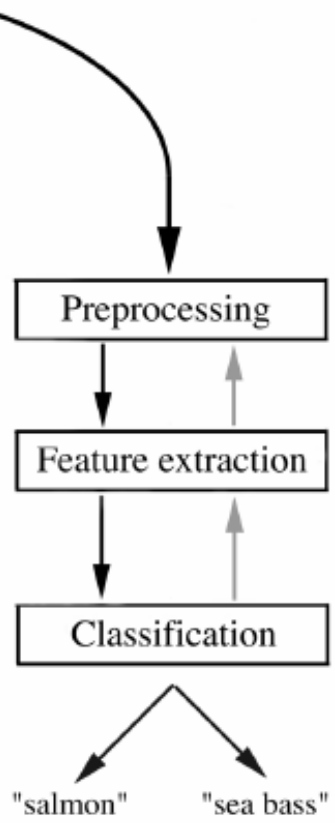
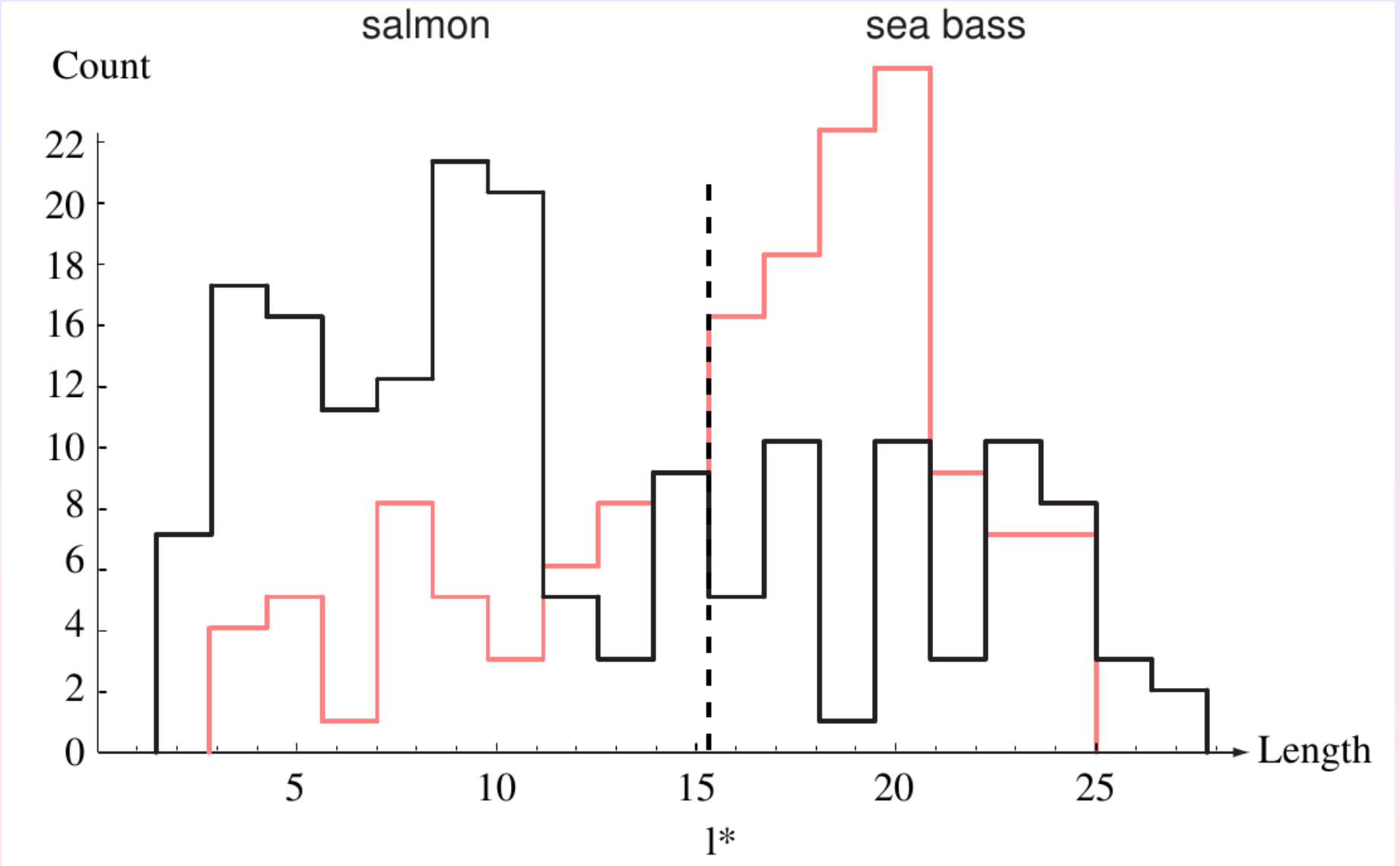
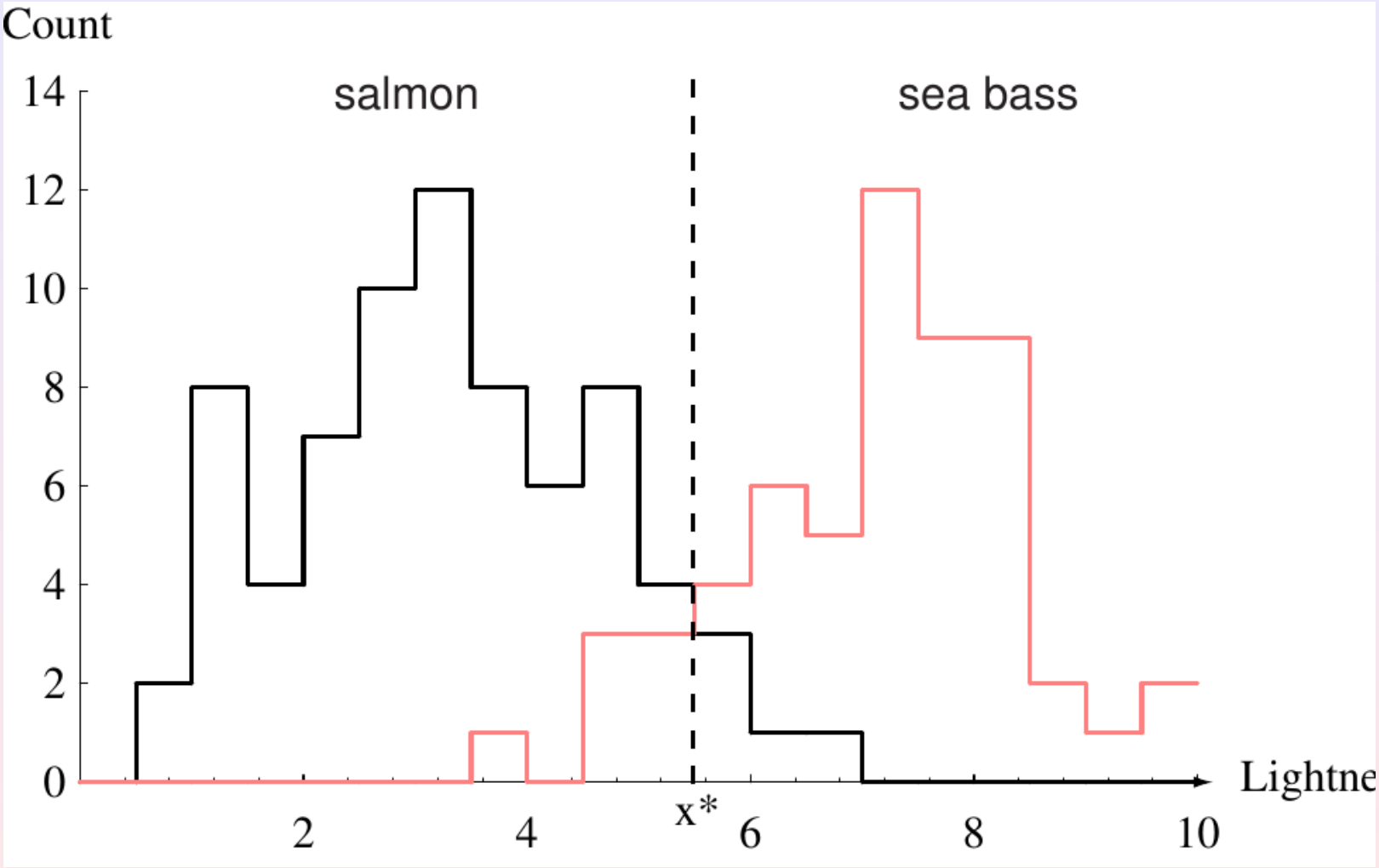
Ejemplo


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

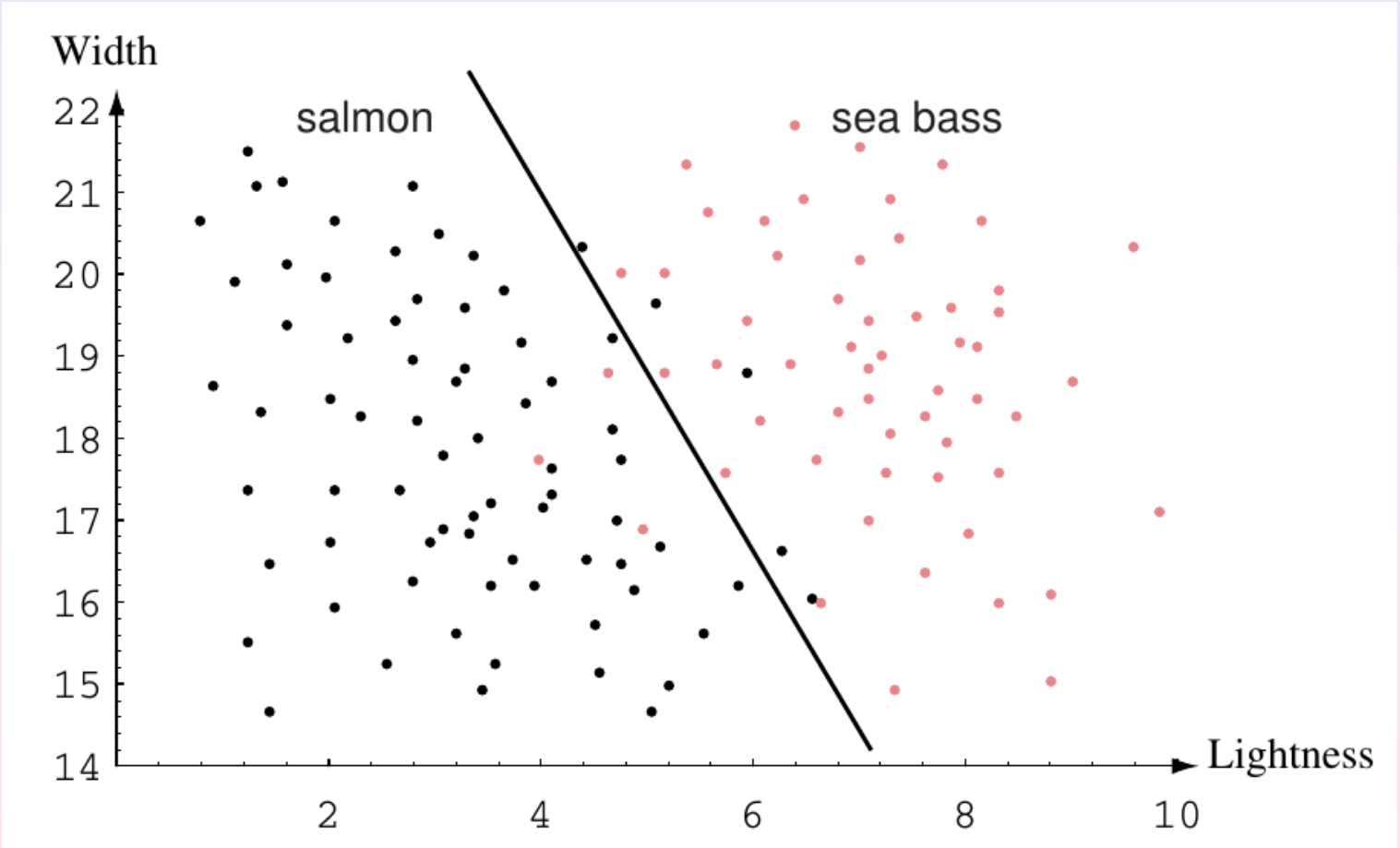
Ejemplo


salmon
sea bass
Count
22
20
18
16
12
10
8
6
4
2
0
5
10
15
20
25
Length
l*

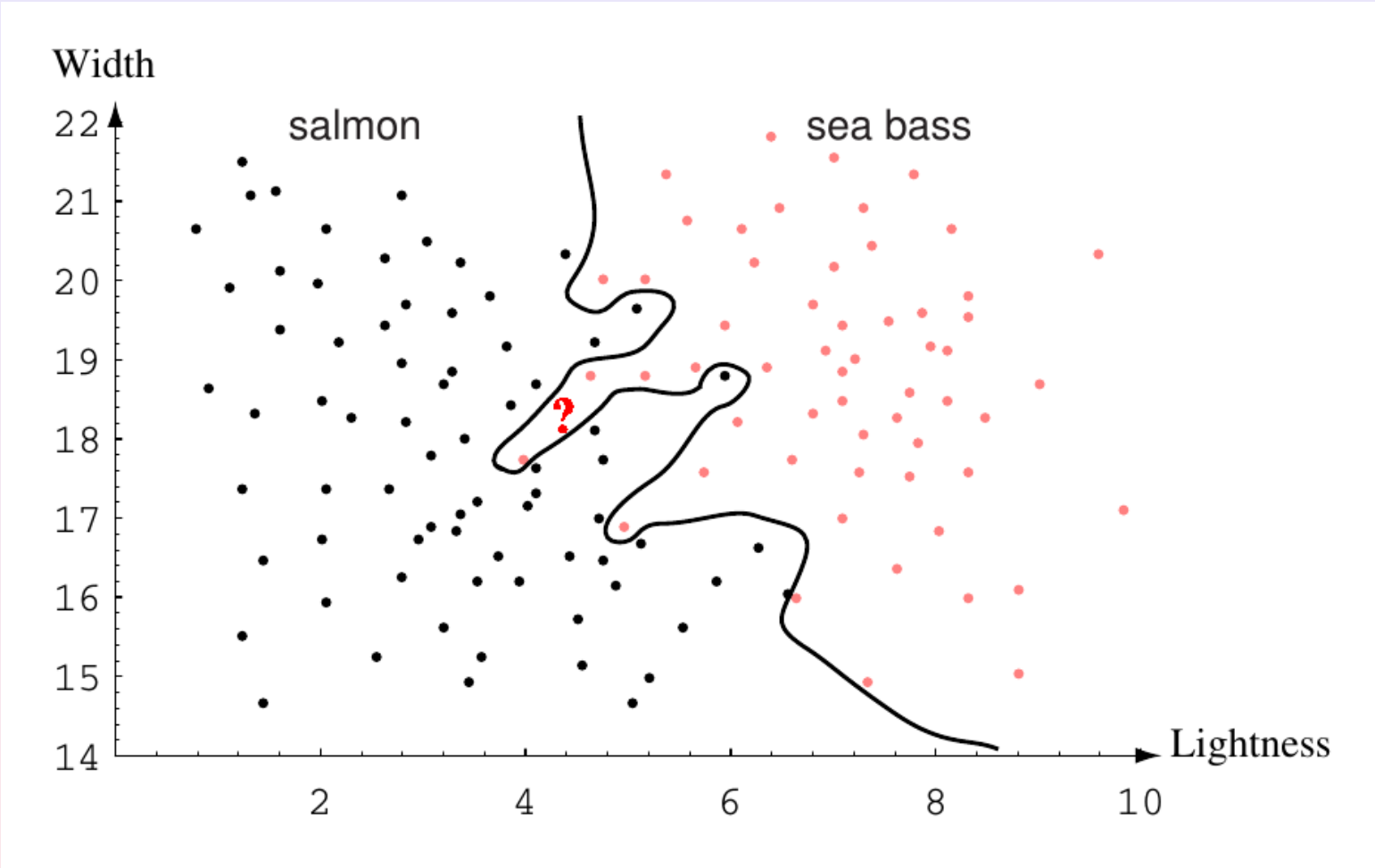
Ejemplo



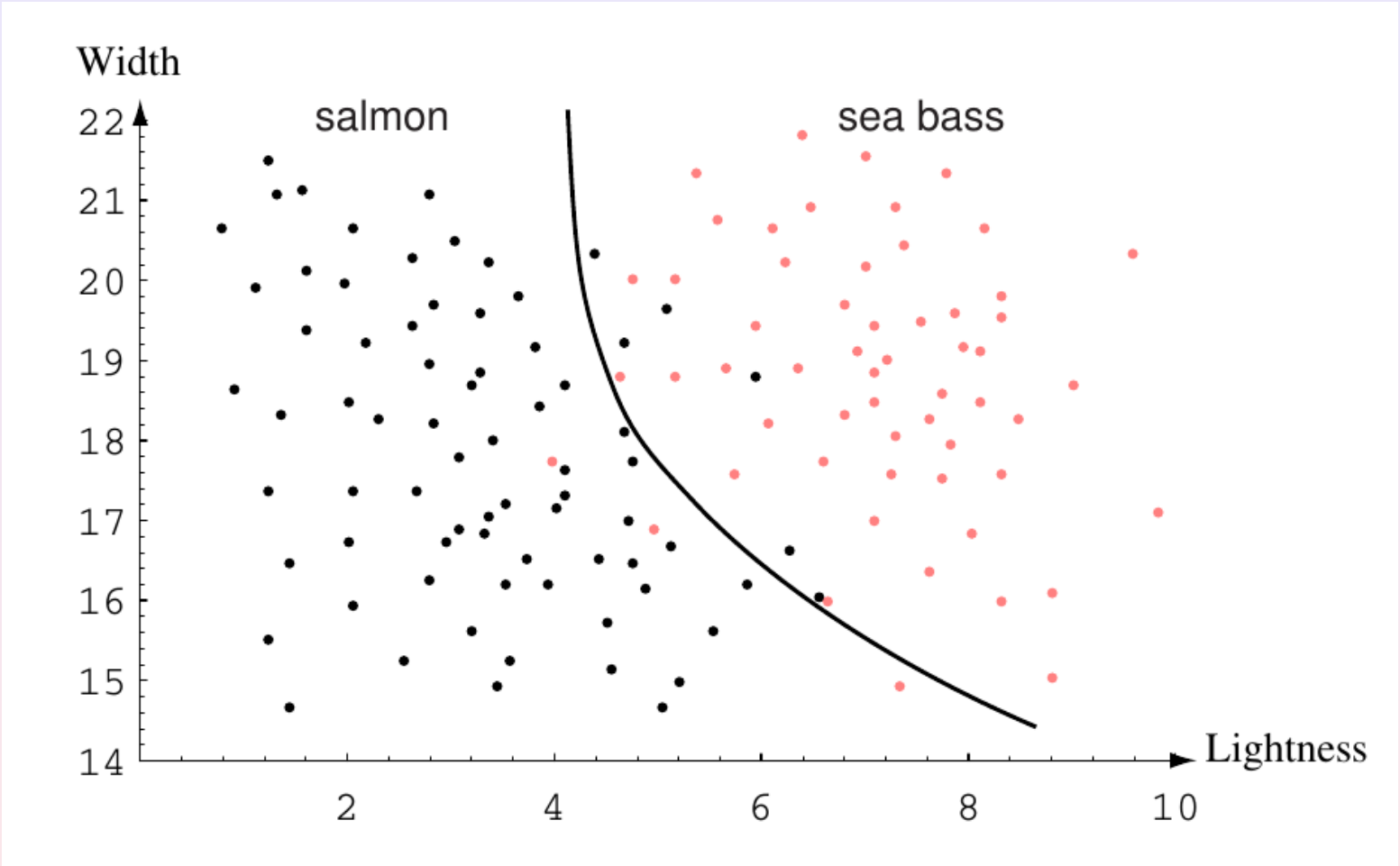
Ejemplo


Width
22
21
20
19
18
17
16
15
14
salmon
sea bass
2
4
6
8
10
Lightness

Ejemplo


Width
salmon
sea bass
?
Lightness
22
21
20
19
18
17
16
15
14
2
4
6
8
10

Ejemplo


Width
salmon
sea bass
22
21
20
19
18
17
16
15
14
2
4
6
8
10
Lightness

Ejemplo

(a) (b) (c)

Figure 1.3 Three types of iris flowers: setosa, versicolor and virginica. Source: `http://www.statlab.uni-heidelberg.de/data/iris/` . Used with kind permission of Dennis Kramb and SIGNA.

Ejemplo

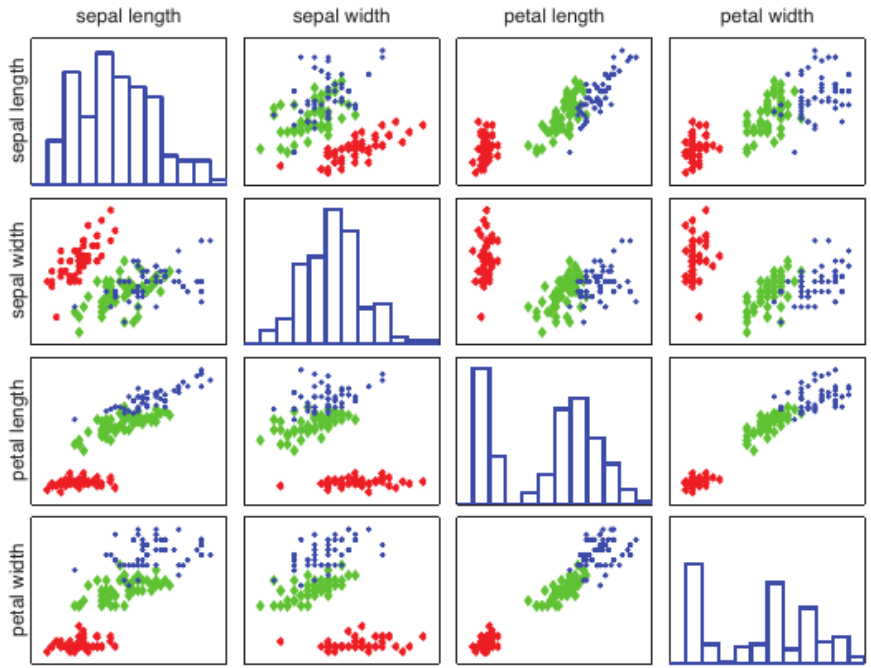


Figure 1.4 Visualization of the Iris data as a pairwise scatter plot. The diagonal plots the marginal histograms of the 4 features. The off diagonals contain scatterplots of all possible pairs of features. Red circle = setosa, green diamond = versicolor, blue star = virginica. Figure generated by `fisheririsDemo`.

Clasificación Bayesiana

Se trata de clasificar objetos en c clases: $\omega_1, \ldots, \omega_c$.
Existe una probabilidad *a priori* para cada una de estas clases:

$$P(\omega_1), \ldots, P(\omega_c)$$

De manera tal que

$$\sum_{i=1}^{c} P(\omega_i) = 1$$

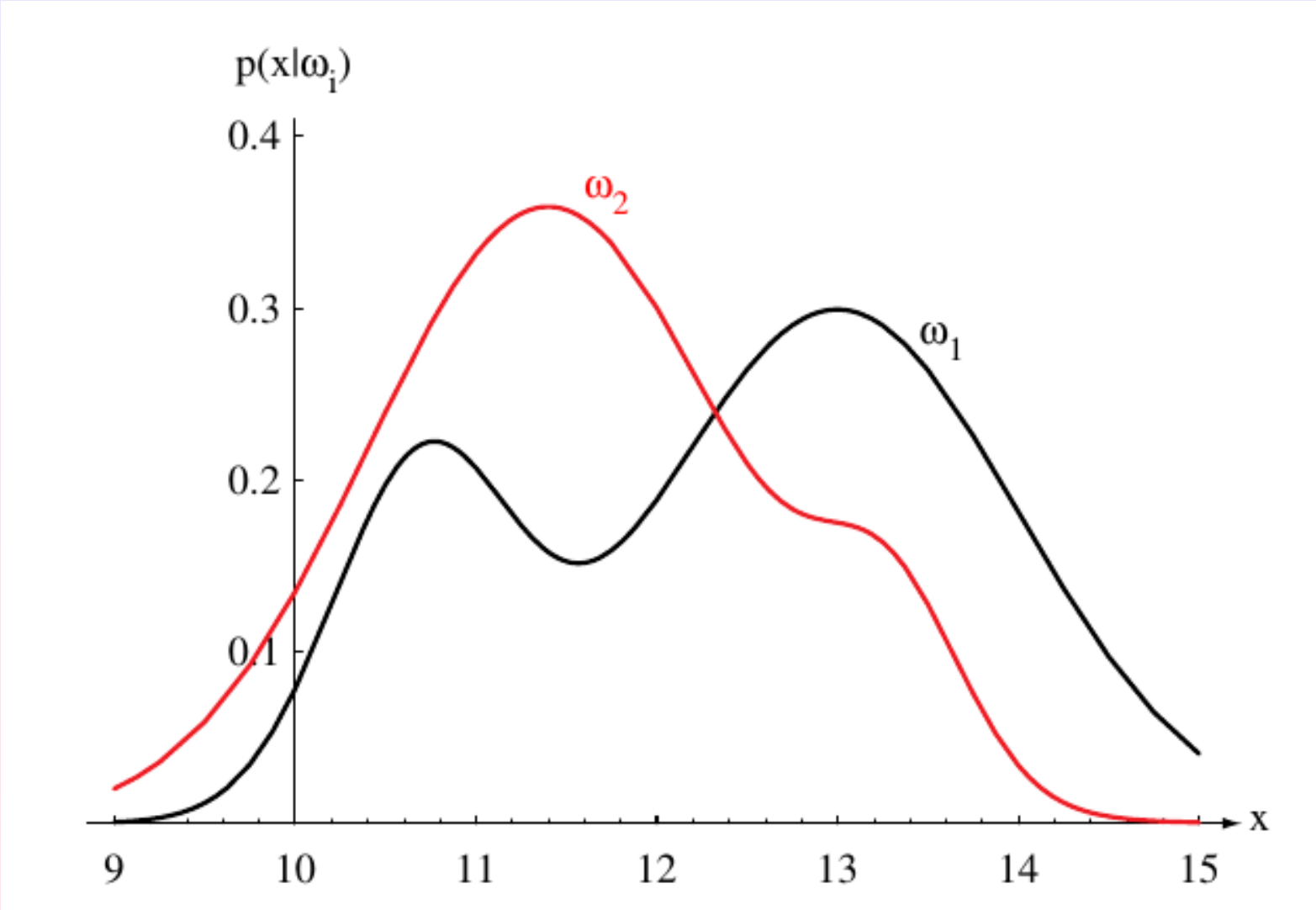
Cuál sería la regla de decisión que deberíamos adoptar para clasificar un objeto dado, de manera de minimizar el error de clasificación?
Claramente: elegir alguna ω_j tal que

$$P(\omega_j) \geq P(\omega_i), \forall i \in \{1, \ldots, c\}$$

Clasificación Bayesiana

Si contamos con información adicional para clasificar los objetos, como el valor medio de gris, o alguna dimensión de los mismos, podríamos mejorar nuestra decisión
Identifiquemos esta información adicional con la variable aleatoria x, para el caso unidimensional, o con el vector aleatorio $\mathbf{x}$ para el caso multidimensional.
Y llamemos $p(x \mid \omega)$, (o $p(\mathbf{x} \mid \omega)$, según el caso) a la densidad de probabilidad para la clase ω.

Clasificación Bayesiana



Clasificación Bayesiana

Ahora, deberíamos modificar la regla de decisión anterior para tener en cuenta esta información adicional.
Propongamos entonces clasificar de acuerdo con la siguiente regla: elegir alguna ω_j tal que

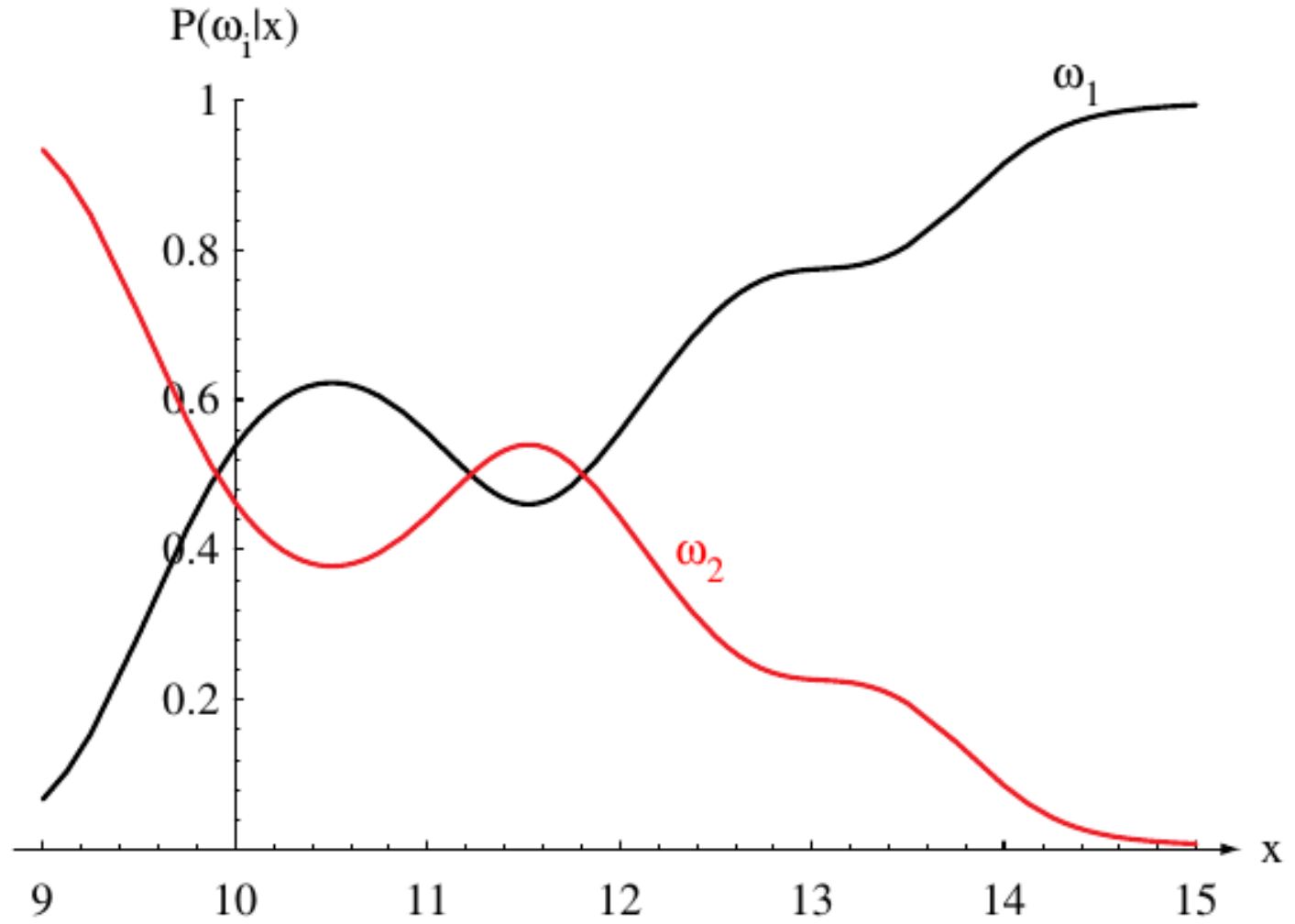
$$P(\omega_j \mid x) \geq P(\omega_i \mid x), \forall i \in \{1, \ldots, c\}$$

Para el caso de tener dos clases: Elegir ω_1 si

$$P(\omega_1 \mid x) \geq P(\omega_2 \mid x),$$

y, si no, elegir ω_2.

Clasificación Bayesiana



Clasificación Bayesiana

Para poder utilizar este criterio debemos utilizar la regla de Bayes

$$P(\omega_i \mid x) = \frac{p(x \mid \omega_i) P(\omega_i)}{p(x)}, \forall i \in \{1, \ldots, c\},$$

o sea,

$$\textit{a posteriori} = \frac{\textit{verosimilitud} \times \textit{a priori}}{\textit{evidencia}},$$

donde

$$p(x) = \sum_{i=1}^{c} p(x, \omega_i) = \sum_{i=1}^{c} p(x \mid \omega_i) P(\omega_i)$$

Clasificación Bayesiana

Pero, como $p(x)$ no es función de ω_i, entonces el criterio anterior ($P(\omega_j \mid x) \geq P(\omega_i \mid x)$) puede expresarse como: elegir ω_j tal que

$$p(x \mid \omega_j)P(\omega_j) \geq p(x \mid \omega_i)P(\omega_i), \forall i \in \{1, \ldots, c\},$$

Distribución Gaussiana
Caso unidimensional

La densidad de la distribución Gaussiana unidimensional está dada por la fórmula

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

donde el valor medio y la varianza, están dados por $\mu = E[x]$ y $\sigma^2 = E[(x-\mu)^2]$, respectivamente.

La densidad de la distribución Gaussiana multidimensional está dada por la fórmula

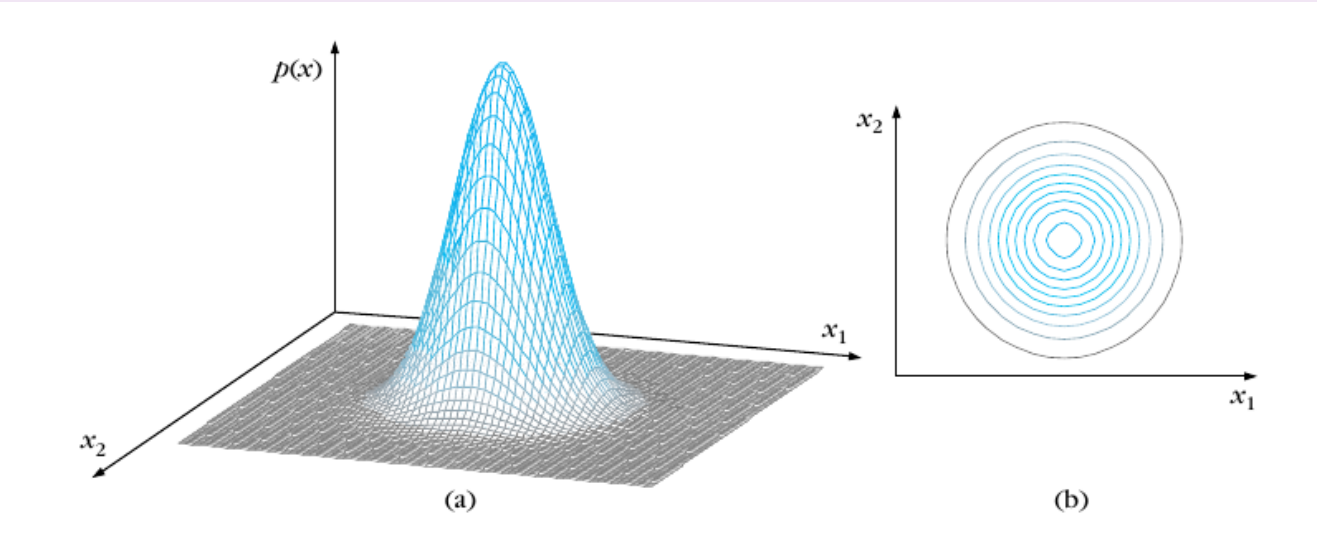
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

donde el vector aleatorio $\mathbf{x}$ es de dimensión d, al igual que el valor medio $\mu = E[\mathbf{x}]$, y $\mathbf{\Sigma} = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^t]$ es la matriz de covarianza de $d \times d$. Caso bidimensional:

$$\mathbf{\Sigma} = E\left[\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}(x_1-\mu_1 \;\; x_2-\mu_2)\right] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

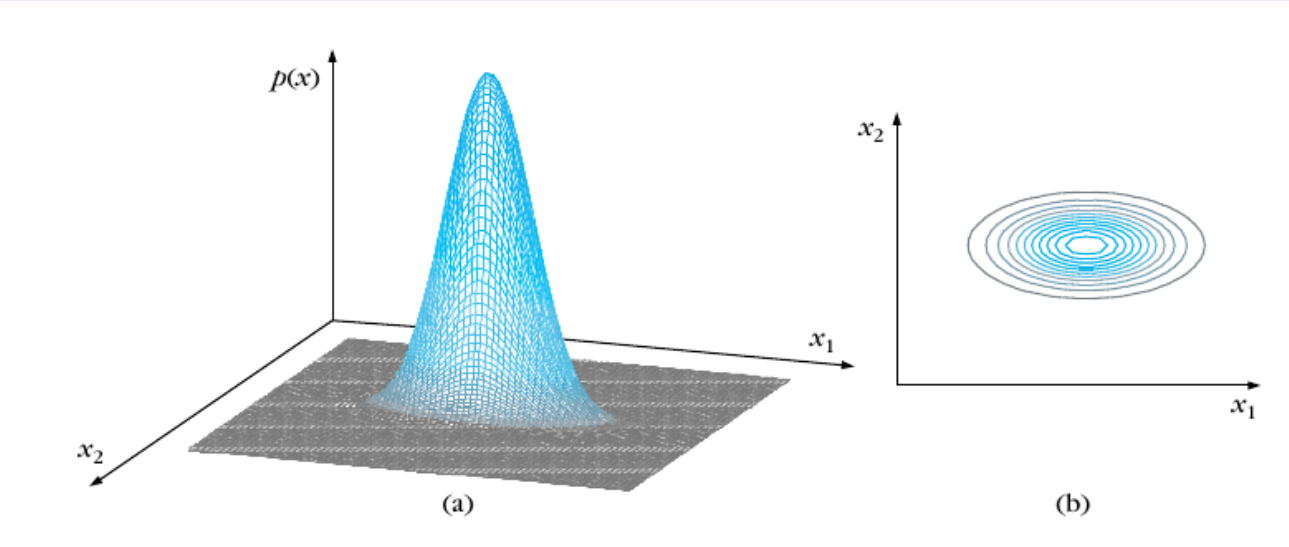
Distribución Gaussiana
Caso multidimensional

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}$$

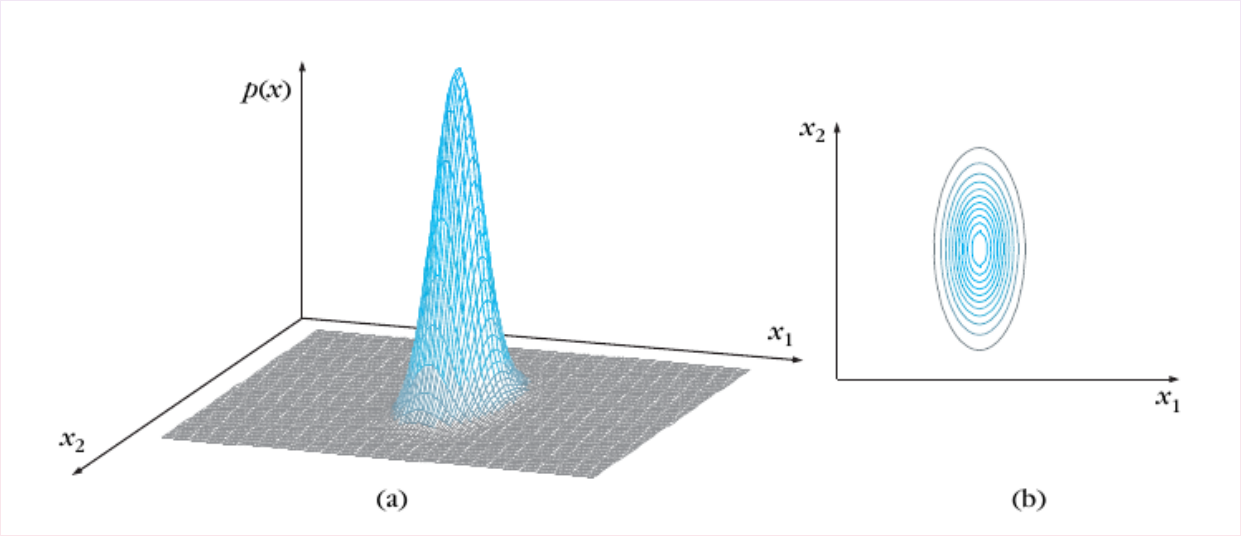


Distribución Gaussiana
Caso multidimensional

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \quad \sigma_1^2 = 15 > \sigma_2^2 = 3$$

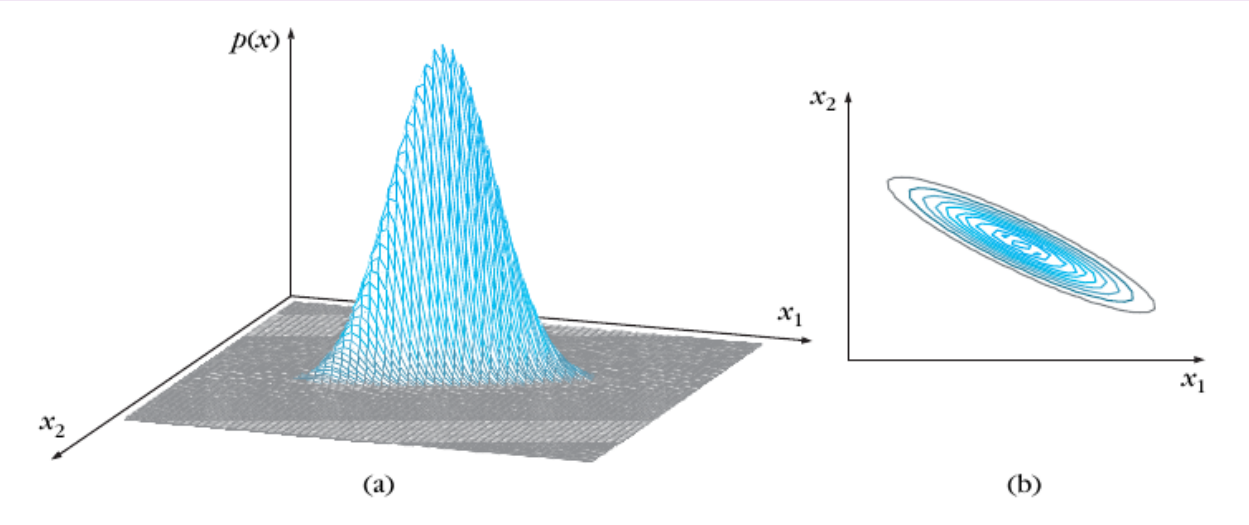


$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \quad \sigma_1^2 = 3 < \sigma_2^2 = 15$$



Distribución Gaussiana
Caso multidimensional

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad \sigma_1^2 = 15 > \sigma_2^2 = 3,\ \sigma_{12} = 6$$



Las curvas de nivel son elipses.Para el caso bidimensional con $\sigma_{12} = \sigma_{21} = 0$ tenemos:

$$\mathbf{x}^t \Sigma^{-1} \mathbf{x} = (x_1 \; x_2) \begin{pmatrix} 1/\sigma_1^2 & 0 \\ 0 & 1/\sigma_2^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = cte.$$

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} = cte.$$

Clasificación Bayesiana: caso Gaussiano

Una herramienta para clasificación Bayesiana es la función discriminante:

$$g_i(\mathbf{x}) = \ln\left(p(\mathbf{x} \mid \omega_i)P(\omega_i)\right) = \ln\left(p(\mathbf{x} \mid \omega_i)\right) + \ln\left(P(\omega_i)\right),$$

para $i = 1, \ldots, c$.
Para el caso Gaussiano tenemos:

$$g_i(\mathbf{x}) = -\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu_i}\right)^t \boldsymbol{\Sigma_i}^{-1}\left(\mathbf{x} - \boldsymbol{\mu_i}\right) + \ln P(\omega_i) + c_i,$$

con $c_i = -(d/s)\ln(2\pi) - (1/2)ln|\Sigma_i|$ y d la dimensión del espacio.

Clasificación Bayesiana: caso Gaussiano

Caso $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$

En este caso, $|\mathbf{\Sigma}_i| = \sigma^{2d}$ y $|\mathbf{\Sigma}_i^{-1}| = 1/\sigma^{2d}$.Como las $\mathbf{\Sigma}_i$ son las mismas para todas las clases, la constante c_i puede ser ignorada.
Tomando $\|.\|$ como la norma Euclídea, tenemos $\|\mathbf{x} - \boldsymbol{\mu_i}\| = (\mathbf{x} - \boldsymbol{\mu_i})^t (\mathbf{x} - \boldsymbol{\mu_i})$, entonces $g_i(\mathbf{x})$ queda:

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu_i}\|}{2\sigma^2} + \ln P(\omega_i).$$

Expandiendo $\|\mathbf{x} - \boldsymbol{\mu_i}\|$ la fórmula anterior queda

$$g_i(\mathbf{x}) = -\frac{-1}{2\sigma^2} \left(\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu_i}^t\mathbf{x} + \boldsymbol{\mu_i}^t\boldsymbol{\mu_i}\right) + \ln P(\omega_i).$$

Como $\mathbf{x}^t\mathbf{x}$ no depende de la clase, entonces la anterior posee la forma

$$g_i(\mathbf{x}) = \mathbf{w_i}^t\mathbf{x} + w_{i0},$$

donde $\mathbf{w_i} = \boldsymbol{\mu_i}/\sigma^2$ y

$$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu_i}^t\boldsymbol{\mu_i} + \ln P(w_i).$$

Las superficies de decisión se encuentran mediante $g_i(\mathbf{x}) = g_j(\mathbf{x})$. Para el caso de dos clases tendremos una ecuación de la forma

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0,$$

Donde

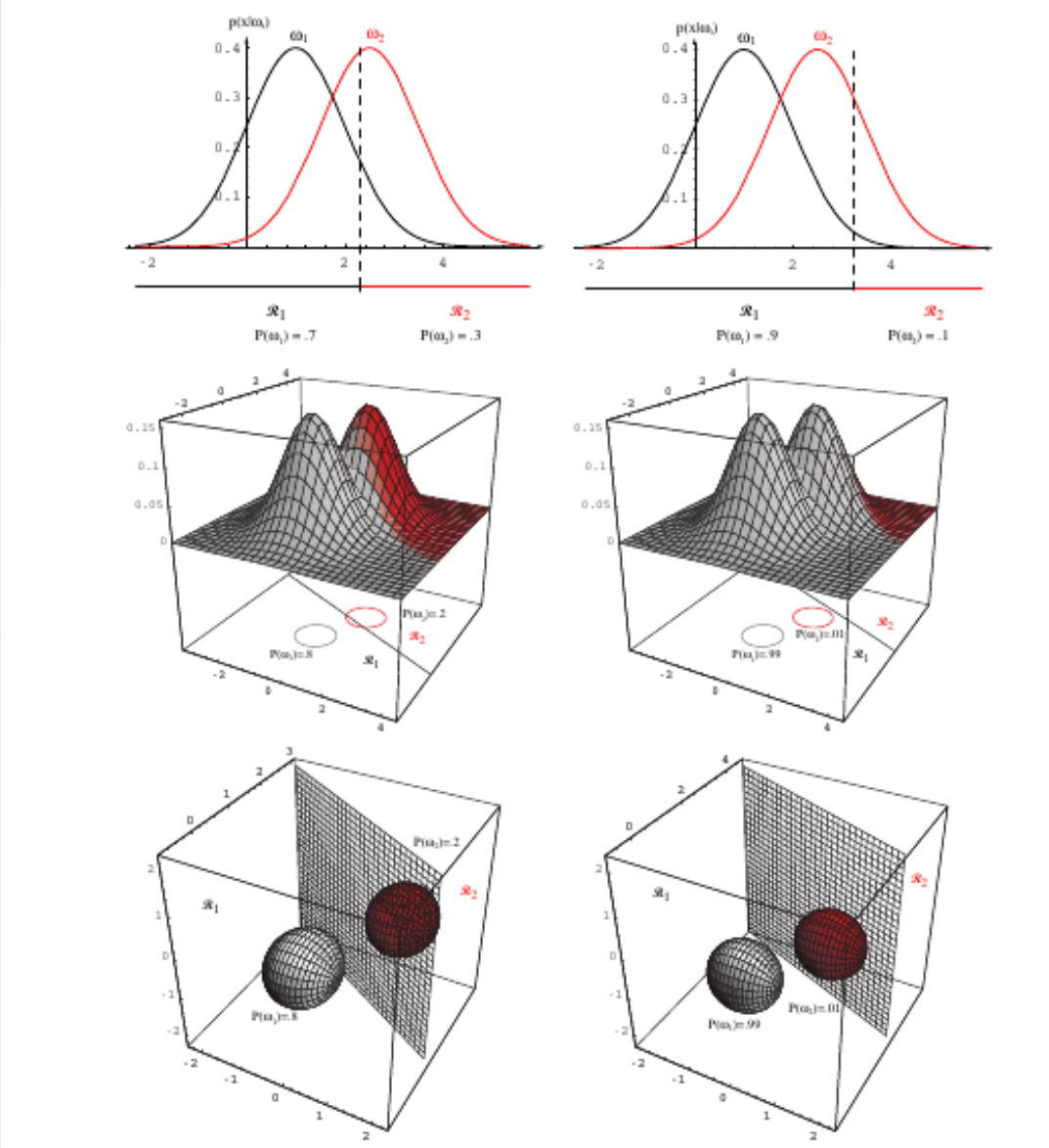
$$\mathbf{w} = \boldsymbol{\mu_i} - \boldsymbol{\mu_j},$$

y

$$\mathbf{x_0} = \frac{1}{2}\left(\boldsymbol{\mu_i} + \boldsymbol{\mu_j}\right) - \frac{\sigma^2}{\|\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\|} \ln \frac{P(w_i)}{P(w_j)} \left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right).$$

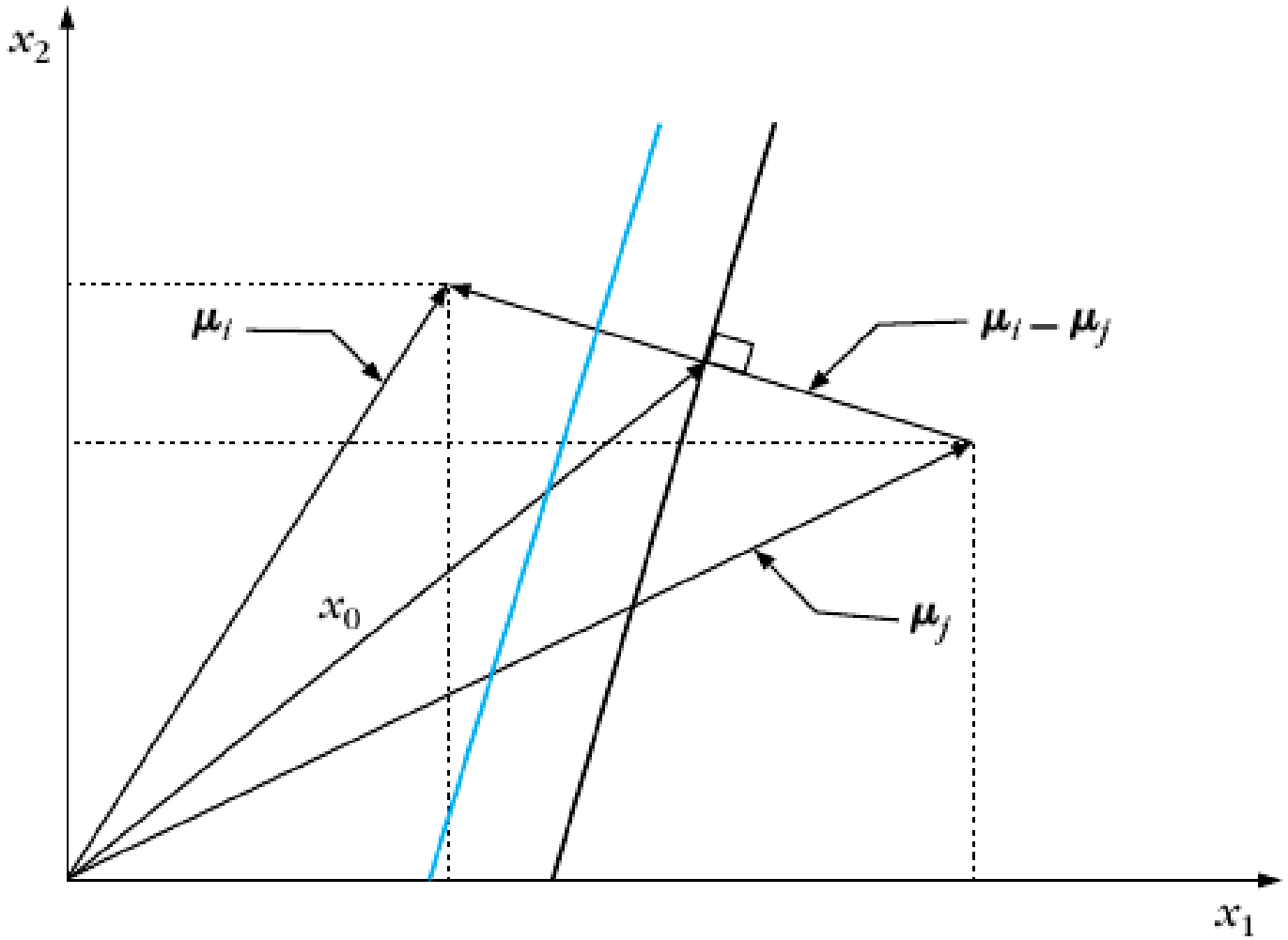
Clasificación Bayesiana: caso Gaussiano

Caso $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$



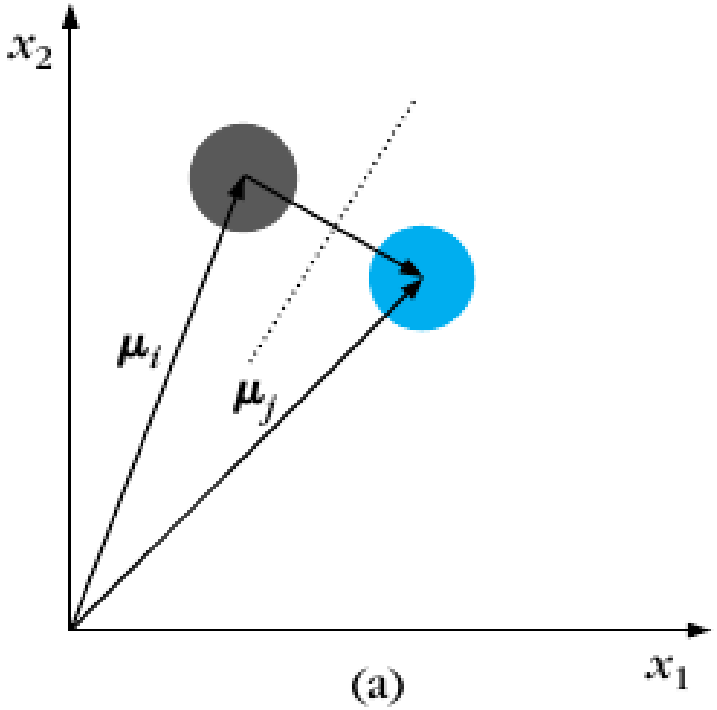
Clasificación Bayesiana: caso Gaussiano

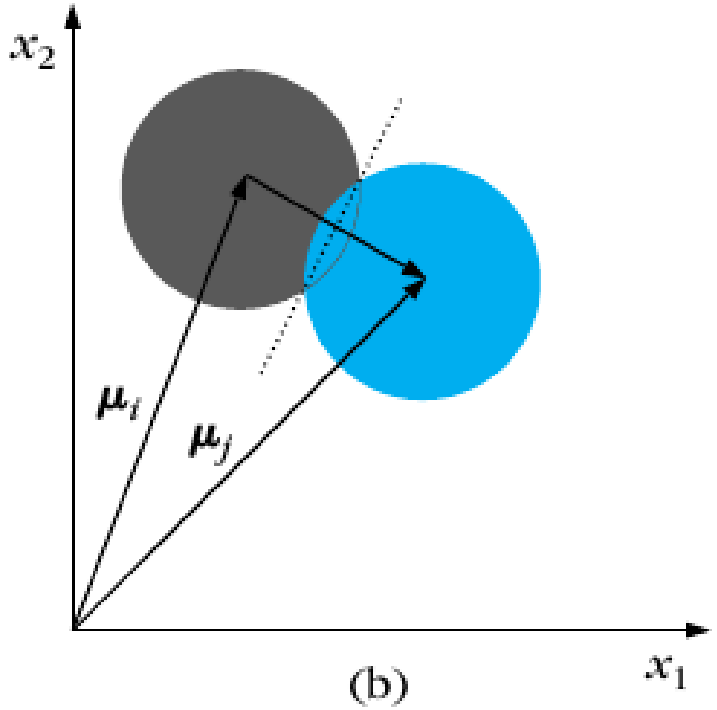
Caso $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$



Clasificación Bayesiana: caso Gaussiano

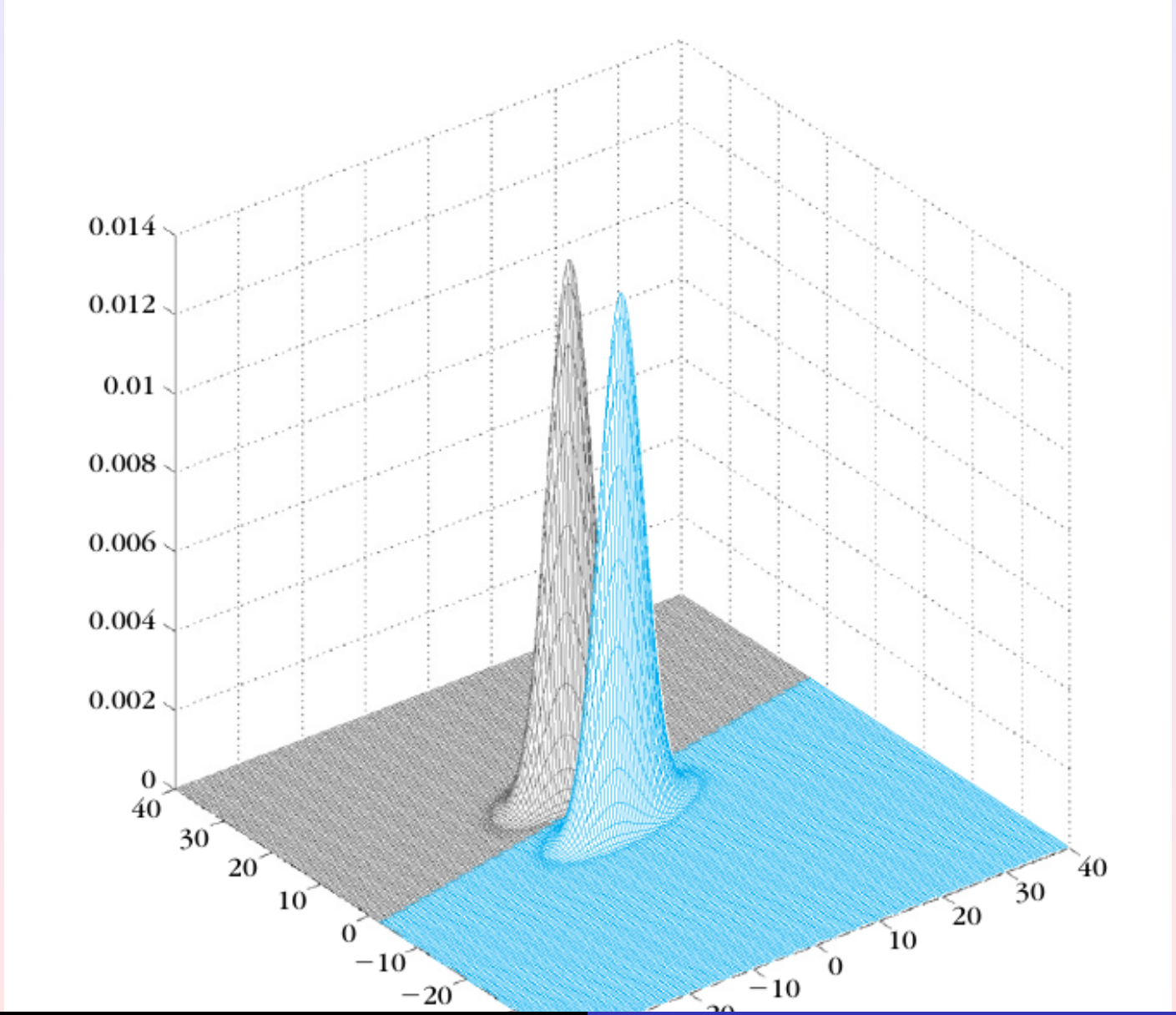
Caso $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$





Clasificación Bayesiana: caso Gaussiano

Caso $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$



Clasificación Bayesiana: caso Gaussiano

Caso $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

En este caso se suponen todas las matrices de covarianza iguales pero no diagonales, o sea $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$. La función discriminante ser'a en este caso

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_i})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu_i}) + \ln P(\omega_i) + c_i,$$

con $c_i = -(d/s)\ln(2\pi) - (1/2)ln|\Sigma_i|$. Expandiendo la anterior obtenemos

$$\begin{aligned} g_i(\mathbf{x}) = \quad & -\tfrac{1}{2}\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x} + \tfrac{1}{2}\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i} - \tfrac{1}{2}\boldsymbol{\mu_i}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i} + \tfrac{1}{2}\boldsymbol{\mu_i}^t\boldsymbol{\Sigma}^{-1}\mathbf{x} \\ & + \ln P(\omega_i) + c_i. \end{aligned}$$

Aquí el primert término $-1/2\mathbf{x}^t\boldsymbol{\Sigma}^{-1}\mathbf{x}$ no depende de i, por lo que puede ser excluido de $g_i(\mathbf{x})$.

Entonces, nuevamente la función discriminante $g_i(\mathbf{x})$ es lineal en $\mathbf{x}$

$$g_i(\mathbf{x}) = \mathbf{w_i}^t \mathbf{x} + w_{i0},$$

donde

$$\mathbf{w_i} = \mathbf{\Sigma}^{-1} \boldsymbol{\mu_i},$$

y

$$w_{i0} = \frac{-1}{2} \boldsymbol{\mu_i}^t \mathbf{\Sigma}^{-1} \boldsymbol{\mu_i} + \ln P(w_i).$$

Nuevamente, resolver $g_i(\mathbf{x}) = g_j(\mathbf{x})$ nos lleva a una ecuación de la forma

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x_0}) = 0,$$

pero con

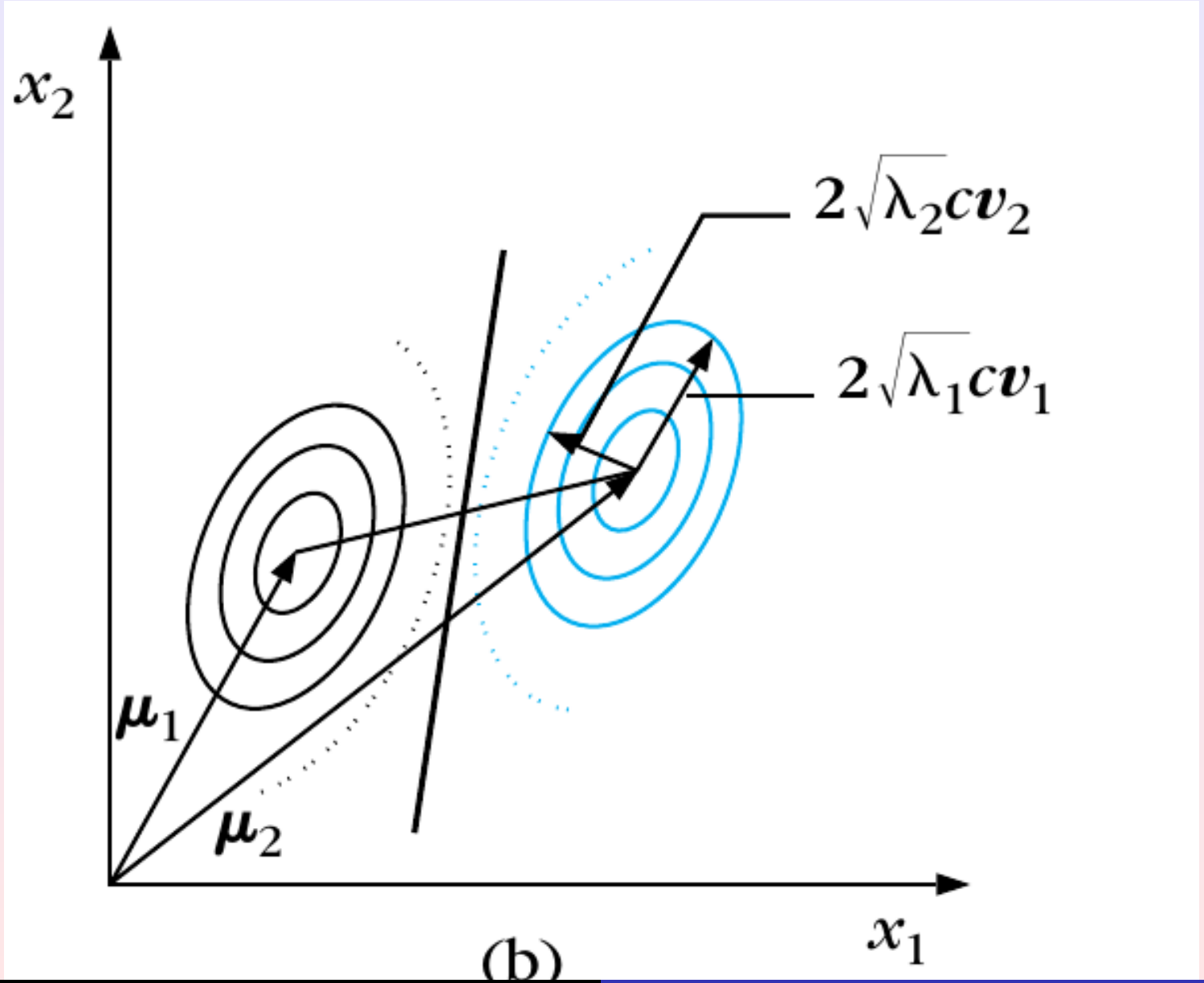
$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right),$$

y

$$\mathbf{x_0} = \frac{1}{2}\left(\boldsymbol{\mu_i} + \boldsymbol{\mu_j}\right) - \frac{\ln\left(P(\omega_i)/P(\omega_j)\right)}{\left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right)^t \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right)}\left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right).$$

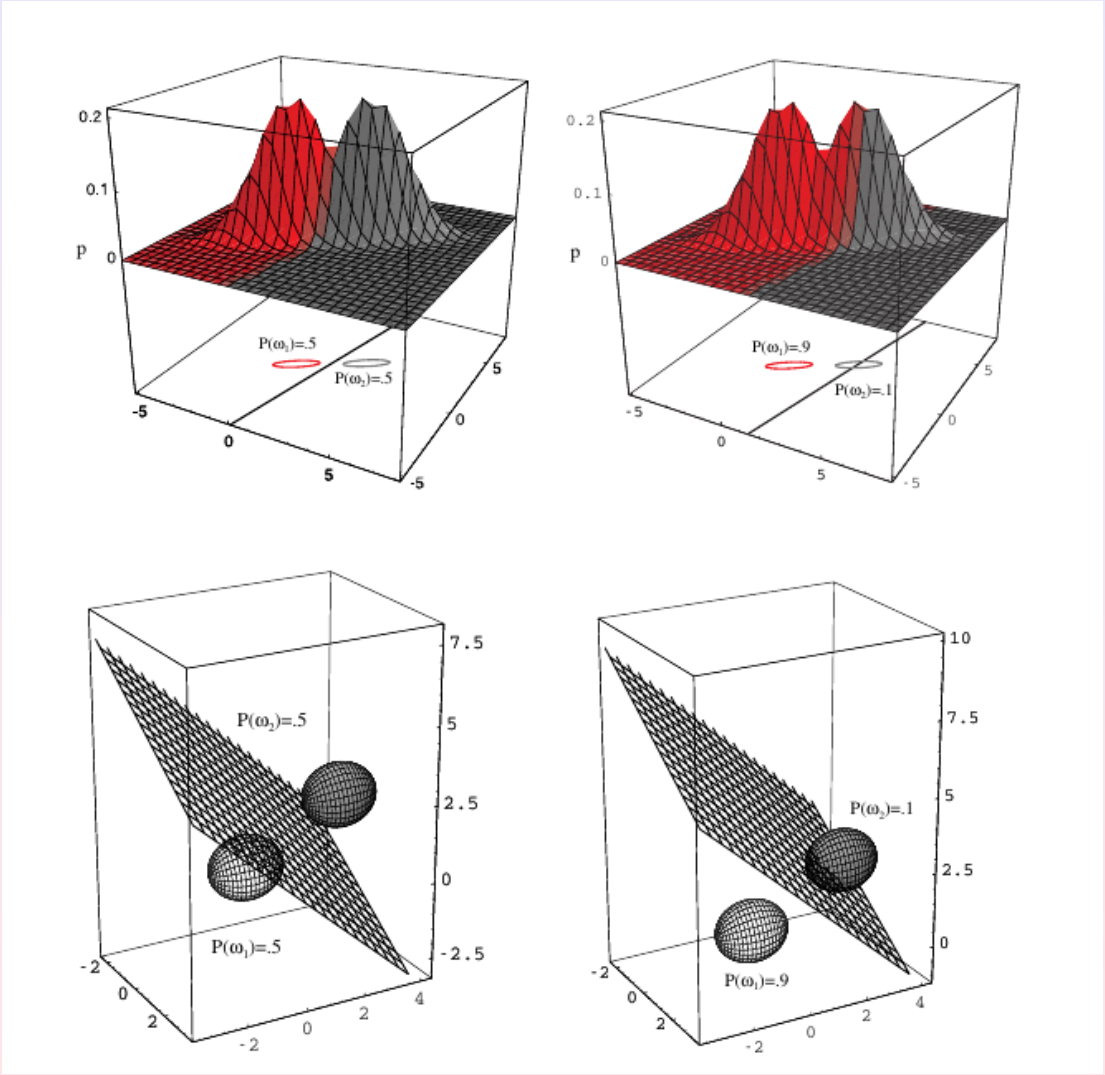
Clasificación Bayesiana: caso Gaussiano

Caso $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$



Clasificación Bayesiana: caso Gaussiano

Caso $\mathbf{\Sigma}_i = \mathbf{\Sigma}$



En este caso, la función discriminante ya no es lineal

$$g_i(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^t\mathbf{\Sigma_i}^{-1}\mathbf{x} + \boldsymbol{\mu_i}^t\mathbf{\Sigma_i}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu_i}^t\mathbf{\Sigma_i}^{-1}\boldsymbol{\mu_i} + \ln P(\omega_i) + c_i,$$

O sea, será de la forma

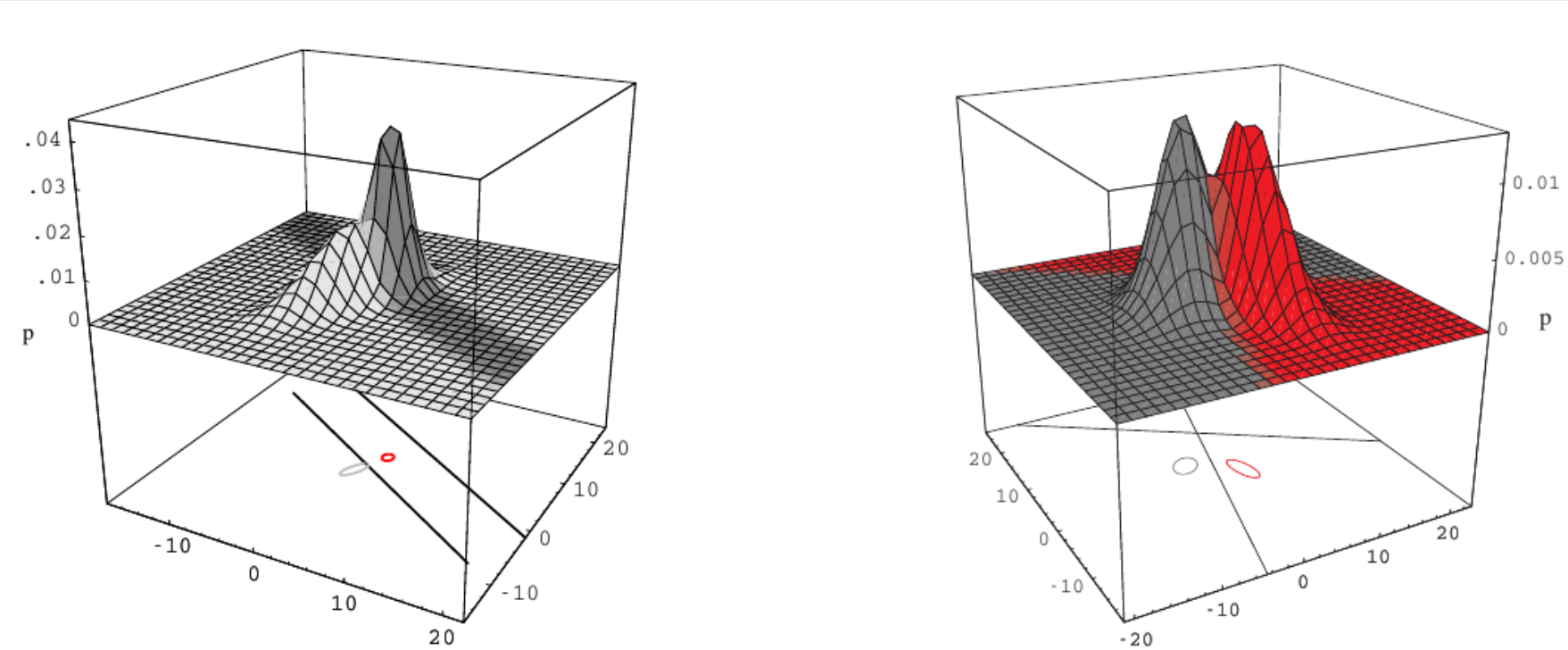
$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W_i}\mathbf{x} + \boldsymbol{w_i}^t\mathbf{x} + w_{i0},$$

donde $\mathbf{W_i} = -1/2\mathbf{\Sigma_i}^{-1}$, $\mathbf{w_i} = \mathbf{\Sigma_i}^{-1}\boldsymbol{\mu_i}$, y

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu_i}^t\mathbf{\Sigma_i}^{-1}\boldsymbol{\mu_i} - \frac{1}{2}|\mathbf{\Sigma_i}| + \ln P(\omega_i),$$

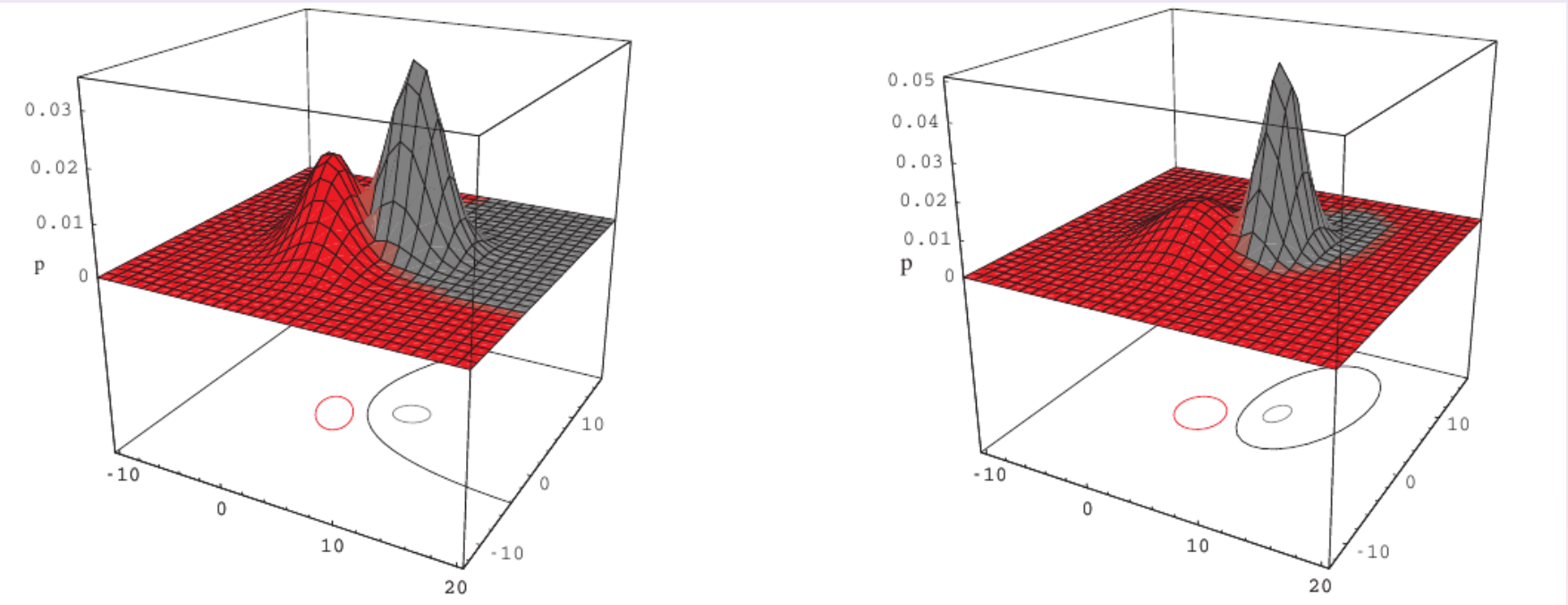
Clasificación Bayesiana: caso Gaussiano

Caso $\boldsymbol{\Sigma}_i$ arbitraria



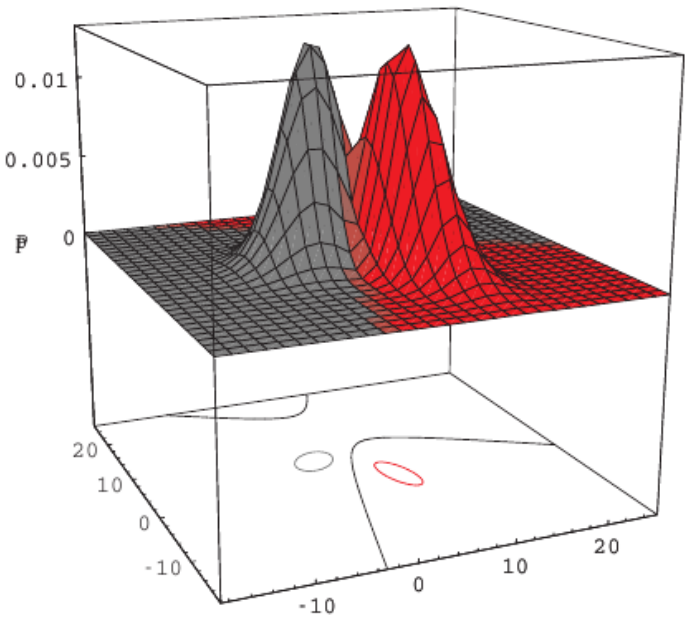
Clasificación Bayesiana: caso Gaussiano

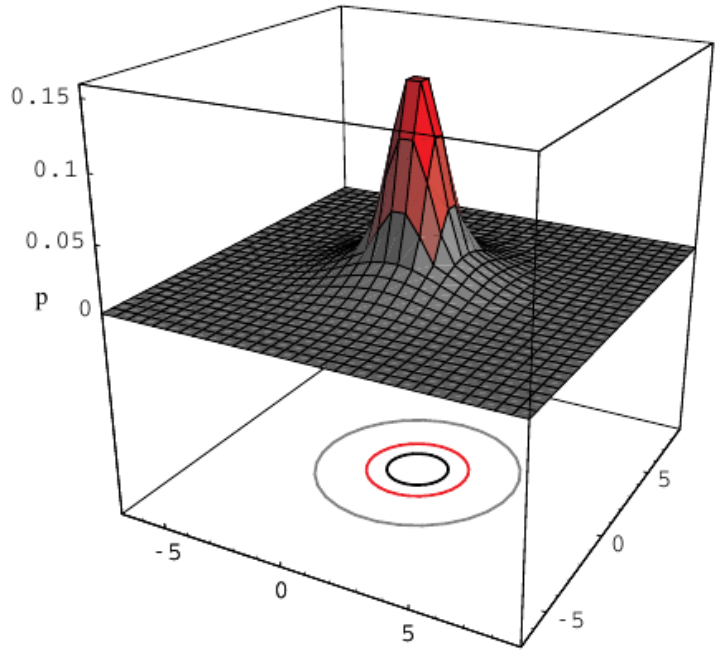
Caso $\boldsymbol{\Sigma}_i$ arbitraria



Clasificación Bayesiana: caso Gaussiano

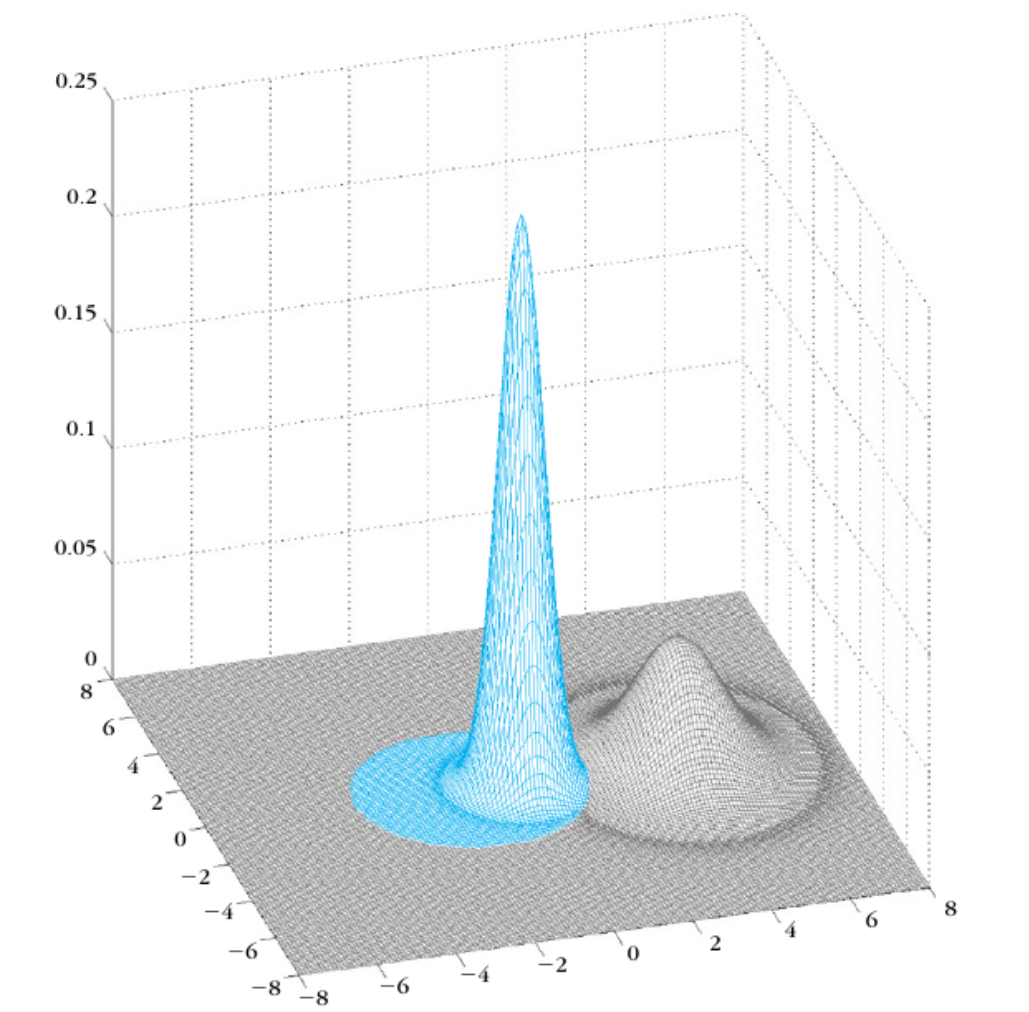
Caso $\boldsymbol{\Sigma}_i$ arbitraria





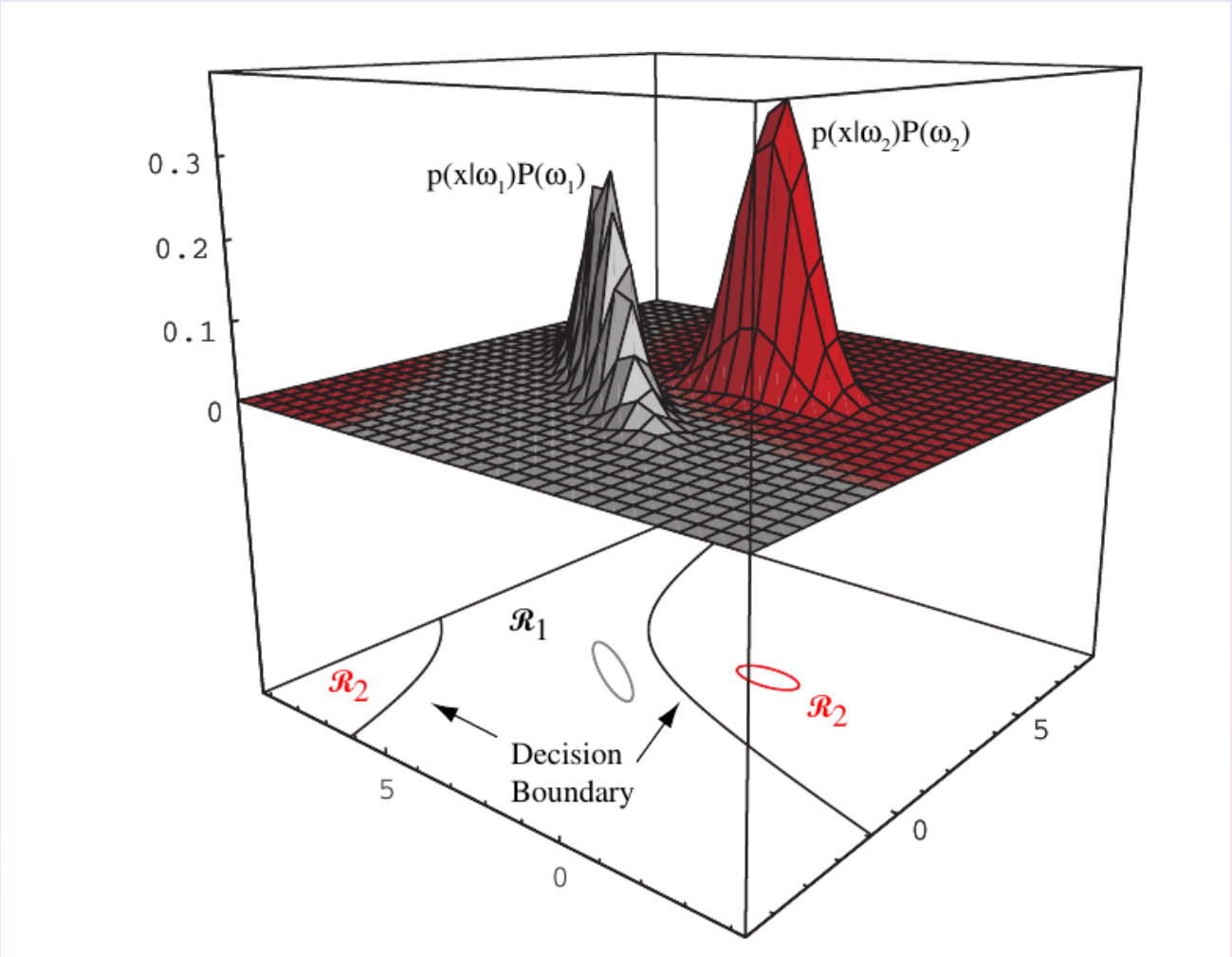
Clasificación Bayesiana: caso Gaussiano

Caso $\boldsymbol{\Sigma}_i$ arbitraria



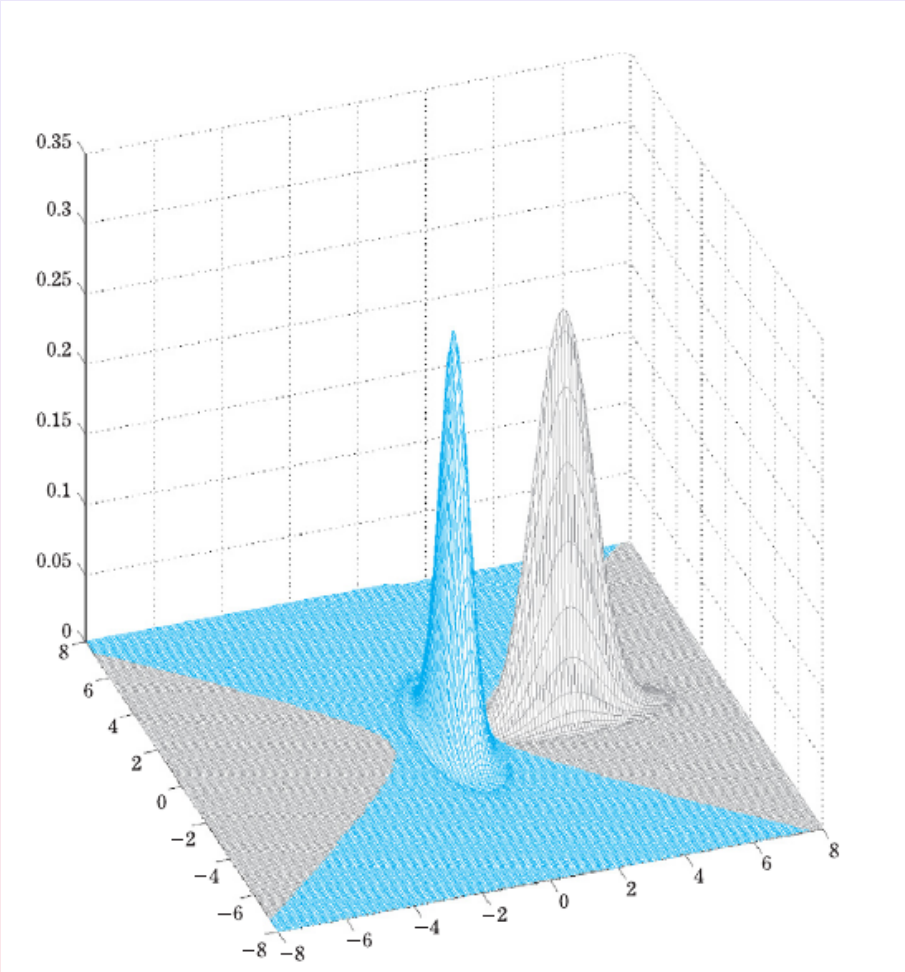
Clasificación Bayesiana: caso Gaussiano

Caso $\boldsymbol{\Sigma}_i$ arbitraria



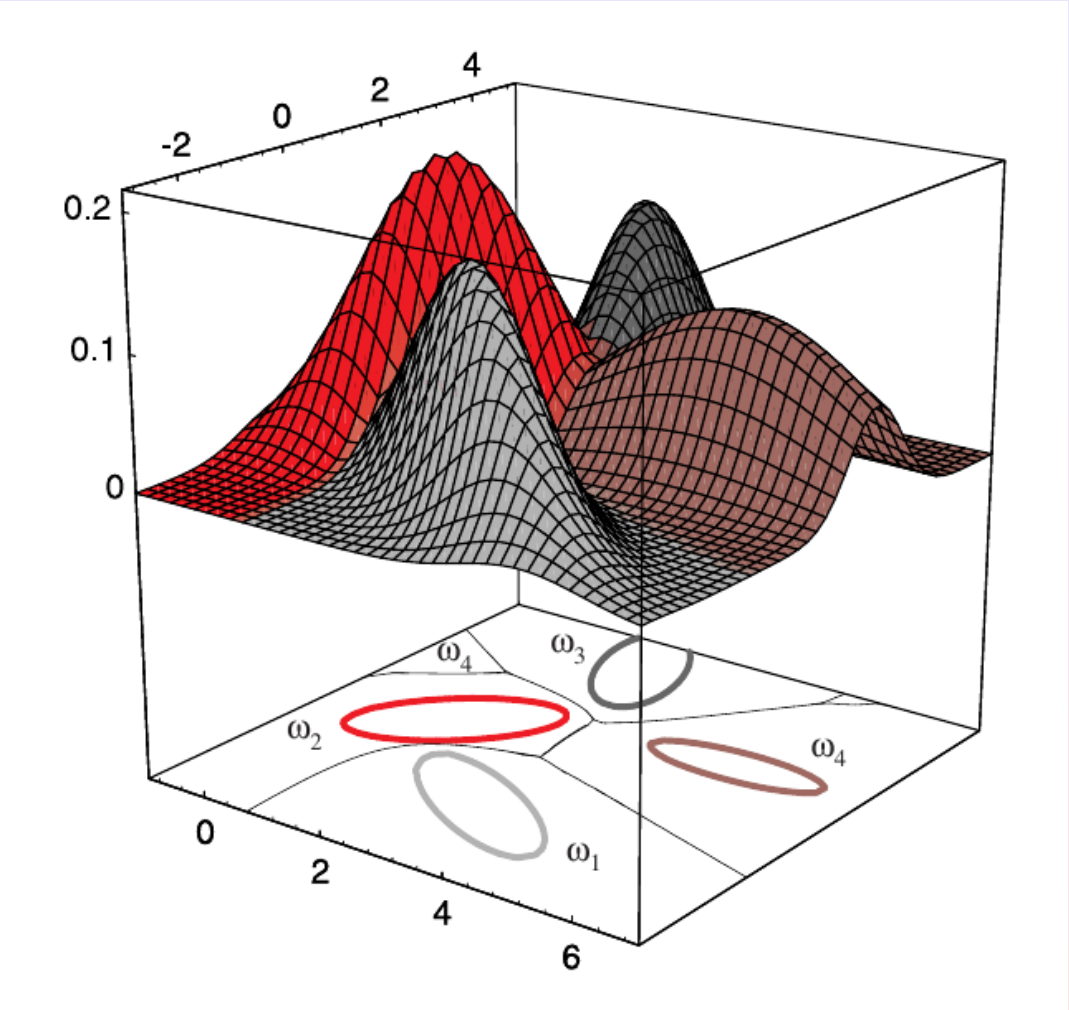
Clasificación Bayesiana: caso Gaussiano

Caso $\boldsymbol{\Sigma}_i$ arbitraria



Clasificación Bayesiana: caso Gaussiano

Caso $\mathbf{\Sigma}_i$ arbitraria



Error y Riesgo

En el caso de dos clases, el error medio es

$$
\begin{aligned}
P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\mathbf{x} \in \mathcal{R}_2 \mid \omega_1) P(w_i) + P(\mathbf{x} \in \mathcal{R}_1 \mid \omega_2) P(w_2) \\
&= P(w_1) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1) d\mathbf{x} + P(w_2) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2) d\mathbf{x}
\end{aligned}
$$

Estimación por Máxima Verosimilitud

Supongamos que tenemos c conjuntos de datos $\mathcal{D}_1, \ldots . \mathcal{D}_c$ pertenecientes a c clases. Supongamos también que los valores de los mismos han sido extraídos de acuerdo a las densidades $p(\mathbf{x} \mid \omega_i)$, con $i \in 1, \ldots, c$, y que estos valores son i.i.d..

Suponemos además que conocemos la "forma" de estas densidades, pero no el valor de sus parámetros.

Por ejemplo: suponemos $p(\mathbf{x} \mid \omega_i) \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, pero desconocemos los valores de $\boldsymbol{\mu}_i$ y de $\boldsymbol{\Sigma}_i$.

En general, llamaremos θ_i al vector de parámetros de estas distribuciones.

Tenemos entonces c problemas separados: omitiendo el subíndice i, estimar el vector de parámetros θ de la densidad $p(\mathbf{x} \mid \theta)$, utilizando el conjunto de datos $\mathcal{D}$.

Para el caso Gaussiano, tendremos $p(\mathbf{x} \mid \theta) = p(\mathbf{x} \mid \mu, \boldsymbol{\Sigma})$.

Estimación por Máxima Verosimilitud

Si nuestros datos son $\mathcal{D} = \{\mathbf{x}_1, \ldots . \mathbf{x}_n\}$, entonces, al ser i.i.d., su densidad será

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k \mid \boldsymbol{\theta}),$$

y el estimador por Máxima Verosimilitud es

$$\hat{\boldsymbol{\theta}}_{MV} = \arg \max_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta}),$$

como el logaritmo es una función creciente, la anterior puede definirse también como

$$\hat{\boldsymbol{\theta}}_{MV} = \arg \max_{\boldsymbol{\theta}} \ln p(\mathcal{D} \mid \boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}),$$

Estimación por Máxima Verosimilitud

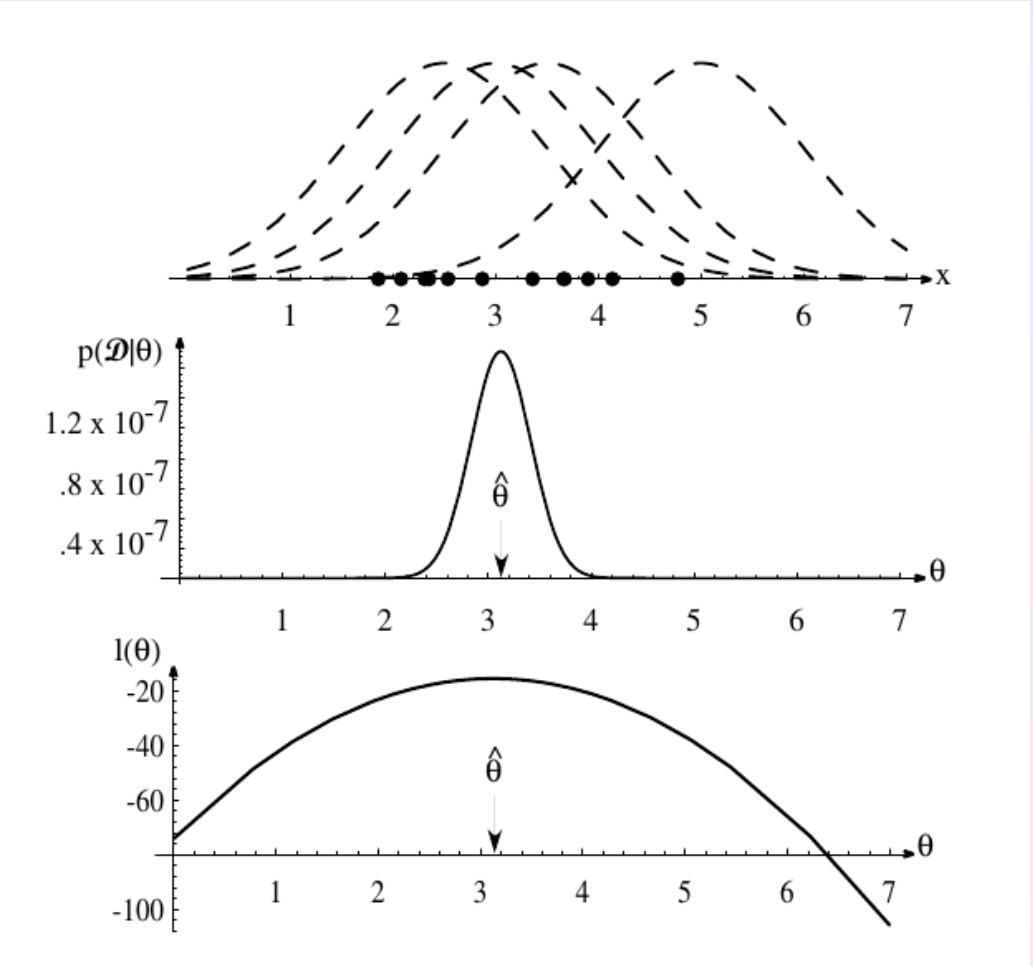
Donde $\sum_{k=1}^{n} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$ es la llamada log-verosimilitud $\ell(\boldsymbol{\theta})$, por lo que la anterior queda

$$\hat{\boldsymbol{\theta}}_{MV} = \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}),$$

si $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_p)$ y definimos el gradiente $\nabla_{\boldsymbol{\theta}}$ como

$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

Estimación por Máxima Verosimilitud



Estimación por Máxima Verosimilitud

Tomando el gradiente de $\ell(\boldsymbol{\theta})$

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta}),$$

$\hat{\boldsymbol{\theta}}_{MV}$ será solución de

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbf{0}$$

En este caso, como

$$\ln p\left(\mathbf{x}_k \mid \boldsymbol{\mu}\right) = -\frac{1}{2}\ln\left((2\pi)^d|\boldsymbol{\Sigma}|\right) - \frac{1}{2}\left(\mathbf{x}_k - \boldsymbol{\mu}\right)^t \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_k - \boldsymbol{\mu}\right),$$

Entonces

$$\nabla_\theta \ln p\left(\mathbf{x}_k \mid \boldsymbol{\mu}\right) = \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_k - \boldsymbol{\mu}\right),$$

y por lo tanto

$$\hat{\boldsymbol{\mu}}_{MV} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{\mu}_k,$$

En el caso unidimensional tenemos $\boldsymbol{\theta}^t = (\mu, \sigma^2)$ y

$$\ln p\left(x_k \mid \mu, \sigma^2\right) = -\frac{1}{2}\ln((2\pi)\sigma^2) - \frac{1}{2\sigma^2}(x_k - \mu)^2$$

y por lo tanto

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln p(x_k \mid \boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2}(x_k - \mu)^2 \\ -\frac{1}{2\sigma^2} + \frac{(x_k-\mu)^2}{2\sigma^2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

De la ecuación anterior obtenemos

$$\hat{\mu}_{MV} = \frac{1}{n}\sum_{k=1}^{n} x_k,$$

y

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \mu)^2.$$

Estimación Bayesiana

En el método de máxima verosimilitud se considera que el vector de parámetros θ es fijo.