Tema 3:

ANALISIS DE COMPONENTES INDEPENDIENTES (ICA)

Febrero-Mayo 2006

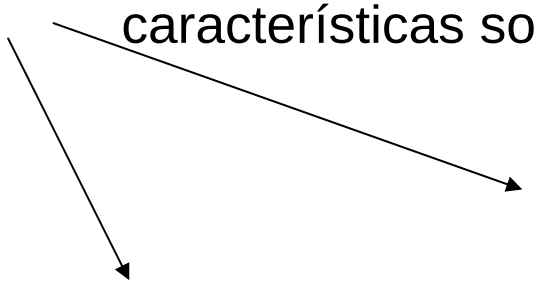
ÍNDICE

3.1 DEFINICIÓN DE ICA

PCA: Búsqueda de nuevas características en las que los vectores de observación quedan mejor representados en el sentido del error cuadrático medio (MMSE)

MDA: Búsqueda de nuevas características en las que las clases quedan más separadas

ICA: Búsqueda de vectores de proyección en los que las características son más independientes entre sí



SEPARACIÓN CIEGA DE FUENTES DE SEÑAL

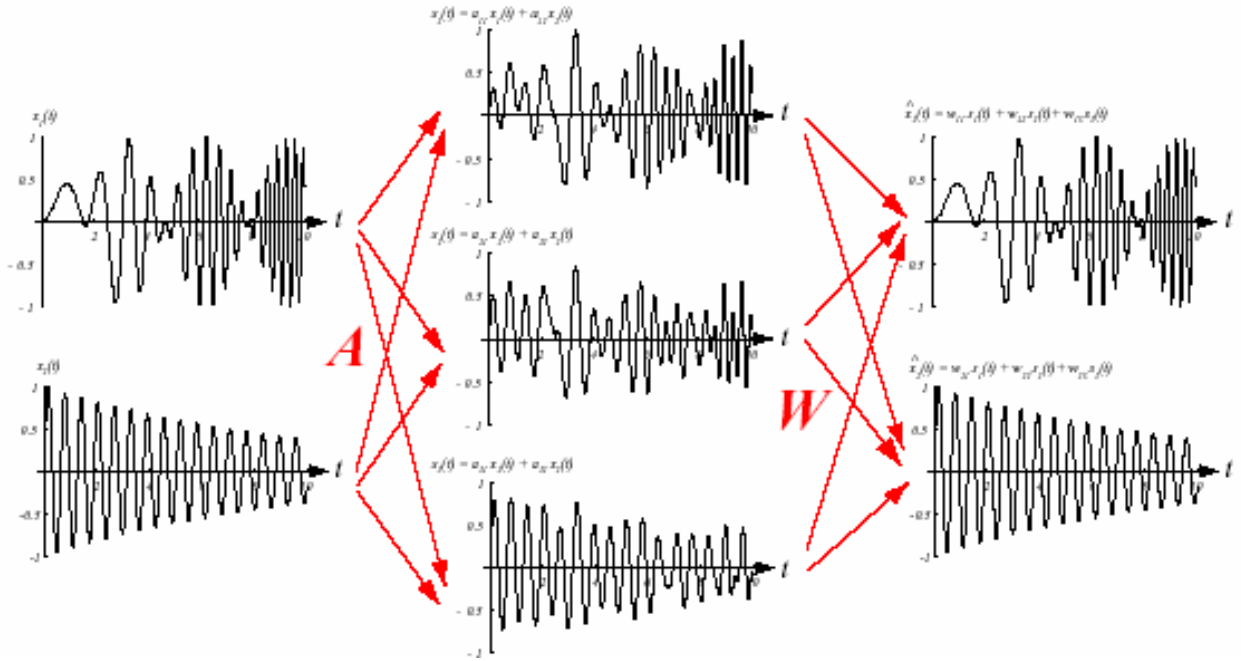
CLASIFICACIÓN

Observamos un conjunto de n señales, $x_1(t),\ldots,x_n(t)$ combinación lineal de otras n señales $s_1(t),\ldots,s_n(t)$ estadísticamente independientes entre si:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{n} \mathbf{a}_i s_i(t) \in \mathbb{R}^{n\times 1}$$

a partir de la observación de $\mathbf{x}(t)$, queremos recuperar $\mathbf{s}(t)$.

Cocktail party problem



RESTRICCIONES EN EL MODELO

1. La matriz de mezcla $\mathbf{A}$ no tiene memoria y es cuadrada
2. Las señales a recuperar $s_i(t)$ son independientes
3. Las señales $s_i(t)$ son no-gausianas
4. Supondremos que las señales son de media cero (siempre puede eliminarse la media y luego reconstruirse una vez separadas las componentes):

$$\mathbf{x}(t) = \mathbf{x}'(t) - E\left\{\mathbf{x}'(t)\right\} = \mathbf{A}\mathbf{s}(t) - \mathbf{A}E\left\{\mathbf{s}(t)\right\}$$

$$\hat{\mathbf{s}}(t) = \mathbf{A}^{-1}\mathbf{x}(t) + \mathbf{A}^{-1}E\left\{\mathbf{x}(t)\right\}$$

AMBIGÜEDADES EN LA SEPARACIÓN

1. Las componentes independientes $s_i(t)$ podrán recuperarse con la ambigüedad de un factor de escala

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) = \sum_{i=1}^{n} \left(\frac{1}{\alpha_i} \mathbf{a}_i \right) \left(\alpha_i s_i(t) \right)$$

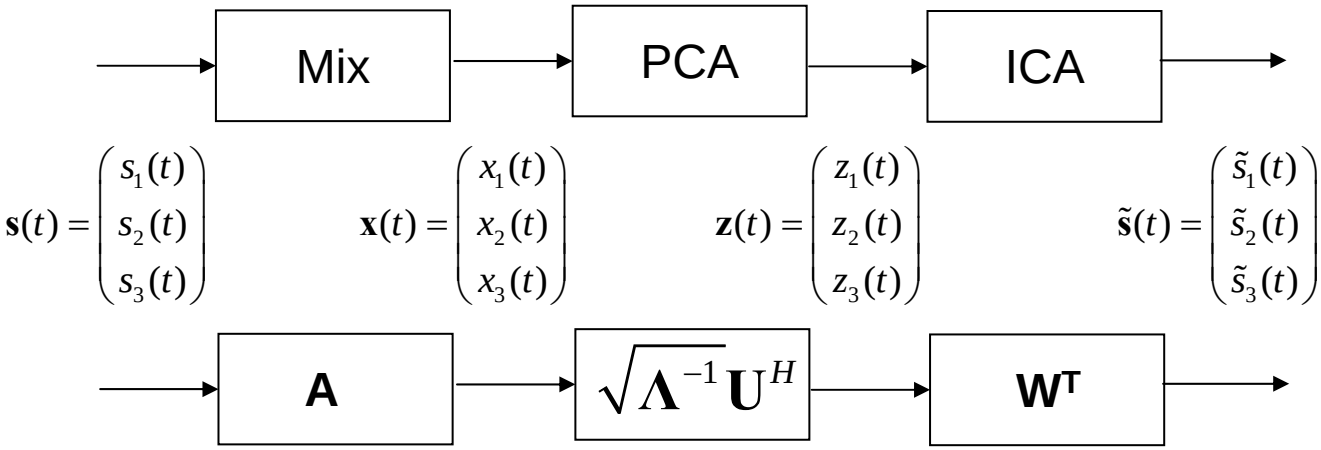
2. No podrá conocerse el orden en que se recuperen las componentes independientes (ambigüedad de una matriz de permutación):

$$\mathbf{x}(t) = \mathbf{A}\mathbf{P}\mathbf{P}^{-1}\mathbf{s}(t)$$

Ejemplo:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

ESQUEMA



$$\tilde{\mathbf{s}}(t) = \mathbf{W}^{\mathbf{T}} \sqrt{\Lambda^{-1}} \mathbf{U}^{\mathbf{H}} \mathbf{A}\mathbf{s}(t) = \mathbf{P}\mathbf{s}(t)$$

3.2 INDEPENDENCIA Y BLANQUEADO

El blanqueado de características siempre es posible, pero no garantiza la independencia.

1.- Señales generadas por las fuentes: Se suponen estadísticamente independientes y por tanto incorreladas:

$$\mathbf{s}(t) = \begin{pmatrix} s_1(t) \\ : \\ s_N(t) \end{pmatrix}; \qquad \mathbf{C}_\mathbf{s} = E\left\{\mathbf{s}(t)\mathbf{s}^T(t)\right\} = \mathbf{I}$$

2.- Proceso de Mezcla: Matriz **A** de rango =N

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t)$$

$$svd(\mathbf{A}) : \mathbf{A} = \mathbf{U}\sqrt{\mathbf{\Lambda}}\mathbf{V}^\mathbf{H}; \quad \mathbf{U}^\mathbf{H}\mathbf{U} = \mathbf{I}; \quad \mathbf{V}^\mathbf{H}\mathbf{V} = \mathbf{I}$$

$$\mathbf{C}_\mathbf{x} = E\left\{\mathbf{x}(t)\mathbf{x}^T(t)\right\} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathbf{H}$$

3.2 INDEPENDENCIA Y BLANQUEADO

3.- Blanqueado o Incorrelación de las señales:

$$\mathbf{z}(t) = \sqrt{\Lambda^{-1}}\mathbf{U}^{\mathbf{H}}\mathbf{x}(t)$$

$$\mathbf{C}_{\mathbf{z}} = E\left\{\mathbf{z}(t)\mathbf{z}^{T}(t)\right\} = \mathbf{I}$$

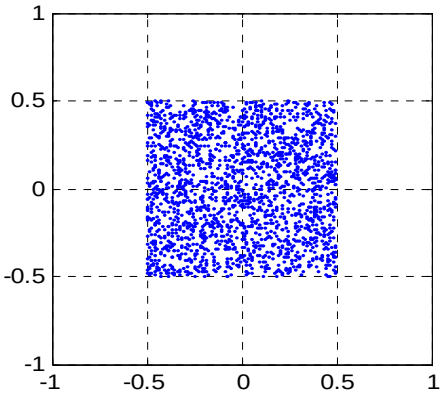
4.- ICA: Transformación Ortogonal. Búsqueda de N(N+1)/2 Incógnitas en lugar de N^2

$$\tilde{\mathbf{s}}(t) = \mathbf{W}^{\mathbf{T}}\mathbf{z}(t)$$

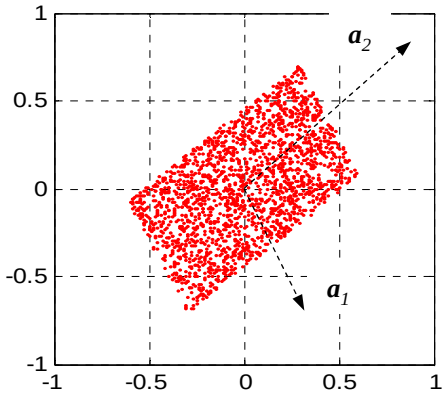
$$\mathbf{C}_{\tilde{\mathbf{s}}} = E\left\{\tilde{\mathbf{s}}(t)\tilde{\mathbf{s}}^{T}(t)\right\} = \mathbf{I} \Rightarrow \mathbf{W}^{\mathbf{T}}\mathbf{W} = \mathbf{I}$$

$$\mathbf{A} = \begin{bmatrix} 0.32 & 0.9 \\ -0.63 & 0.77 \end{bmatrix}$$

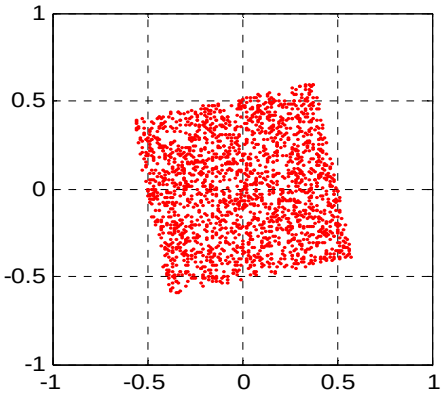
PDF uniforme



Vectores de dos características independientes

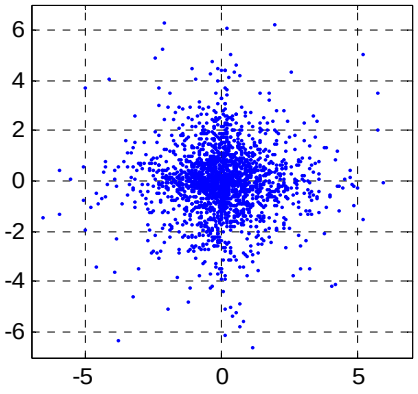


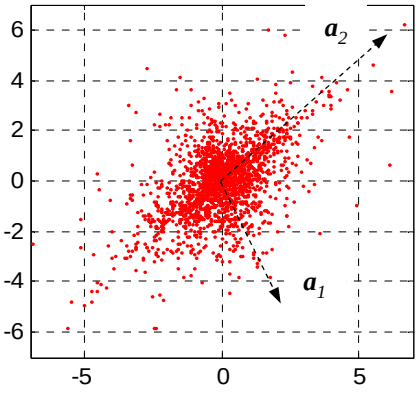
Características mezcladas

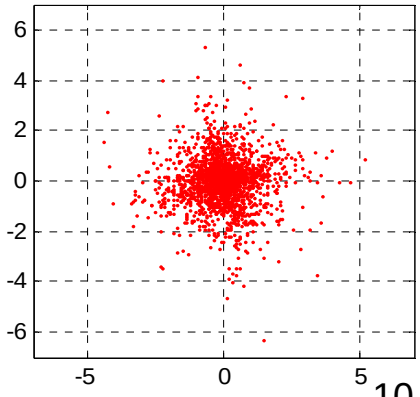


Características blanquedadas (no son aún independientes!)

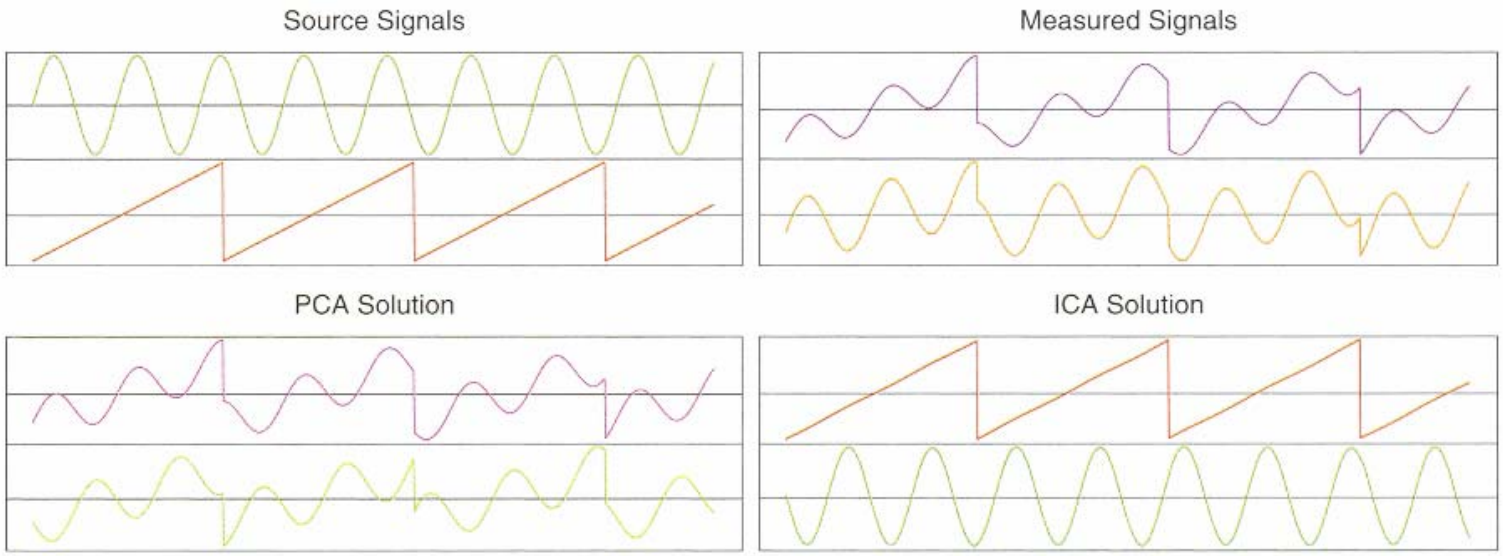
PDF laplaciana







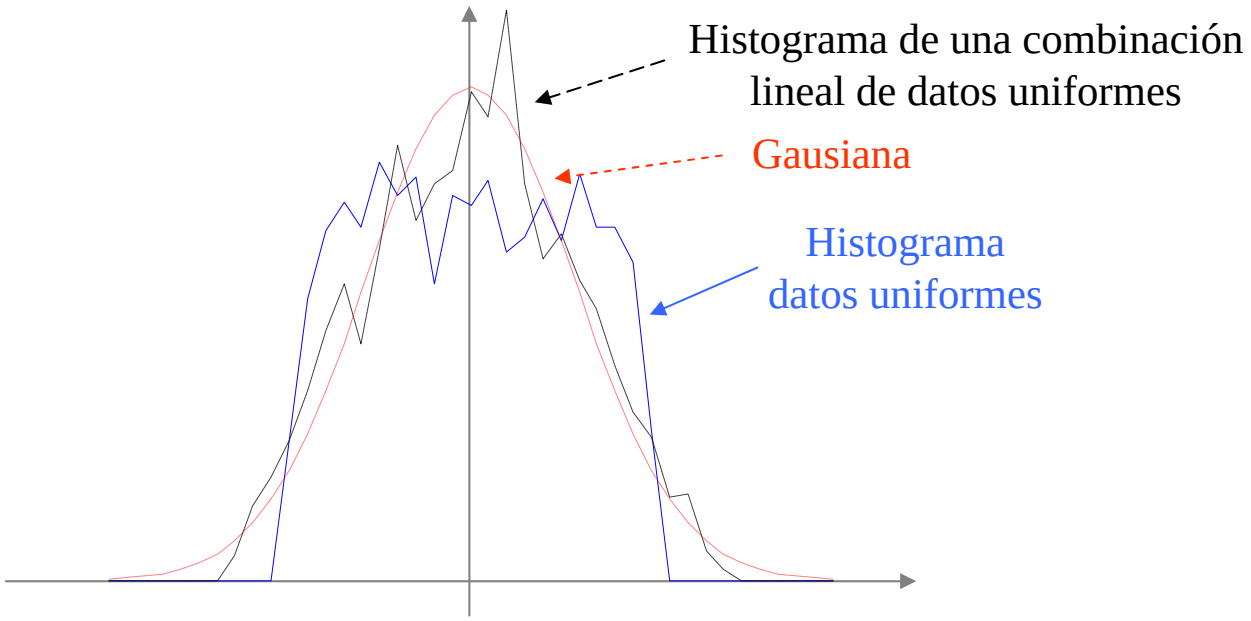
- Para procesos gausianos, el blanqueado implica independencia estadística por lo que no puede aplicarse ningun criterio mas estricto de separación
- En el caso no-gausiano, el blanqueado no es suficiente para separar los procesos



3.3 MAXIMIZACIÓN DE LA NO-GAUSIANIDAD

Si todas las componentes $s_i(t)$ independientes están igualmente distribuidas, su mezcla es más gausiana por el teorema central del límite.

Escogiendo un vector tal que $\mathbf{b}^T \mathbf{A} = \mathbf{e}_i^T$ podemos recuperar $s_i(t)$ sin necesidad de conocer $\mathbf{A}$, únicamente maximizando la no-gausianidad de $\mathbf{b}^T \mathbf{x}(t)$



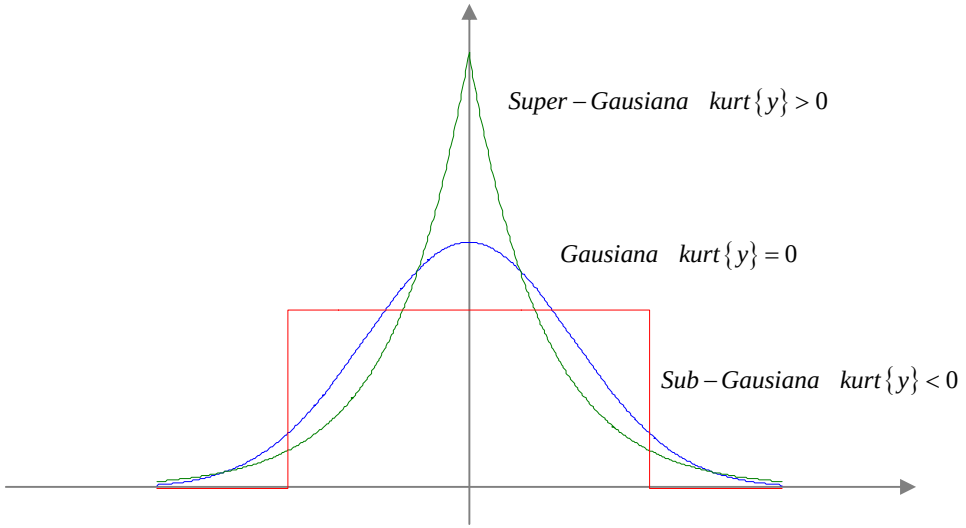
3.3.1 MEDIDA DE NO-GAUSIANIDAD: KURTOSIS

La kurtosis es el cumulante de orden 4, y para una variable aleatoria de media nula se define como:

$$\kappa_4\{y\} = E\{y^4\} - 3\left(E\{y^2\}\right)^2$$

Propiedades

1. Si $y \sim N\left(m, \sigma^2\right) \quad \Rightarrow \quad \kappa_4\{y\} = 0$
2. Puede tomar valores positivos y negativos



3. Para variables aleatorias y_1 , y_2 independientes

$$\kappa_4\{y_1 + y_2\} = \kappa_4\{y_1\} + \kappa_4\{y_2\}$$

4. Escalado

$$\kappa_4\{\alpha y\} = \alpha^4 \kappa_4\{y\}$$

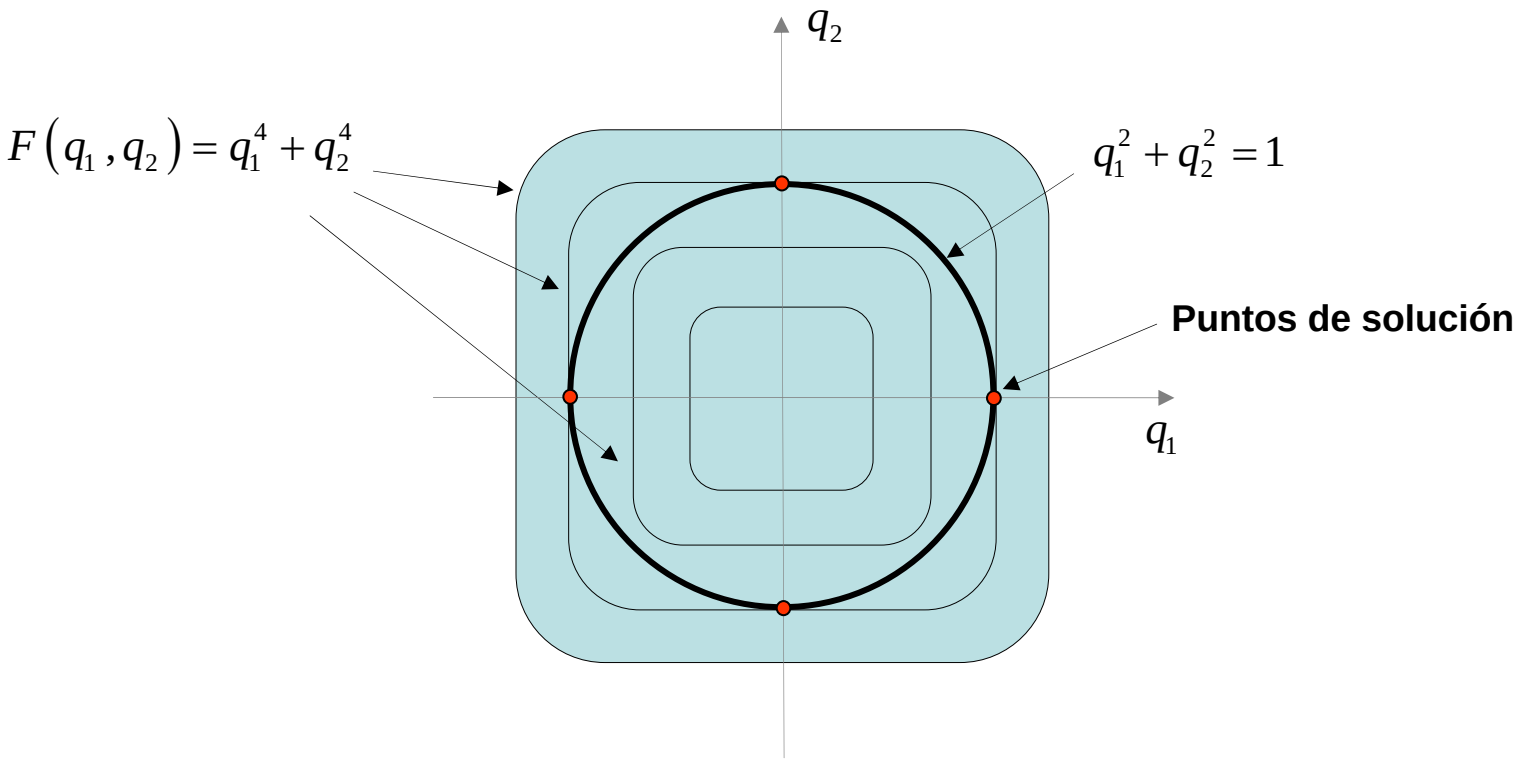
Criterio para separación de fuentes: optimización de la kurtosis con restricción sobre la potencia de la señal separada

$$\hat{s}_i(t) = y(t) = \mathbf{b}^T \mathbf{x}(t) = \mathbf{b}^T \mathbf{A}\mathbf{s}(t) = \mathbf{q}^T \mathbf{s}(t) = \sum_{j=1}^{n} q_j s_j(t)$$

$$\mathbf{b} = \arg\max_{\mathbf{b}} \left|\kappa_4\{y\}\right| = \arg\max_{\mathbf{b}} \left|\sum_{j=1}^{n} q_j^4 \kappa_4\{s_j(t)\}\right| \qquad (1)$$

subject to $\mathbf{q}^T\mathbf{q} = 1$

Ejemplo: si $\kappa_4\left\{s_i(t)\right\}=1 \;\; \forall i=1,2$



La restricción de norma=1 impuesta conduce a que al maximizar la kurtosis se separan las señales:

$$y(t)=q_1 s_1(t)+q_2 s_2(t)_{(q_1,q_2)=(1,0)}=s_1(t)$$

NECESIDAD DE BLANQUEADO

En la ecuación (1), nos aparecen dos variables **q** y **b** relacionadas a través de la matriz de mezcla **A** desconocida. El planteamiento del problema se hace abordable cuando blanqueamos el vector de observaciones **x**:

$$\hat{\mathbf{s}}(t) = \mathbf{W}^T \underbrace{\sqrt{\Lambda^{-1}}\mathbf{U}\mathbf{x}(t)}_{\substack{\mathbf{z}(t):\\ \text{vector}\\ \text{blanqueado}}} = \mathbf{W}^T \sqrt{\Lambda^{-1}}\mathbf{U}\mathbf{A}\mathbf{s}(t) = \mathbf{P}\mathbf{s}(t)$$

$$\mathbf{W} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & : & \mathbf{w}_n \end{bmatrix}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & : & \mathbf{q}_n \end{bmatrix}$$

$$\hat{s}_i(t) = \mathbf{w}_i^T \overbrace{\sqrt{\Lambda^{-1}}\mathbf{U}\mathbf{x}(t)} = \mathbf{w}_i^T \sqrt{\Lambda^{-1}}\mathbf{U}\mathbf{A}\mathbf{s}(t) = \mathbf{q}_i^T \mathbf{s}(t)$$

$$\mathbf{q}_i^T \mathbf{q}_j = \left(\mathbf{w}_i^T \sqrt{\Lambda^{-1}}\mathbf{U}\mathbf{A}\right)\left(\mathbf{A}^T\mathbf{U}^T \sqrt{\Lambda^{-1}}\mathbf{w}_j\right) = \mathbf{w}_i^T \mathbf{w}_j = \delta_{ij}$$

Únicamente aparece la variable matricial **w**

$$\mathbf{w} = \arg\max_{\mathbf{w}} \left|\kappa_4\left\{\mathbf{w}^T\mathbf{z}(t)\right\}\right|$$

$$\text{subject to} \quad \mathbf{w}^T\mathbf{w} = 1$$

ALGORITMO DE GRADIENTE

Si la kurtosis es positiva (negativa) la maximización (minimización) con restricciones puede hacerse mediante una técnica de gradiente:

$$\frac{\partial\left|\kappa_4\left\{\mathbf{w}^T\mathbf{z}(t)\right\}\right|}{\partial\mathbf{w}} = \frac{\partial}{\partial\mathbf{w}}\left|E\left\{\left(\mathbf{w}^T\mathbf{z}(t)\right)^4\right\} - 3E\left\{\left(\mathbf{w}^T\mathbf{z}(t)\right)^2\right\}^2\right| =$$

$$= 4sign\left(\kappa_4\left\{\mathbf{w}^T\mathbf{z}(t)\right\}\right)\left[E\left\{\mathbf{z}\left(\mathbf{w}^T\mathbf{z}(t)\right)^3\right\} - \underbrace{3\mathbf{w}\mathbf{w}^T\mathbf{w}}\right]$$

Este término puede eliminarse ya que sólo cambia la norma de **w**, no su dirección (el vector **w** ha de ser normalizado de todas formas)

Algoritmo de gradiente

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu\ sign\left(\kappa_4\left\{\mathbf{w}_n^T\mathbf{z}(t)\right\}\right)E\left\{\mathbf{z}\left(\mathbf{w}_n^T\mathbf{z}(t)\right)^3\right\}$$

$$\mathbf{w}_{n+1} \leftarrow \mathbf{w}_{n+1}\Big/\left(\mathbf{w}_{n+1}^T\mathbf{w}_{n+1}\right)$$

ASPECTOS PRÁCTICOS

1. Puede aproximarse la esperanza matemática por un valor instantáneo

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu \; sign\left(\kappa_4 \left\{\mathbf{w}_n^T \mathbf{z}(t)\right\}\right) \mathbf{z}\left(\mathbf{w}_n^T \mathbf{z}(t)\right)^3$$

2. Si el signo de la kurtosis de las fuentes independientes no es conocido debe estimarse, pero no puede aplicarse un estimador instantáneo. Es necesario promediar:

$$\gamma(t) = (1-\alpha)\,\gamma(t-1) + \alpha\left(\left(\mathbf{w}_n^T \mathbf{z}(t)\right)^4 - 3\right)$$

ALGORITMO DE PUNTO FIJO (Fast ICA)

El algoritmo de gradiente es de convergencia lenta: depende del paso de adaptación escogido y de la inicialización. Puede converger más rápidamente dándonos cuenta que el gradiente debe apuntar en la dirección del vector **w**. En efecto, en un problema genérico con restricciones esféricas:

$$L(\mathbf{w}) = F(\mathbf{w}) + \lambda(1 - \mathbf{w}^T\mathbf{w})$$

$$\frac{\partial L(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} - \lambda\mathbf{w} = \mathbf{0}$$

El gradiente es proporcional a la solución

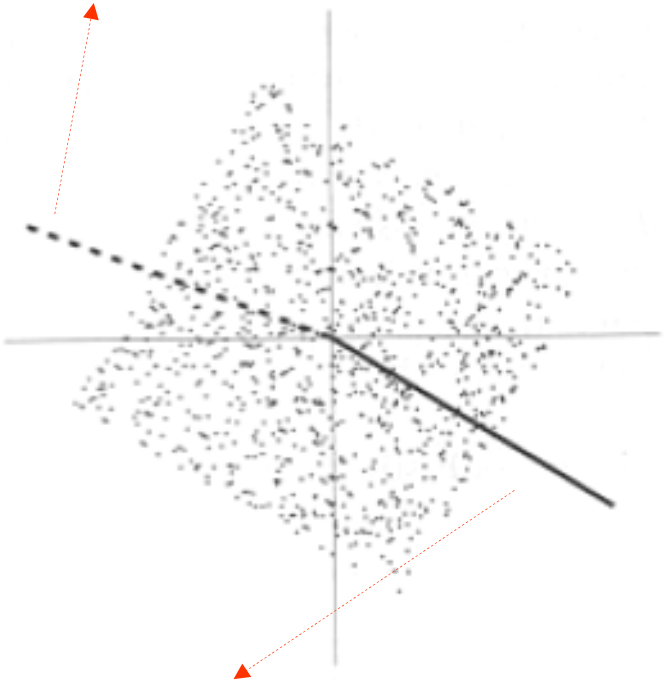
Así pues:

$$\mathbf{w} \leftarrow E\left\{\mathbf{z}\left(\mathbf{w}^T\mathbf{z}(t)\right)^3\right\} - 3\mathbf{w}\mathbf{w}^T\mathbf{w}$$

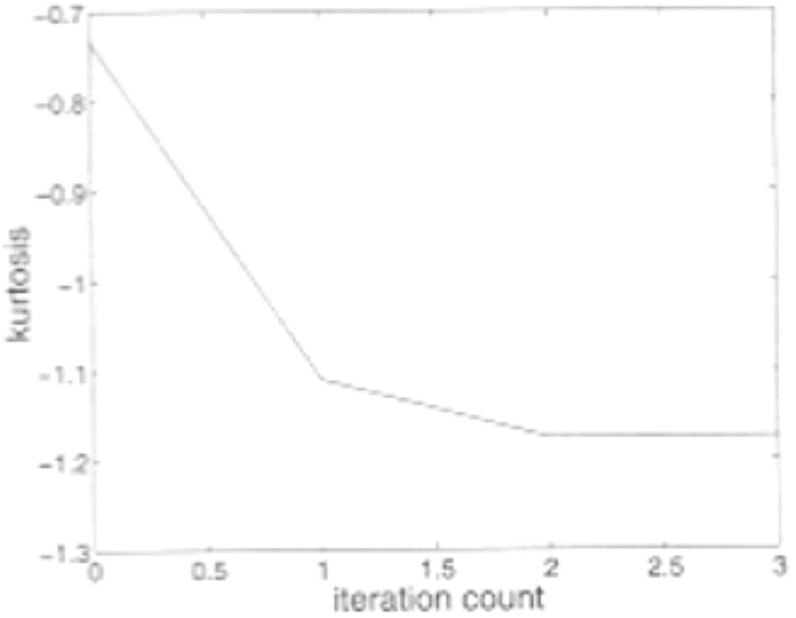
$$\mathbf{w}^T\mathbf{w} = 1$$

Solución no adaptativa, usar promedio temporal

Vector **w** despues de la primera iteración



Vector **w** despues de la segunda iteración



La convergencia es muy rápida

3.3.2 MEDIDA DE NO-GAUSIANIDAD: NEG-ENTROPÍA

La kurtosis es una medida de no-gausianidad poco fiable ya que los valores *outliers* tienen una gran influencia en la estimación. Alternativamente puede usarse la entropía para construir una medida de no-gausianidad.

La entropía es una medida de la información que aporta la observación de los valores de una variable aleatoria sobre la variable aleatoria.

Se define la entropía de una variable aleatoria y como:

$$H(y) = -\int p_y(\eta) \log p_y(\eta) d\eta$$

Fijada la potencia, la pdf gausiana es la que maximiza la entropía:

$$\min_{p_y} \int p_y(\eta) \log p_y(\eta) d\eta \qquad \text{con} \qquad \int \eta^2 p_y(\eta) d\eta = \sigma^2$$

La demostración se apoya en:

Teorema 3.1: Sean $\varphi(\eta)$ y $f(\eta)$ dos fdp. Entonces:

$$-\int \varphi(\eta)\log\varphi(\eta)d\eta \le -\int \varphi(\eta)\log f(\eta)d\eta$$

Demostración:

A partir de la desigualdad $\log(z) \le z-1$ aplicada sobre $z = \frac{f(\eta)}{\varphi(\eta)}$

$$\log\left(\frac{f(\eta)}{\varphi(\eta)}\right) \le \frac{f(\eta)}{\varphi(\eta)} - 1$$

Multiplicando por $\varphi(\eta)$ e integrando

$$\int \varphi(\eta)\log\left(\frac{f(\eta)}{\varphi(\eta)}\right)d\eta \le \int \left(f(\eta)-\varphi(\eta)\right)d\eta = 0$$

Teorema 3.2: El máximo de la entropía de una variable aleatoria de la que se conocen n momentos generalizados:

$$E\{g_i(x)\} = \int g_i(x) f(x) dx = \eta_i \quad i = 1, \ldots, n$$

es exponencial:

$$f(x) = A \exp\{-\lambda_1 g_1(x) - \lambda_2 g_2(x) - \ldots - \lambda_n g_n(x)\}$$

Demostración:

La entropía de $f(x)$ viene dada por

$$H_f = -\int f(x) \log f(x) dx =$$

$$= -\int f(x)[\log A - \lambda_1 g_1(x) - \lambda_2 g_2(x) - \ldots - \lambda_n g_n(x)] dx =$$

$$= \lambda_1 \eta_1 + \lambda_2 \eta_2 + \ldots + \lambda_n \eta_n - \log A$$

Cualquier otra función de densidad $\varphi(x)$ que satisficiera las restricciones daría lugar a una entropía menor:

Teorema 3.1

$$H_\phi = -\int \varphi(x)\log\varphi(x)dx \le -\int \varphi(x)\log f(x)dx =$$

$$= -\int \varphi(x)\left[\log A - \lambda_1 g_1(x) - \lambda_2 g_2(x) - \ldots - \lambda_n g_n(x)\right]dx =$$

$$= \lambda_1\eta_1 + \lambda_2\eta_2 + \ldots + \lambda_n\eta_n - \log A = H_{\max}(x)$$

La gausiana es la pdf que minimiza las suposiciones que se hacen sobre los datos, conocida la potencia.

La medida de no-gausianidad vendrá dada por la **neg-entropía**:

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

El cálculo exacto de la entropía es difícil a partir únicamente de los datos: la estimación de $p_y(y)$ es generalmente poco fiable. En su lugar, estimaremos la entropía a partir de un desarrollo en serie alrededor de una gausiana de media cero y varianza unidad:

$$\phi(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-y^2 / 2\right)$$

$$p_y(y) \approx \phi(y)\left(1+\kappa_3\{y\}\frac{h_3(y)}{3!}+\kappa_4\{y\}\frac{h_4(y)}{4!}\right)$$

$$\kappa_3\{y\} = E\{y^3\}$$

$$\kappa_4\{y\} = E\{y^4\} - 3$$

Polinomios de Hermite

Donde los polinomios de Hermite están definidos como:

$$\frac{\partial \phi^i(y)}{\partial y^i} = h_i(y)\phi(y)$$

y cumplen la propiedad de ortogonalidad según el producto escalar:

$$\int \phi(y) h_i(y) h_j(y) dy = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

Véase el anexo para más detalles

Substituyendo en la expresión de la entropía:

$$H(y) = -\int p_y(\eta)\log p_y(\eta)d\eta = \left\{\log(1+\varepsilon) \approx \varepsilon - \varepsilon^2/2\right\} \approx$$

$$\approx -\int \phi(\eta)\log\phi(\eta)d\eta - \frac{\kappa_3\{y\}^2}{2\times 3!} - \frac{\kappa_4\{y\}^2}{2\times 4!}$$

A tener en cuenta:

1. $H_3(x)$ y $H_4(x)$ son ortogonales a cualquier polinomio de segundo orden, y ortogonales entre si
2. $p_y(y)$ está cerca de ser Gausiana, por lo que los monomios de tercer grado del skewness o la kurtoskis son despreciables respecto a los monomios de segundo grado

Y la neg-entropía viene dada por:

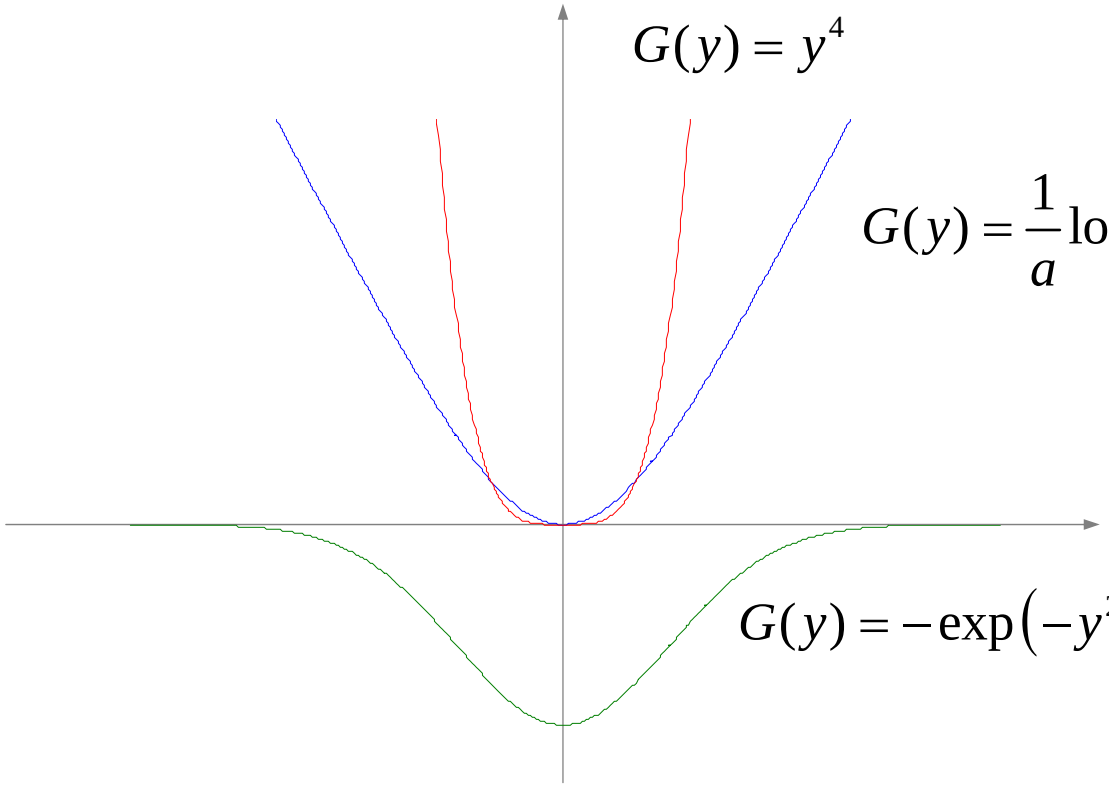
$$J(y) \approx \frac{\kappa_3(y)^2}{12} + \frac{\kappa_4(y)^2}{48} = \left\{\begin{array}{c}\text{si } p_y(y) \text{ es} \\ \text{simétrica}\end{array}\right\} = \frac{\kappa_4(y)^2}{48}$$

Es la expresión que hemos estado maximizando (si la pdf es simétrica, el primer término es cero)

Una alternativa a la kurtosis es el uso de otras funciones pares de variación más suave que la potencia cuarta:

$$J(y) \approx \left[E\{G(y)\} - E\{G(\upsilon)\} \right]^2$$

Variable aleatoria gaussiana



ALGORITMO DE GRADIENTE

Optimización

$$\mathbf{w} = \arg\max_{\mathbf{w}} J\left(\mathbf{w}^T \mathbf{z}(t)\right)$$

$$\text{subject to} \quad \mathbf{w}^T \mathbf{w} = 1$$

Algoritmo de gradiente

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu \nabla_{\mathbf{w}_n} J\left(\mathbf{w}^T \mathbf{z}\right)$$

$$\mathbf{w}_{n+1} \leftarrow \mathbf{w}_{n+1} \Big/ \left(\mathbf{w}_{n+1}^T \mathbf{w}_{n+1}\right)$$

$$g(x) = \frac{dG(x)}{dx}$$

Control de Signo:

$$\nabla_{\mathbf{w}_n} J(\mathbf{w}^T \mathbf{z}) = \gamma E\left\{\mathbf{z} g\left(\mathbf{w}^T \mathbf{z}\right)\right\}$$

$$\gamma = E\left\{G\left(\mathbf{w}^T \mathbf{z}\right)\right\} - E\left\{G(\upsilon)\right\}$$

ASPECTOS PRÁCTICOS

1. Puede aproximarse la esperanza matemática por un valor instantáneo

$$\mathbf{w}_{n+1} = \mathbf{w}_n + \mu\ \gamma \mathbf{z} g\left(\mathbf{w}^T \mathbf{z}(t)\right)$$

2. El término γ juega le mismo papel que el signo de la kurtosis. Es necesario promediar:

$$\gamma(t) = (1-\alpha)\gamma(t-1) + \alpha\left(G\left(\mathbf{w}_n^T \mathbf{z}(t)\right) - E\{G(\upsilon)\}\right)$$

3. Si se conoce a priori el signo de γ (p.e., si todas las señales son supergausianas, como en el caso de señales de voz) puede fijarse

4. También es posible definir un algoritmo de punto fijo o Fast ICA

FAST ICA
ALGORITMO PARA ACELERAR LA CONVERGENCIA

1. Método de Newton: Series de Taylor Multivariables

$$J(\mathbf{w}[n+1]) = J(\mathbf{w}[n]) + \left(\frac{\partial J(\mathbf{w}[n])}{\partial \mathbf{w}[n]}\right)^{\mathbf{T}} (\mathbf{w}[n+1]-\mathbf{w}[n]) + \tfrac{1}{2}(\mathbf{w}[n+1]-\mathbf{w}[n])^{\mathbf{T}} \left(\frac{\partial^2 J(\mathbf{w}[n])}{\partial \mathbf{w}[n]^2}\right)(\mathbf{w}[n+1]-\mathbf{w}[n]) \Rightarrow$$

$$J(\mathbf{w}[n+1]) - J(\mathbf{w}[n]) = \Delta\mathbf{w}^{\mathbf{T}} \nabla J(\mathbf{w}[n]) + \tfrac{1}{2}\Delta\mathbf{w}^{\mathbf{T}} \left(\frac{\partial^2 J(\mathbf{w}[n])}{\partial \mathbf{w}[n]^2}\right)\Delta\mathbf{w}$$

2. Si la matriz Hessiana es definida positiva, la función anterior es de forma parabólica y presenta un mínimo en:

$$\nabla J(\mathbf{w}[n]) + \left(\frac{\partial^2 J(\mathbf{w}[n])}{\partial \mathbf{w}[n]^2}\right)\Delta\mathbf{w} = 0 \Rightarrow \Delta\mathbf{w} = -\left(\frac{\partial^2 J(\mathbf{w}[n])}{\partial \mathbf{w}[n]^2}\right)^{-1} \nabla J(\mathbf{w}[n])$$

3. Ecuación de adaptación de pesos.

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \left(\frac{\partial^2 J(\mathbf{w}[n])}{\partial \mathbf{w}[n]^2}\right)^{-1} \nabla J(\mathbf{w}[n])$$

4. Aproximación de la función kurtosis

$$J(\mathbf{w}) = \left(E\left[G\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\right] - \left[G(\nu)\right]\right)^2 \qquad \nu : \text{v.a. gaussiana de referencia}$$

5. Función a Minimizar

$$E\left[G\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\right] + \lambda\left(\mathbf{w}^{\mathbf{T}}\mathbf{w} - 1\right)$$

6. Obtención del Gradiente

$$\nabla J(\mathbf{w}) = E\left[g\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\mathbf{z}(t)\right] + \lambda\mathbf{w}$$

7. Matriz Hessiana :

$$\left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}^2}\right) = E\left[g'\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\mathbf{z}(t)\mathbf{z}(t)^{\mathbf{T}}\right] + \lambda\mathbf{I} \simeq E\left[g'\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\right]E\left[\mathbf{z}(t)\mathbf{z}(t)^{\mathbf{T}}\right] + \lambda\mathbf{I} = \left(E\left[g'\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\right] + \lambda\right)\mathbf{I} \Rightarrow$$

$$\left(\frac{\partial^2 J(\mathbf{w})}{\partial \mathbf{w}^2}\right)^{-1} \simeq \left(E\left[g'\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}(t)\right)\right] + \lambda\right)^{-1}\mathbf{I}$$

8. Sustituyendo los resultados de 6 y 7 en 3:

$$\mathbf{w}[n+1] = \mathbf{w}[n] - \left(\frac{\partial^2 J(\mathbf{w}[n])}{\partial \mathbf{w}[n]^2} \right)^{-1} \nabla J(\mathbf{w}[n]) = \mathbf{w}[n] - \frac{1}{E\left[g'\left(\mathbf{w}[n]^{\mathrm{T}} \mathbf{z}(t)\right)\right] + \lambda} \left(E\left[g\left(\mathbf{w}[n]^{\mathrm{T}} \mathbf{z}(t)\right)\mathbf{z}(t)\right] + \lambda \mathbf{w}[n] \right) \Rightarrow$$

$$\mathbf{w}[n+1] = E\left[g\left(\mathbf{w}[n]^{\mathrm{T}} \mathbf{z}(t)\right)\mathbf{z}(t)\right] - \left[g'\left(\mathbf{w}[n]^{\mathrm{T}} \mathbf{z}(t)\right)\right]\mathbf{w}[n]$$

Algoritmo Iterativo Fast ICA

1. Inicialización Aleatoria: $\mathbf{w}_i[0]$
2. Adaptación (i=1,..N) $\mathbf{w}_i[n+1] = E\left[g\left(\mathbf{w}_i[n]^{\mathrm{T}} \mathbf{z}(t)\right)\mathbf{z}(t)\right] - \left[g'\left(\mathbf{w}_i[n]^{\mathrm{T}} \mathbf{z}(t)\right)\right]\mathbf{w}_i[n]$
3. Normalización $\mathbf{w}_i[n+1] = \mathbf{w}_i[n+1] / \left\| \mathbf{w}_i[n+1] \right\|$
4. Ortogonalización $\mathbf{w}_i[n+1] = \left(\mathbf{I} - \mathbf{P}_{[\mathbf{w}_1, \mathbf{w}_2, .., \mathbf{w}_{i-1}]} \right) \mathbf{w}_i[n+1]$

   (Proyección en subespacio ortogonal a los anteriores o método similar)
5. Si no converge *go to* 2.

EXTRACCIÓN DE TODAS LAS COMPONENTES

Con las técnicas vistas hasta el momento es posible extraer la componente independiente i a partir del vector $\mathbf{w}_i$. Para extraer todas las demás es necesario garantizar que todos los vectores $\mathbf{w}_i$ serán ortogonales entre si. Pueden aplicarse dos métodos:

1. Ortogonalización deflacionaria: extraídas *d* componentes, el vector $\mathbf{w}_{d+1}$ se restringe a ocupar el espacio ortogonal a $\mathbf{w}_1,\ldots,\mathbf{w}_d$

Los errores cometidos en las primeras componentes extraídas se van acumulando

2. Ortogonalización simétrica: las *n* componentes se extraen simultáneamente: los vectores $\mathbf{w}_1,\ldots,\mathbf{w}_d$ se calculan en paralelo y en cada iteración se ortogonalizan con métodos simétricos:

Método 1

$$\mathbf{W} \leftarrow (\mathbf{W}\mathbf{W}^T)^{-1/2}\,\mathbf{W}$$

Conclusiones/Objetivos

- ICA aplicado a separación de fuentes: Esquema
- Blanqueado fuerza que la matriz de ICA: W sea ortogonal
- Maximizar Kurtosis con restricciones de norma=1 separa señales independientes (Figura).
- Minimizar gaussianidad al maximizar kurtosis (o minimizar k<0) o una aproximación de la Kurtosis a través de una función no lineal G
- Diferencia entre algoritmo de gradiente lento (LMS) y rápido (FastICA)
- Posibilidades de ortogonalización conjunta de todos los vectores.

Método 2 $\mathbf{W}(1) = \mathbf{W}(0) / \left\| \mathbf{W}(0) \right\|$

$$\mathbf{W}(t+1) = \frac{3}{2}\mathbf{W}(t) - \frac{1}{2}\mathbf{W}(t)\mathbf{W}(t)^T \mathbf{W}(t)$$

Se itera hasta que $\mathbf{W}(t)^T \mathbf{W}(t) \approx \mathbf{I}$

REFERENCIAS

Aapo Hyvärinen, Juha Karhunen, Erkki Oja,
“Independent Component Analysis”,
Ed. Wiley Interscience, 2001

Cocktail party demo:
http://www.cis.hut.fi/projects/ica/cocktail/cocktail_en.cgi

ANNEX: Gram-Charlier expansion of a pdf and approximate entropy

The Gram-Charlier expansion is a series that approximate any density function of finite-valued cumulants in the vicinity of a Gaussian density. Let x be a random variable whose cumulants are known. Cumulants are defined as the coefficients of the Taylor series expansion of the second characteristic function:

$$\Phi_x(z) = \ln \Psi_x(z) = -\kappa_1 z + \kappa_2 \frac{z^2}{2!} + \sum_{k=3}^{\infty} \kappa_k \frac{(-z)^k}{k!} \qquad (1.1$$

where κ_i is the i-th order cumulant of *x*. In particular $\kappa_1 = E\{x\} = \bar{x}, \quad \kappa_2 = E\left\{\left|x - E\{x\}\right|^2\right\} = \sigma^2$. Cumulants have two interesting property that can be used in the derivation of the capacity for the interference channel: the cumulant (of any order) of a sum of independent random variables is the sum of cumulants; and the cross cumulant of an ensemble of random variables is zero if one of the random variables is independent of the rest.

$\Psi_x(z)$ is the characteristic function, the Laplace transform of the probability density function:

$$f_x(x) = \int_{c-j\infty}^{c+j\infty} e^{zx} \Psi_x(z) \frac{dz}{2\pi j} \qquad (1.2)$$

a contour integral over the regularity domain $c_1 \le c \le c_2$.
Equation (1.1) can be rewritten as:

$$\Psi_x(z) = \exp\left(-\bar{x}z + \sigma^2 \frac{z^2}{2!}\right) \exp\left(\sum_{k=3}^{\infty} \kappa_k \frac{(-z)^k}{k!}\right) = \exp\left(-\bar{x}z + \sigma^2 \frac{z^2}{2!}\right)\left(1 + \sum_{k=3}^{\infty} c_k (-z)^k\right) \qquad (1.3)$$

where second exponential has been developed in Taylor series in the last equality. The first coefficients are given by:

$$c_3 = \frac{\kappa_3}{3!} \quad c_4 = \frac{\kappa_4}{3!} \quad c_5 = \frac{\kappa_5}{3!} \quad c_6 = \frac{\kappa_6 + 10\kappa_3^2}{6!}$$

By replacing equation (1.3) in (1.2):

$$f_x(x) = \int_{c-j\infty}^{c+j\infty} \exp\left((x-\bar{x})z + \sigma^2 \frac{z^2}{2!}\right)\left(1 + \sum_{k=3}^{\infty} c_k (-z)^k\right) \frac{dz}{2\pi j}$$

Let us evaluate this integral term by term, but first we need to determine the cumulant generating function for the Gaussian case:

$$G(x) = \int_{c-j\infty}^{c+j\infty} \exp\left((x-\bar{x})z + \sigma^2 \frac{z^2}{2!}\right) \frac{dz}{2\pi j} = \frac{1}{\sigma} \phi^{(0)}\left(\frac{x-\bar{x}}{\sigma}\right) \qquad \text{where} \qquad \phi^{(0)}(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Note that a random variable is Gaussian if and only if all cumulants of order greater than 2 are zero. By differenciating *k* times with respect to $\bar{x}$ we obtain:

$$\int_{c-j\infty}^{c+j\infty} (-z)^k \exp\left((x-\bar{x})z + \sigma^2 \frac{z^2}{2!}\right) \frac{dz}{2\pi j} = \frac{1}{\sigma^{k+1}} \phi^{(k)}\left(\frac{x-\bar{x}}{\sigma}\right)$$

$$\phi^{(k)}(x) = (-1)^k \frac{d^k}{dx^k} \phi^{(0)}(x)$$

Therefore we can conclude that

$$f_x(x) = \frac{1}{\sigma} \phi^{(0)}\left(\frac{x-\bar{x}}{\sigma}\right) + \frac{1}{\sigma} \sum_{k=3}^{\infty} \frac{c_k}{\sigma^k} \phi^{(k)}\left(\frac{x-\bar{x}}{\sigma}\right)$$

which is the Gram-Charlier series expansion of the probability density function. The functions $\phi^{(k)}(x)$ can be related to the Hermite polynomials[1] as:

$$\phi^{(k)}(x) = \phi^{(0)}(x) h_k(x)$$

and hence:

$$f_x(x) = \frac{1}{\sigma} \phi^{(0)}\left(\frac{x-\bar{x}}{\sigma}\right)\left[1 + \sum_{k=3}^{\infty} \frac{c_k}{\sigma^k} h_k\left(\frac{x-\bar{x}}{\sigma}\right)\right]$$

This expression can be used to evaluate the entropy of *x*. In particular, truncating the series to the first two terms:

$$\begin{aligned} H(x) &= -\int f_x(\eta) \log f_x(\eta) d\eta = \left\{\log(1+\varepsilon) \approx \varepsilon - \varepsilon^2/2\right\} \approx \\ &\approx -\int \frac{1}{\sigma} \phi^{(0)}\left(\frac{\eta-\bar{x}}{\sigma}\right) \log\left(\frac{1}{\sigma} \phi^{(0)}\left(\frac{\eta-\bar{x}}{\sigma}\right)\right) d\eta - \frac{{\kappa_3}^2}{2\times 3!\times \sigma^6} - \frac{{\kappa_4}^2}{2\times 4!\times \sigma^8} = \\ &= H_G(x) - \frac{{\kappa_3}^2}{2\times 3!\times \sigma^6} - \frac{{\kappa_4}^2}{2\times 4!\times \sigma^8} \end{aligned}$$

where the last equality is obtained by recognising that:

1. $h_3(x)$ and $h_4(x)$ are orthogonal to any 2nd order polynomial
2. The pdf to be approximated will be close to a Gaussian, and hence any third order monomial of κ_3 and κ_4 will be much smaller than second order monomial.

[1] The Hermite polynomials are orthogonal in the scalar product: $\int \phi^{(0)}(y) h_i(y) h_j(y) dy = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. The first Hermite polynomials are: $h_0(y) = 1 \quad h_1(y) = -y \quad h_2(y) = 1 + y^2 \quad h_3(y) = -3y - y^3 \quad h_4(y) = 3 + 6y^2 + y^4$