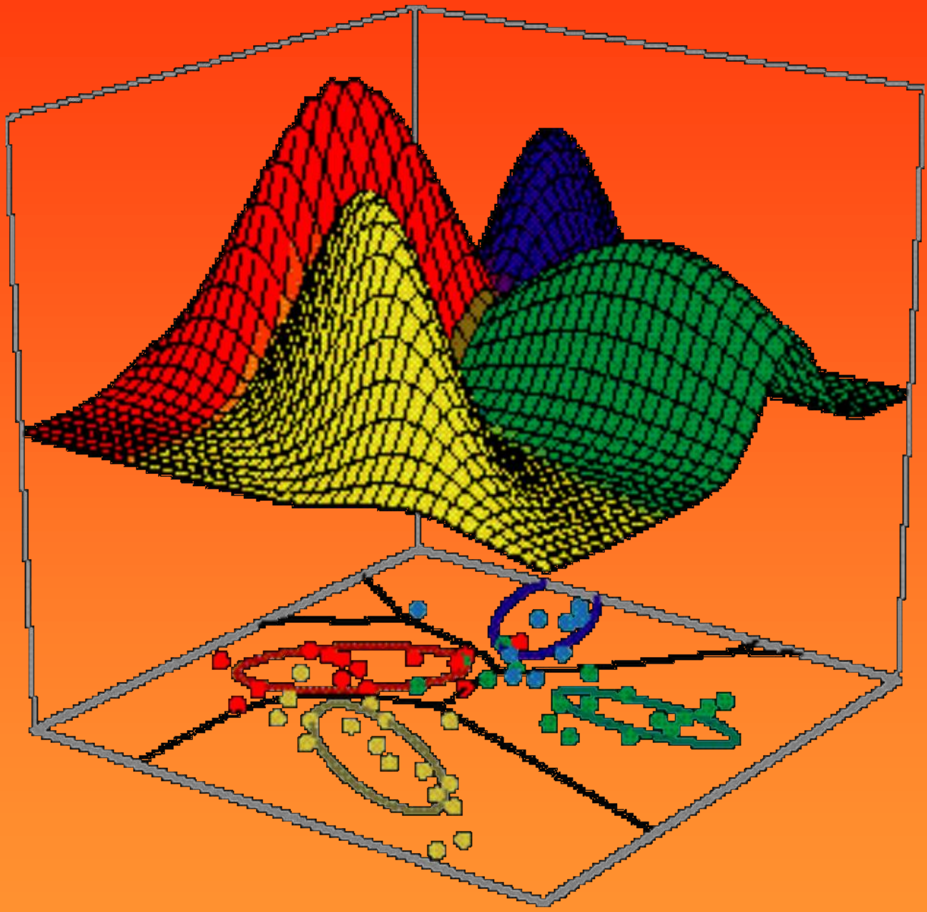


Pattern Classification

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

Normal distribution

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

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

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_i\right| + \ln P(\omega_i) \qquad (49)$$

- Case(1): $\Sigma_i = \sigma^2 I$ (I stands for the identity matrix)
  - Its determinant and inverse can be easily found
  - $|\Sigma_i| = \sigma^{2d}$, $\Sigma_i^{-1} = (1/\sigma^2) I$
- Because both $|\Sigma_i|$ and the $(d/2)\ln 2\pi$ term in Eq. 49 are independent of i, they can be ignored.

- Thus we obtain simple discriminant functions

$$g_i(x) = -\frac{\| x - \mu_i \|^2}{2\sigma^2} + \ln P(\omega_i) \qquad (51)$$

Where $\| x - \mu_i \|^2 = (x - \mu_i)^t (x - \mu_i)$

$$g_i(x) = (-1/2\sigma^2)[x^t x - 2\mu^t_i x + \mu^t_i \mu] + \ln P(\omega_i) \qquad (52)$$

This appears to be quadratic, but ($x^t x$) is same for all i and hence can be ignored. We get,

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

where:

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

(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:

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

- These hyperplanes will be orthogonal to the line joining the respective means
- If the priors for the two classes are same, then the hyperplane passes exactly mid-way between the means.

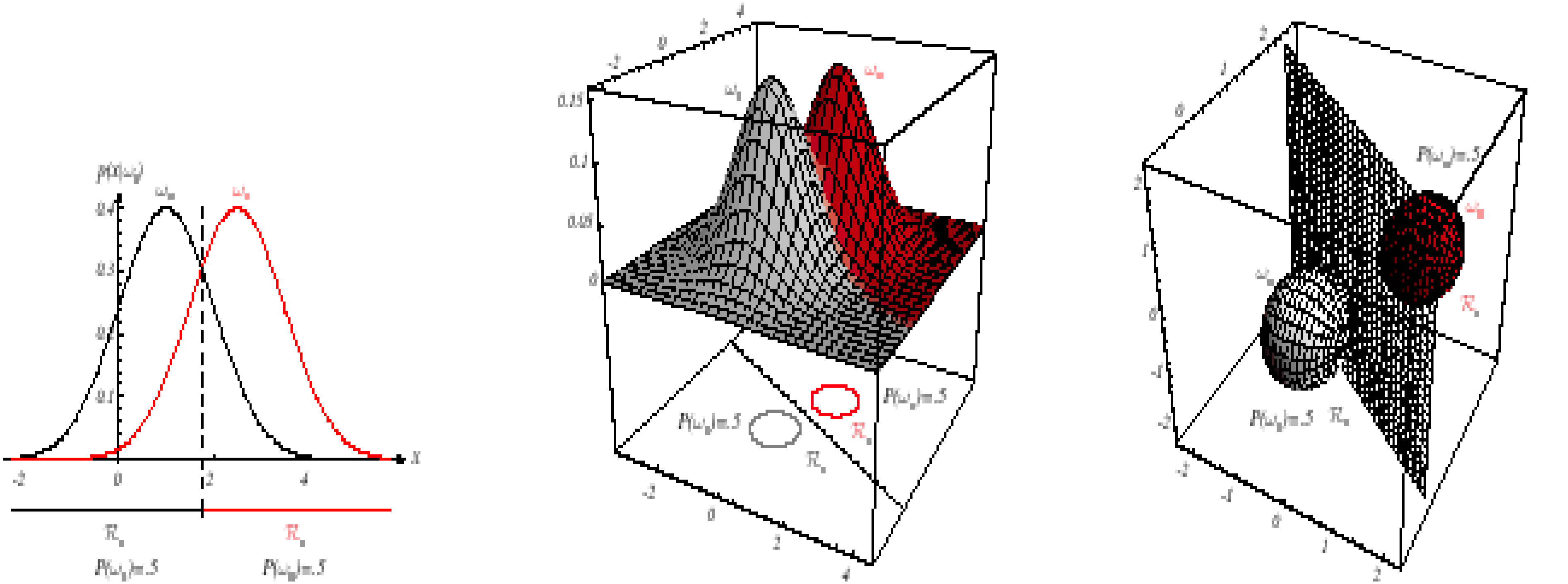
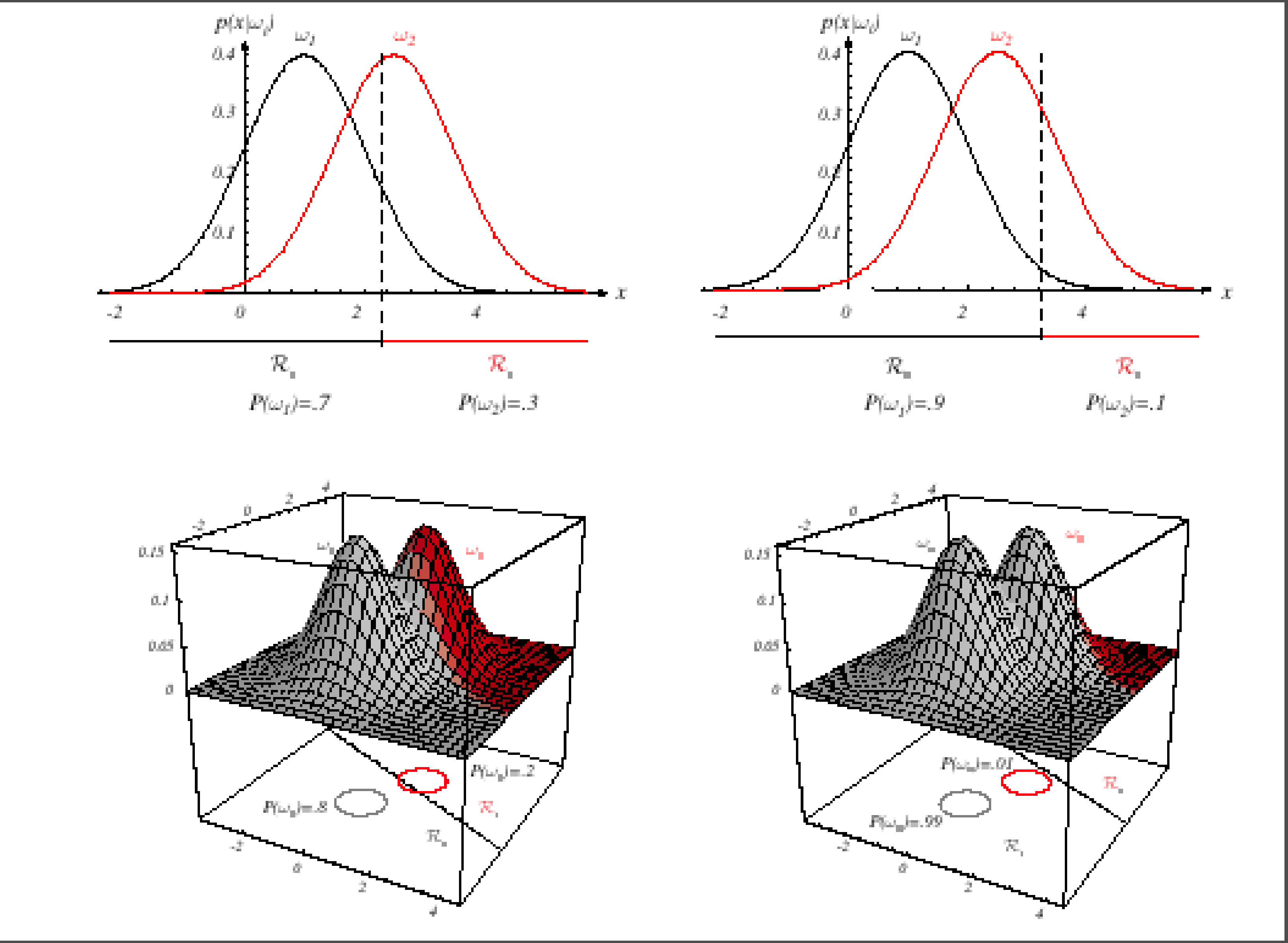


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, 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.

- If the priors are unequal then the hyperplane shifts accordingly.



$p(x|\omega_i)$
ω_1
ω_2
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.7$
$P(\omega_2)=.3$
$P(\omega_1)=.9$
$P(\omega_2)=.1$
$P(\omega_1)=.8$
$P(\omega_2)=.2$
$P(\omega_1)=.99$
$P(\omega_2)=.01$

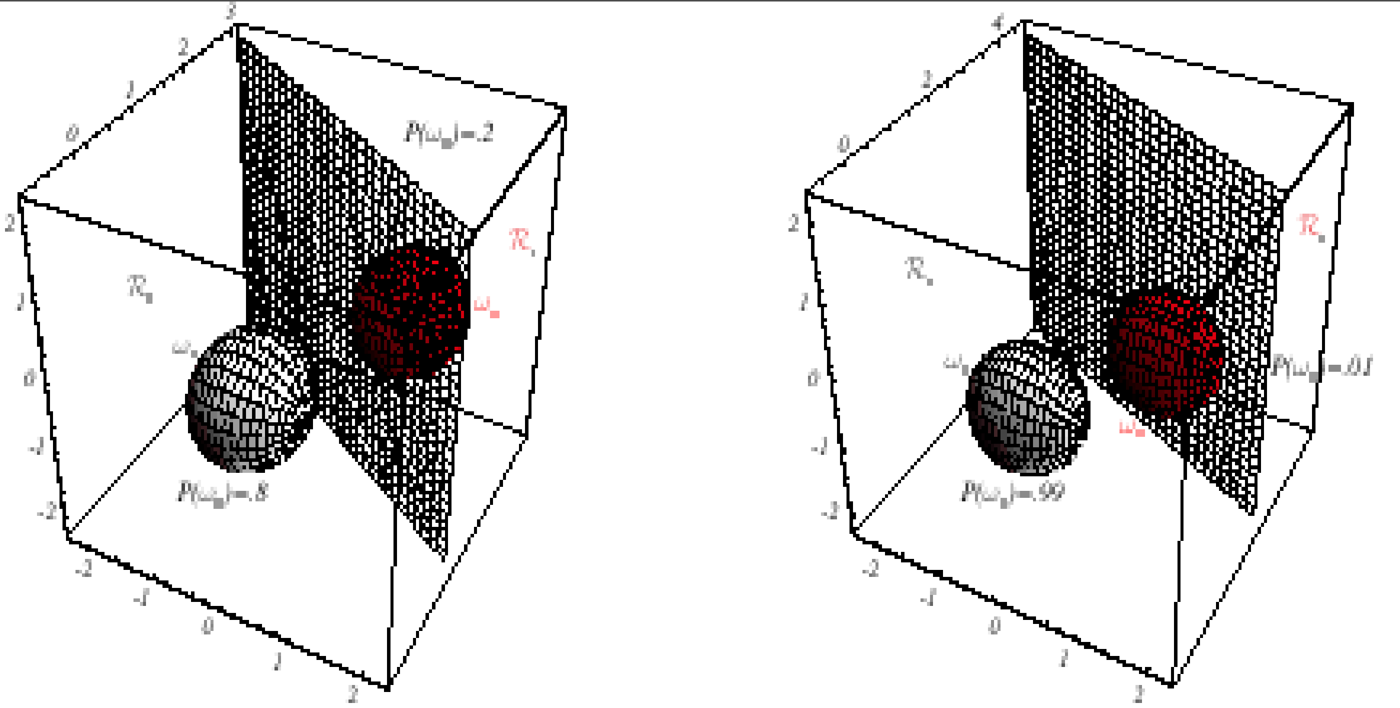


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

Minimum-distance classifier

- If prior probabilities are same for all classes, then *ln P(ω_i)* term is unimportant.
- In this case the classifier is a simple one:
  To classify x, measure the squared Euclidean distance of x with mean of each class i, that is find $|| x - \mu_i ||^2$ and classify to that class for which it is minimum.
- This classifier is called the minimum-distance classifier.

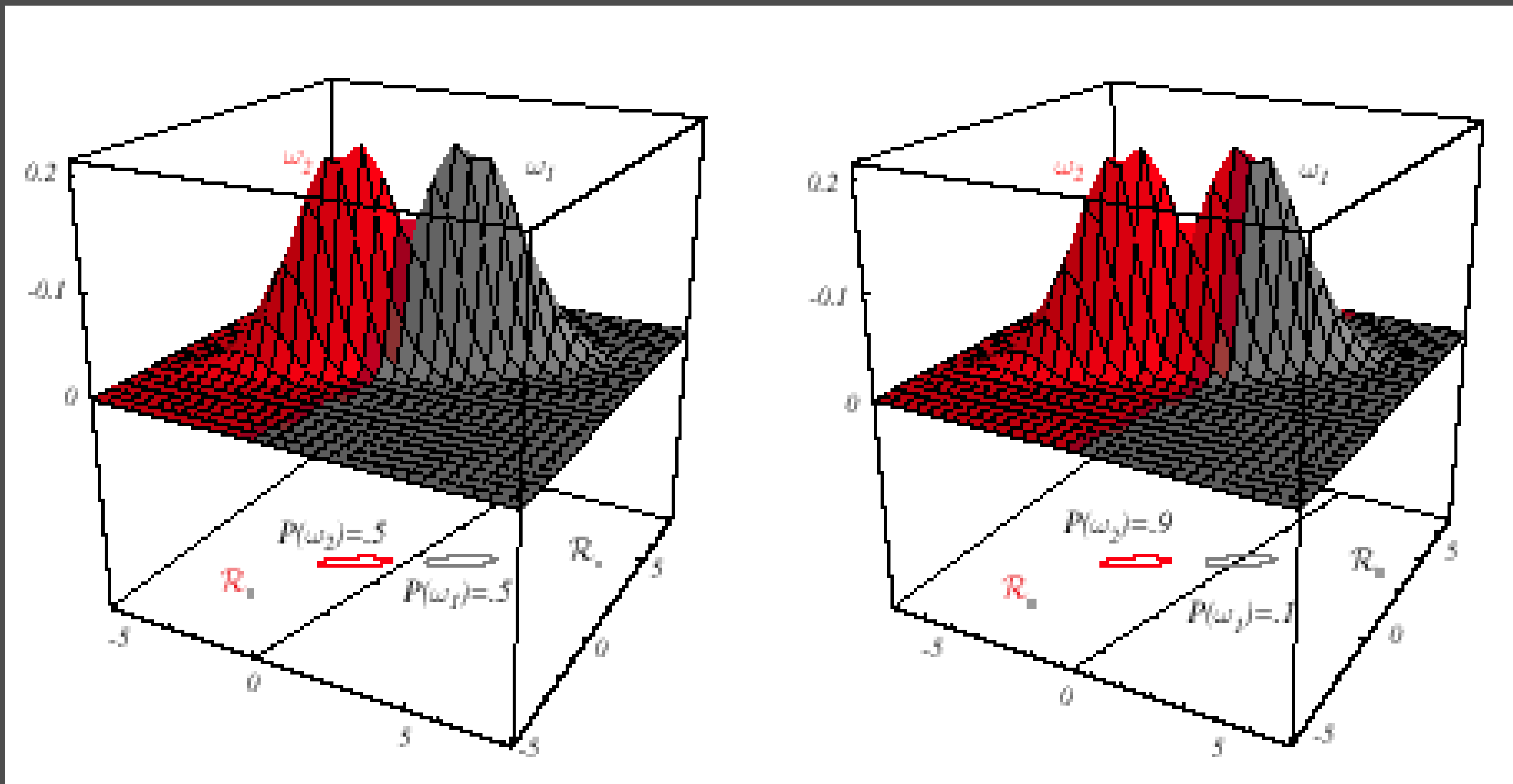
- Case (2) $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

- Because both $|\Sigma_i|$ and the $(d/2)\ln 2\pi$ term in Eq. 49 are independent of i, they can be ignored.

- $g_i(x) = (-1/2)\ \boxed{(x - \mu_i)^t \Sigma^{-1} (x - \mu_i)} + \ln P(\omega_i)$

Squared Mahalanobis distance

- The quadratic term is same for all classes and hence can be ignored.
- We get again linear discriminants.
- But the hyperplane separating the two classes need not be orthogonal to the line joining their means.
- If the priors are equal then the classifier is again a simple one: Find squared Mahalanobis distance with all of the means and classify according the minimum distance.


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

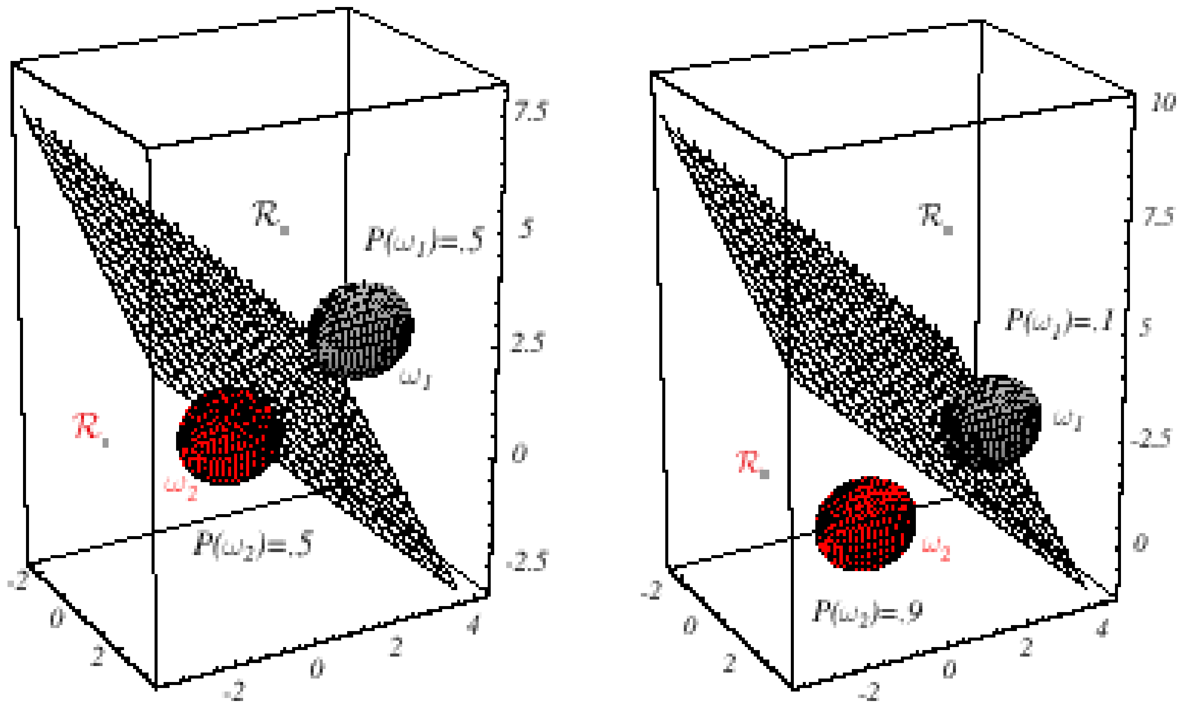


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda, 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

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

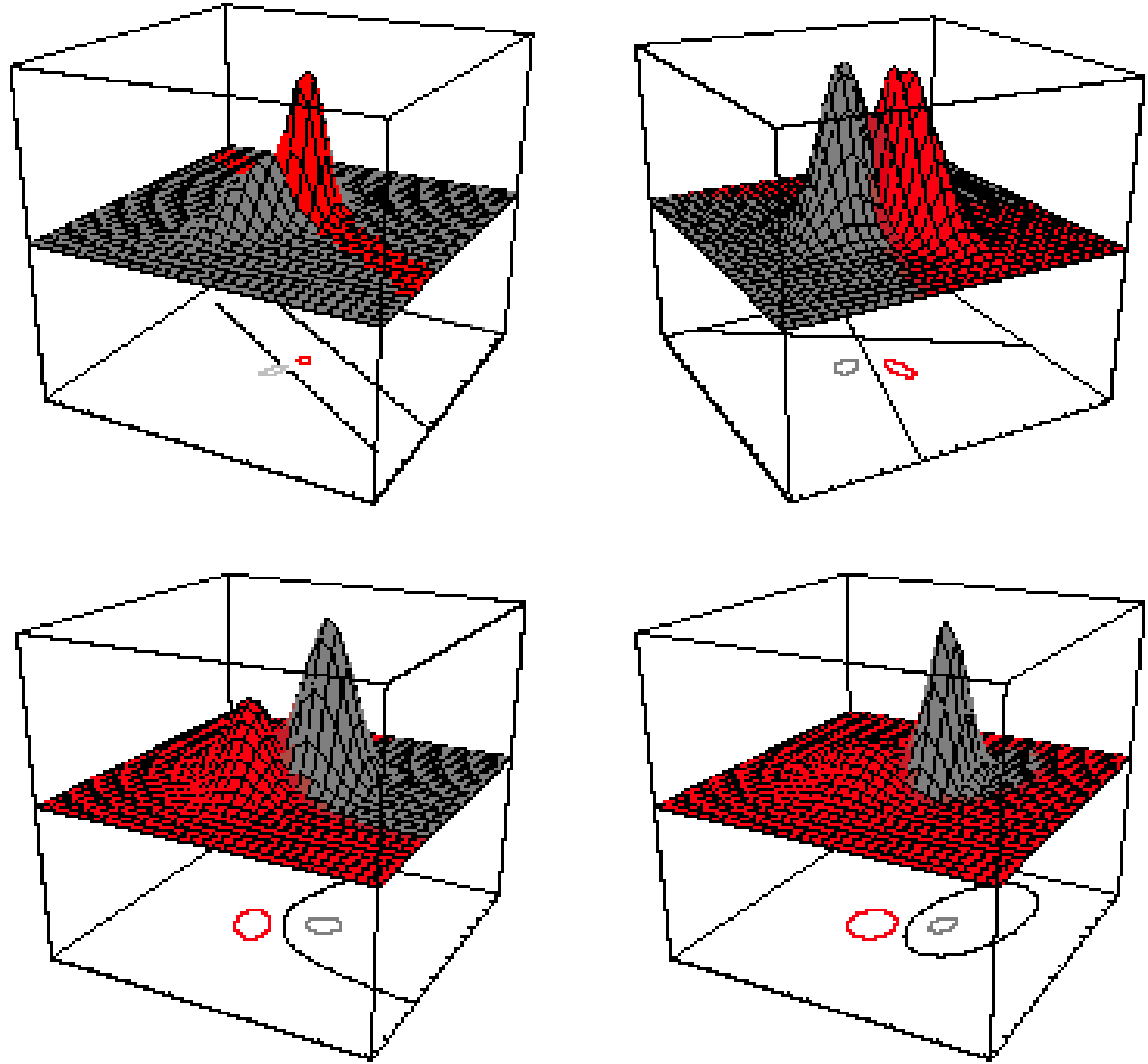
where :

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

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

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

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



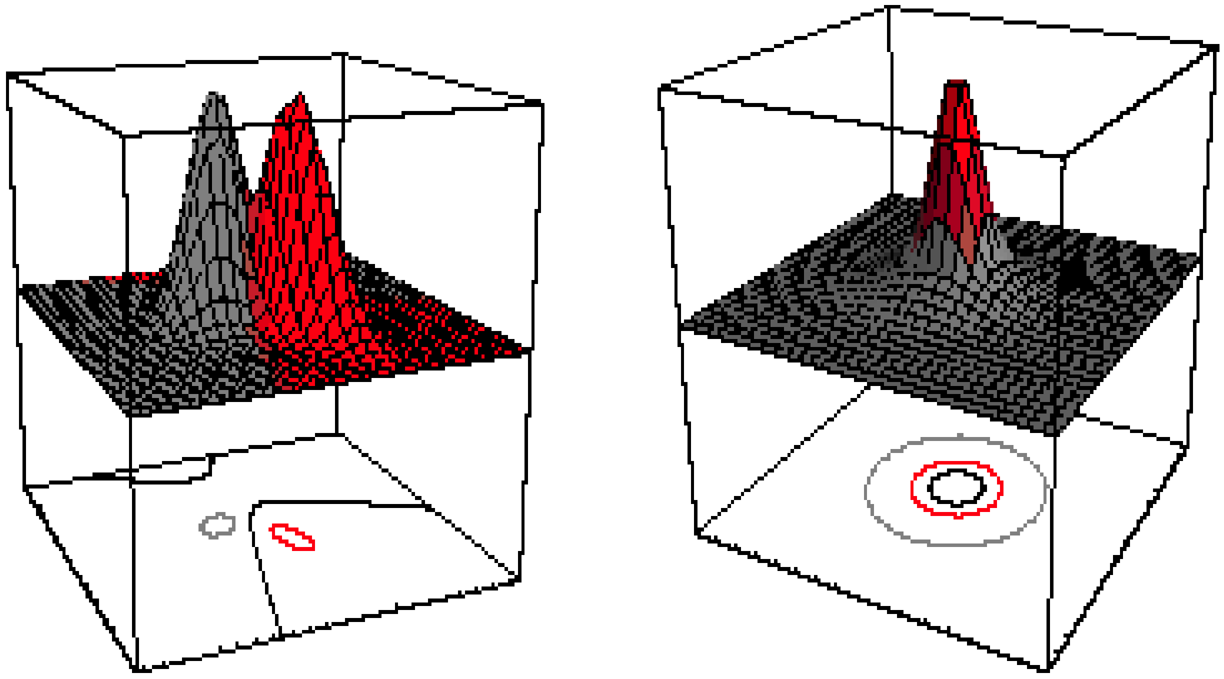


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Next class

- Bayes decision theory with discrete features.