# Análisis de Componentes Independientes

## Lección 23

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

# Contenido

# Introducción

- El Análisis de Componentes Independientes (ACI, o ICA) es similar a PCA en cuanto a que busca una nueva base para los datos.
- Ambos métodos son no supervisados.
- La meta final es muy diferente.
- El PCA busca datos **decorrelados**, el ICA busca datos **independientes**.

# Independencia estadística

- Sean $s_1$ y $s_2$ dos variables aleatorias
- Hay independencia estadística si probabilidad conjunta es igual al producto de marginales:

$$p(s_1,s_2) = p(s_1)p(s_2)$$

o también si probabilidad de un marginal no depende del otro

$$p(s_1|s_2) = p(s_1)$$

# El problema de la fiesta de cocktail (1)

- En el problema de la fiesta de cocktail $n$ fuentes acústicas independientes se capturan con $n$ micrófonos distribuidos en un cuarto.
- Cada micrófono captura combinación lineal de $n$ fuentes.
- Por estar los micrófonos en posiciones espaciales distintas, la combinación de fuentes en cada micrófono es distinta.
- El problema es entonces encontrar la señal de las $n$ fuentes a partir de las señales capturadas con los micrófonos.
- Este es un caso particular de la **separación ciega de fuentes** (*BSS, blind source separation*).
- Aquí asumimos que las grabaciones tienen retardo cero.

# El problema de la fiesta de cocktail (2)

- Sea $\underline{\mathbf{s}} \in \mathbb{R}^n$ la señal generada por las $n$ fuentes. La señal observada es entonces

$$\underline{\mathbf{x}} = \mathbf{A}\underline{\mathbf{s}}$$

con $\mathbf{A} \in \mathbb{R}^{n \times n}$ la **matriz de mezcla**.

- Tenemos $m$ observaciones $\underline{\mathbf{x}}^{(i)}$ y queremos recuperar las fuentes $\underline{\mathbf{s}}^{(i)}$ que dieron origen a los datos.
- En otros términos, queremos encontrar $\mathbf{W} = \mathbf{A}^{-1}$ (matriz de segregación o *unmixing*) tal que

$$\underline{\mathbf{s}}^{(i)} = \mathbf{W}\underline{\mathbf{x}}^{(i)}$$

a pesar de que desconocemos $\mathbf{A}$.

- Denotaremos con $\underline{\mathbf{w}}_j^T$ la $j$-ésima fila de $\mathbf{W}$, encargada de recuperar la $j$-esima fuente $s_j^{(i)} = \underline{\mathbf{w}}_j^T \underline{\mathbf{x}}^{(i)}$

# Ambigüedad de permutación

- Puesto que no tenemos información a priori de las fuentes ni de la matriz de mezcla, el ICA encontrará alguna permutación de las fuentes.
- Sea $\mathbf{P}$ una matriz $n \times n$ de permutación.
- Esto implica que cada fila y columna de $\mathbf{P}$ tienen un único 1:

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

- Dados solo $\underline{\mathbf{x}}^{(i)}$ es imposible distinguir $\mathbf{W}$ de $\mathbf{PW}$

# Ambigüedad de escalamiento

- Dados $\underline{\mathbf{x}}^{(i)} = \mathbf{A}\underline{\mathbf{s}}^{(i)}$, nótese que observaríamos los mismos valores con la matriz de mezcla $\alpha\mathbf{A}$ si las fuentes son ahora $\underline{\mathbf{s}}^{(i)}/\alpha$.
- Más aún, si solo escalamos una columna $\underline{\mathbf{a}}_{:,j}$ de $\mathbf{A}$, basta con escalar la fuente $j$ con el recíproco del factor de escala para obtener lo mismo.
- Nótese que estos cambios de escala solo implicarían un cambio de "volumen" acústico en las señales, y por tanto no son importantes en la aplicación.

# ¿Otras fuentes de ambigüedad?

- Resulta que solo la permutación y el escalamiento son fuentes de ambigüedad en ICA, **siempre y cuando** la fuentes sean **no gaussianas**.
- Con fuentes gaussianas, el problema no es solucionable:
  - Si $\underline{\mathbf{s}} \sim \mathcal{N}(0,\mathbf{I})$ sus curvas equiprobables son círculos/esferoides centrados en cero (simetría rotacional)
  - Se cumple $\text{Cov}(\underline{\mathbf{s}}) = \text{E}[\underline{\mathbf{s}}\underline{\mathbf{s}}^T] = \mathbf{I}$.
  - Si mezclamos $\underline{\mathbf{x}} = \mathbf{A}\underline{\mathbf{s}}$ entonces $\underline{\mathbf{x}} \sim \mathcal{N}(0,\mathbf{\Sigma})$ con $\mathbf{\Sigma} = \text{E}[\underline{\mathbf{x}}\underline{\mathbf{x}}^T] = \text{E}[\mathbf{A}\underline{\mathbf{s}}\underline{\mathbf{s}}^T\mathbf{A}^T] = \mathbf{A}\mathbf{A}^T$
  - Sea $\mathbf{R}$ una matriz de rotación, tal que $\mathbf{R}\mathbf{R}^T = \mathbf{R}^T\mathbf{R} = \mathbf{I}$.
  - Si mezclamos con $\mathbf{A}' = \mathbf{A}\mathbf{R}$ entonces tendremos observaciones $\underline{\mathbf{x}}' \sim \mathcal{N}(0,\mathbf{\Sigma}')$
  - $\mathbf{\Sigma}' = \text{E}[\underline{\mathbf{x}}'(\underline{\mathbf{x}}')^T] = \text{E}[\mathbf{A}'\underline{\mathbf{s}}\underline{\mathbf{s}}^T\mathbf{A}'^T] = \text{E}[\mathbf{A}\mathbf{R}\underline{\mathbf{s}}\underline{\mathbf{s}}^T(\mathbf{A}\mathbf{R})^T] = \mathbf{A}\mathbf{R}\mathbf{R}^T\mathbf{A}^T = \mathbf{A}\mathbf{A}^T = \mathbf{\Sigma}$
  - $\Rightarrow$ Es imposible discernir si $\mathbf{A}$ o $\mathbf{A}'$ mezcló los datos.

# Efecto de transformaciones lineales en densidades

## Caso de escalares

- Sea $s \in \mathbb{R}$ una variable aleatoria tomada de acuerdo a una densidad $p_s(s)$.
- Sea $x \in \mathbb{R}$ otra variable $x = As$ con $A \in \mathbb{R}$.
- ¿Cuál es la densidad $p_x(x)$?
- Podríamos pensar que si calculamos $s = A^{-1}x$ y evaluamos $p_s(s)$ obtenemos $p_x$, pero ¡**no es cierto**!
- Por ejemplo, si $s \sim$ Uniforme[0,1] entonces $p_s(s) = 1\{0 \leq s \leq 1\}$.
- Si elegimos $A = 2$ entonces $x = 2s$ y $x$ se distribuye uniformemente en [0,2] con densidad $p_x(x) = \frac{1}{2}1\{0 \leq x \leq 2\}$, que **no** es igual a $p_s(A^{-1}x) = 1\{0 \leq x \leq 2\}$.
- La fórmula correcta es: $p_x(x) = |A^{-1}|p_s(A^{-1}x)$

# Efecto de transformaciones lineales en densidades

## Caso vectorial

- Si $\underline{\mathbf{s}}$ tiene una distribución vectorial con densidad $p_{\underline{\mathbf{s}}}(\underline{\mathbf{s}})$ y $\underline{\mathbf{x}} = \mathbf{A}\underline{\mathbf{s}}$ con una matriz cuadrada, no singular $\mathbf{A}$, entonces la densidad de $\underline{\mathbf{x}}$ es

$$p_{\underline{\mathbf{x}}}(\underline{\mathbf{x}}) = \frac{1}{|\mathbf{A}|} p_{\underline{\mathbf{s}}}(\mathbf{A}^{-1}\underline{\mathbf{x}})$$

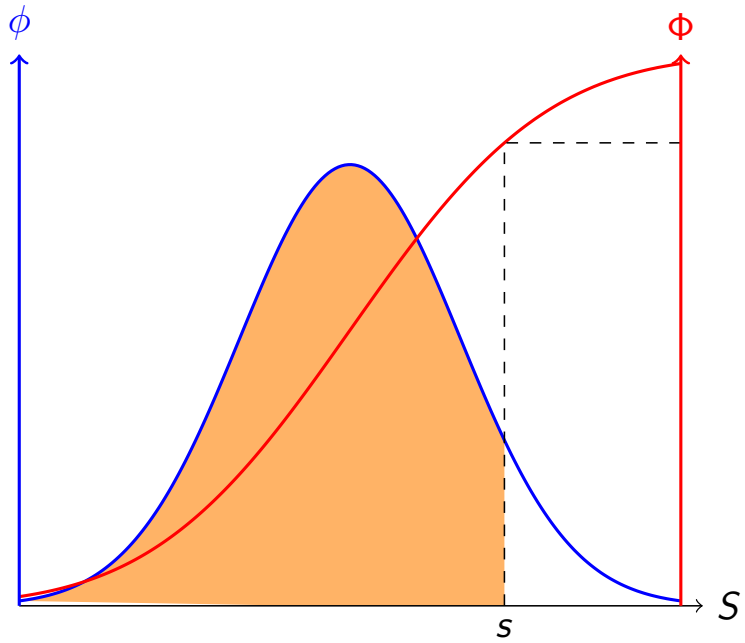
- Nótese que $1/|\mathbf{A}|$ normaliza el (hiper)volumen de la distribución.

# Distribución acumulativa

- Sea $s$ una variable aleatoria con densidad $p_s(s)$
- La función de distribución acumulada (CDF) es

$$F(s) = P(S \leq s) = \int_{-\infty}^{s} p_s(t)\, dt \quad \Rightarrow \quad p_s(s) = \frac{dF(s)}{ds}$$

- Por ejemplo:



# Especificación de variable aleatoria

- Si queremos especificar el comportamiento de una variable aleatoria $s$:
  1. podemos especificar $p_s(s)$ directamente
  2. podemos especificar $F(s)$, pues $p_s(s) = F'(s)$

# Derivación de ICA

- Supondremos que la distribución de cada fuente $s_i$ está dada por una densidad $p_s$ y que la distribución conjunta de las fuentes $\underline{\mathbf{s}} = [s_1, s_2, \ldots, s_n]^T$ es

$$p(\underline{\mathbf{s}}) = \prod_{i=1}^{n} p_s(s_i)$$

  donde hemos asumido para el cálculo que todas las fuentes son independientes.
- Esto implica que con $\underline{\mathbf{x}} = \mathbf{A}\underline{\mathbf{s}} = \mathbf{W}^{-1}\underline{\mathbf{s}}$:

$$p(\underline{\mathbf{x}}) = \prod_{i=1}^{n} p_s(\underline{\mathbf{w}}_i^T \underline{\mathbf{x}})|\mathbf{W}|$$

- Solo queda especificar qué es la densidad de las fuentes individuales $p_s$.

# Definición de la densidad de fuente

- Definiremos las densidades a través de una CDF que **no** puede corresponder a una gaussiana.
- Si CDF se puede determinar experimentalmente o es conocida, pues se utiliza
- En caso contrario, usualmente se selecciona una función que pase lentamente de 0 a 1, tal como

$$g(s) = \frac{1}{1 + e^{-s}}$$

de modo que $p_s(s) = g'(s) = g(s)(1 - g(s))$.

- La PDF asociada a $g(s)$ como CDF decae con $e^{-s}$, más lento que la gaussiana que decae con $e^{-s^2}$
- Buscamos entonces $\mathbf{W}$.

# Máxima verosimilitud

- Dado el conjunto de entrenamiento $\underline{\mathbf{x}}^{(i)}; i = 1 \ldots m$

$$\ell(\mathbf{W}) = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} \ln g' \left( \underline{\mathbf{w}}_j^T \underline{\mathbf{x}}^{(i)} \right) + \ln |\mathbf{W}| \right)$$

- Sabiendo que $\nabla_{\mathbf{W}} |\mathbf{W}| = |\mathbf{W}| (\mathbf{W}^{-1})^T$, entonces calculando el gradiente e igualándolo a cero se obtiene una regla de actualización de ascenso estocástico:

$$\mathbf{W} \leftarrow \mathbf{W} + \alpha \left( \begin{bmatrix} 1 - 2g(\underline{\mathbf{w}}_1^T \underline{\mathbf{x}}^{(i)}) \\ 1 - 2g(\underline{\mathbf{w}}_2^T \underline{\mathbf{x}}^{(i)}) \\ \vdots \\ 1 - 2g(\underline{\mathbf{w}}_n^T \underline{\mathbf{x}}^{(i)}) \end{bmatrix} \underline{\mathbf{x}}^{(i)\,T} + (\mathbf{W}^T)^{-1} \right)$$

con $\alpha$ la tasa de aprendizaje

# Curtosis

- Lo anterior no es la única técnica de solución
- Una técnica fácil de utilizar desde GNU/Octave usa una estimación de **curtosis**
- La curtosis es una medida de qué tan "puntiaguda" es una distribución probabilística.
- Hay varias definiciones. Por ejemplo (*excess curtosis*):

$$\text{Curt}[X] = \text{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

- La curtosis de una distribución gaussiana es 0 (**mesocúrtica**).
- Distribuciones más puntiagudas (como la laplaciana) tienen curtosis mayores a cero (**leptocúrtica**).
- Distribuciones más planas (como la uniforme) tiene curtosis menores a cero (**platicúrtica**).

# ICA en dos líneas (1)

- El siguiente ejemplo de Roweis, Weiss y Simoncelli muestra un concepto:

```
%W=kica(X);
Y=sqrtm(inv(cov(X')))*(X-repmat(mean(X,2),1,size(X,2)));
[W,ss,vv]=svd((repmat(sum(Y.*Y,1),size(Y,1),1).*Y)*Y');
```

- `X` tiene las mediciones mezcladas (una por columna)
- `W` es la matriz de segregación.
- La primera línea hace el enblanquecimiento de los datos, que logra que la matriz de segregación sea ortogonal.
- La segunda línea produce una matriz tal que, si se multiplica por un vector unitario, produce una aproximación de la curtosis en esa dirección.

# ICA en dos líneas (2)

- Los eigenvectores más grandes de esa matriz son las direcciones (ortogonales) de curtosis máxima para fuentes super-gaussianas (más picudas) y son buenas estimaciones de las direcciones de segregación.
- Más detalles en el trabajo de Miettinen, Taskinen, Nordhausen y Oja

# Aplicaciones

- Rescatando señales de EEG
- Análisis financieros
- Análisis de imágenes
- Análisis de datos astronómicos

# Ejemplo

- Ver aquí
- Ver demo audio

# Resumen

*Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux*