Pattern Recognition
Discriminant Functions
Parametric approaches for real classifiers

Francesco Tortorella

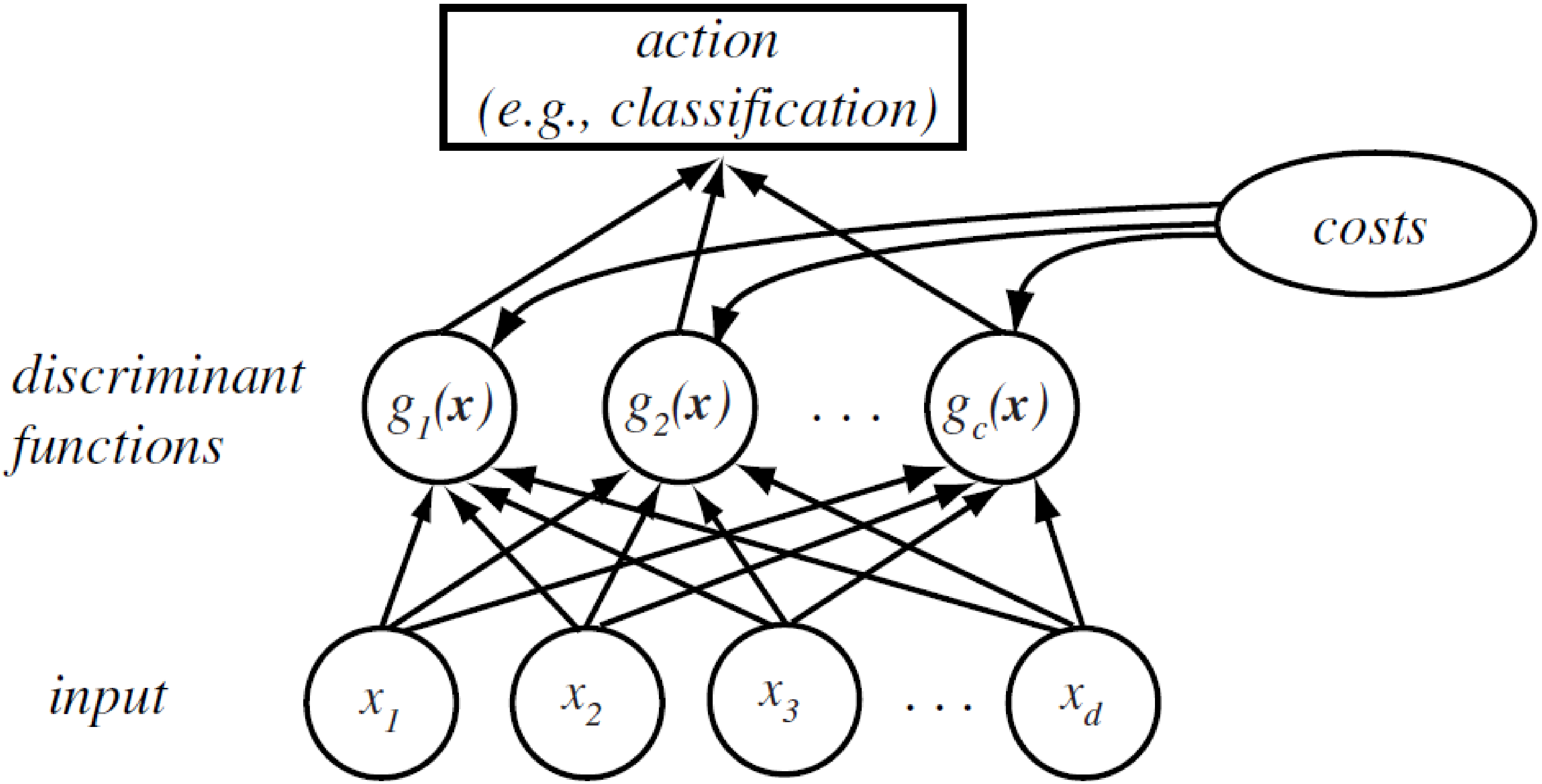
University of Cassino and Southern Latium

Cassino, Italy

Discriminant functions

- An useful representation of a classifier is given in terms of *discriminant functions* $g_i(x)$ $i=1,\ldots,C$.
- A sample x is assigned to the class ω_i iff $g_i(x) > g_j(x) \quad j \neq i$.
- In this way, the classifier is arranged as a system calculating C discriminant funtions and choosing the class with the highest value.

Discriminant functions



Discriminant functions

- A Bayes classifier can be easily represented in terms of discriminant functions and in several ways.
- MAP: $g_i(x) = P(\omega_i|x)$
- Minimum Risk: $g_i(x) = -R(\alpha_i|x)$
- Generally speaking, the choice of the discriminant functions is not unique and every monotonic function of $P(\omega_i|x)$ could be used:
  - $g_i(x) = p(x|\omega_i)\, P(\omega_i)$
  - $g_i(x) = \ln P(\omega_i|x) = \ln p(x|\omega_i) + \ln P(\omega_i)$

Discriminant functions

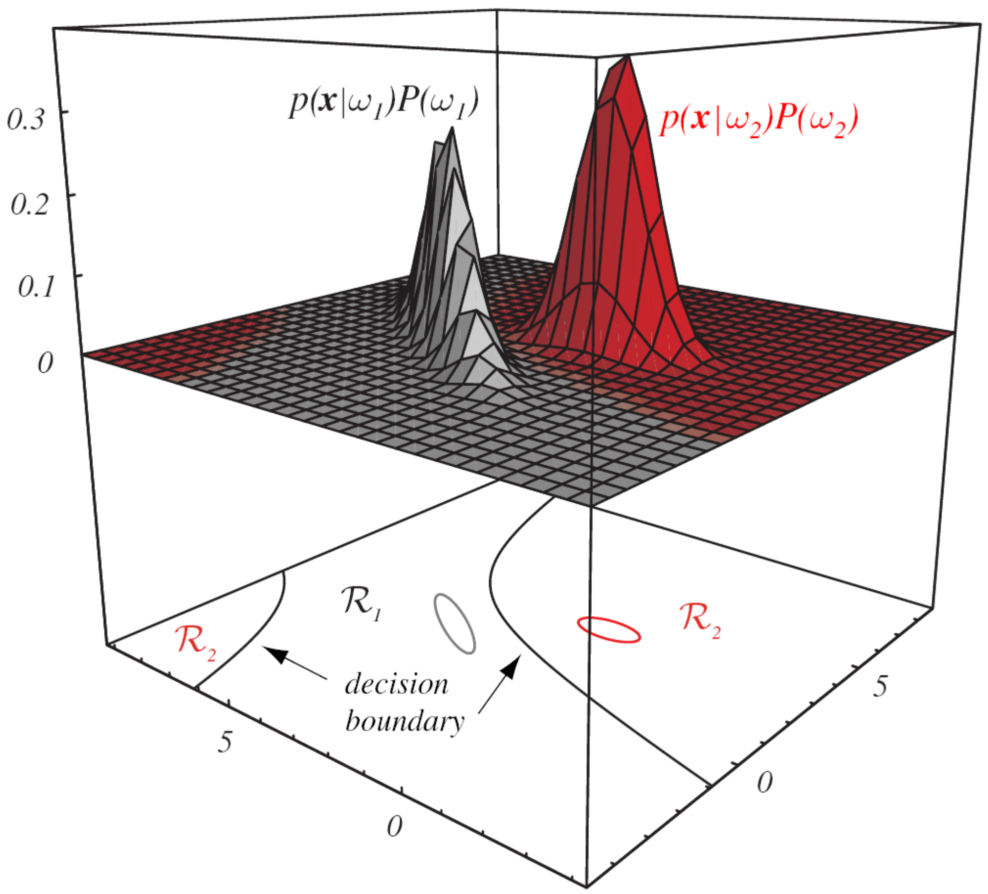
- The decision regions are immediately defined in terms of discriminant regions:

$$\mathcal{R}_i(x) = \{x | g_i(x) > g_j(x), \forall j \neq i\}$$

- While the decision boundary between classes ω_i and ω_j is:

$$\Gamma_{ij}(x) = \{x | g_i(x) = g_j(x), j \neq i\}$$

Discriminant functions

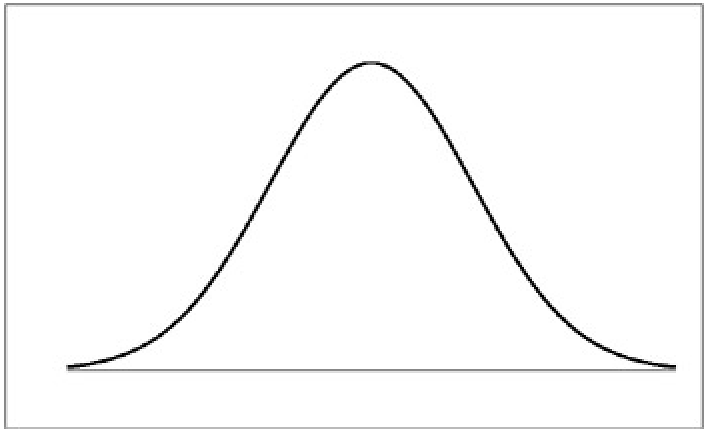


Gaussian densities

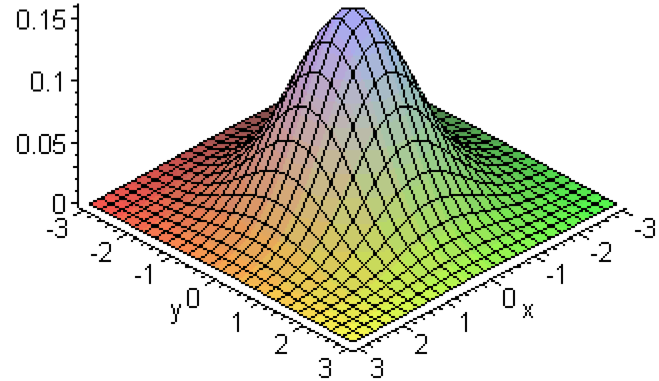
- The structure of a Bayes classifier is determined by the conditional densities $p(x|\omega_i)$ as well as by the prior probabilities $P(\omega_i)$.
- Gaussian (normal) density frequent choice because of
  - its analytical tractability
  - appropriately modelling the f.v. x as the noisy version of a prototype μ_i for the class ω_i

Univariate Gaussian

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma^2}\right)$$



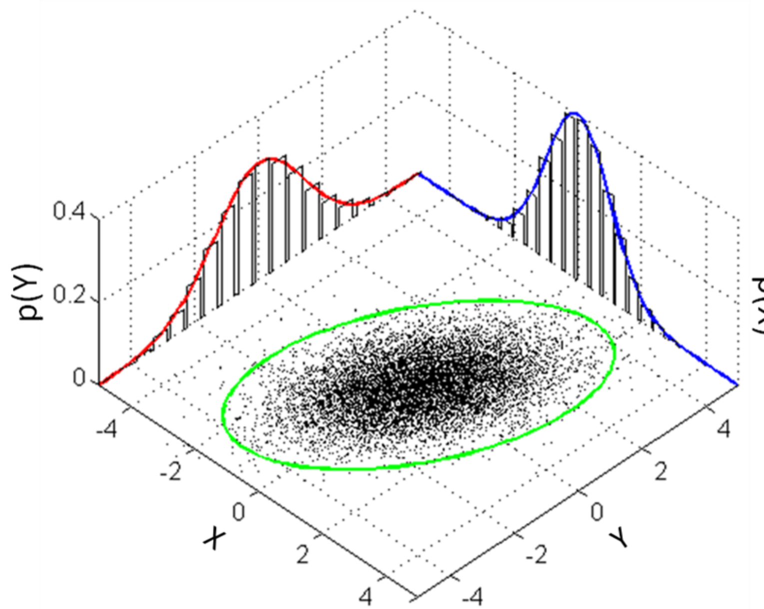
Bivariate Gaussian



$$p(\mathbf{x}|\omega_i) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \qquad \boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \end{bmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{pmatrix}$$



Multivariate Gaussian

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right)$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \boldsymbol{\mu}_i = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \vdots \\ \mu_{id} \end{bmatrix} \quad \Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{12} & \sigma_{22}^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1d} & \sigma_{2d} & \cdots & \sigma_{2d}^2 \end{pmatrix}$$

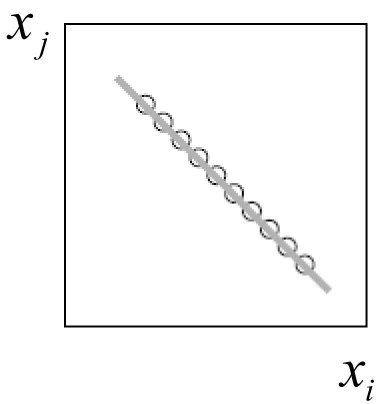
$$\boldsymbol{\mu}_i = E[\mathbf{x}|\omega_i] \qquad \Sigma = E[(\mathbf{x}-\boldsymbol{\mu}_i)(\mathbf{y}-\boldsymbol{\mu}_i)^T|\mathbf{x}\in\omega_i]$$

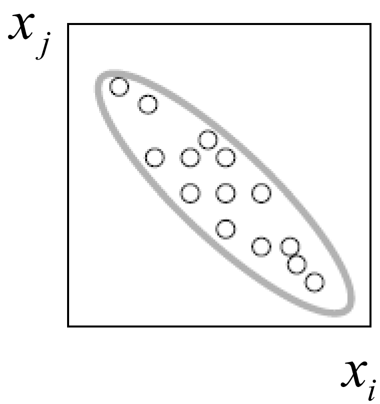
$$\mu_{i,h} = E[x_h|\mathbf{x}\in\omega_i] \quad \sigma_{hk} = E[(x_h-\mu_{i,h})(x_k-\mu_{i,k})|\mathbf{x}\in\omega_i]$$

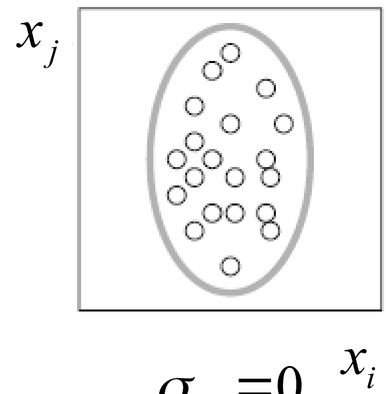
Properties of the covariance matrix

- Symmetric: $\sigma_{ij} = \sigma_{ji}$
- Variances of the components on the diagonal: $\sigma_{ii} = \sigma_i^2$
- The off-diagonal elements are the covariances $\left|\sigma_{ij}\right| \leq \sigma_i \sigma_j$
- If x_i and x_j grow together $\sigma_{ij}>0$.
- If x_i grows when x_j decreases $\sigma_{ij}<0$.
- If x_i and x_j statistically independent $\sigma_{ij}=0$.

Properties of the covariance matrix





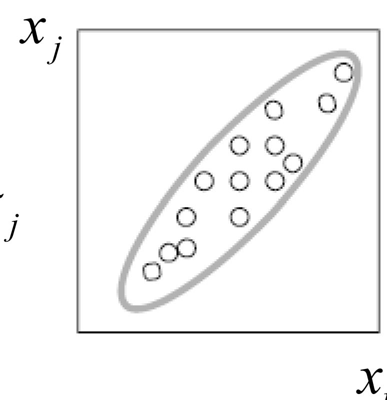


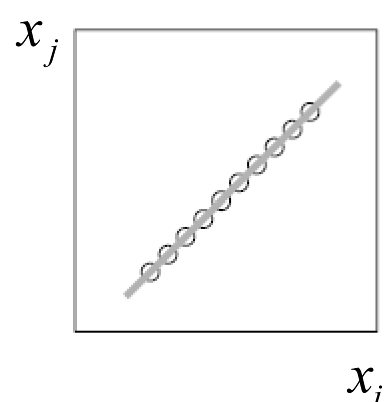
$$\sigma_{ij} = -\sigma_i \sigma_j$$

$$\sigma_{ij} < 0 \quad |\sigma_{ij}| < \sigma_i \sigma_j$$

$$\sigma_{ij} = 0$$

$$\sigma_{ij} > 0 \quad |\sigma_{ij}| < \sigma_i \sigma_j$$

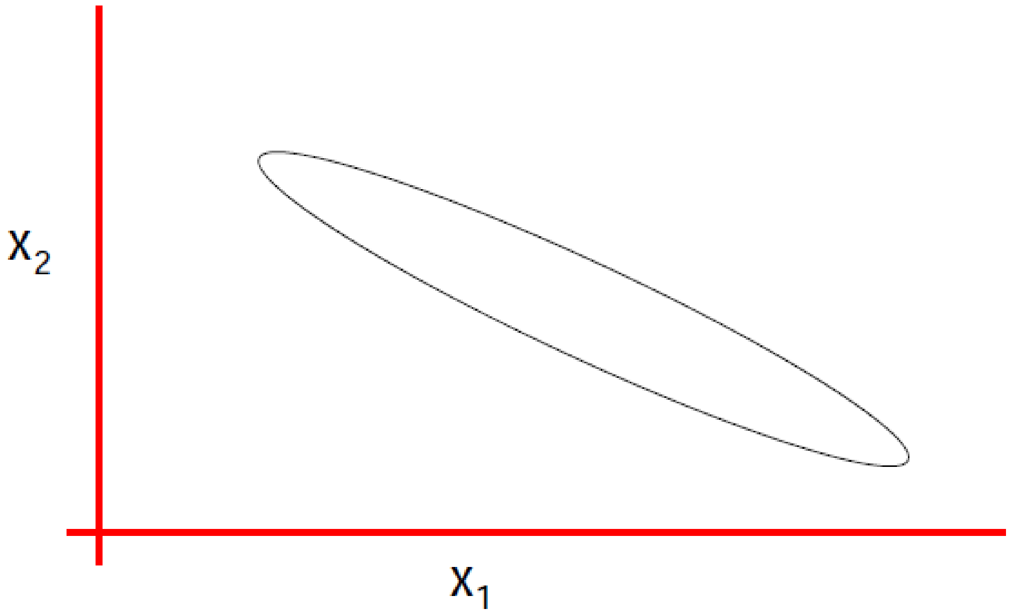




$$\sigma_{ij} = \sigma_i \sigma_j$$

General Gaussians

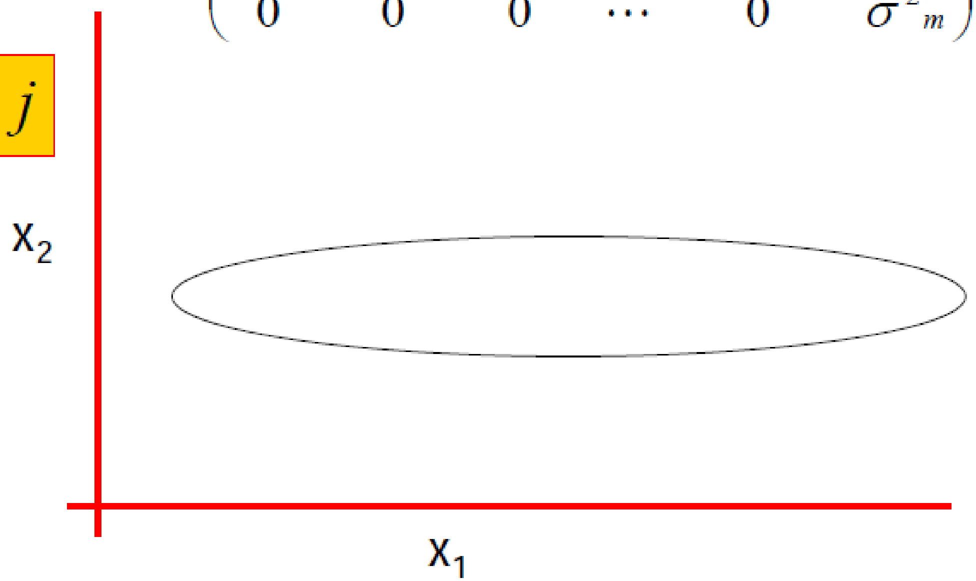
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma^2{}_2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma^2{}_m \end{pmatrix}$$



Axis-Aligned Gaussians

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2{}_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2{}_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2{}_{m-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2{}_m \end{pmatrix}$$

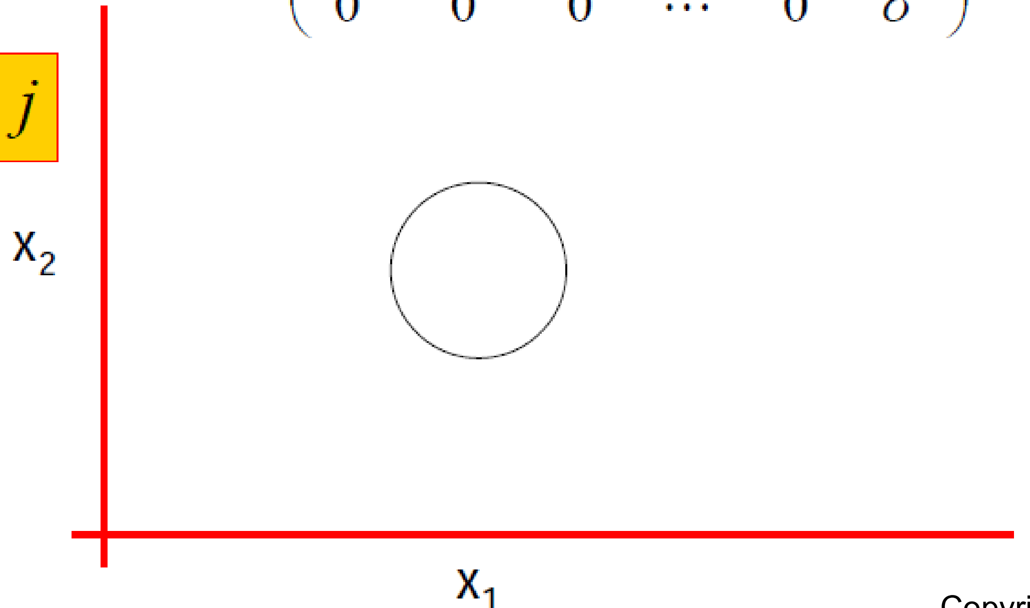
$$\boxed{X_i \perp X_i \text{ for } i \neq j}$$



Spherical Gaussians

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$

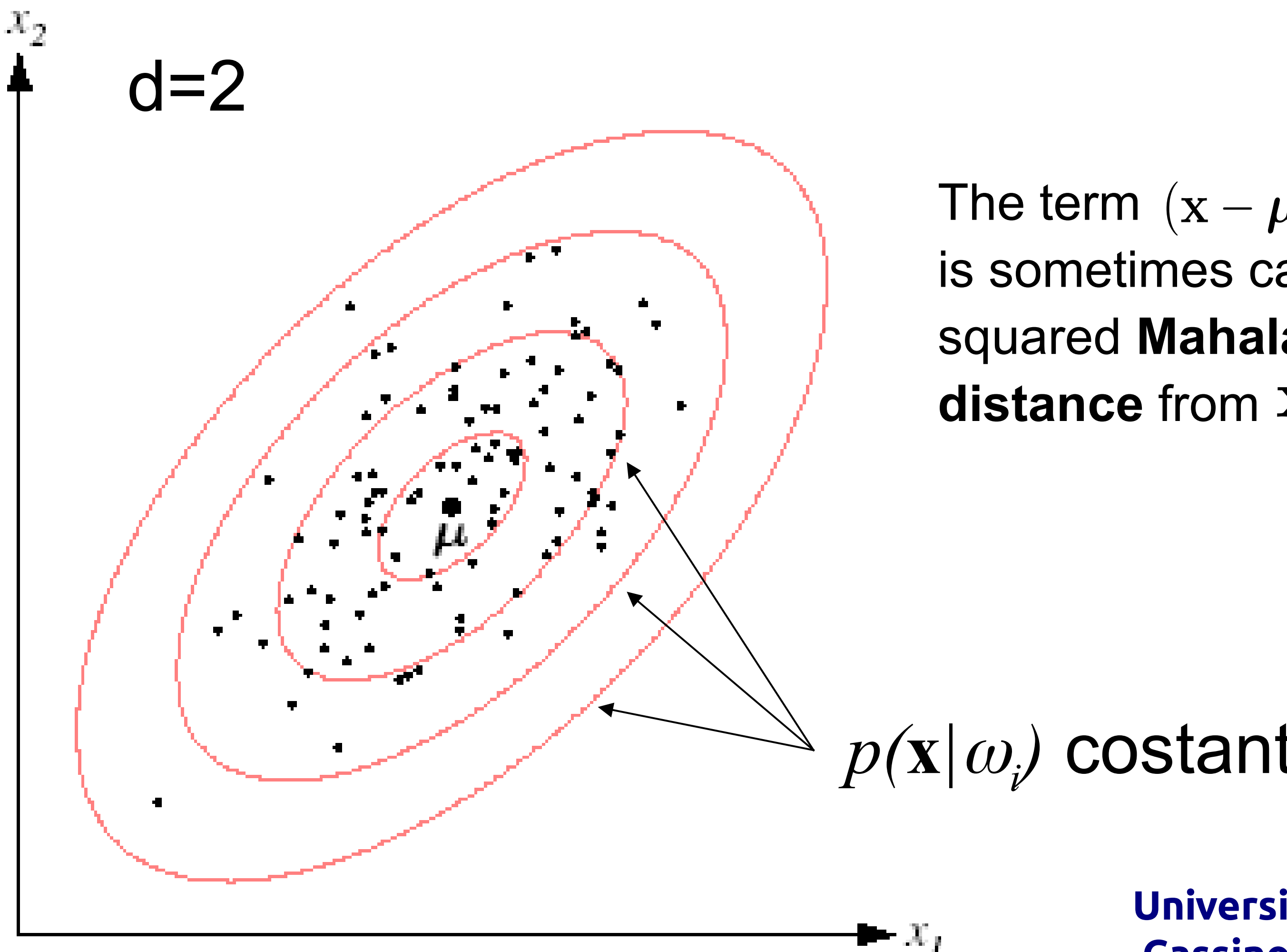
$X_i \perp X_i \text{ for } i \neq j$



Gaussian densities

- Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}_i$
- The shape of the region is defined by the covariance matrix
- Points with the same value for the density lie on curves on which the term $\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)$ is constant.

Gaussian densities



The term $(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ is sometimes called the squared **Mahalanobis distance** from $\mathbf{x}$ to $\boldsymbol{\mu}_i$

From theory to practice

- If we knew all the probability functions related to a particular problem, through the Bayes classifier we could build the optimal decision system.
- Unfortunately, in real problems priors $P(\omega_i)$ and likelihoods $p(x|\omega_i)$ are not known.
- Almost always we have no (or very limited) information about the process that produced the data we want to recognize.

Parametric or non parametric?

- Tipically, what we have is a large (?) set of examples and some knowledge about the problem.
- Thus the only viable option is to learn from the available information how to decide about new samples (*learning by examples*).
- A first approach could be to assume a particular form for the densities (e.g. Gaussian) and evaluate the parameters (e.g. mean and covariance) from the data (***parametric approach***).

Bayes classifier with Gaussian densities

- We saw that, in the case of the MAP rule, the discriminant functions can be defined as:

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

- If we assume Gaussian densities we have:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{d}{2}\ln 2\pi + \ln P(\omega_i)$$

- Without any assumption on Σ_i, the Bayes classifier is a *quadratic classifier*.

Gaussian densities

$\Sigma_i = \sigma^2 \mathbf{I}$

- If the features are statistically independent with the same variance σ^2, the form of $g_i(x)$ becomes simpler:

$$\Sigma_i^{-1} = \frac{1}{\sigma^2}\mathbf{I} \qquad \left|\Sigma_i\right| = \sigma^{2d}$$

$$g_i(\mathbf{x}) = -\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T(\mathbf{x}-\boldsymbol{\mu}_i)}{2\sigma^2} + \ln P(\omega_i) = -\frac{\left\|\mathbf{x}-\boldsymbol{\mu}_i\right\|^2}{2\sigma^2} + \ln P(\omega_i)$$

Euclidean distance

Gaussian densities
$\Sigma_i = \sigma^2 I$

- Let's elaborate $g_i(x)$:

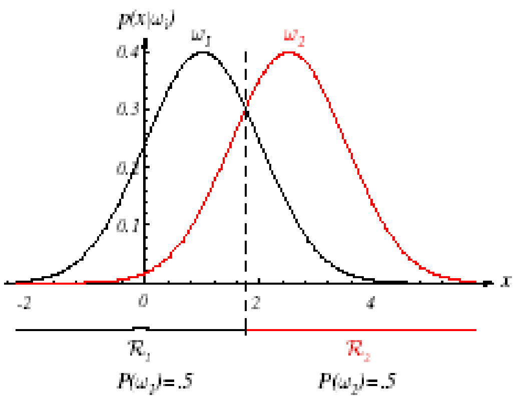
$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^T\mathbf{x} - 2\boldsymbol{\mu}_i^T\mathbf{x} + \boldsymbol{\mu}_i^T\boldsymbol{\mu}_i\right] + \ln P(\omega_i)$$

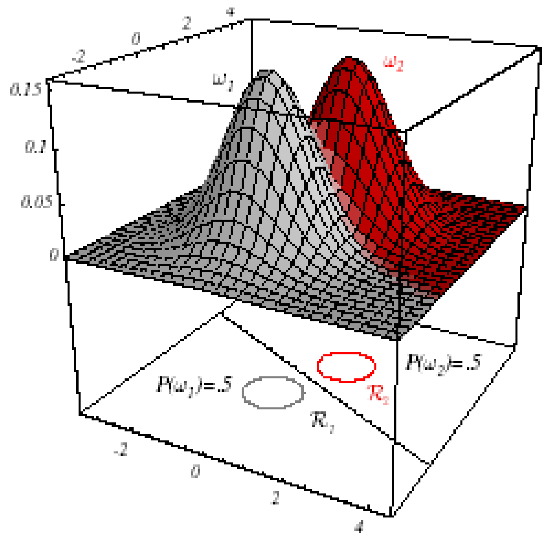
- Since $x^T x$ is independent of the class ω_i, we obtain a *linear classifier* (a.k.a.*linear machine*):

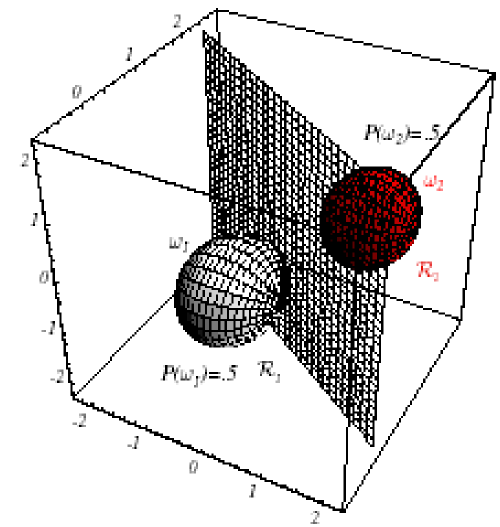
$$g_i(\mathbf{x}) = \frac{\boldsymbol{\mu}_i^T\mathbf{x}}{\sigma^2} - \frac{\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i}{2\sigma^2} + \ln P(\omega_i) = \mathbf{w}_i^T\mathbf{x} + \mathbf{w}_{i0}$$

Gaussian densities

$\Sigma_i = \sigma^2 I$







d = 1 d = 2 d = 3

Gaussian densities $\Sigma_i = \sigma^2 I$

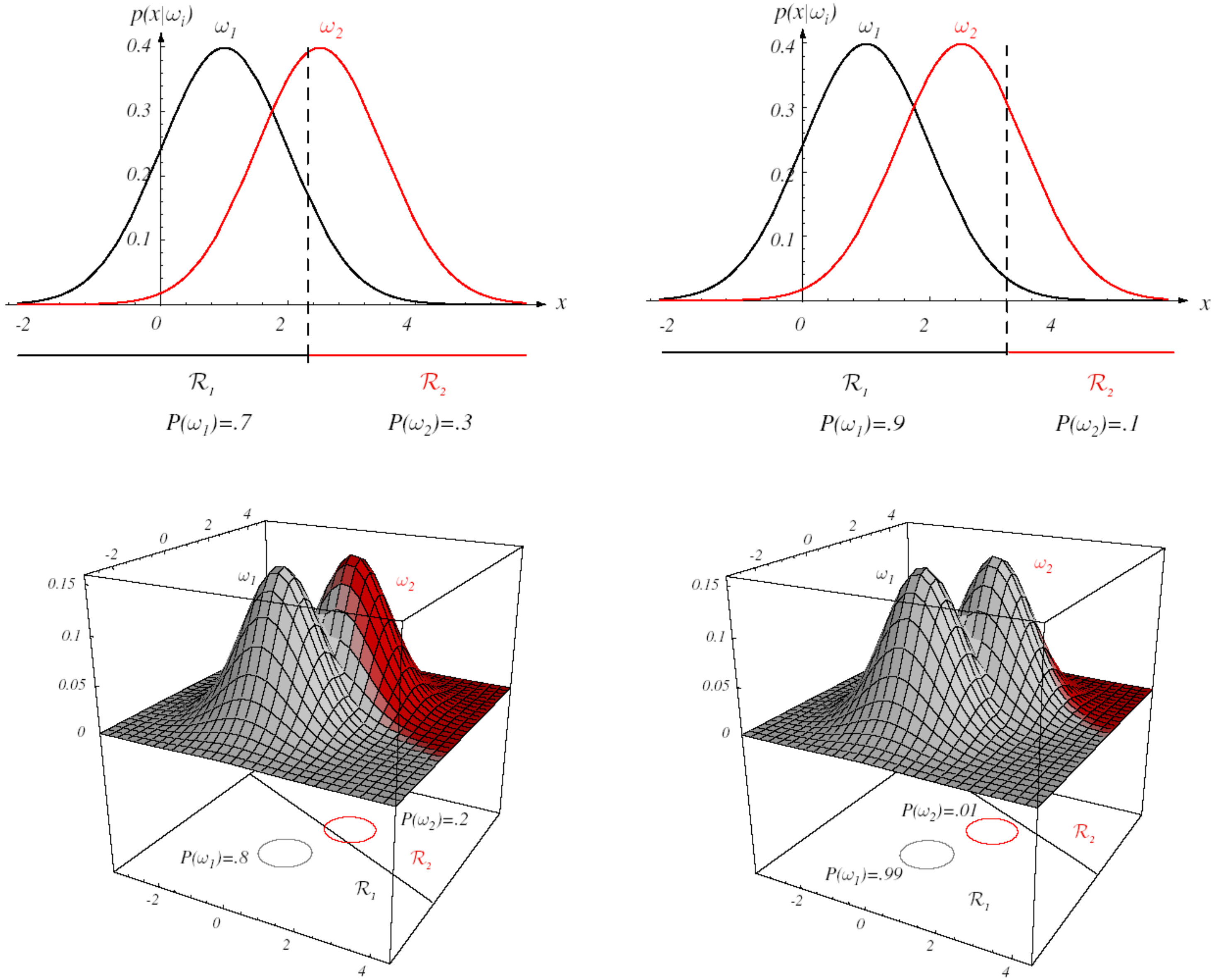
- As for the decision boundary $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = \left(\mathbf{w}_i - \mathbf{w}_j\right)^T \mathbf{x} + \left(w_{i0} - w_{j0}\right) = 0$$

- In this case the equation of the boundary can be written as $\mathbf{w}^T\left(\mathbf{x} - \mathbf{x}_0\right) = 0$ where:

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\mathbf{x}_0 = \frac{1}{2}\left(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j\right) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)$$



The boundary depends on the priors $P(\omega_i)$

Gaussian densities
$\Sigma_i = \Sigma$

- Also in this case $g_i(x)$ becomes simpler:

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i) - \frac{d}{2}\ln 2\pi$$

$$\Downarrow$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln P(\omega_i)$$

$$\Downarrow$$

Mahalanobis distance

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i$$

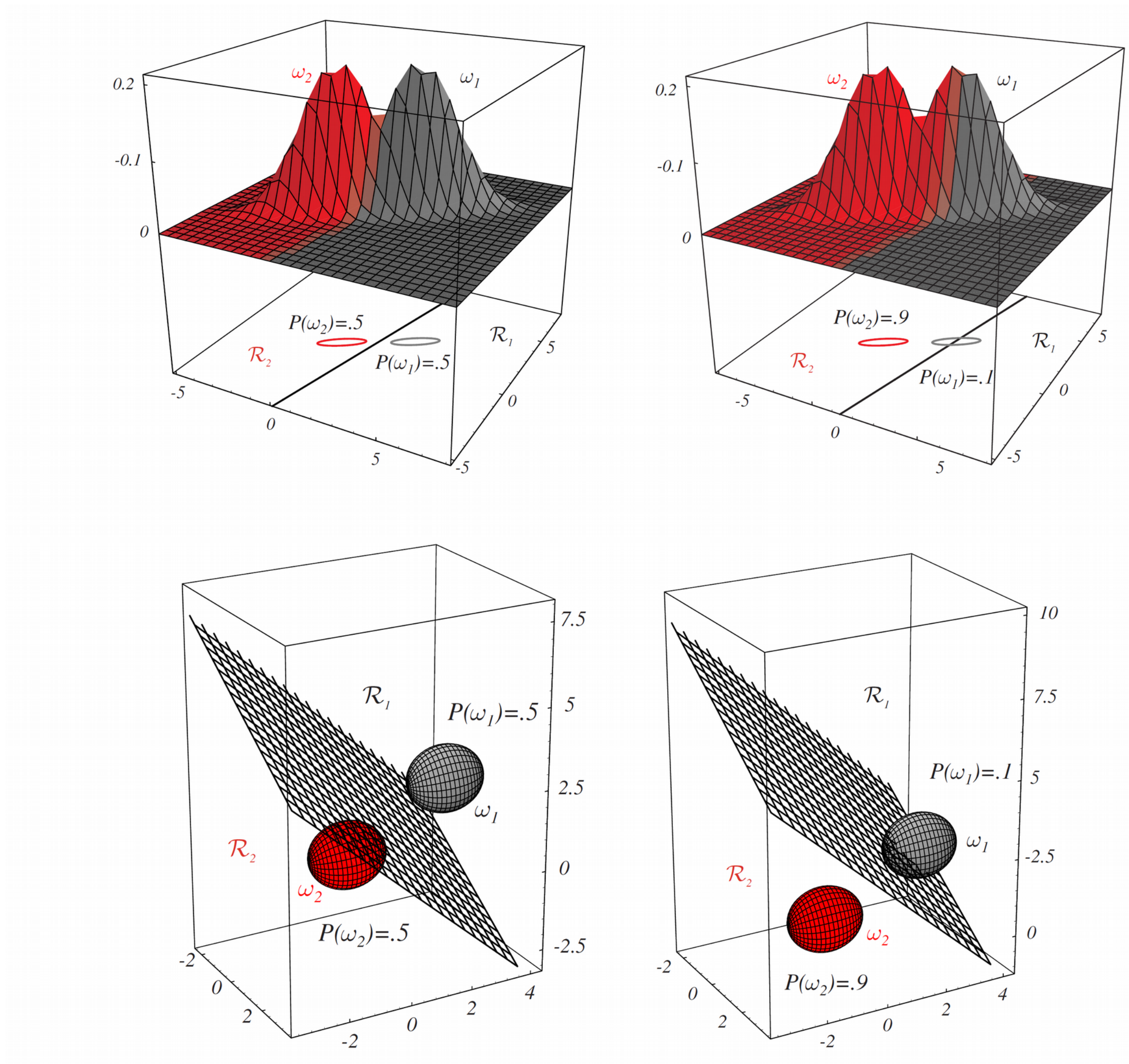
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

Gaussian densities $\Sigma_i = \Sigma$

- Once again the equation of the boundary can be written as $\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$ where:

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

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



The decision boundary needs not be perpendicular to μ_i-μ_j