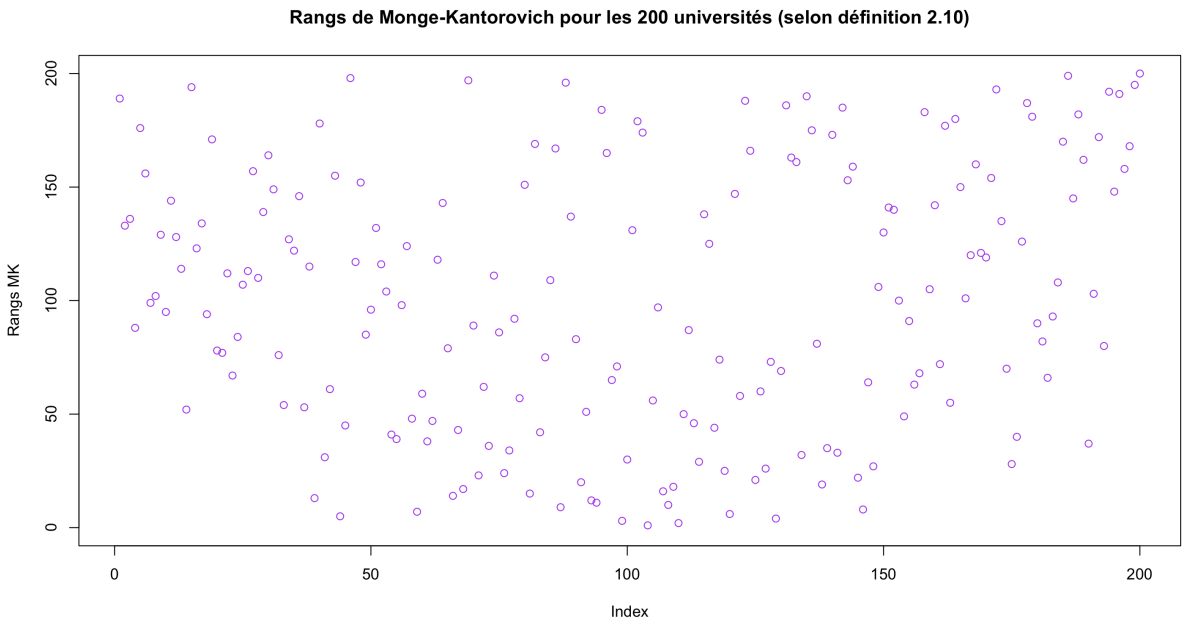


FIGURE 4.7 – Rangs de Monge-Kantorovich (Rg MK)

Nous pouvons remarquer que pour les rangs *Rg HD* et *Rg MK-prof* il y a énormément d'observations ayant le même rang. Regardons cela de plus près.

Dans le logiciel *R*, nous pouvons utiliser la commande *unique* afin de déterminer quels sont les différentes valeurs prises par les rangs. En appliquant cette commande aux rangs de Tukey, on constate qu'il n'y a que 27 rangs différents. Autrement dit, les 200 universités se partagent 27 rangs. C'est le rang 161 qui apparaît le plus souvent. Il apparaît 78 fois, ce qui signifie que plus d'1/3 des universités sont de rang 161. Autrement dit, plus d'1/3 des observations 5-variées se trouvent sur un même contour de profondeur. En procédant de la même façon avec les rangs MK-prof, on obtient qu'il y a 54 rangs différents pour nos observations. Le rang attribué le plus souvent est le rang 170 et celui-ci apparaît 61 fois. Par contre, si l'on considère les rangs MK et qu'on leur applique la commande *unique*, alors on trouve que chaque observation possède un rang différent.

Les classements multivariés obtenus sont bien différents du classement donné par le *Times Higher Education* puisque ces rangs sont établis à partir d'un centre contrairement au classement en dimension 1 qui est établi pour les résultats de la variable *OverallScore* classés de la plus grande à la plus petite valeur. Si nous voulons établir la corrélation entre les rangs multivariés et ces rangs univariés, il faut prendre les rangs univariés calculés en fonction de l'écart par rapport à la médiane de la variable *OverallScore*. Pour les rangs de Tukey, cette corrélation est de 31%, pour les rangs de Monge-Kantorovich *MK-prof*, cette corrélation est d'environ 25% et pour les rangs de MK, cette corrélation est de 20%, ce qui est très faible. Cette forte différence peut s'expliquer par le fait que la variable des scores globaux est construite sur base de l'attribution de différents poids aux différents indicateurs contrairement aux rangs multivariés pour lesquels le poids est le même pour

chaque indicateur.

L'avantage de traiter des rangs multivariés et non des rangs univariés est que l'on considère toutes les variables simultanément contrairement à ce qui est fait en univarié où chaque variable est traitée l'une après l'autre. Toutefois, l'interprétation multivariée est plus complexe car la comparaison et le classement se font par rapport à un certain centre (le point de profondeur maximale), ainsi qu'en terme d'éloignement par rapport à ce centre. On n'a donc pas un classement de l'université la plus performante à la moins performante.

## 4.5 Analyse en composantes principales

Intéressons nous maintenant à la pondération attribuée pour chaque indicateur par le *Times Higher Education*. Rappelons tout d'abord cette pondération. Les indicateurs enseignement, recherche et citations sont les indicateurs possédant le plus d'importance avec un poids de 30% chacun. L'indicateur du financement par les industries représente 2.5% du score global et celui des perspectives internationales compte pour 7.5% du score.

Nous allons effectuer une Analyse en Composantes Principales dans le logiciel *R*, souvent rencontrée dans la littérature sous le nom "*ACP*". Cette technique permet de réduire la dimension d'un problème. Elle permet de se ramener à un problème en dimension 1 ou 2 via ce qu'on appelle les composantes principales. Ces composantes principales ne sont rien d'autre que des combinaisons linéaires des variables aléatoires de départ. Cela permet de visualiser graphiquement certains résultats souhaités. Pour plus d'informations sur l'ACP, le lecteur est renvoyé à l'article "*L'analyse en composantes principales : principes et applications*" de R. Palm [20].

L'idée de faire une ACP vient du fait que dans la littérature, nous trouvons des critiques de la pondération attribuée par le *Times Higher Education*. Les auteurs préconisent de procéder autrement et ils préconisent notamment d'effectuer une ACP. Le lecteur pourra trouver dans l'article *Quality in Higher Education* de Kaycheng Soh [14] une critique de ce classement ainsi qu'une analyse basée sur d'autres méthodes que l'ACP.

Nous allons réduire le problème à une dimension comme cela a été fait avec la variable des scores globaux. Pour cela, nous effectuons une ACP sur l'ensemble des indicateurs grâce à la commande *princomp* du logiciel *R*. La sortie $*loadings* nous donne une expression de la première composante principale. Les coefficients de la combinaison des indicateurs nous fournissant la première composante principale représentent le poids de chaque indicateur. Cette ACP a été réalisée sur la matrice des corrélations car l'hypothèse d'homoscédasticité des variances est rejetée et la matrice de corrélation est moins sensible à la différence de variabilité. Voici ces pondérations :

- Enseignement : 37%
- Recherche : 37%

- Citations : 6%
- Financement par les industries : 14%
- Perspectives internationales : 6%

Nous constatons une très nette différence surtout pour les indicateurs *Citations* et *Financement par les industries* pour lesquels le poids est respectivement divisé par 5 et multiplié par un facteur 6.

En analysant les corrélations entre le classement multivarié obtenu à partir des rangs de Tukey et le classement effectué à partir de la première composante principale et en faisant de même avec les rangs de Monge-Kantorovich (obtenus des deux façons), nous pouvons conclure que la pondération obtenue par l'ACP est plus opportune que la pondération attribuée par le *Times Higher Education*. En effet, les corrélations obtenues avec la première composante principale sont toutes supérieures aux corrélations obtenues lorsque nous comparions le classement univarié aux classements de Tukey et de Monge-Kantorovich. Nous pouvons aussi déterminer la corrélation entre le classement du *Times Higher Education* et le classement obtenu grâce à l'ACP. Cette corrélation est de 93%, ce qui représente une forte corrélation. Cela nous permet de dire que le classement est, malgré cette forte corrélation, sensible à la pondération.

# Conclusion

Tout au long de ce mémoire, nous nous sommes intéressés à la notion de profondeur statistique ainsi qu'à tout ce qui découle de cette notion (quantiles et rangs). La première fonction de profondeur à laquelle nous avons fait référence est la profondeur de Tukey. Notre objectif principal dans ce mémoire était de montrer que cette dernière fonction de profondeur n'était pas entièrement satisfaisante dans le cas où certaines distributions étaient considérées. Ce fait nous permettait alors de justifier l'introduction d'une nouvelle fonction de profondeur définie par Chernozhukov et al. [7] et nommée fonction de profondeur de *Monge-Kantorovich*.

Nous ne pouvions pas parler de profondeur sans définir ce concept et sans essayer de donner l'intuition de ce que cela représente en dimension 1 ainsi que dans le cas d'une distribution à symétrie elliptique, lorsque l'on considère la profondeur de Tukey. De plus, dans ces cas particuliers, la profondeur de Tukey convient car celle-ci décrit parfaitement la distribution. C'est un avantage certain dans le domaine de la statistique non paramétrique où les paramètres de distribution ne sont pas toujours connus. Nous avons ensuite souhaité montrer que cette profondeur n'était pas parfaite en considérant le cas de distributions telles que la distribution de Cauchy bivariée ou encore dans le cas d'une distribution sous forme de banane, pour lesquelles elle ne caractérise plus parfaitement la distribution. Cette dernière constatation vient du fait que les contours de profondeur de Tukey sont convexes et donc ne sauraient pas "coller" correctement à une distribution dont les contours sont non convexes.

Ensuite, nous nous sommes basés sur ce que Chernozhukov et al. ont défini dans leur article *Monge-Kantorovich Depth, Quantiles, Ranks and Signs* [7]. Leur intention est de définir une nouvelle profondeur dont les contours sont capables de coller à n'importe quelle distribution. Pour cela, ils utilisent la notion de transport optimal, que l'on doit à Monge et à Kantorovich, d'une distribution dite de référence sur une distribution d'intérêt.

Comme cette notion de transport n'est pas familière, nous avons choisi de nous étendre sur ce concept en partant de la définition du transport donnée par Villani [28] et en déterminant à quoi cette application du transport fait référence en dimension 1. Le résultat obtenu est que si l'on considère la fonction de répartition habituelle, une application de transport d'une distribution quelconque sur la distribution uniforme sur $[0, 1]$ correspond à la fonction des rangs. Tandis que l'application de transport réciproque correspond

à la fonction quantile en dimension 1. Toutefois, comme la fonction de répartition habituelle suppose que les observations soient classées de la plus petite à la plus grande, nous introduisons une autre fonction de répartition basée sur un classement à partir de la médiane. C'est à partir de là que Chernozhukov et al. définissent les quantiles de Monge-Kantorovich comme étant la fonction qui transporte la distribution uniforme sphérique en dimension $p$ sur une distribution quelconque en dimension $p$ et les rangs comme l'application de transport réciproque.

La fonction de profondeur de Monge-Kantorovich dans le cas d'une distribution de référence uniforme sphérique est définie à partir de la fonction de profondeur de Tukey. De plus, cette nouvelle profondeur est une extension de la profondeur de Tukey dans le sens que celle-ci correspond à la profondeur de Tukey dans les cas particuliers de la dimension 1 et d'une distribution à symétrie elliptique et qu'elle s'adapte mieux à n'importe quel type de distribution et en particulier au cas de distributions non convexes. Pour que cela soit possible, il a fallu relâcher certaines propriétés attribuées aux contours de profondeur en général. En effet, les contours de profondeur de Monge-Kantorovich ne sont pas convexes et ne respectent pas le principe de l'équivariance affine.

La fonction de profondeur de Monge-Kantorovich nous apporte donc satisfaction car ses contours collent parfaitement à tout type de distributions. Cela signifie que dans un cadre pratique où l'on ne connait pas la forme d'une distribution particulière, nous pouvons nous servir de cette profondeur afin de caractériser la distribution en question.

Enfin, déterminer la profondeur de Monge-Kantorovich à la main peut vite devenir pénible lorsque l'ensemble des données sur lequel on travaille est conséquent. Toutefois une procédure existe dans le logiciel *R*. Cette procédure n'est pas encore aussi directe que le calcul de la profondeur de Tukey pour lequel il suffit d'utiliser la fonction *depth* de la libraire *depth*, mais l'algorithme explicité à la Section 3.5 du Chapitre 3 n'est toutefois pas très long à réaliser. L'interprétation des résultats obtenus en utilisant la profondeur de Monge-Kantorovich est la même que pour la fonction de profondeur de Tukey.

# Annexe A

# Annexe

Voici, à la TABLE A.1, la base de données utilisée pour réaliser l'analyse approfondie. Les colonnes 5 à 9 représentent les 5 indicateurs dans l'ordre suivant : "*Teaching*", "*Research*", "*Citations*", "*Industry Income*", "*International Outlook*".

| Rg univarié | Université | Pays | Score | I1 | I2 | I3 | I4 | I5 | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | UniversityofOxford | UnitedKingdom | 94.3 | 86.7 | 99.5 | 99.1 | 63.7 | 95.0 | 161 | 103 | 116 |
| 2 | UniversityofCambridge | UnitedKingdom | 93.2 | 87.8 | 97.8 | 97.5 | 51.5 | 93.0 | 161 | 181 | 167 |
| 3 | CaliforniaInstituteofTechnology | UnitedStates | 93.0 | 90.3 | 97.5 | 99.5 | 92.6 | 59.7 | 161 | 120 | 116 |
| 3 | StanfordUniversity | UnitedStates | 93.0 | 89.1 | 96.7 | 99.9 | 60.5 | 77.6 | 111 | 151 | 167 |
| 5 | MassachusettsInstituteTechnology | UnitedStates | 92.5 | 87.3 | 91.9 | 100.0 | 88.4 | 87.6 | 161 | 174 | 167 |
| 6 | HarvardUniversity | UnitedStates | 91.8 | 84.2 | 98.4 | 99.7 | 46.4 | 79.7 | 161 | 160 | 167 |
| 7 | PrincetonUniversity | UnitedStates | 91.1 | 85.7 | 93.9 | 99.6 | 58.0 | 78.7 | 63 | 56 | 64 |
| 8 | ImperialCollegeLondon | UnitedKingdom | 89.2 | 81.7 | 88.7 | 96.7 | 71.6 | 96.6 | 161 | 107 | 104 |
| 9 | UniversityofChicago | UnitedStates | 88.6 | 85.3 | 90.1 | 99.4 | 39.8 | 69.6 | 161 | 133 | 167 |
| 10 | ETHZurichSwissFederalInstituteTechnology | Switzerland | 87.7 | 76.4 | 92.0 | 94.3 | 60.3 | 98.1 | 161 | 179 | 167 |
| 10 | UniversityofPennsylvania | UnitedStates | 87.7 | 83.7 | 90.1 | 98.5 | 56.9 | 61.3 | 75 | 111 | 104 |
| 12 | YaleUniversity | UnitedStates | 87.6 | 86.7 | 87.0 | 98.4 | 45.1 | 64.6 | 161 | 118 | 116 |
| 13 | JohnsHopkinsUniversity | UnitedStates | 86.5 | 76.1 | 88.1 | 98.4 | 95.8 | 70.6 | 161 | 124 | 167 |
| 14 | ColumbiaUniversity | UnitedStates | 86.0 | 82.2 | 83.3 | 98.8 | 41.3 | 76.6 | 161 | 190 | 167 |
| 15 | UniversityCaliforniaLA | UnitedStates | 85.7 | 80.7 | 88.1 | 97.9 | 48.6 | 59.5 | 90 | 127 | 167 |
| 16 | UniversityCollegeLondon | UnitedKingdom | 85.3 | 74.4 | 88.2 | 94.6 | 41.2 | 94.6 | 161 | 121 | 128 |
| 17 | DukeUniversity | UnitedStates | 85.1 | 80.7 | 80.6 | 98.3 | 100.0 | 62.5 | 161 | 93 | 86 |
| 18 | UniversityCaliforniaBerkeley | UnitedStates | 84.3 | 77.4 | 84.5 | 99.8 | 37.5 | 64.5 | 161 | 101 | 104 |
| 19 | CornellUniversity | UnitedStates | 84.2 | 76.2 | 86.6 | 97.6 | 34.6 | 69.2 | 161 | 117 | 116 |
| 20 | NorthwesternUniversity | UnitedStates | 83.3 | 72.6 | 86.7 | 96.9 | 78.2 | 59.2 | 111 | 45 | 49 |
| 21 | UniversityofMichigan | UnitedStates | 83.1 | 77.2 | 86.3 | 95.7 | 46.2 | 55.8 | 111 | 77 | 86 |
| 22 | NationalUniversitySingapore | Singapore | 82.8 | 77.4 | 88.2 | 81.3 | 61.9 | 95.8 | 161 | 122 | 128 |
| 22 | UniversityofToronto | Canada | 82.8 | 74.6 | 84.8 | 92.6 | 46.5 | 80.1 | 37 | 59 | 58 |
| 24 | CarnegieMellonUniversity | UnitedStates | 81.9 | 65.8 | 83.7 | 99.7 | 50.4 | 79.1 | 161 | 147 | 167 |
| 25 | LondonSchoolofEconomicsandPoliticalScience | UnitedKingdom | 79.4 | 71.8 | 72.0 | 94.9 | 33.7 | 92.2 | 161 | 155 | 167 |
| 25 | UniversityofWashington | UnitedStates | 79.4 | 67.9 | 79.9 | 99.0 | 47.3 | 56.5 | 161 | 119 | 167 |
| 27 | UniversityofEdinburgh | UnitedKingdom | 79.2 | 66.8 | 74.2 | 97.0 | 36.3 | 92.0 | 111 | 128 | 167 |
| 27 | NewYorkUniversity | UnitedStates | 79.2 | 73.7 | 77.4 | 96.5 | 37.0 | 53.4 | 161 | 82 | 74 |
| 27 | PekingUniversity | China | 79.2 | 83.0 | 85.1 | 74.2 | 100.0 | 53.0 | 111 | 186 | 167 |
| 30 | TsinghuaUniversity | China | 79.0 | 80.2 | 93.2 | 71.4 | 99.8 | 41.0 | 161 | 177 | 167 |
| 31 | UniversityofCaliforniaSanDiego | UnitedStates | 78.7 | 62.9 | 79.8 | 98.6 | 96.5 | 51.9 | 161 | 156 | 167 |
| 32 | UniversityofMelbourne | Australia | 77.5 | 64.9 | 74.2 | 90.3 | 70.1 | 92.7 | 27 | 47 | 50 |

| Rg univarié | Université | Pays | Score | I1 | I2 | I3 | I4 | I5 | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 33 | GeorgiaInstituteofTechnology | UnitedStates | 77.0 | 59.7 | 78.9 | 94.3 | 60.2 | 75.0 | 111 | 52 | 47 |
| 34 | UniversityofBritishColumbia | Canada | 76.2 | 61.8 | 72.2 | 93.4 | 42.6 | 92.2 | 44 | 54 | 55 |
| 34 | LMUMunich | Germany | 76.2 | 65.4 | 72.3 | 91.3 | 100.0 | 66.3 | 75 | 97 | 95 |
| 36 | KingsCollegeLondon | UnitedKingdom | 75.6 | 59.3 | 71.2 | 94.4 | 43.9 | 94.5 | 90 | 149 | 167 |
| 37 | UniversityofIllinoisUrbanaChampaign | UnitedStates | 75.4 | 65.8 | 78.9 | 88.9 | 56.0 | 52.8 | 63 | 58 | 61 |
| 38 | EcolePolytechniqueFederaleLausanne | Switzerland | 75.3 | 58.8 | 66.8 | 94.2 | 76.0 | 98.7 | 111 | 157 | 167 |
| 38 | KarolinskaInstitute | Sweden | 75.3 | 57.0 | 74.4 | 95.9 | 71.4 | 70.3 | 111 | 41 | 44 |
| 40 | UniversityofHongKong | HongKong | 75.1 | 68.8 | 77.9 | 74.2 | 54.0 | 99.5 | 161 | 73 | 74 |
| 41 | TechnicalUniversityofMunich | Germany | 73.5 | 60.3 | 71.2 | 88.4 | 100.0 | 66.8 | 75 | 113 | 108 |
| 42 | McGillUniversity | Canada | 73.2 | 63.4 | 70.8 | 84.7 | 39.8 | 87.5 | 53 | 134 | 167 |
| 43 | UniversityWisconsinMadison | UnitedStates | 73.1 | 68.4 | 73.9 | 86.5 | 45.8 | 43.4 | 111 | 165 | 167 |
| 44 | HongKongUniversityScienceandTechnology | HongKong | 72.7 | 55.2 | 68.4 | 93.1 | 58.1 | 83.4 | 14 | 21 | 23 |
| 45 | HeidelbergUniversity | Germany | 72.3 | 63.6 | 62.2 | 94.2 | 56.0 | 64.7 | 11 | 16 | 17 |
| 46 | UniversityofTokyo | Japan | 72.2 | 79.5 | 85.2 | 63.7 | 52.7 | 32.2 | 161 | 123 | 128 |
| 47 | KULeuven | Belgium | 71.8 | 54.2 | 70.9 | 88.7 | 99.9 | 68.3 | 161 | 64 | 70 |
| 48 | AustralianNationalUniversity | Australia | 71.6 | 52.7 | 72.0 | 85.5 | 61.1 | 94.3 | 161 | 85 | 86 |
| 49 | UniversityTexasatAustin | UnitedStates | 71.4 | 60.9 | 68.3 | 95.7 | 48.2 | 36.7 | 161 | 90 | 70 |
| 50 | BrownUniversity | UnitedStates | 70.8 | 63.8 | 57.9 | 96.7 | 34.1 | 59.0 | 111 | 108 | 116 |
| 50 | WashingtonUniversityStLouis | UnitedStates | 70.8 | 60.8 | 59.7 | 99.4 | 36.9 | 52.1 | 90 | 63 | 50 |
| 52 | NanyangTechnologicalUniversity | Singapore | 70.5 | 49.5 | 63.0 | 90.7 | 94.0 | 95.9 | 161 | 169 | 167 |
| 53 | UniversityCaliforniaSantaBarbara | UnitedStates | 70.0 | 50.0 | 61.5 | 98.8 | 82.0 | 65.4 | 111 | 144 | 167 |
| 54 | UniversityCaliforniaDavis | UnitedStates | 69.5 | 60.9 | 64.5 | 86.0 | 53.4 | 63.7 | 1 | 11 | 12 |
| 54 | UniversityManchester | UnitedKingdom | 69.5 | 56.3 | 65.3 | 84.3 | 44.1 | 88.7 | 27 | 69 | 58 |
| 56 | UniversityofMinnesota | UnitedStates | 69.0 | 57.6 | 67.9 | 87.5 | 90.4 | 37.7 | 161 | 161 | 167 |
| 56 | UniversityNorthCarolinaChapelHill | UnitedStates | 69.0 | 57.9 | 61.8 | 96.7 | 39.2 | 41.5 | 161 | 99 | 86 |
| 58 | ChineseUniversityHongKong | HongKong | 68.5 | 57.0 | 64.4 | 80.6 | 56.8 | 86.6 | 16 | 33 | 34 |
| 59 | UniversityofAmsterdam | Netherlands | 68.4 | 51.3 | 62.6 | 91.9 | 49.9 | 72.2 | 5 | 24 | 24 |
| 60 | PurdueUniversity | UnitedStates | 68.2 | 62.3 | 68.9 | 73.4 | 63.5 | 69.9 | 17 | 100 | 99 |
| 61 | UniversityofSydney | Australia | 67.8 | 50.2 | 63.1 | 85.9 | 67.9 | 84.6 | 11 | 13 | 17 |
| 62 | HumboldtUniversityBerlin | Germany | 67.4 | 61.9 | 67.1 | 76.1 | 38.6 | 65.6 | 63 | 116 | 128 |
| 63 | DelftUniversityTechnology | Netherlands | 67.3 | 53.4 | 72.0 | 68.5 | 99.8 | 88.5 | 161 | 145 | 167 |
| 64 | WageningenUniversityResearch | Netherlands | 67.2 | 46.6 | 53.6 | 95.2 | 100.0 | 80.6 | 161 | 163 | 167 |
| 65 | UniversityofQueensland | Australia | 67.0 | 47.6 | 59.4 | 87.8 | 76.2 | 88.8 | 27 | 22 | 24 |
| 66 | UniversityofSouthernCalifornia | UnitedStates | 66.8 | 49.8 | 58.4 | 95.6 | 37.9 | 63.0 | 44 | 18 | 20 |
| 67 | LeidenUniversity | Netherlands | 66.0 | 44.6 | 63.2 | 87.8 | 78.8 | 70.5 | 161 | 115 | 128 |
| 68 | UtrechtUniversity | Netherlands | 65.8 | 45.5 | 63.1 | 89.9 | 72.6 | 59.6 | 161 | 142 | 167 |
| 69 | UniversityofMarylandCollegePark | UnitedStates | 65.7 | 49.6 | 63.0 | 93.5 | 38.4 | 39.0 | 161 | 143 | 116 |
| 70 | BostonUniversity | UnitedStates | 65.4 | 57.1 | 45.9 | 97.2 | 34.0 | 60.0 | 161 | 148 | 167 |
| 70 | OhioStateUniversity | UnitedStates | 65.4 | 55.5 | 56.2 | 88.0 | 52.0 | 56.4 | 2 | 27 | 22 |
| 72 | ErasmusUniversityRotterdam | Netherlands | 65.2 | 38.9 | 57.2 | 96.7 | 51.7 | 80.6 | 161 | 84 | 86 |
| 72 | ParisSciencesLettresPSLResearchUniversityParis | France | 65.2 | 57.0 | 52.0 | 85.0 | 48.2 | 76.7 | 37 | 48 | 47 |
| 74 | KyotoUniversity | Japan | 64.9 | 71.8 | 78.6 | 50.9 | 93.8 | 28.8 | 161 | 171 | 167 |
| 74 | SeoulNationalUniversity | SouthKorea | 64.9 | 69.3 | 71.2 | 60.6 | 79.8 | 34.1 | 75 | 114 | 128 |
| 76 | UniversityofBristol | UnitedKingdom | 64.8 | 45.7 | 51.2 | 94.7 | 39.3 | 84.8 | 32 | 28 | 35 |
| 77 | PennsylvaniaStateUniversity | UnitedStates | 64.6 | 53.6 | 64.8 | 81.6 | 50.0 | 44.3 | 90 | 126 | 116 |
| 78 | McMasterUniversity | Canada | 63.4 | 45.6 | 48.8 | 89.9 | 89.8 | 78.1 | 63 | 74 | 74 |
| 79 | RWTHAachenUniversity | Germany | 63.3 | 53.4 | 62.4 | 72.8 | 99.7 | 56.4 | 53 | 68 | 70 |
| 80 | UniversityofGlasgow | UnitedKingdom | 63.2 | 43.4 | 48.9 | 92.6 | 38.6 | 90.4 | 44 | 31 | 36 |
| 80 | MonashUniversity | Australia | 63.2 | 47.4 | 55.4 | 80.9 | 71.5 | 83.8 | 7 | 4 | 6 |
| 82 | UniversityofFreiburg | Germany | 62.9 | 47.1 | 54.3 | 83.1 | 100.0 | 67.2 | 90 | 182 | 167 |
| 83 | UniversityofGroningen | Netherlands | 62.8 | 40.7 | 54.8 | 89.0 | 77.2 | 74.4 | 63 | 62 | 58 |
| 83 | MichiganStateUniversity | UnitedStates | 62.8 | 55.1 | 53.3 | 82.6 | 35.7 | 61.2 | 27 | 23 | 21 |
| 85 | UniversityofNewSouthWales | Australia | 62.5 | 40.9 | 57.5 | 82.8 | 49.8 | 91.4 | 161 | 106 | 108 |
| 86 | RiceUniversity | UnitedStates | 62.2 | 45.1 | 41.9 | 98.7 | 42.4 | 72.7 | 111 | 65 | 64 |
| 86 | UppsalaUniversity | Sweden | 62.2 | 42.9 | 56.1 | 84.6 | 79.5 | 68.8 | 22 | 25 | 26 |
| 88 | FreeUniversityBerlin | Germany | 62.1 | 58.7 | 66.5 | 60.6 | 39.8 | 71.8 | 161 | 141 | 167 |
| 89 | DartmouthCollege | UnitedStates | 62.0 | 58.5 | 41.9 | 93.4 | 37.9 | 39.1 | 161 | 183 | 167 |
| 90 | UniversityofHelsinki | Finland | 61.7 | 45.7 | 56.6 | 86.9 | 35.5 | 53.3 | 111 | 194 | 167 |
| 91 | UniversityofWarwick | UnitedKingdom | 61.6 | 46.6 | 52.0 | 80.4 | 41.0 | 91.9 | 53 | 49 | 55 |
| 92 | TechnicalUniversityBerlin | Germany | 61.5 | 49.0 | 58.0 | 74.2 | 97.8 | 62.4 | 44 | 158 | 167 |
| 93 | LundUniversity | Sweden | 61.3 | 40.8 | 52.6 | 85.9 | 70.9 | 76.1 | 10 | 9 | 8 |
| 94 | UniversityofTubingen | Germany | 61.2 | 45.7 | 55.4 | 83.0 | 55.4 | 60.8 | 4 | 3 | 2 |
| 95 | UniversityofBasel | Switzerland | 60.9 | 39.8 | 39.9 | 91.0 | 99.9 | 95.9 | 161 | 172 | 167 |
| 95 | KoreaAdvancedInstituteScienceandTechnology | SouthKorea | 60.9 | 56.3 | 59.2 | 70.4 | 100.0 | 35.6 | 161 | 125 | 104 |

| Rg univarié | Université | Pays | Score | I1 | I2 | I3 | I4 | I5 | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 97 | DurhamUniversity | UnitedKingdom | 60.8 | 44.8 | 47.9 | 84.6 | 36.8 | 89.0 | 22 | 26 | 32 |
| 98 | EmoryUniversity | UnitedStates | 60.7 | 47.1 | 40.4 | 98.1 | 42.3 | 53.3 | 161 | 137 | 167 |
| 99 | UniversityCaliforniaIrvine | UnitedStates | 60.6 | 43.4 | 44.9 | 93.6 | 43.9 | 65.2 | 8 | 2 | 2 |
| 100 | UniversityofBonn | Germany | 60.5 | 47.3 | 41.6 | 91.3 | 83.5 | 58.4 | 90 | 96 | 86 |
| 100 | UniversityofColoradoBoulder | UnitedStates | 60.5 | 44.8 | 45.8 | 97.4 | 37.5 | 42.4 | 111 | 199 | 167 |
| 100 | UniversityofPittsburgh | UnitedStates | 60.5 | 45.7 | 48.7 | 94.9 | 39.9 | 36.8 | 161 | 200 | 167 |
| 103 | MaastrichtUniversity | Netherlands | 60.4 | 40.2 | 49.7 | 79.9 | 87.6 | 96.7 | 111 | 87 | 80 |
| 104 | UniversityofSheffield | UnitedKingdom | 60.1 | 42.9 | 45.7 | 86.8 | 43.7 | 85.4 | 3 | 6 | 7 |
| 105 | UniversityofBern | Switzerland | 60.0 | 42.7 | 43.6 | 85.4 | 81.0 | 85.7 | 75 | 109 | 99 |
| 105 | VanderbiltUniversity | UnitedStates | 60.0 | 47.7 | 41.4 | 96.9 | 52.8 | 38.9 | 111 | 192 | 167 |
| 107 | GhentUniversity | Belgium | 59.8 | 44.4 | 56.2 | 77.6 | 84.3 | 56.8 | 53 | 67 | 64 |
| 108 | UniversityofMontreal | Canada | 59.6 | 44.5 | 48.9 | 79.0 | 65.3 | 83.2 | 8 | 8 | 5 |
| 109 | AarhusUniversity | Denmark | 59.4 | 37.3 | 53.6 | 82.7 | 67.6 | 75.3 | 90 | 10 | 14 |
| 109 | UniversityofCopenhagen | Denmark | 59.4 | 44.0 | 39.7 | 90.6 | 44.6 | 79.3 | 37 | 5 | 4 |
| 111 | SungkyunkwanUniversity | SouthKorea | 59.3 | 54.1 | 55.1 | 69.5 | 93.7 | 44.7 | 75 | 50 | 42 |
| 111 | UniversityofWesternAustralia | Australia | 59.3 | 33.5 | 46.1 | 90.6 | 54.5 | 91.9 | 90 | 195 | 167 |
| 113 | UniversityofGottingen | Germany | 59.2 | 47.1 | 50.3 | 82.4 | 33.7 | 58.6 | 63 | 14 | 12 |
| 113 | UniversityofVirginia | UnitedStates | 59.2 | 52.4 | 40.7 | 89.3 | 40.6 | 46.5 | 90 | 180 | 167 |
| 115 | EcolePolytechnique | France | 59.1 | 54.4 | 38.2 | 75.1 | 70.8 | 93.4 | 161 | 173 | 167 |
| 116 | FudanUniversity | China | 58.9 | 59.5 | 57.5 | 65.1 | 53.0 | 38.7 | 75 | 94 | 95 |
| 117 | IndianaUniversity | UnitedStates | 58.7 | 50.5 | 46.5 | 81.9 | 51.5 | 50.5 | 5 | 35 | 31 |
| 117 | TrinityCollegeDublin | Ireland | 58.7 | 45.3 | 44.8 | 79.0 | 41.2 | 92.1 | 53 | 44 | 52 |
| 119 | UniversityofAlberta | Canada | 58.6 | 47.2 | 51.2 | 70.5 | 63.4 | 84.8 | 63 | 61 | 64 |
| 119 | CityUniversityofHongKong | HongKong | 58.6 | 40.8 | 48.6 | 79.4 | 65.9 | 84.4 | 13 | 20 | 19 |
| 121 | QueenMaryUniversityofLondon | UnitedKingdom | 58.5 | 32.8 | 38.2 | 96.7 | 39.1 | 95.5 | 161 | 191 | 167 |
| 122 | RadboudUniversityNijmegen | Netherlands | 58.4 | 33.9 | 50.4 | 89.4 | 44.8 | 68.8 | 161 | 15 | 14 |
| 123 | GeorgetownUniversity | UnitedStates | 58.3 | 51.8 | 37.2 | 87.5 | 66.3 | 49.6 | 161 | 167 | 167 |
| 123 | PierreandMarieCurieUniversity | France | 58.3 | 49.5 | 39.8 | 85.7 | 31.8 | 66.6 | 75 | 72 | 67 |
| 125 | UniversityofMannheim | Germany | 58.2 | 41.8 | 53.0 | 80.5 | 57.8 | 55.9 | 34 | 37 | 29 |
| 126 | ArizonaStateUniversity | UnitedStates | 58.1 | 41.5 | 48.3 | 87.1 | 35.7 | 55.2 | 44 | 159 | 128 |
| 126 | Charite Universitatsmedizin Berlin | Germany | 58.1 | 43.3 | 39.2 | 88.6 | 87.5 | 60.7 | 90 | 51 | 44 |
| 126 | UniversityofSouthampton | UnitedKingdom | 58.1 | 37.9 | 43.7 | 86.0 | 38.5 | 92.0 | 32 | 130 | 128 |
| 129 | UniversiteCatholiquedeLouvain | Belgium | 58.0 | 38.9 | 52.2 | 78.6 | 54.0 | 76.7 | 44 | 1 | 1 |
| 130 | UniversityofExeter | UnitedKingdom | 57.8 | 34.5 | 39.3 | 93.7 | 34.6 | 88.8 | 161 | 102 | 104 |
| 130 | UniversityofGeneva | Switzerland | 57.8 | 35.8 | 43.6 | 82.9 | 68.7 | 98.2 | 90 | 105 | 116 |
| 132 | UniversityofScienceandTechnologyofChina | China | 57.7 | 52.7 | 49.1 | 76.9 | 81.7 | 27.9 | 161 | 197 | 167 |
| 133 | KarlsruheInstituteofTechnology | Germany | 57.6 | 44.6 | 47.7 | 75.7 | 97.9 | 62.9 | 90 | 46 | 52 |
| 134 | UniversityofAdelaide | Australia | 57.5 | 35.9 | 43.1 | 85.3 | 67.3 | 86.6 | 19 | 71 | 70 |
| 134 | StockholmUniversity | Sweden | 57.5 | 33.1 | 48.4 | 89.1 | 34.5 | 72.7 | 161 | 110 | 95 |
| 136 | UniversityofZurich | Switzerland | 57.4 | 42.0 | 36.0 | 87.0 | 43.2 | 90.8 | 90 | 55 | 62 |
| 137 | PohangUniversityofScienceandTechnology | SouthKorea | 57.3 | 47.9 | 49.8 | 76.4 | 99.8 | 34.3 | 161 | 193 | 167 |
| 137 | UniversityofYork | UnitedKingdom | 57.3 | 40.5 | 44.9 | 81.6 | 34.1 | 85.3 | 44 | 34 | 40 |
| 139 | UniversityofLeeds | UnitedKingdom | 57.1 | 42.2 | 46.7 | 77.8 | 38.8 | 82.1 | 34 | 30 | 26 |
| 140 | PompeuFabraUniversity | Spain | 56.9 | 34.7 | 38.9 | 97.1 | 40.0 | 62.3 | 111 | 79 | 78 |
| 141 | UniversityofBirmingham | UnitedKingdom | 56.8 | 36.5 | 38.3 | 89.7 | 37.2 | 86.3 | 32 | 29 | 29 |
| 141 | EindhovenUniversityofTechnology | Netherlands | 56.8 | 41.3 | 49.0 | 72.3 | 100.0 | 73.9 | 161 | 89 | 95 |
| 143 | UniversityofFlorida | UnitedStates | 56.6 | 52.8 | 51.2 | 68.4 | 80.7 | 37.9 | 44 | 95 | 92 |
| 143 | UniversityofStAndrews | UnitedKingdom | 56.6 | 42.9 | 42.5 | 76.6 | 33.6 | 95.5 | 111 | 70 | 86 |
| 145 | UniversityofCologne | Germany | 56.4 | 39.8 | 42.7 | 84.8 | 76.8 | 57.1 | 22 | 32 | 29 |
| 146 | UniversityofOslo | Norway | 56.3 | 36.0 | 42.8 | 86.8 | 41.9 | 73.4 | 14 | 17 | 12 |
| 147 | AutonomousUniversityofBarcelona | Spain | 56.2 | 43.3 | 36.1 | 89.5 | 42.1 | 60.1 | 18 | 12 | 16 |
| 147 | UniversityofNottingham | UnitedKingdom | 56.2 | 40.5 | 41.9 | 80.5 | 38.5 | 84.4 | 27 | 38 | 38 |
| 147 | UniversityofSussex | UnitedKingdom | 56.2 | 34.1 | 38.1 | 89.5 | 33.3 | 91.5 | 161 | 76 | 86 |
| 150 | LancasterUniversity | UnitedKingdom | 55.9 | 37.0 | 40.5 | 83.2 | 33.7 | 91.1 | 90 | 43 | 55 |
| 150 | UniversityofNotreDame | UnitedStates | 55.9 | 49.8 | 37.9 | 83.0 | 35.3 | 50.7 | 90 | 136 | 116 |
| 152 | UniversityofLausanne | Switzerland | 55.7 | 30.4 | 47.4 | 78.5 | 78.4 | 90.8 | 161 | 104 | 116 |
| 153 | TechnicalUniversityofDenmark | Denmark | 55.6 | 35.9 | 39.5 | 81.6 | 63.8 | 91.6 | 44 | 81 | 86 |
| 153 | UniversityofRochester | UnitedStates | 55.6 | 41.7 | 34.2 | 90.2 | 39.5 | 63.7 | 53 | 42 | 38 |
| 155 | ScuolaSuperioreSantAnna | Italy | 55.5 | 41.6 | 36.0 | 88.0 | 87.8 | 48.3 | 161 | 189 | 167 |
| 155 | TUDresden | Germany | 55.5 | 42.8 | 46.9 | 74.2 | 95.0 | 52.6 | 75 | 112 | 116 |
| 155 | UlmUniversity | Germany | 55.5 | 37.1 | 37.4 | 90.0 | 77.3 | 56.5 | 75 | 152 | 128 |
| 158 | CaseWesternReserveUniversity | UnitedStates | 55.2 | 44.4 | 34.9 | 90.1 | 36.7 | 46.1 | 111 | 170 | 167 |
| 159 | UniversityofLeicester | UnitedKingdom | 55.0 | 32.2 | 34.5 | 91.4 | 35.2 | 89.9 | 90 | 132 | 167 |
| 159 | TexasA&MUniversity | UnitedStates | 55.0 | 50.4 | 56.3 | 60.2 | 41.2 | 52.0 | 161 | 80 | 78 |

| Rg univarié | Université | Pays | Score | I1 | I2 | I3 | I4 | I5 | Rg HD | Rg MK | Rg MK-prof |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 161 | UniversityofArizona | UnitedStates | 54,9 | 44,3 | 38,7 | 85,4 | 46,1 | 43,3 | 63 | 78 | 67 |
| 162 | UniversityofCaliforniaSantaCruz | UnitedStates | 54,8 | 31,5 | 35,0 | 99,9 | 38,2 | 51,6 | 161 | 146 | 116 |
| 162 | CardiffUniversity | UnitedKingdom | 54,8 | 35,8 | 39,3 | 84,5 | 35,5 | 80,6 | 19 | 36 | 32 |
| 162 | UniversityofErlangenNuremberg | Germany | 54,8 | 39,8 | 44,0 | 77,6 | 95,9 | 53,0 | 161 | 168 | 167 |
| 165 | UniversityofVienna | Austria | 54,7 | 42,6 | 47,6 | 66,0 | 33,2 | 93,4 | 161 | 176 | 167 |
| 165 | VrijeUniversiteitAmsterdam | Netherlands | 54,7 | 32,5 | 40,7 | 89,1 | 56,3 | 60,7 | 63 | 57 | 43 |
| 165 | UniversityofWurzburg | Germany | 54,7 | 35,8 | 42,8 | 86,5 | 46,1 | 53,7 | 75 | 19 | 10 |
| 168 | UniversityofAlabamaBirmingham | UnitedStates | 54,6 | 39,7 | 32,8 | 95,1 | 67,6 | 34,6 | 161 | 138 | 128 |
| 169 | NanjingUniversity | China | 54,5 | 49,6 | 47,0 | 64,9 | 77,8 | 54,4 | 63 | 86 | 77 |
| 169 | TuftsUniversity | UnitedStates | 54,5 | 44,1 | 33,4 | 88,1 | 36,7 | 52,7 | 63 | 66 | 47 |
| 171 | UniversityofCapeTown | SouthAfrica | 54,4 | 30,5 | 36,2 | 87,0 | 88,5 | 81,1 | 161 | 175 | 167 |
| 172 | RutgersStateUniversityNewJersey | UnitedStates | 54,3 | 43,0 | 45,3 | 79,7 | 36,9 | 39,2 | 161 | 153 | 167 |
| 173 | KTHRoyalInstituteTechnology | Sweden | 54,2 | 42,3 | 46,3 | 67,0 | 50,9 | 83,7 | 75 | 139 | 128 |
| 173 | UniversityofMunster | Germany | 54,2 | 40,3 | 42,1 | 81,2 | 60,8 | 47,7 | 27 | 154 | 167 |
| 175 | UniversiteLibredeBruxelles | Belgium | 54,0 | 32,4 | 42,2 | 80,4 | 48,4 | 83,2 | 44 | 7 | 8 |
| 175 | NewcastleUniversity | UnitedKingdom | 54,0 | 33,4 | 40,8 | 80,6 | 41,6 | 86,7 | 53 | 40 | 36 |
| 177 | UniversityofLiverpool | UnitedKingdom | 53,9 | 32,7 | 34,7 | 86,8 | 37,1 | 89,9 | 53 | 60 | 74 |
| 177 | ZhejiangUniversity | China | 53,9 | 57,0 | 63,7 | 45,1 | 89,4 | 25,2 | 161 | 131 | 167 |
| 179 | UniversityofLuxembourg | Luxembourg | 53,8 | 26,8 | 32,4 | 91,9 | 39,6 | 99,8 | 161 | 188 | 167 |
| 179 | UniversityofTwente | Netherlands | 53,8 | 35,5 | 47,0 | 68,5 | 83,6 | 85,3 | 161 | 98 | 104 |
| 181 | ParisSudUniversity | France | 53,7 | 40,4 | 33,1 | 86,8 | 32,3 | 63,5 | 90 | 196 | 167 |
| 182 | EcoleNormaleSuperieuredeLyon | France | 53,6 | 40,1 | 37,8 | 80,9 | 37,9 | 67,3 | 22 | 39 | 40 |
| 182 | HongKongPolytechnicUniversity | HongKong | 53,6 | 39,1 | 48,1 | 67,7 | 45,8 | 79,9 | 111 | 92 | 95 |
| 184 | ScuolaNormaleSuperiorediPisa | Italy | 53,4 | 53,7 | 33,3 | 75,4 | 37,7 | 49,4 | 161 | 166 | 167 |
| 185 | UniversityofAberdeen | UnitedKingdom | 53,3 | 30,3 | 33,6 | 87,1 | 41,9 | 93,3 | 75 | 53 | 58 |
| 186 | UniversityofMiami | UnitedStates | 53,1 | 42,5 | 24,4 | 90,3 | 37,2 | 67,1 | 161 | 140 | 167 |
| 187 | UniversityofDundee | UnitedKingdom | 52,9 | 26,3 | 32,0 | 94,5 | 43,4 | 79,8 | 161 | 198 | 167 |
| 188 | UniversityofEastAnglia | UnitedKingdom | 52,8 | 29,9 | 31,6 | 90,5 | 32,2 | 85,6 | 111 | 187 | 167 |
| 188 | ShanghaiJiaoTongUniversity | China | 52,8 | 53,5 | 62,5 | 44,1 | 88,9 | 33,4 | 161 | 129 | 167 |
| 190 | AaltoUniversity | Finland | 52,7 | 37,8 | 35,8 | 79,5 | 52,5 | 72,5 | 37 | 83 | 86 |
| 191 | UniversityofMassachusetts | UnitedStates | 52,6 | 36,3 | 32,2 | 88,8 | 51,0 | 54,8 | 90 | 150 | 167 |
| 192 | UniversityofAuckland | NewZealand | 52,5 | 32,5 | 40,7 | 73,8 | 67,3 | 90,6 | 111 | 75 | 80 |
| 193 | NortheasternUniversity | UnitedStates | 52,4 | 35,6 | 26,8 | 91,7 | 35,4 | 70,1 | 111 | 162 | 167 |
| 194 | LomonosovMoscowStateUniversity | RussianFederation | 52,3 | 74,2 | 61,6 | 15,8 | 89,0 | 61,1 | 161 | 185 | 167 |
| 195 | TilburgUniversity | Netherlands | 52,1 | 38,4 | 50,1 | 61,3 | 59,0 | 75,8 | 161 | 88 | 95 |
| 196 | ParisSorbonneUniversityParis4 | France | 52,0 | 44,6 | 34,3 | 77,1 | 32,1 | 57,9 | 161 | 135 | 167 |
| 197 | RoyalHollowayUniversityLondon | UnitedKingdom | 51,9 | 33,9 | 27,9 | 85,0 | 35,0 | 93,0 | 161 | 91 | 104 |
| 198 | UniversityofCaliforniaRiverside | UnitedStates | 51,7 | 30,0 | 30,9 | 92,3 | 37,4 | 63,9 | 161 | 184 | 167 |
| 198 | UniversityofGothenburg | Sweden | 51,7 | 26,7 | 39,2 | 87,5 | 38,9 | 62,5 | 161 | 164 | 116 |
| 198 | NationalTaiwanUniversity | Taiwan | 51,7 | 50,4 | 54,7 | 53,4 | 65,0 | 33,4 | 161 | 178 | 167 |

TABLE A.1 – Classement des 200 meilleures universités du monde publié par le *"Times Higher Education"*

# Table des figures


1 Boîte à moustaches de la variable "*Teaching*" . . . . . 3
2 Classement orienté à partir du centre en dimension 1 . . . . . 3
3 Classement orienté d'observations à partir du centre en dimension 1 . . . . 3
4 Bagplot pour les variables *"Teaching"* et *"Research"* . . . . . 4

1.1 Distribution issue de la famille $\mathcal{P}^1$ . . . . . 10
1.2 Distribution issue de la famille $\mathcal{P}^1$ avec $x \leqslant \mu$ . . . . . 11
1.3 Distribution issue de la famille $\mathcal{P}^1$ avec $x > \mu$ . . . . . 11
1.4 Demi-espaces et contours de densité . . . . . 13
1.5 Demi-espace minimisant (1.1) . . . . . 14
1.6 Exemple de demi-espaces contenant $x^*$ sur leur frontière . . . . . 15
1.7 Distribution à symétrie sphérique et hyperplan qui minimise (1.1) . . . . . 17
1.8 Données ordonnées par ordre croissant . . . . . 17
1.9 Demi-espaces contenant $x_{(i)}$ sur leur frontière . . . . . 18
1.10 Profondeur de Tukey en perspective . . . . . 19
1.11 Profondeur de Tukey . . . . . 19
1.12 Observations bivariées pour les variables *Teaching* et *Research* . . . . . 19
1.13 Représentation de quelques hyperplans dont la frontière passe par l'observation d'intérêt . . . . . 20
1.14 Contours de profondeur de Tukey . . . . . 23
1.15 Contours de profondeur pour $\alpha = 0.05, 0.15, 0.25, 0.35$ et $0.45$ (en lignes pleines) et les contours de densité de la distribution de Cauchy bivariée (en pointillés) [22]. . . . . 30
1.16 Génération d'une distribution sous forme de banane . . . . . 31
1.17 Contours de profondeur de Tukey pour une distribution sous forme de banane, produits à partir d'un échantillon de 400 observations. . . . . 31
1.18 Contours de profondeur de Tukey produits à partir d'un échantillon de 400 observations issues d'une distribution de Cauchy bivariée. . . . . 32

2.1 Fonction de répartition classique . . . . . 36
2.2 Fonction de répartition définie à partir du centre (pris en zéro) pour $x \geqslant 0$ 36

3.1 Comparaison entre la profondeur de Tukey pour les données relatives à la variable *Teaching* avec la profondeur de Monge-Kantorovich de ces mêmes données . . . . . 50
3.2 Représentation des variables *Teaching* et *Research* . . . . . 51

3.3 Contours de profondeur de Tukey pour les variables *Teaching* et *Research* . 51
3.4 Comparaison entre la profondeur de Tukey pour les données relatives aux indicateurs *Teaching* et *Research* avec la profondeur de Monge-Kantorovich de ces mêmes données . . . . . 52
3.5 Génération d'une distribution uniforme sphérique sur la boule unité . . . . 54
3.6 Représentation des contours de profondeur de Monge-Kantorovich . . . . . 54
3.7 Représentation des contours de MK pour une distribution de Cauchy bivariée 55
3.8 Représentation des contours de MK sur les données . . . . . 55

4.1 Boîtes à moustaches pour les cinq indicateurs . . . . . 58
4.2 Bagplot pour les cinq indicateurs, 1ère partie . . . . . 59
4.3 Bagplot pour les cinq indicateurs, 2ème partie . . . . . 60
4.4 Bagplot pour les cinq indicateurs, 3ème partie . . . . . 61
4.5 Rangs de Tukey . . . . . 66
4.6 Rangs de Monge-Kantorovich (Rg MK-prof) . . . . . 66
4.7 Rangs de Monge-Kantorovich (Rg MK) . . . . . 67

# Liste des tableaux


4.1 Résumé à 5 valeurs pour *Teaching* . . . . . 57
4.2 Résumé à 5 valeurs pour *Research* . . . . . 57
4.3 Résumé à 5 valeurs pour *Citations* . . . . . 57
4.4 Résumé à 5 valeurs pour *Industry Income* . . . . . 57
4.5 Résumé à 5 valeurs pour *International Outlook* . . . . . 57
4.6 Résultats des universités belges pour les indicateurs *Enseignement, Recherche* et *Citations* . . . . . 62
4.7 Résultats des universités belges pour les indicateurs *Financement des entreprises* et *Perspectives internationales* . . . . . 62
4.8 Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich, pour les 5 meilleures universités . . . . 63
4.9 Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich pour les 5 moins bonnes universités du classement . . . . . 64
4.10 Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich pour les 5 universités du milieu de classement . . . . . 64
4.11 Comparaisons des rangs univariés et multivariés obtenus par la profondeur de Tukey et de Monge-Kantorovich pour les universités belges . . . . . 65

A.1 Classement des 200 meilleures universités du monde publié par le *"Times Higher Education"* . . . . . 75

# Bibliographie


[1] World University Ranking 2018. https://www.timeshighereducation.com/world-university-rankings/2018/, consulté le 31/05/2018.

[2] Martin Bilodeau et David Brenner. *Theory of Multivariate Statistics*. Springer Texts in Statistics. Springer-Verlag New York, 1999.

[3] Yann Brenier. *Polar Factorization and Monotone Rearrangement of Vector-Valued Functions*. Université de Paris VI, 1990.

[4] W. Bryc. *The Normal Distribution*. Springer, New York, 1995.

[5] Luis A. Caffarelli. Boundary Regularity of Maps with Convex Potentials II. *Annals of Mathematics*, 144(3) : 453–496, 1996.

[6] Cécile Carrère, Didier Lesesvre, et Paul Pegon. *Théorie générale du transport et applications*. Thèse de Master, Ecole Normale Supérieure de Cachan.

[7] Victor Chernozhukov, Alfred Galichon, Marc Hallin, et Marc Henry. Monge-Kantorovich Depth, Quantiles, Ranks And Signs. *The Annals of Statistics*, 45(1) : 223–256, 2017.

[8] Do Chuong B. *The Multivariate Gaussian Distribution*. 2008. https://www.semanticscholar.org/paper/The-Multivariate-Gaussian-Distribution-Do/055df34c304f5ab0d65ccd3f6d6f09d25630346e?tab=abstract, consulté le 31/05/2018.

[9] Subhajit Dutta, Anil Ghosh, et Probal Chaudhuri. Some Intriguing Properties of Tukey's Half-Space Depth. *Bernoulli*, 17(4) : 1420–1434, 2011.

[10] Marc Hallin. On Distribution and Quantile Functions, Ranks and Signs in $R^d$. 2017. https://ideas.repec.org/p/eca/wpaper/2013-258262.html, consulté le 31/05/2018.

[11] Marc Hallin et Davy Paindaveine. Optimal Tests for Multivariate Location Based on Interdirections and Pseudo-Mahalanobis Ranks. *The Annals of Statistics*, 30(4) : 1103–1133, 2002.

[12] Juha Heinonen. *Lectures On Lipschitz Analysis*. 2005. https://www.semanticscholar.org/paper/Lectures-on-Lipschitz-Analysis-Heinonen/be6aaafe9fe3984e6bfbbe22a4319b684a30652b, consulté le 31/05/2018.

[13] Lee Hwi-Young, Park Hyoung-Jin, et Kim Hyoung-Moon. A Clarification of the Cauchy Distribution. *Communications for Statistical Applications and Methods*, 21(2) : 183–191, 2014.

[14] Soh Kaycheng. Times Higher Education 100 under 50 ranking : old wine in a new bottle ? *Quality in Higher Education*, 19(1) : 111–121, 2013.

[15] R. Y. Liu. On a Notion of Data Depth Based on Random Simplicies. *The Annals of Statistics*, 18(1) : 405–414, 1990.

[16] R. Y. Liu et Kesar Singh. A Quality Index Based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association*, 88(421) : 252–260, 1993.

[17] Pierre Mathonet. *Topologie générale*. Notes de cours, Université de Liège, 2013-2014. `http://www.geodiff.ulg.ac.be/topologie/NotesTopologie.pdf`, consulté le 31/05/2018.

[18] Robert J. McCann. Existence and Uniqueness of Monotone Measure-Preserving Maps. *Duke mathematical Journal*, 80(2) :309–323, Novembre 1995.

[19] Samuel Nicolay. *Théorie de la mesure*. Notes de cours, Université de Liège, 2011-2012. `http://www.afaw.ulg.ac.be/mesure/mesure-2011.pdf`, consulté le 31/05/2018.

[20] Rodolphe Palm. L'analyse en composantes principales : principes et applications. *Notes de Statistique et d'Informatique*, 1998. `https://www.tandfonline.com/loi/cqhe20`, consulté le 31/05/2018.

[21] R.Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[22] Peter J. Rousseeuw et Ida Ruts. The Depth Function of a Population Distribution. *Metrika*, 49(3) : 213–244, 1999.

[23] Peter J. Rousseeuw, Ida Ruts, et John W. Tukey. The Bagplot : A Bivariate Boxplot. *The American Statistician*, 53(4) : 382–387, 1999.

[24] Jean Schmets. *Analyse mathématique*. Notes de cours, Université de Liège, 2004-2005. `http://www.anmath.ulg.ac.be/js/ens/am.pdf`, consulté le 31/05/2018.

[25] Luc Thoma. Quantiles en statistique multivariée. Thèse de Master, Université de Liège, 2002-2003.

[26] John W. Tukey. *Mathematics and the Picturing of Data*. volume 2, pages 523–531, Montreal, 1975. Canadian Mathematical Congress.

[27] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley publishing company, 1977.

[28] Cédric Villani. Topics in Optimal Transportation. In *Graduate studies in mathematics*, volume 58. American Mathematical Society, 2003.

[29] Cédric Villani. *Optimal Transport, Old and New*. Springer, 2008.

[30] Yijun Zuo et Robert Serfling. General Notions of Statistical Depth Function. *The Annals of Statistics*, 28(2) : 461–482, 2000.

[31] Yijun Zuo et Robert Serfling. General Properties and Convergence Results for Contours of Sample Statistical Depth Functions. *The Annals of Statistics*, 28(2) : 483–499, 2000.