

# Clasificación de patrones

- Introducción
- Clasificadores paramétricos
- Clasificadores semi-paramétricos
- Reducción de dimensionalidad
- Prueba de significancia

# Clasificadores semiparamétricos

- Mezcla de densidades
- Estimación del parámetro de máxima probabilidad (ML)
- Mezcla de implementaciones
- Maximización de la expectativa (EM)

# Mezcla de densidades

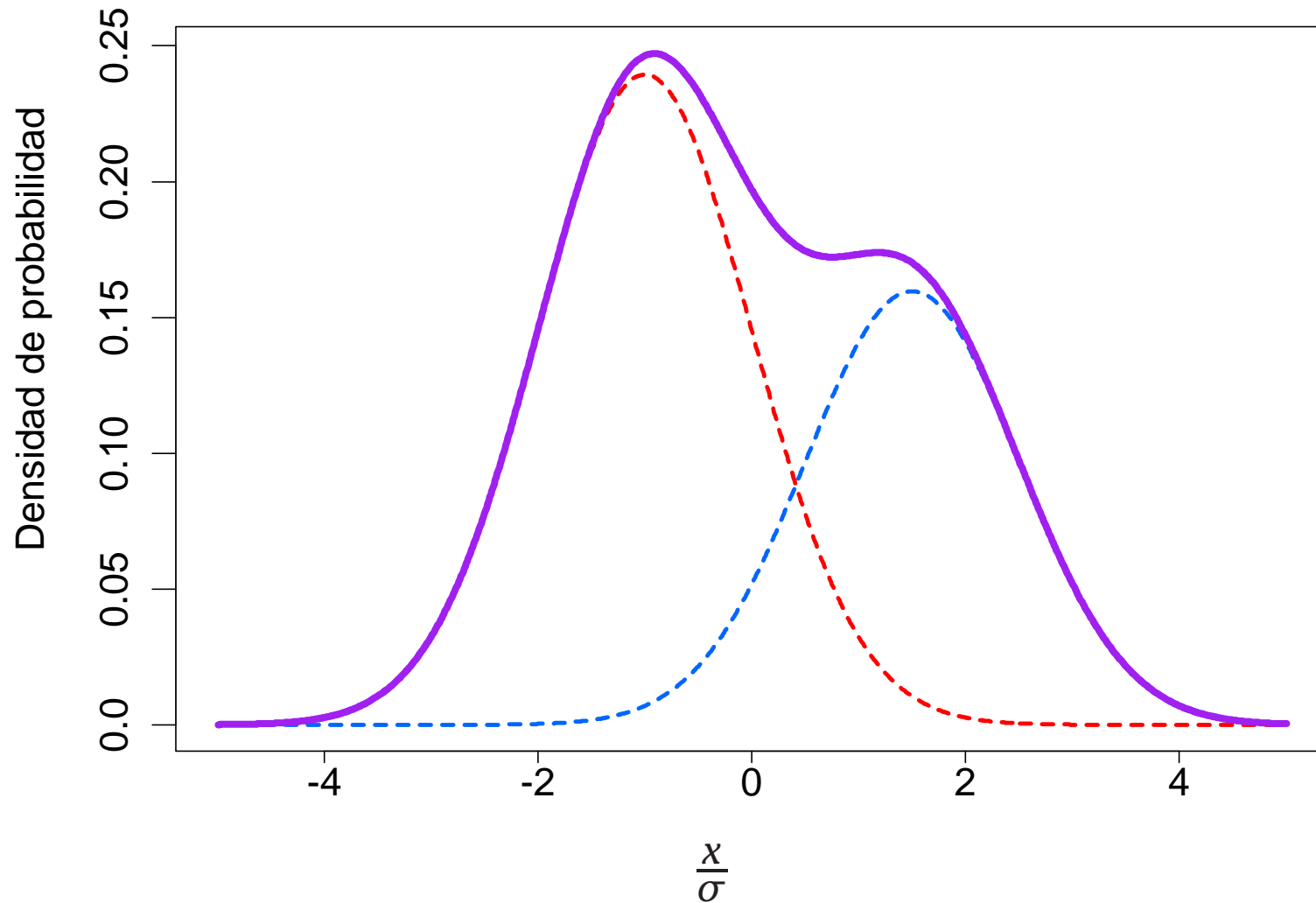
- Un PDF está compuesto por una mezcla de densidades del componente  $m$   $\{\omega_1, \dots, \omega_m\}$ :

$$p(\mathbf{x}) = \sum_{j=1}^m p(\mathbf{x}|\omega_j)P(\omega_j)$$

- Los parámetros PDF del componente y los pesos de mezcla  $P(\omega_j)$  no se conocen normalmente, lo cual convierte a la estimación del parámetro en una forma **de aprendizaje no supervisado**.
- Las mezclas de gaussianas suponen componentes normales:

$$p(\mathbf{x}|\omega_k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

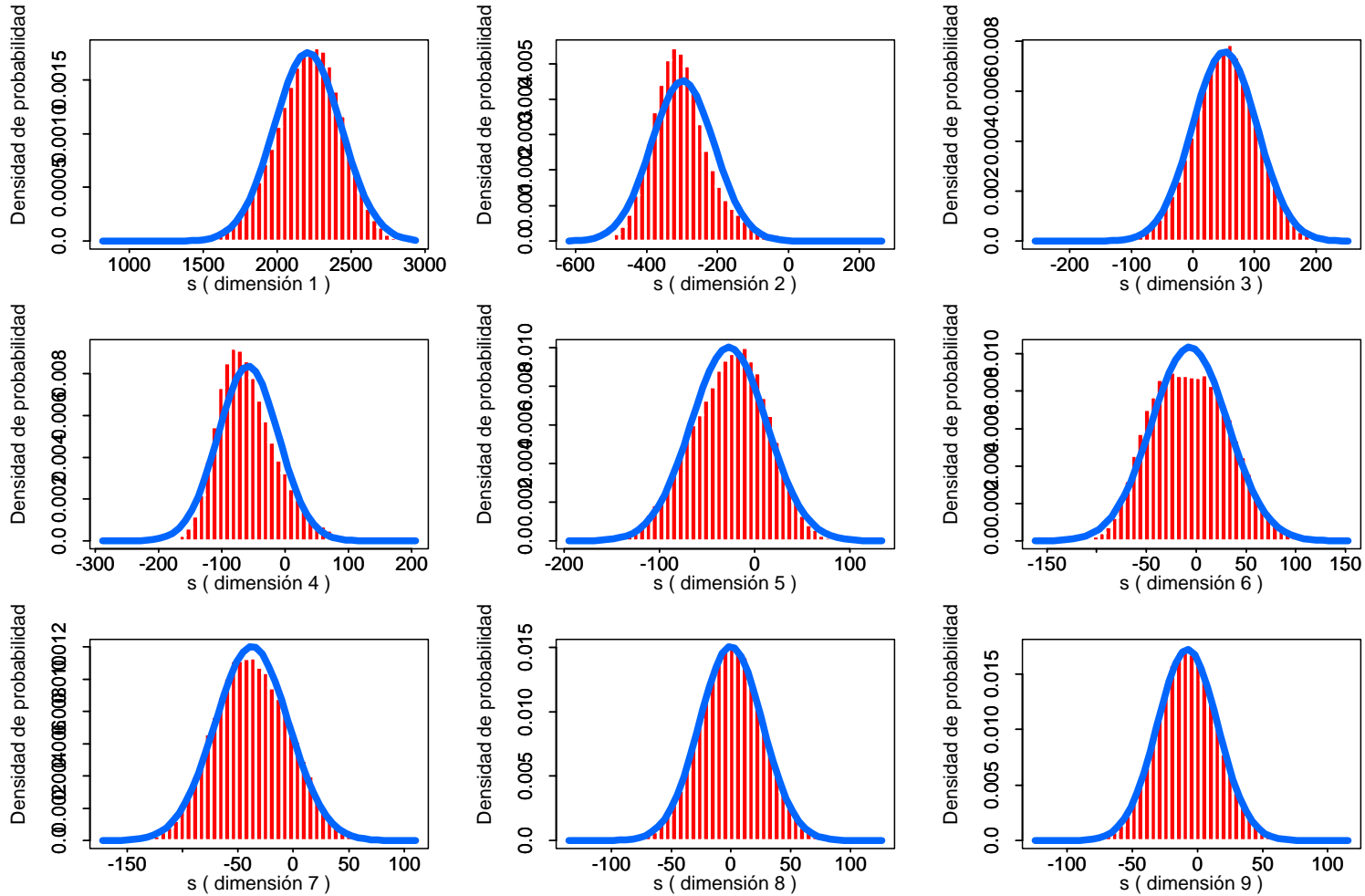
# Ejemplo de mezcla de gaussianas: Una dimensión



$$p(x) = 0.6p_1(x) + 0.4p_2(x)$$

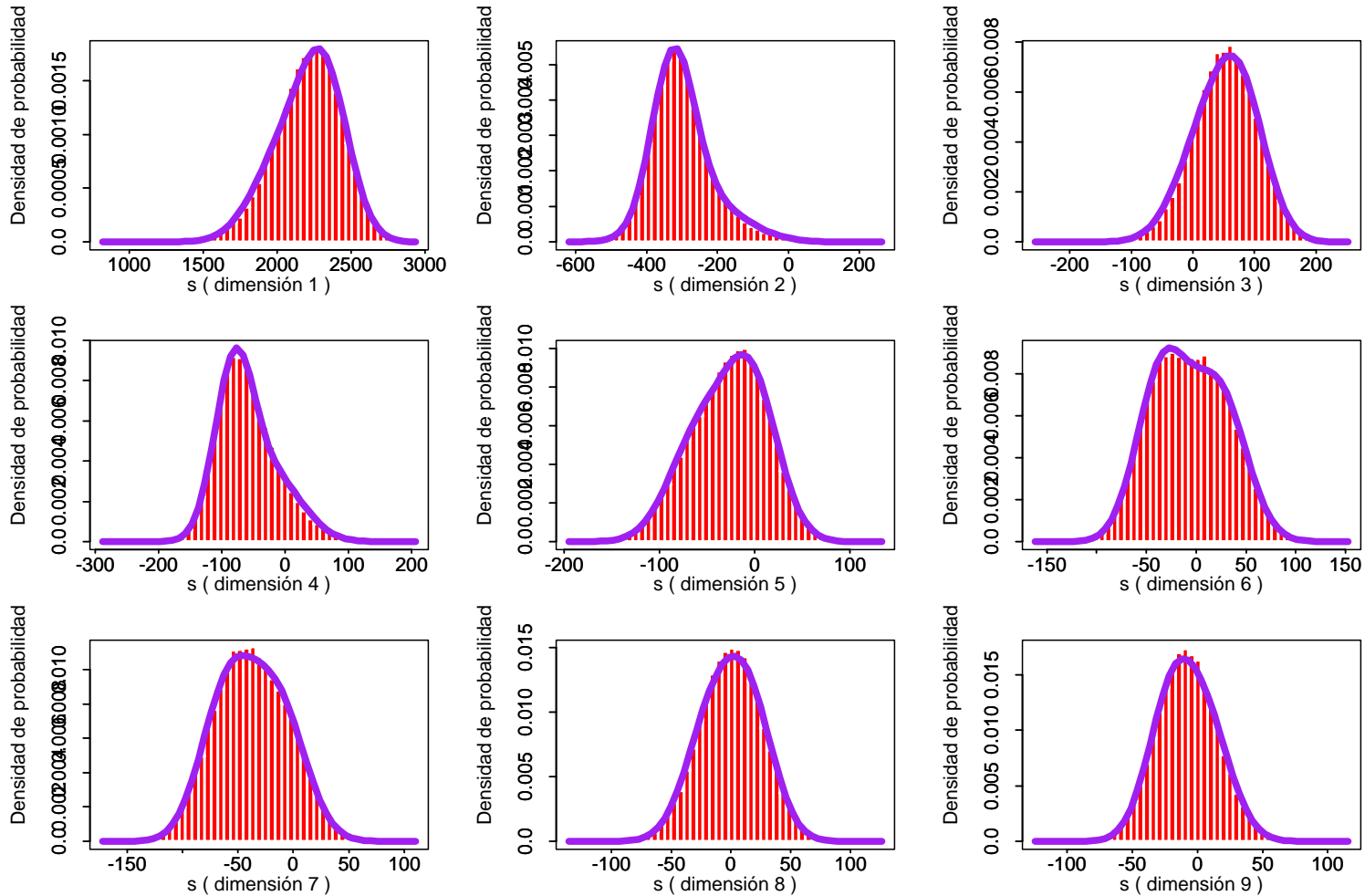
$$p_1(x) \sim N(-\sigma, \sigma^2) \quad p_2(x) \sim N(1.5\sigma, \sigma^2)$$

## 9 primeros MFCC (coeficiente cepstral de frecuencia Mel) de [s]: PDF gaussiano



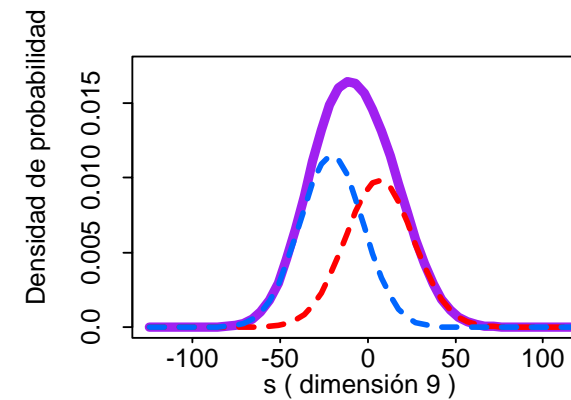
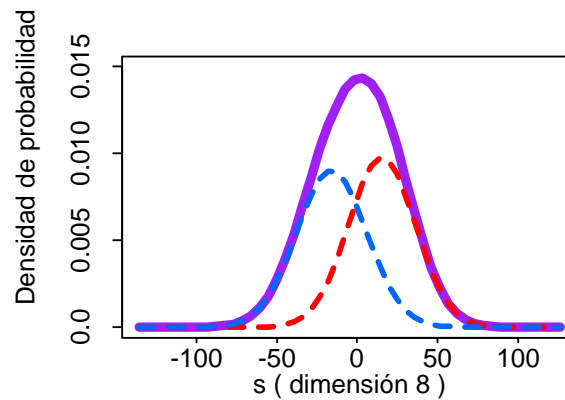
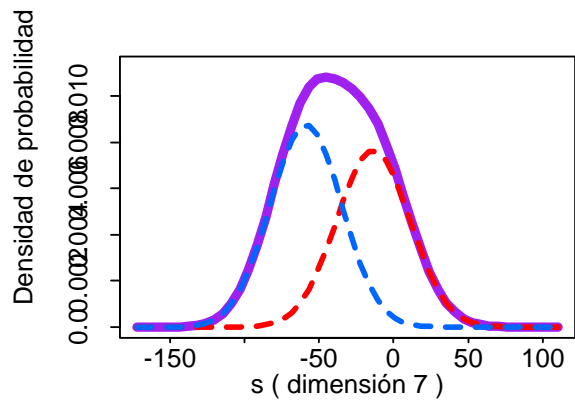
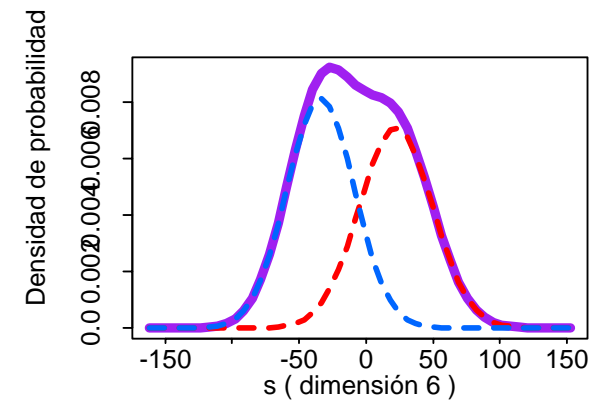
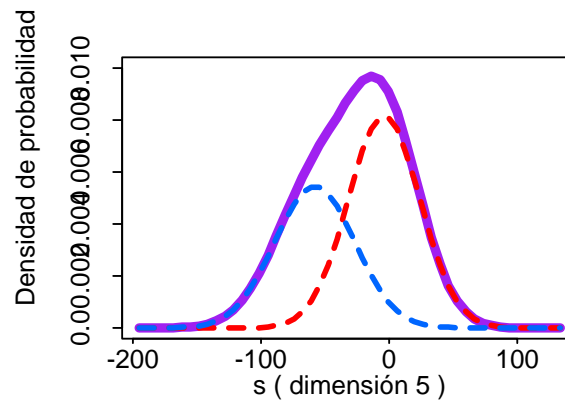
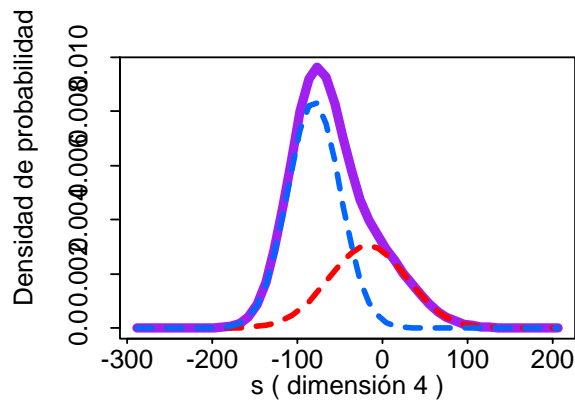
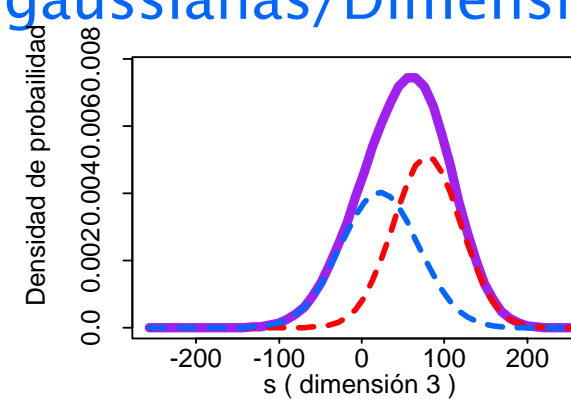
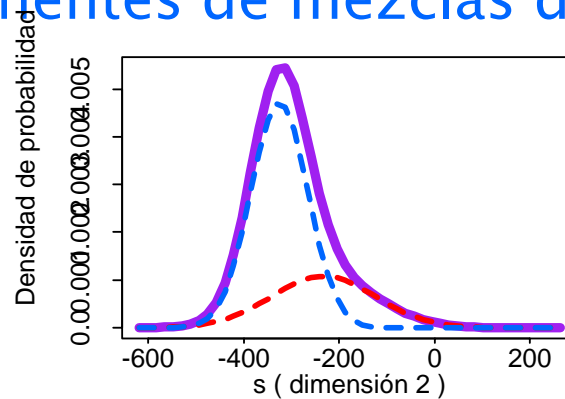
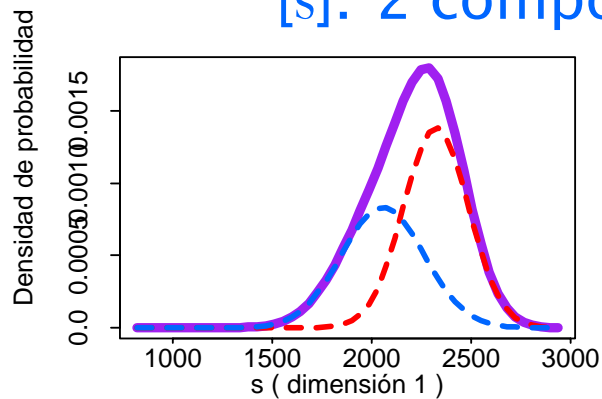
# Mezclas independientes

[s]: 2 componentes de mezclas de gaussianas/Dimensión



# Componentes de mezcla

[s]: 2 componentes de mezclas de gaussianas/Dimensión



# Estimación del parámetro de máxima probabilidad (ML): Medias de mezcla de gaussianas 1D

$$\log L(\mu_k) = \sum_{i=1}^n \log p(x_i) = \sum_{i=1}^n \log \sum_{j=1}^m p(x_i|\omega_j)P(\omega_j)$$

$$\frac{\partial \log L(\mu_k)}{\partial \mu_k} = \sum_i \frac{\partial}{\partial \mu_k} \log p(x_i) = \sum_i \frac{1}{p(x_i)} \frac{\partial}{\partial \mu_k} p(x_i|\omega_k)P(\omega_k)$$

$$\frac{\partial p(x_i|\omega_k)}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}} = p(x_i|\omega_k) \frac{(x_i - \mu_k)}{\sigma_k^2}$$

$$\frac{\partial \log L(\mu_k)}{\partial \mu_k} = \sum_i \frac{P(\omega_k)}{p(x_i)} p(x_i|\omega_k) \frac{(x_i - \mu_k)}{\sigma_k^2} = 0$$

dado  $\frac{p(x_i|\omega_k)P(\omega_k)}{p(x_i)} = P(\omega_k|x_i)$   $\hat{\mu}_k = \frac{\sum_i P(\omega_k|x_i)x_i}{\sum_i P(\omega_k|x_i)}$

Las soluciones de máxima probabilidad poseen la forma de:

$$\hat{\boldsymbol{\mu}}_k = \frac{\frac{1}{n} \sum_i \hat{P}(\omega_k | \mathbf{x}_i) \mathbf{x}_i}{\frac{1}{n} \sum_i \hat{P}(\omega_k | \mathbf{x}_i)}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\frac{1}{n} \sum_i \hat{P}(\omega_k | \mathbf{x}_i) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^t}{\frac{1}{n} \sum_i \hat{P}(\omega_k | \mathbf{x}_i)}$$

$$\hat{P}(\omega_k) = \frac{1}{n} \sum_i \hat{P}(\omega_k | \mathbf{x}_i)$$

- Las soluciones de ML se resuelven normalmente de forma iterativa mediante:
  - La selección de un conjunto de estimaciones iniciales para  $\hat{P}(\omega_k), \hat{\mu}_k, \hat{\Sigma}_k$
  - La utilización de un conjunto de muestras  $n$  para volver a estimar los parámetros de mezcla hasta que se produzca algún tipo de convergencia.
- Los procedimientos de agrupamiento se utilizan normalmente para facilitar las estimaciones iniciales del parámetro.
- Parecido al procedimiento de agrupamiento  $K$ -medias.

## Ejemplo: 4 muestras , 2 densidades

1. Datos:  $\mathcal{X} = \{x_1, x_2, x_3, x_4\} = \{2, 1, -1, -2\}$

2. Inic.:  $p(x|\omega_1) \sim N(1, 1)$   $p(x|\omega_2) \sim N(-1, 1)$   $P(\omega_i) = 0.5$

3. Estimación:

	$x_1$	$x_2$	$x_3$	$x_4$
$P(\omega_1 \mathcal{X})$	0.98	0.88	0.12	0.02
$P(\omega_2 \mathcal{X})$	0.02	0.12	0.88	0.98

$$p(\mathcal{X}) \propto (e^{-0.5} + e^{-4.5})(e^0 + e^{-2})(e^0 + e^{-2})(e^{-0.5} + e^{-4.5})0.5^4$$

4. Recalcular los parámetros de media (sólo mostrado para  $\omega_1$ ):

$$\hat{P}(\omega_1) = \frac{.98+.88+.12+.02}{4} = 0.5$$

$$\hat{\mu}_1 = \frac{.98(2)+.88(1)+.12(-1)+.02(-2)}{.98+.88+.12+.02} = 1.34$$

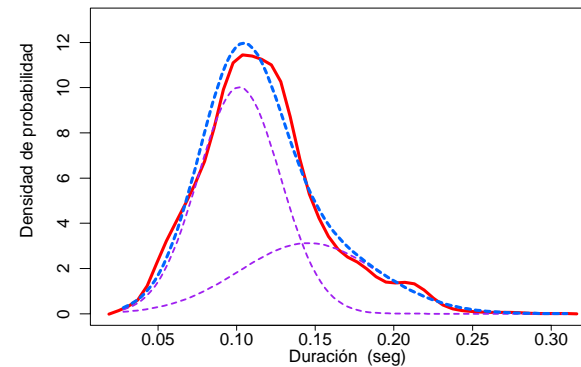
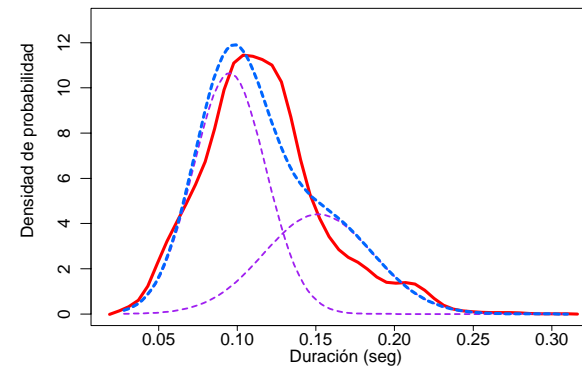
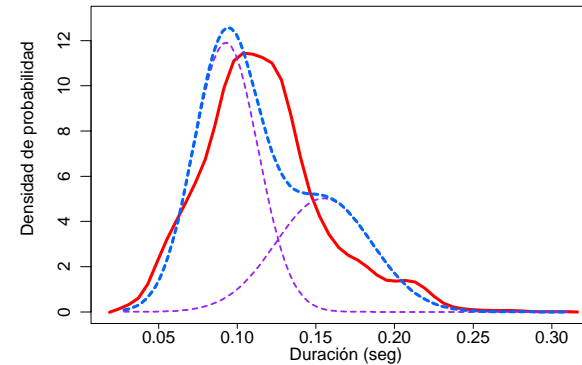
$$\hat{\sigma}_1^2 = \frac{.98(2-1.34)^2+.88(1-1.34)^2+.12(-1-1.34)^2+.02(-2-1.34)^2}{.98+.88+.12+.02} = 0.70$$

5. Repetir los pasos 3,4 hasta que se produzca la convergencia.

## [s] Duración: 2 densidades

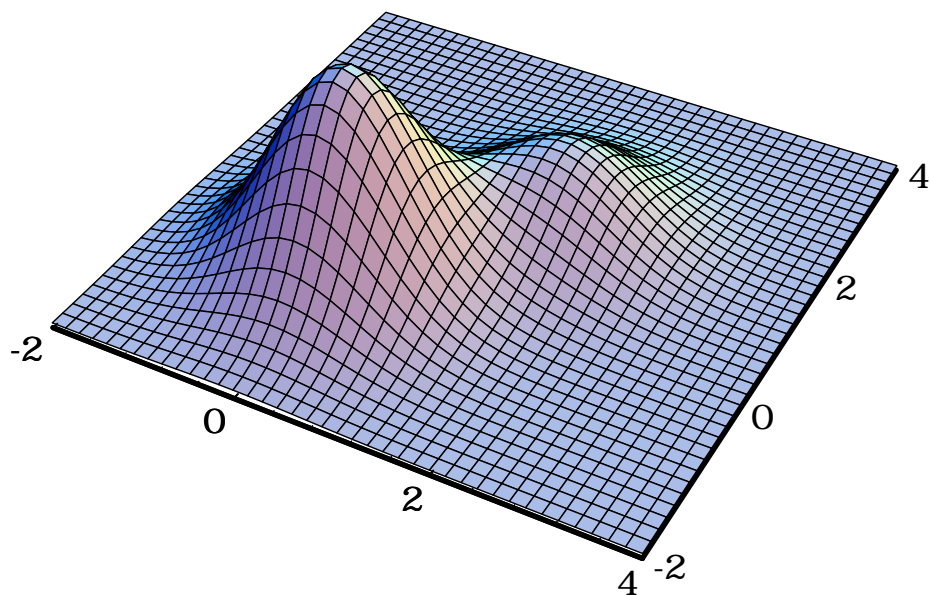
Iter	$\mu_1$	$\mu_2$	$\sigma_1$	$\sigma_2$
0	152	95	35	23
1	150	97	37	24
2	148	98	39	25
3	147	100	41	25
4	146	100	42	26
5	146	102	43	26
6	146	102	44	26
7	145	102	44	26

Iter	$P(\omega_1)$	$P(\omega_2)$	$\log p(\mathcal{X})$
0	.384	.616	2.727
1	.376	.624	2.762
2	.369	.631	2.773
3	.362	.638	2.778
4	.356	.644	2.781
5	.349	.651	2.783
6	.344	.656	2.784
7	.338	.662	2.785

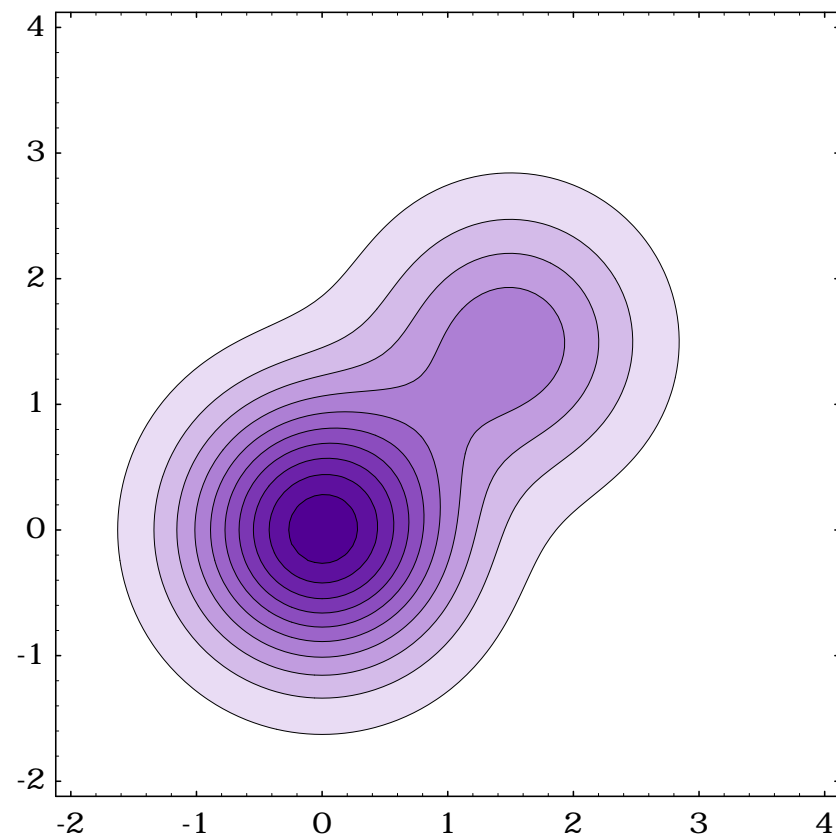


# Ejemplo de mezclas de gaussianas: Dos dimensiones

PDF tridimensional

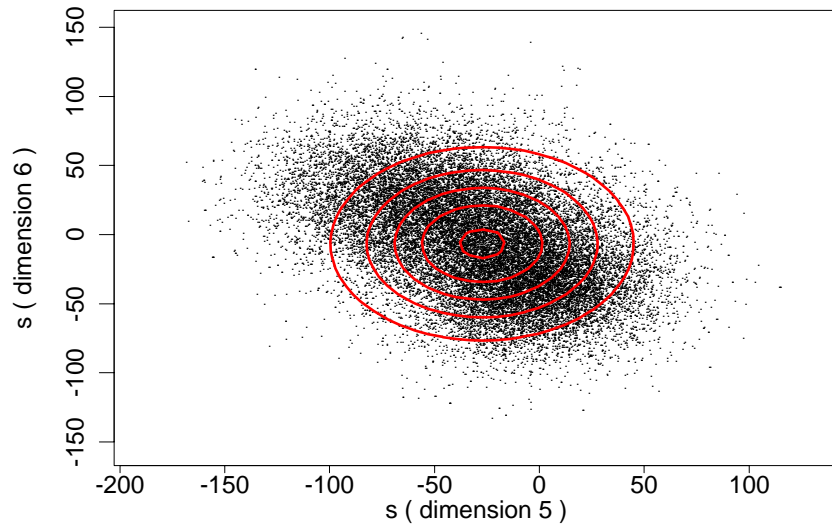


Contorno PDF

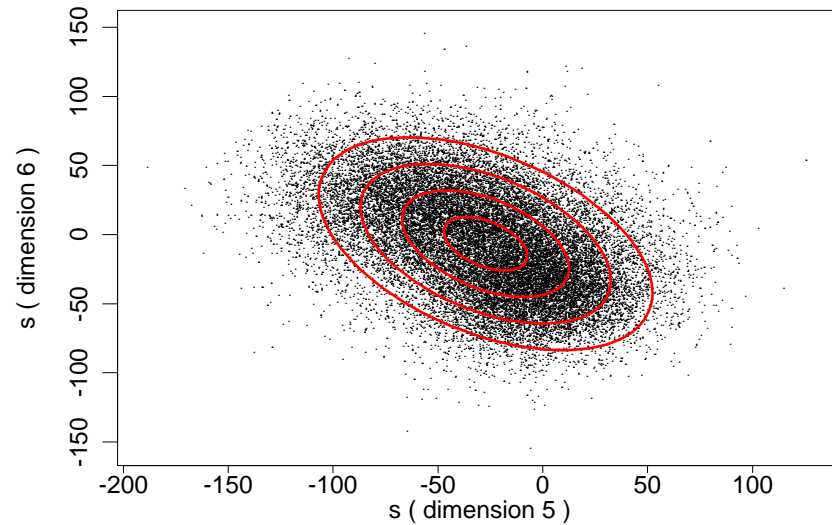


# Mezclas bidimensionales

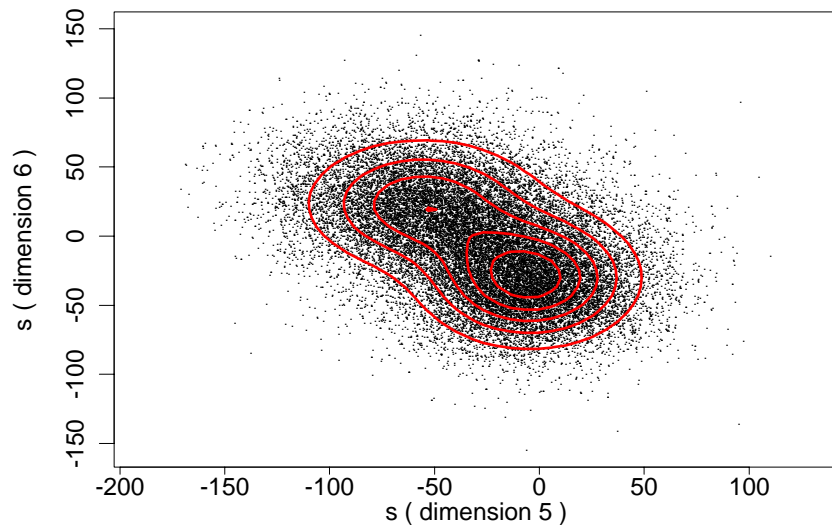
## Covarianza diagonal



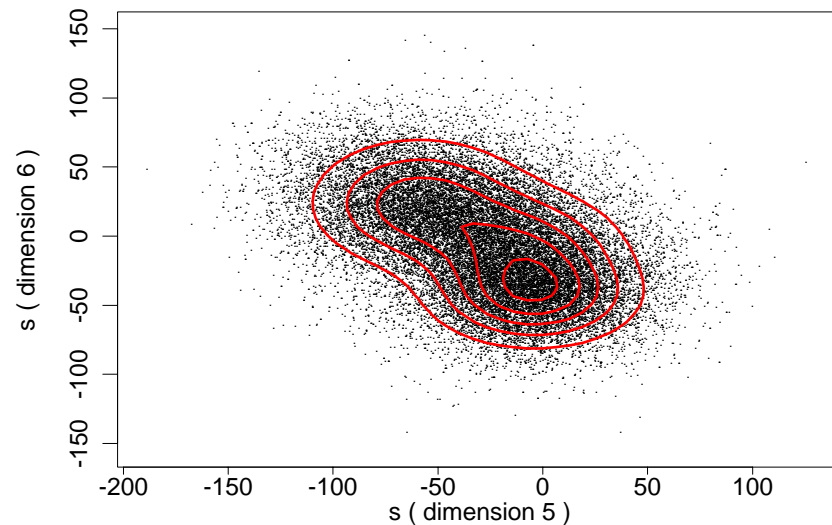
## Covarianza completa



## Dos mezclas

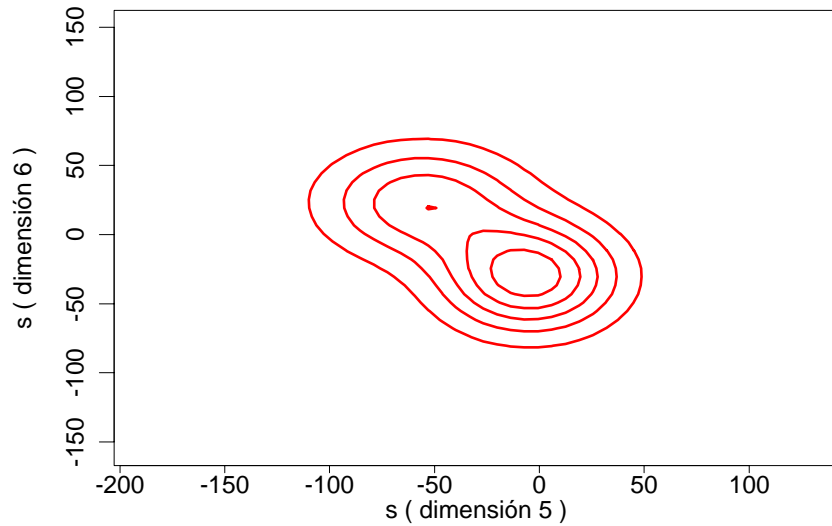


## Tres mezclas

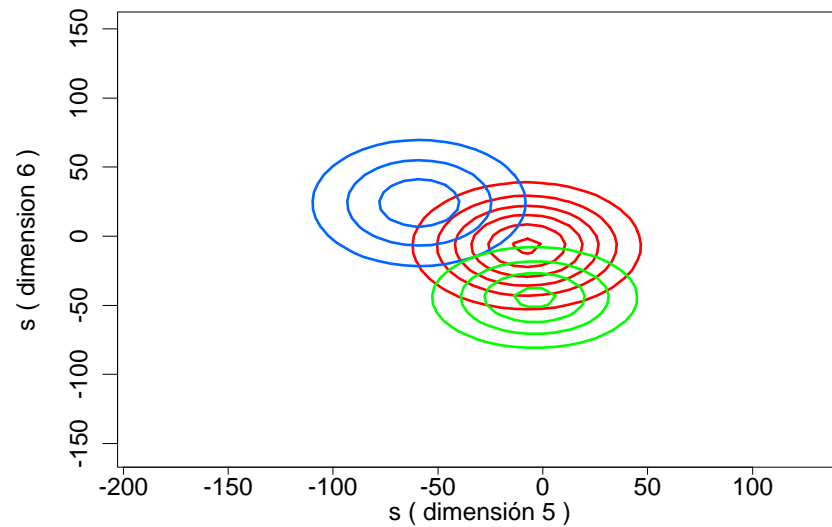
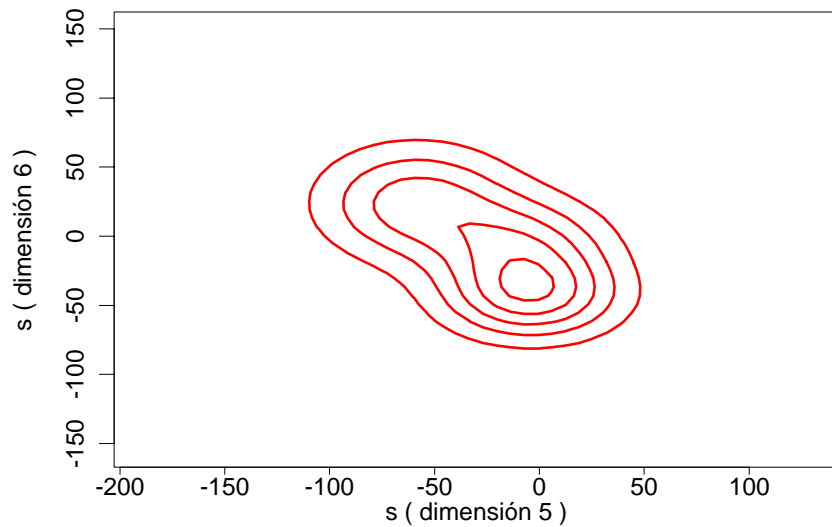
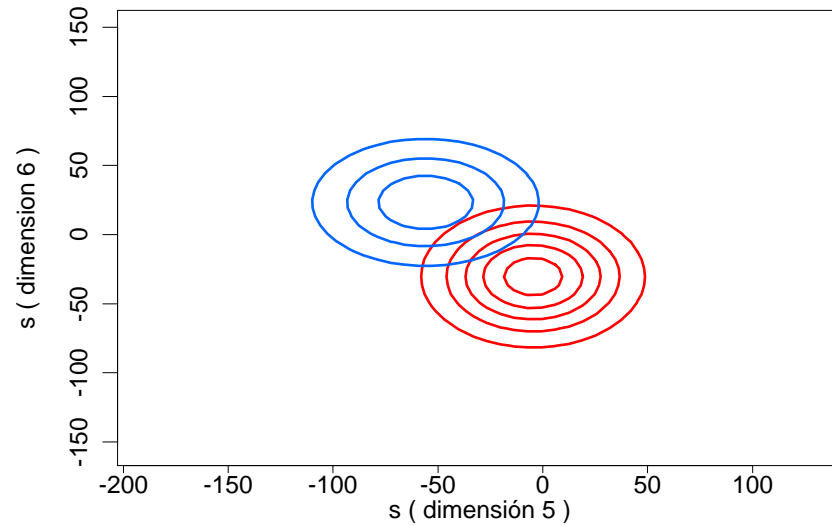


# Componentes bidimensionales

## Mezcla



## Componentes



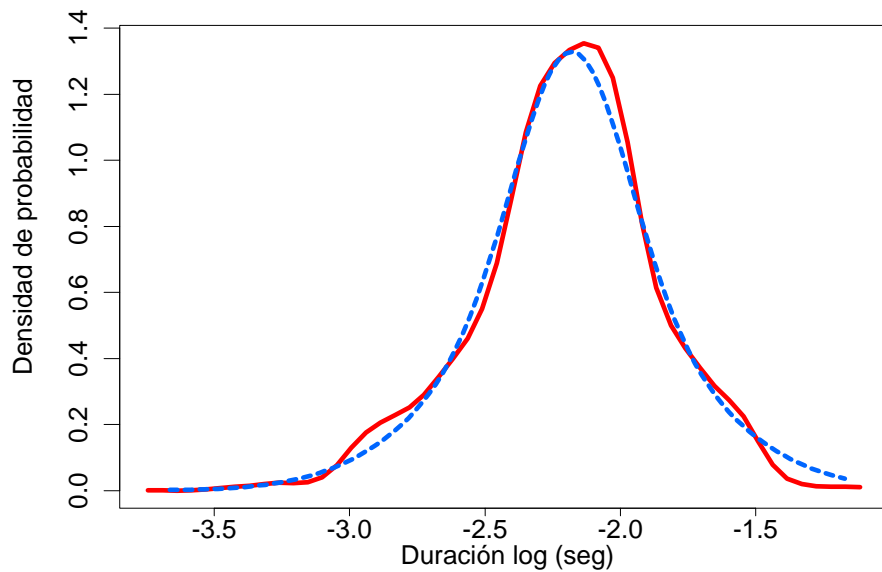
# Mezcla de gaussianas

## Variaciones de implementación

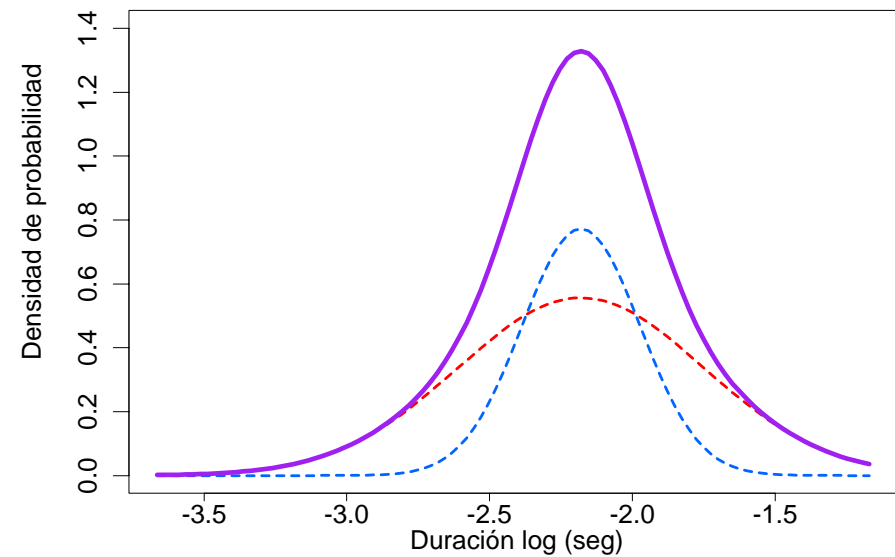
- Se utilizan normalmente las gaussianas diagonales en vez de las de covarianza completa.
  - Pueden reducir el número de parámetros.
  - Pueden modelar el PDF subyacente tan bien como si se estuvieran utilizando suficientes componentes.
- Los parámetros de mezcla normalmente están restringidos a ser los mismos, con el fin de reducir el número de parámetros que deben ser calculados.
  - Las gaussianas de **Richter** comparten la misma media para modelar mejor las colas PDF .
  - **Las mezclas enlazadas** comparten los mismos parámetros gaussianos en *todas* las clases. Únicamente los pesos de mezclas  $P(\omega_i)$  son específicos de la clase . (También conocidos como semi-continuos).

# Mezclas de gaussianas Richter

[s] Duración Log: 2 gaussianas de Richter



Densidad de Richter



Componentes de Richter

# Expectación-Maximización (EM)

- Utilizado para determinar parámetros  $\theta$ , para datos **incompletos**,  $\mathcal{X} = \{\mathbf{x}_i\}$  (p.ej., problemas de aprendizaje no supervisado).
- Introduce la variable  $\mathcal{Z} = \{z_j\}$  para **completar** datos de forma que  $\theta$  pueda resolverse utilizando técnicas de máxima probabilidad (ML).

$$\log L(\theta) = \log p(\mathcal{X}, \mathcal{Z} | \theta) = \sum_{i,j} \log p(\mathbf{x}_i, z_j | \theta)$$

- En realidad  $z_j$  únicamente puede ser estimado por  $P(z_j | \mathbf{x}_i, \theta)$ , por lo tanto, sólo podemos calcular la **expectación** de  $\log L(\theta)$

$$\mathcal{E} = E(\log L(\theta)) = \sum_i \sum_j P(z_j | \mathbf{x}_i, \theta) \log p(\mathbf{x}_i, z_j | \theta)$$

- Las soluciones de EM se calculan iterativamente hasta que se produzca la convergencia.
  1. Cálculo de la **expectación** de  $\log L(\theta)$
  2. Cálculo de los valores  $\theta_l$ , los cuales **maximizan**  $\mathcal{E}$

# Estimación del parámetro de máxima probabilidad (ML): Medias de mezclas de gaussianas 1D

- Sea  $z_i$  el componente id  $\{\omega_j\}$ , cuya  $x_i$  pertenece a

$$\mathcal{E} = E(\log L(\theta)) = \sum_i \sum_j P(z_j|x_i, \theta) \log p(x_i, z_j|\theta)$$

- Convertir a notación del componente mezcla:

$$\mathcal{E} = E(\log L(\mu_k)) = \sum_i \sum_j P(\omega_j|x_i) \log p(x_i, \omega_j)$$

- Diferenciar con respecto a  $\mu_k$ :

$$\frac{\partial \mathcal{E}}{\partial \mu_k} = \sum_i P(\omega_k|x_i) \frac{\partial}{\partial \mu_k} \log p(x_i, \omega_k) = \sum_i P(\omega_k|x_i) \left( \frac{x_i - \mu_k}{\sigma_k^2} \right) = 0$$

$$\hat{\mu}_k = \frac{\sum_i P(\omega_k|x_i) x_i}{\sum_i P(\omega_k|x_i)}$$

# Propiedades de EM

- Cada iteración de EM **aumentará** la probabilidad de  $\mathcal{X}$

$$\begin{aligned} \log \frac{p(\mathcal{X}|\theta')}{p(\mathcal{X}|\theta)} &= \sum_i \log \frac{p(\mathbf{x}_i|\theta')}{p(\mathbf{x}_i|\theta)} = \sum_i \sum_j P(z_j|\mathbf{x}_i, \theta) \log \frac{p(\mathbf{x}_i|\theta')}{p(\mathbf{x}_i|\theta)} \\ &= \sum_i \sum_j P(z_j|\mathbf{x}_i, \theta) \left( \log \frac{p(\mathbf{x}_i|\theta')}{p(\mathbf{x}_i, z_j|\theta')} \frac{p(\mathbf{x}_i, z_j|\theta)}{p(\mathbf{x}_i|\theta)} + \log \frac{p(\mathbf{x}_i, z_j|\theta')}{p(\mathbf{x}_i, z_j|\theta)} \right) \end{aligned}$$

- Se utiliza la regla de Bayes y la distancia métrica de Kullback-Liebler:

$$\frac{p(\mathbf{x}_i, z_j|\theta)}{p(\mathbf{x}_i|\theta)} = P(z_j|\mathbf{x}_i, \theta) \quad \sum_j P(z_j|\mathbf{x}_i, \theta) \log \frac{P(z_j|\mathbf{x}_i, \theta)}{P(z_j|\mathbf{x}_i, \theta')} \geq 0$$

- Dado que  $\theta'$  estaba definido para maximizar  $E(\log L(\theta))$ :

$$\sum_i \sum_j P(z_j|\mathbf{x}_i, \theta) \log \frac{p(\mathbf{x}_i, z_j|\theta')}{p(\mathbf{x}_i, z_j|\theta)} \geq 0$$

- Se combinan estas dos propiedades:  $p(\mathcal{X}|\theta') \geq p(\mathcal{X}|\theta)$

## Reducción de la dimensionalidad

- Dado un grupo de entrenamiento, la estimación del parámetro PDF se hace menos robusto a medida que aumenta la dimensionalidad.
- Las crecientes dimensiones pueden dificultar el obtención de una visión de la estructura subyacente.
- Existen técnicas analíticas que pueden transformar un espacio de muestreo en un grupo distinto de dimensiones.
  - Si las dimensiones originales guardan correlación, la misma información puede requerir menos dimensiones.
  - El espacio transformado tendrá por lo general más distribuciones normales que el espacio original.
  - Si las nuevas dimensiones son ortogonales, podría ser más fácil modelar el espacio transformado.

## Análisis de los componentes principales

- Transformación lineal del vector de dimensión  $d$ ,  $\mathbf{x}$ , al vector de dimensión  $d'$ ,  $\mathbf{y}$ , mediante vectores **ortonormales**,  $\mathbf{W}$

$$\mathbf{y} = \mathbf{W}^t \mathbf{x} \quad \mathbf{W} = \{\mathbf{w}_1, \dots, \mathbf{w}_{d'}\} \quad \mathbf{W}^t \mathbf{W} = \mathbf{I}$$

- Si  $d' < d$ ,  $\mathbf{x}$  sólo puede reconstruirse parcialmente a partir de  $\mathbf{y}$

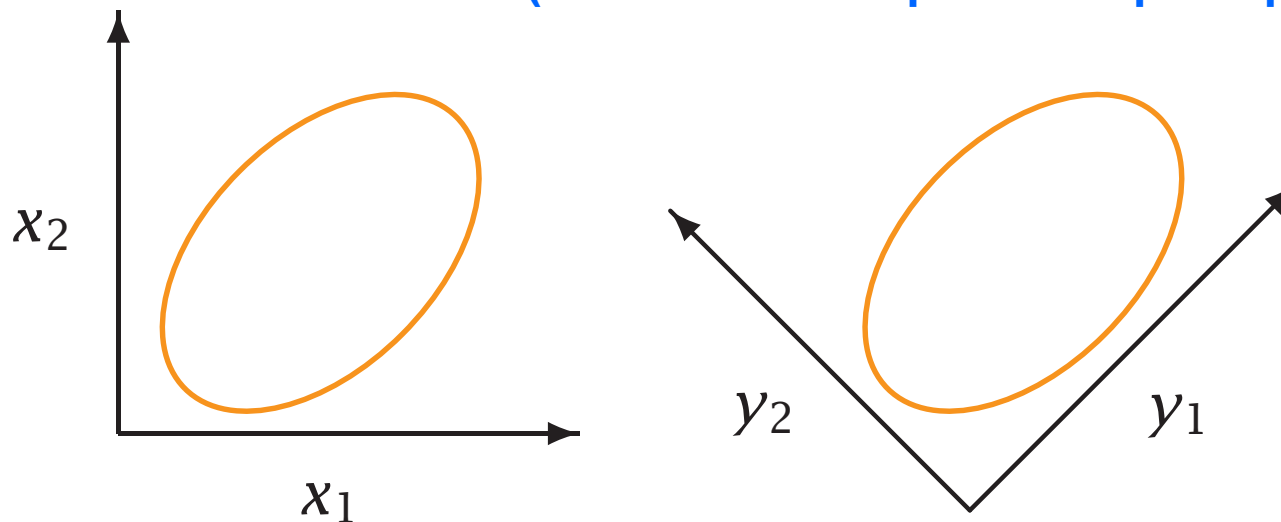
$$\hat{\mathbf{x}} = \mathbf{W}\mathbf{y}$$

- Los **componentes principales**  $\mathbf{W}$ , minimizan la distorsión  $\mathcal{D}$ , entre  $\mathbf{x}$  y  $\hat{\mathbf{x}}$  en el entrenamiento de datos,  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$

$$\mathcal{D} = \sum_{i=1}^n \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2$$

- Conocido también como la expansión de Karhunen-Loève (K-L)  
(Los  $\mathbf{w}_i$  son sinusoides para algunos procesos estocásticos).

# Computación del PCA (Análisis de componentes principales)



- $W$  corresponde a los primeros **autovectores**  $d_l$ ,  $P$ , de  $\Sigma$

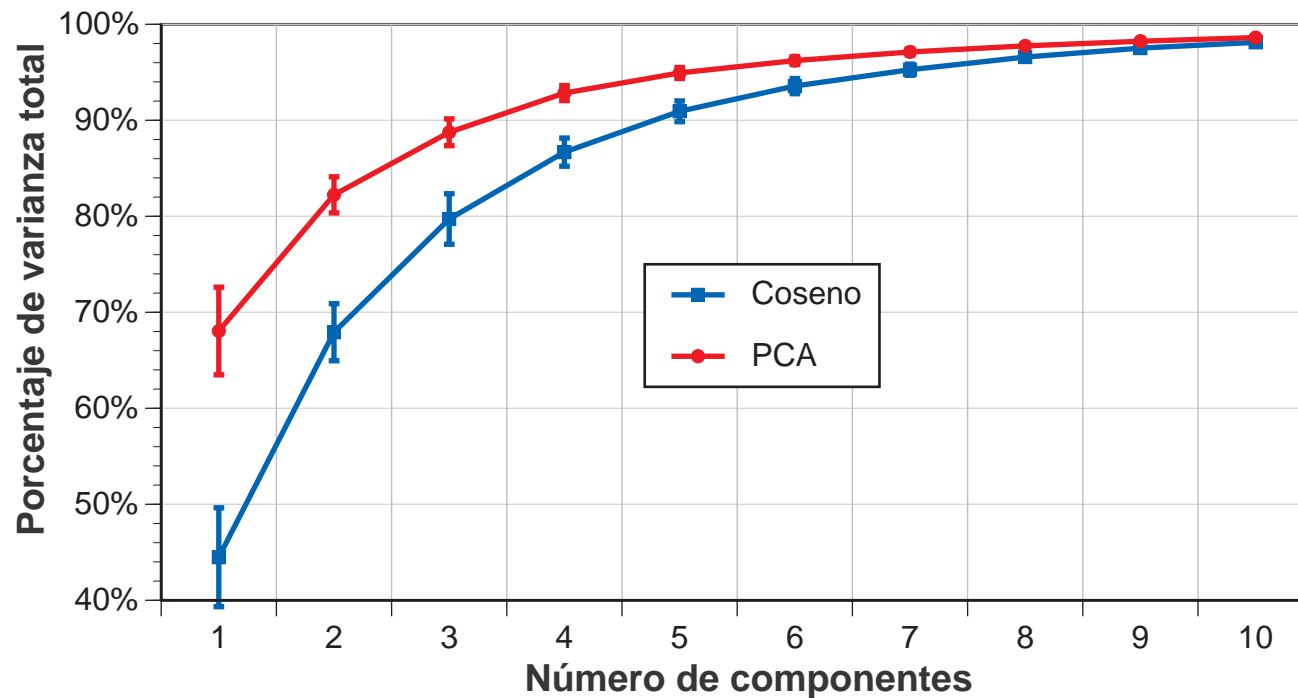
$$P = \{e_1, \dots, e_d\} \quad \Sigma = P \Lambda P^t \quad w_i = e_i$$

- La estructura de covarianza completa del espacio original  $\Sigma$ , se transforma en una estructura de covarianza diagonal,  $\Lambda$
- Los **autovalores**  $\{\lambda_1, \dots, \lambda_{d'}\}$ , representan las varianzas en  $\Lambda$
- Los ejes del espacio  $d_l$  contienen la máxima cantidad de varianza.

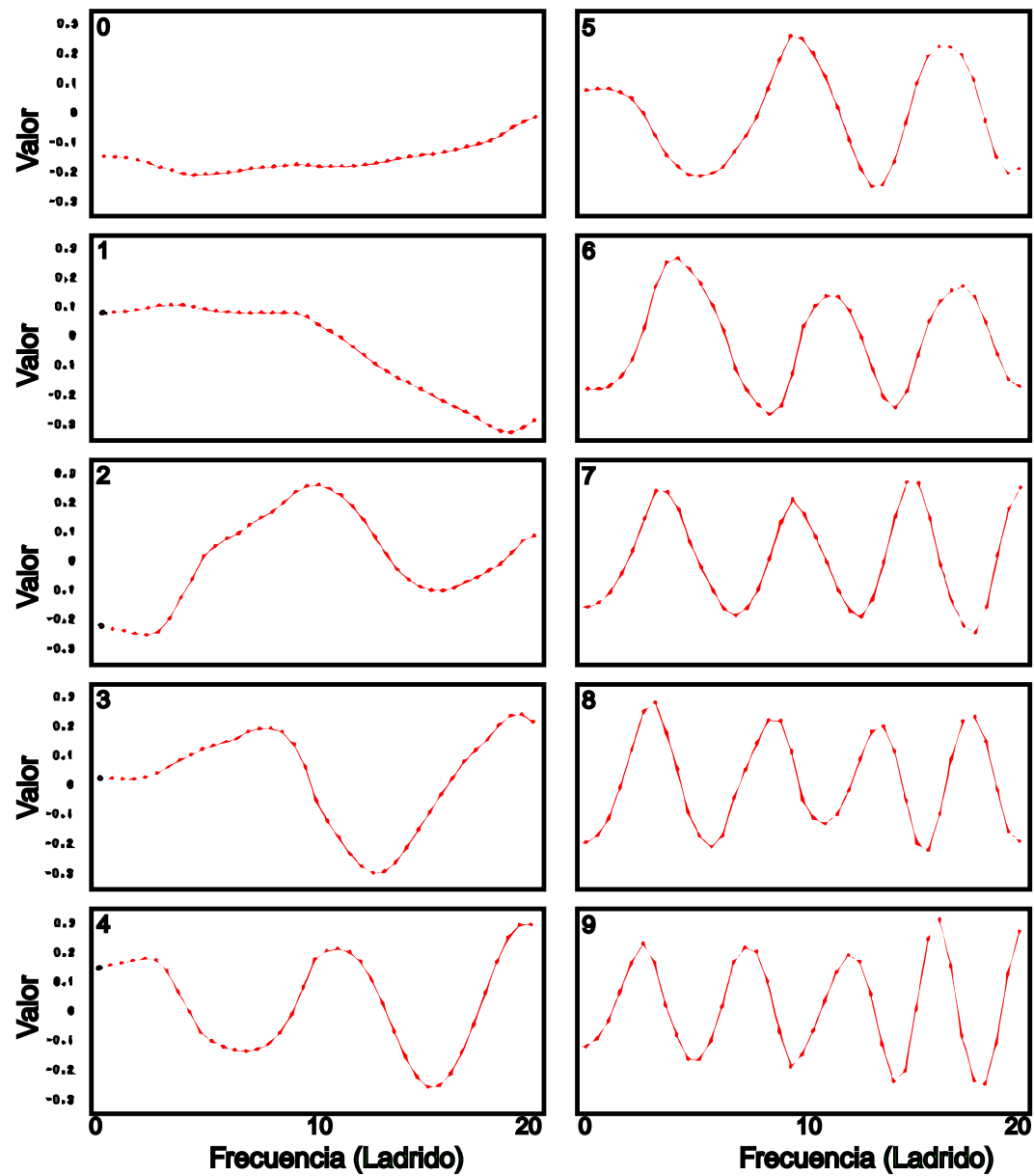
$$D = \sum_{i=d'+1}^d \lambda_i$$

# Ejemplo del PCA

- Velocidad original de respuesta de la media del vector característico ( $d = 40$ )
- Datos obtenidos de 100 hablantes del corpus TIMIT.
- Los 10 primeros componentes explican el 98% de la varianza total.

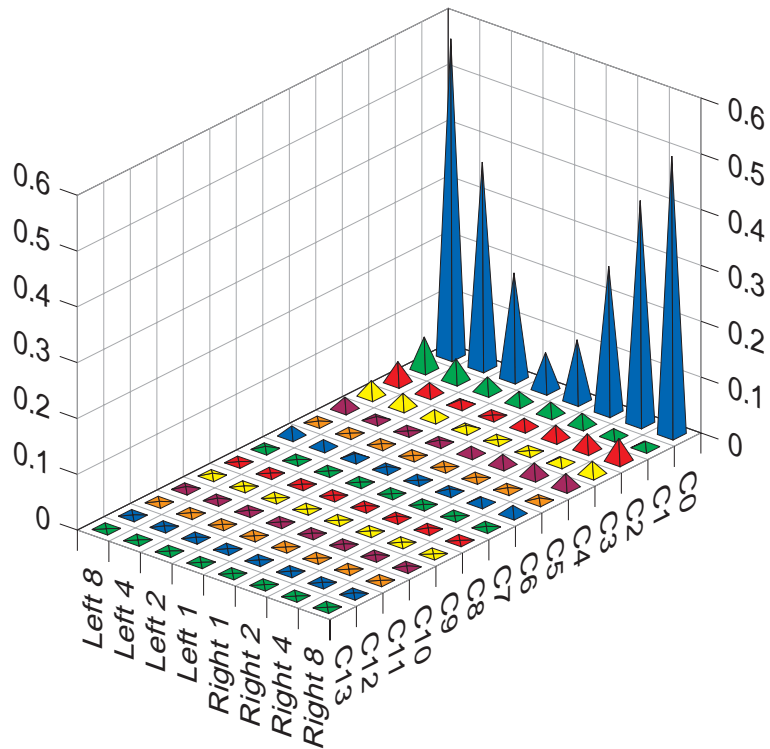


## Ejemplo de PCA

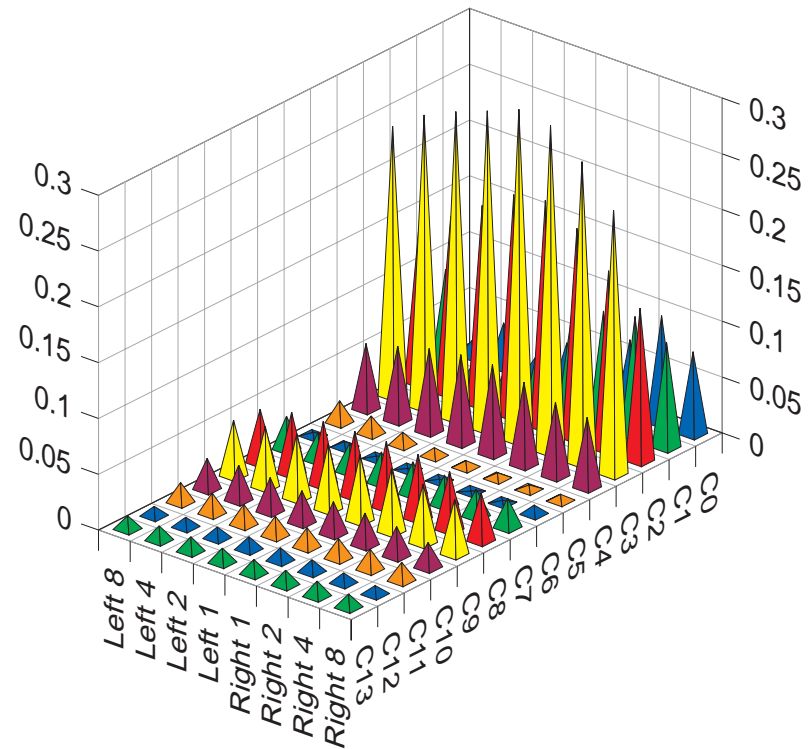


# PCA para la clasificación de límites

- Ocho promedios no uniformes de 14 MFCC (coeficientes cepstrales de frecuencia Mel).
- Las primeras 50 dimensiones utilizadas para la clasificación.



Segundo componente



Séptimo componente

# Cuestiones del PCA

- El PCA se puede llevar a cabo mediante:
  - Covarianzas  $\Sigma$
  - Matriz de coeficientes de correlación  $\mathcal{P}$

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \quad |\rho_{ij}| \leq 1$$

- $\mathcal{P}$  se prefiere normalmente cuando las dimensiones de entrada poseen rangos considerablemente distintos.
- El PCA se puede utilizar para normalizar o **blanquear** el espacio de la dimensión  $d$ , con el fin de simplificar el procesamiento posterior.

$$\Sigma \implies \mathcal{P} \implies \Lambda \implies \mathbf{I}$$

- La operación de blanqueamiento se puede realizar en un paso:  $\mathbf{z} = \mathbf{V}^t \mathbf{x}$

# Prueba de significancia

- Para comparar correctamente los resultados de los distintos algoritmos clasificadores,  $A_1$  y  $A_2$ , es necesario llevar a cabo las siguientes pruebas de significancia.
  - Las grandes diferencias pueden ser insignificantes para los pequeños grupos de pruebas.
  - Las pequeñas diferencias pueden ser significativas para los grupos de pruebas grandes.
- Las pruebas de significancia general evalúan la hipótesis basada en que la probabilidad  $p_i$  de ser correcto, es la misma para ambos algoritmos.
- Las comparaciones más potentes se pueden realizar mediante la práctica de un entrenamiento común más corpórea de prueba, y la aplicación de un mismo criterio de evaluación.
  - Los resultados reflejan diferencias en los algoritmos, más que diferencias fortuitas en los grupos de pruebas.
  - Las pruebas de significancia pueden ser más exactas cuando se utilizan datos **idénticos**, dado que en vez de aplicarse a todas las muestras, se aplicarán sólo a aquellas mal clasificadas por uno de los algoritmos.

# Prueba de significancia de McNemar

- Cuando los algoritmos  $A_1$  y  $A_2$  se prueban sobre datos idénticos, podemos colapsar los resultados en una **matriz de recuentos de 2x2**

$A_1/A_2$	Correcto	Incorrecto
Correcto	$n_{00}$	$n_{01}$
Incorrecto	$n_{10}$	$n_{11}$

- Para comparar algoritmos, probamos la hipótesis nula  $\mathcal{H}_0$  basada en que

$$p_1 = p_2, \text{ o } n_{01} = n_{10}, \text{ or } q = \frac{n_{01}}{n_{01} + n_{10}} = \frac{1}{2}$$

- Dado  $\mathcal{H}_0$ , la probabilidad de observar  $K$  muestras, clasificadas de forma asimétrica de entre  $n = n_{01} + n_{10}$ , presenta una PMF (función de probabilidad) de la binomial.

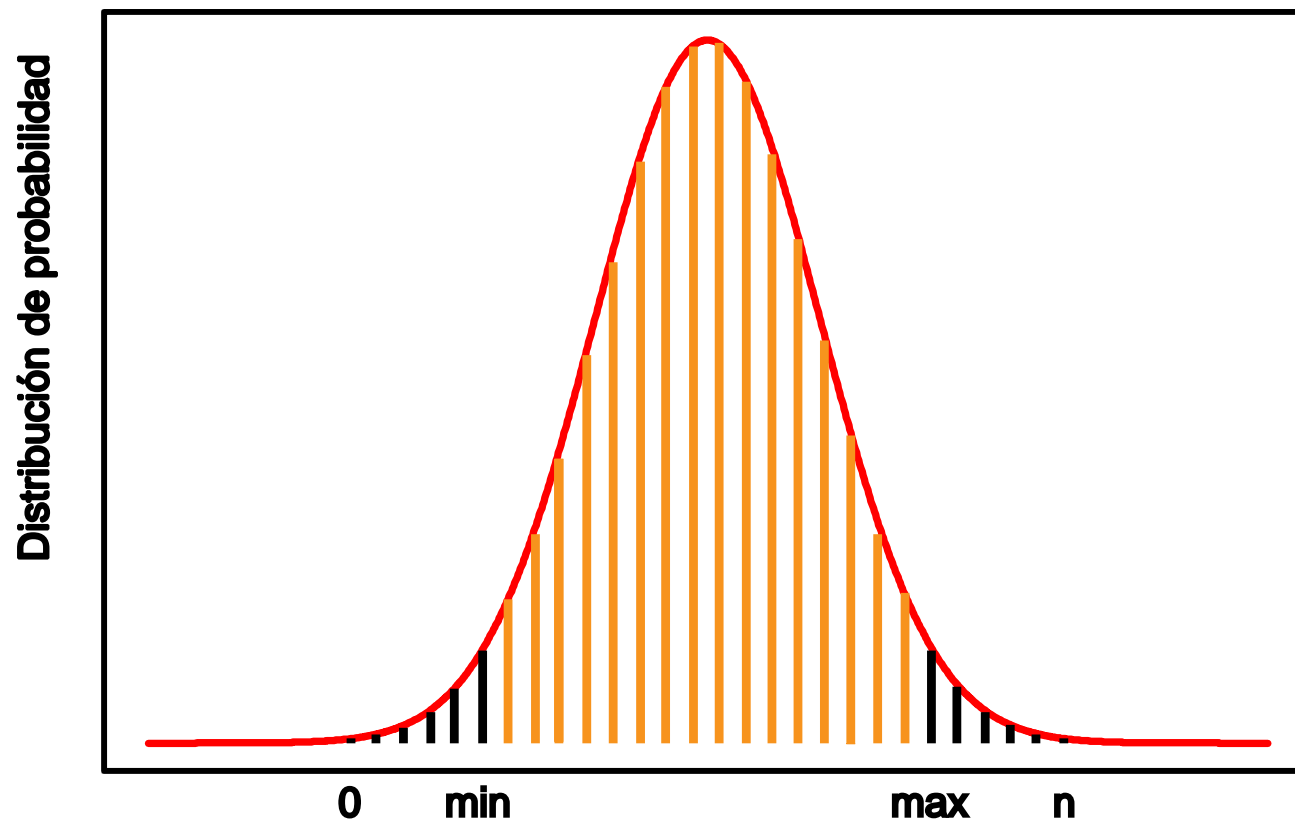
$$P(k) = \binom{n}{k} \left(\frac{1}{2}\right)^n$$

- El test de McNemar mide la probabilidad  $P$  de todos los casos que satisfacen o exceden la distribución asimétrica observada, y prueba  $P < \alpha$

# Prueba de significancia de McNemar (continuación)

- La probabilidad  $P$  se calcula sintetizando las colas de la PMF.

$$P = \sum_{k=0}^l P(k) + \sum_{k=m}^n P(k) \quad l = \min(n_{01}, n_{10}) \quad m = \max(n_{01}, n_{10})$$



- Para  $n$  grandes, se supone normalmente una distribución normal.

## Ejemplo de la prueba de significancia (Gillick y Cox, 1989)

- Grupo de pruebas comunes de 1400 muestras.
- Los algoritmos  $A_1$  y  $A_2$  dan 72 y 62 errores.
- ¿Son **significativas** las diferencias?

		$A_2$		
$A_1$	1266	62	$n = 134$	$m = 72$
	72	0		

		$A_2$		
$A_1$	1325	3	$n = 16$	$m = 13$
	13	59		

		$A_2$		
$A_1$	1328	0	$n = 10$	$m = 10$
	10	62		

# MIT

## Referencias

- Huang, Acero y Hon, *Spoken Language Processing*, Prentice-Hall, 2001.
- Duda, Hart y Stork, *Pattern Classification*, John Wiley & Sons, 2001.
- Jelinek, *Statistical Methods for Speech Recognition*. MIT Press, 1997.
- Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, 1995.
- Gillick y Cox, Some Statistical Issues in the Comparison of Speech Recognition Algorithms, *Proc. ICASSP*, 1989.