1 Principe et introduction

On définit les variables et les hypothèses suivantes :

— S matrice de taille $n_1 \times T$: il y a n sources, échantillonnées sur T périodes.

$$S = (s_1, \dots, s_{n_1})^\top, \quad s_i = (s_{i,1}, \dots, s_{i,T}). \tag{1}$$

Aucun de ces signaux est gaussien, il sont tous centrés et normés.

— X sont les signaux observés, il y'en a n_2 échantionnés sur T périodes.

$$X = (x_1, \dots, x_{n_2})^\top, \quad x_i = (x_{i,1}, \dots, x_{i,T}). \tag{2}$$

— Il y a une relation linéaire (sans bruit) pour chacun des signaux observés par rapport aux signaux sources, i.e., pour tout i,

$$x_i = \sum_{d=1}^{n_1} a_{i,d} s_d, \tag{3}$$

ce qui se réécrit

$$X = AS, \tag{4}$$

où $A = (a_{i,j})_{i=1,j=1}^{n_2,n_1}$. Cette relation est indépendante de la période t considérée, la combinaison linéaire est stable dans le temps.

— On suppose ici que $n_1 = n_2 = n$, (à voir comment faire pour la suite).

— On définit $W = A^{-1}$

2 ICA : estimation par non gaussianité

Comme expliqué précedemment, on peut écrire $X = AS$, ou encore $S = WX$, et ici l'assertion principale est la suivante (sous les conditions de non gaussianité de S) : **La somme de variables indépendantes est toujours plus gaussienne que n'importe quelle variable qui la compose** On note également quelques propriétés importantes, puisque S centré et normé, on remarque que pour $w \in \mathbb{R}^n$: $y = w^\top X = w^\top AS$, ainsi si w est une ligne de W on a que $var(y) = var(s_i) = 1$, on peut donc supposer également que A, W sont des matrices orthogonales.

Algorithm 1: le nom de mon algo...projection into Stiefel manifold ?

input : An initial matrix W, a precision parameter $\varepsilon > 0$
output: An output matrix W
$W^{(0)} \leftarrow W$
while $||W^{(k+1)} - W^k|| > \varepsilon$ **do**
  $\nabla J(W_k) = [\text{sgn}(\text{kurt}(w_i^{k,T}X))\mathbb{E}[(w_i^{k,T}X)^3X]]_{i=1}^n$
  $\hat{\nabla} J(W_k) \leftarrow W_k W_k^\top \nabla J(W^k) - W_k \nabla J(W_k)^\top W_k$
  $W^{k+1} \leftarrow W^k + \eta \hat{\nabla} J(W^k)$
end
$S \leftarrow WX$
return S

2.1 Mesures de gaussianité

2.1.1 kurtosis

Soit y une variable aléatoire réelle,

$$\text{kurt}(y) = \mathbb{E}[y^4] - 3\mathbb{E}[y^3]^2. \tag{5}$$

Ainsi pour tout i, on cherche w_i (vecteur ligne de W) tel que $w_i X$ est moins gaussien que X (car cela vaut s_i et que s_i est moins gaussien que x_j pour tout j). Il s'agit donc de maximiser la non gaussianité $w_i^\top X$ Si on utilise cette mesure de non gaussianité, et l'algorithme de descente de gradient usuel on cherche à maximiser ($X(t)$ désigne la colonne t de X)

$$J(w_i) = \sum_{t=1}^{T} |\text{kurt}(w_i^\top X(t))| = \sum_{t=1}^{T} |\mathbb{E}[w_i^\top X(t)]^4 - 3\mathbb{E}[(w_i^\top X(t))^2]^2|, \tag{6}$$

avec la contrainte W orthogonale. On maximise en réalité $\sum_{i=1}^n J(w_i)$. On calcule alors la dérivée par rapport à w_i et on obtient l'algorithme 2. (ici X pour désigner la somme sur t des $X(t)$)

On peut faire le même algorithme en le faisant par vecteur w_i et en utilisant une opérationde gram schmidt pour que les w_i ne convergent pas vers le même vecteur.

2.1.2 neguentropie

On applique un algorithme de descente de gradient similaire au précédent en utilisant cette fois comme mesure la neguentropie. Définissons d'abord l'entropie par :

$$H(y) = -\mathbb{E}(\log(f(Y))) \tag{7}$$

f densité de Y. La neguentropie est définie par :

$$J(y) = H(y_{gauss}) - H(y) \tag{8}$$

ou y_{gauss} est une variable gaussienne avec la même matrice de covariance que y. C'est une estimation couteuse de l anon gaussianité car il faut estimer la fonction de densité de y de manière non paramétrique probablement. On se concentre sur le cas ou y est une variable réelle, cela suffit pour notre usage. On considère l'estimation suivante de la neguentropie :

$$J(y) \propto [\mathbb{E}(G(y)) - \mathbb{E}(G(v))]^2, \tag{9}$$

avec v variable normale centrée réduite et G fonction non quadratique. La descente de gradient donne alors :

La constante gamma joue un rôle d'autodétermination du taux d'apprentissage, elle donne également la direction de l'optimisation. Ici aussi X signifie qu'on fera la somme sur toutes les colonnes de X. **Procédé d'orthonormalisation de gram schmidt** : $w_i = w^{k+1}$ et

$$w_i = w_i - \sum_{j=1}^{i-1} (w_i^\top w_j) w_j$$

$$w_i = \frac{w_i}{||w_i||}$$

$w^{k+1} = w_i$ **end while Compute S**

$$S = WX$$

end A noter ici que comme précédemment on aurait pu faire l'inverse : faire une itération complète de l'algo puis orthogonaliser et symmétriser W, réitérer.

Algorithm 2: le nom de mon algo...projection into Stiefel manifold ?

input : An initial matrix W, a precision parameter $\varepsilon > 0$
output: An output matrix W
$W^{(0)} \leftarrow W$
for $i = 1$ **to** n **do**
  $w_i^{(0)} \leftarrow w$
  while $||w_i^{(k+1)} - w_i^{(k)}|| > \varepsilon$ **do**
    $w_i^{(k)} \leftarrow w$
    $\gamma \leftarrow \mathbb{E}(G(w_i^{(k)\top} X)) - \mathbb{E}(G(v))$
    $\nabla J(w_i) \leftarrow \gamma \mathbb{E}(X g(w_i^{(k)\top} X))$
    $w_i^{k+1} \leftarrow w_i^{(k)} + \nabla J(w_i^{(k)})$
  end
end
$S \leftarrow WX$
return S

2.1.3 FastICA

Mêmes métriques mais on utlise le point fixe au lieu d'utiliser la descente de gradient. En effet en regardant la mise à jour lors des itérations on cherche en réalité (en ignorant γ et η car ils seront supprimés par normalisation). On a alors :

— kurtosis apprcche : Point fixe equation en utilisant que $w \propto \mathbb{E}(X(w^\top W)^3) - 3||w||^2 w$ et en dérivant on obtient :

$$w = \mathbb{E}(X(w^\top X)^3) - 3w$$

. On applique alors récursivement cette équation en prennant soin de normaliser à chaque fois.

— neguentropie approche : Point fixe equation en utilisant que $w = \mathbb{E}(X g(W^\top W))$ et en dérivant (on ignore le dénominateur car normalisation) on obtient alors :

$$w = \mathbb{E}(X g(w^\top X) - \mathbb{E}(g'(w^\top X) W$$

3 ICA : estimation par maximum de vraissemblance

L'ICA s'écrit $X = AS$ ou encore $S = WX$. On note p_i la densité du signal indépendant s_i et p_x la densité jointe des signaux $x_1, \ldots, x_n$. Puisque $s_i = w_i X$,

$$p_x(x) = |\det(W)| \prod_{i=1}^{n} p_i(w_i^\top X)$$

ici il s'agit de la densité empirique et X désigne qu'on approxime en utilisant les T observations de chaque signal. Ainsi on peut noter la likelihood en fonction de la matrice W :

$$L(W) = \prod_{t=1}^{T} \prod_{i=1}^{n} p_i(w_i^\top X(t)) |\det(W)| \tag{10}$$

soit une loglikelihood sous la forme :

$$\log L(W) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log(p_i(w_i^\top X(t)) + T \log |\det(W)| \tag{11}$$

en utilisant que $\mathbb{E}(\log(p_i(w_i X)) = \frac{1}{T} \sum_{t=1}^{T} (\log(p_i(w_i^\top X(t))$, on réecrit l'équation précédente en utilisant cette espérience empirique et donc :

$$\frac{1}{T} \log(L(W)) = \mathbb{E}(\sum_{i=1}^{n} \log(p_i(w_i^\top X)) + \log(|\det(W)|)) \tag{12}$$

Ainsi si on a les densités p_i ceci revient à une estimation paramétrique du paramètre W. Il reste à déterminer une estimation de p_i, il y a deux options :

- A partir de connaissances métier/ préalables, on peut détermikner un prior sur p_i, reste à montrer qu'une erreur raisonnable sur le prior ne change pas considérablement le résultat sur W.
- Déterminer une estimation paramétrique sur p_i, par exemple supposer que p_i peut être égale à f_1 ou f_2.

En réalité, l'exemple donné ci dessus est suffisant.

Théorème : Notons $\tilde{p_i}$ la densité supposée du signal propre s_i, alors on définit :

$$g_(si_i) = \frac{\partial}{\partial s_i} \log(\tilde{p_i}(s_i) = \frac{\tilde{p_i}'(s_i)}{\tilde{p_i}(s_i)} \tag{13}$$

On rappelle que l'estimateur de s_i noté $y_i = \hat{w}_i^\top X$ a la même contrainte que les s_i c'est à dire qu'ils sont non corrélés (ou indépendants à voir) et de variance unitaire. Alors MLE est localement consistant si on suppose que $\forall i, \mathbb{E}(s_i g_i(s_i) - g_i'(s_i)) > 0$ ***Idée de preuve :*** En effet en dérivant la logvraissemblance (sans le detcar par la condition de d'orthogonalité de W pour avoir que less s_i sont bien de variance unitaire, on a $|\det(W)| = 1$) on obtient :

$$\sum_{i=1}^{n} \mathbb{E}(\frac{X p_i'(w_i^\top X}{p_i(w_i^\top X}) = \sum_{i=1}^{n} \mathbb{E}(X g_i(s_i)) \tag{14}$$

ainsi une première condition de consistance est donnée âr $\sum_{i=1}^{n} \mathbb{E}(X g_i(s_i)) = 0$, par linéarité $\sum_{i=1}^{n} w_i^\top \mathbb{E}(X g_i(s_i)) = 0$ soit $\mathbb{E}(s_i g_i(s_i) = 0$ et en ajoutant la contrainte de concavité de la fonction de logvraissemblance $\mathbb{E}(g'(s_i)) < 0$ d'où le résultat. [A VERIFIER]

On note que si $\tilde{p_i}$ change peu (comment quantifier ? ? ?) alors cette assertion reste vraie et le MLE reste consistant.

Corollaire Il suffit donc de savoir si $\tilde{p_i}$ est sous ou sur gaussienne. En effet définissons

$$\log(p_i^+(s)) = \alpha_1 - 2\log(\cosh(s)), \qquad \log(p_i^-(s)) = \alpha_2 - \frac{s^2}{2} + \log(\cosh(s)) \tag{15}$$

Les paramètres α_1, α_2 jouent le rôle de constante de normalisation pour les distributions. Elles ne sont pas utiles pour vérifier l'assertion du théorème. On note que les constantes $2, 0.5$ peuvent ête changées (et donc les constantes α_1, α_2 changent en fonction). La première densité approche celle de laplace, on est dans le cas sur gaussien, la seconde densité est proche d'une loi normale applatie par le $los(\cosh(.))$ et donc sous gaussienne. On a alors pour le cas sur-gaussien (en ignorant les constantes) :

$$g_{i,+}(s_i) = -\tanh(s_i), \quad g_{i,+}'(s_i) = -(1 - \tanh^2(s_i)) \tag{16}$$

$$\mathbb{E}(s_i g_{i,+}(s_) - g_{i,+}'(s_i)) = \mathbb{E}(-\tanh(s_i)s_i + (1 - \tanh(s_i)^2) \tag{17}$$

Pour le cas sous gaussien :

$$g_{i,-}(s_i) = -s_i + \tanh(s_i), g_{i,-}'(s_i) = -\tanh^2(s_i), \tag{18}$$

$$\begin{aligned} \mathbb{E}(s_i g_{i,+}(s_) - g_{i,+}'(s_i)) &= \mathbb{E}(-\tanh(s_i)s_i - s_i^2 + \tanh(s_i)^2) \\ &= \mathbb{E}(\tanh(s_i)s_i - (1 - \tanh(s_i)^2) \end{aligned} \tag{19}$$

Ainsi pour chaque i en calculant ces quantités, on choisit si $\tilde{p}_i = p_{i,+}$ ou si $\tilde{p}_i = p_{i,-}$ en vérifiant quelle condition empirique vérifie l'inégalité du théorème et on a alors une estimation semi paramétrique de W par MLE.

3.1 Algorithme de MLE par descente de gradient

3.1.1 Bell Sejnowski

On a montré que

$$\begin{aligned}\frac{1}{T}\log(L(W)) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}(\log(p_i(w_i^\top X(t))) + \log(|\det(W)|) \\ &= \mathbb{E}(\sum_{i=1}^{n}\log(p_i(w_i^\top X)) + \log(|\det(W)|)\end{aligned} \tag{20}$$

en dérivant cette expression et en remarquant :

$$\partial_{W_{ij}}\log(|\det(W)|) = \frac{cof(W_{i,j})}{\det(W)} = \frac{adj(W_{i,j})^\top}{\det(W^\top)} = (W^\top)^{-1}, \tag{21}$$

$$\frac{1}{T}\frac{\partial\log(L(W))}{\partial W} = (W^\top)^{-1} + \mathbb{E}(\sum_{i=1}^{n} g_i(w_i^\top X)X^\top) \tag{22}$$

en notant $g(y) = (g_1(y_1), .., g_n(y_n))^\top$,

$$\frac{1}{T}\frac{\partial\log(L(W))}{\partial W} = (W^\top)^{-1} + \mathbb{E}(g(WX)X^\top) \tag{23}$$

(On ne simplifie pas $(W^\top)^{-1}$ car on ne va pas procéder à l'orthogonalisation de W à chaque itération ? ?) : On peut utiliser la version stochastique, on peut omettre l'espérance et utiliser un point à la fois.

$$\partial W = (W^\top)^{-1} + g(WX)X^\top \tag{24}$$

3.1.2 Descente de gradient naturelle

On repart de l'incrément précédent ∂W et on multiplie par $W^\top W$ à droite :

$$\partial W = [Id + \mathbb{E}(g(WX)(WX)^\top)]W \tag{25}$$

Ici on montre que l'algorithme converge lorsque $\mathbb{E}(g(y)y^\top) = Id$ c'est à dire lorsque y_i et $g_i(y_i)$ sont non corrélés pour i, j différents.

4 ICA : estimation mutual information

4.1 Mutual information et non gaussianité

Définissons d'abord l'entropie, soit Y une variable de $\mathbb{R}^n$ de densité p :

$$H(y) = -\mathbb{E}(\log(p(Y)) \tag{26}$$

On définit la mutual information entre $y_1, \ldots, y_n$ par :

$$I(y_1, \ldots, y_n) = \sum_{i=1}^{n} H(y_i) - H(y). \tag{27}$$

On peut voir cette quantité comme une mesure de dépendance, elle ne considère pas que la corrélation/ covariance, elle est positive et est nulle si et seulement si les variables sont statistiquement indépendantes. Comme défini précedemment on note p_i la densité du signal s_i et p_x la densité jointe des signaux x_i. Ecrivons maintenant $y = f(x)$ vecteur de taille n, où f est une transformation invertible. On a alors =

$$p_y(y) = p_x(f^{-1}(y))|\det(J_{f(f^{-1}(y))}|^{-1} \tag{28}$$

et donc

$$\begin{aligned} H(y) &= \mathbb{E}(\log[p_x(f^{-1}(y))|\det J_{f(f^{-1}(y)}|^{-1}]) \\ &= \mathbb{E}(\log[p_x(x)|\det J_{f(x)}|^{-1}) \\ &= \mathbb{E}(\log(p_x(x))) - \mathbb{E}(\log(|\det(J_{f(x)}|) \end{aligned} \tag{29}$$

On obtient donc la relation suivante entre les deux entropies :

$$H(y) = H(x) + \mathbb{E}(\log(|\det(J_{f(x)}|) \tag{30}$$

En se ramenant maintenant à l'information mutuelle on obtient et en se ramenant au cas linéaire ou $y = Wx$ (on ignore l'indice t pour le moment) :

$$I(y) = \sum_{i=1}^{n} H(y_i) - H(x) - \log(|\det(W)|) \tag{31}$$

Mais on a supposé que les s_i étaient non correlés et de variance unitaire, il en va de même pour leur estimé y_i et donc $A = \det(W)\det(XX^\top)\det(W^\top)$

ainsi même sans le procédé d'orthonormalisation, $\det(W)$ est constant. On obtient :

$$\begin{aligned} I(y_1, \ldots, y_n) &= \sum_{i=1}^{n} H(y_i) - H(x) + cste \\ &= \sum_{i=1}^{n} -J(y_i) + H(y_{gauss}) - H(x) \\ &= cste' - \sum_{i=1}^{n} J(y_i), \end{aligned} \tag{32}$$

où $cste'$ ne dépend pas de W. ***Conclusion 1*** Estimation par analyse en composantes indépendantes (ICA) par minimisation d'information mutuelle équivaut à maximiser la somme des non-gaussianités des estimations des composantes indépendantes, lorsque les estimations sont contraintes d'être sans corrélation. Il existe cependant des différences importantes entre ces deux critères :

- L'estimation déflationniste, c'est-à-dire une par une, des composantes indépendantes est disponible en négentropie ($H(y_i)$) puisque l'on est capable de rechercher les maxima de non-gaussianité d'une seule projection $w_i^\top X$. Mais c'est impossible avec une information mutuelle.
- En utilisant la non-gaussianité (la négentropie ou le), nous forçons à décorréler les estimations des composantes indépendantes. Ce n'est pas nécessaire puisqu'on peut utiliser la forme $I(y) = \sum_{i=1}^{n} H(y_i) - H(x) - \log(|\det(W)|)$ directement.

4.2 Mutual information et maximum de vraissemblance

En reprenant la forme

$$I(y) = \sum_{i=1}^{n} H(y_i) - H(x) - \log(|\det(W)|), \tag{33}$$

Et en remarquant que si $w_i^\top x$ a pour loi p_i alors $\mathbb{E}(\log(p_i(w_i^\top X)) = -H(y_i)$, ainsi en notant $G_i(y_i) = \log(p_i(y_i))$, on peut écrire l'information mutuelle comme :

$$I(y_1, .., y_n) = -\sum_{i=1}^{n} \mathbb{E}(G_i(y_i)) - \log(|\det(W)|) - H(x) \tag{34}$$