Modelos de Mixtura Gaussiana

K-Means y clustering jerárquico

- K-Means y clustering jerárquico son heurísticos
- Las asignaciones de cluster son “duras”
  - Un punto de dato solo puede pertenecer a 1 clase

Modelos de Mixtura

- Asumen que la data es generada por un proceso estadístico que es una combinación de componentes
- Clustering: determinar qué componente está generando el data point.
- La clasificación es “suave”
  - Un punto de dato puede pertenecer a más de 1 clase
  - El resultado es una lista de la probabilidad de que pertenezca a una clase determinada

Recordatorio de Distribución Normal

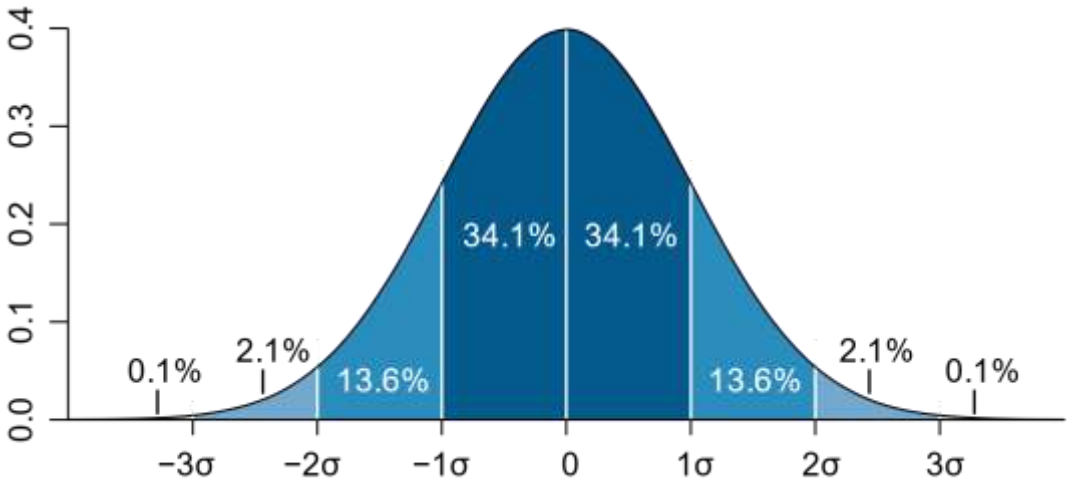
- Para 1D La función de densidad de la probabilidad (PDF) está definida como:

- $PDF(x) = \mathcal{N}(x|\mu, \sigma) =$

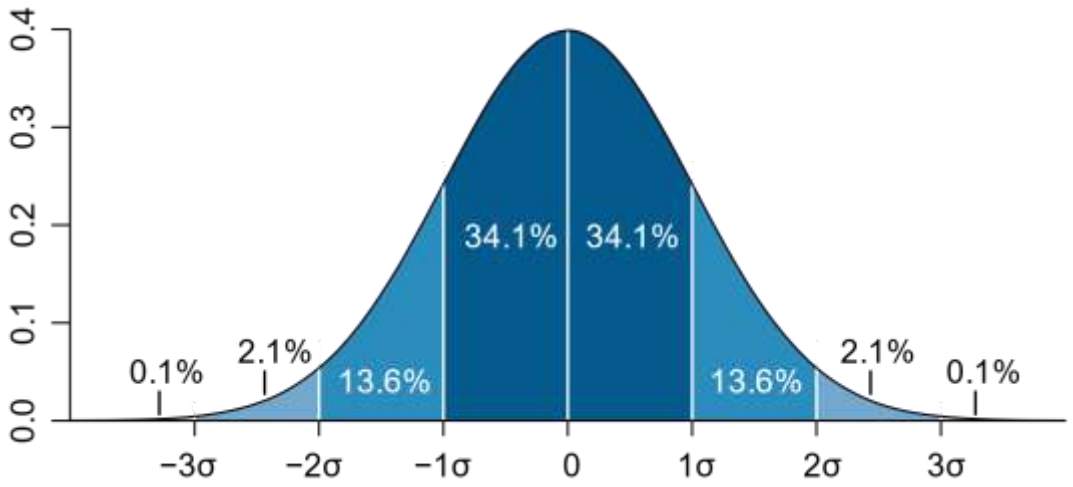
$$\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{x-\mu}{2\sigma}\right)^2$$

Constante para que el área bajo la curva sea 1

Da forma a la "campana"



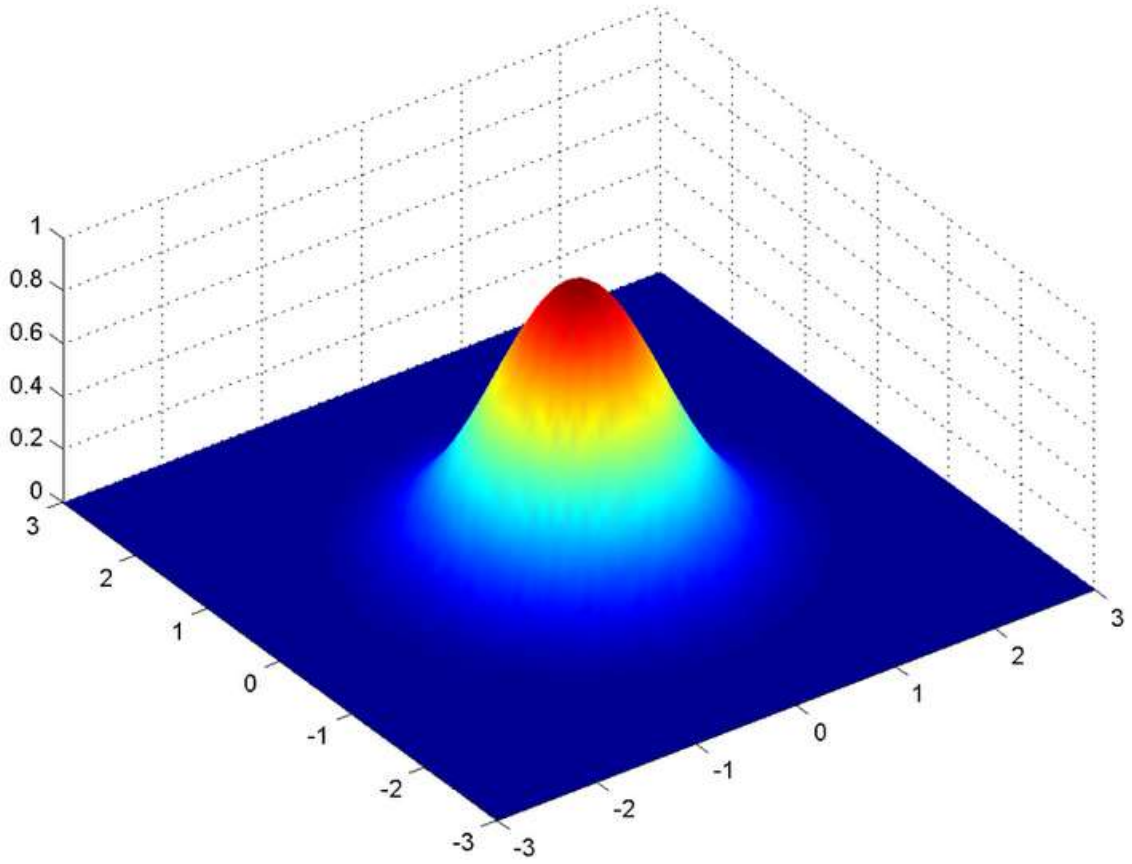
Desviación Estándar σ (sigma)



- Es la raíz cuadrada de la varianza
- Indica la variabilidad con respecto a la media μ

- Para un PDF Gaussiano
  - El 68% del área bajo la curva se encuentra dentro de **una** desviación estándar de la media
  - 95% se encuentra dentro de **dos** desviaciones estándar
  - 99% se encuentra dentro de **tres**.

Múltiples dimensiones



Múltiples Dimensiones

$$PDF(x) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

Constante para que el área bajo la curva sea 1

Da forma a la "campana"

x es el vector, o el punto de dato de D dimensiones

μ es el vector media de D dimensiones

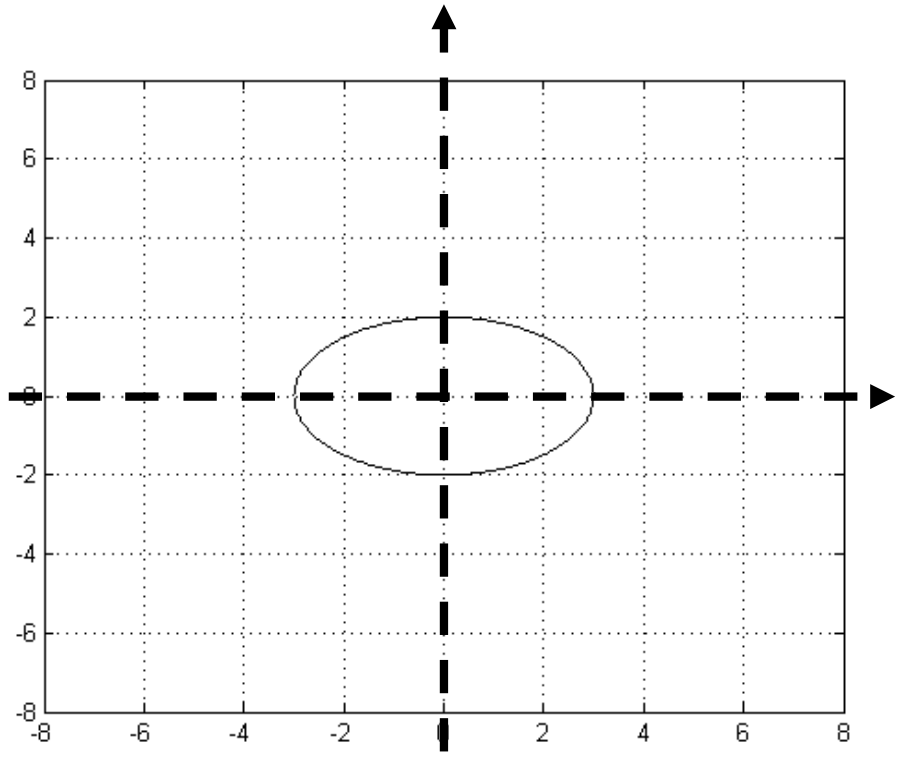
Σ es la matriz de covarianza de $D \times D$ dimensiones

Visualizando la covarianza para 2D

- Podemos dibujar un contorno de radio 1 desviación estándar.

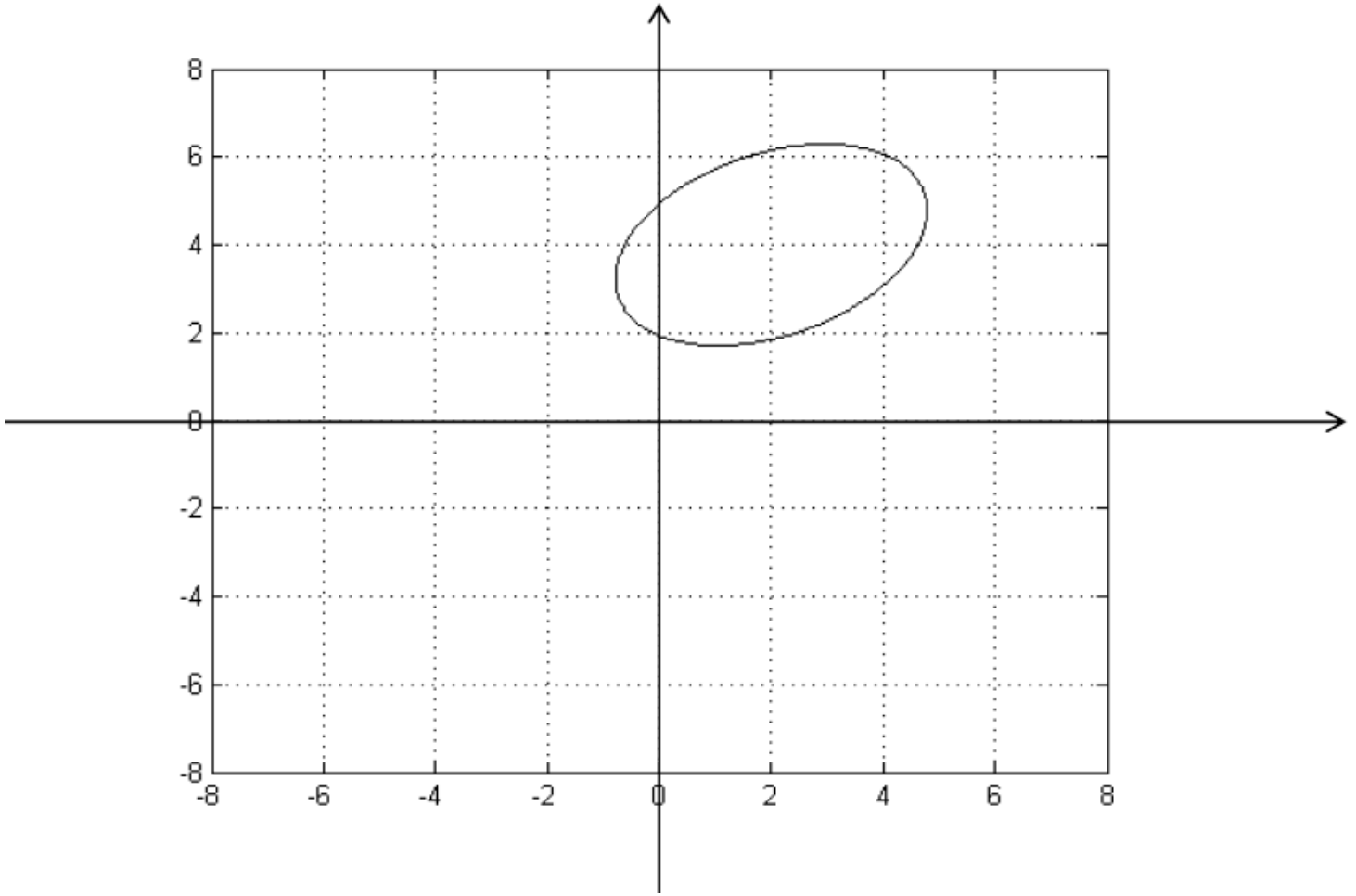
- $\Sigma = \begin{bmatrix} 9 & 0 \\ 0 & 4 \end{bmatrix}$
- $\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

- $\Sigma = \begin{bmatrix} var\ (x_1) & cov(x_1, x_2) \\ cov(x_2, x_1) & var\ (x_2) \end{bmatrix}$



Visualizando la covarianza para 2D

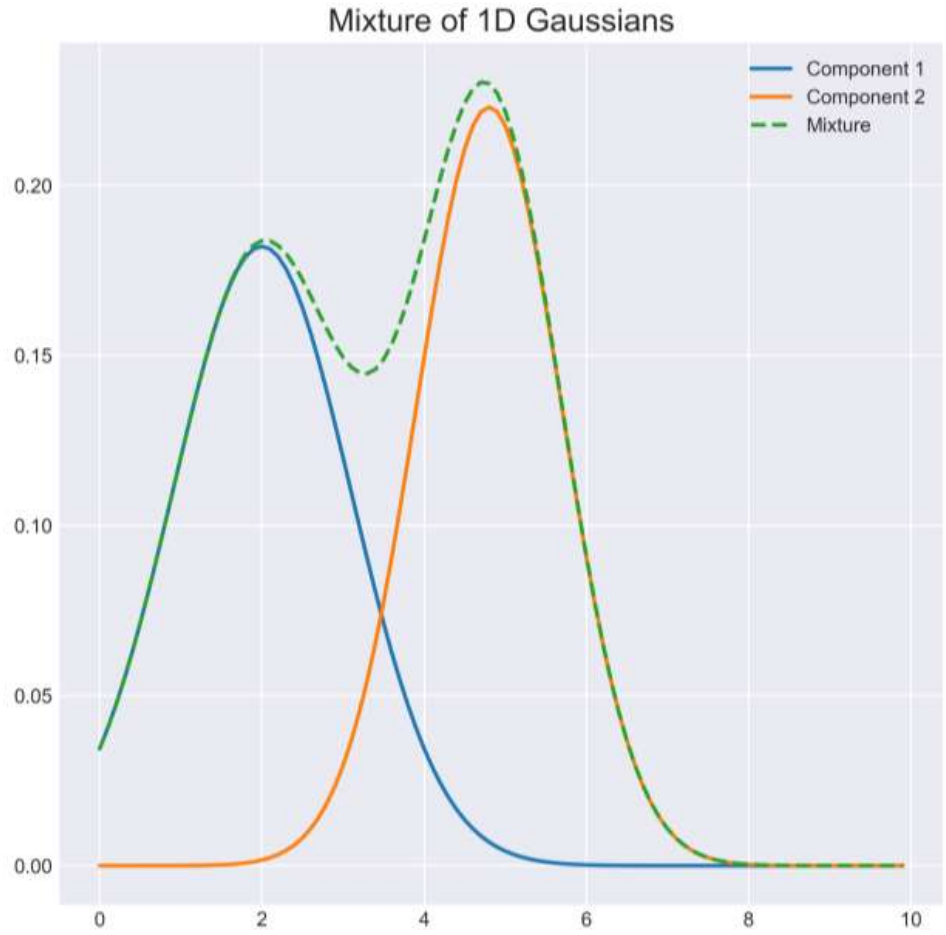
- $\Sigma = \begin{bmatrix} 7.75 & 2.17 \\ 2.17 & 5.25 \end{bmatrix}$
- $\mu = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$



Modelo de Mixtura Gaussiana

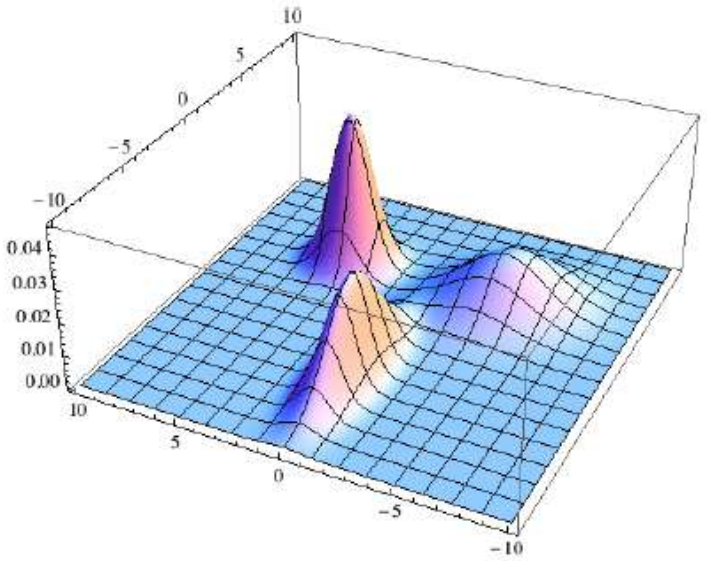
Modelo de Mixtura Gaussiana

- El modelo más común que asume que la data está siendo generada por varias distribuciones normales, cada una con media μ y desviación estándar σ

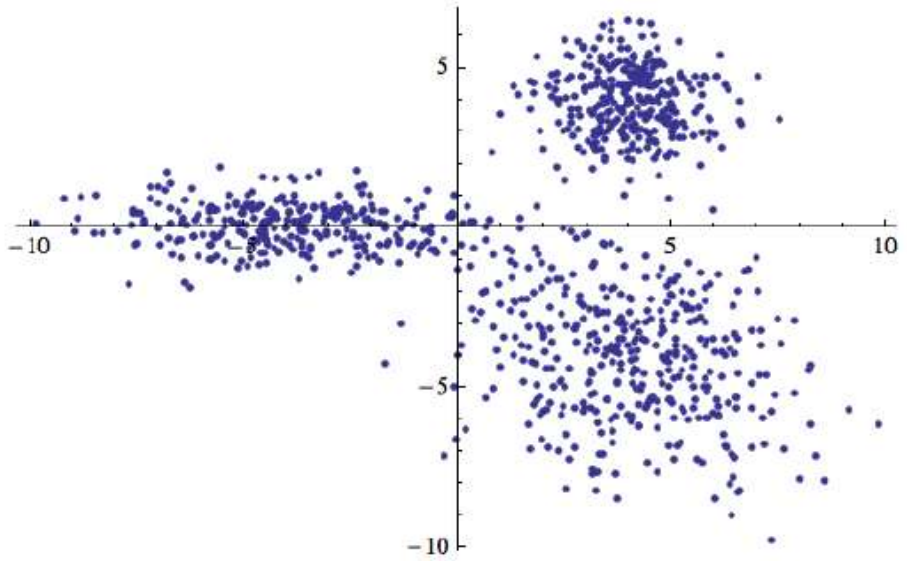


Modelo de Mixtura Gaussiana

Para dataset con D features, tendremos K distribuciones gaussianas, cada una con un vector media μ de D dimensiones y una matriz de covarianza Σ de $D \times D$ dimensiones



(a) A probability distribution on $\mathbb{R}^2$.



(b) Data sampled from this distribution.

Modelo de Mixtura Gaussiana (Normal)

Para K distribuciones gaussianas (K clusters) la función de densidad de la totalidad de los datos está definida por:

$$PDF(x) = \sum_{k=1}^{K} \pi_k * \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

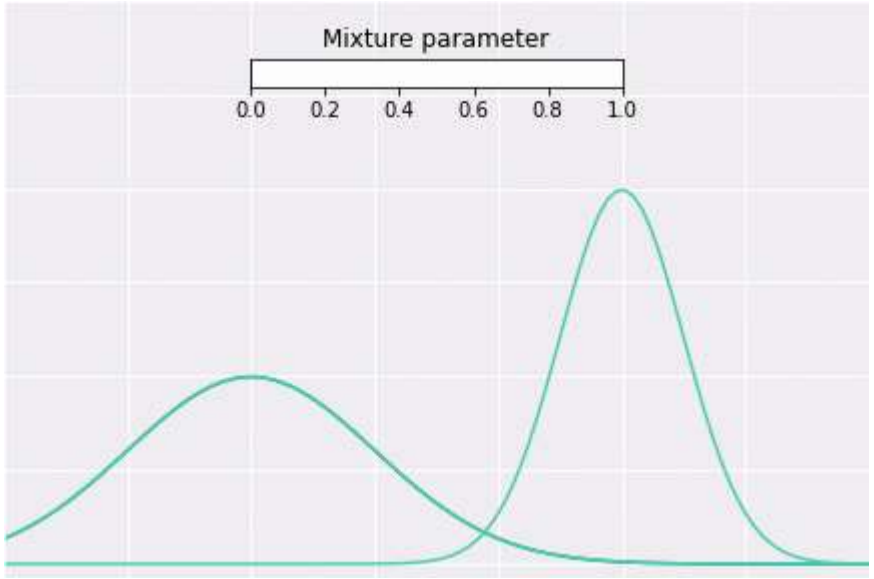
Peso asociado a cada componente gaussiano

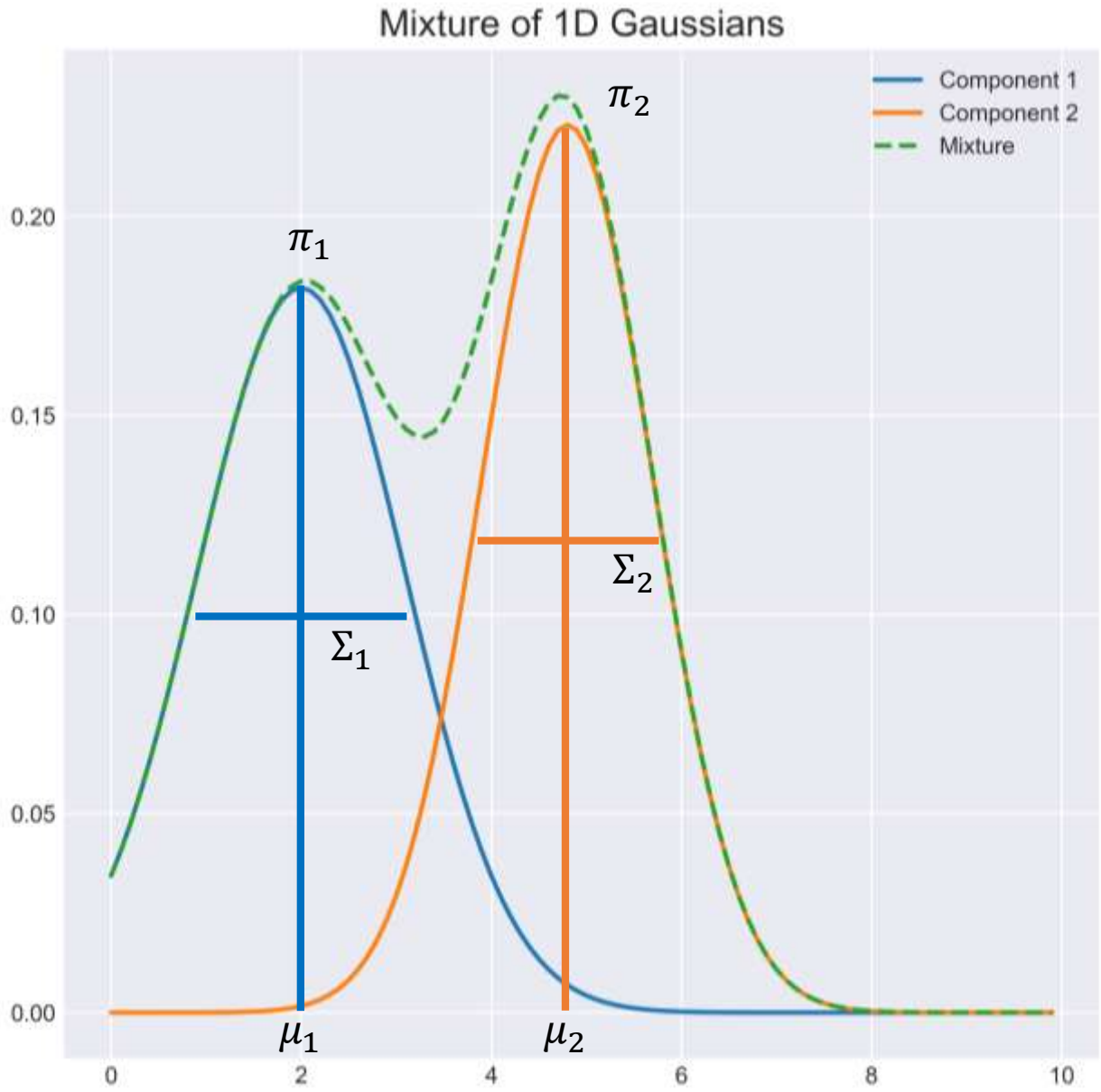
Función de densidad normal con media μ_k y matriz de covarianza Σ_k

La suma de los π_k es igual a 1. Representa la "densidad" de los puntos del cluster K

Gráficamente

https://lukapopijac.github.io/gaussian-mixture-model/





https://medium.com/swlh/gaussian-mixture-models-visually-explained-f54f761d304f

Relación con los modelos K-Means

- En ambos casos se modela la data usando centroides / media

- En clustering no existe un parámetro que indique que tan ajelados están los puntos con respecto al centro.
- En los modelos de mixtura, sí. Se usa la matriz de covarianza

- En clustering, el punto de data se le asigna un solo cluster.
- En modelos de mixtura se asigna a todos los clusters con probabilidades.

¿Como encontramos los parámetros?

- Para cada una de las K distribuciones de probabilidad tenemos que encontrar el vector μ y la matriz de covarianza Σ y el π correspondiente

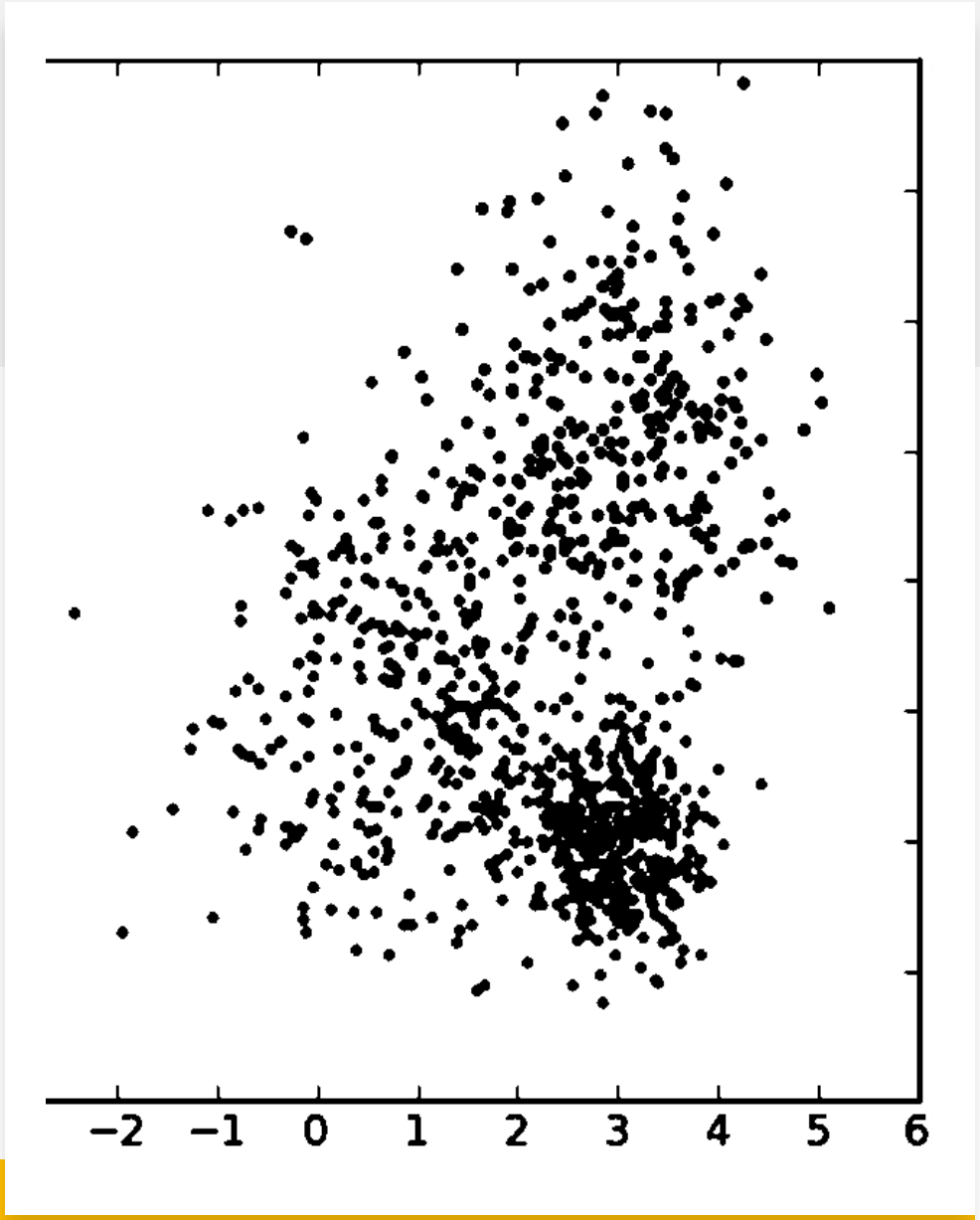
$$\mu_1, \mu_2, \dots, \mu_K$$

$$\Sigma_1, \Sigma_2, \dots, \Sigma_K$$

$$\pi_1, \pi_2, \dots, \pi_K$$

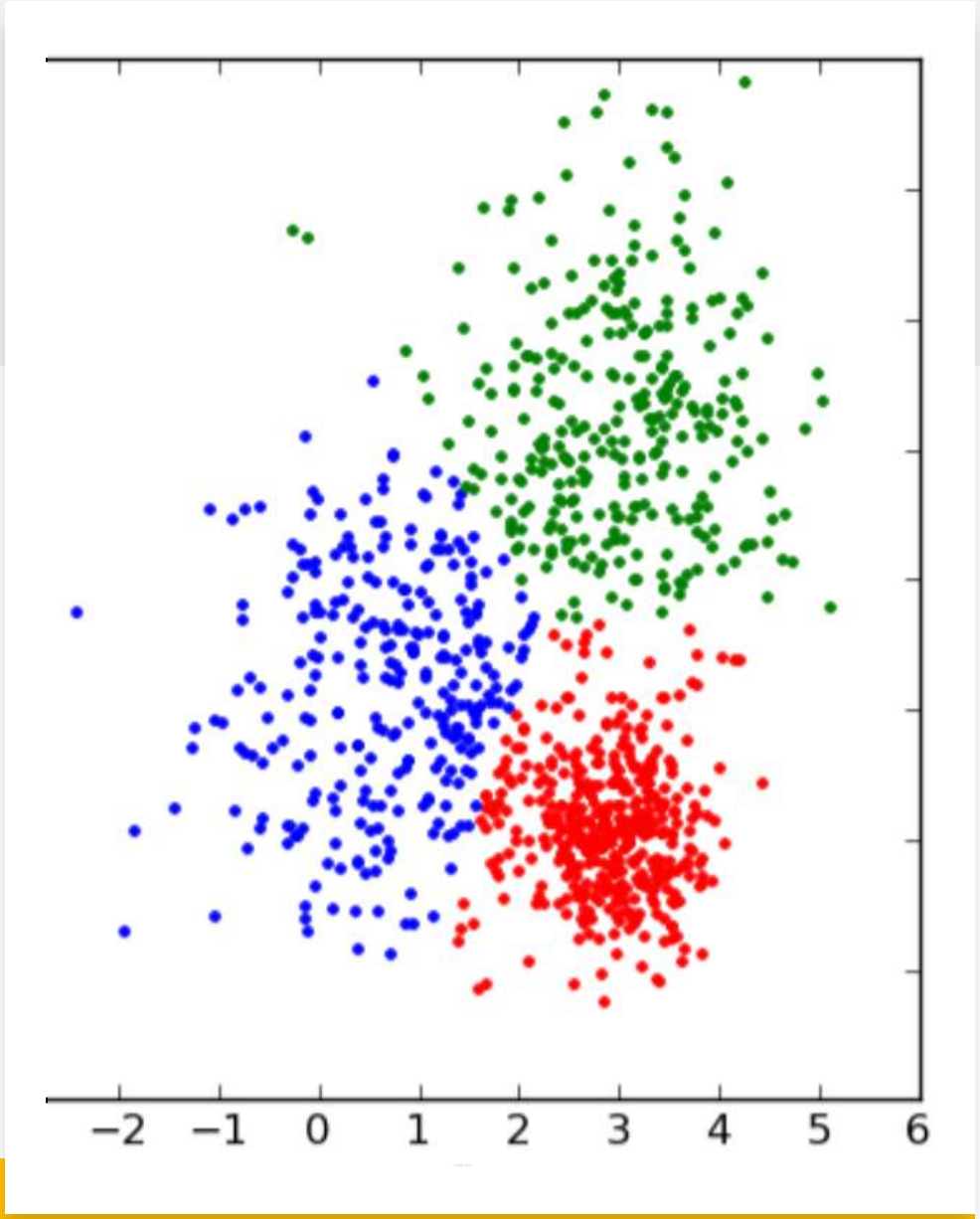
Idealmente

- Si por cada data point supiéramos de qué distribución proviene, sería fácil determinar los parámetros.
  - μ es la media de todos los puntos que pertenecen a la distribución k
  - Σ es la matriz de covarianza de todos los puntos que pertenecen a la distribución k
  - π es la proporción de datapoints que pertenecen a la distribución k
- Pero NO sabemos a qué distribución pertenece cada datapoint



Idealmente

- Si por cada distribución supiéramos μ, Σ y π sabríamos a qué distribución k pertenece cada data point
- Pero NO sabemos los parámetros de las distribuciones



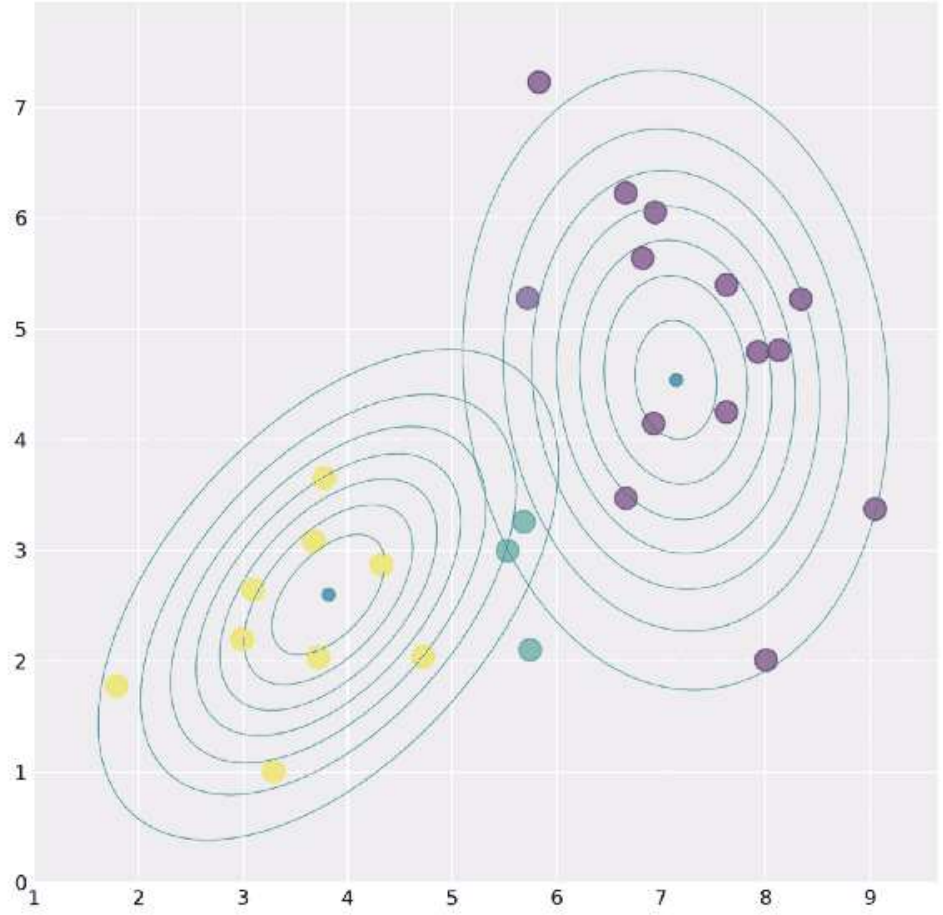
Tenemos que fijar algo

1. Podemos asumir parámetros
   - Aleatoriamente definimos los valores de μ, Σ y π para cada distribución k

2. Vemos que tan bien encajan los datapoints con esos parámetros (la probabilidad de pertenencia a una distribución específica)
   - Dado x_i, qué tan probable es que haya sido originado por k?

3. En base a lo observado (paso 2), redefinimos los modelos para que se ajusten lo mejor posible las observaciones
   - Dado los x_i puntos que pertenecen a k, cuál es la mejor distribución k para que genere dichos puntos? (μ, Σ y π)

Algoritmo Expectation-Maximization



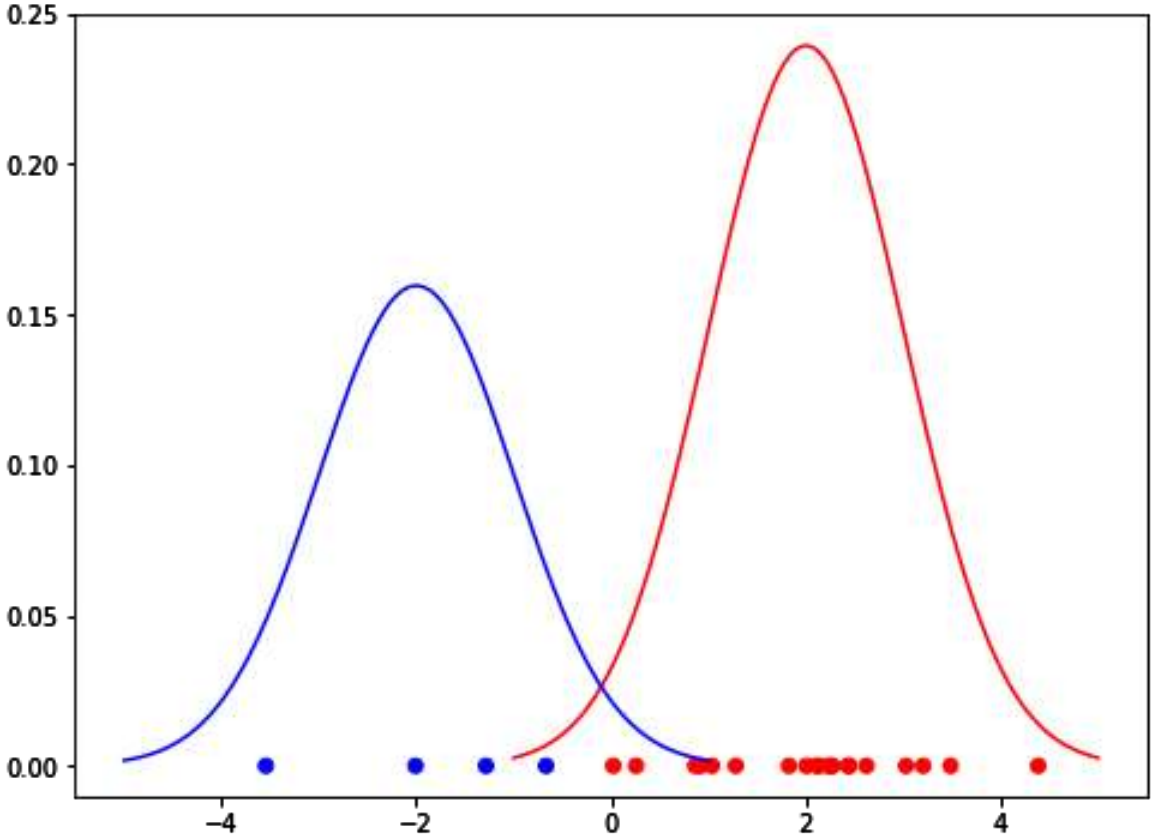
Temas iniciales

- Definir una variable z de tal forma que es la asignación del punto x_i al cluster k
- La variable z es una "variable latente". Explica la diferencia en las distribuciones

- Por lo tanto, $P(z_i = k) = \pi_k$

Probabilidad de que un punto cualquiera sea del cluster k

Porcentaje de elementos del cluster k con respecto al total. La "densidad" de los puntos



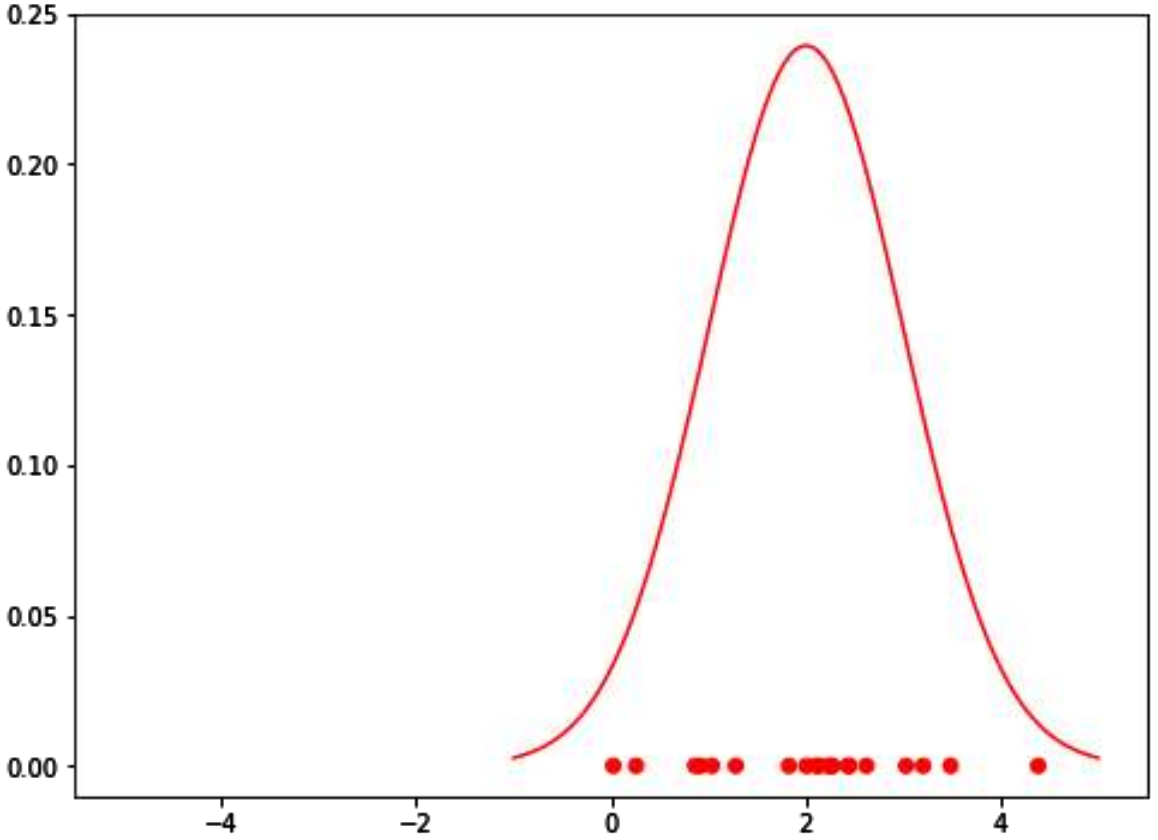
Temas iniciales

Dado a que la observación x_i es del cluster k, qué tan probable es observar x_i

$$P(x_i|z_i = k, \mu_k, \Sigma_k) = \mathcal{N}(x_i|\, \mu_k, \Sigma_k)$$

Probabilidad de observar x_i dado a que pertenece a la distribución k con media μ_k y matriz de covarianza Σ_k

Probabilidad de observar x_i dado una distribución normal con media μ_k y matriz de covarianza Σ_k



Expectation

- Asumimos que las distribuciones son verdad y …
- Determinamos para cada punto la probabilidad de que haya sido generado por k, dado a que "sabemos" los parámetros de k.
- La "responsabilidad" de cada distribución para generar x_i

Teorema de Bayes

$$r_{i,k} = P(z_i = k|x_i) = \frac{P(x_i|z_i = k) * P(z_i = k)}{P(x_i)}$$

Expectation

$\mathcal{N}(x_i|\,\mu_k, \Sigma_k)$ π_k

$$r_{i,k} = P(z_i = k|x_i) = \frac{P(x_i|z_i = k) * P(z_i = k)}{P(x_i)}$$

$$PDF(x) = \sum_{k=1}^{K} \pi_k * \mathcal{N}(x_i|\,\mu_k, \Sigma_k)$$

Probabilidad de que x_i pertenezca a k

$$P(z_i = k|x_i) = \frac{\pi_k * \mathcal{N}(x_i|\,\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k * \mathcal{N}(x_i|\,\mu_k, \Sigma_k)}$$

Probabilidad de x_i

Expectation

$$r_{i,k} = P(z_i = k|x_i) = \frac{\pi_k * \mathcal{N}(x_i|\,\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k * \mathcal{N}(x_i|\,\mu_k, \Sigma_k)}$$

Ejemplo:

$$dato\ 1 = \begin{bmatrix} 0.04 \\ 0.5 \\ 0.46 \end{bmatrix} \qquad dato\ 2 = \begin{bmatrix} 0.80 \\ 0.15 \\ 0.05 \end{bmatrix} \qquad dato\ 2 = \begin{bmatrix} 0.50 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$x_i = \begin{bmatrix} r_{i,1} \\ r_{i,2} \\ r_{i,3} \end{bmatrix}$$

Maximización

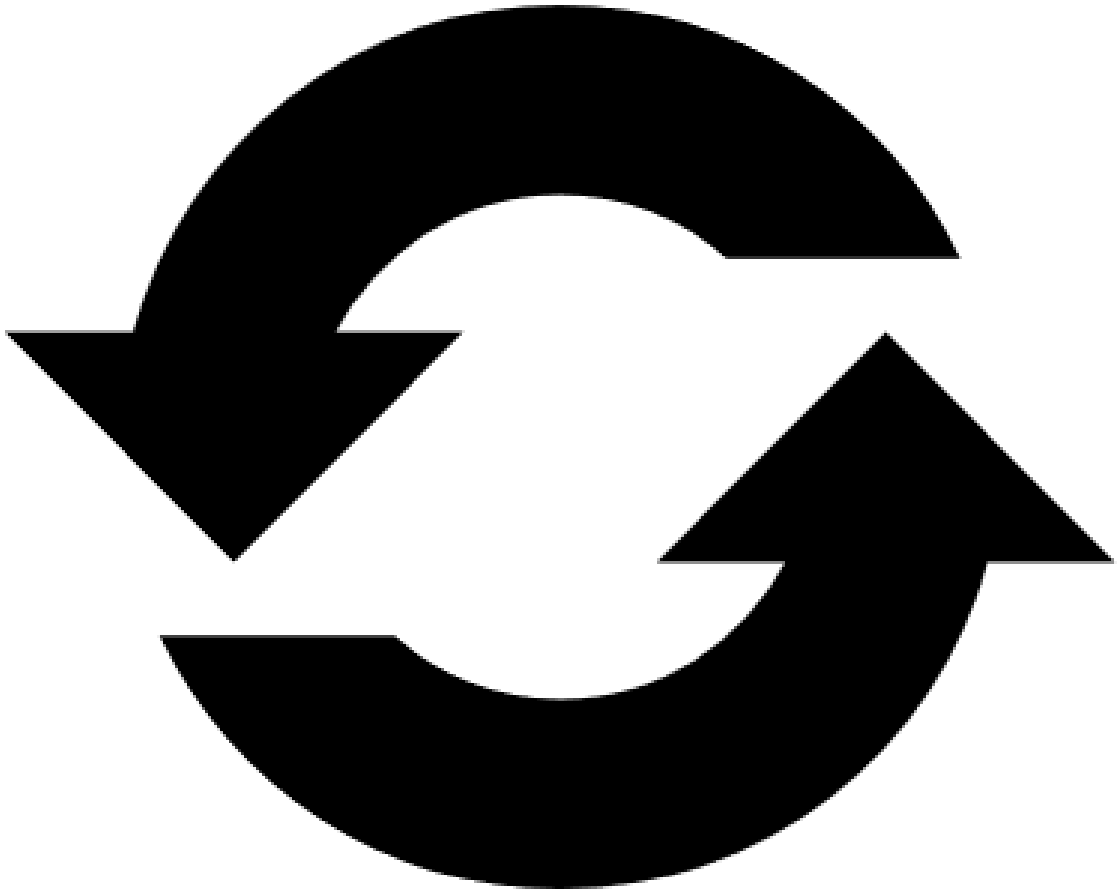
- Asumimos que los puntos son verdad y…
- Redefinimos las distribuciones de tal manera que maximicen la pertenencia de los puntos

$$\pi_k = \frac{puntos\ "asignados"\ al\ cluster\ k}{total\ de\ puntos} = \frac{\sum r_{i,k}}{N}$$

$$\mu_k = \frac{\sum(r_{i,k} * x_i)}{puntos\ "asignados"\ al\ cluster\ k} = \frac{\sum(r_{i,k} * x_i)}{\sum r_{i,k}}$$

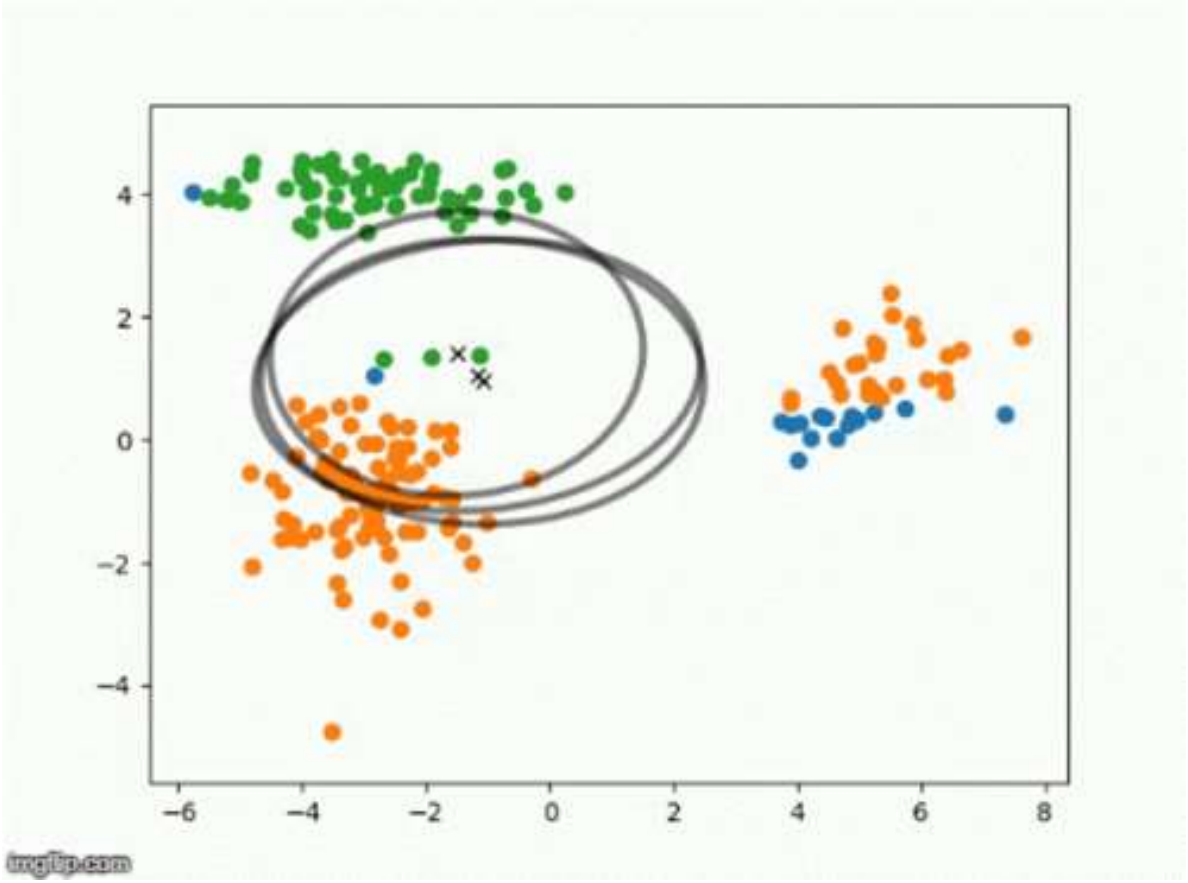
$$\Sigma_k = \frac{\sum(r_{i,k} * (x_i - \mu_k)(x_i - \mu_k)^T)}{puntos\ "asignados"\ al\ cluster\ k} = \frac{\ldots}{\sum r_{i,k}}$$

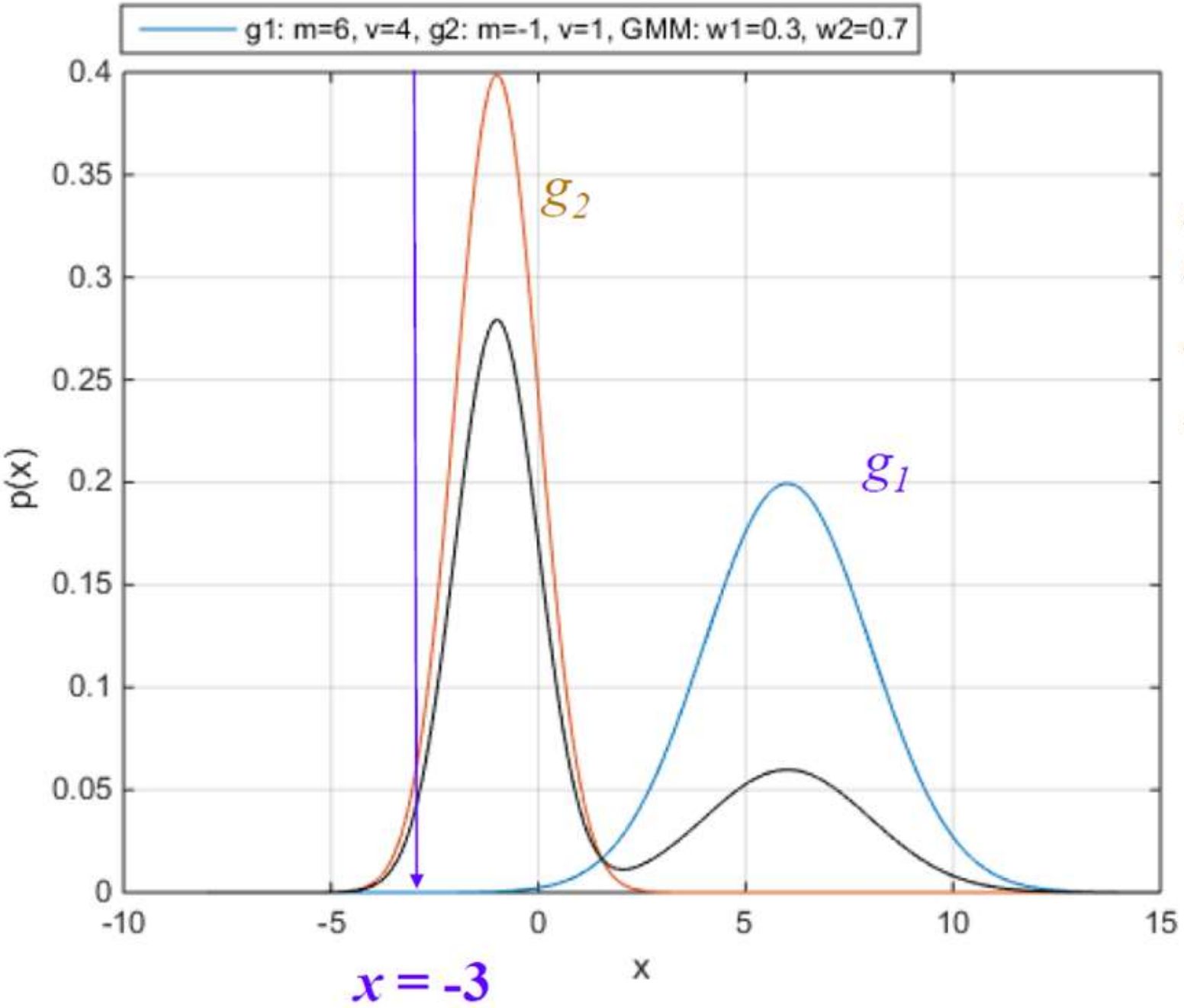
Y repetimos hasta converger



Iterpretación

- Iniciamos asumiendo una hipótesis (inicialización aleatoria)
- Observamos el experimento y vemos que tanto se ajusta a la hipotesis (Expectation)
- En base al experimento, deducimos una nueva hipótesis (Maximization)



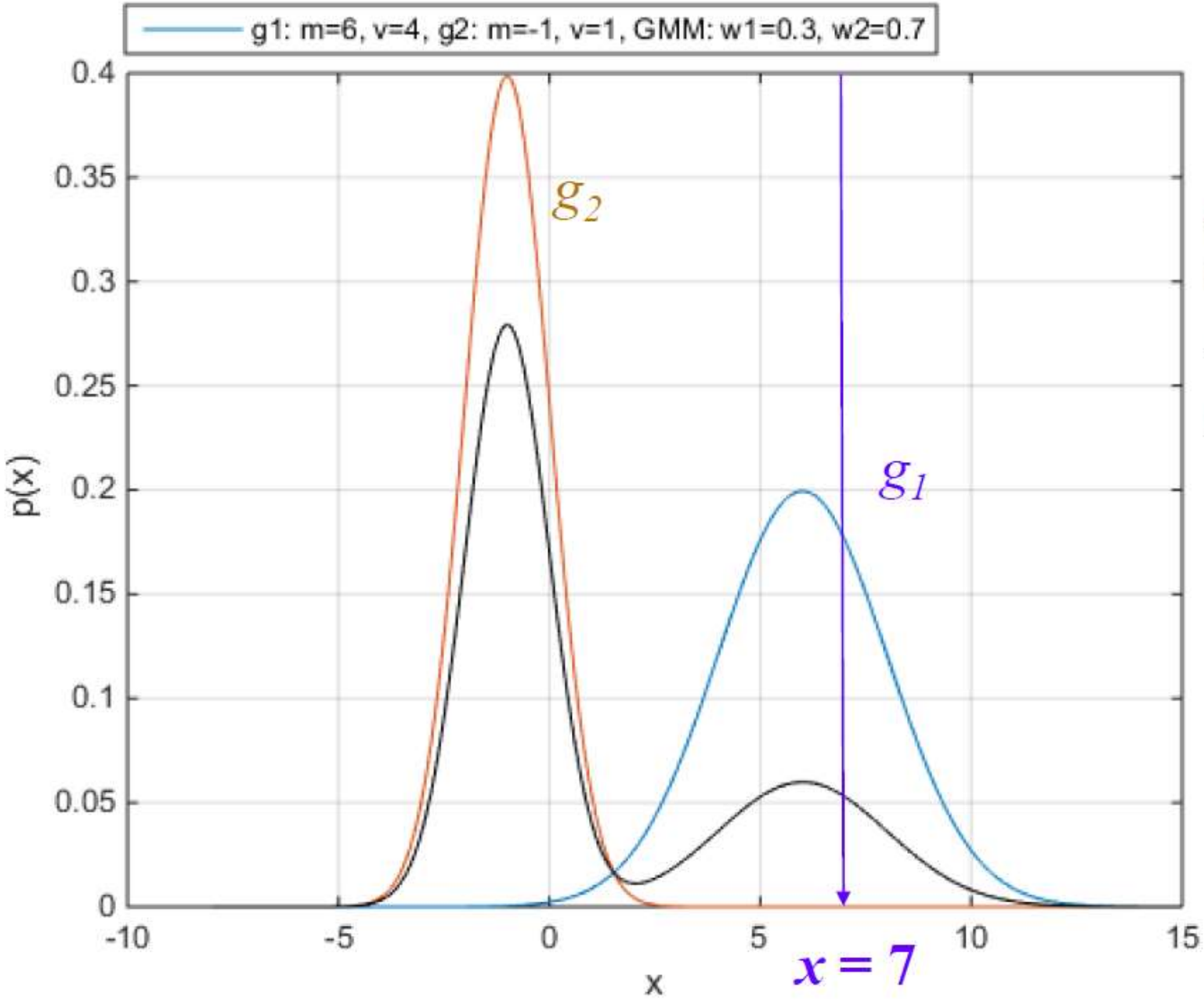


$$g_1(x) \approx 0$$

$$g_2(x) = 0.054$$

$$P(1 \mid x) \approx \frac{0 \times 0.3}{0 \times 0.3 + 0.054 \times 0.7} = 0$$

$$P(2 \mid x) \approx 1$$

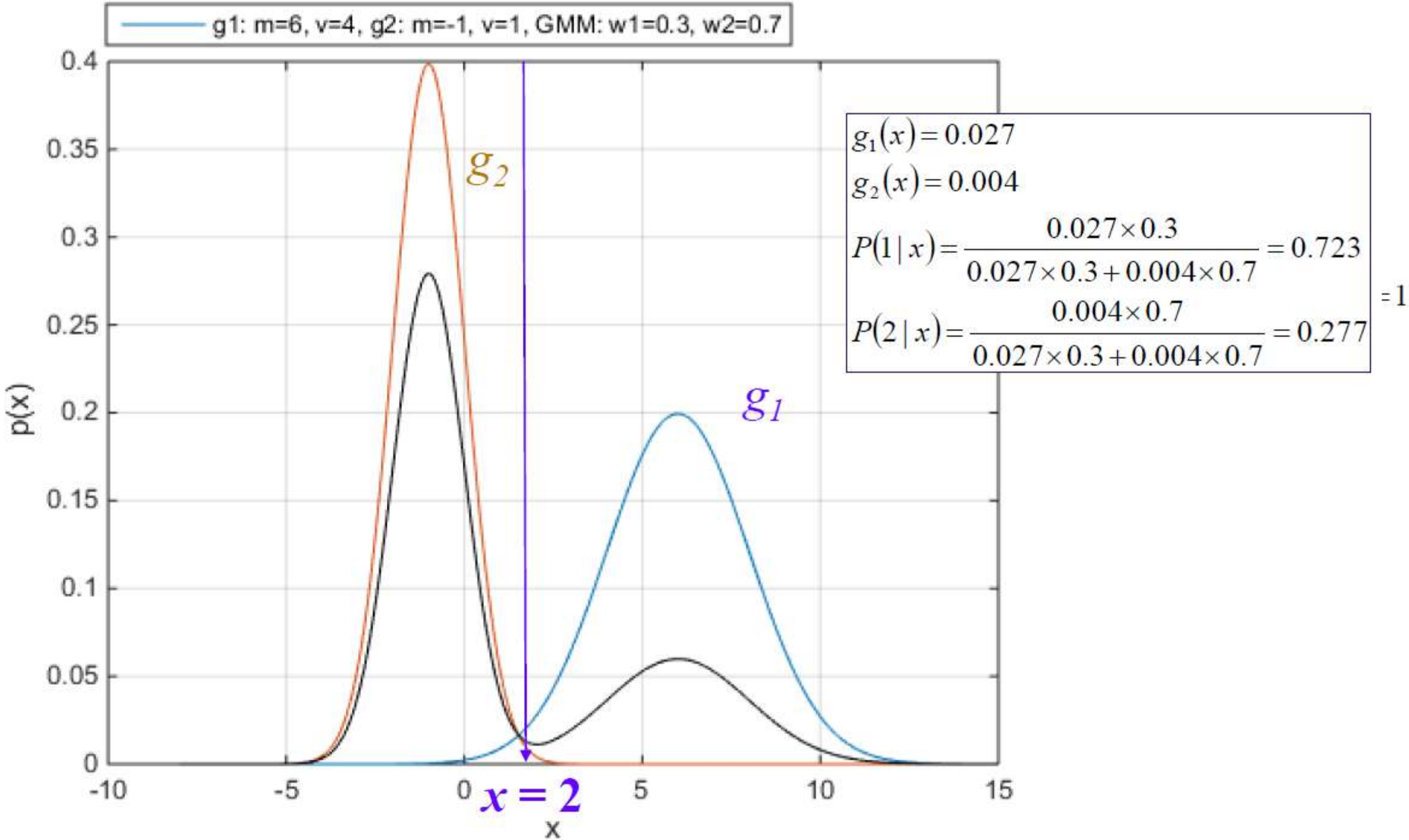


$$g_1(x) = 0.176$$

$$g_2(x) \approx 0$$

$$P(1 \mid x) \approx \frac{0.176 \times 0.3}{0.176 \times 0.3 + 0 \times 0.7} = 1$$

$$P(2 \mid x) \approx 0$$

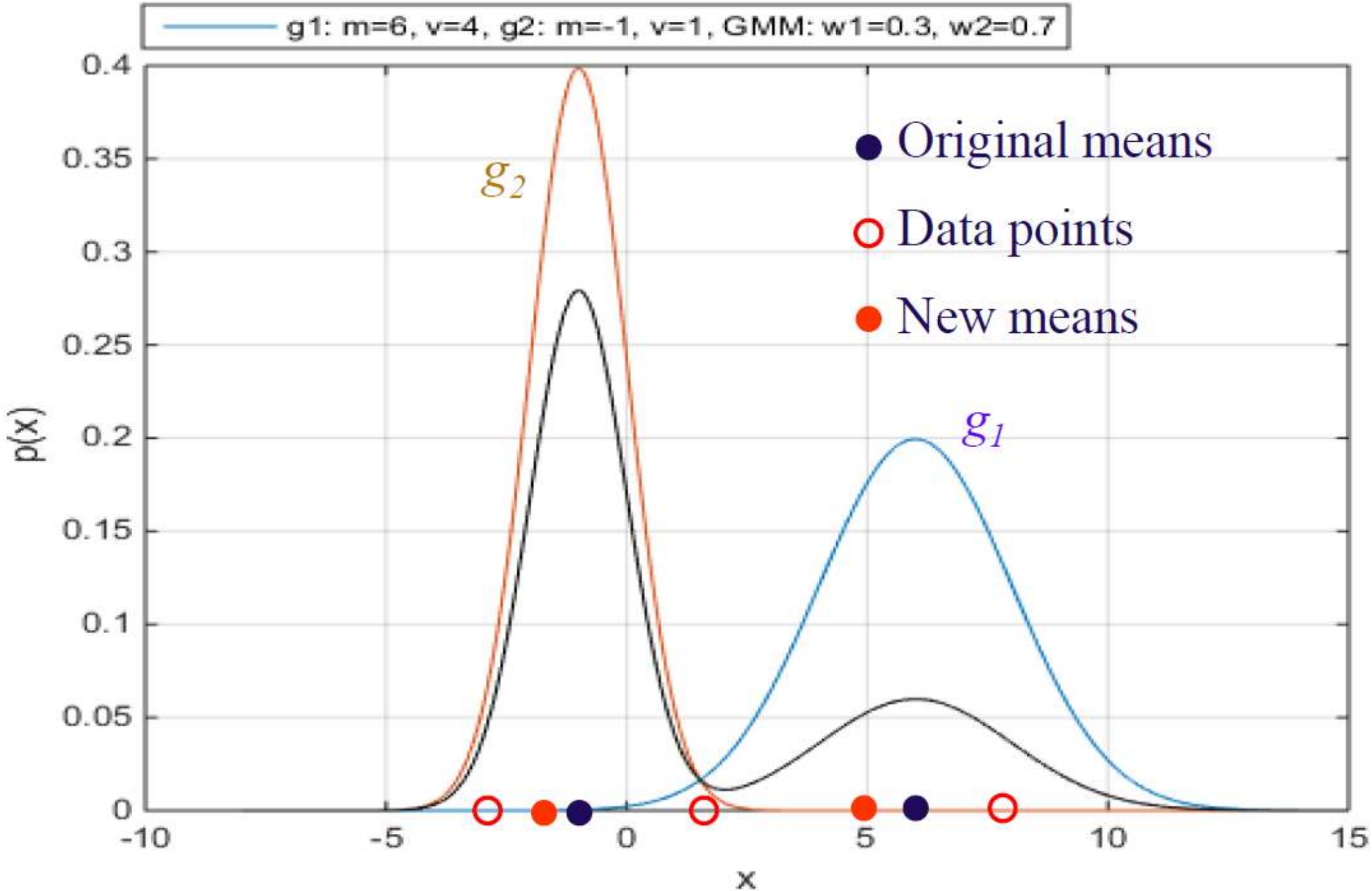

g1: m=6, v=4, g2: m=-1, v=1, GMM: w1=0.3, w2=0.7
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
p(x)
-10
-5
0
5
10
15
x
g_2
g_1
$x = 2$
$g_1(x) = 0.027$
$g_2(x) = 0.004$
$P(1|x) = \frac{0.027 \times 0.3}{0.027 \times 0.3 + 0.004 \times 0.7} = 0.723$
$P(2|x) = \frac{0.004 \times 0.7}{0.027 \times 0.3 + 0.004 \times 0.7} = 0.277$

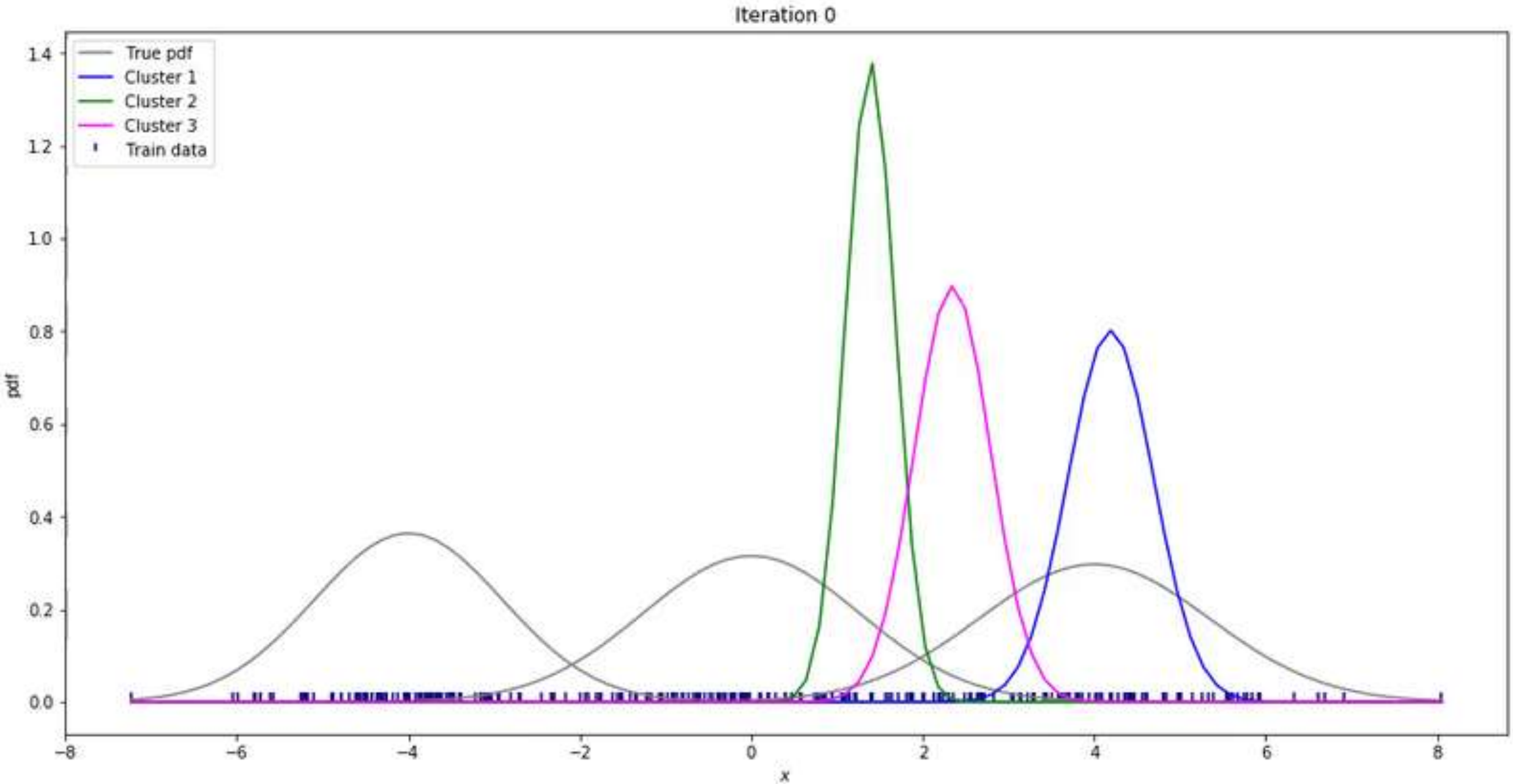
- Dado los datapoints $X = \{-3,2,7\}$

- Las nuevas medias son

$$\overline{m}_1 = \frac{0 \times x_1 + 0.723 \times x_2 + 1 \times x_3}{0 + 0.723 + 1} = \frac{0 \times (-3) + 0.723 \times 2 + 1 \times 7}{1.723} = 4.9$$

$$\overline{m}_2 = \frac{1 \times x_1 + 0.277 \times x_2 + 0 \times x_3}{1 + 0.277 + 0} = \frac{1 \times (-3) + 0.277 \times 2 + 0 \times 7}{1.277} = -1.92$$


g1: m=6, v=4, g2: m=-1, v=1, GMM: w1=0.3, w2=0.7
Original means
Data points
New means
g_2
g_1
p(x)
x
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-10
-5
0
5
10
15


Iteration 0
True pdf
Cluster 1
Cluster 2
Cluster 3
Train data
pdf
x
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−8
−6
−4
−2
0
2
4
6
8

En Resumen

- GMM encuentra K distribuciones gaussianas que se ajustan a los datos.
  - Cada una con una media, covarianza y responsabilidad respectiva.
- Usa el algoritmo E-M para hallarlo
  - Asume una distribución
  - Determina la participación (responsabilidad) de cada distribución para los puntos
  - Ajusta las distribuciones para que refleje dichas participaciones.
  - Repetir hasta converger