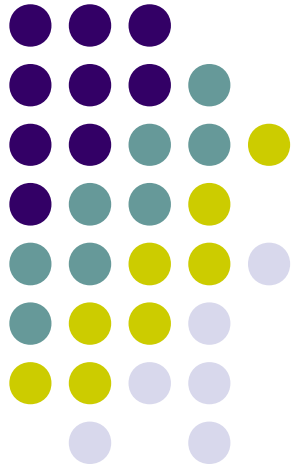
Separación de Audio usando *ICA*

Introducción al Reconocimiento de Patrones . iie . fing

Segundo semestre 2008 / feb.2009

Haldo Spontón



Cocktail-Party Problem

Imaginemos estar en un cuarto donde dos personas hablan simultáneamente. Se tienen dos micrófonos en distintos lugares de la habitación. Se grabarían dos señales $x_1(t)$ y $x_2(t)$, una por micrófono, que son sumas ponderadas de cada señal emitida por las personas, llamadas $s_1(t)$ y $s_2(t)$. Se tiene:

$$\begin{cases} x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \\ x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \end{cases}$$

Cocktail-Party Problem (2)

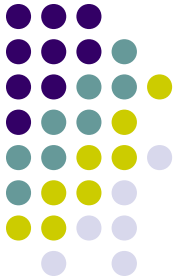
Sería útil poder estimar $s_1(t)$ y $s_2(t)$ a partir de las señales grabadas $x_1(t)$ y $x_2(t)$. Esto es llamado *Cocktail-Party Problem*.

En general: $x = \left[x_1(t), x_2(t), \ldots, x_n(t)\right]^T$

$$s = \left[s_1(t), s_2(t), \ldots, s_m(t)\right]^T \Rightarrow \quad x = As$$

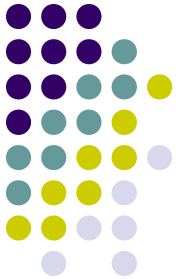
Recuperar las componentes independientes desde la observación de una combinación de las mismas es denominado *Blind Signal Separation*.

Definición de *ICA*



- En el modelo *ICA*, asumimos que cada mezcla x_j y cada componente independiente s_k es una variable aleatoria.
- Si esto se cumple, la ecuación $x=As$ es llamada *modelo ICA*. Dicha ecuación muestra como son generados los datos observados mediante la mezcla de las fuentes independientes.
- Las fuentes no pueden ser directamente observadas.
- También se asume desconocida la matriz A.

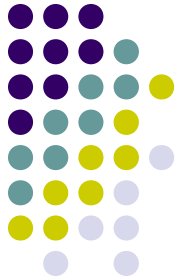
Definición de *ICA* (2)



- El punto de partida de *ICA* es asumir que los componentes s_i son estadísticamente independientes.
- Se busca estimar la inversa de la matriz de mezcla *A*, llamémosle *W*, de manera de obtener las fuentes independientes en función de los datos observados:

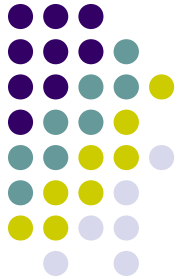
$$s = Wx$$

Principios de la estimación *ICA*



- Intuitivamente $\rightarrow$ No-gaussianidad.
- El Teorema Central del Límite dice que la distribución de una suma de variables aleatorias independientes tiende a una gaussiana. Por lo tanto, la suma de dos variables aleatorias independientes es “más gaussiana” que cada una de las variables aleatorias originales.
- Para estimar uno de los componentes independientes, consideremos una combinación lineal de los datos observados: $y = w^T x = \sum_i w_i x_i$

Principios de la estimación *ICA* (2)

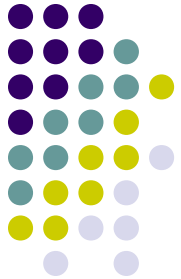


- Si *w* fuese una de las filas de la inversa de *A*, la combinación lineal sería igual a uno de los componentes independientes.
- La cuestión está en como usar el TCL para determinar *w* de manera que se aproxime a una de las filas de la inversa de *A*.
- Cambio de variable: $z = A^T w \Rightarrow y = w^T x = w^T As = z^T s$
- Con esto, *y* es combinación lineal de los s_i ponderados por z_i.

Principios de la estimación *ICA* (3)

- Dado que la suma de dos variables aleatorias independientes es más gaussiana que las originales, $z^T s$ es más gaussiana que cualquiera de las s_i, y se vuelve menos gaussiana cuando es, de hecho, igual a alguna s_i.
- En este caso, obviamente solo uno de los elementos de z es distinto de cero.
- Tomamos vector w que maximice la no-gaussianidad de $w^T x$.

Medida de la no-gaussianidad



- Debemos tener una medida cuantitativa de la no-gaussianidad de una variable aleatoria (llámese *y*).
- Se asume que *y* está centrada (tiene media nula) y varianza unitaria (se verá más adelante un procedimiento para siempre lograr esta simplificación).
- Varias medidas de no-gaussianidad.

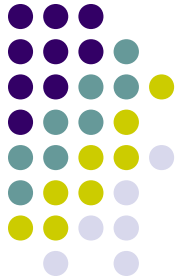
Kurtosis

- Medida clásica de no-gaussianidad.

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

- Para una *y* gaussiana el *kurtosis* se anula.
- Típicamente la no-gaussianidad se mide con el valor absoluto del *kurtosis*, el cual es nulo para una variable gaussiana, y mayor que cero cuanto “menos gaussiana” sea la variable.

Kurtosis (2)



- Ventajas: simplicidad computacional y teórica. El análisis teórico se simplifica dado que el *kurtosis* es lineal.

- Desventajas: muy sensible a *outliers*, lo que hace que el *kurtosis* no sea una medida robusta de la no-gaussianidad.

Negentropía

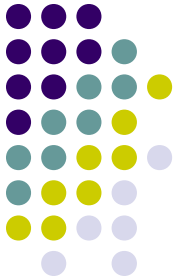
- Basada en el concepto clásico de la teoría de la información; la entropía.
- La entropía de una variable aleatoria puede ser interpretada como el grado de información que aporta la observación de una variable.
- Cuanto más aleatoria, impredecible e inestructurada sea la variable, mayor será su entropía.

Negentropía (2)

- Un resultado fundamental en teoría de la información es que una variable gaussiana posee la mayor entropía entre todas las variables aleatorias de igual varianza.
- Esto significa que la entropía puede ser usada como medida de no-gaussianidad.
- Para obtener una medida de no-gaussianidad, que se anule para una variable gaussiana, y sea siempre no negativa, se usa una versión modificada de la definición de entropía diferencial, llamada negentropía:

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

Negentropía (3)



- Ventajas: fuerte justificación teórica.

- Desventajas: alto costo computacional, dado que requiere estimar la pdf, probablemente no paramétrica.

- Entonces, aproximaciones de la negentropía serán muy útiles.

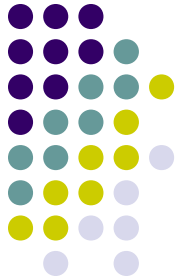
Aproximaciones de la negentropía

- El clásico método para aproximar la negentropía es usando momentos de alto orden:

$$J(y) \approx \frac{1}{12}\mathrm{E}\{y^3\}^2 + \frac{1}{48}kurt(y)^2$$

- Sin embargo, la validez de esta aproximación puede ser bastante limitada. En particular, esta aproximación sufre los mismos problemas de robustez que el *kurtosis*.

Aproximaciones de la negentropía (2)



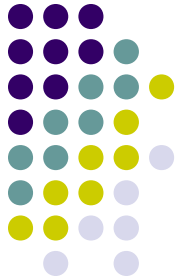
- Para evitar estos problemas, se desarrollaron nuevas aproximaciones basadas en el principio de máxima entropía:

$$J(y) \approx \sum_{i=1}^{p} k_i \left[\mathrm{E}\{G_i(y)\} - \mathrm{E}\{G_i(v)\} \right]^2$$

- En el caso que usemos solo una función no-cuadrática *G* , la aproximación es:

$$J(y) \propto \left[\mathrm{E}\{G(y)\} - \mathrm{E}\{G(v)\} \right]^2$$

Aproximaciones de la negentropía (3)



- El punto esta en elegir adecuadamente *G*, para obtener mejores aproximaciones de la negentropía. En particular, eligiendo *G* que no crezca muy rápido, uno obtiene estimadores más robustos. Las siguientes funciones son útiles:

$$G_1(u) = \frac{1}{a_1}\log\cosh a_1 u \qquad G_2(u) = -e^{-\frac{u^2}{2}}$$

donde $1 \le a_1 \le 2$ es una constante adecuada.

Pre-procesamiento para *ICA*

- Antes de aplicar *ICA* sobre los datos, es muy útil realizar cierto pre-procesado. En esta sección veremos ciertas técnicas de procesamiento para hacer el problema de estimación *ICA* más simple y mejor condicionado.

Centrado

- El más básico y necesario pre-procesamiento es centrar *x* restándole el vector medio *E{x}*, haciendo el vector *x* de media nula.
- Esto implica que también *s* es de media nula.
- Este pre-procesamiento es únicamente para simplificar los algoritmos *ICA*, no significa que las medias no puedan estimarse.

Whitening (Blanqueado)

- Implica transformar linealmente el vector *x* para obtener un nuevo vector *blanco* (con componentes no correlacionadas y varianzas unitarias).
- En otras palabras, la matriz de covarianza del nuevo vector es igual a la identidad.
- La transformación de blanqueado es siempre posible. Un método popular para hacerlo es usar la descomposición en valores propios (EVD) de la matriz de covarianza de *x*.

Whitening (Blanqueado) (2)

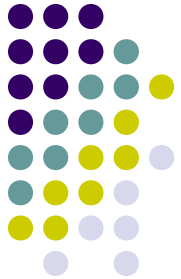
$$\mathrm{E}\left\{xx^T\right\} = EDE^T$$

- *E* es la matriz ortogonal de vectores propios de la matriz de covarianza de *x* y *D* es la matriz diagonal con sus vectores propios.
- El blanqueado puede ahora realizarse:

$$\tilde{x} = ED^{-1/2}E^T x$$

donde: $D^{-1/2} = diag\left(d_1^{-1/2}, \cdots, d_n^{-1/2}\right)$

Whitening (Blanqueado) (3)



- El blanqueado transforma la matriz de mezcla en una nueva $\tilde{A}$.

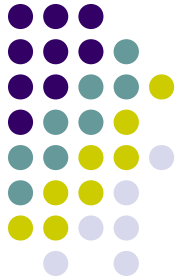
$$\tilde{x} = ED^{-1/2}E^T As = \tilde{A}s$$

- La utilidad del blanqueado reside en que la nueva matriz de mezcla es ortogonal:

$$\mathrm{E}\left\{\tilde{x}\tilde{x}^T\right\} = \tilde{A}\mathrm{E}\left\{ss^T\right\}\tilde{A}^T = \tilde{A}\tilde{A}^T = \mathrm{I}$$

- Vemos que el blanqueado reduce el número de parámetros a ser estimados.

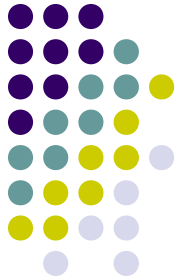
El algoritmo *FastICA*



- Hasta ahora introdujimos diferentes medidas de no-gaussianidad, que son funciones objetivo para la estimación *ICA*.
- En la práctica también se necesita un algoritmo para maximizar la función de contraste adecuada para esta tarea.

- Se asume de aquí en adelante que los datos están centrados y blanqueados.

FastICA para una unidad



- Por una “unidad” nos referimos a una unidad computacional, eventualmente una neurona artificial con un vector de pesos *w*, el cual dicha neurona es capaz de entrenar.
- La regla de aprendizaje de *FastICA* encuentra una dirección, es decir un vector *w* tal que la proyección $w^T x$ maximiza la no-gaussianidad.
- Aquí la no-gaussianidad es medida por la aproximación de la negentropía vista anteriormente.

FastICA para una unidad (2)

- Forma básica del algoritmo *FastICA*:

1. Elegir un vector inicial *w*, por ejemplo aleatorio.
2. Se calcula: $w^+ = \mathrm{E}\{xg(w^T x)\} - \mathrm{E}\{g'(w^T x)\}w$
3. Se actualiza *w*: $w = \frac{w^+}{\|w^+\|}$
4. Si no converge, volver a 2.

FastICA para varias unidades

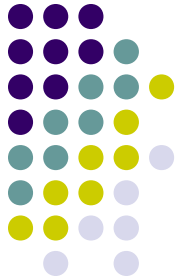
- El algoritmo para una unidad antes descrito estima solo uno de los componentes independientes.
- Para estimar varias componentes independientes, necesitamos correr el algoritmo *FastICA* de una unidad pero usando varias unidades (neuronas) con vectores de pesos $w_1,\ldots,w_n$.
- *Para evitar que diferentes vectores converjan al mismo máximo debemos decorrelacionar las salidas* $w_1^T x,\ldots,w_n^T x$ luego de cada iteración.

FastICA para varias unidades (2)

- Para ello estimamos los componentes independientes uno a uno.

- Cuando tenemos estimados *p* componentes independientes, corremos el algoritmo de una unidad para w_{p+1}, y luego de cada paso iterativo le restamos al mismo las "proyecciones" $w^T_{p+1} w w_j$, $j=1 \ldots p$.

FastICA para varias unidades (3)



- Luego de cada iteración:

1. Calculamos: $$w_{p+1} = w_{p+1} - \sum_{j=1}^{p} w_{p+1}^T w_j w_j$$
2. Normalizamos: $$w_{p+1} = \frac{w_{p+1}}{\sqrt{w_{p+1}^T w_{p+1}}}$$

Propiedades del algoritmo *FastICA*

- Bajo la asunción de validez del modelo *ICA*, la convergencia es cúbica (o al menos cuadrática).
- No hay ningún parámetro a seleccionar.
- Encuentra directamente componentes independientes de casi cualquier distribución no-gaussiana.
- Las componentes independientes pueden ser estimadas una a una.

Escuchemos los resultados…