Ordonner $\mathbb{R}^d$, $d \geq 2$

fonctions de répartition, fonctions quantiles et transports de mesures

Marc Hallin

ECARES et Département de Mathématique
Université libre de Bruxelles
mhallin@ulb.ac.be

Résumé. Contrairement à la droite réelle, $\mathbb{R}^d$, pour $d \geq 2$, n'est pas ordonné de façon canonique. Une conséquence de cette absence d'un ordre canonique est que des concepts aussi fondamentaux que ceux de fonction de répartition ou de fonction quantile ne sont pas davantage définis de façon canonique. La définition usuelle, fondée sur les odres marginaux, ne jouit d'aucune des propriétés désirables pour une fonction de distribution; en particulier, son inverse fournit une notion de quantile sans grande signification. Nous montrons comment une caractérisation de type transport de mesures permet d'étendre ces notions à $\mathbb{R}^d$, construisant ainsi un ordre spécifique à chaque loi (dans la population), induit par les observations (dans léchantillon). A la différence des nombeuses propositions faites dans la littérature, les concepts ainsi obtenus possèdent toutes les propriétés qui font des rangs et des quantiles univariés des outils fondamentaux pour l'analyse des données aussi bien que pour l'inférence.

Abstract. Unlike the real line, the real space $\mathbb{R}^d$, for $d \geq 2$, is not canonically ordered. As a result, such fundamental concepts as distribution and quantile functions, in dimension $d \geq 2$, are not canonically defined. The classical definition of a multivariate distribution function, based on marginal orderings, does not enjoy any of the quintessential properties of the univariate concept; in particular, its inverse defines an essentially meaningless quantile function. We show how a measure transportation characterization of the univariate concept readily extends to $\mathbb{R}^d$, leading to distribution-specific orderings in population, data-driven ones in samples. Contrary to the many concepts that can be found in the literature, the corresponding distribution and quantile functions possess all the properties that make their univariate counterparts successful tools for data analysis and statistical inference.

Ordonner $\mathbb{R}^d$, $d \geq 2$

fonctions de répartition, fonctions quantiles et transports de mesures

Marc Hallin

ECARES et Département de Mathématique
Université libre de Bruxelles
mhallin@ulb.ac.be

Contrairement à la droite réelle, l'espace réel $\mathbb{R}^d$, en dimension $d \geq 2$, n'est pas ordonné de façon canonique. Une conséquence de cette absence d'ordre est que des concepts aussi fondamentaux que ceux de fonction de répartition ou de fonction quantile, intimement liés à la notion d'ordre, ne sont pas davantage définis de façon canonique en dimension $d \geq 2$.

On peut protester, bien sûr, que la notion familière de fonction de répartition multivariée, fondée sur les ordres marginaux, et étroitement liée à la notion de transformation copule[1], satisfait aux besoins de l'analyse multivariée. Ce n'est pas le cas. En particulier, la fonction quantile (l'inverse de la fonction de répartition) associée à cette notion traditionnelle ne présente aucune des propriétés faisant l'intérêt de la notion univariée de quantile: contours ouverts, étroite dépendance au choix d'un système de coordonnées, très mauvaises propriétés d'équivariance, fût-ce par rotation, ... (voir, par exemple, Genest et Rivest (2001)). La transformation copule, semblablement, ne présente que de façon marginale la propriété, essentielle dans le cas des fonctions de répartition univariées, de réaliser une *probability integral transformation*[2].

Cette absence d'une notion satisfaisante de fonction de répartition (donc de fonction quantile) s'étend au cas empirique, où elle est sans doute encore plus regrettable. Les contreparties empiriques de ces notions sont en effet la fonction de répartition empirique (liée à la fonction de répartition population par le Théorème de Glivenko-Cantelli) et la collection des quantiles empiriques. Tous deux (la fonction de répartition empirique via les *rangs*) sont des outils statistique fondamentaux, tant en analyse descriptive qu'en inférence.

L'objectif de cet exposé est de montrer comment, de façon très naturelle, une réinterprétation de la notion univariée de fonction de répartition du point de vue des transports de mesures s'étend immédiatement au cas multivarié, et conduit, tant dans le cadre de la population que dans celui de l'échantillon, à une définition qui, contrairement aux

[1] Il s'agit de la transformation $\mathbf{X} = (X_1, \ldots, X_d) \mapsto (F_1(X_1), \ldots, F_d(X_d))$, où les F_i sont des fonctions de répartition marginales.

[2] Si X est de loi (absolument continue) P, $F(X)$ est uniforme sur l'intervalle $[0, 1]$; dans le langage des transports de mesures, la fonction de répartition F d'une loi P "pousse" celle-ci vers l'uniforme $\mathrm{U}_{[0,1]}$ sur l'intervalle unité, ce qui se note $F\#\mathrm{P} = \mathrm{U}_{[0,1]}$.

nombreuses propositions faites dans la littérature (parmi lesquelles les diverses notions de *profondeur statistique*), conduit aux propriétés (Glivenko-Cantelli, "distribution-freeness", préservation de l'efficacité semiparamétrique, invariance/équivariance par rapport un groupe de transformations "order-preserving" cohérent, ...) qui font le succès des concepts univariés classiques.

Une structure d'ordre sur $\mathbb{R}^d$ ne peut pas prendre la forme de l'ordre total, "de la gauche vers la droite" sur lequel est construite la fonction de répartition univariée usuelle F. Un premier pas consiste à substituer à F la fonction de répartition "center-outward" $\mathbf{F}_\pm := 2F - 1$, à valeurs dans la boule unité $[-1, 1]$ de $\mathbb{R}$. Clairement, $\mathbf{F}_\pm$ contient la même information que F, mais ordonne les points depuis le "centre de la loi" (la médiane) vers l'extérieur. Tout comme F, $\mathbf{F}_\pm$ est monotone croissante, et constitue une *probability integral transformation*[3].

A ce stade, on ne voit pas bien clairement le lien qui pourrait exister entre la définition de $\mathbf{F}_\pm$ et les transports de mesure. Un transport de mesure optimal, en effet, est la solution d'un problème d'optimisation sous contrainte (problème de Monge-Kantorovich). Ce problème ne fait sens que sous des lois pour lesquelles les coûts des transports considérés sont finis—c'est-à-dire sous des conditions de moments. Et une définition des notions de fonction de répartition et de fonction quantile exigeant l'existence de moments finis ne semble guère pertinente.

En l'occurrence, les coûts de transport considérés ici sont des coûts quadratiques—le carré de la distance (euclidienne) parcourue. Ce choix (qui n'était pas celui de Monge) n'est nullement lié à une approche de type L^2 ou Gaussienne des problèmes statistiques qui nous occupent—d'autant qu'un quantile est un objet de nature L^1—mais aux propriétés de monotonicité des transports optimaux correspondants. Un remarquable résultat de Yann Brenier (1987, 1991) montre en effet que (limitons-nous, pour la commodité, à des lois absolument continues) les solutions du problème de Monge-Kantorovich, essentiellement, appartiennent à la classe des gradients de fonctions convexes; en dimension 1, cette classe est celle des fonctions monotones croissantes. Malheureusement, le résultat de Brenier exige des moments d'ordre 2 finis, une hypothèse qui n'est pas bienvenue dans la définition d'une notion de fonction de répartition ou d'un quantile.

Le contexte considéré par Brenier est celui du problème de Monge-Kantorovich—un problème d'optimisation de nature analytique, qui s'inscrit tout naturellement dans un cadre de type Sobolev. Cette vision analytique a été considérablement élargie grâce à l'intuition de Robert McCann (1995), qui a vu, derrière le résultat de Brenier, une propriété de nature plus géométrique. Son résultat (limitons-nous, toujours pour la commodité, à des lois absolument continues) montre que, quelles que soient les lois P_1 et P_2

[3]Si X est de loi (absolument continue) P, $\mathbf{F}_\pm(X)$ est uniforme sur la boule unité $[-1, 1]$; dans le langage des transports de mesures, la fonction de répartition center-outward $\mathbf{F}_\pm$ d'une loi P "pousse" celle-ci vers $\mathrm{U}_{[-1,1]}$, ce qui se note $\mathbf{F}_\pm \# \mathrm{P} = \mathrm{U}_{[-1,1]}$.

sur $\mathbb{R}^d$, il existe dans la classe des gradients de fonctions convexes (de $\mathbb{R}^d$ vers $\mathbb{R}$) un élément (essentiellement) unique $\nabla\Psi$ tel que, ssi $\mathbf{X} \sim \mathrm{P}_1$, alors $\nabla\Psi\mathbf{X} \sim \mathrm{P}_2$, c'est-à-dire $\nabla\Psi\#\mathrm{P}_1 = \mathrm{P}_2$. Ce résultat ne fait référence à aucun problème d'optimisation, et ne requiert aucune condition sur l'existence de moments. Si, toutefois, P_1 possède des moments d'ordre 2 finis, $\nabla\Psi$ est un transport optimal au sens de Monge, Kantorovich et Brenier.

Revenons au cas univarié: les fonctions F et $\mathbf{F}_\pm$ sont monotones croissantes, et transportent la loi P vers les lois uniformes sur l'intervalle $[0,1]$ et sur la boule unité $[-1,1]$, respectivement. En vertu du résultat de McCann, F et $\mathbf{F}_\pm$ sont donc les uniques gradients de fonction convexe réalisant ce transport. Cette caractérisation de $\mathbf{F}_\pm$ (sinon celle de F elle-même) s'étend sans modification au cas de $\mathbb{R}^d$. Notons U_d la loi uniforme sur la boule unité de $\mathbb{R}^d$.[4] Nous définissons la fonction de répartition "center-outward" $\mathbf{F}_\pm$ d'une loi absolument continue P sur $\mathbb{R}^d$ (d'un vecteur aléatoire $\mathbf{X}$ de loi P) comme l'unique gradient de fonction convexe tel que

$$\mathbf{F}_\pm\#\mathrm{P} = \mathrm{U_d} \quad \text{ou, de façon équivalente,} \quad \mathbf{F}_\pm(\mathbf{X}) \sim \mathrm{U_d};$$

un résultat de Figalli (2018) établit en effet l'unicité d'une telle fonction, et montre qu'elle constitue un homéomorphisme de $\mathbb{R}^d \setminus \mathbf{F}_\pm^{-1}(\{0\})$ vers la boule unité privée de l'origine[5].

L'inverse $\mathbf{F}_\pm^{-1}$ de $\mathbf{F}_\pm$ définit une fonction quantile. L'image par $\mathbf{F}_\pm^{-1}$ des sphères emboîtées de la boule unité constitue une collection de contours quantiles continus, fermés et connexes, dont le contenu de probabilité est le rayon de la sphère correspondante: $\|\mathbf{F}_\pm(\mathbf{X})\|$ est donc le contenu de probabilité du contour quantile passant par $\mathbf{X}$—son ordre quantile.

Une version empirique $\mathbf{F}_\pm^{(n)}$ de $\mathbf{F}_\pm$ est également définie (Hallin 2018), via un couplage optimal (un transport discret) de l'échantillon avec une grille régulière de la boule unité, dont le calcul se fait aisément par programmation linéaire. Cette version empirique définit des rangs (ceux des modules $\|\mathbf{F}_\pm^{(n)}\|$) et des signes multivariés $\mathbf{F}_\pm^{(n)}/\|\mathbf{F}_\pm^{(n)}\|$ (directions dans $\mathbb{R}^d$). On peut démontrer que ces rangs et ces signes sont "distribution-free", possèdent les propriétés d'invariance maximale qui, dans le cas univarié, garantissent (Hallin et Werker 2003) la préservation de l'efficacité semiparamétrique, et que $\mathbf{F}_\pm^{(n)}$ est lié à $\mathbf{F}_\pm$ par une propriété de Glivenko-Cantelli.

Ces définitions permettent d'équiper $\mathbb{R}^d$ d'un ordre comparable à celui, fondé sur les sphères emboîtées, qui structure la boule unité. Dans le cas sphérique, il coïncide avec l'ordre directionnel sur les rayons issus du centre de la loi. Il ne s'agit toutefois plus d'un ordre canonique, mais d'un ordre spécifique à la loi considérée. De même que chaque

[4] Nous définissons la loi uniforme sur la boule unité dans $\mathbb{R}^d$ comme une uniforme sphérique, c'est-à-dire le produit d'une uniforme sur la sphère unité et d'une uniforme sur la distance à l'origine. Pour $d = 1$, uniforme sphérique et uniforme au sens de Lebesgue coïncident, mais pas pour $d \geq 2$.

[5] En outre, $\mathbf{F}_\pm^{-1}(\{0\})$ est de mesure de Hausdorff nulle dans $\mathbb{R}^d$.

loi P induit un ordre via $\mathbf{F}_{\pm}$, un ordre empirique est induit par l'échantillon via $\mathbf{F}_{\pm}^{(n)}$, qui asymptotiquement reconstruit l'ordre-population.

Ceci ouvre la voie à une multitude d'applications et, notamment, à la construction de procédures fondées sur les rangs (tests de rangs et R-estimateurs) pour l'ensemble des problèmes d'Analyse Multivariée (MANOVA, régression "multiple output", MANOCOVA, etc.) dont la validité s'étend à des lois multivariées (absolument continues) quelconques, ainsi qu'à des méthodes de régression quantile "multiple output", des concepts de composantes principales non linéaires, des détections d'outliers et des Values-at-Risk multivariées, etc.

References

[1] del Barrio, E., Cuesta-Albertos, J., Hallin, M., and Matrán, C. (2018). Smooth cyclically monotone interpolation and empirical center-outward distribution functions, available at `http://arxiv.org/abs/1806.01238`.

[2] Brenier, Y. (1987). Décomposition polaire et réarrangement monotone des champs de vecteurs, *Comptes Rendus de l'Académie des Sciences de Paris Série I Mathémathique* **305 19**, 805–808.

[3] Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions, *Communications in Pure and Applied Mathematics* **44**, 375–417.

[4] Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks, and signs, *Annals of Statistics* **45**, 223–256.

[5] Figalli, A. (2017). On the continuity of center-outward distribution and quantile functions, available at `http://arxiv.org/abs/1805.04946`.

[6] Genest, Chr. and Rivest, P.L. (2001). On the multivariate probability integral transformation, *Statistics & Probability Letters* **53**, 391–399.

[7] Hallin, M. (2018). On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$: a measure transportation approach, available at `http://arxiv.org/abs/1806.01238`.

[8] Hallin, M. and Werker, B.J.M. (2003). Semiparametric efficiency, distribution-freeness, and invariance, *Bernoulli* **9**, 137–165.

[9] McCann, R. J. (1995). Existence and uniqueness of monotone measure-preserving maps, *Duke Mathematical Journal* **80**, 309–324.