Entropie conditionnelle de Rényi et segmentation

Olivier MICHEL[1], Patrick FLANDRIN[2], Alfred HERO[3]

[1]Laboratoire d'Astrophysique, UMR 6525-CNRS
Université de Nice-Sophia Antipolis, Parc Valrose, 06108 Nice cedex 2

[2]Laboratoire de Physique, UMR 5672-CNRS
ENS-Lyon, 46 allée d'Italie, 69364 Lyon Cedex 7

[3]Dept. of EECS, University of Michigan, Ann Arbor, USA

olivier.michel@lunice.fr, patrick.flandrin@ens-lyon.fr
hero@eecs.umich.edu

Résumé – Après avoir rappelé la démarche axiomatique proposée par A. Rényi pour la définition de l'entropie, nous montrons que cette dernière s'étend facilement aux notions d'entropie conditionnelle et d'information mutuelle. L'expression de l'entropie d'une distribution conjointe en fonction des entropies conditionnelles, généralisant l'expression dite 'chain rule' pour l'entropie de Shannon est proposée. Dans la dernière partie, ces résultats sont appliqués et discutés sur un exemple de séparation de mélange, pour lequel l'estimation des entropies de Rényi est conduite à l'aide d'outils issus de la théorie des graphes.

Abstract – The Rényi axiomatic derivation of entropy is briefly introduced. The generalized mean which is the key point in this derivation is then used to extend the Shannon chain rule for conditional entropies. A similar rule is derived for Rényi entropy. This new expression is applied and discussed in the context of entropy based mixture separation; the entropies are estimated with newly introduced methods, based on properties of minimal spanning trees.

1 Introduction

Dans son article fondateur de 1961 [11], A. Rényi propose une dérivation axiomatique de l'entropie, s'appuyant très largement sur les notions de distributions incomplètes et de moyenne généralisée. Nous proposons d'exploiter cette dernière notion pour établir une loi de composition des entropies de Rényi conditionnelles, généralisant la loi de composition bien connue pour les entropies conditionnelles de Shannon. Nous montrons que la loi proposée possède la propriété d'additivité des entropies quand les lois sont indépendantes et permet de retrouver la loi de composition de Shannon comme un cas particulier. Dans une deuxième partie, après avoir fait quelques rappels sur les liens récemment établis entre la longueur de graphes acycliques minimaux (MST, pour *Minimal Spanning Trees*) et l'entropie de Rényi de la distribution des sommets du graphe, un nouvel algorithme de segmentation est proposé. Nous développons une approche utilisant à la fois la théorie des graphes de représentation minimaux pour l'estimation de l'entropie de Rényi et la nouvelle loi de composition des entropies conditionnelles ; nous montrons que le problème de segmentation peut alors être abordé comme un problème de minimisation d'entropie. Cette approche permet de donner une justification théorique à de nombreuses méthodes de segmentations ou d'identification faisant appel aux graphes, et généralement présentées comme des méthodes *ad hoc*[1][9][10]. Dans la dernière partie, l'algorithme de segmentation est discuté dans des cas où les méthodes usuelles de segmentation directe sont prises en défaut, en particulier si les composantes à séparer sont imbriquées (confusion des centres de masses des différentes composantes).

2 Axiomatique de Rényi, loi de composition des entropies conditionnelles

Nous ne rappellerons dans cet article que les deux propriétés imposées dans la démarche de Rényi. Les axiomes de symétrie, continuité et normalisation sont en effet commun aux différentes approches axiomatiques proposées (voir e.g. [5]).

– *Additivité.*
Soient x et y deux variables aléatoires discrètes définies respectivement sur les ensembles d'événements finis $\Omega_x = (x_i, i = 1, \ldots, n)$ et $\Omega_y = (y_i, i = 1, \ldots, m)$, et $P = (p_1, p_2, \ldots, p_n)$ et $Q = (q_1, q_2, \ldots, q_m)$ leur lois de probabilité respectives. Ces lois pouvent être incomplètes, i.e.

$$W_P = \sum_{i=1}^{n} p_i \leq 1$$
$$W_Q = \sum_{i=1}^{m} q_i \leq 1$$

Soit $(P \times Q)$ la loi de probabilité conjointe des variables x et y ; si ces dernières sont indépendantes, alors

$$(P \times Q) = (p_1 q_1, \ldots, p_1 q_m, p_2 q_1, \ldots, p_2 q_m, \ldots, p_n q_1, \ldots, p_n q_m) \quad (1)$$

et l'entropie H de la loi de probabilité conjointe doit vérifier

$$H(P \times Q) = H(P) + H(Q)$$

– *Moyenne.*
Pour tout couple (P, Q) de lois de probabilité, tel que

$$W_P + W_Q \leq 1$$

on introduit

$$(P, Q) \equiv (p_1, p_2, \ldots, p_n, q_1, q_2, \ldots, q_m)$$

L'entropie de la loi de probabilité (P, Q) s'exprime comme moyenne généraliée au sens de Kolmogorov-Nagumo [6] selon

$$H(P, Q) = \psi^{-1}\left(\frac{W_P \psi(H(P)) + W_Q \psi(H(Q))}{W_P + W_Q}\right) \quad (2)$$

où $\psi(.)$ est une fonction continue, strictement croissante, d'inverse $\psi(.)^{-1}$.

Il est établi que les seules solutions admissibles [11] pour $\psi(.)$ (au sens où toutes les propriétés précédentes sont vérifiées) sont :

$$\psi_{r,\alpha(x)} = 2^{(\alpha-1)x} \quad , 0 < \alpha \neq 1$$

et

$$\psi_s(x) = ax + b \quad , (a, b) \in \mathbb{R}^2$$

La fonction H_α obtenue avec la fonction $\psi_{r,\alpha}$ définit l'entropie de Rényi d'ordre α et prend la forme suivante [11] :

$$H_\alpha(P) = \frac{1}{1-\alpha} \log_2 \left(\frac{1}{W_P} \sum_{i=1}^{n} p_i^\alpha\right) \quad (3)$$

On établit par ailleurs facilement que la propriété de convexité de H_α est obtenue pour $0 < \alpha < 1$. On ne considérera plus que ce dernier cas dans la suite. L'entropie de Shannon de la loi de probabilité P, notée $H_s(P)$, s'obtient comme limite de l'entropie de Rényi d'ordre α pour $\alpha \to 1$ (voir e.g. [2]). La fonction $\psi_r(x)$ tend alors vers la fonction

$$\psi_{r,\alpha\to 1}(x) = 1 + [(1-\alpha)\log(2)]x$$

qui est de la forme $\psi_s(x)$ introduite ci-dessus. l'expression de la moyenne généralisée (2) dans laquelle $\psi = \psi_s$ conduit à l'expression de l'entropie de Shannon.

2.1 Entropies conditionnelles

Les lois P et Q ne sont pas indépendantes en général et la loi conjointe n'est pas définie par l'équation (1). D'après l'égalité de Bayes,

$$(P \times Q)_{i,j} = (Q/x_i)_j . p_i$$

La règle de composition bien connue pour les entropies conditionnelles de Shannon [3] est alors :

$$H_s(P \times Q) = \sum_i p_i H_s(Q/x_i) + H_s(P) \quad (4)$$

Le premier terme n'est autre que l'espérance mathématique de l'entropie conditionnelle $H_s(Q/x_i)$ sous la loi P, ou encore la valeur moyenne de l'entropie conditionnelle $H_s(Q/x_i)$ pondérée par les p_i.

L'extension de la notion de moyenne généralisée vue au paragraphe précédent conduit à établir le thórème suivant :

Théorème
Soient x et y deux variables aléatoires discrètes définies respectivement sur $\Omega_x = \{x_i, i = 1, \ldots, n\}$ et $\Omega_y = \{y_i, i = 1, \ldots, m\}$, et $P = (p_1, p_2, \ldots, p_n)$ et $Q = (q_1, q_2, \ldots, q_m)$ leur lois de probabilité respectives. Soit $(P \times Q)$ la loi de probabilité conjointe des variables x et y. L'entropie de Rényi d'ordre α de la loi de probabiblité conjointe s'exprime en fonction des entropies de Rényi conditionnelles selon

$$H_\alpha(P \times Q) = \psi_{r,\alpha}^{-1}\left(\frac{\sum_i p_i^\alpha \psi_{r,\alpha}(H_\alpha(Q/x_i))}{\sum_i p_i^\alpha}\right) + H_\alpha(P) \quad (5)$$

où

$$\psi_{r,\alpha}^{-1}(x) = \frac{1}{1-\alpha}\log_2(x)$$

Preuve

$$\begin{aligned} H(P \times Q) &= \\ &\frac{1}{1-\alpha}\log_2\left(\frac{\sum_i p_i^\alpha 2^{(1-\alpha)H_\alpha(Q|x_i)}}{\sum_i p_i^\alpha}\right) + H_\alpha(P) \\ &= \frac{-1}{1-\alpha}\log_2\left(\sum_i p_i^\alpha\right) + H_\alpha(P) \\ &+ \frac{1}{1-\alpha}\log_2\left(\sum_i p_i^\alpha 2^{(1-\alpha)\left(\frac{1}{1-\alpha}\log_2 \sum_j (P\times Q)_{i,j}^\alpha / p_i^\alpha\right)}\right) \\ &= \frac{1}{1-\alpha}\log_2\left(\sum_i \sum_j p_i^\alpha \frac{(P \times Q)_{i,j}^\alpha}{p_i^\alpha}\right) \\ &= \frac{1}{1-\alpha}\log_2\left(\sum_i \sum_j (P \times Q)_{i,j}^\alpha\right) \quad \text{q.e.d.} \end{aligned}$$

En conséquence, l'entropie de Rényi conditionnelle a pour expression

$$H_\alpha(Q|P) = H_\alpha(P \times Q) - H_\alpha(P) \quad (6)$$

avec

$$H_\alpha(Q|P) = \psi_{r,\alpha}^{-1}\left(\frac{\sum_i p_i^\alpha \psi_{r,\alpha}(H_\alpha(Q/x_i))}{\sum_i p_i^\alpha}\right) \quad (7)$$

Notons que l'équation (4) apparaît comme la limite lorsque $\alpha \to 1$ de l'équation (5). L'indépendance des lois P et Q dans l'équation (5) conduit naturellement à la propriété d'additivité. Ces résultats s'étendent sans difficulté aux notions d'entropie différentielle de Rényi, mettant en oeuvre non plus des lois de probabilité discrètes mais des fonctions de densité de probabilité λ.

2.2 Information mutuelle de Rényi

L'information mutuelle obtenue à partir de la notion d'entropie de Rényi d'ordre α étend naturellement la notion d'information mutuelle "classique", qui repose sur

l'entropie de Shannon. Un calcul rapide permet d'établir les égalités suivantes

$$\begin{aligned} I_\alpha(P,Q) &= H_\alpha(P) + H_\alpha(Q) - H_\alpha(P \times Q) \\ &= H_\alpha(Q) - H_\alpha(Q|P) \\ &= H_\alpha(P) - H_\alpha(P|Q) \end{aligned} \quad (8)$$

dans lesquelles $H_\alpha(.|.)$ est exprimé par l'equation (7). La limite pour $\alpha \to 1$ des équations (8) conduit aux égalités bien connues pour l'information mutuelle construite à patir de l'entropie de Shannon. L'information mutuelle de Rényi est, comme l'information mutuelle de Shannon, une quantité positive ou nulle, pour $0 < \alpha < 1$. On retrouve ici la condition de convexité de H_α.

3 MST et segmentation

3.1 MST

Soit $X = \{x_1, \ldots, x_n\}$ une réalisation de n vecteurs aléatoires indépendants et identiquement distribués, où chaque $x_i \in \mathbb{R}^d$ suit une distribution notée P, de densité de Lebesgue f. Un arbre de représentation est un graphe T non dirigé, défini par un ensemble de sommets X et un ensemble de liens $e_{i,j} = (x_i, x_j)$ connectant les sommets entre eux. La longueur totale d'ordre γ du graphe est définie par

$$L_{n,\gamma} = \sum_{e_{i,j} \in T} |e_{i,j}|^\gamma$$

Le graphe acyclique minimal de représentation (MST, pour Minimal Spanning Tree) est parmi tous les graphes totalement connectés, le graphe noté $T^\star$ dont la longueur $L_{n,\gamma}$ est minimale. $T^\star$ peut être calculé de façon exacte à l'aide d'algorithmes dont le coût varie en $n \log n$.
Dans [7][8] nous avons établi et étudié la possibilité d'estimer l'entropie de Rényi d'une distribution à l'aide de graphes de représentation minimaux, par la relation :

$$\hat{H}_\alpha(\lambda) = \frac{1}{1-\alpha} \log_2(n^{-\alpha} L^\alpha_{n,\gamma}) + \beta(\alpha, d) \quad (9)$$

où $\gamma = (1-\alpha)d$, $0 < \alpha < 1$, et donc $0 < \gamma < d$. β est une constante, dépendant de d et γ, mais non de la densité λ.

3.2 Entropie conditionnelle de Rényi et segmentation

Soit un ensemble de vecteurs aléatoires $(x_i, i = 1, \ldots, n)$ i.i.d. de densité λ de support $A_0 \subset \mathbb{R}^d$, et $\Pi_{A_0} = \{A_{01}, A_{02}\}$ une partition de A_0. Soient $\pi = (p_{A_0}, p_{A_1})$ la distribution de probabilité discrète associée aux évènements $x \in A_{01}$ ou $x \in A_{02}$ respectivement.
À partir de l'équation (5), l'expression de l'entropie de Rényi d'ordre α de la distribution conjointe $(\lambda \times \pi)$ s'exprime

$$\begin{aligned} H_\alpha(\lambda \times \pi) &= H_\alpha(\pi) + \\ &\psi^{-1}_{r,\alpha}\left(\frac{p_{A_{01}} \psi_{r,\alpha}(H_\alpha(\lambda|A_{01})) + p_{A_02} \psi_{r,\alpha}(H_\alpha(\lambda|A_0 2))}{p^\alpha_{A_{01}} + p^\alpha_{A_{02}}}\right) \\ &= H_\alpha(\pi) - \frac{1}{1-\alpha} \log_2(p^\alpha_{A_{01}} + p^\alpha_{A_{02}}) + \\ &\frac{1}{1-\alpha} \log_2\left(p^\alpha_{A_{01}} 2^{(1-\alpha)[\hat{H}_\alpha(\lambda|A_{01})]} + p^\alpha_{A_{02}} 2^{(1-\alpha)[\hat{H}_\alpha(\lambda|A_{02})]}\right) \end{aligned} \quad (10)$$

La première ligne de cette dernière expression est évidemment nulle. À partir de l'équation (9), on a

$$\begin{aligned} \hat{H}_\alpha(\lambda|A_{01}) &= \frac{1}{1-\alpha} \log_2(n^{-\alpha}_{A_{01}} L^\alpha_{n_{A_{01}},\gamma}) + \beta(\alpha, d) \\ \hat{H}_\alpha(\lambda|A_{02}) &= \frac{1}{1-\alpha} \log_2(n^{-\alpha}_{A_{02}} L^\alpha_{n_{A_{02}},\gamma}) + \beta(\alpha, d) \end{aligned} \quad (11)$$

où $n_{A_{01}}$ et $n_{A_{02}}$ sont les nombres de points de X dans A_{01} et A_{02} respectivement. D'autre part des estimateurs simples des probabilités $p_{A_{01}}$ et $p_{A_{02}}$ sont respectivement

$$p_{A_{01}} = \frac{n_{A_{01}}}{n} \quad , \quad p_{A_{02}} = \frac{n_{A_{02}}}{n} \quad (12)$$

En substituant (11) et (12) dans l'équation (10), on a après un court calcul

$$\hat{H}_\alpha(\lambda \times \pi) = \frac{1}{1-\alpha} \log_2\left(\frac{L_{n_{A_{01}},\gamma} + L_{n_{A_{02}},\gamma}}{n^\alpha}\right) + \beta(\gamma, d) \quad (13)$$

D'autre part la propriété de superadditivité [13]

$$L_{n=(n_{A_{01}}+n_{A_{02}}),\gamma} > L_{n_{A_{01}},\gamma} + L_{n_{A_{02}},\gamma}$$

conduit naturellement à

$$\hat{H}_\alpha(\lambda \times \pi) \leq \hat{H}(\lambda|A_0) = \frac{1}{1-\alpha} \log_2(n^{-\alpha}_{A_0} L^\alpha_{n_{A_0},\gamma}) + \beta(\alpha, d) \quad (14)$$

En l'absence d'a-priori sur la partition Π_{A_0}, et en observant que la suppression d'un lien dans un MST définit deux sous ensembles de X, de support disjoint, donc une partition de A_0, l'équation (14) conduit à formuler le théorème suivant :

Théorème
Soit $T^\star$ un MST défini sur une réalisation $X = \{x_1, \ldots, x_n\}$ de n vecteurs i.i.d., de densité de Lebesgues λ, et Π une partition de X. La partition qui minimise l'entropie de Rényi d'ordre α conjointe $H_\alpha(\lambda \times \Pi)$ est obtenue en suprimant le lien e_{ij} le plus long de $T^\star$.

La démonstration est une conséquence immédiate des équations (9)(11)(13) et (14).

Un algorithme simple de segmentation cherchant à minimiser l'entropie de Rényi de la distribution peut alors être proposé :

1. Estimer T^*, MST construit sur l'ensemble des n réalisations.
2. Trouver le segment de longueur maximale dans T^*, soit e_{max}.
3. Définir $T^*_{n_{A_{01}}}, T^*_{n_{A_{02}}}$ en appliquant la coupure $C(e_{max})$ sur T^*, définissant ainsi les deux sous-ensembles de réalisations et donc la partition recherchée.

Le coût de calcul de cet algorithme varie en $O(n \log n)$).

3.3 Exemples

Deux exemples sont traités, permettant d'illustrer les performances mais aussi les défauts de cette approche.

Soient deux distribution (figure 1) sur R^2 dont les supports sont disjoints mais concentriques, et de centre de masse confondus. L'identification (et la segmentation) des

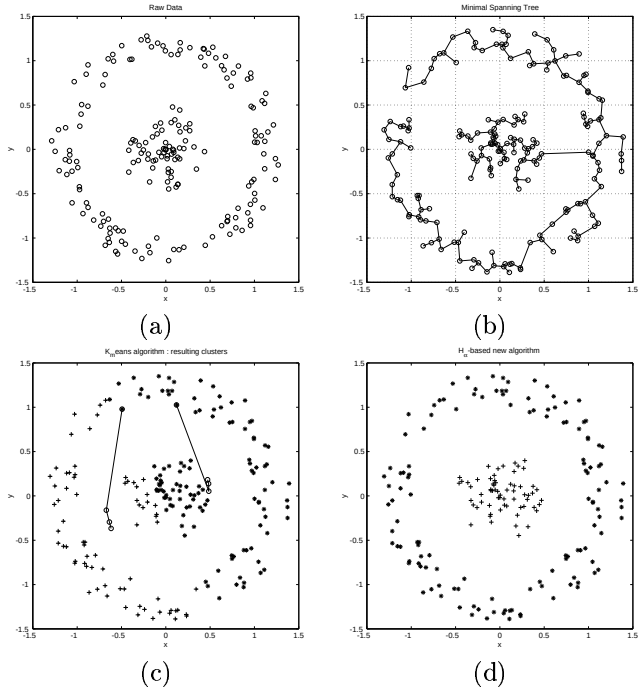

FIG. 1 – Exemple de partition en deux sous ensembles non convexes, de barycentres confondus. (a) Données ; (b) MST ; (c) Segmentation par K-Means ; (d) Segmentation par minimisation d'entropie de Rényi.

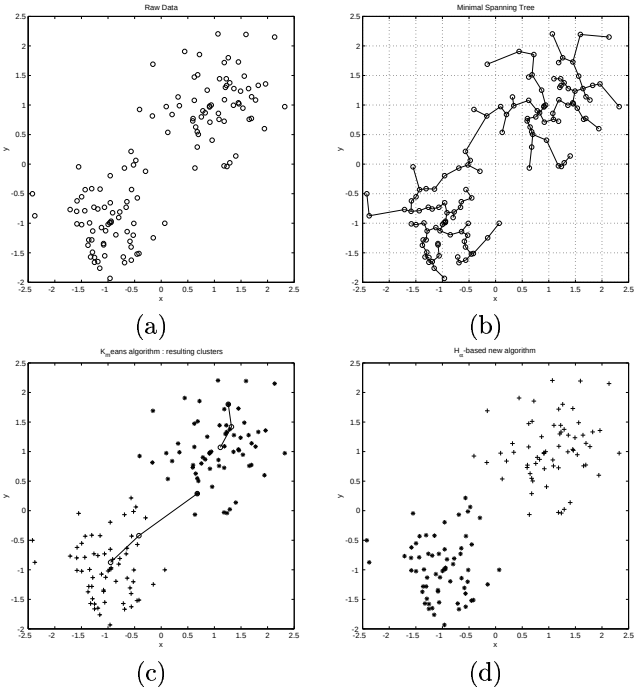

FIG. 2 – Les données contiennet deux réalisations de 64 points de distributions gaussiennes bi-dimensionnelles iid de variance .25 and moyennes (-1,-1) and (+1,+1) respectivement ; (a) données (b) MST (c) Segmentation par K-means (d) segmentation par minimisation d'entropie

deux composantes ne peut être résolu par la méthode usuelle des *K-Means* [4], et à notre connaissance, seules des méthodes reposant sur un grand nombre d'a-priori (sur la topologie des supports) ou sur la notion d'apprentissage présentent une alternative. Le second exemple (figure 2) est un exemple pour lequel la méthode des K-means est adaptée : composantes concaves et centres de masses bien distincts. Sur un tel exemple, on percoit immédiatement la limitation de l'algorithme proposé : dès que les support des deux composantes se rapprochent, il peut exister des segments de $T^\star$ plus importants à l'intérieur d'une même composante que ne l'est le segment qui relie les composantes. Une étude statistique précise des performances de notre algorithme est en cours, en fonction du nombre n de réalisations observées.

Références

[1] D. Banks, "The minimal spanning tree for nonparametric regression and structure discovery," in *Book of Abstracts of the 1996 Meeting of the Classification Society of North America*, p. 54, 1996.

[2] , M. Basseville : "Distances measures for Signal Processing and Pattern Recognition.", *Signal Processing*, vol. 18, pp.349-369, 1989.

[3] Thomas M. Cover, Joy A. Thomas : "Elements of Information Theory.", Wiley Series in Telecommunications, New York, 1991.

[4] R.O. Duda, P.E. Hart :*Pattern Classification and Scene Analysis.*Wiley, N.Y. 1973.

[5] D.K. Fadeev : "Zum Begriff der Entropie einer endlichen Wahrscheinlichkeitsschemas.", in *Arbeiten zur Informationtheorie I* pp. 85-90, Deutscher Verlag der Wissenschaften, Berlin, 1957.

[6] G.H.Hardy, J.E.Littlewood, G. Polya : *Inequalities"*, Cambridge Univ. Press, Cambridge, 1934.

[7] A.O.Hero, O.Michel : "Asymptotic theory of greedy approximations to minimal K-point random graphs.", *IEEE Trans. on Information theory*, vol.45, No.6, pp.1921-1938, septembre 1999.

[8] A.O.Hero, O.Michel :"Estimation of Rényi information divergence via pruned minimal spanning trees.", proc. of IEEE SP Workshop on higher Order statistics, Ceasarea, Israel, pp.264-268, 1999.

[9] O.Michel, A.O.Hero : "Pruned MST's for entropy estimation and outlier rejection.", IEEE-IT workshop on DECI, Detection, Classification and Imaging, Santa-Fe, NM, USA., Feb 99.

[10] O.Michel, A.O.Hero, P.Flandrin :"MST et divergences informationelles : applications.", en cours de publication *Traitement du Signal*, Mars 2000.

[11] A. Rényi : "On measures of entropy and information.", Proc 4th Berkeley Symp. Math. Stat. Proba (1960), vol. 1, pp.547-561, Univ. of Calif. Press, Berkeley, 1961.

[12] C.E. Shannon, W. Weaver : *The Mathematical Theory of Communication*, Univ. of Illinois Press, Urbana, 1949.

[13] J.E.Yukich :"Probability Theory of Classical Euclidean Optimization problems.", *Lecture Notes in mathematics*, 1675, Springer, 1998.