Dependence, Correlation and Gaussianity in Independent Component Analysis

Cardioso – Résumé

De Lara, Tilquin, Vidal

1 Définitions préalables et propriétés

1.1 Définitions

Pour une variable aléatoire $Y \in \mathbb{R}^n$, on notera Y_i sa i-ième composante. Jusqu'à la fin, pour X une variable aléatoire à valeur dans $\mathbb{R}^n$, on notera $P(X)$ sa densité de probabilité et Σ_X sa matrice de covariance.

On posera aussi $\mathcal{G}$ l'ensemble des distributions gaussiennes, $\mathcal{P}$ l'ensemble des distributions « produits » (indiquant une indépendance des composantes). Pour une distribution P, on définira alors $P^{\mathcal{G}}$, $P^{\mathcal{P}}$ et $P^{\mathcal{G}\wedge\mathcal{P}}$ les distributions respectivement gaussiennes, produit et gaussienne produit minimisant la valeur de leur divergence par rapport à P. On verra ces distributions comme des projections sur ces différents espaces.

On définit alors les grandeurs suivantes.

La **divergence de Kullback–Leibler** de la distribution Q par rapport à P :

$$\mathrm{K}(P \parallel Q) = \int_{\mathbb{R}^n} P(x) \log \frac{P(x)}{Q(x)} \mathrm{d}x. \tag{1}$$

L'**entropie** de Y :

$$\mathrm{H}(P) = -\int_{\mathbb{R}^n} P(x) \log P(x) \mathrm{d}x. \tag{2}$$

L'**information mutuelle**, que l'on prendra pour mesure d'indépendance :

$$\mathrm{I}(Y) = \mathrm{K}\Big(P(Y) \parallel \Pi_i P_i(Y_i)\Big) = \mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{P}}\Big). \tag{3}$$

La **Non-Gaussianité** de Y :

$$\mathrm{G}(Y) = \mathrm{K}\Big(Y \parallel \mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\Sigma_Y\right)\Big) = \mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{G}}\Big). \tag{4}$$

La **Corrélation** de Y :

$$\begin{aligned} \mathrm{C}(Y) &= \mathrm{K}\Big(\mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\Sigma_Y\right) \parallel \mathcal{N}\left(\mathbb{E}[\,Y\,]\,,\,\mathrm{Diag}\,\Sigma_Y\right)\Big) \\ &= \mathrm{K}\Big(P(Y)^{\mathcal{G}} \parallel P(Y)^{\mathcal{P}\wedge\mathcal{G}}\Big) \\ &= \frac{1}{2} \log \frac{\det\left(\mathrm{Diag}(\Sigma_Y)\right)}{\det\left(\Sigma_Y\right)}. \end{aligned} \tag{5}$$

Dans la suite, on se permettra, pour éviter des notation trop lourde, d'écrire l'information mutuelle, la non-gaussianité et la corrélation d'une distribution.

1.2 Propriétés

Par propriété, si $\mathrm{K}(P \parallel Q) = 0$, les distributions P et Q sont égales sur les espaces de mesures non nulles. Remarquons la propriété suivante, si T est une matrice inversible et μ est un vecteur quelconque, on a :

$$\mathrm{K}\Big(P(Y) \parallel P(Z)\Big) = \mathrm{K}\Big(P(\mu + TY) \parallel P(\mu + TZ)\Big)\,. \tag{6}$$

Par définition, $\mathrm{I}(Y) = 0$, on aura $P(Y) = \prod_i P_i(Y_i)$ et les composantes de Y seront indépendantes. On a la relation, d'où découle l'égalité de la définition de l'indépendance mutuelle

$$\mathrm{K}\left(P(Y) \parallel \prod_i Q_i\right) = \mathrm{I}(Y) + \sum_i \mathrm{K}\Big(P(Y_i) \parallel Q_i\Big)\,.$$

Si T est une matrice inversible et μ est un vecteur quelconque, on a l'égalité suivante :

$$\mathrm{G}\Big(P(Y)\Big) = \mathrm{G}\Big(P(\mu + TY)\Big)\,. \tag{7}$$

1.3 Relations

On a les relations suivantes, découlant de la définition de la divergence de Kullback–Leibler ou du « théorème de Pythagore » :

$$\mathrm{I}(Y) = \sum_i \mathrm{H}(P_i) - \mathrm{H}(P)\,, \tag{8}$$

$$\mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{P}\wedge\mathcal{G}}\Big) = \mathrm{I}(Y) + \sum_i \mathrm{G}(Y_i)\,,$$

$$\mathrm{K}\Big(P(Y) \parallel P(Y)^{\mathcal{P}\wedge\mathcal{G}}\Big) = \mathrm{G}(Y) + \mathrm{C}(Y)\,,$$

$$\boxed{\mathrm{I}(Y) + \sum_i \mathrm{G}(Y_i) = \mathrm{G}(Y) + \mathrm{C}(Y)\,.} \tag{9}$$

Cette dernière relation nous donne que minimiser la grandeur $I\,(Y)$ revient à minimiser $\mathrm{C}(Y) - \sum_i \mathrm{G}(Y_i)$, la Non-gaussianité de Y étant indépendante ici du changement de référentiel que l'on recherche.

2 Géométrie

2.1 Théorème de Pythagore

On considère ici les deux variétés $\mathcal{P}$ et $\mathcal{G}$, comme définies précédemment. Remarquons que ces deux variétés sont toutes les deux des familles exponentielles et ainsi, elles vérifient toutes les deux la propriétés suivantes. Si on y prend deux distributions p et q, alors toutes les distributions du segment exponentiel qu'elles définissent y appartiennent aussi, le segment exponentiel étant défini par

$$w_\alpha(x) = p(x)^{1-\alpha} q(x)^\alpha \mathrm{e}^{-\psi(\alpha)},$$

avec $\psi(\alpha)$ un coefficient de normalisation.

On peut montrer, par ailleurs, que la famille exponentielle $\mathcal{G}$ est de dimension $L_\mathcal{G} = n + \frac{1}{2}n(n+1)$ dans le sens où on peut trouver une mesure de référence $g(x)$ (non nécessairement une distribution) et une « base » de L fonctions scalaire $S_l(x)$ telle que toute distribution s'écrive sous la forme :

$$p_\alpha(x) = g(x) \exp\left(\sum_l \alpha_l S_l(x) - \psi(\alpha)\right),$$

pour $\psi(\alpha)$ une fonction de normalisation et $\alpha \in \mathbb{R}^L$.

Remarquons que dans le cas de $\mathcal{G}$, on peut prendre comme « base » les fonctions $y \mapsto y_i$ et $y \mapsto y_i y_j$ pour i et j dans $[\![1, n]\!]$.

Le Théorème de Pythagore s'énonce de la manière suivante :
Si $\mathcal{E}$ est une famille exponentielle et P est une distribution quelconque, il existe alors une unique distribution $P^{\mathcal{E}}$ de $\mathcal{E}$ vérifiant :

$$\forall Q \in \mathcal{E}, \qquad \mathrm{K}(P \parallel Q) = \mathrm{K}\big(P \parallel P^{\mathcal{E}}\big) + \mathrm{K}\big(P^{\mathcal{E}} \parallel Q\big)\,.$$

On peut voir $P^{\mathcal{E}}$ comme la « projection orthogonale » de P sur la famille $\mathcal{E}$. Par positivité de la divergence de Kullback-Leibler, on peut voir $P^{\mathcal{E}}$ comme la distribution de $\mathcal{E}$ minimisant sa divergence par rapport à P.

On peut alors remarquer que l'équation 9 peut être interprété comme suit : on obtient le même résultat en projetant P d'abord sur $\mathcal{G}$ puis sur $\mathcal{P} \wedge \mathcal{G}$ ou d'abord sur $\mathcal{P}$ puis sur $\mathcal{P} \wedge \mathcal{G}$. Remarquons que tous les termes de cette équations sont invariants par translation et changement d'échelle indépendamment sur les différentes coordonnées. L'espace $\mathcal{G} \wedge \mathcal{P}$ étant de dimension $2n$ (chaque coordonnée correspond à une gaussienne avec 2 paramètres), qui correspond exactement à la dimension de l'ensemble des transformations précédentes, tout se ramène exactement au cas d'une seule distribution dans $\mathcal{G} \wedge \mathcal{P}$.

2.2 Structures marginales

On va ici s'intéresser à l'espace « union » des deux variétés $\mathcal{G}$ et $\mathcal{P}$ que l'on notera $\mathcal{G} \vee \mathcal{P}$ qui correspondra à la plus petite famille exponentielle qui contient $\mathcal{G}$ et $\mathcal{P}$, ce qui revient à considérer exactement toutes les distributions de la forme

$$p(y) = \phi(y) \exp\left(\sum_{l=1}^{L_{\mathcal{G}}} \alpha_l S_l(y) + \sum_{i=1}^{n} r_i(y_i) - \psi\right), \tag{10}$$

avec S_l une « base » de $\mathcal{G}$, r_i des fonctions réelles (puisque l'on a pas de « base » finie pour $\mathcal{P}$), ψ un coefficient de normalisation et $\phi(y)$ une distribution normale homogène ($\mathcal{N}(0,\ I_n)$).

Par le théorème de Pythagore, on trouve que pour toute distribution Q de $\mathcal{G}$ ou de $\mathcal{P}$, on a :

$$\mathrm{K}(P \parallel Q) = \mathrm{K}\big(P \parallel P^{\mathcal{G}\vee\mathcal{P}}\big) + \mathrm{K}\big(P^{\mathcal{G}\vee\mathcal{P}} \parallel Q\big)\,. \tag{11}$$

Ainsi, la divergence minimum sur $\mathcal{G}$ par rapport à P et par rapport à $P^{\mathcal{G}\vee P}$ est atteinte au même point $P^{\mathcal{G}}$, ce qui revient à dire que :

$$\big(P^{\mathcal{G}\vee\mathcal{P}}\big)^{\mathcal{G}} = P^{\mathcal{G}},$$

de même pour $P^{\mathcal{P}}$.

Ainsi, en reprenant l'équation 11 et en l'utilisant avec $Q = P^{\mathcal{P}}$ et avec $Q = P^{\mathcal{G}}$, on trouve les relations suivantes :

$$\begin{array}{rcl} \mathrm{I}(P) & = & \mathrm{K}\big(P \parallel P^{\mathcal{G}\vee\mathcal{P}}\big) + \mathrm{I}\big(P^{\mathcal{G}\vee\mathcal{P}}\big) \\ \mathrm{G}(P) & = & \mathrm{K}\big(P \parallel P^{\mathcal{G}\vee\mathcal{P}}\big) + \mathrm{G}\big(P^{\mathcal{G}\vee\mathcal{P}}\big) \end{array}$$

En supposant que les valeurs des divergences sont suffisamment petites, on peut supposer que la figure formé par $P^{\mathcal{G}\vee\mathcal{P}}$, $P^{\mathcal{G}}$, $P^{\mathcal{P}}$ et $P^{\mathcal{G}\wedge\mathcal{P}}$ et un rectangle. L'égalité des longueurs nous donne alors :

$$\mathrm{I}\big(P^{\mathcal{P}\vee\mathcal{G}}\big) \simeq \mathrm{C}(P) \qquad \text{et} \qquad \mathrm{G}\big(P^{\mathcal{P}\vee\mathcal{G}}\big) \simeq \sum_i \mathrm{G}(P_i)\,.$$

Cela donne, avec les équations précédentes :

$$\begin{array}{rcl} \mathrm{I}(P) & \simeq & \mathrm{K}\big(P \,\|\, P^{\mathcal{G}\vee\mathcal{P}}\big) + \mathrm{C}(P) \\ \mathrm{G}(P) & \simeq & \mathrm{K}\big(P \,\|\, P^{\mathcal{G}\vee\mathcal{P}}\big) + \sum_i \mathrm{G}(P_i) \end{array}$$

La distribution $P^{\mathcal{G}\vee\mathcal{P}}$ peut être interprété comme la distribution la plus simple approchant la distribution P, dans le sens où elle capture la structure marginale de P et sa structure de premier et second ordre (voir équation 10).

3 Cumulant et géométrie locale

Pour rappel, les cumulants κ_n d'une variable aléatoire X sont définis avec la fonction génératrice des cumulants :

$$g(t) = \log \mathbb{E}\big[\,\mathrm{e}^{tX}\,\big] = \sum_{n=1}^{\infty} \frac{\kappa_n}{n!} t^n \tag{12}$$

On s'intéresse ici aux distributions au voisinage de $P^{\mathcal{G}\wedge\mathcal{P}}$, ce qui correspond aux distributions faiblement corrélées et faiblement non-gaussiennes. On assimilera les variétés $\mathcal{P}$ et $\mathcal{G}$ à leur plan tangent et la divergence de Kullback-Leibler à une mesure quadratique.

3.1 Construction des plans tangents

Pour deux distributions $p(x)$ et $n(x)$, on définit la fonction

$$e_p(x) = \frac{p(x)}{n(x)} - 1,$$

qui sera alors d'espérance nulle selon $n(x)$: $\mathbb{E}_{X\sim n}\big[\,e_p(X)\,\big] = 0$.

On cherche alors à identifier les distributions p proches de n aux « petites » fonctions d'espérance nulle selon n. Si on considère une distribution proche de n, la fonction e_p sera *petite* et d'espérance nulle. Réciproquement, pour une fonction e_p données, on considérera alors $p(x) = n(x)\,(e_p(x)+1)$ qui sera bien une distribution proche de $n(x)$.

Ainsi, on peut identifier l'espace vectoriel des variables aléatoires d'espérance nulle et de variance finie au plan tangent à la variété des distributions au point n.

Si on considère deux distributions p et q proches de n, on peut exprimer K_n l'expansion au second ordre suivant e_p et e_q de $\mathrm{K}(p \,\|\, q)$ de la manière suivante :

$$\mathrm{K}_n(p \,\|\, q) = \frac{1}{2}\mathbb{E}_{X\sim n}\Big[\,(e_p(X) - e_q(X))^2\,\Big] \tag{13}$$

3.2 Expansion de Gram-Charlier

3.3 Base d'Hermite

3.4 Approximation de la divergence par des petits cumulants

3.5 Décomposition en quatres parties

3.6 Divergences et objectifs