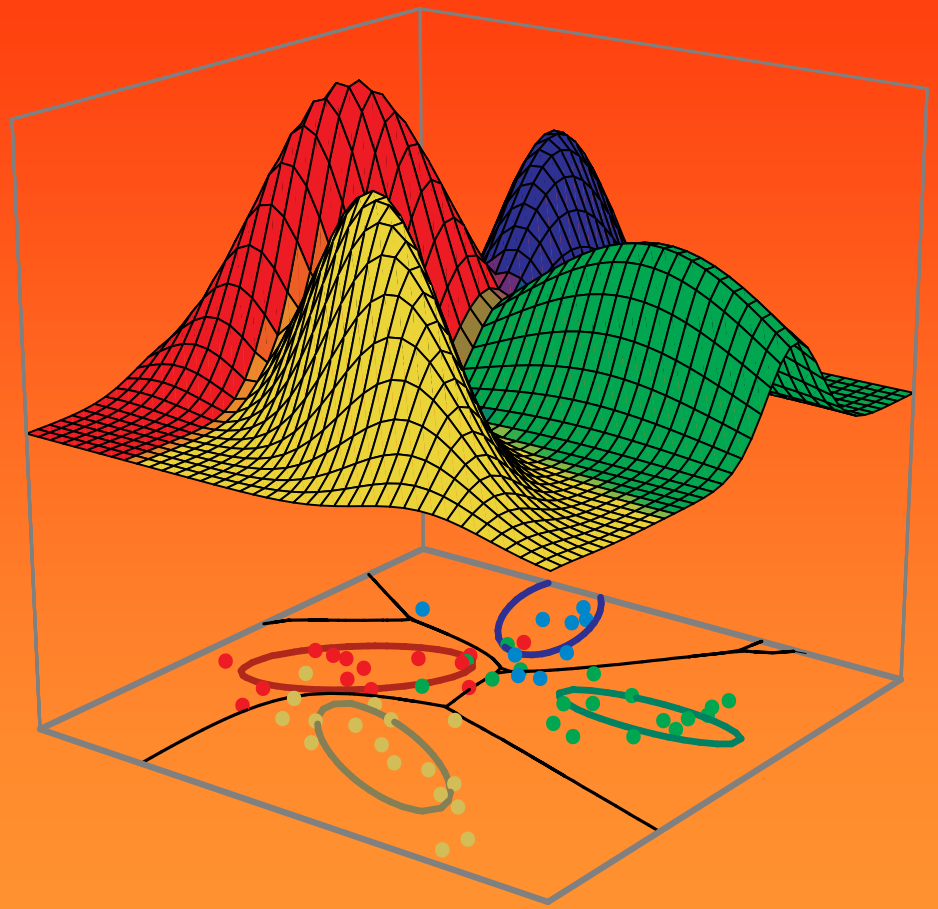


Pattern Classification

All materials in these slides were taken from Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

Chapter 2 (part 3)
Bayesian Decision Theory
(Sections 2-6,2-9)

- Discriminant Functions for the Normal Density
- Bayes Decision Theory – Discrete Features

Discriminant Functions for the Normal Density

- We saw that the minimum error-rate classification can be achieved by the discriminant function

  $g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$

- Case of multivariate normal

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Case $\Sigma_i = \sigma^2 \mathrm{I}$ (I stands for the identity matrix)

- What does "$\Sigma_i = \sigma^2 \mathrm{I}$" say about the dimensions?
- What about the variance of each dimension?

Note : both $|\Sigma_i|$ and $(d/2)\ln\pi$ are independent of i in

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Thus we can simplify to :

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

where $\|\cdot\|$ denotes the Euclidean norm

- We can further simplify by recognizing that the quadratic term $\boldsymbol{x}^t\boldsymbol{x}$ implicit in the Euclidean norm is the same for all *i*.

$$g_i(x) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \text{ (linear discriminant function)}$$

where :

$$\mathbf{w}_i = \frac{\boldsymbol{\mu}_i}{\sigma^2};\ w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

(ω_{i0} is called the threshold for the ***i***th category!)

- A classifier that uses linear discriminant functions is called "a linear machine"

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

$$g_i(x) = g_j(x)$$

The equation can be written as:

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$

- The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

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

always orthogonal to the line linking the means!

$$\textit{if} \quad P(\omega_i) = P(\omega_j) \ \text{ then } \ \mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$

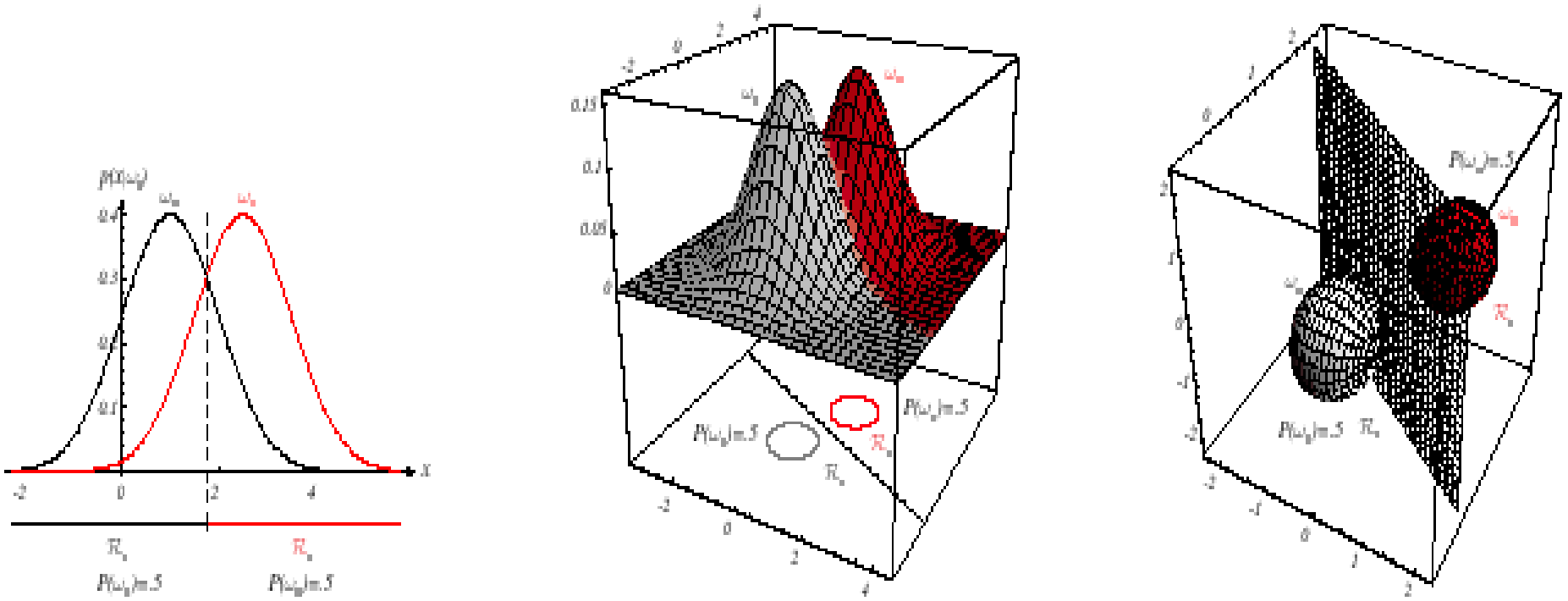
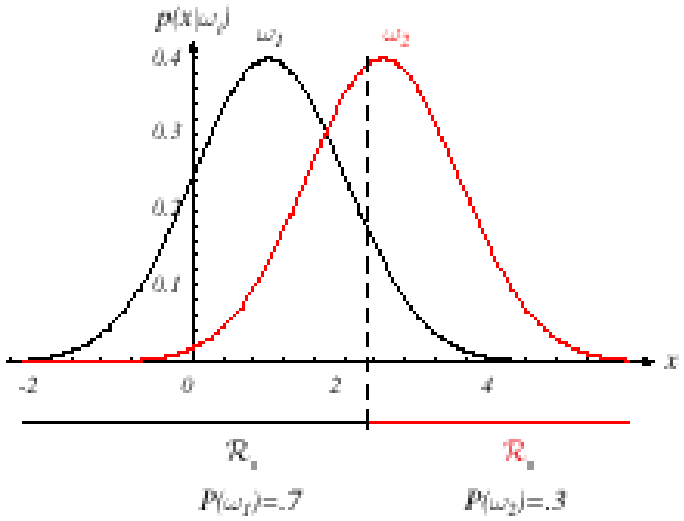
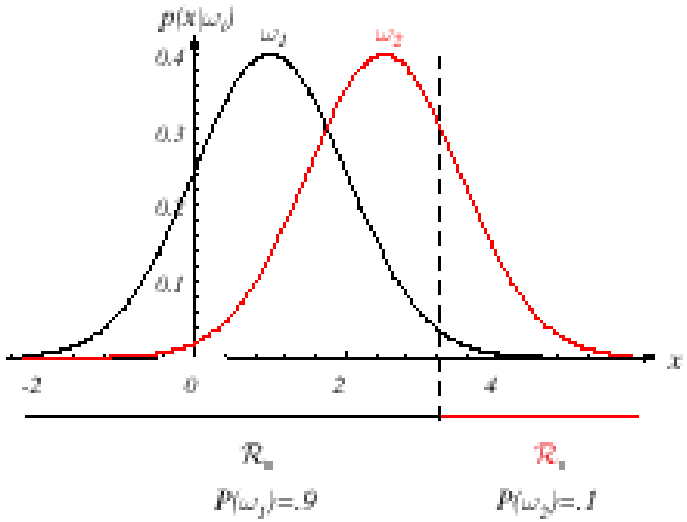
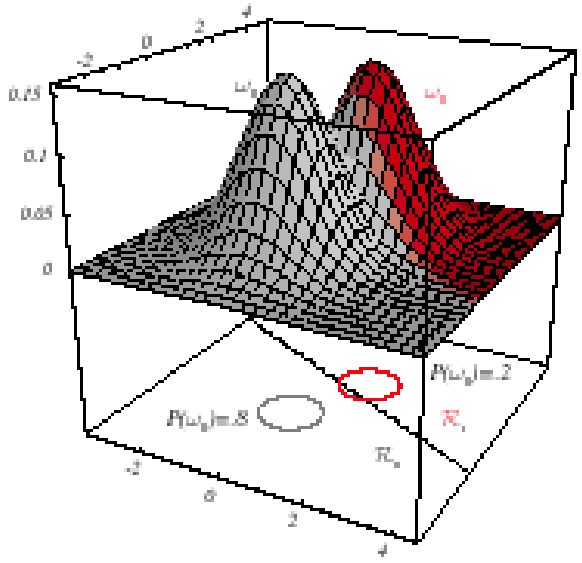
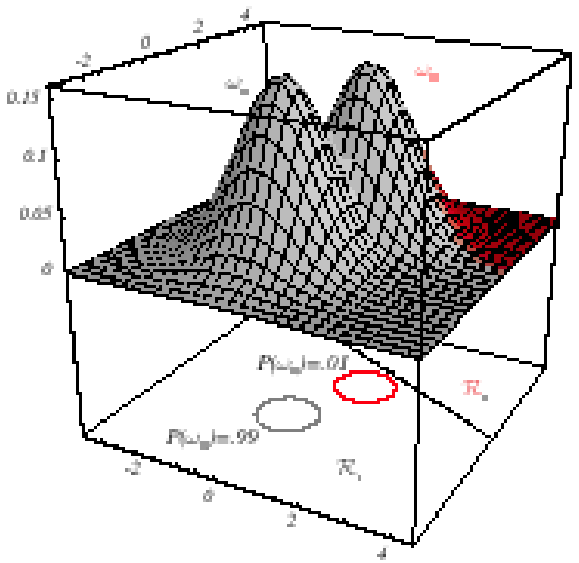


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.


p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
R1
R2
P(ω1)=.7
P(ω2)=.3


p(x|ωj)
ω1
ω2
0.4
0.3
0.2
0.1
x
-2
0
2
4
R1
R2
P(ω1)=.9
P(ω2)=.1


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8
R1
R2


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99
R1
R2

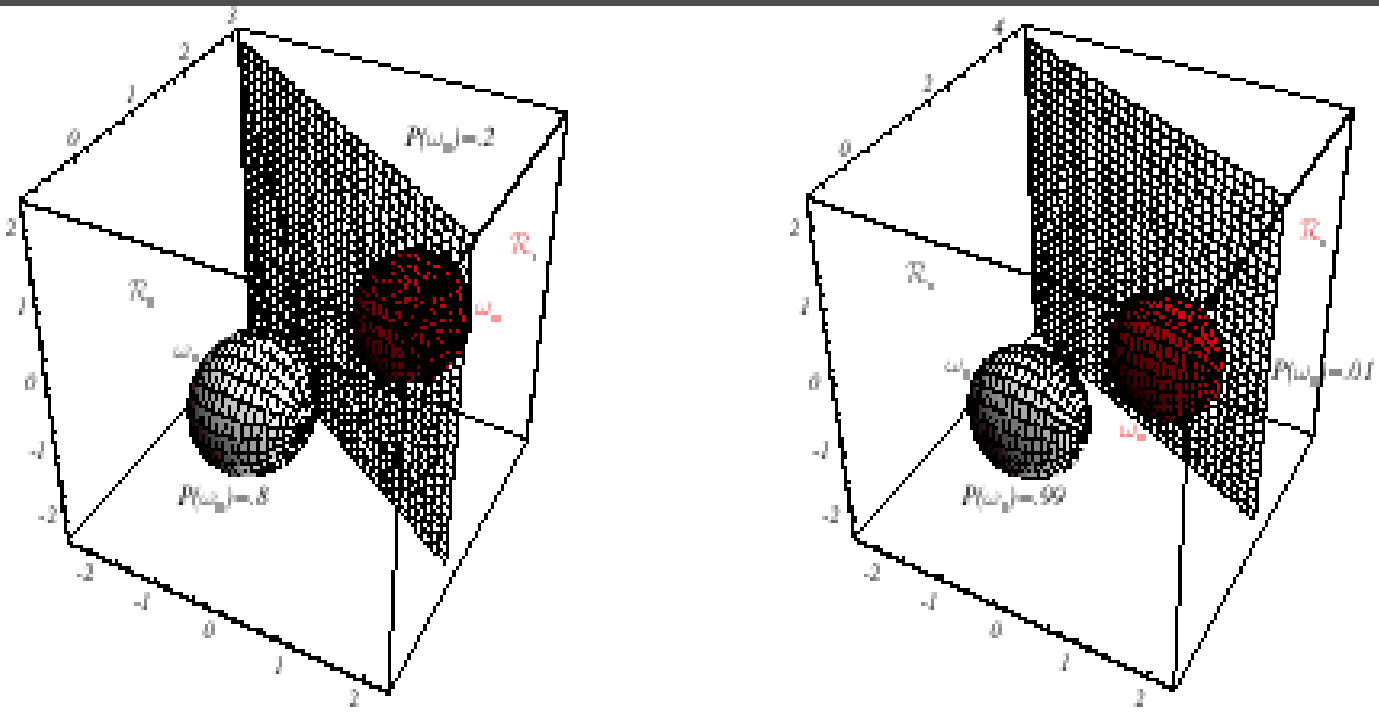


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

  Hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

Has the equation

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0$$

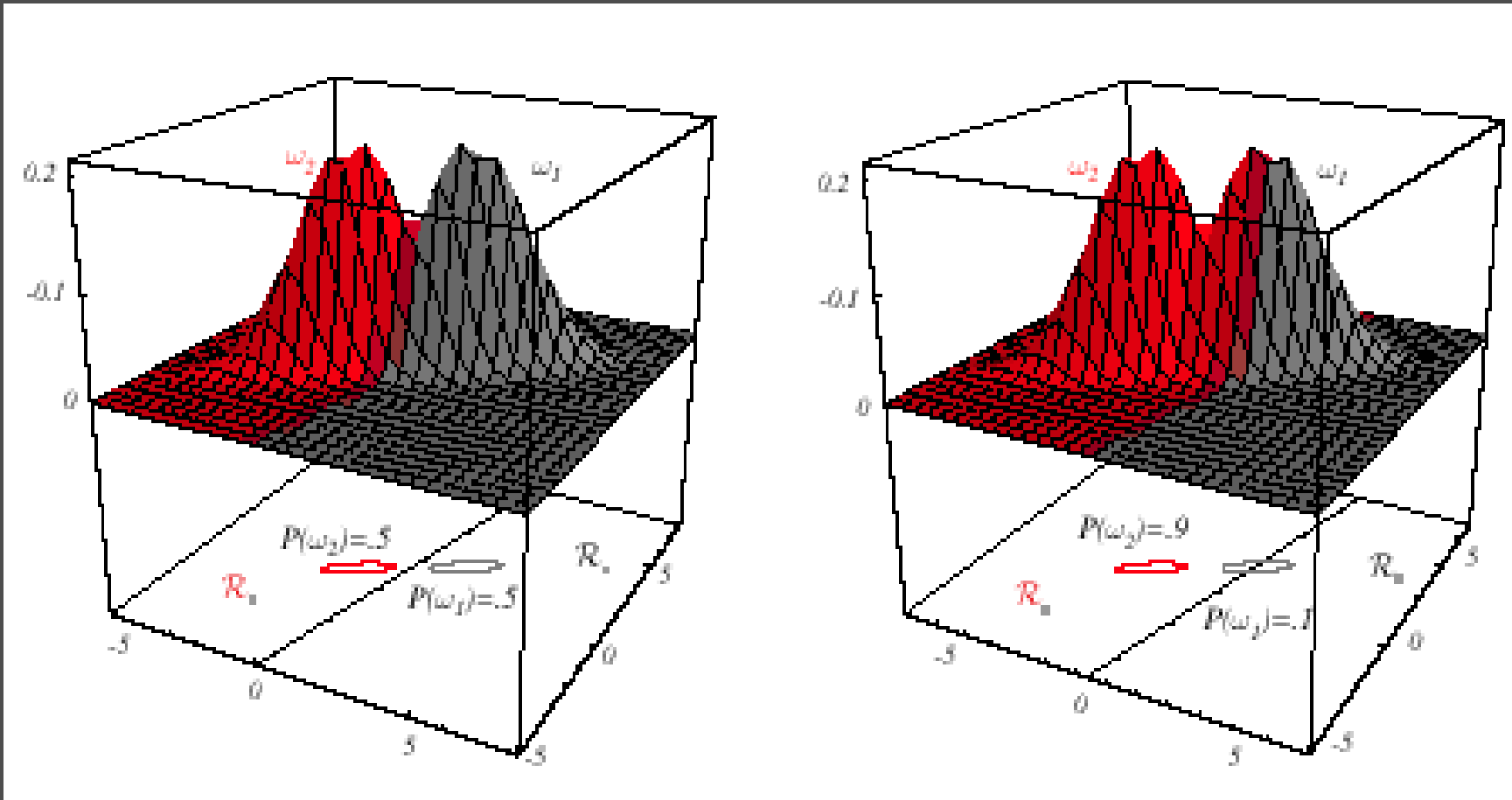
Where

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

and

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}.(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

(the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line between the means!)


0.2
-0.1
0
ω2
ω1
P(ω2)=.5
P(ω1)=.5
R1
R2
-5
0
5
P(ω2)=.9
P(ω1)=.1

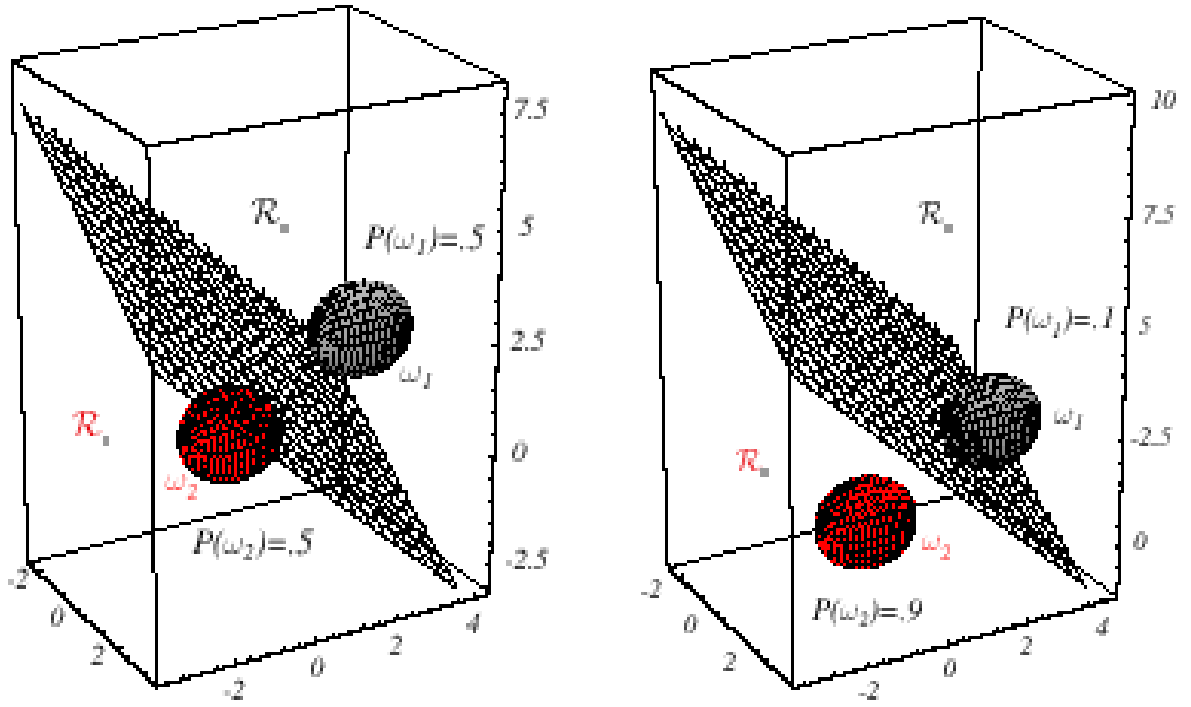


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Case Σ_i = arbitrary
  - The covariance matrices are different for each category

$$g_i(x) = x^t W_i x + w_i^t x = w_{i0}$$

where :

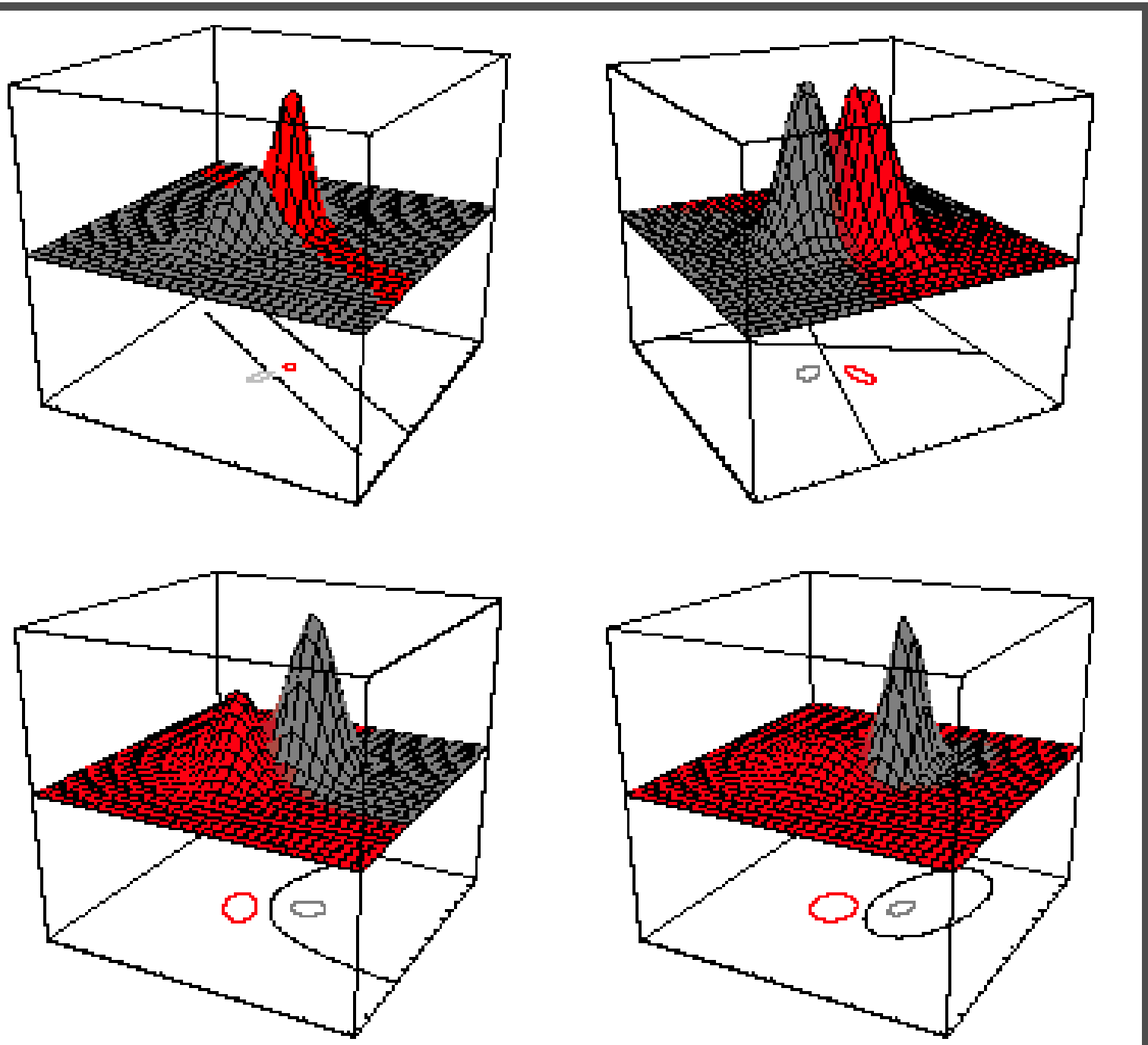
$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

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

The decision surfaces are hyperquadratics

(Hyperquadrics are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)



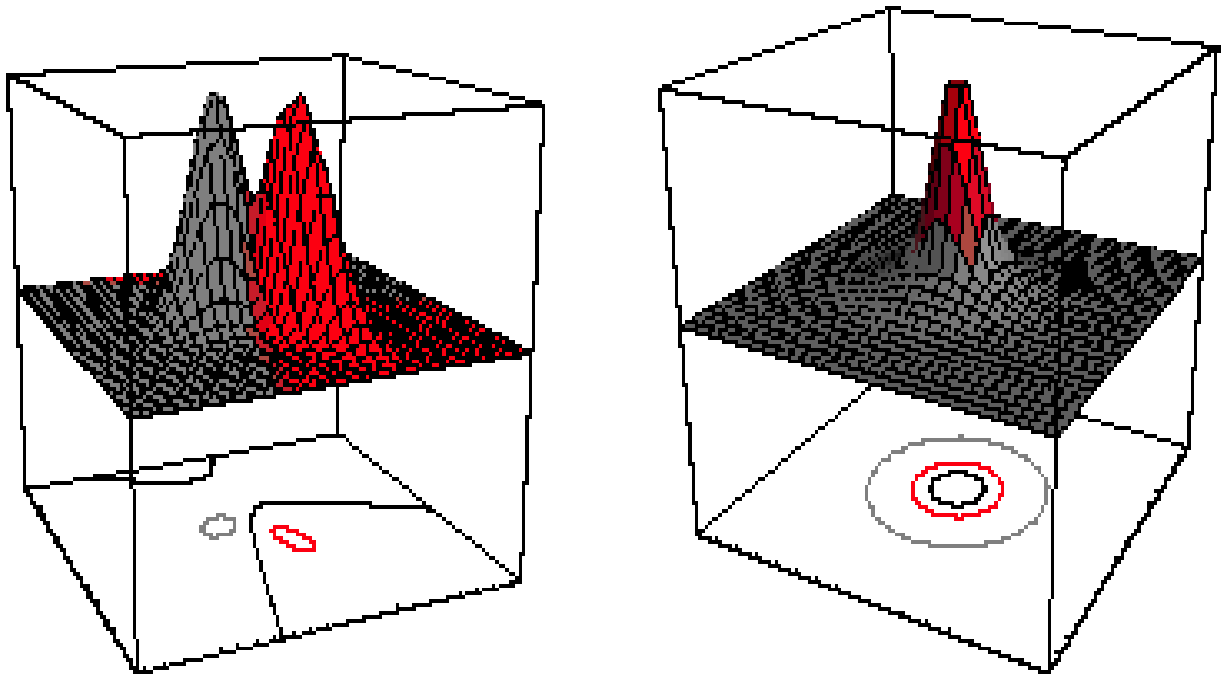


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Decision Theory – Discrete Features

- Components of x are binary or integer valued, x can take only one of m discrete values

$$v_1, v_2, \ldots, v_m$$

➔ concerned with probabilities rather than probability densities in Bayes Formula:

$$P(\omega_j \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \omega_j) P(\omega_j)}{P(\mathbf{x})}$$

where

$$P(\mathbf{x}) = \sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j) P(\omega_j)$$

Bayes Decision Theory – Discrete Features

- Conditional risk is defined as before: R(α|**x**)
- Approach is still to minimize risk:

$$\alpha^* = \arg\min_i R(\alpha_i \mid \mathbf{x})$$

Bayes Decision Theory – Discrete Features

- Case of independent binary features in 2 category problem

  Let $x = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1)$$
$$q_i = P(x_i = 1 \mid \omega_2)$$

Bayes Decision Theory – Discrete Features

- Assuming conditional independence, $P(\mathbf{x}|\omega_i)$ can be written as a product of component probabilities:

$$P(\mathbf{x}\,|\,\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

and

$$P(\mathbf{x}\,|\,\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$$

yielding a likelihood ratio given by :

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

Bayes Decision Theory – Discrete Features

- Taking our likelihood ratio

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

and plugging it into Eq. 31

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)} + \ln \frac{p(\omega_1)}{p(\omega_2)}$$

yields:

$$g(\mathbf{x}) = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1-x_i) \ln \frac{1-p_i}{1-q_i} \right] + \ln \frac{p(\omega_1)}{p(\omega_2)}$$

- The discriminant function in this case is:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

where :

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\ldots,d$$

and :

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

$$\textbf{decide } \omega_1 \textbf{ if } g(x) > 0 \textbf{ and } \omega_2 \textbf{ if } g(x) \leq 0$$