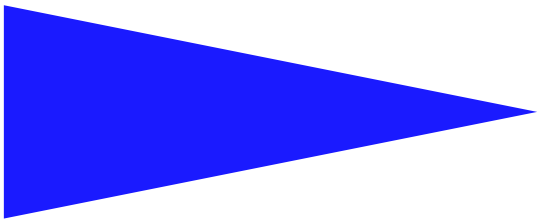
I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1020



INFORMATION: ENTROPIES, DIVERGENCES ET MOYENNES

MICHÈLE BASSEVILLE

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

Tél. : (33) 99 84 71 00 – Fax : (33) 99 84 71 71

Information: entropies, divergences et moyennes

Michèle Basseville*

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet AS

Publication interne n ° 1020 — Mai 1996 — 74 pages

Résumé : Après avoir rappelé l'importance, en traitement du signal, de concepts issus de la théorie de l'information, on répertorie des procédés constructifs entre des notions telles que entropies, divergences, moyennes, projections. On introduit deux grandes classes de divergences et on en discute l'intersection.

Mots-clé : Information, entropie, divergence, distance, métrique associée, valeur moyenne, axiomatique.

(Abstract: pto)

Ces notes de lecture résultent partiellement de discussions avec Jean-François Cardoso (Télécom. Paris/Cnrs).

* Irisa/Cnrs – basseville@irisa.fr.

Centre National de la Recherche Scientifique
(ura 227) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Information: entropies, divergences and mean values

Abstract: The design concepts of divergences are of interest because of the key role they play in statistical inference and signal processing. We distinguish two different classes of divergences built on entropies. They are compared using the associated quadratic differential metrics, mean values and projections.

Key-words: Information, entropy, divergence, distance, associated metric, mean value, axiomatic.

Notations et opérateurs

Les opérateurs concernant les fonctions h (ou f), ψ, ϕ qui interviennent dans les différentes constructions (sections 2.1 et 2.2), les équations fonctionnelles caractéristiques (chapitre 3), et les propriétés des entropies et divergences, sont [154] :

- *conjugaison* : $u \to 1-u$,
- *translation* : $u \to u+\delta$,
- *dilatation* : $u \to \beta u$,
- **fonction dilatée** :
$$\underline{f}(u) = \gamma\, f(u) + \delta \tag{0.1}$$
induisant une propriété d'invariance des entropies non intégrales (section 2.4),
- **fonction translatée** :
$$\check{f}(u) = f(u) + \gamma\, u + \delta \tag{0.2}$$
induisant une propriété d'invariance des f-divergences intégrales (section 2.5),
- **fonction miroir** :
$$\breve{f}(u) = u\, f\left(\frac{1}{u}\right) \tag{0.3}$$
intervenant dans la propriété de symétrie des f-divergences intégrales (section 2.5); à noter aussi que :
$$\phi(u) = -\frac{h(u)}{u} = -\breve{h}\left(\frac{1}{u}\right)$$
intervient dans les caractérisations fonctionnelles des entropies (chapitre 3), et que ses propriétés influent sur les algorithmes de reconstruction;
- **fonctions d'information** [3, 109] :
$$\bar{h}(u) = -h(u) - h(1-u) \tag{0.4}$$
et :
$$\bar{f}(u,v) = v\, f\left(\frac{u}{v}\right) + (1-v)\, f\left(\frac{1-u}{1-v}\right) \tag{0.5}$$
vérifiant des équations fonctionnelles caractéristiques [3, 109] (chapitre 3),
- **autres fonctions** [105, 154] :
$$\begin{aligned} a(u) &= (2-u)\, f\left(\frac{u}{2-u}\right) \\ b(u) &= \frac{1}{2}\,(2-u)\, f\left(\frac{2+u}{2-u}\right) \end{aligned}$$
- **différence de Jensen** :
$$\mathbf{J}_h^{\beta}(u,v) = h(\beta u + (1-\beta)v) - \beta h(u) - (1-\beta)h(v)$$
intervenant dans la définition de l'information mutuelle (section 2.3), un procédé constructif de divergences (section 2.2) et ses extensions (section 4.2), et les équations fonctionnelles (chapitre 3).

– **différentielle de Gâteaux** (ou dérivée directionnelle) :

$$\delta\mathbf{H}(P:R) = \frac{d}{dt}\mathbf{H}(P+tR)\bigg|_{t=0} = \lim_{t\to 0}\ \frac{\mathbf{H}(P+tR)-\mathbf{H}(P)}{t} = \mathbf{H}'(P)\cdot R \tag{0.6}$$

intervenant dans la définition de la distance de Bregman (section 2.2),

Transformée de Legendre-Fenchel

Pour toute fonction $f : \mathbb{R}^m \to]-\infty, +\infty[$, on définit la transformée de Lengendre (ou Fenchel) [138, 23] par :

$$f^*(u) = \sup_{t\in\mathbb{R}^m} \left(< t, u > -f(t)\right) \tag{0.7}$$

f^* est convexe sur $\mathbb{R}^m$, et est dite aussi duale convexe de f.
Si f est convexe et semi-continue inférieurement, alors $f^{**} = f$.
Si f est différentiable, on a de plus :

$$f^*(u) =< f'^{-1}(u), u > -f(f'^{-1}(u))$$

$$f^{*\prime} = f'^{-1}$$

$$f(u) + f^*(f'(u)) - < u, f'(u) >= 0 \tag{0.8}$$

Chapitre 1

Introduction

On donne, en guise de motivation, quelques exemples de concepts de la théorie de l'information intervenant dans la formalisation et la résolution de problèmes de traitement de signal. Puis on indique comment le document est organisé.

1.1 Motivations: information et signal

Citons, en vrac et de manière non exhaustive, quelques problèmes où interviennent de manière cruciale des notions d'information.

Les statistiques exhaustives (qui conservent l'information, de Fisher, mais aussi bien d'autres) jouent un rôle-clef en estimation et en détection (tests d'hypothèses), et il en est de même de la question des distances (ou divergences). Il serait d'ailleurs intéressant de mettre en évidence de manière explicite les propriétés de ces notions qui sont propres au cas Gaussien, propres aux paramètres de translation, ou générales. L'information de Kullback (ou entropie relative) est sous-jacente à l'identification par maximum de vraisemblance, à des critères de sélection de modèles, à la reconstruction par maximum d'entropie. L'information de Fisher fournit la précision optimale de tout estimateur. L'information (ou distance) de Kullback intervient dans les probabilités d'erreur des détecteurs par rapport de vraisemblance, et aussi dans le retard à la détection optimal en détection séquentielle de ruptures; pour cette raison, elle est utilisée pour définir la détectabilité d'un changement de modèle.

L'entropie joue par ailleurs un rôle-clef dans des méthodes d'analyse spectrale. Elle intervient dans la résolution de problèmes inverses, et l'information est omni-présente dans les techniques bayésiennes de régularisation. Évidemment la quantification de l'information est au coeur des problèmes de compression et aussi de fusion de données.

1.2 Organisation du document

On examine les deux principales notions d'*information* concernant des lois de probabilité :

1. l'*entropie* ou mesure d'information, notée **H**,
2. la *divergence* ou gain d'information, notée **I** ou **D**,

ainsi que les liens qui les unissent. On étudie en particulier les relations entre plusieurs procédés constructifs de divergences à partir d'entropies (et réciproquement). On décrit aussi l'*information mutuelle*.

La plupart des concepts décrits valent dans un contexte *non-paramétrique*. Leur application à des lois paramétrées donne lieu à des objets (e.g. distances entre vecteurs de paramètres) et propriétés (e.g. la métrique et son éventuelle invariance par changement de paramétrisation) supplémentaires.

Le chapitre 2 est consacré à l'introduction des deux classes de divergences. On commence par distinguer, à la section 2.1, deux formes fonctionnelles d'entropies, intégrales et non intégrales, et une classe de divergences, dites f-divergences, notées **I**, dont on peut dire qu'elles sont formellement identiques aux entropies (entropies relatives). De nombreux exemples d'entropies et de f-divergences sont décrits en détail dans les sections 2.4 et 2.5, ainsi que le lien crucial avec l'information de Fisher et l'exhaustivité.

Ensuite, à la section 2.2, on décrit les $\mathbf{J}, \mathbf{C}, \mathbf{D}, \mathbf{K}, \mathbf{L}$-divergences qui se déduisent d'entropies à partir de procédés constructifs distincts. Ce sont la $\mathbf{J}$-divergence basée sur la différence de Jensen, la distance à la Chernoff $\mathbf{C}$ basée sur une maximisation de celle-ci, la distance de Bregman $\mathbf{D}$ basée sur la différentielle de Gâteaux, et les $\mathbf{K}$ et $\mathbf{L}$-divergences basées sur des analogies avec des expressions alternatives de l'information de Kullback. Puis, à la section 2.3, on s'intéresse à l'information mutuelle.

Les caractérisations axiomatiques et les équations fonctionnelles liées à toutes ces notions sont discutées dans le chapitre 3.

Le chapitre 4 discute de notions de moyennes (ou mélanges) et d'extensions possibles de divergences. À la section 4.1, on s'intéresse aux moyennes sous-jacentes explicitement aux entropies et f-divergences, et aux moyennes définies comme des projections au sens de distances telles que f-divergences et distances de Bregman. Ensuite, on utilise ces moyennes généralisées pour construire, à la section 4.2, des extensions de la différence de Jensen et de la distance de Bregman. Ce jeu permet permet d'introduire le rayon d'information et une notion de capacité.

Le chapitre 5 contient des aspects géométriques. Les propriétés géométriques des entropies et des divergences, au sens de la géométrie différentielle de l'espace des lois, sont principalement la métrique différentielle quadratique qui peut leur être associée, ses éventuelles propriétés d'invariance, et donc le lien des entropies et divergences avec l'information de Fisher. Elles sont discutées à la section 5.1. Avec les moyennes et projections de la section 4.1, elles sont un des outils pour étudier l'intersection – non vide – des différentes classes de divergences, ainsi qu'on le montre dans la section 5.2.

Enfin, de nombreux exemples d'entropies et de divergences concernant les familles exponentielles et les cas Gaussiens (variables scalaires ou vectorielles et processus) sont donnés en annexe.

Chapitre 2

Entropies et construction de deux classes de divergences

On définit deux classes de divergences, toutes deux construites à partir de fonctionnelles d'entropies, bien qu'à partir de propriétés différentes. La première classe contient les divergences définies comme étant des entropies relatives. La deuxième contient des distances dont la construction repose sur la concavité de la fonctionnelle d'entropie considérée.

2.1 Entropies et f-divergences (Classe 1)

On introduit plusieurs types d'entropies et les f-divergences (ou entropies relatives) qui leur sont associées.

2.1.1 Entropies

On considère deux formes fonctionnelles d'entropies, intégrales et non intégrales, et on donne trois exemples.

Entropies intégrales et non intégrales Remarquons d'emblée que les entropies sont toujours définies d'une manière *relative à une mesure de référence* ν. On considère les formes fonctionnelles intégrales et non intégrales suivantes [3].

Les entropies intégrales sont définies par[1] :

$$\mathbf{H}_{h,\nu}(\mu) = -\int h\left(\frac{d\mu}{d\nu}\right)\, d\nu \tag{2.1}$$

$$\mathbf{H}_{g,\nu}(\mu) = \int \frac{d\mu}{d\nu}\, g\left(-\ln \frac{d\mu}{d\nu}\right)\, d\nu \tag{2.2}$$

où h est le plus souvent convexe, et h et g sont liées par :

$$\begin{aligned} h(u) &= -u\, g(-\ln u) \\ g(u) &= -e^u\, h(e^{-u}) \\ &= -\breve{h}(e^u) \end{aligned}$$

où $\breve{h}(u) = u\, h\left(\frac{1}{u}\right)$ est la fonction miroir de h.

1. Noter que [141] considère une forme intégrale différente:

$$\mathbf{H}_{S,\nu}(\mu) = -\int \frac{d\mu}{d\nu}\, S\left(\ln \frac{d\mu}{d\nu}\right)\, d\nu$$

où S est un opérateur linéaire, par exemple l'identité pour Shannon, et la dérivée seconde pour Fisher. Noter aussi la définition de la concentration d'une mesure [79] :

$$q_{h,\nu}(\mu) = \int h\left(|\psi_\mu(t)|\right)\, d\nu(t)$$

où ψ_μ est la fonction caractéristique de μ.

Les entropies non intégrales sont définies par :

$$\mathbf{H}_{\phi,\nu}(\mu) = -\ln \phi^{-1}\left(\int \frac{d\mu}{d\nu}\,\phi\left(\frac{d\mu}{d\nu}\right)\,d\nu\right) = -\ln \mathbf{G}_{\phi,\nu}(\mu) \tag{2.3}$$

$$\mathbf{H}_{\psi,\nu}(\mu) = \psi^{-1}\left(\int \frac{d\mu}{d\nu}\,\psi\left(-\ln\frac{d\mu}{d\nu}\right)\,d\nu\right) \tag{2.4}$$

où ϕ et ψ sont strictement monotones, et liées par :

$$\begin{aligned}\phi(u) &= \psi(-\ln u)\\ \psi(u) &= \phi(e^{-u})\\ \psi^{-1}(u) &= -\ln\ \phi^{-1}(u)\end{aligned}$$

Le passage d'une forme intégrale à une forme non intégrale peut être effectué à l'aide de :

$$\begin{aligned}\phi(u) &= -\frac{h(u)}{u}\\ &= -\breve{h}\left(\frac{1}{u}\right)\\ \psi(u) &= g(u)\end{aligned}$$

modulo les propriétés d'invariance des formes fonctionnelles d'entropies par transformation des fonctions f, g, ϕ, ψ évoquées plus loin. On en montre un exemple pour les entropies d'ordre α pour lesquelles ces relations sont différentes.

Exemples Les quatre formes fonctionnelles valent en particulier pour l'*entropie de Shannon* :

$$\mathbf{H}_{1,\nu}(\mu) = -\int \frac{d\mu}{d\nu}\,\ln\frac{d\mu}{d\nu}\,d\nu \tag{2.5}$$

qui correspond à :

$$\begin{aligned}h(u) = h_1(u) &= u\,\ln u\\ g(u) = g_1(u) &= u\\ \psi(u) = \psi_1(u) &= u\\ \phi(u) = \phi_1(u) &= -\ln u = \breve{h}_1(u) = -\breve{h}_1\left(\frac{1}{u}\right)\end{aligned}$$

L'exemple le plus connu d'entropie intégrale est l'*entropie d'ordre α de Havrda-Charvát* :

$$\mathbf{H}_{\alpha,\nu}(\mu) = \frac{1}{\alpha-1}\,\left(1 - \int\left(\frac{d\mu(x)}{d\nu(x)}\right)^{\alpha}\,d\nu(x)\right) \tag{2.6}$$

($\alpha \neq 1$) qui correspond à :

$$h(u) = h_\alpha(u) = \frac{1}{\alpha-1}\,(u^\alpha - u) \tag{2.7}$$

$$g(u) = g_\alpha(u) = \frac{1}{\alpha-1}\,\left(1 - e^{-(\alpha-1)u}\right) \tag{2.8}$$

Elle admet comme cas particulier l'entropie de Shannon :

$$\lim_{\alpha\to 1} \mathbf{H}_{\alpha,\nu}(\mu) = \mathbf{H}_{1,\nu}(\mu)$$

Pour $\alpha = 2$, il s'agit de l'indice de diversité de Gini [122].

L'exemple le plus connu d'entropie non intégrale est l'*entropie d'ordre α de Rényi* :

$$\widetilde{\mathbf{H}}_{\alpha,\nu}(\mu) = -\frac{1}{\alpha-1}\,\ln\int\left(\frac{d\mu(x)}{d\nu(x)}\right)^{\alpha}\,d\nu(x) \tag{2.9}$$

TAB. 2.1 – *Les trois fonctions h, ψ, ϕ et les entropies d'ordre α.*

Cas général	Entropies d'ordre α
$\phi(u) = \psi(-\ln u)$	$\phi_\alpha(u) = \psi_\alpha(-\ln u)$
$\phi(u) = -\frac{h(u)}{u}$ $= -\breve{h}\left(\frac{1}{u}\right)$ $\phi_1(u) = -h_1'(u) + 1$	$\phi_\alpha(u) = -(\alpha-1)\ \left(-\frac{h_\alpha(u)}{u}\right) + 1$ $= (\alpha-1)\ \breve{h}_\alpha\left(\frac{1}{u}\right) + 1$ $\phi_\alpha(u) = \frac{1}{\alpha}\ [(\alpha-1)\ h_\alpha'(u) + 1]$
$\psi(u) = -e^u h(e^{-u})$ $= -\breve{h}(e^u)$	$\psi_\alpha(u) = -(\alpha-1)\ (-e^u h_\alpha(e^{-u})) + 1$ $= (\alpha-1)\ \breve{h}_\alpha(e^u) + 1$
$h(u) = -u\ \phi(u)$ $h(u) = -u\ \psi(-\ln u)$	$h_\alpha(u) = -\frac{1}{\alpha-1}\ (-u\ \phi_\alpha(u)) - \frac{u}{\alpha-1}$ $h_\alpha(u) = -\frac{1}{\alpha-1}\ (-u\ \psi_\alpha(-\ln u)) - \frac{u}{\alpha-1}$

($\alpha \neq 1$) qui correspond à :

$$\psi(u) = \psi_\alpha(u) \quad = \quad e^{-(\alpha-1)u} \tag{2.10}$$

$$\phi(u) = \phi_\alpha(u) \quad = \quad u^{\alpha-1} \tag{2.11}$$

Elle admet comme cas particulier l'entropie de Shannon :

$$\lim_{\alpha \to 1} \widetilde{\mathbf{H}}_{\alpha,\nu}(\mu) = \mathbf{H}_{1,\nu}(\mu)$$

Son importance, en particulier pour le codage, est résumée par la propriété :

$$-\alpha\ \widetilde{\mathbf{H}}_{\alpha+1,\nu}(P) = \ln\ \mathbf{E}_\nu\left(e^{\alpha \ln P(X)}\right)$$

En d'autres termes, l'entropie de Rényi n'est autre que le logarithme de la fonction caractéristique (ou fonction génératrice des moments m.g.f.) ou encore la fonction génératrice des cumulants de $\ln P(X)$.

Le tableau 2.1 met en évidence les relations entre les formes d'entropies d'ordre α.

On donne d'autres exemples d'entropies à la section 2.4.

2.1.2 f-divergences

Les f-divergences ont été proposées vers 1965 indépendamment par Ali-Silvey [11] et Csiszár [51], puis redécouvertes par Akaïke [8] et Zakaï-Ziv [165] dans la décennie suivante. Les f-divergences entre une mesure ν et une mesure μ, que l'on peut introduire pour chaque forme fonctionnelle d'entropie (2.1)-(2.3), sont définies [11, 51, 6, 165, 53, 117] par :

$$\mathbf{I}(\mu, \nu) \triangleq -\mathbf{H}_\nu(\mu) \tag{2.12}$$

On notera $\mathbf{I}_h$ (ou $\mathbf{I}_f$), $\mathbf{I}_\psi$, $\mathbf{I}_\phi$ les divergences ainsi associées aux entropies $\mathbf{H}_h$, $\mathbf{H}_\psi$, $\mathbf{H}_\phi$. Pour certaines divergences d'ordre α, cette définition vaut à un coefficient multiplicatif près. On en décrit à la section 2.5 de très nombreux exemples, recouvrant la plupart des distances utilisées. On peut d'ailleurs s'étonner, au vu de l'identité formelle (2.12) entre entropies et f-divergences, qu'il y ait eu dans la littérature si peu d'entropies introduites en regard du nombre de f-divergences.

Dans le cas de lois admettant une densité, on peut évidemment aussi écrire une f-divergence entre deux lois P et Q en considérant leurs densités p et q par rapport à une même mesure de référence λ. On verra plus loin que les f-divergences, qui sont homogènes de degré 1, ne dépendent pas du choix de cette mesure de référence λ, contrairement aux entropies dont elles découlent. On verra aussi que, dans le cas paramétrique, les f-divergences sont étroitement liées à l'information de Fisher. Enfin, on montrera que la dualité convexe joue un rôle à deux niveaux pour les f-divergences intégrales : une propriété de dualité sur f caractérisant des f-divergences particulières, une propriété de dualité sur $\mathbf{I}_f$ en tant que fonction sur l'espace des mesures.

f-divergences intégrales Une f-divergence intégrale, associée à une entropie intégrale $\mathbf{H}_h$, s'écrit :

$$\mathbf{I}_h(\mu,\nu) = \int h\left(\frac{d\mu}{d\nu}\right)\ d\nu$$

La forme usuelle d'une f-divergence [52] :

$$\mathbf{I}_f(P,Q) = \int f\left(\frac{p}{q}\right)\ q\ d\lambda = \mathbf{E}_Q\left(f\left(\frac{p}{q}\right)\right) \tag{2.13}$$

correspond donc bien à :

$$f(u) = h(u)$$

On impose en général la condition $f(1) = 0$ pour assurer que $\mathbf{I}_f(P,P) = 0$.

L'exemple le plus connu de f-divergence intégrale est l'*information de Kullback* ou *entropie relative*, qui se déduit de l'entropie de Shannon (2.5) par la définition (2.12) :

$$\bar{\mathbf{K}}(\mu,\nu) = \int \frac{d\mu}{d\nu}\ \ln\frac{d\mu}{d\nu}\ d\nu \tag{2.14}$$

et qui correspond à :

$$h(u) = h_1(u) = u\ \ln u$$

En terme des densités :

$$\bar{\mathbf{K}}(P,Q) = \int p\ \ln\frac{p}{q}\ d\lambda$$

Un autre exemple connu est la $\mathbf{I}_\alpha$-divergence de Csiszár, appelée aussi χ^2-divergence d'ordre α, définie par [105, 154, 117] :

$$\mathbf{R}_\alpha(P,Q) = \frac{1}{\alpha(\alpha-1)}\int \left(p^\alpha q^{1-\alpha} - \alpha p - (1-\alpha)q\right)\ d\lambda \tag{2.15}$$

$$= \frac{1}{\alpha(\alpha-1)}\int \left(p^\alpha q^{1-\alpha} - 1\right)\ d\lambda \tag{2.16}$$

et qui correspond à :

$$f(u) = r_\alpha(u) = \begin{cases} \frac{1}{\alpha(\alpha-1)}\left(u^\alpha - \alpha u + \alpha - 1\right) & \alpha \neq 0,1 \\ -\ln u + u - 1 & \alpha = 0 \\ u\ln u - u + 1 & \alpha = 1 \end{cases} \tag{2.17}$$

Du constat de ce que :

$$r_\alpha(u) = \frac{1}{\alpha}(h_\alpha(u) - u + 1) \quad (\alpha \neq 0)$$

on déduit que l'on pourrait évidemment appeler aussi cette divergence : information d'ordre α de Havrda-Charvát par la correspondance (2.12) avec l'entropie (2.6).

f-divergences non intégrales De même, une f-divergence non intégrale, associée à une entropie non intégrale $\mathbf{H}_\psi$, s'écrit :

$$\mathbf{I}_\psi(\mu,\nu) = -\psi^{-1}\left(\int \frac{d\mu}{d\nu}\ \psi\left(-\ln\frac{d\mu}{d\nu}\right)\ d\nu\right)$$

$$\mathbf{I}_\psi(P,Q) = -\psi^{-1}\left(\mathbf{E}_Q\left(\frac{p}{q}\ \psi\left(-\ln\frac{p}{q}\right)\right)\right)$$

L'exemple le plus connu de f-divergence non intégrale est l'*information de Rényi*, qui se déduit de l'entropie de Rényi (2.9) par la définition (2.12) :

$$\widetilde{\mathbf{R}}_\alpha(\mu,\nu) = \frac{1}{\alpha-1}\ \ln \int \left(\frac{d\mu(x)}{d\nu(x)}\right)^\alpha\ d\nu(x) \tag{2.18}$$

et qui correspond à :

$$\psi(u) = \psi_\alpha(u) = e^{-(\alpha-1)u}$$

et :

$$\phi(u) = \phi_\alpha(u) = u^{\alpha-1}$$

En termes des densités, l'information de Rényi s'écrit [2] :

$$\widetilde{\mathbf{R}}_\alpha(P,Q) = \frac{1}{\alpha-1}\ \ln \int p^\alpha(x)\ q^{1-\alpha}(x)\ d\lambda(x)$$

Dans certaines références, et exceptionnellement dans la suite, l'information de Rényi est définie avec un facteur multiplicatif $1/\alpha$ supplémentaire.

On notera [52, 56] qu'elle est liée à la fonction caractéristique (fonction génératrice des moments) du logarithme du rapport de vraisemblance $\ln \frac{P(X)}{Q(X)}$ par :

$$\mathbf{E}_P\left(e^{\alpha \ln \frac{P(X)}{Q(X)}}\right) = e^{\alpha\ \widetilde{\mathbf{R}}_{1+\alpha}(P,Q)} \tag{2.19}$$

On en déduit que :

$$\ln\ \mathbf{E}_P\left(e^{\alpha \ln \frac{P(X)}{Q(X)}}\right) = \alpha\ \widetilde{\mathbf{R}}_{1+\alpha}(P,Q)$$

autrement dit que l'information de Rényi n'est autre, à une constante près, que la seconde fonction caractéristique, ou fonction génératrice des cumulants, du logarithme du rapport de vraisemblance $\ln \frac{P(X)}{Q(X)}$. Ceci laisse entrevoir son rôle en test d'hypothèses et classification pour le calcul de probabilités d'erreur [28, 155].

De nombreux autres exemples de telles f-divergences intégrales et non intégrales sont discutés en détail à la section 2.5.

2.2 Autres divergences associées à une entropie (Classe 2)

Suivant [41], on définit maintenant des procédés constructifs de divergences à partir d'une entropie $\mathbf{H}$, distincts de (2.12). On définit ainsi quatre divergences *entre lois*: $\mathbf{D}, \mathbf{J}, \mathbf{K}, \mathbf{L}$. Les divergences $\mathbf{D}$ et $\mathbf{J}$ sont définies à partir de la différentielle de Gâteaux et de la différence de Jensen respectivement. Elles ne requièrent pas que l'entropie $\mathbf{H}$ soit de la forme intégrale (2.1). Les divergences $\mathbf{K}$ et $\mathbf{L}$, définies par analogie avec des expressions alternatives de Kullback, ne valent que pour des entropies intégrales.

L'identité de $\mathbf{D}$ et $\mathbf{J}$ est une caractérisation de l'entropie quadratique [125]. De plus, on montre à la section 5.2, en utilisant les métriques associées, que certaines divergences peuvent être caractérisées par l'identité de ce type de construction avec les f-divergences, comme indiqué au tableau 5.1. En outre, comme déjà indiqué à la section 2.4, l'identité de la f-divergence $\mathbf{I}_f$ et de la $\mathbf{L}$-divergence permet de construire de nouvelles entropies, par la solution (5.27) de (5.24).

Remarquons pour commencer que la différence des entropies :

$$\mathbf{H}(Q) - \mathbf{H}(P)$$

ne définit pas un gain d'information (ou divergence) satisfaisant, pour la simple raison qu'elle n'est pas toujours positive.

2.2.1 Différentielle de Gâteaux et différence de Jensen

On définit d'abord la différentielle de Gâteaux et la différence de Jensen nécessaires à la définition des divergences $\mathbf{D}$ et $\mathbf{J}$.

2. La quantité $\int p^\alpha(x)\ q^{1-\alpha}(x)\ d\lambda(x)$ est appelée *intégrale de Hellinger*. Cette intégrale, et donc l'information de Rényi, s'évalue aisément dans le cas de martingales (processus et champs aléatoires) [155].

2.2.1.1 Différentielle de Gâteaux

La différentielle de Gâteaux est définie par:

$$\delta\mathbf{H}(P:R) = \left.\frac{d}{dt}\mathbf{H}(P+tR)\right|_{t=0} = \lim_{t\to 0}\ \frac{\mathbf{H}(P+tR)-\mathbf{H}(P)}{t} = \mathbf{H}'(P)\cdot R \tag{2.20}$$

Dans le cas d'une fonctionnelle d'entropie intégrale (2.40):

$$\delta\mathbf{H}_h(P:R) = -\int h'(p(x))\ r(x)\ d\lambda(x) \tag{2.21}$$

2.2.1.2 Différence de Jensen

Soit $0 \leq \beta \leq 1$:

$$\mathbf{J}_H^{(\beta)}(P,Q) = \mathbf{H}(\beta P + (1-\beta)Q) - \beta\ \mathbf{H}(P) - (1-\beta)\ \mathbf{H}(Q) \tag{2.22}$$

Dans le cas d'une fonctionnelle d'entropie intégrale, à cause de (2.41), on a:

$$\text{pour } \check{h}(u) = \gamma\ h(u) + \delta\ u + \epsilon, \qquad \mathbf{J}_{\check{h}}^{(\beta)}(P,Q) = \gamma\ \mathbf{J}_h^{(\beta)}(P,Q) \tag{2.23}$$

2.2.2 Divergences associées à des entropies non nécessairement intégrales

On définit maintenant la distance de Bregman et la **J**-divergence, qui sont définies pour toutes les formes d'entropies, intégrales ou non. On montre le rôle joué par l'opérateur de moyenne arithmétique dans le lien entre ces deux divergences.

2.2.2.1 Distance de Bregman ou divergence dirigée

Cette classe de distances a été introduite par Bregman pour la programmation convexe en 1967 [38].

Définition Pour une fonctionnelle d'entropie suffisamment régulière, la distance de Bregman ou divergence dirigée est définie par [125] :

$$\mathbf{D}_H(P,Q) = \mathbf{H}(Q) - \mathbf{H}(P) + \delta\mathbf{H}(Q:P-Q) \tag{2.24}$$

où $\delta\mathbf{H}(Q:P-Q)$ est définie en (2.20) et (2.21).

Remarquer que cette divergence est identique à la α-divergence d'Amari [13].

Noter que $\mathbf{D}_H(P,Q)$ n'est pas nécessairement symétrique[3].

Dans le cas d'une entropie intégrale, elle est appelée divergence dirigée par Rao; voir plus loin.

Tout comme l'entropie, la distance de Bregman dépend de la mesure de référence, et sera notée $\mathbf{D}_{H,\lambda}(P,Q)$ si besoin est.

Pour l'entropie de Shannon $\mathbf{H} = \mathbf{H}_1$, on retrouve l'information de Kullback :

$$\mathbf{D}_{H_1}(P,Q) = \bar{\mathbf{K}}(P,Q) \tag{2.25}$$

Propriétés

Bregman comme dérivée à l'origine de Jensen

$$\mathbf{D}_H(P,Q) = \left.\frac{\partial}{\partial\beta}\ \mathbf{J}_H^{(\beta)}(P,Q)\right|_{\beta=0} \tag{2.26}$$

Jensen comme moyenne de Bregman [125] Lorsque **H** est strictement concave, du fait de la linéarité de l'opérateur différentiel de Gâteaux, la différence de Jensen s'écrit comme la moyenne des distances de Bregman entre chaque loi et le barycentre des lois :

$$\mathbf{J}_H^{(\beta)}(P,Q) = \beta\ \mathbf{D}_H(P,\beta P+(1-\beta)Q) + (1-\beta)\ \mathbf{D}_H(Q,\beta P+(1-\beta)Q) \tag{2.27}$$

Il est important de souligner que le barycentre des lois qui intervient ici est celui qui correspond à la moyenne *arithmétique*, et pas à la moyenne Σ_h ou Σ_ϕ sous-jacente à l'entropie **H**.

3. En fait, la symétrie de $\mathbf{D}_H$ caractérise les entropies quadratiques [125].

Autres relations [125]

$$\begin{aligned}\frac{\partial}{\partial\beta}\,\mathbf{J}_H^{(\beta)}(P,Q) &= \mathbf{D}_H((1+\beta)P-\beta Q,\beta P+(1-\beta)Q)\\ &\quad -\mathbf{H}(\beta P+(1-\beta)Q)+\mathbf{H}((1+\beta)P-\beta Q)+\mathbf{H}(Q)-\mathbf{H}(P)\\ \frac{\partial}{\partial\beta}\,\mathbf{J}_{H_1}^{(\beta)}(P,Q) &= \bar{\mathbf{K}}(P,\beta P+(1-\beta)Q)-\bar{\mathbf{K}}(Q,\beta P+(1-\beta)Q)\end{aligned}$$

2.2.2.2 J-divergence de Rao

D'usage courant en anthropologie, génétique, biologie, elle est définie par Rao [122] :

$$\mathbf{J}_H(P,Q)=\mathbf{J}_H^{(1/2)}(P,Q)=\mathbf{H}\left(\frac{P+Q}{2}\right)-\frac{1}{2}\,\mathbf{H}(P)-\frac{1}{2}\,\mathbf{H}(Q)$$

Elle est appelée aussi *rayon d'information* [146, 56], car elle est égale à ce rayon d'information dans le cas particulier où l'on considère $n=2$ lois; et ceci en donne une autre justification, puisque le rayon d'information n'est autre que l'exposant de Gallager qui joue un rôle important en codage. Le rayon d'information est défini plus loin lorsque l'on discute les extensions de cette **J**-divergence.

2.2.3 Divergences associées à des entropies intégrales

2.2.3.1 Distance de Bregman ou distorsion projective

La distorsion projective est définie en [84] par :

$$\mathbf{D}_h(P,Q)=\int d_h(p,q)\;d\lambda(x)$$

où:

$$d_h(p,q)=h(p)-h(q)-h'(q)\;(p-q)$$

et donc, d'après (2.21):

$$\mathbf{D}_h(P,Q)=\mathbf{H}_h(Q)-\mathbf{H}_h(P)+\delta\mathbf{H}_h(Q:P-Q) \tag{2.28}$$

où $\mathbf{H}_h$ est définie en (2.40). Elle est identique à la distance de Bregman $\mathbf{D}_H$ (2.24).

Elle est invariante par transformation linéaire ;

$$\mathbf{D}_{\gamma h(u)+\delta u+\epsilon}=\gamma\;\mathbf{D}_h$$

Remarque $\mathbf{D}=\mathbf{D}_h$ satisfait l'égalité triangulaire:

$$\begin{aligned}\mathbf{D}(R,P) &= \mathbf{D}(R,\bar{P})+\mathbf{D}(\bar{P},P)\quad(\forall P\in\mathcal{P})\\ \text{où }\;\bar{P} &= \arg\min_{P\in\mathcal{P}}\;\mathbf{D}(R,P)\end{aligned}$$

Ceci est faux [84] pour les distances **D** de la forme f-divergence non intégrale (2.88) dans le cas où $g(u)\neq u$, et donc en particulier pour l'information de Rényi (2.90), la distance de Bhattacharrya ou celle de Matusita.

2.2.3.2 J, K, L-divergences de Rao

Elles ont été introduites par Rao [41] :

- **J-divergence** : elle s'écrit maintenant :

$$\mathbf{J}_h(P,Q)=\int\left(\frac{1}{2}\,h(p)+\frac{1}{2}\,h(q)-h\left(\frac{p+q}{2}\right)\right)\;d\lambda(x) \tag{2.29}$$

- **K-divergence** : elle est définie par analogie avec (2.70) où $h(u)=u\ln u$:

$$\mathbf{K}_h(P,Q)=\int(p-q)\left(\frac{h(p)}{p}-\frac{h(q)}{q}\right)\;d\lambda(x) \tag{2.30}$$

– **L-divergence**: elle est définie par analogie avec (2.69) où $h(u) = u \ln u$:

$$\mathbf{L}_h(P,Q) = \int \left(p\ h\left(\frac{q}{p}\right) + q\ h\left(\frac{p}{q}\right)\right)\ d\lambda(x) \tag{2.31}$$

La notation n'est pas complètement stabilisée puisque, outre les perturbations introduites en [106], en [44] la **L**-divergence est appelée **K**-divergence.

Voir [42] pour propriétés (positivité, convexité, inégalités) des $\mathbf{J}, \mathbf{K}, \mathbf{L}$-divergences en fonction des propriétés de h. En particulier:

$$4\ \mathbf{J}_h \leq \mathbf{K}_h \Leftrightarrow -\frac{h(u)}{u} \quad \text{convexe}$$

Noter qu'une **L**-divergence est une f-divergence, pour f donnée en (5.24), ainsi qu'on le montre à la section 5.2.

2.2.3.3 Distance de Bregman et J,K,L-divergences d'ordre α

Pour $h = h_\alpha$, on note $\mathbf{D}_\alpha$, et $\mathbf{J}_\alpha, \mathbf{K}_\alpha, \mathbf{L}_\alpha$ (dites $\mathbf{J}, \mathbf{K}, \mathbf{L}$-divergences d'ordre α):

$$\mathbf{D}_\alpha(P,Q) = \int q^\alpha\ d\lambda(x) + \frac{1}{\alpha-1}\int p^\alpha\ d\lambda(x) - \frac{\alpha}{\alpha-1}\int p q^{\alpha-1}\ d\lambda(x) \tag{2.32}$$

$$\mathbf{J}_\alpha(P,Q) = \frac{1}{\alpha-1}\int \left[\frac{p^\alpha+q^\alpha}{2} - \left(\frac{p+q}{2}\right)^\alpha\right]\ d\lambda(x) \tag{2.33}$$

$$\mathbf{K}_\alpha(P,Q) = \frac{1}{\alpha-1}\int (p-q)(p^{\alpha-1}-q^{\alpha-1})\ d\lambda(x)$$

$$\mathbf{L}_\alpha(P,Q) = \frac{1}{\alpha-1}\left(\int \left(p^\alpha q^{1-\alpha} + q^\alpha p^{1-\alpha}\right)\ d\lambda(x)\ - 2\right)$$

$$\begin{aligned}\mathbf{D}_1(P,Q) &= \bar{\mathbf{K}}(P,Q)\\ \mathbf{K}_1(P,Q) &= \mathbf{K}(P,Q)\\ \mathbf{L}_1(P,Q) &= \mathbf{K}(P,Q)\end{aligned}$$

$$\alpha\ \mathbf{K}_\alpha(P,Q) = \mathbf{D}_\alpha(P,Q) + \mathbf{D}_\alpha(Q,P) \tag{2.34}$$

$$(\alpha-1)\ \mathbf{L}_\alpha = e^{\alpha(\alpha-1)\left(\widetilde{\mathbf{R}}_\alpha + \widetilde{\mathbf{R}}_{(1-\alpha)}\right)} - 2 \tag{2.35}$$

où $\widetilde{\mathbf{R}}_\alpha$ est l'information d'ordre α de Rényi (2.90).

On note que $\mathbf{L}_\alpha$ et $\widetilde{\mathbf{R}}_\alpha$ sont homogènes de degré 1 tout comme les f-divergences $\mathbf{I}_f$ (2.63), et que, pour $\alpha \neq 1$, $\mathbf{D}_\alpha, \mathbf{J}_\alpha, \mathbf{K}_\alpha$ ne le sont pas. Ceci a une conséquence sur les métriques associées; voir plus loin.

2.2.3.4 Divergence et entropie quadratique

[125] Pour l'entropie quadratique $\mathbf{Q}$ (2.62), on a les relations:

$$\begin{aligned}\mathbf{J}_Q^{(\beta)}(P,Q) &= 4\ \beta\ (1-\beta)\ \mathbf{J}_Q(P,Q)\\ \mathbf{D}_Q(P,Q) &= 4\ \mathbf{J}_Q(P,Q)\end{aligned} \tag{2.36}$$

autrement dit la distance de Bregman et la **J**-divergence sont identiques.

Réciproquement, ceci fournit une caractérisation de l'entropie quadratique:

$$\mathbf{J}_H^{(\beta)}(P,Q) = g(\beta, 1-\beta)\ \mathbf{J}_H(P,Q) \ \text{ où } \ g \text{ symétrique } \Leftrightarrow \mathbf{H} = \mathbf{Q}$$

On donnera une autre caractérisation de l'entropie quadratique au chapitre 3.

2.3 Information mutuelle

[51, 3, 49, 74, 75, 154].

2.3.1 Information mutuelle de Shannon

[142, 154] L'information mutuelle ou information de Shannon entre 2 variables aléatoires X_1 et X_2 est l'entropie relative (ou information de Kullback) entre la loi jointe $P(x_1, x_2)$ et le produit des lois $P_1(x_1) \times P_2(x_2)$:

$$\begin{aligned} \mathcal{I}_1(X_1, X_2) &= \bar{\mathbf{K}}(P, P_1 \times P_2) && (2.37) \\ &= \mathbf{H}_1(P_1) + \mathbf{H}_1(P_2) - \mathbf{H}_1(P) \\ &= \mathbf{H}_1(X_1) + \mathbf{H}_1(X_2) - \mathbf{H}_1(X_1, X_2) && (2.38) \\ &= \mathbf{H}_1(X_2) - \mathbf{H}_1(X_2|X_1) \\ &= \mathbf{H}_1(X_1) - \mathbf{H}_1(X_1|X_2) && (2.39) \end{aligned}$$

2.3.2 Extensions

On peut imaginer étendre la définition précédente :

- Soit en changeant d'entropie dans (2.38):

$$\mathcal{I}_H(X_1, X_2) = \mathbf{H}(X_1) + \mathbf{H}(X_2) - \mathbf{H}(X_1, X_2)$$

  - noter qu'avec cette définition, on a $\mathcal{I}_H(X, X) = \mathbf{H}(X)$, i.e. [75] *l'entropie d'une variable aléatoire est l'information qu'elle porte sur elle-même*, ce qui est bien confortable;

- Soit [51, 154] en changeant de divergence dans (2.37):

$$\mathcal{I}_D(X_1, X_2) = \mathbf{D}(P, P_1 \times P_2)$$

  Pour $\mathbf{D} = \mathbf{V}$ (distance en variation), on obtient le coefficient de Höffding, et pour $\mathbf{D} = \mathbf{R}$ (χ^2-divergence) on obtient le « mean square contingency » de Pearson [154, 126]; pour $\mathbf{D} = \mathbf{I}_f$, on obtient la f-information de Csiszár [51, 52]; pour $\mathbf{D} = \widetilde{\mathbf{R}}_\alpha$, on a l'information mutuelle d'ordre α de Csiszár [52, 56].

- Soit même [152] en changeant de définition de l'entropie conditionnelle dans (2.39).

- ou encore [52] en définissant:

$$\mathbf{H}_f(X) = \mathcal{I}_{I_f}(X, X)$$

  et:

$$\mathcal{I}_{H_f}(X_1, X_2) = \mathbf{H}_f(X_2) - \mathbf{H}_f(X_2|X_1)$$

Noter que

$$\mathcal{I}_{\widetilde{H}_\alpha}(X_1, X_2) \neq \mathcal{I}_{\widetilde{R}_\alpha}(X_1, X_2) \quad [56]$$

2.3.3 Conditionnement et additivité

L'information de Kullback possède la propriété de décomposition additive par conditionnement (dite de chain rule [49]) :

$$\bar{\mathbf{K}}(P_{X_1,X_2}, Q_{X_1,X_2}) = \bar{\mathbf{K}}(P_{X_1}, Q_{X_1}) + \mathbf{E}_{X_1}(\bar{\mathbf{K}}(P_{X_2|X_1}, Q_{X_2|X_1}))$$

La seule autre f-divergence qui possède cette propriété semble être l'information de Rényi (2.90) dans le cas Gaussien (voir à la fin). En effet, pour les f-divergences intégrales, une condition nécessaire est que :

$$f\left(\frac{q_{x_2|x_1}}{p_{x_2|x_1}}\,\frac{q_{x_1}}{p_{x_1}}\right) = f\left(\frac{q_{x_1}}{p_{x_1}}\right) + f\left(\frac{q_{x_2|x_1}}{p_{x_2|x_1}}\right)$$

i.e. compte tenu de (2.64) :

$$f(u) = -\alpha \ln u \quad (\alpha > 0)$$

Pour les f-divergences non intégrales, il faut de même que :

$$\begin{aligned} \psi(u+v) &= \psi(u)\ \psi(v) \\ \psi^{-1}(uv) &= \psi^{-1}(u) + \psi^{-1}(v) \end{aligned}$$

soit $\psi(u) = e^{-\alpha u}$.

2.4 Exemples d'entropies

On considère des fonctionnelles d'entropie $\mathbf{H}_\nu(P)$, pas nécessairement concaves. La dépendance vis-à-vis de la mesure de référence ν sera omise lorsqu'il n'en résultera pas d'ambiguïté.

Comme indiqué en introduction, on distingue trois formes fonctionnelles d'entropie, l'une intégrale, les deux autres pas.

2.4.1 Fonctionnelle d'entropie intégrale

2.4.1.1 Définition et propriétés d'invariance

Soit h une fonction convexe, souvent de classe $\mathcal{C}^2$: $T_h \to \mathbb{R}$, où $T_0 = [0,1] \subset T_h \subset \mathbb{R}^+$.

L'entropie intégrale d'une loi de probabilité P relative à une mesure ν est définie par :

$$\mathbf{H}_{h,\nu}(P) = -\int h(p(x))\ d\nu(x) \tag{2.40}$$

où la densité $p = \frac{dP}{d\nu}$ est à valeurs dans T_h.

En vertu de (4.31), on a la propriété :

$$\text{pour } \breve{h}(u) = \gamma\ h(u) + \delta\ u, \qquad \mathbf{H}_{\breve{h},\nu}(P) = \gamma\ \mathbf{H}_{h,\nu}(P) - \delta \tag{2.41}$$

2.4.1.2 Exemple : entropie d'ordre α de Havrda-Charvát

Entropie d'ordre 1 et dualité L'entropie de Shannon, dite aussi entropie ou mesure d'information d'ordre 1, s'écrit :

$$\mathbf{H}_{1,\nu}(P) = -\int p(x) \ln p(x)\ d\nu(x) \tag{2.42}$$

$$= \mathbf{E}_P\left(\ln \frac{1}{p(X)}\right) \tag{2.43}$$

Elle est de forme intégrale pour :

$$h(u) = h_1(u) = u \ln u$$

et l'effet du changement de mesure de référence s'exprime par :

$$\mathbf{H}_{1,\nu'}(P) = \mathbf{H}_{1,\nu}(P) + \mathbf{E}_P\left(\ln \frac{d\nu'}{d\nu}\right)$$

Enfin, un calcul direct de Lagrangien permet de montrer que, considérée comme fonction d'une densité de probabilité, *l'opposée de l'entropie de Shannon est la fonction duale d'un log. m.g.f.* (logarithme d'une fonction génératrice des moments) [59] :

$$(-\mathbf{H}_{1,\nu})^*(q) = \ln \int e^q\ d\nu \tag{2.44}$$

Entropie d'ordre α L'entropie d'ordre α de Havrda-Charvát, dite aussi mesure d'information d'ordre α, est l'entropie intégrale correspondant à la fonction h_α [78] :

$$h_\alpha(u) = \begin{cases} \frac{1}{\alpha-1}\ (u^\alpha - u) & (\alpha \neq 1) \\ u \ln u & (\alpha = 1) \end{cases} \tag{2.45}$$

définie pour $\alpha \in \mathbb{R}$. Cette fonction satisfait :

$$h''_\alpha(u) = \alpha u^{\alpha-2}$$

$$\breve{h}_\alpha(u) = \frac{1}{\alpha-1}\ (u^{1-\alpha} - 1) \tag{2.46}$$

$$\ln\Big(1 + (\alpha-1)\ \breve{h}_\alpha(u)\Big) = (1-\alpha) \ln u \tag{2.47}$$

$$h^*_\alpha(u) = \left(\frac{(\alpha-1)u + 1}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \tag{2.48}$$

La fonction g associée est :

$$g_\alpha(u) = \frac{1}{\alpha - 1}\left(1 - e^{-(\alpha-1)u}\right)$$

Pour simplifier, on note $\mathbf{H}_{\alpha,\nu} = \mathbf{H}_{h_\alpha,\nu}$ cette entropie qui s'écrit :

$$\mathbf{H}_{\alpha,\nu}(P) = \begin{cases} \frac{1}{\alpha-1}\left(1 - \int p^\alpha(x)\, d\nu(x)\right) & (\alpha \neq 1) \\ -\int p(x) \ln p(x)\, d\nu(x) & (\alpha = 1) \end{cases} \tag{2.49}$$

Pour $\alpha = 1$, on retrouve l'entropie de Shannon $\mathbf{H}_1$, et pour $\alpha = 2$, il s'agit de l'indice de diversité de Gini [122].

Cette entropie admet une autre expression [109] :

$$\mathbf{H}_{\alpha,\nu}(P) = \int F^\alpha(x)\, \bar{h}_\alpha\left(\frac{p(x)}{F(x)}\right)\, d\nu(x)$$

où :

$$F(x) = \int_{-\infty}^{x} p(y)\, d\nu(y)$$

est la fonction de répartition, et $\bar{h}_\alpha$ est la fonction d'information associée à h_α :

$$\bar{h}_\alpha(u) = h_\alpha(u) + h_\alpha(1-u) = \frac{1}{\alpha - 1}\left[u^\alpha + (1-u)^\alpha - 1\right]$$

Cette expression peut être utile en fiabilité pour des durées de vie.

Opérateur de moyenne L'opérateur de moyenne (4.8) associé à l'entropie de Havrda-Charvát est :

$$\Sigma_{h_\alpha}^{(\beta)}(p) = \frac{1}{\alpha - 1}\left(1 - \sum_{i=1}^{n} \beta_i\, p_i^{\alpha-1}\right) \tag{2.50}$$

où l'on suppose toujours les poids normalisés : $\sum_{i=1}^n \beta_i = 1$.

2.4.1.3 Entropies déduites des f-divergences intégrales

L'écriture des formes intégrales d'entropies (2.1) et de f-divergences associées (2.12) permet de définir, à partir d'une f-divergence intégrale quelconque entre une loi μ et une mesure de référence ν, une entropie intégrale de la loi μ par :

$$\mathbf{H}_{f,\nu}(\mu) = -\mathbf{I}_f(\mu, \nu)$$

soit encore :

$$\mathbf{H}_{f,\nu}(P) = -\int f(p(x))\, d\nu(x)$$

pour l'une quelconque des fonctions f listées plus loin.

Par exemple, l'entropie qui correspond à la distance (ou divergence) de Hellinger est :

$$\mathbf{H}_{\mathcal{H},\nu}(P) = \int \sqrt{p(x)}\, d\nu(x) - 1$$

qui n'est autre que la moitié de l'entropie de Havrda-Charvát d'ordre 1/2. De même, l'entropie qui correspond à la χ^2-divergence est :

$$\mathbf{H}_{r,\nu}(P) = \frac{1}{2}\int (p(x) - 1)^2\, d\nu(x)$$

qui n'est autre que la moitié de l'entropie de Havrda-Charvát d'ordre 2.

2.4.2 Fonctionnelles d'entropie non intégrales

On considère maintenant les formes non intégrales (2.4)-(2.3).

2.4.2.1 Définition et propriétés d'invariance

On définit des entropies non intégrales d'une loi P relativement à une mesure de référence ν par [52, 3] :

$$\mathbf{H}_{\psi,\nu}(P) = \psi^{-1}\left(\int p(x)\ \psi(-\ln p(x))\ d\nu(x)\right) \tag{2.51}$$

$$\mathbf{H}_{\phi,\nu}(P) = -\ln\ \phi^{-1}\left(\int p(x)\ \phi(p(x))\ d\nu(x)\right) \tag{2.52}$$

où ψ et ϕ sont strictement monotones, et souvent de classe $\mathcal{C}^2$, et :

$$\phi(u) = \psi(-\ln u)$$

On déduit de (4.29)-(4.30) que les entropies $\mathbf{H}_{\psi,\nu}(P), \mathbf{H}_{\phi,\nu}(P)$ sont inchangées par dilatation de ψ, ϕ :

$$\text{pour } \underline{\psi}(u) = \gamma\ \psi(u) + \delta, \qquad \mathbf{H}_{\underline{\psi},\nu}(P) = \mathbf{H}_{\psi,\nu}(P) \tag{2.53}$$

$$\text{pour } \underline{\phi}(u) = \gamma\ \phi(u) + \delta, \qquad \mathbf{H}_{\underline{\phi},\nu}(P) = \mathbf{H}_{\phi,\nu}(P) \tag{2.54}$$

2.4.2.2 Exemple: entropie d'ordre α de Rényi

Comme on l'a déjà indiqué, l'entropie de Shannon (2.42) peut être écrite sous ces formes fonctionnelles non intégrales à l'aide de :

$$\psi(u) = \psi_1(u) = u$$

et :

$$\phi(u) = \phi_1(u) = -\ln u$$

L'entropie d'ordre α de Rényi [130, 52, 49] est l'entropie de forme fonctionnelle (2.51) correspondant à la fonction ψ :

$$\psi_\alpha(u) = e^{-(\alpha-1)u} \tag{2.55}$$

et à la fonction ϕ :

$$\phi_\alpha(u) = u^{\alpha-1} \tag{2.56}$$

Pour simplifier, on note $\widetilde{\mathbf{H}}_{\alpha,\nu} = \mathbf{H}_{\psi_\alpha,\nu} = \mathbf{H}_{\phi_\alpha,\nu}$ cette entropie qui s'écrit :

$$\widetilde{\mathbf{H}}_{\alpha,\nu}(P) = \begin{cases} -\frac{1}{\alpha-1}\ \ln \int p^\alpha(x)\ d\nu(x) & (\alpha \in \mathbb{R}_+^*, \neq 1) \\ -\int p(x) \ln p(x)\ d\nu(x) & (\alpha = 1) \\ \ln\ \nu(\{x : p(x) > 0\}) & (\alpha = 0) \\ -\ln \max_x p(x) & (\alpha = \infty) \end{cases} \tag{2.57}$$

L'entropie d'ordre 0 a été introduite par Hartley en 1920 et semble avoir été la première mesure d'information [4]. Selon Rényi et Csiszár [130, 52], l'entropie d'ordre α a été introduite par Schützenberger en 1953 [141]; elle a été reprise par Rényi en 1961. Néanmoins Parzen et Vajda [117, 155] affirment que l'information de Rényi n'est autre qu'une divergence proposée par Bhattacharrya en 1943 [30]. Pour $\alpha \neq 1$, cette entropie n'est *pas* de la forme (2.40). Pour $0 < \alpha \leq 1$, $\widetilde{\mathbf{H}}_\alpha$ est strictement convexe [28]. Par contre, pour $\alpha > 1$, $\widetilde{\mathbf{H}}_\alpha$ n'est ni concave ni convexe en général [28, 122]. Cependant, pour tout α, $\widetilde{\mathbf{H}}_\alpha$ est pseudo-concave.

L'entropie de Rényi possède l'importante propriété d'être liée à la fonction caractéristique de $\ln P(X)$ (ce qui laisse entrevoir son rôle en codage) par [46, 52, 56] :

$$\mathbf{E}_\nu\left(e^{\alpha \ln P(X)}\right) = e^{-\alpha\ \widetilde{\mathbf{H}}_{\alpha+1,\nu}(P)} = \psi_{\alpha+1}\left(\widetilde{\mathbf{H}}_{\alpha+1,\nu}(P)\right) \tag{2.58}$$

Autrement dit, *l'opposé de l'entropie de Rényi – et non pas sa fonction duale – est un log. m.g.f.* :

$$-\alpha\ \widetilde{\mathbf{H}}_{\alpha+1,\nu}(P) = \ln\ \mathbf{E}_\nu\left(e^{\alpha \ln P(X)}\right)$$

L'entropie de Rényi est liée à l'entropie de Havrda-Charvát par :

$$(1-\alpha)\ \widetilde{\mathbf{H}}_{\alpha,\nu}(P) = \ln[1 - (\alpha-1)\ \mathbf{H}_{\alpha,\nu}(P)]$$

soit :

$$\mathbf{H}_{\psi_\alpha,\nu}(\mu) = \psi_\alpha^{-1}\ (1 - (\alpha-1)\mathbf{H}_{h_\alpha,\nu}(\mu))$$

Opérateurs de moyenne Les opérateurs de moyenne (4.7) et (4.9) correspondant à l'entropie de Rényi sont :

$$\Sigma_{\psi_\alpha}^{(\beta)}(u) = -\frac{1}{\alpha-1}\ \ln\ \sum_{i=1}^{n} \beta_i\ e^{-(\alpha-1)u_i} \tag{2.59}$$

et :

$$\Sigma_{\phi_\alpha}^{(\beta)}(p) = \left(\sum_{i=1}^{n} \beta_i\ p_i^{\alpha-1}\right)^{\frac{1}{\alpha-1}} \tag{2.60}$$

tandis que la ϕ_α-average probability (4.10) est :

$$\mathbf{G}_\alpha(p) = \left(\sum_{i=1}^{n}\ p_i^{\alpha}\right)^{\frac{1}{\alpha-1}} \tag{2.61}$$

Les relations établissant le lien avec l'opérateur de moyenne (2.50) sous-jacent à l'entropie de Havrda-Charvát diffèrent des relations (4.12)-(4.13) du cas général, à cause du lien entre les trois fonctions $h_\alpha, \psi_\alpha, \phi_\alpha$ résumé au tableau 2.1, qui est distinct du cas général discuté dans l'introduction.

Ces relations s'écrivent dans le cas présent :

$$\begin{aligned}\Sigma_{\psi_\alpha}^{(\beta)}(-\ln p, \ldots, -\ln p_n) &= \psi_\alpha^{-1}\left(1-(\alpha-1)\ \Sigma_{h_\alpha}^{(\beta)}(p)\right)\\ \Sigma_{\phi_\alpha}^{(\beta)}(p) &= \phi_\alpha^{-1}\left(1-(\alpha-1)\ \Sigma_{h_\alpha}^{(\beta)}(p)\right)\end{aligned}$$

Par contre, le lien entre ψ et ϕ étant inchangé, la relation (4.20) vaut toujours :

$$\Sigma_{\psi_\alpha}^{(\beta)}(-\ln p, \ldots, -\ln p_n) = -\ln\ \Sigma_{\phi_\alpha}^{(\beta)}(p, \ldots, p_n)$$

Lorsque α tend vers l'infini, l'entropie de Rényi correspond à l'opérateur de moyenne particulier qu'est le sup [146]. (Ceci remarque permet de donner une interprétation de la distance en variation de Kolmogorov en termes d'information). On utilisera ces notions de moyenne sous-jacentes aux entropies d'ordre α pour définir, à la section 4.2, le rayon d'information comme extension de la différence de Jensen.

2.4.2.3 Entropies déduites des f-divergences non intégrales

L'écriture des formes non intégrales d'entropies (2.4)-(2.3) et de f-divergences associées (2.12) permet de définir, à partir d'une f-divergence non intégrale quelconque entre une loi μ et une mesure de référence ν, une entropie non intégrale de la loi μ par :

$$\begin{aligned}\mathbf{H}_{\psi,\nu}(\mu) &= -\mathbf{I}_\psi(\mu,\nu)\\ \mathbf{H}_{\phi,\nu}(\mu) &= -\mathbf{I}_\phi(\mu,\nu)\end{aligned}$$

soit encore :

$$\begin{aligned}\mathbf{H}_{\psi,\nu}(P) &= \psi^{-1}\left(\int p(x)\ \psi(-\ln p(x))\ d\nu(x)\right)\\ \mathbf{H}_{\phi,\nu}(P) &= -\ln\phi^{-1}\left(\int p(x)\ \phi(p(x))\ d\nu(x)\right)\end{aligned}$$

pour l'une quelconque des fonctions ψ, ϕ listées dans le tableau 2.3.

Par exemple, l'entropie qui correspond à la distance (ou divergence) de Bhattacharrya ($\phi(u) = -1/\sqrt{u}$) est :

$$\mathbf{H}_{\phi,\nu}(P) = 2\ln\int \sqrt{p(x)}\ d\nu(x)$$

qui n'est autre que l'entropie de Rényi d'ordre 1/2. De même, l'entropie qui correspond à l'information d'ordre α de Rényi est :

$$\mathbf{H}_{\psi_\alpha,\nu}(P) = -\frac{1}{\alpha-1}\ \ln\int p^\alpha(x)\ d\nu(x)$$

qui n'est autre que l'entropie de Rényi d'ordre α.

2.4.3 Généralisations

On discute plusieurs types de généralisations.

2.4.3.1 Par l'axiomatique pure

De nombreuses entropies ont été introduites dans la littérature [3, 109, 152]. Par exemple, les entropies :

$$\begin{aligned}
\mathbf{H}_{\alpha}^{\beta}(P) &= -\frac{1}{\alpha-1}\left(1-\frac{\int p^{\alpha+\beta-1}\,d\nu(x)}{\int p^{\beta}\,d\nu(x)}\right) \\
\widetilde{\mathbf{H}}_{\alpha}^{\beta}(P) &= -\frac{1}{\alpha-1}\,\ln\,\frac{\int p^{\alpha+\beta-1}\,d\nu(x)}{\int p^{\beta}\,d\nu(x)} \\
\check{\mathbf{H}}^{\beta}(P) &= -\int p^{\beta+1}\,\ln p\,d\nu(x)
\end{aligned}$$

sont extraites des 25 entropies différentes listées dans [152]. Selon [52], ces entropies n'ont jamais donné lieu à autre chose que des justifications purement axiomatiques, et non pragmatiques ou opérationnelles, et n'ont donc qu'un intérêt réduit.

2.4.3.2 Entropie quadratique

Par contre, l'entropie quadratique :

$$\mathbf{Q}(P) = -\int_{\mathcal{X}\otimes\mathcal{X}} q(x,y)\;p(x)\;p(y)\;d\nu(x)\;d\nu(y) \tag{2.62}$$

où $q(x,x)=0$, a été introduite par Rao [125, 124, 102]. Elle est intéressante :

- par au moins trois de ses caractérisations (chapitre 3) : d'une part, l'identité de deux procédés constructifs de divergences (par différentielle de Gâteaux et par différence de Jensen, i.e. **D** et **J**), et la symétrie de **D**; d'autre part, la positivité de toutes les différences de Jensen successives;
- parce que la métrique différentielle quadratique qui lui est associée est, comme celle de Fisher, invariante par changement sur le paramètre et sur la variable.

Sa concavité est donnée par la condition :

$$q \text{ est un noyau défini positif } \Leftrightarrow \mathbf{Q} \text{ concave}$$

À titre d'exemple, pour $\mathcal{X}=\mathbb{R}$, $q(x,y)=(x-y)^2$ redonne la variance comme mesure de diversité [123] :

$$\int_{\mathcal{X}\otimes\mathcal{X}} (x-y)^2\;p(x)\;p(y)\;d\nu(x)\;d\nu(y) = 2\;\mathrm{var}(P)$$

Un autre exemple concerne $\mathcal{X}=\mathbb{R}^2$, $q(x,y)=(x_1-y_1)^2+(x_2-y_2)^2$.

2.4.3.3 Entropie associée à une f-divergence symétrique

On peut aussi définir des entropies à partir de f-divergences particulières, par des procédés autres que la définition (2.12). Nous en donnons deux exemples.

Par identité de procédés constructifs de divergences Comme on le montre plus loin en cherchant les f-divergences qui sont aussi une **L**-divergence de Rao, à toute f-divergence telle que la fonction convexe f satisfasse (5.25), on peut associer une famille d'entropies par résolution de (5.24) :

$$h(u) = \frac{1}{2}\,f(u) + \kappa\,\sqrt{u}\,\left(\underline{h}(u) - \underline{h}\left(\frac{1}{u}\right)\right)$$

où $\underline{h}$ est n'importe quelle fonction convexe définie sur $[0,1[$ et nulle en dehors.

2.4.3.4 Entropie associée à une information mutuelle

On peut aussi imaginer définir une entropie à partir d'une information mutuelle par la relation [52, 19] :

$$\mathbf{H}_f(X) = \mathcal{I}_{I_f}(X, X)$$

C'est ce que propose Arimoto [19] pour une entropie définie par un inf, qui lui fait retrouver le rayon d'information de Sibson (section 4.2).

2.5 Exemples de f-divergences

On décrit en détail les f-divergences (2.12), intégrales ou non, proposées dans la littérature. En particulier, on distingue un certain nombre de divergences d'ordre α. On donne ensuite un contre-exemple de divergence d'ordre α qui n'est *pas* une f-divergence. On examine ensuite, dans le cas paramétrique, la relation étroite entre les f-divergences et l'information de Fisher et l'exhaustivité. On termine par un certain nombre d'inégalités entre ces distances, permettant d'affirmer quand l'une est plus fine que l'autre.

2.5.1 f-divergences intégrales

2.5.1.1 Définition

Dans le cas de lois de probabilité P, Q admettant des densités p, q par rapport à une mesure de référence λ, on les définit [11, 51, 6, 165, 53, 117] par :

$$\mathbf{I}_f(P, Q) = \mathbf{E}_Q\left(f\left(\frac{p}{q}\right)\right) = \int f\left(\frac{p}{q}\right)\, q\, d\lambda(x) \tag{2.63}$$

où f est continue et convexe sur $[0, +\infty[$, souvent de classe $\mathcal{C}^2$. On impose de plus :

$$f(1) = 0 \qquad \text{pour garantir } \mathbf{I}_f(P, P) = 0 \tag{2.64}$$

$$f''(1) > 0 \qquad \text{pour la métrique} \tag{2.65}$$

Il convient de noter que les f-divergences – qui sont homogènes de degré 1 – *ne dépendent pas de la mesure* λ *de référence*. De ce point de vue, on peut dire que la comparaison de telles mesures de distance se réduit à la comparaison de fonctions convexes.

Comme déjà annoncé en introduction, les f-divergences possèdent les propriétés d'invariance suivantes :

$$\text{pour } \check{f}(u) = f(u) + \gamma\, u + \delta, \qquad \mathbf{I}_{\check{f}}(P, Q) = \mathbf{I}_f(P, Q) + \gamma + \delta \tag{2.66}$$

$$\text{pour } \check{f}(u) = uf\left(\frac{1}{u}\right), \qquad \mathbf{I}_{\check{f}}(P, Q) = \mathbf{I}_f(Q, P) \tag{2.67}$$

Enfin, comme on l'a déjà indiqué à la section 1 en (2.12), une f-divergence intégrale (entre deux mesures n'admettant pas nécessairement de densité) se déduit d'une entropie intégrale par :

$$\mathbf{I}_h(\mu, \nu) = -\,\mathbf{H}_{h,\nu}(\mu)$$

à un coefficient multiplicatif près; voir le tableau 2.1.

2.5.1.2 Exemples

On présente les exemples de f-divergences résumés au tableau 2.2.

Distance en variation de Kolmogorov Elle correspond à $f(u) = v(u)$ où :

$$v(u) = \frac{1}{2}\,|u - 1|$$

$$\mathbf{I}_v(P, Q) = \mathbf{V}(P, Q) = \frac{1}{2}\int |p - q|\, d\lambda(x) \tag{2.68}$$

Une interprétation en termes de notion d'information est donnée en [146]; on l'indique en section 4.2 lors des extensions de la différence de Jensen.

TAB. 2.2 – *f-divergences intégrales.*

nom	notation pour $f(u)$	$\mathbf{I}_f(P,Q)$
Kolmogorov	$v(u)$	$\mathbf{V}(P,Q) = \frac{1}{2} \int \lvert p-q \rvert \, d\lambda(x)$
info. Kullback	$\bar{k}(u)$	$\bar{\mathbf{K}}(P,Q) = \int p \ln \frac{p}{q} \, d\lambda(x)$
div. Kullback	$k(u)$	$\mathbf{K}(P,Q) = \int (p-q)(\ln p - \ln q) \, d\lambda(x)$
K-div. Lin	$\kappa(u)$	$K(P,Q) = \int p \ln \frac{2p}{p+q} \, d\lambda(x)$
L-div. Lin	$l(u)$	$L(P,Q) = \int \left(p \ln p + q \ln q - (p+q) \ln \frac{p+q}{2} \right) \, d\lambda(x)$
Hellinger	$\frac{1}{2} (\sqrt{u}-1)^2$ $1 - \sqrt{u} = \frac{1}{2} h_{1/2}(u) - u + 1$	$\mathcal{H}^2(P,Q) = \frac{1}{2} \int (\sqrt{p} - \sqrt{q})^2 \, d\lambda(x)$ $= 1 - \int \sqrt{p\,q} \, d\lambda(x) = \frac{1}{2} \mathbf{I}_{h_{1/2}}(P,Q)$
moy. harmon.	$t(u)$	$\mathbf{T}(P,Q) = \int \left(p - \frac{2\,p\,q}{p+q} \right) \, d\lambda(x)$
χ^2-divergence	$r(u)$ $= \frac{1}{2} h_2(u) + \frac{1-u}{2}$	$\mathbf{R}(P,Q) = \frac{1}{2} \int \frac{(p-q)^2}{q} \, d\lambda(x)$ $= \frac{1}{2} \mathbf{I}_{h_2}(P,Q)$
χ^α-div. Vajda	w_α	$\mathbf{W}_\alpha(P,Q) = \int \lvert p-q \rvert^\alpha q^{1-\alpha} \, d\lambda(x)$ $\mathbf{W}_1(P,Q) = 2\,\mathbf{V}(P,Q)$ $\mathbf{W}_2(P,Q) = 2\,\mathbf{R}(P,Q)$
χ^2-div. α	$r_\alpha(u)$ $\frac{1}{\mathcal{Q}} \, (h_\alpha(u) - (u-1))$ $\bar{k}(u)$ $\bar{k}(u)$ $r(u)$ $2(\sqrt{u}-1)^2$	$\mathbf{R}_\alpha(P,Q) = \frac{1}{\alpha\,(\alpha-1)} \, \left(\int p^\alpha \, q^{1-\alpha} \, d\lambda(x) - 1 \right)$ $= \frac{1}{\alpha} \mathbf{I}_{h_\alpha}(P,Q)$ $\mathbf{R}_0(P,Q) = \bar{\mathbf{K}}(Q,P)$ $\mathbf{R}_1(P,Q) = \bar{\mathbf{K}}(P,Q)$ $\mathbf{R}_2(P,Q) = \mathbf{R}(P,Q)$ $\mathbf{R}_{1/2}(P,Q) = 4\,\mathcal{H}^2(P,Q)$
Info α Parzen	$z_\alpha(u)$ $\breve{\bar{k}}(u)$ $\frac{\bar{k}}{u}$	$\mathbf{Z}_\alpha(P,Q)$ $\mathbf{Z}_1(P,Q) = \bar{\mathbf{K}}(Q,P)$ $\mathbf{Z}_0(P,Q) = \mathbf{R}_2(P,Q) - \bar{\mathbf{K}}(P,Q)$ $\mathbf{Z}_{1/2}(P,Q) = 2\,\bar{\mathbf{K}}(P,Q) - 4\,K(P,Q)$
Info.α Bose-Einstein	$b_\alpha(u)$	$\mathbf{B}_\alpha(P,Q) = \int \left(p \ln p + \alpha q \ln q \right.$ $\left. -(\alpha q + p) \, \ln \frac{\alpha q + p}{\alpha + 1} \right) \, d\lambda(x)$
Info.α Fermi-Dirac	$f_\alpha(u)$	$\mathbf{F}_\alpha(P,Q) = \int \left(p \ln p - \alpha q \ln q \right.$ $\left. +(\alpha q - p) \, \ln \frac{\alpha q - p}{\alpha - 1} \right) \, d\lambda(x)$

Information de Kullback ou entropie relative Elle est dite aussi gain d'information d'ordre 1 [146] ou divergence dirigée [145] ou entropie relative de Shannon. Elle correspond à la fonction $f(u) = u \ln u$ ou $f(u) = \bar{k}(u)$ où :

$$\bar{k}(u) = u \ln u - (u-1)$$

$$\begin{aligned} \mathbf{I}_{\bar{k}}(P,Q) = \bar{\mathbf{K}}(P,Q) &= \int p \ln \frac{p}{q} \, d\lambda(x) \\ &= \int \left(\frac{p}{q} \ln \frac{p}{q} - \frac{p}{q} + 1 \right) \, q \, d\lambda(x) \\ &= -\int \left(\ln \frac{q}{p} - \frac{q}{p} + 1 \right) \, p \, d\lambda(x) \end{aligned}$$

Divergence de Kullback Elle est dite aussi divergence de Jeffreys-Kullback-Leibler. C'est la symétrisée de la précédente, elle correspond donc à la fonction $f(u) = k(u)$ où :

$$k(u) = \bar{k}(u) + u \, \bar{k}\left(\frac{1}{u}\right) = (u-1) \, \ln u$$

$$\mathbf{K}(P,Q) = \bar{\mathbf{K}}(P,Q) + \bar{\mathbf{K}}(Q,P) = \int \left(p \ln \frac{p}{q} + q \ln \frac{q}{p} \right) \, d\lambda(x) \tag{2.69}$$

$$= \mathbf{I}_k(P,Q) = \int (p-q)(\ln p - \ln q) \, d\lambda(x) \tag{2.70}$$

K, L-divergences de Lin Elles sont d'introduction plus récente [106]. La K-divergence de Lin correspond à la fonction $f(u) = \kappa(u) = u \ln u - u \ln \frac{1+u}{2}$:

$$\mathbf{I}_k(P,Q) = K(P,Q) = \bar{\mathbf{K}}\left(P, \frac{P+Q}{2}\right) = \int p \ln \frac{2p}{p+q} \, d\lambda(x) \tag{2.71}$$

et la L-divergence de Lin à la fonction symétrisée $f(u) = l(u)$ où :

$$l(u) = \kappa(u) + u \, \kappa\left(\frac{1}{u}\right) = u \ln u - (1+u) \, \ln \frac{1+u}{2}$$

$$\mathbf{I}_l(P,Q) = L(P,Q) = \bar{\mathbf{K}}\left(P, \frac{P+Q}{2}\right) + \bar{\mathbf{K}}\left(Q, \frac{P+Q}{2}\right) \tag{2.72}$$

$$= \int \left(p \ln p + q \ln q - (p+q) \ln \frac{p+q}{2} \right) \, d\lambda(x) \tag{2.73}$$

$$= 2 \, \mathbf{J}_1(P,Q) \tag{2.74}$$

La L-divergence de Lin est donc une différence de Jensen $\mathbf{J}_1 \stackrel{\Delta}{=} \mathbf{J}_{H_1}^{(1/2)}$ définie en (4.33).

Distance de Hellinger La distance de Hellinger a été introduite en 1909 dans le cas discret. Elle joue un rôle crucial en inférence statistique. L'un de ses intérêts est qu'elle est souvent calculable explicitement. Par contre, elle est moins fine que l'information de Kullback (voir les inégalités à la fin de la présente section).

Elle correspond à la fonction $f(u) = \frac{1}{2} \, (\sqrt{u} - 1)^2$:

$$\mathcal{H}^2(P,Q) = \frac{1}{2} \, \int (\sqrt{q} - \sqrt{p})^2 \, d\lambda(x)$$

mais peut aussi être écrite sous la forme :

$$\mathcal{H}^2(P,Q) = 1 - \int \sqrt{p \, q} \, d\lambda(x)$$

Ces deux écritures donnent lieu à deux définitions possibles de divergence de Hellinger d'ordre α. La première consiste à remplacer, dans la dernière expression, les puissances $1/2$ et $1/2$ par α et $1-\alpha$, pour donner lieu à la χ^2-divergence d'ordre α. La deuxième consiste à remplacer, dans la première expression, les puissances $1/2$ par $\alpha/2$, pour donner lieu à ce qui est appelé la distance de Hellinger d'ordre α, mais n'est *pas* une f-divergence. On décrit ces deux distances plus loin.

On déduit de la deuxième expression :

$$\mathcal{H}^2(P,Q) = \mathbf{I}_{\check{h}_{1/2}}(P,Q)$$

où:

$$\check{h}_{1/2}(u) = 1 - \sqrt{u} = \frac{1}{2}\, h_{1/2}(u) - u + 1$$

et $h_{1/2}$ est définie en (2.45). En vertu de (2.66):

$$\mathcal{H}^2(P,Q) = \frac{1}{2}\, \mathbf{I}_{h_{1/2}}(P,Q)$$

Autrement dit, la distance de Hellinger n'est autre que la f-divergence associée à l'entropie de Havrda-Charvát d'ordre $1/2$.

Distance de moyenne harmonique de Toussaint Elle correspond à la fonction [109] $f(u) = t(u)$ où:

$$t(u) = u\, \frac{u-1}{u+1}$$

$$\mathbf{I}_t(P,Q) = \mathbf{T}(P,Q) = \int \left(p - \frac{2\, p\, q}{p+q} \right)\, d\lambda(x)$$

χ^2-divergence ou W-divergence de Kagan Elle est discutée par exemple en [117, 89, 86, 151], et correspond à $f(u) = r(u)$ où:

$$r(u) = \frac{1}{2}\,(u-1)^2 = \frac{1}{2}\,(u^2-u) + \frac{1}{2}\,(1-u) = \frac{1}{2}\, h_2(u) + \frac{1}{2}\,(1-u)$$

$$\mathbf{I}_r(P,Q) = \mathbf{R}(P,Q) = \frac{1}{2} \int \frac{(p-q)^2}{q}\, d\lambda(x) = \frac{1}{2}\, \mathbf{I}_{h_2}(P,Q) \tag{2.75}$$

Autrement dit, la χ^2-divergence n'est autre que la f-divergence associée à l'entropie de Havrda-Charvát d'ordre 2 (ou indice de diversité de Gini [122]).

C'est une des fonctions de contraste de [121]. Cette distance conduit à une mesure d'information mutuelle proposée par Pearson. La limite (au sens local) en a été proposée par Kagan comme mesure d'information alternative à Fisher (pour une borne de Cramer-Rao) dans le cas de densités non différentiables par rapport au paramètre.

χ^α-divergence de Vajda Elle a été introduite en 1971 [153] avec une notion d'information de Fisher d'ordre α. On y revient plus loin lorsque l'on discute le lien des f-divergences avec l'exhaustivité.

Elle correspond à $f(u) = w_\alpha(u)$ pour $\alpha \geq 1$, où:

$$w_\alpha(u) = |u-1|^\alpha$$

$$\mathbf{I}_{w_\alpha}(P,Q) = \mathbf{W}_\alpha(P,Q) = \int |p-q|^\alpha q^{1-\alpha}\, d\lambda(x) \tag{2.76}$$

Elle admet comme cas particuliers :

$$\begin{aligned} \mathbf{W}_1(P,Q) &= 2\, \mathbf{V}(P,Q) \\ \mathbf{W}_2(P,Q) &= 2\, \mathbf{R}(P,Q) \end{aligned}$$

χ^2-divergence d'ordre α ou $\mathbf{I}_\alpha$-divergence de Csiszár Elle correspond [105, 154, 117] à la fonction $f(u) = r_\alpha(u)$ où :

$$r_\alpha(u) = \begin{cases} \frac{1}{\alpha\ (\alpha-1)}\ (u^\alpha - 1 - \alpha\ (u-1)) & (\alpha \neq 0,1) \\ -\ln u + u - 1) = u\ \bar{k}\left(\frac{1}{u}\right) & (\alpha = 0) \\ u \ln u - u + 1 = \bar{k}(u) & (\alpha = 1) \\ \frac{1}{2}(u-1)^2 = r(u) = \frac{1}{2}\ h_2(u) + \frac{1}{2}\ (1-u) & (\alpha = 2) \\ 2(\sqrt{u}-1)^2 = 2\ h_{1/2}(u) + 2\ (1-u) & (\alpha = \frac{1}{2}) \end{cases} \tag{2.77}$$

qui est telle que: $r_\alpha(1) = 0, r''_\alpha(1) = 1, r''_\alpha(u) = u^{\alpha-2}$ et :

$$r_\alpha(u) = \frac{1}{\alpha}\ (h_\alpha(u) - (u-1))$$

Donc :

$$\mathbf{I}_{r_\alpha}(P,Q) = \mathbf{R}_\alpha(P,Q) = \begin{cases} \frac{1}{\alpha\ (\alpha-1)}\ \left(\int p^\alpha\ q^{1-\alpha}\ d\lambda(x) - 1\right) & (\alpha \neq 0,1) \\ \bar{\mathbf{K}}(Q,P) & (\alpha = 0) \\ \bar{\mathbf{K}}(P,Q) & (\alpha = 1) \\ \mathbf{R}(P,Q) & (\alpha = 2eq - div) \\ 4\ \mathcal{H}^2(P,Q) & (\alpha = \frac{1}{2}) \end{cases} \tag{2.78}$$

et :

$$\mathbf{R}_\alpha(P,Q) = \frac{1}{\alpha}\ \mathbf{I}_{h_\alpha}(P,Q) = \mathbf{R}_{1-\alpha}(Q,P)$$

La χ^2-divergence d'ordre α est la $\mathbf{I}_\alpha$-divergence de Csiszár [51, 52, 154]. Elle est identique à la « power divergence statistics » de [127]. On pourrait, pour des raisons évidentes, l'appeler aussi information d'ordre α de Havrda-Charvát (car elle correspond à cette entropie), et distance de Hellinger d'ordre α [153]. Cependant cette dernière terminologie est utilisée pour une autre divergence en [44]; on l'introduit comme exception un peu plus loin.

Information d'ordre α de Parzen D'introduction plus récente, par analogie avec l'expression de l'information de Rényi dans le cas de changement de variance en Gaussien [117], elle correspond à $f(u) = z_\alpha(u)$ où:

$$z_\alpha(u) = \begin{cases} \frac{1}{\alpha\ (\alpha-1)}\ [-(\alpha-1)\ \ln u - \ln(1-(\alpha-1)(u-1))_+] & (\alpha \neq 0,1) \\ -(\ln u - u + 1) = \breve{k}(u) & (\alpha = 1) \\ \ln u - 1 + \frac{1}{u} = \frac{r_0(u)}{u} = \frac{\bar{k}(u)}{u} & (\alpha = 0) \end{cases} \tag{2.79}$$

qui est telle que : $z_\alpha(1) = 0, f''_\alpha(1) = 1$. Donc :

$$\mathbf{I}_{z_\alpha}(P,Q) = \mathbf{Z}_\alpha(P,Q) = \begin{cases} & (\alpha \neq 0,1) \\ \bar{\mathbf{K}}(Q,P) & (\alpha = 1) \\ \mathbf{R}_2(P,Q) - \bar{\mathbf{K}}(P,Q) & (\alpha = 0) \\ 2\ \bar{\mathbf{K}}(P,Q) - 4\ \bar{\mathbf{K}}\left(P, \frac{P+Q}{2}\right) = 2\ \bar{\mathbf{K}}(P,Q) - 4\ K(P,Q) & (\alpha = \frac{1}{2}) \end{cases} \tag{2.80}$$

Noter que:

$$\mathbf{Z}_{1/2}(P,Q) + \mathbf{Z}_{1/2}(Q,P) = 2\ \mathbf{K}(P,Q) - 8\ \mathbf{J}_1(P,Q) = 2\ \mathbf{K}(P,Q) - 4\ L(P,Q) \geq 0$$

où $\mathbf{J}_1$ est définie en (4.32).

« Information d'ordre α de Bose-Einstein» Elle a été introduite très récemment par Knockaert [95] par le biais du remplacement de la loi multinômiale de la physique statistique de Maxwell-Boltzman qui intervient dans la définition de l'information de Kullback, par la loi de Bose-Einstein; ceci pour la résolution de problèmes de minimisation sous contraintes d'observation. (Noter que le lien entre le principe du maximum d'entropie et les distributions de Fermi-Dirac et Bose-Einstein est discuté en [68, 5] – voir plus haut en section 2.4).

Elle correspond à la fonction $f(u) = b_\alpha(u)$ pour $\alpha \in \mathbb{R}_+$, où :

$$b_\alpha(u) = u \ \ln u - (\alpha + u) \ \ln \frac{\alpha + u}{\alpha + 1} \tag{2.81}$$

$$\mathbf{I}_{b_\alpha}(P,Q) = \mathbf{B}_\alpha(P,Q) = \int \left(p \ln p + \alpha q \ln q - (\alpha q + p) \ \ln \frac{\alpha q + p}{\alpha + 1} \right) \ d\lambda(x) \tag{2.82}$$

C'est une différence de Jensen (2.22) :

$$\mathbf{B}_\alpha(P,Q) = (1+\alpha) \ \mathbf{J}_{H_1}^{(1/(1+\alpha))}(P,Q)$$

$$\begin{aligned}
\mathbf{B}_1(P,Q) &= L(P,Q) = 2 \ \mathbf{J}_1(P,Q) \\
\lim_{\alpha \to \infty} \mathbf{B}_\alpha(P,Q) &= \bar{\mathbf{K}}(P,Q) \quad \text{(voir (2.26) et (2.25))} \\
\mathbf{B}_{1/\alpha}(P,Q) &= \frac{1}{\alpha} \ \mathbf{B}_\alpha(P,Q) \\
\frac{\partial}{\partial \alpha} \mathbf{B}_\alpha(P,Q) &= \bar{\mathbf{K}} \left(Q, \frac{P + \alpha \, Q}{1+\alpha} \right) \\
\mathbf{B}_\alpha(P,Q) &\approx \bar{\mathbf{K}}(Q,P) \quad (\alpha \text{ petit})
\end{aligned}$$

Dans le cas d'un processus Gaussien centré, on retrouve la distance de Chernoff et l'information d'ordre α de Rényi (introduites avec les f-divergences non intégrales) :

$$\mathbf{B}_\alpha(S_1,S_2) = (1+\alpha) \ \mathbf{C}_{\alpha/(1+\alpha)}(S_1,S_2) = \frac{\alpha}{1+\alpha} \ \widetilde{\mathbf{R}}_{\alpha/(1+\alpha)}(S_1,S_2)$$

« Information d'ordre α de Fermi-Dirac » Elle a été introduite encore plus récemment par Knockaert [96] par le biais du remplacement de la loi multinômiale de la physique statistique de Maxwell-Boltzman par la loi de Fermi-Dirac et de l'utilisation des grandes déviations.

Elle correspond à la fonction $f(u) = f_\alpha(u)$ pour $\alpha \geq 1$, où :

$$f_\alpha(u) = u \ \ln u - (\alpha - u) \ \ln \frac{\alpha - u}{\alpha - 1} \tag{2.83}$$

$$\mathbf{I}_{f_\alpha}(P,Q) = \mathbf{F}_\alpha(P,Q) = \int \left(p \ln p - \alpha q \ln q - (\alpha q - p) \ \ln \frac{\alpha q - p}{\alpha - 1} \right) \ d\lambda(x) \tag{2.84}$$

f-divergences « naturelles » [96] On dit qu'une fonction $f : \mathbb{R}_+ \to \mathbb{R}$ est naturelle si $f(1) = 0$ et si la fonction miroir satisfait :

$$\breve{f}(u) = u f\left(\frac{1}{u}\right) = g_\alpha^*(u) \tag{2.85}$$

où :

$$\begin{aligned}
g_\alpha(u) &= \alpha \ g(u) \\
g(u) &= \ln B(e^u) \\
B(z) &= \gamma \ A(\beta z) = \gamma \ \sum_k \mu_k \beta^k z^k
\end{aligned}$$

Une telle fonction est strictement convexe et peut s'écrire sous la forme :

$$f(u) = \alpha u \ h^*\left(\frac{1}{\alpha u}\right) - \alpha \ h^*\left(\frac{1}{\alpha}\right) - \alpha \ (\ln \gamma) \ (u-1) \tag{2.86}$$

où :

$$\begin{aligned} h(u) &= \ln A(e^u) \\ &= \frac{1}{\alpha}\,(g_\alpha(u-\ln\beta)-\alpha\ \ln\gamma) \\ &= g(u-\ln\beta)-\ln\gamma \\ \ln\beta &= \alpha\, h^*\left(\frac{1}{\alpha}\right)-\alpha\ \ln\gamma \end{aligned}$$

Une condition nécessaire pour que f soit naturelle est que :

$$f(u)\approx au+b-\frac{1}{q}\ \ln u+O(u^{-1/q})\quad \text{quand}\ \ u\to\infty$$

Une f-divergence intégrale est dite naturelle si elle correspond à une fonction f naturelle.

De la condition nécessaire précédente, on déduit des contre-exemples, parmi les f-divergences : *la distance en variation de Kolmogorov* $\mathbf{V}$ *et la distance de Hellinger* $\mathcal{H}^2$ *ne sont pas des f-divergences naturelles.*

Exemples

– *Maxwell-Boltzman :*

$$\begin{aligned} A(z) &= e^z \\ h(u) &= \ln A(e^u)=e^u \\ h^*(u) &= u\ln u-u \\ f(u) &= -\ln u\quad (\text{pas de }\alpha!) \\ \mathbf{I}_f &= \bar{\mathbf{K}} \end{aligned}$$

$$\begin{aligned} g^*_\alpha(u) &= \breve{f}(u)=u\ln u \\ g_\alpha(u) &= e^{u-1}=h(u-1) \end{aligned}$$

– *Bose-Einstein :*

$$\begin{aligned} A(z) &= \frac{1}{1-z} \\ h(u) &= \ln A(e^u)=-\ln(1-e^u) \\ h^*(u) &= u\ln u-(u+1)\ \ln(u+1) \\ uh^*\left(\frac{1}{u}\right) &= h^*(u) \\ f(u) &= \alpha u\ \ln(\alpha u)-(\alpha u+1)\ \ln(\alpha u+1)-\alpha\ \ln\alpha+(\alpha+1)\ \ln(\alpha+1)-\alpha\ \ln\frac{\alpha}{\alpha+1}\,(u-1) \\ \mathbf{I}_f &= \mathbf{B}_\alpha \end{aligned}$$

– *Fermi-Dirac :*

$$\begin{aligned} A(z) &= 1+z \\ h(u) &= \ln A(e^u)=\ln(1+e^u) \\ h^*(u) &= u\ln u+(1-u)\ \ln(1-u)=\bar{h}_1(u) \\ uh^*\left(\frac{1}{u}\right) &= (u-1)\ln(u-1)-u\ln u \\ f(u) &= (\alpha u-1)\ \ln(\alpha u-1)-\alpha u\ \ln(\alpha u)-(\alpha-1)\ \ln(\alpha-1)+\alpha\ \ln\alpha-\alpha\ \ln\frac{\alpha-1}{\alpha}\,(u-1) \\ \mathbf{I}_f &= \mathbf{F}_\alpha \end{aligned}$$

2.5.1.3 Dualité

Une f-divergence peut, en tant que fonction d'une densité de probabilité, s'écrire comme une duale convexe. C'est du moins vrai pour l'information de Kullback. Le calcul de la fonction duale d'une f-divergence quelconque ne conduit pas à une formule explicite.

Duale de l'information de Kullback [59] Un calcul direct de Lagrangien permet de montrer que l'information de Kullback écrite sous la forme :

$$\bar{\mathbf{K}}(p) = \int p \ln p \, d\nu$$

(où $p = \frac{d\mu}{d\nu}$) admet comme duale convexe la fonction :

$$\bar{\mathbf{K}}^*(\phi) = \ln \int e^{\phi} \, d\nu$$

Autrement dit, l'information de Kullback admet la formulation variationnelle :

$$\bar{\mathbf{K}}(P,Q) = \sup_{\phi} \left(\mathbf{E}_P(\phi) - \ln \mathbf{E}_Q \left(e^{\phi} \right) \right)$$

due à Donsker-Varadhan.

Duale d'une f-divergence Plus généralement, le même calcul permet de montrer que toute f-divergence :

$$\mathbf{I}_f(p) = \int f(p) \, d\nu$$

admet comme fonction duale :

$$\mathbf{I}_f^*(\phi) = \int f^*(\phi - \lambda^*) \, d\nu + \lambda^*$$

où le multiplicateur λ^* est donné par :

$$\int f'^{-1}(\phi - \lambda^*) \, d\nu = 1$$

ou encore :

$$\int f^{*\prime}(\phi - \lambda^*) \, d\nu = 1$$

Dans le cas de l'information de Kullback, on a :

$$f(u) = u \ln u$$

et

$$f^*(u) = f'^{-1}(u) = e^{(u-1)} = f^{*\prime}(u)$$

ce qui permet un calcul explicite de λ^* et de $\mathbf{I}_f^*(\phi)$.

2.5.2 f-divergences non intégrales

2.5.2.1 Définition

On considère plus généralement comme distance toute fonction croissante d'une f-divergence intégrale :

$$\mathbf{I}_f(P,Q) = g\left(\mathbf{E}_Q \left(f\left(\frac{p}{q} \right) \right) \right) = g\left(\int q \, f\left(\frac{p}{q} \right) \, d\lambda(x) \right) \tag{2.87}$$

où f est continue convexe sur $[0, +\infty[$, et g est croissante sur $\mathbb{R}$. On impose de plus:

$$\begin{aligned} g(f(1)) &= 0 \qquad \text{pour garantir } \mathbf{I}_f(P,P) = 0 \\ g'(f(1)) \, f''(1) &> 0 \qquad \text{pour la métrique (5.13)} \end{aligned}$$

TAB. 2.3 – *Fonctions ψ , et leurs fonctions duale et inverse.*

nom	$\psi(u)$	$\psi^*(u)$	$\psi^{-1}(u)$
Bhattacharyya	$e^{u/2}$		$2\ln u$
Rényi ψ_α	$e^{-(\alpha-1)u}$	$\frac{u}{1-\alpha}\left(\ln\frac{u}{1-\alpha}-1\right)$	$\frac{1}{1-\alpha}\ln u$
Matusita	$\lvert e^{\alpha u}-1\rvert^{1/\alpha}$		$\frac{1}{\alpha}\ln(1-v^\alpha)$

TAB. 2.4 – *Fonctions ϕ , et leurs fonctions duale et inverse.*

nom	$\phi(u)$	$\phi^*(u)$	$\phi^{-1}(u)$
Bhattacharyya	$1/\sqrt{u}$		u^{-2}
Rényi ϕ_α	$u^{\alpha-1}$	$(\alpha-2)\left(\frac{u}{\alpha-1}\right)^{\frac{\alpha-1}{\alpha-2}}$	$u^{\frac{1}{\alpha-1}}$
Matusita	$\lvert u^{-\alpha}-1\rvert^{1/\alpha}$		$(1-v^\alpha)^{-1/\alpha}$

Cependant, comme on l'a déjà indiqué plus haut en (2.12), une f-divergence non intégrale se déduit d'une entropie non intégrale par :

$$\begin{aligned}\mathbf{I}_\psi(\mu,\nu) &= -\mathbf{H}_{\psi,\nu}(\mu)\\ \mathbf{I}_\phi(\mu,\nu) &= -\mathbf{H}_{\phi,\nu}(\mu)\end{aligned}$$

et ceci fournit des formes fonctionnelles plus pertinentes que la précédente :

$$\mathbf{I}_\psi(P,Q) = -\psi^{-1}\left(\int \frac{p}{q}\,\psi\left(-\ln\frac{p}{q}\right)\,q\,d\lambda(x)\right) \tag{2.88}$$

$$\mathbf{I}_\phi(P,Q) = \ln\phi^{-1}\left(\int \frac{p}{q}\,\phi\left(\frac{p}{q}\right)\,q\,d\lambda(x)\right) \tag{2.89}$$

Les relations entre f,ψ,ϕ sont données en général par la colonne de gauche du tableau 2.1.

Il en résulte qu'en fait les conditions imposées dans la première définition – à savoir f convexe et g croissante – sont trop fortes; il suffit de ψ (ou ϕ) strictement monotone. D'autre part, $\mathbf{I}_\psi(P,P)=\mathbf{I}_\phi(P,P)=0$ sans condition supplémentaire sur ψ,ϕ.

2.5.2.2 Exemples

On présente les exemples de f-divergences non intégrales résumés au tableau 2.5 et correspondant aux fonctions ψ,ϕ données aux tableaux 2.3 et 2.4 avec leurs fonctions miroir et duale convexe.

Distance et coefficient (ou affinité) de Bhattacharyya Le coefficient (ou affinité) de Bhattacharyya est défini par :

$$\rho(P,Q)=\int\sqrt{p\,q}\,d\lambda(x)=1-\mathcal{H}^2(P,Q)$$

et vérifie :

$$1-\rho^2=\mathcal{H}^2\,(1-\mathcal{H}^2)$$

La distance de Bhattacharyya est définie en [90, 92] comme :

$$\mathcal{B}=-\ln\rho=-\ln(1-\mathcal{H}^2)$$

TAB. 2.5 – *f-divergences non intégrales.*

nom	$f(u) = u\psi(-\ln u)$	$\psi(u)$	$g(u)$	$\mathbf{I}_f(P,Q)$
Bhatta-charyya	$\sqrt{u}$ $\sqrt{u}$	$e^{u/2}$ $e^{u/2}$	$\psi^{-1}(u)$ $\arccos u$	$\mathcal{B}(P,Q) = -2\ln \int \sqrt{p\,q}\; d\lambda(x)$ $\mathcal{B}(P,Q) = \arccos \int \sqrt{p\,q}\; d\lambda(x)$
Rényi	u^α	$e^{-(\alpha-1)u}$	$\frac{1}{\alpha}\psi_\alpha^{-1}(u)$	$\widetilde{\mathbf{R}}_\alpha(P,Q) = \frac{1}{\alpha\,(\alpha-1)}\ \ln \int p^\alpha\; q^{1-\alpha} d\lambda(x)$ $\widetilde{\mathbf{R}}_1(P,Q) = \bar{\mathbf{K}}(P,Q)$ $\widetilde{\mathbf{R}}_0(P,Q) = \bar{\mathbf{K}}(Q,P)$ $\widetilde{\mathbf{R}}_{-1}(P,Q) = \frac{1}{2}\ \ln(1+2\,\mathbf{R}(P,Q))$ $\widetilde{\mathbf{R}}_{1/2}(P,Q) = -4\ \ln(1-\mathcal{H}^2(P,Q))$
Chernoff	u^α	$e^{-(\alpha-1)u}$	$(\alpha-1)\ \psi_\alpha^{-1}(u)$	$\mathbf{C}_\alpha(P,Q) = -\ln \int p^\alpha\; q^{1-\alpha}\; d\lambda(x)$ $\mathbf{C}(P,Q) = \sup_\alpha \mathbf{C}_\alpha(P,Q)$
Matusita	$\lvert u^\alpha - 1\rvert^{1/\alpha}$	$\lvert e^{\alpha u}\rvert^{1/\alpha}$	u^α	$\mathbf{M}_\alpha(P,Q) = \left(\int \lvert p^\alpha - q^\alpha\rvert^{1/\alpha}\, d\lambda(x)\right)^\alpha$ $\mathbf{M}_1(P,Q) = \mathbf{V}(P,Q)$ $\mathbf{M}_2(P,Q) = \sqrt{2}\,\mathcal{H}(P,Q)$

i.e. sous la forme (2.88) avec $\psi(u) = e^{u/2}$.

Cependant en [61] cette distance est définie par :

$$\mathcal{B}(P,Q) = \arccos\left(\int \sqrt{p\,q}\; d\lambda(x)\right) = \arccos(1-\mathcal{H}^2(P,Q)) = \arccos \rho(P,Q)$$

Information d'ordre α de Rényi [130, 154, 117] Elle est dite aussi gain d'information d'ordre α [146]. Elle est définie, soit exactement sous la forme fonctionnelle (2.88) avec $\psi = \psi_\alpha$ donnée en (2.55), soit avec $g = \frac{1}{\alpha}\ \psi_\alpha^{-1}$:

$$\widetilde{\mathbf{R}}_\alpha(P,Q) = \begin{cases} \frac{1}{\alpha\,(\alpha-1)}\ \ln \int p^\alpha\; q^{1-\alpha}\; d\lambda(x) & (\alpha \neq 0,1) \\ \bar{\mathbf{K}}(P,Q) & (\alpha = 1) \\ \bar{\mathbf{K}}(Q,P) & (\alpha = 0) \\ \frac{1}{2}\ \ln(1+2\,\mathbf{R}(P,Q)) & (\alpha = 2) \\ -4\ \ln(1-\mathcal{H}^2(P,Q)) = -4\ \ln\left(1-\frac{1}{4}\,\mathbf{R}_{1/2}(P,Q)\right) = 4\,\mathcal{B}(P,Q) & (\alpha = \frac{1}{2}) \end{cases} \tag{2.90}$$

Cette définition avec le facteur multiplicatif supplémentaire $1/\alpha$, est celle de Parzen [117]; elle est justifiée par le cas Gaussien, pour lequel on retrouve la distance de Bhattacharrya (voir l'annexe).

On peut aussi l'écrire comme :

$$\widetilde{\mathbf{R}}_\alpha(P,Q) = \frac{1}{\alpha\,(\alpha-1)}\ \ln\ |\mathbf{I}_f(P,Q)|$$

pour :

$$f(u) = u^\alpha\ \mathrm{sgn}(\alpha-1)$$

L'information d'ordre α de Rényi est liée à la χ^2-divergence d'ordre α par (voir (2.47)) :

$$\widetilde{\mathbf{R}}_\alpha \quad = \quad \frac{1}{\alpha\,(\alpha-1)}\ \ln\left[1+\alpha\,(\alpha-1)\,\mathbf{R}_\alpha\right]$$

$$\begin{aligned}\mathbf{R}_\alpha &= \frac{1}{\alpha\,(\alpha-1)}\,\left(e^{\alpha\,(\alpha-1)\,\widetilde{\mathbf{R}}_\alpha}-1\right)\\ \widetilde{\mathbf{R}}_\alpha(P,Q) &= \widetilde{\mathbf{R}}_{1-\alpha}(Q,P)\end{aligned}$$

Enfin, rappelons que l'information et l'entropie d'ordre α de Rényi sont liées par :

$$\widetilde{\mathbf{R}}_\alpha(\mu,\nu) = -\widetilde{\mathbf{H}}_{\alpha,\nu}(\mu)$$

On notera [52, 56] qu'elle est liée à la fonction caractéristique du logarithme du rapport de vraisemblance $\ln\frac{P(X)}{Q(X)}$, c'est-à-dire à la fonction génératrice des moments, par :

$$\mathbf{E}_P\left(e^{\alpha\,\ln\frac{P(X)}{Q(X)}}\right) = e^{\alpha\,(1+\alpha)\,\widetilde{\mathbf{R}}_{1+\alpha}(P,Q)} \tag{2.91}$$

On en déduit que :

$$\ln\,\mathbf{E}_P\left(e^{\alpha\,\ln\frac{P(X)}{Q(X)}}\right) = \alpha\,(1+\alpha)\,\widetilde{\mathbf{R}}_{1+\alpha}(P,Q)$$

autrement dit que l'information de Rényi n'est autre, à une constante près, que la seconde fonction caractéristique, ou fonction génératrice des cumulants, du logarithme du rapport de vraisemblance $\ln\frac{P(X)}{Q(X)}$. Ceci laisse entrevoir son rôle en test d'hypothèses et classification pour le calcul de probabilités d'erreur [28].

Distance et coefficient (ou nombre d'information) de Chernoff La distance de Chernoff correspond [49] à $g(u) = (\alpha-1)\,\psi_\alpha^{-1}(u)$ $(0<\alpha<1)$, et s'écrit donc :

$$\begin{aligned}\mathbf{C}_\alpha(P,Q) &= -\ln\int p^\alpha\,q^{1-\alpha}\,d\lambda(x)\\ &= \alpha\,(1-\alpha)\,\widetilde{\mathbf{R}}_\alpha(P,Q)\end{aligned}$$

Pour $\alpha = 1/2$, il s'agit de la distance de Bhattacharya, et sa dérivée à l'origine (en α) est l'information de Kullback :

$$\begin{aligned}\mathbf{C}_{1/2}(P,Q) &= -\ln(1-\mathcal{H}^2(P,Q)) = \mathcal{B}(P,Q)\\ \frac{\partial}{\partial\alpha}\,\mathbf{C}_\alpha(P,Q)\Big|_{\alpha=0} &= \bar{\mathbf{K}}(Q,P)\quad[91,\ 92]\end{aligned}$$

Noter que l'argument du logarithme est aussi appelé arc de Hellinger par Le Cam.

Le coefficient de Chernoff est défini par [52, 49] :

$$\mathbf{C}(P,Q) = \sup_{0<\alpha<1}\,\mathbf{C}_\alpha(P,Q)$$

et intervient dans les performances des algorithmes de classification à $m>2$ classes [52].

Distance d'ordre α de Matusita Elle a en fait été introduite par Jeffreys en 1948! Elle correspond à $g(u) = u^\alpha; f(u) = m_\alpha(u)$ $(\alpha\le 1)$, où:

$$m_\alpha(u) = |u^\alpha - 1|^{1/\alpha} \tag{2.92}$$

est stable par miroir. Elle n'est donc pas de la forme (2.88). Cette distance s'écrit :

$$\mathbf{I}_{m_\alpha}(P,Q) = \mathbf{M}_\alpha(P,Q) = \begin{cases}\left(\int|p^\alpha - q^\alpha|^{1/\alpha}\,d\lambda(x)\right)^\alpha & (\alpha\neq 1,1/2)\\ \mathbf{V}(P,Q) & (\alpha=1)\\ \sqrt{2}\,\mathcal{H}(P,Q) & (\alpha=1/2)\end{cases} \tag{2.93}$$

2.5.3 Exception : divergence de Hellinger d'ordre α

En [44], la divergence de Hellinger est définie comme étant :

$$\mathcal{H}^2_\psi(P,Q) = \int (\psi(q) - \psi(p))^2 \; d\lambda(x) \tag{2.94}$$

pour n'importe quelle fonction ψ de classe $\mathcal{C}^2$.

Le calcul de la métrique associée (5.15) permet de montrer que cette divergence de Hellinger généralisée n'est une f-divergence que si $\psi(u) = \sqrt{2u}$, i.e. pour la distance de Hellinger usuelle.

Toujours en [44], la *divergence de Hellinger d'ordre* α est définie comme correspondant à la fonction :

$$\psi(u) = \frac{1}{\alpha\sqrt{2}} \; u^{\alpha/2}$$

soit :

$$\mathcal{H}^2_\alpha(P,Q) = \frac{1}{2\,\alpha^2} \int (q^{\alpha/2} - p^{\alpha/2})^2 \; d\lambda(x) \tag{2.95}$$

Bien qu'elle ne soit pas une f-divergence, elle présente l'intérêt de *fournir la même métrique que l'entropie d'ordre* α *de Havrda-Charvát*, ainsi qu'on en discute à la section 5.1.

2.5.4 Lien avec l'information de Fisher et l'exhaustivité

On montre maintenant en quoi les f-divergences sont, dans le cas de densités de probabilité *paramétrées*, liées étroitement à l'exhaustivité et à l'information de Fisher. On commence par définir celle-ci.

2.5.4.1 Information de Fisher

Dans le cas d'une densité de probabilité pour laquelle on note $l_\theta(x) = \ln p_\theta(x)$, l'information de Fisher est définie par :

$$\begin{aligned} g_{ij}(\theta) &= + \mathbf{E}_\theta(\partial_i l_\theta(x) \; \partial_j l_\theta(x)) \qquad (2.96) \\ &= \int p_\theta \; (\partial_i \ln p_\theta) \; (\partial_j \ln p_\theta) \; d\lambda(x) \\ &= \int \frac{1}{p_\theta} \; (\partial_i p_\theta) \; (\partial_j p_\theta) \; d\lambda(x) \\ &= 4 \int (\partial_i \sqrt{p_\theta}) \; (\partial_j \sqrt{p_\theta}) \; d\lambda(x) \qquad (2.97) \end{aligned}$$

et peut aussi s'écrire comme

$$g_{ij}(\theta) = - \mathbf{E}_\theta(\partial_i \partial_j l_\theta(x))$$

Dans le cas d'une densité spectrale :

$$g_{ij}(\theta) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \partial_i \ln S(\omega,\theta) \; \partial_j \ln S(\omega,\theta) \; d\omega$$

2.5.4.2 Exhaustivité et information

Localement, pour $P = P_\theta$ et $Q = P_{\theta+\delta\theta}$, on retrouve la métrique de Fisher (5.1) [86, 35, 89]:

$$\mathbf{R}(\theta,\theta+\delta\theta) \approx \bar{\mathbf{K}}(\theta,\theta+\delta\theta) \approx 4\,\mathcal{H}^2(\theta,\theta+\delta\theta) \approx \frac{1}{2} \sum_{i,j} g_{ij}(\theta) \; \delta\theta_i \; \delta\theta_j$$

Une statistique exhaustive conserve l'information de Fisher (Fisher dixit) et les distances de Kullback [100, 33], de Hellinger et du χ^2 (Le Cam). Plus généralement, elle conserve toute f-divergence [11, 97] et l'information mutuelle $\mathcal{I}_f$ associée [52].

En fait, on montrera plus loin à la section 5.1 que *la métrique associée à toute* f*-divergence est la métrique de Fisher*, qui possède l'importante propriété d'être invariante par changement de paramétrisation et par changement de variable.

2.5.4.3 Information et sensibilité - Information de Fisher d'ordre α

Dans le cas d'un paramètre scalaire, a été notée [153] l'importance de la prise en compte du signe de la petite déviation : ce point de vue (de Rao 1962) interprète Fisher comme mesure d'information à l'aide de la sensibilité d'une loi paramétrée à des déviations faibles du paramètre, sensibilité exprimée en terme de divergence. De ce point de vue, toutes les f-divergences ne sont *pas* équivalentes; par exemple la distance en variation de Kolmogorov a le mauvais goût de permettre des changements de signe.

Pour pallier ce genre de limitation, a été introduite en [153] l'information de Fisher d'ordre $\alpha \geq 1$ qui est définie par :

$$\begin{aligned} g^{(\alpha)}(\theta) &= \lim \inf_{\delta\theta \to 0} \frac{\mathbf{W}_\alpha(\theta, \theta+\delta\theta)}{|\delta\theta|^\alpha} \\ &= \mathbf{E}_\theta(|\partial l_\theta|^\alpha) \end{aligned}$$

Une statistique exhaustive conserve cette information [153].

Une borne de Cramer-Rao généralisée en résulte sous la forme :

$$\left(g^{(\alpha)}(\theta)\right)^{1/\alpha} \left(\mathbf{E}_\theta\left(|T - b(\theta)|^{\alpha/(\alpha-1)}\right)\right)^{(\alpha-1)/\alpha} \geq b'(\theta)$$

2.5.5 Inégalités

[109, 58, 106, 105]

$$\begin{aligned}
e^{-\bar{\mathbf{K}}/2} &\leq \rho \\
\mathcal{H}^2 &\leq \mathbf{V} \leq \mathcal{H}\ \sqrt{2-\mathcal{H}^2} \\
\frac{1}{4}\, e^{\bar{\mathbf{K}}} &\leq 1 - \mathbf{V} \leq \rho \\
\bar{\mathbf{K}} &\geq \mathbf{V} - \ln(1+\mathbf{V}) \\
\bar{\mathbf{K}} &\geq \frac{\mathbf{V}^2}{2} + \frac{\mathbf{V}^4}{36} + \frac{\mathbf{V}^6}{288} \qquad (0 \leq \mathbf{V} \leq 2) \\
\bar{\mathbf{K}} &\geq \ln \frac{2+\mathbf{V}}{2-\mathbf{V}} - \frac{2\,\mathbf{V}}{2+\mathbf{V}} \qquad (0 \leq \mathbf{V} \leq 2) \\
\mathbf{K} &\geq \mathbf{V}\ \ln \frac{2+\mathbf{V}}{2-\mathbf{V}} \qquad (0 \leq \mathbf{V} \leq 2) \\
\mathbf{K} &\geq 4\,\mathcal{H}^2 \\
\mathbf{K} &\geq 4\,\mathcal{B} \\
\mathbf{T} &\geq e^{-\mathbf{K}/2} \\
\mathbf{T} &\geq 1 - \mathbf{K}/4 \\
\mathbf{K} &\geq 2\,\sqrt{1-\mathbf{T}}\ \ln \frac{1+\sqrt{1-\mathbf{T}}}{1-\sqrt{1-\mathbf{T}}} \\
1 - \mathbf{V}^2 &\leq \mathbf{T} \leq 1 - \mathbf{V} \\
\rho^2 &\leq \mathbf{T} \leq \rho \\
\mathbf{J}_1 &\leq \mathbf{K}/4 \\
\mathbf{J}_1 &\leq \mathbf{V} \\
\mathbf{R}_\alpha &\leq \frac{2}{\alpha(1-\alpha)}\,\mathcal{H}^2 \\
\mathcal{H}^2 &\leq \frac{\alpha}{2}\,\mathbf{R}_\alpha \qquad \left(\alpha \geq \frac{1}{2}\right)
\end{aligned}$$

où $\mathbf{J}_1$ est la $\mathbf{J}$-divergence (4.32) correspondant à l'entropie de Shannon.

On constate en particulier que *la distance de Kullback est plus fine que celle de Hellinger.* Il en est de même des χ^2-divergences d'ordre α ou $\mathbf{I}_\alpha$-divergences de Csiszár d'ordre $\alpha \geq 1/2$.

Chapitre 3

Axiomatique - Caractérisation - Équations fonctionnelles

On renvoie à [130, 3, 109, 55, 152], [154, chap10]; et aussi [1, 2].

3.1 Axiomes

Les propriétés que l'on peut souhaiter imposer comme axiomes concernant les entropies sont nombreuses [3, 109]. Notant $I_n(p_1, \ldots, p_n)$ l'entropie d'une distribution discrète à n masses ($\sum_i p_i = 1$), citons seulement :

- symétrie (ou invariance par permutation),
- normalisation :
$$I_2\left(\frac{1}{2}, \frac{1}{2}\right) = 1$$
- expansiblité :
$$I_n(p_1, \ldots, p_n) = I_{n+1}(p_1, \ldots, p_k, 0, p_{k+1}, \ldots, p_n)$$
- récursivité :
$$I_n(p_1, \ldots, p_n) = I_{n-1}(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2) I_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)$$
- additivité forte :
$$I_{mn}(p_1 q_{11}, \ldots, p_1 q_{1n}, \ldots, p_m q_{m1}, \ldots, p_m q_{mn}) = I_m(p_1, \ldots, p_m) + \sum_{j=1}^{m} p_j I_n(q_{j1}, \ldots, q_{jn})$$
- additivité :
$$I_{mn}(p_1 q_1, \ldots, p_1 q_n, \ldots, p_m q_1, \ldots, p_m q_n) = I_m(p_1, \ldots, p_m) + I_n(q_1, \ldots, q_n)$$
- sous-additivité :
$$I_{mn}(p_{11}, \ldots, p_{1n}, \ldots, p_{m1}, \ldots, p_{mn}) \leq I_m\left(\sum_{k=1}^{n} p_{1k}, \ldots, \sum_{k=1}^{n} p_{mk}\right) + I_n\left(\sum_{j=1}^{m} p_{j1}, \ldots, \sum_{j=1}^{m} p_{jn}\right)$$
- maximalité :
$$I_n(p_1, \ldots, p_n) \leq I_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$
- monotonie : $I_2(1-p, p)$ non-décroissante sur $[0, 1/2]$,

– propriété de branchement : il existe une suite de fonctions J_n telle que :

$$I_n(p_1,\ldots,p_n) - I_{n-1}(p_1+p_2,p_3,\ldots,p_n) = J_n(p_1,p_2)$$

Noter qu'il suffit de choisir :

$$J_n(p_1,p_2) = (p_1+p_2)\; I_2\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right)$$

Si en particulier J_n est de la forme $J_n(p_1,p_2) = g(p_1)+g(p_2)-g(p_1+p_2)$ et $I_2(1-p,p) = g(1-p)+g(p)$, alors on a la propriété suivante :

– propriété de somme (ou forme intégrale) ;

$$I_n(p_1,\ldots,p_n) = \sum_{k=1}^{n} g(p_k)\,.$$

– propriété de quasilinéarité (voir les moyennes généralisées) : il existe une fonction w positive sur $]0,1[$ et une fonction strictement monotone ψ telles que :

$$I_n(p_1,\ldots,p_n) = \psi^{-1}\left(\frac{\sum_{k=1}^{n} w(p_k)\psi(-\ln p_k)}{\sum_{k=1}^{n} w(p_k)}\right)$$

Comme expliqué longuement en [3], il est difficile, à partir de sous-ensembles de ces axiomes, de ne pas retrouver l'entropie de Shannon.

Il est à noter cependant que certaines formes de fonctions de coût permettent de mettre en évidence l'intérêt de l'entropie de Rényi pour le codage [46, 27].

Voir [4] pour une caractérisation intéressante des combinaisons linéaires de Shannon et Hartley (ou Rényi d'ordre 0).

3.2 Équations fonctionnelles concernant les entropies intégrales

On considère ici essentiellement l'entropie d'ordre α de Havrda-Charvát, associée à la fonction h_α définie en (2.45), dont on remarque qu'elle vérifie :

$$\begin{aligned} h_\alpha(uv) &= uh_\alpha(v) + vh_\alpha(u) + (\alpha-1)h_\alpha(u)h_\alpha(v) \\ h_1(uv) &= uh_1(v) + vh_1(u) \end{aligned}$$

3.2.1 Sur la fonction miroir

[27] On s'intéresse d'abord aux équations fonctionnelles concernant la fonction :

$$\breve{h}_\alpha(u) = u\; h_\alpha\left(\frac{1}{u}\right)$$

qui vérifie donc :

$$\breve{h}_\alpha(uv) = \breve{h}_\alpha(u) + \breve{h}_\alpha(v) + (\alpha-1)\breve{h}_\alpha(u)\breve{h}_\alpha(v)$$

On en déduit que la fonction :

$$\breve{h}_\alpha(u) = 1 + (\alpha-1)\breve{h}_\alpha(u)$$

vérifie :

$$\breve{h}_\alpha(uv) = \breve{h}_\alpha(u)\breve{h}_\alpha(v)$$

et donc :

$$\breve{h}_\alpha(u) = u^\kappa$$

où nécessairement $\kappa = 1-\alpha$. D'où résulte la propriété (2.47).

3.2.2 Sur la fonction d'information

[3, 109] On considère maintenant les équations fonctionnelles concernant la fonction :

$$\bar{h}(p) = \mathbf{H}(p, 1-p)$$

où $\mathbf{H}(p, 1-p)$ est une entropie de lois discrètes à 2 masses; autrement dit, la fonction :

$$\bar{h}(u) = -h(u) - h(1-u)$$

Remarquer qu'avec la convention de (2.41) :

$$\bar{\bar{h}}(u) = \gamma\ \bar{h}(u) - \delta$$

Entropie de Shannon Typiquement, la fonction d'information :

$$\bar{h}(u) = \bar{h}_1(u) = -h_1(u) - h_1(1-u) = -u \ln u - (1-u)\ \ln(1-u)$$

est, en vertu de la propriété de symétrie de l'information mutuelle, la solution de l'équation fonctionnelle (dite [3] *équation fondamentale de l'information*) :

$$\bar{h}(u) + (1-u)\ \bar{h}\left(\frac{v}{1-u}\right) = \bar{h}(v) + (1-v)\ \bar{h}\left(\frac{u}{1-v}\right) \tag{3.1}$$

qui vérifie :

$$\bar{h}(0) = \bar{h}(1), \quad \bar{h}\left(\frac{1}{2}\right) = \ln 2 \tag{3.2}$$

Il est à noter [3] que la fonction :

$$\mathcal{H}_1(u, v) = (u+v)\ \bar{h}_1\left(\frac{v}{u+v}\right)$$

qui redonne $\bar{h}_1$ par l'opération :

$$\bar{h}_1(u) = \mathcal{H}_1(1-u, u)$$

vérifie :

$$\mathcal{H}_1(u, v) = h_1(u+v) - h_1(u) - h_1(v)$$

et plus généralement, la fonction :

$$\mathcal{H}_1^{(\beta)}(u, v) = (\beta u + (1-\beta)v)\ \bar{h}_1\left(\frac{(1-\beta)v}{\beta u + (1-\beta)v}\right)$$

est une différence de Jensen :

$$\mathcal{H}_1^{(\beta)}(u, v) = h_1(\beta u + (1-\beta)v) - \beta\ h_1(u) - (1-\beta)\ h_1(v)$$

Les autres propriétés de la fonction $\mathcal{H}_1(u, v)$ sont :

$$\begin{aligned} \mathcal{H}_1(u, v)) &= \mathcal{H}_1(v, u) \\ \mathcal{H}_1(u+v, w) + \mathcal{H}_1(u, v) &= \mathcal{H}_1(u, v+w) + \mathcal{H}_1(v, w) \\ \mathcal{H}_1(wu, wv) &= w\ \mathcal{H}_1(u, v) \end{aligned} \tag{3.3}$$

ainsi que la normalisation :

$$\mathcal{H}_1\left(\frac{1}{2}, \frac{1}{2}\right) = \ln 2$$

Réciproquement [3], une fonction g est solution de l'équation fonctionnelle (3.1) si et seulement s'il existe une fonction ϕ telle que :

$$\begin{aligned} \phi(uv) &= \phi(u) + \phi(v), \quad \phi\left(\frac{1}{2}\right) = \ln 2 \\ g(u) &= u\ \phi(u) + (1-u)\ \phi(1-u) \end{aligned}$$

En effet, la fonction :

$$\mathcal{G}(u,v) = (u+v)\ g\left(\frac{v}{u+v}\right)$$

qui redonne g par :

$$g(u) = \mathcal{G}(1-u,u)$$

a les propriétés (3.3) et peut donc s'écrire sous la forme :

$$\mathcal{G}(u,v) = h(u+v) - h(u) - h(v)$$

Il suffit alors de prendre :

$$\phi(u) = -\frac{h(u)}{u}$$

Plus généralement [109], l'équation fonctionnelle généralisée :

$$F(u) + (1-u)\ G\left(\frac{v}{1-u}\right) = H(v) + (1-v)\ K\left(\frac{u}{1-v}\right) \tag{3.4}$$

admet comme solution:

$$\begin{aligned} F(u) &= A\,\bar{h}_1(u) + B_1 u + D \\ G(u) &= A\,\bar{h}_1(u) + B_2 u + B_1 - B_4 \\ H(u) &= A\,\bar{h}_1(u) + B_3 u + B_1 + B_2 - B_3 - B_4 + D \\ K(u) &= A\,\bar{h}_1(u) + B_4 u + B_3 - B_2 \end{aligned}$$

Entropie de Havrda-Charvát De même, la fonction :

$$\bar{h}(u) = \bar{h}_\alpha(u) = -h_\alpha(u) - h_\alpha(1-u) = -\frac{1}{\alpha-1}\left(u^\alpha + (1-u)^\alpha - 1\right)$$

qui vérifie [27] :

$$\begin{aligned} \bar{h}_3(u) &= \frac{3}{4}\,\bar{h}_2(u) \\ \bar{h}_\alpha(u) &\neq \bar{h}_{\alpha-1}(u) \quad (\alpha \geq 4) \end{aligned}$$

est la solution de l'équation fonctionnelle :

$$\bar{h}(u) + (1-u)^\alpha\ \bar{h}\left(\frac{v}{1-u}\right) = \bar{h}(v) + (1-v)^\alpha\ \bar{h}\left(\frac{u}{1-v}\right)$$

qui vérifie les mêmes conditions initiales (3.2). La solution générale de cette équation fonctionnelle est :

$$A\,\bar{h}_\alpha(u) + B\ u^\alpha$$

Il est à noter [3] que la fonction :

$$\mathcal{H}_\alpha(u,v) = (u+v)^\alpha\ \bar{h}_\alpha\left(\frac{v}{u+v}\right)$$

qui redonne $\bar{h}_\alpha$ par l'opération :

$$\bar{h}_\alpha(u) = \mathcal{H}_\alpha(1-u,u)$$

vérifie :

$$\mathcal{H}_\alpha(u,v) = h_\alpha(u+v) - h_\alpha(u) - h_\alpha(v)$$

Plus généralement [109], l'équation fonctionnelle généralisée :

$$F(u) + (1-u)^\alpha\ G\left(\frac{v}{1-u}\right) = H(v) + (1-v)^\alpha\ K\left(\frac{u}{1-v}\right)$$

admet comme solution:

$$\begin{aligned} F(u) &= A\,\bar{h}_\alpha(u) + d_1\,u^\alpha - c_2\,(1-u)^\alpha + c_1 \\ G(u) &= A\,\bar{h}_\alpha(u) + d_2\,u^\alpha + c_2 \\ H(u) &= A\,\bar{h}_\alpha(u) + d_2\,u^\alpha - c_4\,(1-u)^\alpha + c_1 \\ K(u) &= A\,\bar{h}_\alpha(u) + d_1\,u^\alpha + c_4 \end{aligned}$$

Noter que cette équation est trompeuse et d'intérêt limité: voir plus loin l'équation fonctionnelle associée aux f-divergences qui peut être résolue à l'aide de la solution de (3.4).

3.3 Caractérisation de l'entropie quadratique

[102] Pour $\mathcal{X}$ Hausdorff dénombrable:

$$\mathbf{Q} \text{ entropie quadratique} \Leftrightarrow (\forall k)\ \ \mathbf{Q}_k \text{ non-négative}$$

où $\mathbf{Q}_k$ est la différence de Jensen (2.22) d'ordre k de Q:

$$\mathbf{Q}_1 = \mathbf{J}_Q^{(1/2)},\ \ \mathbf{Q}_k = \mathbf{J}_{Q_{k-1}}^{(1/2)}$$

Remarque: dans le cas multinomial, l'entropie $\mathbf{H}_\alpha$ (2.49) de Havrda-Charvát n'a, selon α, que deux différences de Jensen non-négatives [41, 123]. La seule valeur de α pour laquelle plus de deux différences de Jensen non-négatives est $\alpha = 2$ [43], qui correspond à l'indice de diversité de Gini, et à la concentration de Lorenz et la χ^2-divergence (voir section 2.4).

3.4 Équations fonctionnelles concernant les f-divergences intégrales

[130, 3, 109, 83, 145, 84, 55, 152].

Les équations fonctionnelles données en [3, 109] concernent la fonction:

$$\bar{f}(p,q) = \mathbf{D}(p, 1-p; q, 1-q)$$

où $\mathbf{D}(p, 1-p; q, 1-q)$ est une divergence entre lois discrètes à 2 masses.

Pour $\mathbf{D} = \mathbf{I}_f$, il s'agit de la fonction:

$$\bar{f}(u,v) = v\ f\left(\frac{u}{v}\right) + (1-v)\ f\left(\frac{1-u}{1-v}\right)$$

Remarquer que, avec les conventions de (2.66) et (2.67):

$$\begin{aligned} \bar{\breve{f}}(u,v) &= \bar{f}(u,v) + \gamma + \delta \\ \bar{\breve{f}}(u,v) &= \bar{f}(v,u) \end{aligned}$$

et donc noter que, puisque $f = \breve{h}$, on a $\bar{f}(u,v) = \bar{h}(v,u)$.

3.4.1 Information de Kullback

Typiquement, pour:

$$f(u) = u \ln u = h_1(u)$$

la fonction $\bar{f}$ s'écrit:

$$\bar{f}(u,v) = u\ \ln\frac{u}{v} + (1-u)\ \ \ln\frac{1-u}{1-v}$$

et est la solution de l'équation fonctionnelle [3, 109]:

$$\begin{aligned} \bar{f}(u,x) + (1-u)\,\bar{f}\left(\frac{v}{1-u}, \frac{y}{1-x}\right) &= \bar{f}(v,y) + (1-v)\,\bar{f}\left(\frac{u}{1-v}, \frac{x}{1-y}\right) \\ &= \bar{f}(u+v, x+y) + (u+v)\,\bar{f}\left(\frac{v}{u+v}, \frac{y}{x+y}\right) \end{aligned} \tag{3.5}$$

qui vérifie :

$$\bar{f}(0,0) = \bar{f}(1,1), \quad \bar{f}\left(1, \frac{1}{2}\right) = \ln 2$$

La résolution de (3.5) se ramène à celle de (3.4) en considérant, pour x, y fixés, les fonctions :

$$\begin{aligned} F(u) &= \bar{f}(u,x) \\ G\left(\frac{v}{1-u}\right) &= \bar{f}\left(\frac{v}{1-u}, \frac{y}{1-x}\right) \\ H(v) &= \bar{f}(v,y) \\ H\left(\frac{u}{1-v}\right) &= \bar{f}\left(\frac{u}{1-v}, \frac{x}{1-y}\right) \end{aligned}$$

3.4.2 χ^2-divergence d'ordre α (Havrda-Charvát)

De même, pour :

$$f(u) = h_\alpha(u)$$

la fonction $\bar{f}$ est, à un coefficient multiplicatif près :

$$\bar{f}_\alpha(u,v) = u^\alpha v^{1-\alpha} + (1-u)^\alpha (1-v)^{1-\alpha} - 1$$

et est la solution de l'équation fonctionnelle [109] :

$$\bar{f}(u,x) + (1-u)^\alpha (1-x)^{1-\alpha}\ \bar{f}\left(\frac{v}{1-u}, \frac{y}{1-x}\right) = \bar{f}(v,y) + (1-v)^\alpha (1-y)^{1-\alpha}\ \bar{f}\left(\frac{u}{1-v}, \frac{x}{1-y}\right)$$

qui vérifie :

$$\bar{f}(0,0) = \bar{f}(1,1), \quad \bar{f}\left(0, \frac{1}{2}\right) = \bar{f}\left(1, \frac{1}{2}\right) = 2^{-\alpha} - 1$$

Plus généralement, l'équation fonctionnelle généralisée :

$$F(u,x) + (1-u)^\alpha (1-x)^\gamma\ F\left(\frac{v}{1-u}, \frac{y}{1-x}\right) = F(v,y) + (1-v)^\alpha (1-y)^\gamma\ F\left(\frac{u}{1-v}, \frac{x}{1-y}\right)$$

admet comme solution :

$$F(u,v) = bu^\alpha v^\gamma - A(u^\alpha v^\gamma + (1-u)^\alpha (1-v)^\gamma - 1)$$

Pour $\gamma = 1-\alpha$, on obtient $F = \bar{f}_\alpha$ ci-dessus définie.

On renvoie à [130, 3, 109, 55, 152] pour les f-divergences non intégrales.

Chapitre 4

Moyennes, mélanges et extensions de divergences

On décrit à la section 4.1 les moyennes sous-jacentes aux entropies et divergences, en distinguant pour ces dernières des définitions explicites et implicites de moyennes. On montre ensuite à la section 4.2 qu'il est possible d'étendre des procédés constructifs de divergences en jouant sur ces notions de moyennes sous-jacentes, et que ceci constitue un moyen d'introduire le rayon d'information ou l'exposant de codage de Gallager.

4.1 Moyennes généralisées et projections

On met d'abord en évidence les moyennes arithmétique et géométrique sous-jacentes à l'entropie de Shannon, et les deux moyennes d'ordre α sous-jacentes à l'entropie de Rényi. On introduit ensuite les moyennes généralisées sous-jacentes aux quatre formes fonctionnelles d'entropie, et on discute leurs propriétés d'invariance. On explicite enfin les relations entre les quatre formes fonctionnelles. Puis on montre que les moyennes généralisées explicitement sous-jacentes aux quatre formes fonctionnelles de f-divergences sont les mêmes que pour les entropies. On donne ensuite la définition de moyennes dites entropiques associées implicitement aux f-divergences intégrales. On donne aussi une définition implicite des moyennes généralisées non intégrales, qui est basée sur la distance de Bregman. Enfin on discute les propriétés d'invariance de ces diverses moyennes.

Dans ce qui suit, on désigne une suite finie $a_1, \ldots, a_n$ par la notation abrégée a lorsque celle-ci n'est pas ambigüe. D'autre part, on considère des poids β_i strictement positifs et normalisés : $\sum_{i=1}^n \beta_i = 1$

4.1.1 Moyennes sous-jacentes aux entropies

On considère d'abord des lois de probabilité discrètes à n masses $p, \ldots, p_n$ ($\sum_{i=1}^n p_i = 1$).

4.1.1.1 Entropie de Shannon et moyennes

L'écriture de l'entropie de Shannon sous la forme :

$$\mathbf{H}_1(p) = \sum_{i=1}^n p_i \ \ln \frac{1}{p_i}$$

montre que la forme intégrale (2.5) correspond à la moyenne arithmétique des $\ln \frac{1}{p_i}$ pondérés par les p_i [130]. Autrement dit, l'opérateur de moyenne qui, *dans l'espace des logarithmes des probabilités*, est sous-jacent à l'entropie de Shannon, est le barycentre [130, 146, 52] :

$$\Sigma_1^{(\beta)}(u) = \Sigma_{h_1}^{(\beta)}(u) = \sum_{i=1}^n \beta_i \ u_i \stackrel{\Delta}{=} A^{(\beta)}(u) \tag{4.1}$$

Le cas de poids égaux $\beta_i = \frac{1}{n}$ correspond à la *moyenne arithmétique* :

$$\Sigma_1^{(1/n)}(u) = \frac{1}{n} \ \sum_{i=1}^n \ u_i$$

De même, l'écriture sous la forme :

$$\mathbf{H}_1(p) = -\ln \prod_{i=1}^{n} {p_i}^{p_i}$$

montre que, dans la forme (2.3), la quantité $\mathbf{G}_{\phi_1}$ s'écrit :

$$\mathbf{G}_{\phi_1}(p) = \prod_{i=1}^{n} {p_i}^{p_i}$$

et correspond à la moyenne géométrique des p_i pondérés aussi par les p_i eux-mêmes. Autrement dit, l'opérateur de moyenne qui, *dans l'espace des probabilités*, est sous-jacent à l'entropie de Shannon est :

$$\Sigma_{\phi_1}^{(\beta)}(p) = \prod_{i=1}^{n} p_i^{\beta_i} \triangleq G^{(\beta)}(p) \tag{4.2}$$

Le cas de poids égaux $\beta_i = \frac{1}{n}$ correspond à la *moyenne géométrique* :

$$G^{(1/n)}(p) = \sqrt[n]{\prod_{i=1}^{n} p_i}$$

4.1.1.2 Entropie de Rényi et moyennes

L'écriture de l'entropie de Rényi sous la forme :

$$\widetilde{\mathbf{H}}_\alpha(p) = \frac{1}{1-\alpha} \ln \sum_{i=1}^{n} p_i^\alpha = -\frac{1}{\alpha-1} \ln \sum_{i=1}^{n} p_i \, e^{-(\alpha-1) \ln \frac{1}{p_i}}$$

montre que la forme non intégrale (2.9) correspond à la moyenne exponentielle des $\ln \frac{1}{p_i}$ pondérés par les p_i [130]. Autrement dit, l'opérateur de moyenne qui, *dans l'espace des logarithmes des probabilités*, est sous-jacent à l'entropie de Rényi, est [130, 146, 52] :

$$\Sigma_{\psi_\alpha}^{(\beta)}(u) = -\frac{1}{\alpha-1} \ln \sum_{i=1}^{n} \beta_i \, e^{-(\alpha-1) u_i} \tag{4.3}$$

Cette moyenne admet comme cas particuliers la moyenne arithmétique et le supremum :

$$\begin{aligned} \lim_{\alpha \to 1} \Sigma_{\psi_\alpha}^{(\beta)}(u) &= \Sigma_1^{(\beta)}(u) = \sum_{i=1}^{n} \beta_i \, u_i \\ \lim_{\alpha \to \infty} \Sigma_{\psi_\alpha}^{(\beta)}(u) &= \max_i u_i \end{aligned}$$

De même, l'écriture sous la forme :

$$\widetilde{\mathbf{H}}_\alpha(p) = -\ln \left(\sum_{i=1}^{n} {p_i}^\alpha \right)^{1/\alpha - 1}$$

montre que, dans la forme (2.3), la quantité $\mathbf{G}_\alpha$ s'écrit :

$$\mathbf{G}_\alpha(p) = \left(\sum_{i=1}^{n} p_i \, p_i^{\alpha-1} \right)^{1/(\alpha-1)}$$

et correspond au $(\alpha-1)$-mélange des p_i pondérés aussi par les p_i eux-mêmes. Autrement dit, l'opérateur de moyenne qui, *dans l'espace des probabilités*, est sous-jacent à l'entropie de Rényi est :

$$\Sigma_{\phi_\alpha}^{(\beta)}(p) = \left(\sum_{i=1}^{n} \beta_i \, p_i^{\alpha-1} \right)^{1/(\alpha-1)} \tag{4.4}$$

4.1.1.3 Moyenne d'ordre α

On appelle *moyenne d'ordre* α :

$$\Sigma_{\alpha}^{(\beta)}(p) \stackrel{\Delta}{=} \Sigma_{\phi_{\alpha+1}}^{(\beta)}(p) = \left(\sum_{i=1}^{n} \beta_i \, p_i^{\alpha} \right)^{1/\alpha} \tag{4.5}$$

Cette moyenne admet comme cas particuliers les moyennes arithmétique, géométrique, harmonique et « root mean square » :

$$\begin{aligned}
\Sigma_{1}^{(\beta)}(p) &= A^{(\beta)}(p) = \sum_{i=1}^{n} \beta_i \, p_i \\
\lim_{\alpha \to 0} \Sigma_{\alpha}^{(\beta)}(p) &= G^{(\beta)}(p) = \prod_{i=1}^{n} p_i^{\beta_i} \\
\Sigma_{-1}^{(\beta)}(p) &= H^{(\beta)}(p) = \left(\sum_{i=1}^{n} \frac{\beta_i}{p_i} \right)^{-1} \\
\Sigma_{1/2}^{(\beta)}(p) &= R^{(\beta)}(p) = \left(\sum_{i=1}^{n} \beta_i \, \sqrt{p_i} \right)^{2}
\end{aligned}$$

4.1.1.4 Moyennes généralisées

Des discussions précédentes, il ressort que plus généralement, à partir des quatre formes fonctionnelles (2.1)-(2.4), il convient d'introduire deux types de moyennes.

Le premier type de moyennes, qui travaille sur des quantités $u_i = \ln p_i$ qui ne sont ni positives ni normalisées, est défini de la manière suivante :

$$\Sigma_{g}^{(\beta)}(u) = \sum_{i=1}^{n} \beta_i \, g(u_i) \tag{4.6}$$

$$\Sigma_{\psi}^{(\beta)}(u) = \psi^{-1} \left(\sum_{i=1}^{n} \beta_i \, \psi(u_i) \right) \tag{4.7}$$

et indique de quelle façon on doit calculer la moyenne de quantités telles que des entropies ou des divergences. L'utilité de cette notion apparaîtra dans la section 4.2.

Le deuxième type de moyennes, qui travaille sur des quantités p_i qui, elles, sont strictement positives et normalisées, est défini par :

$$\Sigma_{h}^{(\beta)}(p) = \sum_{i=1}^{n} \beta_i \left(-\frac{h(p_i)}{p_i} \right) \tag{4.8}$$

$$\Sigma_{\phi}^{(\beta)}(p) = \phi^{-1} \left(\sum_{i=1}^{n} \beta_i \, \phi(p_i) \right) \tag{4.9}$$

et indique de quelle façon on doit calculer la moyenne (ou le *mélange*) de (lois de) probabilités. Pour cette raison, et malgré les liens entre g et h d'une part, et entre ψ et ϕ d'autre part, la définition de ce deuxième type de moyennes n'est pas redondante avec les définitions précédentes. L'utilité de cette notion apparaîtra dans la section 4.2.. La quantité (4.9) est connue sous le nom de *moyenne généralisée* [98, 112, 67, 1, 77] (et la fonction ϕ est alors appelée *fonction de Kolmogorov-Nagumo* [130]), et aussi sous le nom de *collecteur d'information* [22, 88]. D'autre part, la quantité :

$$\mathbf{G}_{\phi}(p) \stackrel{\Delta}{=} \Sigma_{\phi}^{(p)}(p) = \phi^{-1} \left(\sum_{i=1}^{n} p_i \, \phi(p_i) \right) \tag{4.10}$$

est appelée *ϕ-average probability* [3] (Voir comment distinguer en Français « mean » et « average » !)

Cependant, l'utilisation de ces moyennes pour le calcul du mélange de *densités* de probabilité p_i (par rapport à une mesure de référence λ) requiert une normalisation :

$$\begin{aligned}\widetilde{\Sigma}_h^{(\beta)}(p) &= \frac{\sum_{i=1}^n \beta_i \ \left(-\frac{h(p_i)}{p_i}\right)}{\int \sum_{i=1}^n \beta_i \ \left(-\frac{h(p_i(x))}{p_i(x)}\right) \ d\lambda(x)} \\ \widetilde{\Sigma}_\phi^{(\beta)}(p) &= \frac{\phi^{-1}\left(\sum_{i=1}^n \beta_i \ \phi(p_i)\right)}{\int \phi^{-1}\left(\sum_{i=1}^n \beta_i \ \phi(p_i(x))\right) \ d\lambda(x)}\end{aligned}$$

pour que le mélange soit, lui aussi, une densité. Par exemple, on peut définir une moyenne normalisée d'ordre α par :

$$\widetilde{\Sigma}_\alpha^{(\beta)}(p) = \frac{\left(\sum_{i=1}^n \beta_i \ p_i^\alpha\right)^{1/\alpha}}{\int \left(\sum_{i=1}^n \beta_i \ p_i^\alpha(x)\right)^{1/\alpha} \ d\lambda(x)} \tag{4.11}$$

Ce mélange intervient dans la définition des extensions de la différence de Jensen (section 4.2) et l'introduction du rayon d'information [146, 56]. La relation ;

$$\left(\widetilde{\Sigma}_\alpha^{(\beta)}(p)\right)^\alpha = \text{cste} \times \sum_{i=1}^n \beta_i \ p_i^\alpha$$

fait apparaître le lien avec les α-familles discutées en [13].

4.1.1.5 Relations entre les moyennes généralisées

La construction de ces moyennes fournit le moyen de passer d'une forme intégrale à une forme non intégrale :

$$\Sigma_\psi^{(\beta)}(u) = \psi^{-1}\left(\Sigma_g^{(\beta)}(u)\right) \tag{4.12}$$

$$\Sigma_\phi^{(\beta)}(p) = \phi^{-1}\left(\Sigma_h^{(\beta)}(p)\right) \tag{4.13}$$

pour :

$$\psi(u) = g(u) \tag{4.14}$$

$$\phi(u) = -\frac{h(u)}{u} = -\breve{h}\left(\frac{1}{u}\right) \tag{4.15}$$

et, réciproquement, d'une forme non intégrale à une forme intégrale :

$$\Sigma_g^{(\beta)}(u) = \psi\left(\Sigma_\psi^{(\beta)}(u)\right) \tag{4.16}$$

$$\Sigma_h^{(\beta)}(p) = \phi\left(\Sigma_\phi^{(\beta)}(p)\right) \tag{4.17}$$

pour :

$$g(u) = \psi(u) \tag{4.18}$$

$$h(u) = -u \ \phi(u) \tag{4.19}$$

On a aussi évidemment :

$$\Sigma_\psi^{(\beta)}(-\ln p, \ldots, -\ln p_n) = -\ln \ \Sigma_\phi^{(\beta)}(p) \tag{4.20}$$

4.1.2 Moyennes sous-jacentes aux f-divergences

Dans cette sous-section et la suivante, on introduit deux types de définition de moyennes associées à des f-divergences (et aussi à des distances de Bregman pour ce qui est du deuxième type). Dans un premier temps, on donne une définition *explicite* analogue à celle donnée pour les moyennes (4.6)-(4.9) sous-jacentes aux entropies. Dans un deuxième temps, on considère une définition *implicite* plus récente [29] de deux classes de moyennes de quantités strictement positives; d'une part une classe de moyennes dites *moyennes entropiques*, qui contient non seulement celles des moyennes précédentes qui sont homogènes, mais aussi, par extension au cas de

variables aléatoires, toutes les mesures de centralité (espérance, médiane, quantiles, ...); d'autre part, la classe des moyennes généralisées (4.9) qui sont non homogènes. L'interprétation géométrique de ces deux moyennes implicites, qui sont définies par résolution d'un problème de minimisation, est une simple projection.

Pour ce qui est de la définition explicite, en procédant comme pour les entropies, on met en évidence les moyennes arithmétique et géométrique sous-jacentes à la divergence de Kullback, et on montre ainsi que les moyennes généralisées sous-jacentes aux quatre formes fonctionnelles de f-divergences sont identiques aux moyennes généralisées sous-jacentes aux quatre formes fonctionnelles d'entropies déjà introduites, à ceci près qu'elles travaillent dans l'espace du rapport de vraisemblance (ou de son logarithme) et non dans celui de la probabilité (ou de son logarithme).

On considère des lois de probabilité discrètes à n masses : $p, \ldots, p_n$ ($\sum_{i=1}^n p_i = 1$) et $q_1, \ldots, q_n$ ($\sum_{i=1}^n q_i = 1$). L'information de Kullback s'écrit :

$$\begin{aligned} \mathbf{K}(p;q) &= \sum_{i=1}^n p_i \ln \frac{p_i}{q_i} \\ &= -\ln \prod_{i=1}^n \left(\frac{q_i}{p_i}\right)^{p_i} \end{aligned}$$

Les moyennes arithmétique et géométrique sous-jacentes à l'information de Kullback sont donc bien celles définies pour l'entropie de Shannon en (4.1)-(4.2), si l'on remplace les p_i par les $\frac{q_i}{p_i}$.

Le même raisonnement montre que les moyennes sous-jacentes à toutes les f-divergences, intégrales ou non, sont identiques aux moyennes (4.6)-(4.9) sous-jacentes aux entropies correspondantes. Ceci est utilisé pour l'obtention des extensions de la différence de Jensen discutées à la section 4.2.

4.1.3 Moyennes et projections

Considérons maintenant la définition implicite de moyennes (de quantités a_i supposées seulement strictement positives) définies par résolution d'un problème de minimisation [29] :

$$\mathbf{M}_d^{(\beta)}(a) = \arg\min_b \sum_{i=1}^n \beta_i \, d(b, a_i) \tag{4.21}$$

où d est une distance au sens où $d(b,a) > 0$ pour $b \neq a$ et $d(a,a) = 0$.

Une telle moyenne est donc une projection, au sens de la distance d, des nombres a_i sur la demi-droite $x_1 = \ldots = x_n > 0$ [57].

4.1.3.1 Moyennes entropiques ou $\mathbf{I}_f$-projections

En choisissant une distance d du type f-divergence intégrale :

$$d(u,v) = v \; h\left(\frac{u}{v}\right)$$

où h est une fonction strictement convexe et différentiable [1], on peut ainsi définir des moyennes, dites *entropiques*, par la projection correspondante :

$$\mathbf{M}_{I,h}^{(\beta)}(a) = \arg\min_b \sum_{i=1}^n \beta_i \; a_i \; h\left(\frac{b}{a_i}\right) \tag{4.22}$$

La définition implicite de ces moyennes est donc :

$$\sum_{i=1}^n \beta_i \; h'\left(\frac{b}{a_i}\right) = 0 \tag{4.23}$$

On donne au tableau 4.1 plusieurs exemples de moyennes connues, qui sont obtenues par résolution de cette équation pour une fonction h appropriée [2], et qui sont donc des moyennes entropiques.

D'autre part, une même moyenne peut correspondre, par ce procédé, à plusieurs fonctions h différentes. Par exemple, la moyenne géométrique de $n = 2$ valeurs a_i de poids $\beta_i = 1/2$ correspond aussi bien à $u \ln u - u + 1$ qu'à $(-2 \ln u + u^2 - 1)/3$. On reviendra sur ce point en étudiant les propriétés d'invariance des moyennes.

1. Cette dernière hypothèse peut être relâchée : voir l'exemple des quantiles au tableau 4.2.
2. La fonction $r_\alpha(u)$ apparaissant dans ce tableau est définie en (2.17).

TAB. 4.1 – *Moyennes entropiques.*

$h(u)$	moyenne $\mathbf{M}_{I,h}^{(\beta)}(a)$	
$r_0(u)$	$\sum_{i=1}^n \beta_i\, a_i$	$A^{(\beta)}(a) = \Sigma_1^{(\beta)}(a)$
$r_1(u)$	$\prod_{i=1}^n a_i^{\beta_i}$	$G^{(\beta)}(a) = \Sigma_0^{(\beta)}(a)$
$r_2(u)$	$\left(\sum_{i=1}^n \frac{\beta_i}{a_i}\right)^{-1}$	$H^{(\beta)}(a) = \Sigma_{-1}^{(\beta)}(a)$
$r_{1/2}(u)$	$\left(\sum_{i=1}^n \beta_i\, \sqrt{a_i}\right)^2$	$R^{(\beta)}(a) = \Sigma_{1/2}^{(\beta)}(a)$
$r_{1-\alpha}(u)$	$\left(\sum_{i=1}^n \beta_i\, a_i^\alpha\right)^{1/\alpha}$	$\Sigma_\alpha^{(\beta)}(a)$
$\frac{u^{1-\alpha}-1}{1-\alpha} - \frac{u^{1-\gamma}-1}{1-\gamma}$	$\left(\frac{\sum_{i=1}^n \beta_i\, a_i^\gamma}{\sum_{i=1}^n \beta_i\, a_i^\alpha}\right)^{1/\gamma-\alpha}$	$\mathrm{Gini}(\alpha,\gamma), \gamma \geq 0 > \alpha$ ou $\gamma > 0 \geq \alpha$
$-\frac{2}{3}\ln u + \frac{u^2}{3} - \frac{1}{3}$	$\left(\frac{\sum_{i=1}^n \beta_i\, a_i}{\sum_{i=1}^n \frac{\beta_i}{a_i}}\right)^{1/2}$	composition $G^{(1/2)}(A^{(\beta)}(a), H^{(\beta)}(a))$

TAB. 4.2 – *Moyennes entropiques d'une variable aléatoire X de densité p.*

$h(u)$	moyenne $\mathbf{M}_{I,h}(X)$	
$-\ln u + u - 1 = r_0(u)$	$\mathbf{E}(X)$	espérance
$u\ln u - u + 1 = r_1(u)$	$e^{\mathbf{E}(X)}$	espérance géométrique
$(u-1)^2 = r_2(u)$	$1/\mathbf{E}(1/X)$	espérance harmonique
$\frac{u^{1-\alpha}-\alpha}{\alpha-1} + u = r_{1-\alpha}(u)$	$\left(\int x^\alpha p(x)\, d\lambda(x)\right)^{1/\alpha}$	moy.d'ordre $\alpha > 0$
$\begin{cases} (1-\alpha)(u-1) & (u>1) \\ \alpha(1-u) & (0<u\leq 1) \end{cases}$	$F^{-1}(\alpha)$, où $F' = p$	α-ième quantile

Par extension, on appelle *moyenne entropique d'une variable aléatoire* positive X, de support $[x_1, x_2] \subset \mathbb{R}^+$, de densité p par rapport à la mesure λ, la quantité :

$$\mathbf{M}_{I,h}(X) = \arg\min_b \int_{x_1}^{x_2} x\; h\left(\frac{b}{x}\right) p(x)\; d\lambda(x)$$

qui est donc solution de l'équation :

$$\int_{x_1}^{x_2} h'\left(\frac{b}{x}\right) p(x)\; d\lambda(x) = 0$$

On donne au tableau 4.2 plusieurs exemples de mesures de centralité (espérance, moments, médiane, quantiles...), qui sont obtenues par résolution de cette équation pour une fonction h appropriée.

On verra à la section 4.2.1 que le barycentre (ou α-mélange ou α-moyenne) qui sert à la définition du rayon d'information de Sibson est aussi solution d'un problème de minimisation du type (4.21) où d est l'information de Rényi, à ceci près que la moyenne arithmétique considérée en (4.21) est remplacée par la moyenne d'ordre α sous-jacente à cette divergence.

4.1.3.2 Moyennes généralisées ou D-projections

Pour des raisons de non homogénéité (voir plus loin), les moyennes généralisées (4.9) :

$$\Sigma_{\phi}^{(\beta)}(a) = \phi^{-1}\left(\sum_{i=1}^{n} \beta_i \ \phi(a_i)\right) \tag{4.24}$$

ne sont pas, en général, solution d'une équation de la forme (4.22), et donc ne sont pas des moyennes entropiques. Par contre, elles sont solution d'un problème de minimisation du type (4.21) :

$$\mathbf{M}_{D,h}^{(\beta)}(a) = \arg\min_{b} \sum_{i=1}^{n} \beta_i \ d_h(b, a_i) \tag{4.25}$$

pour une *distance de Bregman* :

$$d(u,v) = d_h(u,v) = h(u) - h(v) - h'(v)(u-v) \tag{4.26}$$

où h est déduite de ϕ par :

$$h' = \phi \tag{4.27}$$

(Remarquer que si $h(u)$ convient, $ah(u) + bu + c$ convient aussi). En effet, de telles moyennes vérifient :

$$\mathbf{M}_{D,h}^{(\beta)}(a) = \arg\min_{b}\left(h(b) - b \ \sum_{i=1}^{n} \beta_i \ h'(a_i)\right)$$

et la définition implicite correspondante en est donc :

$$h'(b) = \sum_{i=1}^{n} \beta_i \ h'(a_i) \tag{4.28}$$

qui est bien identique à (4.24) en vertu de (4.27).

Par exemple, pour :

$$h(u) = u \ln u - (1+u)\ln(1+u)$$

(voir f-divergences naturelles en section 2.5), la moyenne correspondante est une composition de moyennes géométriques :

$$\mathbf{M}_{D,h}^{(\beta)}(a) = \frac{G^{(\beta)}(a)}{G^{(\beta)}(1+a) - G^{(\beta)}(a)}$$

On peut aussi étendre ces moyennes au cas de variables aléatoires [29].

4.1.3.3 Moyenne d'ordre α

De ces deux types de moyennes implicites, de la petite remarque sur les distances d_h de (4.26) :

$$d_{r_\alpha}(u,v) = \frac{1}{\alpha} d_{h_\alpha}(u,v) = \frac{1}{\alpha(\alpha-1)} d_{\phi_{\alpha+1}}(u,v)$$

et de la discussion sur le rayon d'information en 4.2.1 [146], il résulte que la moyenne d'ordre α (4.5) possède plusieurs définitions fonctionnelles implicites, à savoir :

$$\begin{aligned}
\Sigma_{\alpha}^{(\beta)} \triangleq \Sigma_{\phi_{1+\alpha}}^{(\beta)} &= \mathbf{M}_{I,r_{1-\alpha}}^{(\beta)} \\
&= \mathbf{M}_{D,h_{1+\alpha}}^{(\beta)} = \mathbf{M}_{D,r_{1+\alpha}}^{(\beta)} = \mathbf{M}_{D,\phi_{2+\alpha}}^{(\beta)} \\
\widetilde{\Sigma}_{\alpha}^{(\beta)} &= \mathcal{M}_{\psi_\alpha, \widetilde{\mathbf{R}}_\alpha}^{(\beta)}
\end{aligned}$$

où :

$$\mathcal{M}_{\psi,d}^{(\beta)}(a) = \arg\min_{b} \ \Sigma_{\psi}^{(\beta)}\left(d(a_i, b)\right)$$

Les α-moyennes, qui sont les mélanges sous-jacents à l'entropie de Rényi, sont donc à la fois des Bregman-projections et des f-div.-projections, et on verra plus loin, par invariance, que ce sont les seules moyennes généralisées qui possèdent cette propriété.

Pour comparer des (classes de) distances, on ne doit donc pas seulement se préoccuper des distances elles-mêmes, mais aussi des projections qui résultent de l'usage des dites distances. (La remarque précédente sur les α-moyennes, n'apporte pas le même éclairage que celui qui résulte de ce que l'intersection des distances de Bregman et des f-divergences est réduite à l'information de Kullback (discutée à la section 5.2)).

4.1.4 Invariance des moyennes

On donne des propriétés d'invariance des moyennes soit par transformation (permutation,translation, dilatation) des nombres, soit par transformation des fonctions h, ψ, ϕ. On indique aussi que toutes les moyennes généralisées sont localement équivalentes à la moyenne arithmétique.

4.1.4.1 Invariance par transformation des nombres

On considère les invariances par permutation, translation et dilatation.

Symétrie Dans le cas de poids égaux $\beta_i = \frac{1}{n}$, *toutes* les moyennes, qu'elles soient définies explicitement par (4.6)-(4.9), ou implicitement par (4.21), sont invariantes par permutation des nombres u_i, c'est-à-dire *symétriques*.

Translation D'autre part, il est montré en [112, 77], par résolution d'équations fonctionnelles, que les seules moyennes *généralisées* (4.7) qui soient invariantes par translation ou par dilatation sont les moyennes d'ordre α sous-jacentes à l'entropie de Rényi.

Plus précisément, les seules moyennes généralisées invariantes par translation, i.e. vérifiant :

$$\Sigma_{\psi}^{(\beta)}(u_1 + l, \ldots, u_n + l) = \Sigma_{\psi}^{(\beta)}(u_1, \ldots, u_n) + l$$

sont les moyennes Σ_{ψ_α} définies en (4.3), et qui admettent comme cas particulier la moyenne arithmétique.

Dilatation De même, les seules moyennes généralisées Σ_ϕ (4.24) ou $\mathbf{M}_{D,h}$ (4.25) qui soient invariantes par dilatation (ou changement d'échelle), i.e. vérifiant :

$$\Sigma_{\phi}^{(\beta)}(l u_1, \ldots, l u_n) = l\ \Sigma_{\phi}^{(\beta)}(u_1, \ldots, u_n)$$

et qui sont dites aussi *homogènes*, sont les Σ_α, moyennes d'ordre α définies en (4.5), et qui admettent comme cas particulier la moyenne géométrique.

Les autres moyennes généralisées sont non homogènes.

Par contre, *toutes* les moyennes *entropiques* (4.22) sont homogènes.

4.1.4.2 Invariance par transformation des fonctions

Les moyennes généralisées (4.7) et (4.9) possèdent des propriétés d'invariance par transformation de ψ, ϕ :

$$\text{pour } \underline{\psi}(u) = \gamma\ \psi(u) + \delta, \qquad \Sigma_{\underline{\psi}}^{(\beta)}(u) = \Sigma_{\psi}^{(\beta)}(u) \tag{4.29}$$

$$\text{pour } \underline{\phi}(u) = \gamma\ \phi(u) + \delta, \qquad \Sigma_{\underline{\phi}}^{(\beta)}(p) = \Sigma_{\phi}^{(\beta)}(p) \tag{4.30}$$

Par contre :

$$\text{pour } \underline{h}(u) = \gamma\ h(u) + \delta\ u, \qquad \Sigma_{\underline{h}}^{(\beta)}(p) = \gamma\ \Sigma_{h}^{(\beta)}(p) - \delta \tag{4.31}$$

4.1.4.3 Équivalence locale des moyennes

Les moyennes entropiques (4.23) sont toutes équivalentes à la moyenne arithmétique à l'infini [29]. Il en est de même des moyennes généralisées (4.9) correspondant à des fonctions ϕ telles que ${\phi^{-1}}'$ et ${\phi^{-1}}' \circ \phi$ satisfassent certaines conditions [34].

4.1.4.4 Comparaison des moyennes entropiques

S'il existe une constante c telle que :

$$c\ h'(u) \leq f'(u)$$

alors :

$$\mathbf{M}_{I,h}^{(\beta)}(a) \geq \mathbf{M}_{I,f}^{(\beta)}(a)$$

4.2 Moyennes généralisées et extensions de divergences

On utilise les différentes moyennes généralisées décrites à la section 4.1 pour imaginer des extensions de la différence de Jensen et de la distance de Bregman. Ce jeu n'est pas complètement formel, en ce sens qu'il permet de retrouver des notions connues, dont le rayon d'information, qui n'est qu'un cas particulier de capacité de canal [56].

4.2.1 Extensions de la différence de Jensen et rayon d'information

La différence de Jensen qui est, par définition, la différence entre l'entropie du mélange (linéaire) et la moyenne (arithmétique) des deux entropies individuelles, s'écrit :

$$\mathbf{J}_H^{(\beta)}(P,Q) = \mathbf{H}\left(\beta P + (1-\beta)Q\right) - \beta\,\mathbf{H}(P) - (1-\beta)\,\mathbf{H}(Q)$$

Dans le cas de l'entropie de Shannon $\mathbf{H} = \mathbf{H}_1$, elle s'écrit aussi :

$$\begin{aligned} \mathbf{J}_{H_1}^{(\beta)}(P,Q) &= \mathbf{H}_1\left(\beta P + (1-\beta)Q\right) - \beta\,\mathbf{H}_1(P) - (1-\beta)\,\mathbf{H}_1(Q) && (4.32)\\ &= \beta\,\bar{\mathbf{K}}\left(P, \beta P + (1-\beta)Q\right) + (1-\beta)\,\bar{\mathbf{K}}\left(Q, \beta P + (1-\beta)Q\right) && (4.33)\end{aligned}$$

c'est-à-dire comme la moyenne (arithmétique) des divergences entre chaque loi et leur mélange (linéaire).

Cette relation permet d'imaginer de fabriquer d'autres divergences en changeant de divergence en (4.33). Plus généralement, on peut imaginer de jouer sur les notions de moyennes sous-jacentes soit aux entropies comme en (4.32), soit aux divergences comme en (4.33). Se pose alors la question de savoir si réellement on crée ainsi d'autres types de divergences, distincts de ceux que l'on a déjà introduits, à savoir $\mathbf{I}_f, \mathbf{D}, \mathbf{J}, \mathbf{K}, \mathbf{L}$.

4.2.1.1 En jouant sur la notion de divergence

On se propose donc d'abord de fabriquer d'autres divergences en remplaçant l'information de Kullback $\bar{\mathbf{K}}$ par une autre divergence $\mathbf{D}$ [152]. Les divergences ainsi fabriquées ne peuvent appartenir à la classe des f-divergences que si la métrique qui leur est associée, qui est identique à celle de $\mathbf{D}$, est proportionnelle à celle de Fisher.

Il est à noter que ce jeu sur la divergence ne permet pas toujours d'obtenir de nouvelles divergences. Il convient en effet de se rappeler que pour toute entropie, non nécessairement intégrale, la différence de Jensen s'écrit comme moyenne des distances de Bregman entre chaque loi et le barycentre des lois :

$$\mathbf{J}_H^{(\beta)}(P,Q) = \beta\,\mathbf{D}_H(P, \beta P + (1-\beta)Q) + (1-\beta)\,\mathbf{D}_H(Q, \beta P + (1-\beta)Q)$$

et que cette propriété vaut en particulier pour l'information de Rényi.

4.2.1.2 En jouant sur la notion de moyenne

Pour procéder à d'autres extensions, on peut aussi imaginer de jouer sur la notion de moyenne sous-jacente à la définition de l'entropie ou de la divergence considérée, comme on l'a expliqué en 4.1. En effet, que ce soit en (4.32) ou en (4.33), on constate qu'interviennent deux types de moyenne :

- une moyenne $\widetilde{\Sigma}^{(\beta)}$ ou barycentre des lois P et Q,
- une moyenne $\Sigma^{(\beta)}$ des entropies de chaque loi ou des divergences entre chaque loi et le barycentre.

Dans le cas présent de l'entropie de Shannon et de l'information de Kullback, ces deux moyennes sont identiques à la moyenne pondérée usuelle [130, 146, 52] :

$$\widetilde{\Sigma}_1^{(\beta)}(u) = \Sigma_1^{(\beta)}(u) = \sum_{i=1}^{n} \beta_i u_i$$

où les poids β_i peuvent toujours être supposés normalisés : $\sum_{i=1}^n \beta_i = 1$.

On en déduit que l'on peut étendre les définitions (4.32) et (4.33), réécrites sous la forme :

$$\begin{aligned} \mathbf{J}_1^{(\beta)}(P,Q) &= \mathbf{H}_1\left(\widetilde{\Sigma}_1^{(\beta)}(P,Q)\right) - \Sigma_1^{(\beta)}\left(\mathbf{H}_1(P), \mathbf{H}_1(Q)\right) && (4.34)\\ &= \Sigma_1^{(\beta)}\left(\bar{\mathbf{K}}\left(P, \widetilde{\Sigma}_1^{(\beta)}(P,Q)\right), \bar{\mathbf{K}}\left(Q, \widetilde{\Sigma}_1^{(\beta)}(P,Q)\right)\right) && (4.35)\end{aligned}$$

en :

- considérant un nombre fini quelconque n de lois P_i,
- adoptant comme moyennes les moyennes correspondant à l'entropie $\mathbf{H}$ ou la divergence $\mathbf{D}$ considérée (voir section 1).

De manière précise, on étend (4.34) en[3] :

$$\mathcal{J}_H^{(\beta)}(P_1,\ldots,P_n) = \mathbf{H}\left(\widetilde{\Sigma}_H^{(\beta)}(P_1,\ldots,P_n)\right) - \Sigma_H^{(\beta)}\left(\mathbf{H}(P_1),\ldots,\mathbf{H}(P_n)\right) \tag{4.36}$$

et on étend (4.35) en :

$$\mathcal{J}_D^{(\beta)}(P_1,\ldots,P_n) = \Sigma_D^{(\beta)}\left(\mathbf{D}\left(P_1,\widetilde{\Sigma}_D^{(\beta)}(P_1,\ldots,P_n)\right),\ldots,\mathbf{D}\left(P_n,\widetilde{\Sigma}_D^{(\beta)}(P_1,\ldots,P_n)\right)\right) \tag{4.37}$$

On examine maintenant ces extensions pour les différentes formes fonctionnelles d'entropies et de f-divergences considérées, en insistant particulièrement sur le cas de l'information de Rényi.

Le cas des f-divergences On a vu à la section 1 que les moyennes généralisées sous-jacentes aux quatre formes fonctionnelles d'entropies (2.1)-(2.3) sont :

$$\begin{aligned}
\Sigma_g^{(\beta)}(u) &= \sum_{i=1}^n \beta_i\, g(u_i) \\
\Sigma_\psi^{(\beta)}(u) &= \psi^{-1}\left(\sum_{i=1}^n \beta_i\, \psi(u_i)\right) \\
\widetilde{\Sigma}_h^{(\beta)}(p) &= \frac{\sum_{i=1}^n \beta_i \left(-\frac{h(p_i)}{p_i}\right)}{\int \sum_{i=1}^n \beta_i \left(-\frac{h(p_i(x))}{p_i(x)}\right) d\lambda(x)} \\
\widetilde{\Sigma}_\phi^{(\beta)}(p) &= \frac{\phi^{-1}\left(\sum_{i=1}^n \beta_i\, \phi(p_i)\right)}{\int \phi^{-1}\left(\sum_{i=1}^n \beta_i\, \phi(p_i(x))\right) d\lambda(x)}
\end{aligned}$$

Autrement dit :

$$\begin{aligned}
&\text{Pour } \mathbf{H} = \mathbf{H}_h, \qquad \Sigma_H = \Sigma_g,\ \widetilde{\Sigma}_H = \widetilde{\Sigma}_h \\
&\text{Pour } \mathbf{H} = \mathbf{H}_\psi, \qquad \Sigma_H = \Sigma_\psi,\ \widetilde{\Sigma}_H = \widetilde{\Sigma}_\phi
\end{aligned}$$

où $h(u) = -ug(-\ln u)$ et où $\phi(u) = \psi(-\ln u)$. On a vu aussi que les moyennes généralisées sous-jacentes aux f-divergences, intégrales ou non, sont identiques aux moyennes sous-jacentes aux entropies, c'est-à-dire :

$$\text{Pour } \mathbf{D} = \mathbf{I}_h, \mathbf{I}_\psi, \quad \Sigma_D = \Sigma_H,\ \widetilde{\Sigma}_D = \widetilde{\Sigma}_H$$

Les extensions (4.36)-(4.37) de la différence de Jensen sont alors :

$$\mathcal{J}_{H_h}^{(\beta)}(P_1,\ldots,P_n) = \mathbf{H}_h\left(\widetilde{\Sigma}_h^{(\beta)}(P_1,\ldots,P_n)\right) - \Sigma_g^{(\beta)}\left(\mathbf{H}_h(P_1),\ldots,\mathbf{H}_h(P_n)\right) \tag{4.38}$$

$$\mathcal{J}_{I_h}^{(\beta)}(P_1,\ldots,P_n) = \Sigma_g^{(\beta)}\left(\mathbf{I}_h\left(P_1,\widetilde{\Sigma}_h^{(\beta)}(P_1,\ldots,P_n)\right),\ldots,\mathbf{I}_h\left(P_n,\widetilde{\Sigma}_h^{(\beta)}(P_1,\ldots,P_n)\right)\right) \tag{4.39}$$

dans le cas intégral, et :

$$\mathcal{J}_{H_\psi}^{(\beta)}(P_1,\ldots,P_n) = \mathbf{H}_\psi\left(\widetilde{\Sigma}_\phi^{(\beta)}(P_1,\ldots,P_n)\right) - \Sigma_\psi^{(\beta)}\left(\mathbf{H}_\psi(P_1),\ldots,\mathbf{H}_\psi(P_n)\right) \tag{4.40}$$

$$\mathcal{J}_{I_\psi}^{(\beta)}(P_1,\ldots,P_n) = \Sigma_\psi^{(\beta)}\left(\mathbf{I}_\psi\left(P_1,\widetilde{\Sigma}_\phi^{(\beta)}(P_1,\ldots,P_n)\right),\ldots,\mathbf{I}_\psi\left(P_n,\widetilde{\Sigma}_\phi^{(\beta)}(P_1,\ldots,P_n)\right)\right) \tag{4.41}$$

dans le cas non intégral. (En fait, cette définition vient de la généralisation du rayon d'information défini ci-dessous).

On étudie maintenant le cas des entropies d'ordre α, intégrales ou non, i.e. Havrda-Charvát et Rényi, et les f-divergences associées, i.e. χ^2-divergence d'ordre α et information de Rényi.

3. Mais ceci n'a de sens que si l'entropie $\mathbf{H}$ considérée est effectivement concave par rapport à cette notion de mélange!

Le cas de l'information de Rényi: le rayon d'information On considère le cas de l'information (ou divergence) d'ordre α de Rényi $\mathbf{D} = \mathbf{I}_\psi = \widetilde{\mathbf{R}}_\alpha$ (2.18).

Définition du rayon d'information Dans le cas de lois P_i admettant une densité p_i par rapport à une mesure λ, l'extension (4.41) a été proposée par Sibson sous le nom de *rayon d'information* [146] :

$$\mathcal{S}_\alpha^{(\beta)}(P_1, \ldots, P_n) = \mathcal{J}_{\widetilde{R}_\alpha}^{(\beta)}(P_1, \ldots, P_n)$$

La moyenne correspondant à l'information d'ordre α de Rényi $\mathbf{D} = \widetilde{\mathbf{R}}_\alpha$ (2.18) est :

$$\Sigma_{\psi_\alpha}^{(\beta)}(u) = \frac{1}{\alpha - 1} \ln \sum_{i=1}^n \beta_i \, e^{(\alpha-1)u_i}$$

et la moyenne normalisée servant à la définition du mélange des lois est :

$$\widetilde{\Sigma}_\alpha^{(\beta)}(p) = \frac{\left(\sum_{i=1}^n \beta_i \, p_i^\alpha\right)^{1/\alpha}}{\int \left(\sum_{i=1}^n \beta_i \, p_i^\alpha(x)\right)^{1/\alpha} d\lambda(x)}$$

Le rayon d'information n'est donc autre que :

$$\begin{aligned} \mathcal{S}_\alpha^{(\beta)}(P_1, \ldots, P_n) &= \Sigma_{\psi_\alpha}^{(\beta)}\left(\widetilde{\mathbf{R}}_\alpha(P_1, \widetilde{\Sigma}_\alpha^{(\beta)}(P_1, \ldots, P_n)), \ldots, \widetilde{\mathbf{R}}_\alpha(P_n, \widetilde{\Sigma}_\alpha^{(\beta)}(P_1, \ldots, P_n))\right) \\ &= \frac{\alpha}{\alpha - 1} \ln \int \left(\sum_{i=1}^n \beta_i \, p_i^\alpha(x)\right)^{1/\alpha} d\lambda(x) \end{aligned} \tag{4.42}$$

Noter que les extensions (4.40) et (4.41) ne coïncident pas dans ce cas, autrement dit :

$$\mathcal{S}_\alpha^{(\beta)}(P_1, \ldots, P_n) \neq \mathcal{J}_{\widetilde{H}_\alpha}^{(\beta)}(P_1, \ldots, P_n)$$

où :

$$\begin{aligned} \mathcal{J}_{\widetilde{H}_\alpha}^{(\beta)}(P_1, \ldots, P_n) &= \widetilde{\mathbf{H}}_\alpha\left(\widetilde{\Sigma}_\alpha^{(\beta)}(P_1, \ldots, P_n)\right) - \Sigma_{\psi_\alpha}^{(\beta)}\left(\widetilde{\mathbf{H}}_\alpha^{(\beta)}(P_1), \ldots, \widetilde{\mathbf{H}}_\alpha^{(\beta)}(P_n)\right) \\ &= -\frac{1}{\alpha - 1}\left[\ln \int \left(\sum_{i=1}^n \beta_i p_i^\alpha(x)\right) d\lambda(x) - \ln \sum_{i=1}^n \frac{\beta_i}{\int p_i^\alpha(x) \, d\lambda(x)}\right] \end{aligned}$$

Voir plus loin la remarque de [56] sur les informations mutuelles : $\mathcal{I}_{\widetilde{H}_\alpha} \neq \mathcal{I}_{\widetilde{R}_\alpha}$.

Cas particuliers et variantes

- *Cas particuliers :* Pour $\alpha = 1$ et $n = 2$, on retrouve (4.33) et donc la différence de Jensen (4.32) :

$$\mathcal{S}_1^{(\beta)}(P, Q) = \mathbf{J}_{H_1}^{(\beta)}(P, Q)$$

  Pour $\alpha \to \infty$, le rayon d'information devient :

$$\mathcal{S}_\infty^{(\beta)}(P_1, \ldots, P_n) = \ln \int \max_i p_i(x) \, d\lambda(x)$$

  et la moyenne correspondante (non considérée par Rényi) est :

$$\Sigma_\infty^{(\beta)}(u) = \max_i \, u_i$$

  On en déduit que, pour $n = 2$:

$$\mathcal{S}_\infty^{(\beta)}(P, Q) = \ln(1 + \mathbf{V}(P, Q))$$

  ce qui donne à la distance en variation $\mathbf{V}$ de Kolmogorov une interprétation en termes de notion d'information [146].

- *Variantes :* On peut introduire [146] des contraintes supplémentaires dans la définition du barycentre. Par exemple, pour des lois multivariées de matrice de covariance définie positive, on peut imposer que le barycentre soit recherché parmi les lois Gaussiennes. Pour $\alpha = 1$ et $n = 2$, on peut obtenir ainsi dans le cas Gaussien homoscédastique d'autres généralisations de la distance de Bhattacharrya que celles qui découlent de la classe des f-divergences; voir exemples à la fin.

Autre propriété On peut montrer – c'est même la définition dans [146] – que le rayon d'information est tel que :

$$\begin{aligned}\mathcal{S}_\alpha^{(\beta)}(P_1,\ldots,P_n) &= \inf_{Q>>\sum_{i=1}^n \beta_i P_i} \mathcal{S}_\alpha^{(\beta)}(P_1,\ldots,P_n|Q)\\ &= \mathcal{S}_\alpha^{(\beta)}\left(P_1,\ldots,P_n|\widetilde{\Sigma}_\alpha^{(\beta)}(P_1,\ldots,P_n)\right)\end{aligned}$$

où :

$$\mathcal{S}_\alpha^{(\beta)}(P_1,\ldots,P_n|Q) \triangleq \Sigma_{\psi_\alpha}^{(\beta)}\left(\widetilde{\mathbf{R}}_\alpha(P_1,Q),\ldots,\widetilde{\mathbf{R}}_\alpha(P_n,Q)\right)$$

De sorte que le mélange d'ordre α apparaît comme une moyenne de type entropique (4.21), i.e. minimum d'une moyenne de distances, mais où la moyenne arithmétique est remplacée par la moyenne $\Sigma_{\psi_\alpha}^{(\beta)}$:

$$\widetilde{\Sigma}_\alpha^{(\beta)}(P_1,\ldots,P_n) = \arg\min_Q \Sigma_{\psi_\alpha}^{(\beta)}\left(\widetilde{\mathbf{R}}_\alpha(P_1,Q),\ldots,\widetilde{\mathbf{R}}_\alpha(P_n,Q)\right)$$

Pour montrer ce résultat, il suffit, mais le calcul est un peu long, de montrer que :

$$\mathcal{S}_\alpha^{(\beta)}(P_1,\ldots,P_n|Q) - \mathcal{S}_\alpha^{(\beta)}\left(P_1,\ldots,P_n|\widetilde{\Sigma}_\alpha^{(\beta)}(P_1,\ldots,P_n)\right) = \widetilde{\mathbf{R}}_\alpha\left(\widetilde{\Sigma}_\alpha^{(\beta)}(P_1,\ldots,P_n),Q\right)$$

(pour la définition (2.18)).

D'autre part, la définition du rayon d'information a été reprise et étendue récemment par Csiszár [56] qui la relie d'une part à des probabilités d'erreur en classification et d'autre part à une notion de capacité de canal. En effet, le rayon d'information $\mathcal{S}$ vérifie :

$$\mathcal{S}_\alpha^{(\beta)} = \frac{\alpha}{1-\alpha}\,\mathcal{G}_{(1-\alpha)/\alpha}^{(\beta)}$$

où $\mathcal{G}$ est la fonction de Gallager qui intervient dans le « random coding exponent » et le « sphere packing exponent » [70].

4.2.2 Extensions de la distance de Bregman

On définit une notion de distance de Bregman plus adaptée au cas de fonctionnelles non intégrales, en remplaçant la définition (2.24), dans laquelle intervient la différentielle de Gâteaux, par la propriété de dérivée à l'origine (2.26) où l'on remplace la différence de Jensen ordinaire par une de ses extensions.

On considère les distances de Bregman généralisées définies par :

$$\mathcal{D}_\psi(P,Q) = \frac{\partial}{\partial\beta}\,\mathcal{J}_\psi^{(\beta)}(P,Q)\Big|_{\beta=0} \tag{4.43}$$

où $\mathcal{J}_\psi^{(\beta)}$ est une des extensions de la différence de Jensen introduites plus haut.

Dans le cas du rayon d'information de Sibson, on obtient :

$$\frac{\partial}{\partial\beta}\,\mathcal{S}_\alpha^{(\beta)}(P,Q)\Big|_{\beta=0} = \alpha\,\mathbf{R}_\alpha(P,Q) \tag{4.44}$$

autrement dit la dérivée à l'origine de moyennes pondérées de l'information de Rényi est la χ^2-divergence d'ordre α (ou information de Havrda-Charvát). Autrement dit, la χ^2-divergence d'ordre α qui est une f-divergence, s'écrit aussi comme une distance de Bregman.

Chapitre 5

Divergences et métriques

On décrit d'abord les métriques associées aux entropies et divergences, et ensuite on discute l'intersection des deux classes de divergences introduites au chapitre 2.

5.1 Métriques associées aux entropies et divergences

Où l'on voit [41, 125, 124] comment associer une métrique différentielle quadratique à une fonctionnelle d'entropie (ou à une divergence) en considérant son Hessien [41, 124]. (En [44] développements techniques – noyaux reproduisants, transformations pseudo-conformes – montrant que la métrique de Fisher (5.1) est la métrique de Bergman, et suggérant que les courbures scalaire et de Ricci associées à cette métrique peuvent donner, pour l'inférence, des résultats analogues à ceux obtenus par Efron avec la courbure Gaussienne).

5.1.1 Métrique de Fisher

$\theta \in \mathbb{R}^n$.

$$ds^2(\theta) = \sum_{i,j=1}^{n} g_{ij}(\theta)\, d\theta_i\, d\theta_j \tag{5.1}$$

$(g_{ij}(\theta))_{ij}$ = Fisher; $(g^{ij}(\theta))_{ij}$ = Fisher^{-1}. Les formules sont données en (2.96).

On peut considérer d'autres métriques, correspondant à une autre entropie que celle de Shannon – voir (5.6), (5.8) et (5.10) [41, 124].

5.1.2 Entropies et métriques

5.1.2.1 Métrique associée à une fonctionnelle d'entropie

On considère le Hessien:

$$\Delta_R \mathbf{H}(P) = \delta^2 \mathbf{H}(P : R, R) \tag{5.2}$$

où la différentielle de Gâteaux est définie en (2.20).

$$\Delta_R \mathbf{H}(P) \leq 0 \Leftrightarrow \mathbf{H}(P) \text{ concave}$$

Alors l'opposé du Hessien le long d'une direction dans le plan tangent à l'espace paramétrique est une forme définie positive sur le plan tangent qui peut être considérée comme une métrique différentielle d'une géométrie Riemannienne, ce que l'on nomme *métrique de* $\mathbf{H}$*-entropie*:

$$\begin{aligned} ds^2_H(\theta) &= -\Delta_\theta \mathbf{H}(P_\theta) \\ &= \sum_{i,j=1}^{n} g^{(H)}_{ij}(\theta)\, d\theta_i\, d\theta_j \\ g^{(H)}_{ij}(\theta) &= -\frac{\partial^2 \mathbf{H}(P_\theta)}{\partial\theta_i\, \partial\theta_j} = -\partial_i \partial_j \mathbf{H}(P_\theta) \end{aligned} \tag{5.3}$$

5.1.2.2 Entropie intégrale

Pour $\mathbf{H}$ définie par (2.40), la différentielle de Gâteaux est calculée en (2.21), et le Hessien vaut :

$$\begin{aligned} \delta^2\mathbf{H}_h(P:R,S) &= -\int h''(p(x))\; r(x)\; s(x)\; d\lambda(x) \\ \Delta_R\mathbf{H}_h(P) &= \delta^2\mathbf{H}_h(P:R,R) = -\int h''(p(x))\; r^2(x)\; d\lambda(x) \end{aligned} \tag{5.4}$$

$$\Delta_R\mathbf{H}_h(P) \leq 0 \Leftrightarrow h \text{ convexe } \Leftrightarrow \mathbf{H}_h(P) \text{ concave}$$

Métrique de h-entropie:

$$\begin{aligned} ds_h^2(\theta) &= -\Delta_\theta\mathbf{H}_h(P_\theta) \\ &= \sum_{i,j=1}^{n} g_{ij}^{(h)}(\theta)\; d\theta_i\; d\theta_j \end{aligned} \tag{5.5}$$

$$g_{ij}^{(h)}(\theta) = \int h''(p_\theta(x))\; (\partial_i p_\theta)\; (\partial_j p_\theta)\; d\lambda(x) \tag{5.6}$$

Identique à la métrique de Fisher (2.96) si $uh''(u) = \text{cste}$:

$$\text{pour } \; uh''(u) = h''(1) = \gamma, \qquad ds_h^2(\theta) = \gamma\; ds^2(\theta) \tag{5.7}$$

Entropie d'ordre α de Havrda-Charvát [78] Pour $\alpha \in \mathbb{R}$, elle est définie en (2.49).
Métrique d'entropie d'ordre α:

$$g_{ij}^{(\alpha)}(\theta) = g_{ij}^{(H_\alpha)}(\theta) = \; \alpha \int p_\theta^\alpha\; (\partial_i \ln p_\theta)\; (\partial_j \ln p_\theta)\; d\lambda(x) \tag{5.8}$$

Pour $\alpha = 1$, $(g_{ij}(\theta))$=Fisher.

5.1.2.3 Entropie non intégrale

Entropie d'ordre α de Rényi [130, 49] Elle est définie en (2.57). L'opposé du Hessien:

$$\begin{aligned} g_{ij}^{\left(\widetilde{H}_\alpha\right)}(\theta) &= e^{(\alpha-1)\widetilde{\mathbf{H}}_\alpha(P_\theta)} \\ &\left[g_{ij}^{(\alpha)}(\theta) + \frac{\alpha^2}{\alpha-1}\, e^{(\alpha-1)\widetilde{\mathbf{H}}_\alpha(P_\theta)} \left(\int p_\theta^\alpha\; (\partial_i \ln p_\theta)\; d\lambda(x)\right)\; \left(\int p_\theta^\alpha\; (\partial_j \ln p_\theta)\; d\lambda(x)\right)\right] \end{aligned} \tag{5.9}$$

ne fournit pas en général une métrique $ds^2_{\widetilde{H}}(\theta)$ (i.e. non défini positif), car pour $\alpha > 1$, $\widetilde{\mathbf{H}}_\alpha$ n'est ni concave ni convexe en général; elle est cependant pseudo-concave [28].

5.1.2.4 Entropie quadratique

[125, 124] Elle est définie en (2.62). La métrique associée est:

$$g_{ij}^{(Q)}(\theta) = -2 \int q(x,y)\; (\partial_i \ln p_\theta(x))\; (\partial_j \ln p_\theta(y))\; d\lambda(x)\; d\lambda(y) \tag{5.10}$$

Elle est, comme la métrique associée à Fisher, invariante par rapport au paramètre et à la variable. Voir aussi les commentaires de Rao en [89].

5.1.3 Divergences et métriques

[41, 124, 47]

5.1.3.1 Distance géodésique associée à une métrique

[124, 21, 89]

$$d(\theta_1,\theta_2) = \inf_{\mathcal{C}_{\theta_1,\theta_2}} \left| \int_{\mathcal{C}_{\theta_1,\theta_2}} \sqrt{ds^2(\theta)}\ d\theta \right| \tag{5.11}$$

où $ds^2(\theta)$ est définie en (5.1). La courbe géodésique qui réalise le minimum peut être calculée par résolution des $\dim(\theta)$ équations différentielles du second ordre d'Euler-Lagrange, ce qui peut être lourd, de même que le calcul de l'intégrale en (5.11). Autres méthodes [21]: soit calculer la courbe géodésique par résolution des $2\ \dim(\theta)$ équations différentielles du premier ordre d'Hamilton, soit calculer directement (5.11) par résolution de l'équation aux dérivées partielles nonlinéaire de Hamilton-Jacobi. Voir exemples à la fin.

5.1.3.2 Métrique associée à une divergence

Une divergence $\mathbf{D}$ entre lois peut être utilisée pour définir une métrique différentielle quadratique sur l'espace paramétrique en considérant deux distributions voisines. Ceci est obtenu aisément en considérant, comme pour on l'a vu pour une entropie, le Hessien le long du plan tangent à l'espace paramétrique :

$$\begin{aligned} \delta^2\mathbf{D}(\theta) &\triangleq ds^2_D(\theta) \\ &= \sum_{i,j=1}^{n} g^{(D)}_{ij}(\theta)\ d\theta_i\ d\theta_j \\ g^{(D)}_{ij}(\theta) &= \left.\frac{\partial^2 \mathbf{D}(\theta,\phi)}{\partial\theta_i\ \partial\phi_j}\right|_{\phi=\theta} = \partial_i\ \partial_j\ \mathbf{D}(\theta,\phi=\theta) \end{aligned} \tag{5.12}$$

f-divergence, entropie et métrique

f-divergences intégrales Pour $\mathbf{D} = \mathbf{I}_f$ (2.87), on a [6, 124]:

$$ds^2_{I,f}(\theta) = g'(f(1))\ f''(1)\ ds^2(\theta) \tag{5.13}$$

f-divergences non intégrales Pour l'expression $\mathbf{D} = \mathbf{I}_\psi$ (2.88), où $f(u) = \psi(\ln u)$, on a :

$$ds^2_{I,\psi}(\theta) = g'(\psi(0))\ (\psi''(0) - \psi'(0))\ ds^2(\theta) \tag{5.14}$$

5.1.3.3 Divergence de Hellinger d'ordre α

La métrique associée à la divergence de Hellinger « généralisée » (2.94) est :

$$g^{(\psi)}_{ij}(\theta) = 2 \int (\psi'(p_\theta))^2\ (\partial_i p_\theta)(\partial_j p_\theta)\ d\lambda(x) \tag{5.15}$$

Celle associée à la divergence de Hellinger d'ordre α (2.95) est :

$$g^{(\alpha)}_{ij}(\theta) = 4 \int p^\alpha_\theta\ (\partial_i p_\theta)(\partial_j p_\theta)\ d\lambda(x)$$

qui n'est autre que la métrique d'entropie d'ordre α (5.8).

5.1.3.4 Différence de Jensen et distance de Bregman

Pour $\mathbf{D} = \mathbf{J}^{(\beta)}_H$, on a [124]:

$$\begin{aligned} g^{\left(J^{(\beta)}_H\right)}_{ij}(\theta) &= -\left.\frac{\partial^2 \mathbf{H}(\beta\ P_\theta + (1-\beta)\ P_\phi)}{\partial\theta_i\ \partial\phi_j}\right|_{\phi=\theta} \\ &= -\beta\ (1-\beta)\ \frac{\partial^2 \mathbf{H}(P_\theta)}{\partial\theta_i\ \partial\theta_j} \\ ds^2_{J,H,\beta}(\theta) &= \beta\ (1-\beta)\ ds^2_H(\theta) \end{aligned} \tag{5.16}$$

où $ds^2_H(\theta)$ est définie en (5.3); et donc, à cause de (2.26) :

$$ds^2_{D,H}(\theta) = ds^2_H(\theta)$$

5.1.3.5 D et J, K, L-divergences intégrales

Pour **H** de la forme (2.40):

$$\begin{aligned} ds^2_{D,h}(\theta) &= ds^2_h(\theta) \\ ds^2_{J,h,\beta}(\theta) &= \beta\,(1-\beta)\,ds^2_h(\theta) \\ ds^2_{J,h}(\theta) &= \frac{1}{4}\,ds^2_h(\theta) \\ ds^2_{K,h}(\theta) &= 2\int \left(\frac{h(p_\theta)}{p_\theta}\right)' (\partial_i p_\theta)(\partial_j p_\theta)\,d\lambda(x) \\ ds^2_{L,h}(\theta) &= 2\,h''(1)\,ds^2(\theta) \end{aligned} \tag{5.17}$$

où $ds^2_h(\theta)$ est définie en (5.6) et $ds^2(\theta)$ est la métrique de Fisher (5.1).

En particulier, pour $h = h_\alpha$:

$$\begin{aligned} ds^2_{D_\alpha}(\theta) &= ds^2_\alpha(\theta) \\ ds^2_{J_\alpha}(\theta) &= \frac{1}{4}\,ds^2_\alpha(\theta) \\ ds^2_{K_\alpha}(\theta) &= \frac{2}{\alpha}\,ds^2_\alpha(\theta) \end{aligned} \tag{5.18}$$

$$ds^2_{L_\alpha}(\theta) = 2\,\alpha\,ds^2(\theta) \tag{5.19}$$

$$ds^2_{\widetilde{R}_\alpha}(\theta) = ds^2(\theta) \tag{5.20}$$

où $ds^2_\alpha(\theta)$ est définie en (5.8).

Autrement dit, *les métriques associées aux divergences* $\mathbf{D}_\alpha, \mathbf{J}_\alpha, \mathbf{K}_\alpha$ *sont proportionnelles à la métrique d'entropie d'ordre* α *(*$\alpha \neq 1$*). Les métriques associées à l'information de Rényi* $\widetilde{\mathbf{R}}_\alpha$ *(2.90) et aux divergences* $\mathbf{L}_\alpha, \mathbf{D}_1, \mathbf{J}_1, \mathbf{K}_1$ *sont proportionnelles à la métrique de Fisher.*

5.2 Intersections des deux classes de divergences

On étudie les relations entre les deux classes de divergences introduites en 2.1 et 2.2.

Un premier point de vue possible est celui des moyennes généralisées et projections associées aux divergences et décrites à la section 4.1 : on compare les distances par le biais des projections qui résultent de leur usage. On se contente ici de renvoyer aux commentaires faits en 4.1 sur les α-moyennes.

Un deuxième point de vue est celui de la métrique différentielle quadratique associée à toute divergence et décrite à la section 5.1 : en effet, la métrique associée à toute f-divergence étant la métrique de Fisher (5.13), pour trouver celles des autres divergences qui sont aussi des f-divergences, il suffit de chercher celles qui ont la métrique de Fisher comme métrique associée.

On considère d'abord les formes fonctionnelles intégrales pour les entropies (2.40) et pour les f-divergences (2.63) dans le cas où g est l'identité, et on cherche sous quelles conditions les divergences dérivées d'entropie (**J**, **C**, **D**, **K**, **L**-divergences) sont aussi une f-divergence.

Puis on considère les autres formes fonctionnelles pour les entropies (2.51) et pour les f-divergences (2.88), et on se pose les mêmes questions.

5.2.1 Formes intégrales

On considère donc les formes (2.40) et (2.63) :

$$\mathbf{H}_h(P) = -\int h(p(x))\,d\lambda(x) \tag{5.21}$$

$$\mathbf{I}_f(P,Q) = \int q\,f\left(\frac{p}{q}\right)\,d\lambda(x) \tag{5.22}$$

Dans ce cas, outre l'identité formelle des expressions de divergences considérées, l'argument-clef est la prise en compte des métriques associées. En effet, la métrique associée à toute f-divergence est la métrique de Fisher (5.13).

On montre les résultats donnés par le tableau 5.1.

f-divergence et différence de Jensen En considérant les métriques associées (5.13) et (5.16)-(5.6), on obtient :

$$\begin{aligned} uh''(u) &= \text{cste} = \gamma = \frac{f''(1)}{\beta(1-\beta)} \\ h(u) &= \gamma\, u \ln u + \delta u + \epsilon \end{aligned} \tag{5.23}$$

Puis, en résolvant:

$$q\, f\left(\frac{p}{q}\right) = \beta\, h(p) + (1-\beta)\, h(q) - h(\beta\, p + (1-\beta)\, q)$$

il est facile de voir qu'une f-divergence ne peut être une différence de Jensen $\mathbf{J}_h$ que si:

$$f(u) = b_\alpha(u) \quad \text{(voir (2.82))}$$

où $\alpha = \frac{\beta}{1-\beta}$.

Autrement dit, en vertu de (2.23):

$$\mathbf{I}_f = \mathbf{J}_H^{(\alpha/(1+\alpha))} \Leftrightarrow \mathbf{H} = \mathbf{H}_1, \quad \mathbf{I}_f = \frac{1}{1+\alpha}\, \mathbf{B}_\alpha$$

(Condition suffisante donnée dans [95, 96]).

Une f-divergence ne peut être une différence de Jensen que de l'entropie de Shannon, et seulement si elle est proportionnelle à l'information d'ordre α de Bose-Einstein ou de Fermi-Dirac.

Par contre, la divergence de Kullback n'est pas une différence de Jensen. L'information de Kullback est la dérivée à l'origine de la différence de Jensen, voir (2.25)-(2.26).

f-divergence et J-divergence Cas précédent pour $\mathbf{J}^{(1/2)}$.

$$\mathbf{I}_f = \mathbf{J}_H^{(1/2)} \Leftrightarrow \mathbf{H} = \mathbf{H}_1, \quad \mathbf{I}_f = \frac{1}{2}\, L$$

Une f-divergence ne peut être une $\mathbf{J}$-divergence que de l'entropie de Shannon, et seulement si elle est la moitié de la L-divergence de Lin.

TAB. 5.1 – *Quelles divergences sont des f-divergences intégrales?*

Si f-divergence	Alors doit être	pour $\mathbf{H}$
$\mathbf{J}_H^{(\beta)}$ différence de Jensen	Bose-Einstein d'ordre $\frac{\beta}{1-\beta}$	$\mathbf{H}_1$
$\mathbf{B}_H$ distance de Bregman	information de Kullback	$\mathbf{H}_1$
$\mathbf{C}_H$ à la Chernoff	$\mathbf{C}_{H_1}$	$\mathbf{H}_1$
$\mathbf{J}$-divergence de Rao	L-divergence de Lin	$\mathbf{H}_1$
$\mathbf{K}$-divergence de Rao	divergence de Kullback	$\mathbf{H}_1$
$\mathbf{L}$-divergence de Rao	divergences de Kullback, Lin	$\mathbf{H}_1$
	distance de Hellinger	$\mathbf{H}_{1/2}$
symétrique!	$f(u) = h(u) + u\, h\left(\frac{1}{u}\right)$	...

f-divergence et K-divergence La condition:

$$q\ f\left(\frac{p}{q}\right) = (p-q)\left(\frac{h(p)}{p} - \frac{h(q)}{q}\right)$$

implique:

$$\frac{h(p)}{p} - \frac{h(q)}{q} = k\left(\frac{p}{q}\right)$$

où $k(u) = \frac{f(u)}{1-u}$. Par dérivations:

$$\begin{aligned} k'(u) &= -uk''(u) \\ k(u) &= \gamma\ \ln u \\ f(u) &= \gamma\ (u-1)\ \ln u \end{aligned}$$

$$u\ h''(u) = \gamma$$

L'entropie sous-jacente est donc encore nécessairement de la forme (5.23). D'autre part, l'identité des métriques associées implique:

$$2\left(\frac{h(u)}{u}\right)' = \frac{f''(1)}{u}$$

et $\epsilon = 0$ et $\gamma = \frac{f''(1)}{2}$.

Autrement dit:

$$\mathbf{I}_f = \mathbf{K}_h \Leftrightarrow \mathbf{H} = \gamma\ \mathbf{H}_1, \quad \mathbf{I}_f = \gamma\ \mathbf{K}$$

Une f-divergence ne peut être une **K***-divergence de Rao que pour l'entropie de Shannon, et seulement si elle est proportionnelle à la divergence de Kullback.*

f-divergence et L-divergence Cette fois, la comparaison des métriques ne fournit plus de condition sur h, puisque la métrique associée à une **L**-divergence est identique à celle de Fisher, quelle que soit l'entropie sous-jacente.

La condition:

$$q\ f\left(\frac{p}{q}\right) = p\ h\left(\frac{q}{p}\right) + q\ h\left(\frac{p}{q}\right)$$

implique:

$$f(u) = h(u) + u\ h\left(\frac{1}{u}\right) \quad [42] \tag{5.24}$$

et donc:

$$f(u) = u\ f\left(\frac{1}{u}\right) \tag{5.25}$$

Une f-divergence ne peut être une **L***-divergence de Rao que si elle est symétrique.* Réciproquement toute f-divergence symétrique est une **L**-divergence de Rao pour toute entropie h satisfaisant (5.24). Il est à noter que l'on obtient ainsi, par résolution de (5.24), un moyen de définir une entropie h, associée à f. Une solution évidente à cette équation est $h(u) = \frac{1}{2}\ f(u)$. La solution générale est $h(u) = \frac{1}{2}\ f(u) + \kappa\ h_0(u)$ où h_0 est la solution de:

$$h(u) + u\ h\left(\frac{1}{u}\right) = 0 \tag{5.26}$$

On en déduit aisément que:

$$h(u) = \frac{1}{2}\ f(u) + \kappa\ \sqrt{u}\ \left(\breve{h}(u) - \breve{h}\left(\frac{1}{u}\right)\right) \tag{5.27}$$

où $\breve{h}$ est n'importe quelle fonction convexe définie sur $[0,1[$ et nulle en dehors.

Parmi toutes celles listées précédemment, les seules candidates possibles sont les *divergences de Kullback, de Hellinger, et de Lin* (L-divergence). En effet, la seule χ^2-divergence d'ordre α – ou entropie d'ordre α de Havrda-Charvát – qui satisfasse (5.25) est la χ^2-divergence d'ordre $\frac{1}{2}$ (i.e. Hellinger), et la seule divergence de Bose-Einstein d'ordre α qui satisfasse (5.25) est la divergence de Bose-Einstein d'ordre 1 (i.e. Lin).

Pour Kullback et Lin, l'entropie associée est celle de Shannon. Pour Hellinger, l'entropie associée à $f(u) = \frac{1}{2}\ (\sqrt{u}-1)^2$ par $h(u) = \frac{1}{2}\ f(u)$ est l'entropie d'ordre $\frac{1}{2}$ de Havrda-Charvát.

Les **J** et **K**-divergences de Rao que l'on obtient à partir de ces entropies (5.27) ne sont pas en général des f-divergences (sauf dans le cas de l'entropie de Shannon).

f-divergence et distance de Bregman

Point de vue de la métrique En considérant là encore les métriques associées et résolvant:

$$q\ f\left(\frac{p}{q}\right) = h(p) - h(q) + (q-p)h'(q)$$

on obtient de même:

$$f(u) = \gamma\ (u\ \ln u - u + 1)$$

où $\gamma = f''(1)$. Autrement dit:

$$\mathbf{I}_f = \mathbf{D}_H \Leftrightarrow \mathbf{H} = \gamma\ \mathbf{H}_1, \quad \mathbf{I}_f = \gamma\ \bar{\mathbf{K}}$$

Une f-divergence ne peut être une distance de Bregman que pour l'entropie de Shannon, et seulement si elle est proportionnelle à l'information de Kullback (ou entropie relative) [57].

Point de vue de l'antisymétrie En utilisant le caractère antisymétrique de la dérivée première de Bregman :

$$\frac{\partial}{\partial P}\mathbf{D}_H(P,Q) = \frac{\partial H}{\partial P}(Q) - \frac{\partial H}{\partial P}(P)$$

et la dérivée de la f-divergence :

$$\frac{\partial}{\partial P}\mathbf{I}_f(P,Q) = \int f'\left(\frac{p}{q}\right)\ d\lambda(x)$$

on obtient la condition :

$$f'(u) + f'\left(\frac{1}{u}\right) = 0$$

(Noter que $f(u) = u\ \ln u - u + 1$ la satisfait). Elle entraîne :

$$\begin{aligned} f'(u) &= \phi\left(u - \frac{1}{u}\right) \quad \text{avec} \quad \phi(-u) = -\phi(u) \\ f'(u) &= \psi(u) - \psi\left(\frac{1}{u}\right) \quad \text{avec} \quad \psi \text{ quelconque} \end{aligned}$$

5.2.2 Formes non intégrales

On considère donc les formes (2.51) et (2.88) :

$$\mathbf{H}_\psi(P) = \psi^{-1}\left(\int p(x)\ \psi(-\ln p(x))\ d\lambda(x)\right) \tag{5.28}$$

$$\mathbf{I}_\psi(P,Q) = -\psi^{-1}\left(\int \frac{p(x)}{q(x)}\ \psi\left(-\ln\frac{p(x)}{q(x)}\right)\ q(x)\ d\lambda(x)\right) \tag{5.29}$$

et :

$$\mathbf{H}_\phi(P) = -\ln\phi^{-1}\left(\int p(x)\ \phi(p(x))\ d\lambda(x)\right) \tag{5.30}$$

$$\mathbf{I}_\phi(P,Q) = \ln\phi^{-1}\left(\int \frac{p(x)}{q(x)}\ \phi\left(\frac{p(x)}{q(x)}\right)\ q(x)\ d\lambda(x)\right) \tag{5.31}$$

Dans le cas des entropies de cette forme, les seules divergences dérivées de ces entropies qui soient calculables sont la différence de Jensen (et la **J**-divergence) et la distance de Bregman, puisque les **K** et **L**-divergences supposent que l'entropie soit sous forme intégrale.

D'autre part, on ne peut plus maintenant considérer commodément l'argument de la métrique associée. La métrique correspondant aux f-divergences de ce type est toujours celle de Fisher (5.14). Mais les calculs faits plus haut pour les « divergences» de Rényi montrent que les calculs de Hessien se passent mal pour les entropies de ce type.

On se base donc plutôt sur les extensions de la différence de Jensen et de la distance de Bregman discutées précédemment.

f-divergence et distance de Bregman On considère les distances de Bregman généralisées définies par :

$$\mathcal{D}_\psi(P,Q) = \frac{\partial}{\partial\beta}\,\mathcal{J}_\psi^{(\beta)}(P,Q)\bigg|_{\beta=0} \tag{5.32}$$

où $\mathcal{J}_\psi^{(\beta)}$ est une des extensions de la différence de Jensen introduites plus haut.

Le cas du rayon d'information de Sibson (4.44) montre que la χ^2-divergence d'ordre α $\mathbf{R}_\alpha$ est à la fois une f-divergence et une distance de Bregman généralisée. Le fait que $\mathbf{R}_\alpha$ soit une distance de Bregman est connu dans le cas des α-familles, où $\mathbf{R}_{(1-\alpha)/2} = \mathbf{A}_\alpha$ [13].

Annexe - Exemples : familles exponentielles et cas Gaussiens

A1. Cas des familles exponentielles

$$p_\theta(x) = e^{\theta x - \psi(\theta)} \tag{6.33}$$

On suppose que la mesure de référence λ est celle de Lebesgue; important pour le calcul des entropies, mais pas pour celui des divergences considérées qui sont homogènes de degré 1 (sauf la distance géodésique).

Entropies

Shannon

$$\begin{aligned} \mathbf{H}_1(\theta) &= \psi(\theta) - \theta\ \psi'(\theta) \\ &= -\psi^*(\psi'(\theta)) \end{aligned}$$

Havrda-Charvát

$$\mathbf{H}_\alpha(\theta) = \frac{1}{\alpha - 1}\ \left(1 - e^{\psi(\alpha\theta) - \alpha\ \psi(\theta)}\right)$$

Rényi

$$\widetilde{\mathbf{H}}_\alpha(\theta) = \frac{1}{\alpha - 1}\ [\alpha\ \psi(\theta) - \psi(\alpha\theta)]$$

Divergences

Kullback [105]

$$\bar{\mathbf{K}}(\theta_1, \theta_2) = \psi(\theta_2) - \psi(\theta_1) + (\theta_1 - \theta_2)\ \psi'(\theta_1) = -\mathbf{D}_\psi(\theta_2, \theta_1)$$

Rényi [105]

$$\widetilde{\mathbf{R}}_\alpha(\theta_1, \theta_2) = \frac{1}{\alpha\ (\alpha - 1)}\ \mathbf{J}_\psi^{(\alpha)}(\theta_1, \theta_2)$$

A2. Cas Gaussien scalaire $\mathcal{N}(\theta, \sigma^2)$

Entropies

– Par rapport à la mesure $d\lambda(x) = \frac{\sigma}{\sqrt{2\pi}}\ e^{-x^2/2}\ dx$, avec $\psi(\theta) = \frac{\theta^2}{2\ \sigma^2}$:

$$\begin{aligned} \mathbf{H}_1(\theta) &= -\frac{\theta^2}{2\ \sigma^2} \\ \mathbf{H}_\alpha(\theta) &= \frac{1}{\alpha - 1}\left(1 - e^{\alpha\ (\alpha - 1)\ \frac{\theta^2}{2\ \sigma^2}}\right) \\ \widetilde{\mathbf{H}}_\alpha(\theta) &= -\alpha\ \frac{\theta^2}{2\ \sigma^2} \\ &= \alpha\ \mathbf{H}_1(\theta) \end{aligned}$$

– Par rapport à la mesure de Lebesgue :

$$\begin{aligned}
\mathbf{H}_1(\theta) &= \frac{1}{2}\ln(2\ \pi\ e\ \sigma^2) \\
\mathbf{H}_\alpha(\theta) &= \frac{1}{\alpha-1}\left(1-\frac{1}{\sqrt{\alpha\ (2\ \pi\ \sigma^2)^{\alpha-1}}}\right) \\
\widetilde{\mathbf{H}}_\alpha(\theta) &= -\frac{1}{2}\ln\theta+\frac{1}{2}\ln(2\ \pi)+\frac{1}{2(\alpha-1)}\ \ln\alpha \\
&= \mathbf{H}_1(\theta)-\frac{1}{2}\left(1+\frac{\ln\alpha}{1-\alpha}\right)
\end{aligned}$$

Divergences

Kullback

$$\bar{\mathbf{K}}(\theta_1,\theta_2)=\frac{(\theta_1-\theta_2)^2}{2\ \sigma^2}$$

Rényi

$$\widetilde{\mathbf{R}}_\alpha(\theta_1,\theta_2)=\frac{(\theta_1-\theta_2)^2}{2\ \sigma^2}$$

Géodésique

$$d(\theta_1,\theta_2)=\frac{|\theta_1-\theta_2|}{\sigma}$$

A3. Cas Gaussien scalaire $\mathcal{N}(\mu,1/\theta)$

Entropies Par rapport à la mesure de Lebesgue :

$$\begin{aligned}
\mathbf{H}_1(\theta) &= -\frac{1}{2}\ln\theta+\frac{1}{2}\ln(2\ \pi\ e) \\
\mathbf{H}_\alpha(\theta) &= \frac{1}{\alpha-1}\left(1-\frac{1}{\sqrt{\alpha\ (2\ \pi\ \sigma^2)^{\alpha-1}}}\right) \\
\widetilde{\mathbf{H}}_\alpha(\theta) &= -\frac{1}{2}\ln\theta+\frac{1}{2}\ln(2\ \pi)+\frac{1}{2(\alpha-1)}\ \ln\alpha \\
&= \mathbf{H}_1(\theta)-\frac{1}{2}\left(1+\frac{\ln\alpha}{1-\alpha}\right)
\end{aligned}$$

Divergences

Kullback

$$\begin{aligned}
\bar{\mathbf{K}}(\theta_1,\theta_2) &= -\frac{1}{2}\ (\ln\kappa-\kappa+1) \\
&= \frac{1}{2}\ r_1(\kappa)=\frac{1}{2}\ z_1(\kappa)
\end{aligned}$$

où $\kappa=\frac{\theta_2}{\theta_1}$ et où r_α et z_α sont définies en (2.77) et (2.79).

Rényi

$$\begin{aligned}
\widetilde{\mathbf{R}}_\alpha(\theta_1,\theta_2) &= \frac{1}{2\ \alpha\ (\alpha-1)}\ [-(\alpha-1)\ \ln\kappa-\ln(1-(\alpha-1)(\kappa-1))_+] \\
&= \frac{1}{2}\ z_\alpha(\kappa) \\
&= \frac{1}{\alpha\ (\alpha-1)}\ \mathbf{J}^{(\alpha)}_{H_1}(\theta_1,\theta_2)
\end{aligned}$$

Or :

$$\mathbf{J}_{H_1}^{(\alpha)}(\theta_1,\theta_2) = \mathbf{J}_{\widetilde{H}_\alpha}^{(\alpha)}(\theta_1,\theta_2)$$

Géodésique

$$d(\theta_1,\theta_2) = \frac{\sqrt{2}}{2}\,|\ln\theta_1 - \ln\theta_2|$$

a4. Cas Gaussien vectoriel $\mathcal{N}_r(\mu,\Sigma)$

Entropies Par rapport à la mesure de Lebesgue :

$$\mathbf{H}_1(\mu,\Sigma) = -\frac{1}{2}\ln\left|\Sigma^{-1}\right| + \frac{r}{2}\ln(2\,\pi\,e) \tag{6.34}$$

$$\mathbf{H}_\alpha(\mu,\Sigma) = \frac{1}{\alpha-1}\left(1 - \frac{1}{\sqrt{\alpha}}\,(2\pi)^{r(1-\alpha)/2}\,|\Sigma|^{(1-\alpha)/2}\right) \tag{6.35}$$

$$\widetilde{\mathbf{H}}_\alpha(\mu,\Sigma) = -\frac{1}{2}\ln\left|\Sigma^{-1}\right| + \frac{r}{2}\ln(2\,\pi) - \frac{1}{2(1-\alpha)}\,\ln\alpha \tag{6.36}$$

$$= \mathbf{H}_1(\mu,\Sigma) - \frac{1}{2}\left(1 + \frac{\ln\alpha}{1-\alpha}\right) \tag{6.37}$$

Divergences

Kullback

$$\bar{\mathbf{K}}(\mu_1,\Sigma_1,\mu_2,\Sigma_2) = \frac{1}{2}\left(\,(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_1\,\Sigma_2^{-1} - I_r) - \ln\left|\Sigma_1\,\Sigma_2^{-1}\right|\right)$$

$$\mathbf{K}(\mu_1,\Sigma_1,\mu_2,\Sigma_2) = \frac{1}{2}\left(\,(\mu_2-\mu_1)^T\left(\Sigma_1^{-1}+\Sigma_2^{-1}\right)(\mu_2-\mu_1) + \mathrm{tr}(\Sigma_1\,\Sigma_2^{-1} + \Sigma_2\,\Sigma_1^{-1} - 2I_r)\right)$$

En particulier :

- Pour $\mu_1 = \mu_2$:

$$\bar{\mathbf{K}}(\Sigma_1,\Sigma_2) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_1\,\Sigma_2^{-1} - I_r) - \ln\left|\Sigma_1\,\Sigma_2^{-1}\right|\right)$$

  Cette distance entre matrices a été utilisée, pour qualifier une méthode d'identification, en [16] par exemple. Pour $\Sigma_2 = I_r$:

$$\bar{\mathbf{K}}(\Sigma,I_r) = \frac{1}{2}\left(\mathrm{tr}(\Sigma - I_r) - \ln|\Sigma|\right)$$

- Pour $\Sigma_1 = \Sigma_2 = \Sigma$:

$$\bar{\mathbf{K}}(\mu_1,\mu_2) = \frac{1}{2}\,\mathcal{M}^2(\mu_1,\mu_2) = 4\,\mathcal{B}(\mu_1,\mu_2)$$

  où :

$$\mathcal{M}^2(\mu_1,\mu_2) = (\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)$$

  est la distance de Mahalanobis. En fait, dans ce cas Gaussien homoscédastique, il est facile de montrer, à l'aide du changement de variable :

$$y = (x-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1)/\mathcal{M}(\mu_1,\mu_2)$$

  que toute f-divergence (2.63) est une fonction croissante de $\mathcal{M}$ [11]. Le rayon d'information de Sibson (4.42), calculé sous la contrainte de barycentre Gaussien (voir plus haut), est aussi une fonction croissante de $\mathcal{M}$ [146] :

$$\mathcal{S}_1 = \frac{1}{2}\,\ln\left(1 + \frac{1}{4}\,\mathcal{M}^2\right)$$

  A rapprocher de l'expression de l'information d'ordre -1 de Rényi en fonction de la χ^2-divergence (2.90).

Rényi

$$\begin{aligned}\widetilde{\mathbf{R}}_\alpha(\mu_1,\Sigma_1,\mu_2,\Sigma_2) &= \frac{1}{2\ \alpha\ (\alpha-1)}\ \left(\alpha\ \ln\left|\Sigma_1^{-1}\right| + (1-\alpha)\ \ln\left|\Sigma_2^{-1}\right| - \ln\left|\alpha\ \Sigma_1^{-1} + (1-\alpha)\ \Sigma_2^{-1}\right|\right)\\ &\quad -\frac{1}{2}(\mu_2-\mu_1)^T\ \left((1-\alpha)\ \Sigma_1 + \alpha\ \Sigma_2\right)^{-1}\ (\mu_2-\mu_1)\\ &= \frac{1}{\alpha\ (\alpha-1)}\ \mathbf{J}_{H_1}^{(\alpha)}(\Sigma_1,\Sigma_2) + \frac{1}{2}(\mu_2-\mu_1)^T\ \left((1-\alpha)\ \Sigma_1 + \alpha\ \Sigma_2\right)^{-1}\ (\mu_2-\mu_1)\end{aligned}$$

où $\mathbf{H}_1$ est donnée en (6.34).

En particulier :

- Pour $\mu_1 = \mu_2$:

$$\widetilde{\mathbf{R}}_\alpha(\Sigma_1,\Sigma_2) = \frac{1}{\alpha\ (\alpha-1)}\ \mathbf{J}_{H_1}^{(\alpha)}(\Sigma_1,\Sigma_2)$$

- Pour $\Sigma_1 = \Sigma_2 = \Sigma$: la dépendance en α disparaît :

$$\widetilde{\mathbf{R}}_\alpha(\mu_1,\mu_2) = \frac{1}{2}(\mu_2-\mu_1)^T\Sigma^{-1}(\mu_2-\mu_1) = \bar{\mathbf{K}}(\mu_1,\mu_2) = \frac{1}{2}\ \mathcal{M}^2(\mu_1,\mu_2) = 2\ \mathcal{B}(\mu_1,\mu_2)$$

Géodésique Soit :

$$\check{\Sigma}_i = \begin{pmatrix} \Sigma_i + \mu_i\mu_i^T & \mu_i \\ \mu_i^T & 1 \end{pmatrix}$$

et :

$$\check{\Sigma} = \begin{pmatrix} \Sigma_1^{-1/2}\ \left(\Sigma_2 + (\mu_2-\mu_1)(\mu_2-\mu_1)^T\right)\ \Sigma_1^{-1/2} & \Sigma_1^{-1/2}\ (\mu_2-\mu_1) \\ (\mu_2-\mu_1)^T\ \Sigma_1^{-1/2} & 1 \end{pmatrix}$$

Alors une borne inférieure (invariante) pour la distance géodésique d est [45] :

$$\begin{aligned}\underline{d}(\check{\Sigma}_1,\check{\Sigma}_2) &= \sqrt{\mathrm{tr}(\ln\check{\Sigma}\ \ln\check{\Sigma}^T)/2}\\ &= \sqrt{\sum_{i=1}^{r+1}(\ln\sigma_i)^2/2}\end{aligned}$$

où les σ_i sont les valeurs propres de $\check{\Sigma}$.

En particulier :

- Pour $\mu_1 = \mu_2$:

$$\underline{d}(\Sigma_1,\Sigma_2) = d(\Sigma_1,\Sigma_2) = \sqrt{\sum_{i=1}^{r}(\ln\sigma_i)^2/2}$$

où les σ_i sont les valeurs propres de $\Sigma_1^{-1}\ \Sigma_2$ [21, 124].

- Pour $\Sigma_1 = \Sigma_2 = \Sigma$:

$$\begin{aligned}\underline{d}(\mu_1,\mu_2) &= \arg\cosh\left(\frac{1}{2}\ \mathcal{M}(\mu_1,\mu_2) + 1\right) \quad [45]\\ d(\mu_1,\mu_2) &= \mathcal{M}(\mu_1,\mu_2) \quad [124]\end{aligned}$$

A5. Cas Gaussien bivarié de corrélation ϱ

Information de Rényi [154]

$$\mathcal{I}_{\widetilde{R}_\alpha}(X_1,X_2) = \begin{cases} \frac{1}{\alpha\ (\alpha-1)}\left(\sqrt{\frac{(1-\varrho)^{1-\alpha}}{1-(1-\alpha)^2\ \varrho}} - 1\right) & (\alpha \neq 0,1)\\ -\frac{1}{2}\ \ln(1-\varrho) & (\alpha = 1)\\ -\frac{1}{2}\ \ln(1-\varrho) + \frac{\varrho}{1-\varrho} & (\alpha = 0)\end{cases}$$

A6. Cas Gaussien: processus centrés de densité spectrale S

Entropies

Shannon

$$\mathbf{H}_1(S) = \frac{1}{4\,\pi}\,\int_{-\pi}^{\pi} \ln S(\omega)\;d\omega + \frac{1}{2}\;\ln(2\,\pi\,e) \tag{6.38}$$

Divergences

Kullback

$$\begin{aligned}
\bar{\mathbf{K}}(S_1,S_2) &= \frac{1}{4\,\pi}\,\int_{-\pi}^{\pi}\left(\frac{S_1}{S_2} - \ln\frac{S_1}{S_2} - 1\right)\,d\omega \\
\mathbf{K}(S_1,S_2) &= \frac{1}{2}\left\|\sqrt{\frac{S_1}{S_2}} - \sqrt{\frac{S_2}{S_1}}\right\|_2^2
\end{aligned}$$

$\bar{\mathbf{K}}$ est la distance d'Itakura-Saïto [25, 151]; $\mathbf{K}$ est la moitié de la distance de Hellinger entre les deux rapports de spectres.

Rényi [151]

$$\widetilde{\mathbf{R}}_\alpha(S_1,S_2) = \frac{1}{\alpha\,(1-\alpha)}\;\mathbf{J}_{H_1}^{(1-\alpha)}(S_1,S_2)$$

où $\mathbf{H}_1$ est donnée en (6.38).

Pour les processus vectoriels à temps continu, voir [91, 92].

Bibliographie

[1] J. Aczél (1948). On mean values. *Bull. Amer. Math. Soc.*, vol.54, pp.392-400.

[2] J. Aczél (1966). *Lectures on Functional Equations and Their Applications*. Academic Press, Mathematics in Science and Engineering, vol.019.

[3] J. Aczél and Z. Daròczy (1975). *On Measures of Information and Their Characterizations*. Academic Press, Mathematics in Science and Engineering, vol.115.

[4] J. Aczél, B. Forte and C.T. Ng (1974). Why the Shannon and Hartley entropies are « natural »?. *Adv. Appl. Proba.*, vol.6, pp.131-146.

[5] J. Aczél and B. Forte (1986). Generalized entropies and the maximum entropy principle. In *Bayesian Entropy and Bayesian Methods in Applied Statistics* (J.H. Justice, Ed.), Cambridge Univ. Press, pp.95-100.

[6] H. Akaïke (1971). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Intern. Symposium on Information Theory*, Tsahkadsor, Arménie, pp.267-281.

[7] H. Akaïke (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Intern. Symposium on Information Theory*, Budapest, B.N. Petrov and F. Caski (eds), pp.267-281.

[8] H. Akaïke (1974). A new look at statistical model identification. *IEEE Trans. Automatic Control*, vol.AC-19, no 6, pp.716-723.

[9] H. Akaïke (1976). Canonical correlation analysis of time series and the use of an information criterion. In *System Identification : Advances and Case Studies* (R.K. Mehra and D.G. Lainotis, eds.), pp.27-96.

[10] H. Akaïke (1977). On entropy maximization principle. In *Application of Statistics* (P.R. Krishnaiah, Ed.), pp.27-41, North Holland.

[11] S.M. Ali and D. Silvey (1966). A general class of coefficients of divergence of one distribution from another. *Jal Royal Stat. Soc. B*, vol.28, no 1, pp.131-142.

[12] S-I. Amari (1982). Differential geometry of curved exponential families – Curvatures and information loss. *Annals Statistics*, vol.10, no 2, pp.357-385.

[13] S-I. Amari (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics, vol.28, Springer-Verlag.

[14] S-I. Amari (1987). Differential geometry of a parametric family of invertible linear systems – Riemannian metric, dual affine connections, and divergence. *Math. Syst. Theory*, vol.20, pp.53-82.

[15] S-I. Amari (1987). Differential geometrical theory of Statistics. In *Inst. Math. Stat., Lecture Notes - Monograph Series*, vol.10, S. Gupta (Ed.), pp.19-94.

[16] B.D.O. Anderson, J.B. Moore and R.M. Hawkes (1978). Model approximations via prediction error identification. *Automatica*, vol.14, pp.615-622.

[17] T.W. Anderson (1971). *The Statistical Analysis of Time Series*. Series in Probability and Mathematical Statistics, Wiley, New York.

[18] E. Arikan (1996). An inequality on guessing and its application to sequential decoding. *IEEE Trans. Information Theory*, vol.IT-42, no. 1, pp.99-105.

[19] S. Arimoto (1971). Information-theoretical consideration on estimation problems. *Information and Control*, vol.19, no 3, pp.181-194.

[20] S. Arimoto (1975). Information measures and capacity of order alpha for discrete memoryless channels. In *Topics in Information Theory: Colloquia Mathematica Societatis Janos Bolyaï, 16*, I. Csiszár and P. Elias, Eds. August 1975, Keszthely, Hungary, pp.41-52.

[21] C. Atkinson and A.F.S. Mitchell (1981). Rao's distance measure. *Sankhyā A*, vol.43, no 3, pp.345-365.

[22] J.A. Baker, B. Forte and L.F. Lam (1972). On the existence of a collector for a class of information measures. *Utilitas Mathematica*, vol.2, pp.219-239.

[23] O.E. Barndorff-Nielsen (1978). *Information and Exponential Families in Statistical Theory*. Wiley Series in Probability and Mathematical Statistics.

[24] O.E. Barndorff-Nielsen and D.R. Cox (1989). *Asymptotic Methods for Use in Statistics*. Chapman and Hall Monographs on Statistics and Applied Probability.

[25] M. Basseville (1989). Distance measures for signal processing and pattern recognition. *Signal Processing*, vol.18, no 4, pp.349-369.

[26] M. Basseville and J.F. Cardoso (1995). On entropies, divergences and mean values. *IEEE Int. Symp. Information Theory*, ISIT'95, Whistler, B.C. Canada.

[27] M. Behara and P. Nath (1973). Additive and non-additive entropies of finite measurable partitions. In *Probability and Information Theory*. Lecture Notes in Mathematics, vol.296, Springer-Verlag, pp.102-138.

[28] M. Ben-Bassat and J. Raviv (1978). Rényi's entropy and the probability of error. *IEEE Trans. Information Theory*, vol.IT-24, no 2, pp.324-331.

[29] A. Ben-Tal, A. Charnes and M. Teboulle (1989). Entropic means. *Jal Math. Anal. Appl.*, vol.139, pp.537-551.

[30] A. Bhattacharrya (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, vol.35, pp.99-109.

[31] P. Billingsley (1965). *Ergodic Theory and Information*. Wiley.

[32] R.E. Blahut (1974). Hypothesis testing and information theory. *IEEE Trans. Information Theory*, vol.IT-20, no 4, pp.405-417.

[33] R.E. Blahut (1987). *Principles and Practice of Information Theory*. Addison-Wesley, Reading, MA.

[34] R.P. Boas and J.L. Brenner (1987). The asymptotic behavior of inhomogeneous means. *Jal Math. Anal. Appl.*, vol.123, pp.262-264.

[35] A.A. Borovkov (1987). *Statistique Mathématique - Estimation et Tests d'Hypothèses*, Mir, Paris.

[36] J.M. Borwein and A.S. Lewis (1991). Duality relationships for entropy-like minimization problems. *SIAM Jal Control and Optimization*, vol.29, no 2, pp.325-338.

[37] H. Bozdogan (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics, Part A: Theory and Methods*, vol.19, no 1, pp.221-278.

[38] L.M. Bregman (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Comput. Math. and Math. Phys.*, vol.7, pp.200-217.

[39] T.A. Brown (1963). Entropy and conjugacy. *Ann. Math. Stat.*, vol.34, pp.226-232.

[40] S.P. Bruzzone and M. Kaveh (1984). Information tradeoffs in using the sample autocorrelation function in ARMA parameter estimation. *IEEE Trans. Acoustics, Speech, Signal Processing*, vol.ASSP-32, no 4, pp.701-715.

[41] J. Burbea and C.R. Rao (1982a). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *Jal Multivariate Anal.*, vol.12, pp.575-596.

[42] J. Burbea and C.R. Rao (1982b). On the convexity of some divergence measures based on entropy functions. *IEEE Trans. Information Theory*, vol.IT-28, no 3, pp.489-495.

[43] J. Burbea and C.R. Rao (1982c). On the convexity of higher order Jensen differences based on entropy functions. *IEEE Trans. Information Theory*, vol.IT-28, no 6, pp.961-963.

[44] J. Burbea and C.R. Rao (1984). Differential metrics in probability spaces. *Proba. and Math. Stat.*, vol.3, no 2, pp.241-258.

[45] M. Calvo and J.M. Oller (1990). A distance between multivariate normal distributions based in an embedding into the Siegel group. *Jal Multivariate Analysis*, vol.35, no 2, pp.223-242.

[46] L.L. Campbell (1965). A coding theorem and Rényi's entropy. *Information and Control*, vol.8, pp.423-429.

[47] L.L. Campbell (1985). The relation between information theory and the differential geometry approach to statistics. *Information Sciences*, vol.35, pp.199-210.

[48] T.M. Cover and B. Gopinath (eds.) (1987). *Open Problems in Communication and Computation.* Springer.

[49] T.M. Cover and J.A. Thomas (1991). *Elements of Information Theory.* Wiley Series in Telecommunications.

[50] T.M. Cover, P. Gács and R.M. Gray (1989). Kolmogorov's contributions to information theory and algorithmic complexity. *Annals of Probability*, vol.17, no 3, pp.840-865.

[51] I. Csiszár (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarium Mathematicarum Hungarica*, vol.2, pp. 299-318.

[52] I. Csiszár (1974). Information measures: a critical survey. *Proc. 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp.73-86.

[53] I. Csiszár (1975). I-divergence geometry of probability distributions and minimization problems. *Annals Probability*, vol.3, pp.146-158.

[54] I. Csiszár and J. Körner (1981). *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press Series on Probability and Mathematical Statistics.

[55] I. Csiszár (1991). Why least-squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Annals Statistics*, vol.19, no 4, pp.2032-2066.

[56] I. Csiszár (1995a). Generalized cutoff rates and Rényi's information measures. *IEEE Trans. Information Theory*, vol.IT-41, no 1, pp.26-34.

[57] I. Csiszár (1995b). Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, vol.68, no 1-2, pp.161-185.

[58] D. Dacunha-Castelle (1977). Inégalités sur les couples de probabilités. *Ecole d'Eté de Saint-Flour*, chap.3.

[59] A. Dembo and O. Zeitouni (1993). *Large Deviations Techniques and Applications.* Jones and Bartlett Books in Mathematics.

[60] A. Dold and B. Eckmann (eds.) (1974). *Théories de l'Information.* Lecture Notes in Mathematics, vol.398, Springer.

[61] B. Efron (1980). A distance theorem for exponential families. *Proba. and Math. Stat.*, vol.1, no 1, pp.95-98.

[62] P.E. Ferreira (1981). Extending Fisher's measure of information. *Biometrika*, vol.68, no 3, pp.695-698.

[63] T.S. Han (1975). Linear dependence structure of the entropy space. *Information and Control*, vol.29, pp.337-368.

[64] T.S. Han (1978). Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, vol.36, no 2, pp.133-156.

[65] S.E. Fienberg and D.V. Hinkley (1980). *R.A. Fisher: an Appreciation*. Lectures Notes in Statistics, vol.1.

[66] R.A. Fisher (1925). Theory of statistical estimation. *Proc. Cambridge Philosophical Society*, vol.22, pp.700-725.

[67] B. de Finetti (1931). Sul concetto di media. *Giornale dell'Istituto Italiano degli Attuari*, vol.2, pp.369-396.

[68] B. Forte and C. Sempi (1976). Maximizing conditional entropies: A derivation of quantal statistics. *Rend. Matematica*, vol.6, pp.551-566.

[69] P. Gács and J. Körner (1973). Common information is far less than mutual information. *Problems of Control and Information Theory*, vol.2, no 2, pp.149-162.

[70] R.G. Gallager (1968). *Information Theory and Reliable Communication*. Wiley.

[71] I.M. Gelfand and A.M. Yaglom (1959). Calculation of the amount of information about a random function contained in another such function. *Am. Math. Soc. Transl.*, Series 2, vol.12, pp.199-246.

[72] A.H. Gray and J.D. Markel (1976). Distance measures for speech processing. *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.ASSP-24, no 5, pp.380-391.

[73] R.M. Gray, A. Buzo, A.H. Gray and Y. Matsuyama (1980). Distortion measures for speech processing. *IEEE Trans. Acoustics, Speech and Signal Processing*, vol.ASSP-28, no 4, pp.367-376.

[74] R.M. Gray (1988). *Probability, Random Processes, and Ergodic Properties*. Springer-Verlag, N.Y.

[75] R.M. Gray (1990). *Entropy and Information Theory*. Springer-Verlag, N.Y.

[76] R.W. Hamming (1980). *Coding and Information Theory*. Prentice Hall.

[77] G.H. Hardy, J.E. Littlewood and G. Pólya (1952). *Inequalities*, 2nd ed, chap.3 et 6. Cambridge Univ. Press.

[78] M.E. Havrda and F. Charvát (1967). Quantification method of classification processes: concept of structural α-entropy. *Kybernetica*, vol.3, pp.30-35.

[79] W. Hengartner and R. Theodorescu (1974). Concentration and information. Dans [60].

[80] C.C. Heyde (1989). Fisher lecture: quasi-likelihood and optimality for estimating functions : some current unifying themes. *Proc. Int.Stat.Inst., 47th Session*, Paris, pp.19-29.

[81] I.A. Ibragimov and R.Z. Khasminskii (1973). On the information in a sample about a parameter. *Proc. 2nd Intern. Symposium on Information Theory*, Budapest, B.N. Petrov and F. Caski (eds.), pp.295-309.

[82] H. Jeffreys (1948). *Theory of Probability*. Oxford University Press.

[83] R.W. Johnson (1979). Axiomatic characterization of the directed divergences and their linear combinations. *IEEE Trans. Information Theory*, vol.IT-25, no 6, pp.709-716.

[84] L.K. Jones and C.L. Byrne (1990). General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Trans. Information Theory*, vol.IT-36, no 1, pp.23-30.

[85] L.K. Jones and V. Trutzer (1989). Computationally feasible high-resolution minimum-distance procedures which extend the maximum-entropy method. *Inverse Problems*, vol.5, pp.749-766.

[86] A.M. Kagan (1963). On the theory of Fisher amount of information. *Soviet. Math. Dokl.*, vol.4, no 4, pp.991-993.

[87] A.M. Kagan, Iu.V. Linnik and C.R. Rao (1973). *Characterization Problems in Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics.

[88] J. KAMPÉ DE FERIET (1974). La théorie généralisée de l'information et la mesure subjective de l'information. Dans [60].

[89] R.E. KASS (1989). The geometry of asymptotic inference. *Statistical Science*, vol.4, no 3, pp.188-234.

[90] D. KAZAKOS and P. PAPANTONI-KAZAKOS (1980). Spectral distance measures between Gaussian processes. *IEEE Trans. Automatic Control*, vol.AC-25, no 5, pp.950-959.

[91] D. KAZAKOS (1982). Spectral distance measures between continuous-time vector Gaussian processes. *IEEE Trans. Information Theory*, vol.IT-28, no 4, pp.679-681.

[92] D. KAZAKOS and P. PAPANTONI-KAZAKOS (1990). *Detection and Estimation*. Computer Science Press.

[93] M.G. KENDALL and A. STUART (1973). *The Advanced Theory of Statistics - vol 1: Distribution Theory*. Charles Griffin and Co. Ltd., London.

[94] A.I. KHINCHIN (1957). *Mathematical Fundations of Information Theory*. Dover Books on Intermediate and Advanced Mathematics.

[95] L. KNOCKAERT (1993). A class of statistical and spectral distance measures based on Bose-Einstein statistics. *IEEE Trans. Signal Processing*, vol.SP-41, no 11, pp.3171-3174.

[96] L. KNOCKAERT (1994). Statistical thermodynamics and natural f-divergences. *Submitted to IEEE Trans. Information Theory*.

[97] H. KOBAYASHI and J.B. THOMAS (1967). Distance measures and related criteria. *Proc. 5th Allerton Conference on Circuit and System Theory*, pp.491-500.

[98] A. KOLMOGOROV (1930). Sur la notion de moyenne. *Atti Accad. Naz. Lincei*, vol.6, no 12, pp.388-391.

[99] S. KULLBACK and R.A. LEIBLER (1951). On information and sufficiency. *Annals Math. Statistics*, vol.22, pp.79-86.

[100] S. KULLBACK (1959). *Information Theory and Statistics*. Wiley, New York (also Dover, New York, 1968).

[101] S. KULLBACK, J.C. KEEGEL and J.H. KULLBACK (1987). *Topics in Statistical Information Theory*. Lecture Notes in Statistics, vol.42.

[102] K.-S. LAU (1985). Characterization of Rao's quadratic entropies. *Sankhyā A*, vol.47, no 3, pp.295-309.

[103] BING LI (1993). A deviance function for the quasi-likelihood method. *Biometrika*, vol.80, no 4, pp.741-753.

[104] MING LI and PAUL VITÁNYI (1993). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag Texts and Monographs in Computer Science.

[105] F. LIESE and I. VAJDA (1987). *Convex Statistical Distances*. Teubner-Texte zur Mathematik, Band 95, Leipzig.

[106] J. LIN (1991). Divergence measures based on the Shannon entropy. *IEEE Trans. Information Theory*, vol.IT-37, no 1, pp.145-151.

[107] R.J. MCELIECE (1977). *The Theory of Information and Coding: a Mathematical Framework for Communication*. Reading, Mass.: Addison-Wesley.

[108] D.L. MCLEISH and C.G. SMALL (1988). *The Theory and Applications of Statistical Inference Functions*. Lecture Notes in Statistics, vol.44, Springer-Verlag.

[109] A.M. MATHAI and P.N. RATHIE (1975). *Basic Concepts in Information Theory and Statistics*. Wiley Eastern Ltd, New Delhi.

[110] D. MIDDLETON (1960). *Introduction to Statistical Communication Theory*. International Series in Pure and Applied Physics.

[111] M.K. MURRAY and J.W. RICE (1993). *Differential Geometry and Statistics*. Chapman and Hall.

[112] V.M. Nagumo (1930). Über eine klasse der mittelwerte. *Japanese Jal Mathematics*, pp.71-79.

[113] P. Nath (1975). On a coding theorem connected with Rényi's entropy. *Information and Control*, vol.29, no 3, pp.234-242.

[114] J.M. Oller (1993). On an intrinsic analysis of statistical estimation. In *Multivariate Analysis: Future Directions 2*, C.M. Cuadras and C.R. Rao (eds.), Elsevier, pp.421-437.

[115] J.M. Oller and J.M. Corcuera (1995). Intrinsic analysis of statistical estimation. *Annals Statistics*, vol.23, no 5, pp.1562-1581.

[116] Papaioannu and Kempthorne (1971). *On Statistical Information Theory and Related Measures of Information.*

[117] E. Parzen (1990). Time series, statistics and information. *New Directions in Time Series Analysis I*, Springer, IMA vol. 45, pp.265-286.

[118] E. Parzen (1993). Stationary time series analysis using information and spectral analysis. In *Developments in Time Series Analysis* (T.S. Rao, Ed.), pp.139-148, Chapman and Hall.

[119] A. Perez (1984). Barycenter of a set of probability measures and its application in statistical decision. In *COMPSTAT: Lectures in Computational Statistics*, vol.3, J.M. Chambers, J. Gordesch, A. Klas, L. Lebart and P.P.Sint (Eds.), Physica-Verlag, Vienna.

[120] M.S. Pinsker (1964). *Information and Information Stability of Random Variables and Processes.* Holden Day, San Francisco.

[121] H.V. Poor (1980). Robust decision design using a distance criterion. *IEEE Trans. Information Theory*, vol.IT-26, no 5, pp.575-587.

[122] C.R. Rao (1982). Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, vol.21, no 1, pp.24-43.

[123] C.R. Rao (1986). Rao's axiomatization of diversity measures. *Encyclop. Stat. Sciences*, vol.7 (S. Kotz and N.L. Johnson, eds.), pp.614-617.

[124] C.R. Rao (1987). Differential metrics in probability spaces. In *Inst. Math. Stat., Lecture Notes – Monograph Series*, vol.10, S. Gupta (Ed.), pp.217-240.

[125] C.R. Rao and T.K. Nayak (1985). Cross entropy, dissimilarity measures, and characterizations of quadratic entropy. *IEEE Trans. Information Theory*, vol.IT-31, no 5, pp.589-593.

[126] J.C.W. Rayner and D.J. Best (1989). *Smooth Tests of Goodness of Fit.* Oxford University Press.

[127] T.R.C. Read and N.A.C. Cressie (1988). *Goodness of Fit Statistics for Discrete Multivariable Data.* Springer.

[128] E. Regazzini (1992). Concentration comparisons between probability measures. *Sankhyā B*, vol.54, part 2, pp.129-149.

[129] A. Rényi (1959). On measures of dependence. *Acta Mathematica*, vol.10, pp.441-451.

[130] A. Rényi (1961). On measures of entropy and information. *4th Berkeley Symp. Math. Stat. and Proba.*, vol.I, pp.547-561, Univ. California Press.

[131] A. Rényi (1967). On some basic problems of statistics from the point of view of information theory. *5th Berkeley Symp. Math. Stat. and Proba.*, vol.I, pp.531-543, Univ. California Press.

[132] F.M. Reza (1961). *An Introduction to Information Theory.* McGraw-Hill Electrical and Electronic Engineering Series.

[133] J. Rissanen (1978). Modeling by shortest data description. *Automatica*, vol.14, pp.465-471.

[134] J. Rissanen (1986). Stochastic complexity and modeling. *Annals of Statistics*, vol.14, no 3, pp.1080-1100.

[135] J. Rissanen (1987). Stochastic complexity. *Jal Royal Statistical Society*, B, vol.49, no 3, pp.223-239.

[136] J. Rissanen (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.

[137] J. Rissanen (1996). Fisher information and stochastic complexity. *IEEE Trans. Information Theory*, vol.IT-42, no. 1, pp 40-47.

[138] R.T. Rockafellar (1970). *Convex Analysis*. Princeton University Press.

[139] P.K. Sahoo and A.K.C. Wong (1988). Generalized Jensen difference based on entropy functions. *Kybernetika*, vol.24, no 4, pp.241-250.

[140] C.E. Shannon (1957). Some geometrical results in channel capacity. *Nachrichtentechnische Zeitschrift*, vol.10, no 1, pp.1-4.

[141] M.P. Schützenberger (1953). *Contribution aux Applications Statistiques de la Théorie de l'Information*. Thèse d'État, Inst. Stat. Univ. Paris.

[142] C.E. Shannon (1949). *The Mathematical Theory of Communication*. Urbana, University of Illinois Press.

[143] C.E. Shannon (1993). *Claude Elwood Shannon : Collected Papers*. Edited by N.J.A. Sloane and A.D. Wyner, New York : IEEE Press.

[144] R. Shibata (1989). Statistical aspects of model selection. In *From data to model*, (J.C. Willems, Ed.), Springer, pp.215-240.

[145] J.E. Shore and R.W. Johnson (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. Information Theory*, vol.IT-26, no 1, pp.26-37.

[146] R. Sibson (1969). Information radius. *Z. Wahrscheinlichkeitsth. Werw. Gebiete*, vol.14, no 2, pp.149-160.

[147] H.J. Skaug and D. Tjostheim (1996). Testing for serial independence using measures of distance between densities. In *Hannan Volume*.

[148] D. Slepian (Ed.) (1974). Key papers in the development of information theory. *IEEE Press Selected Reprint Series*.

[149] A.A. Stoorvogel and J.H. Van Schuppen (1994). An $\mathcal{H}_\infty$ parameter estimator and its interpretation. *SYSID'94, 10th IFAC Symposium on System Identification*, Copenhague.

[150] A.A. Stoorvogel and J.H. Van Schuppen (1995). System identification with information theoretic criteria. *CWI Report BS-R9513*.

[151] S. Sugimoto and T. Wada (1988). Spectral expressions of information measures of Gaussian time series and their relation to AIC and CAT. *IEEE Trans. Information Theory*, vol.IT-34, no 4, pp.625-631.

[152] I.J. Taneja (1989). On generalized information measures and their applications. *Advances in Electronics and Electron Physics*, vol.76, pp.327-413.

[153] I. Vajda (1971). χ^α-divergence and generalized Fisher's information. *Proc. 6th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, pp.873-886.

[154] I. Vajda (1989). *Theory of Statistical Inference and Information*. Kluwer.

[155] I. Vajda (1990). Generalization of discrimination-rate theorems of Chernoff and Stein. *Kybernetika*, vol.26, no 4, pp.273-288.

[156] S.M. Veres (1990a). Relations betwen information criteria for model-structure selection. Part 1: The role of bayesian model order estimation. *Int. Jal Control*, vol.52, no 2, pp.389-408.

[157] S.M. Veres (1990b). Relations betwen information criteria for model-structure selection. Part 2: Modeling by shortest data description. *Int. Jal Control*, vol.52, no 2, pp.409-421.

[158] S.M. Veres (1991). *Structure Selection of Stochastic Dynamic Systems, the Information Criterion Approach*. Stochastic Monographs, Gordon and Breach Science Publishers, NY.

[159] S. Watanabe (1960). Information theoretic analysis of multivariate correlation. *IBM Jal Research and Development*, vol.4, no 1, pp.66-81.

[160] S. Watanabe (1969). *Knowing and Guessing: a Quantitative Study of Inference and Information*. Wiley.

[161] H.L. Weidemann and E.B. Stear (1969). Entropy analysis of parameter estimation. *Information and Control*, vol.14, no 6, pp.493-506.

[162] P. Whittle (1953). Estimation and information in stationary time series. *Arkiv for Matematik*, vol.2, pp.423-434.

[163] J. Wolfowitz (1978). *Coding Theorems in Information Theory*. Springer, NY. (3rd ed.).

[164] G. Zhang and M. Taniguchi (1994). Discrimination analysis for stationary time series. *Jal Time Series Analysis*, vol.15, pp.117-126.

[165] J. Ziv and M. Zakaï (1973). On functionals satisfying a data processing theorem. *IEEE Trans. Information Theory*, vol.IT-19, pp.275-282.

[166] W.H. Zurek (Ed.) (1990). *Complexity, Entropy and the Physics of Information*. A Proceedings volume in the Santa Fe Institute Studies in the Sciences of Complexity, vol. VIII, Addison Wesley Publishing Co.

Table des matières