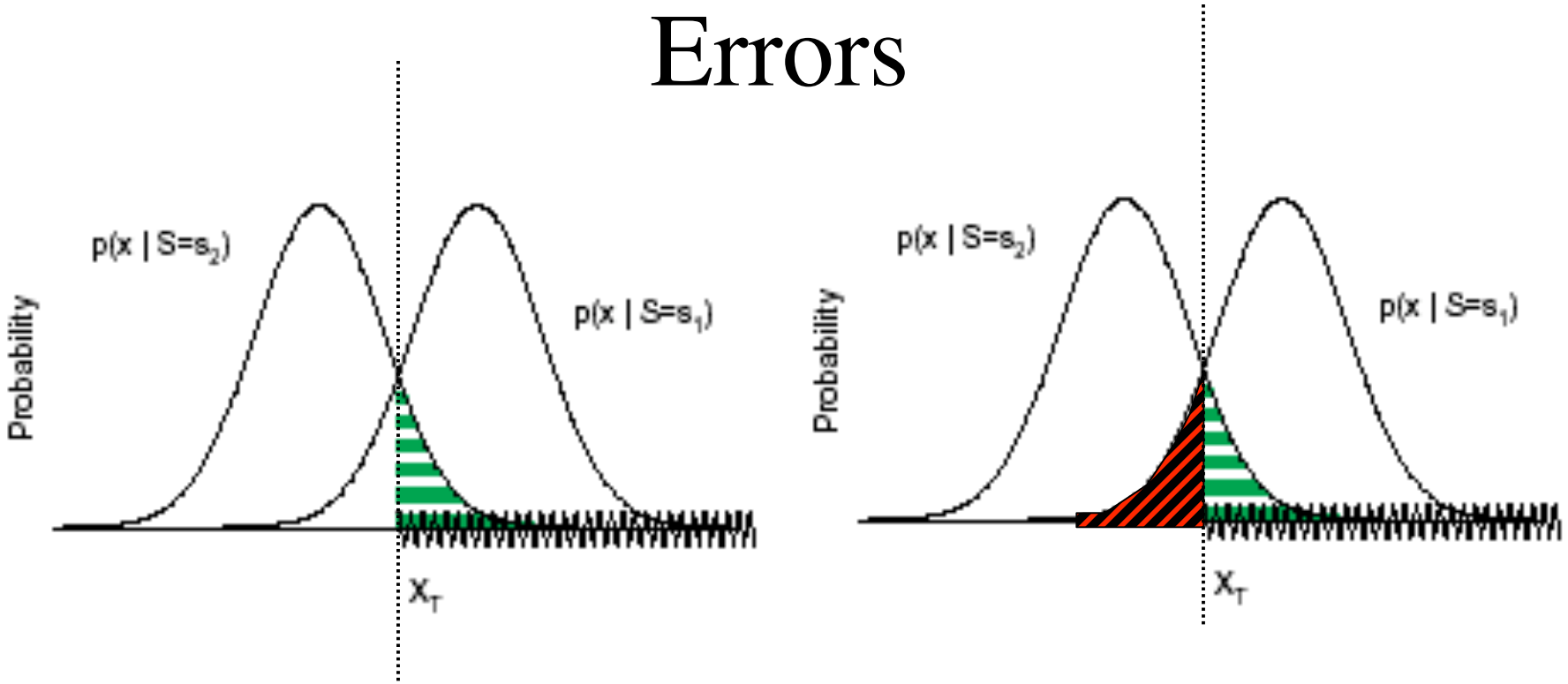
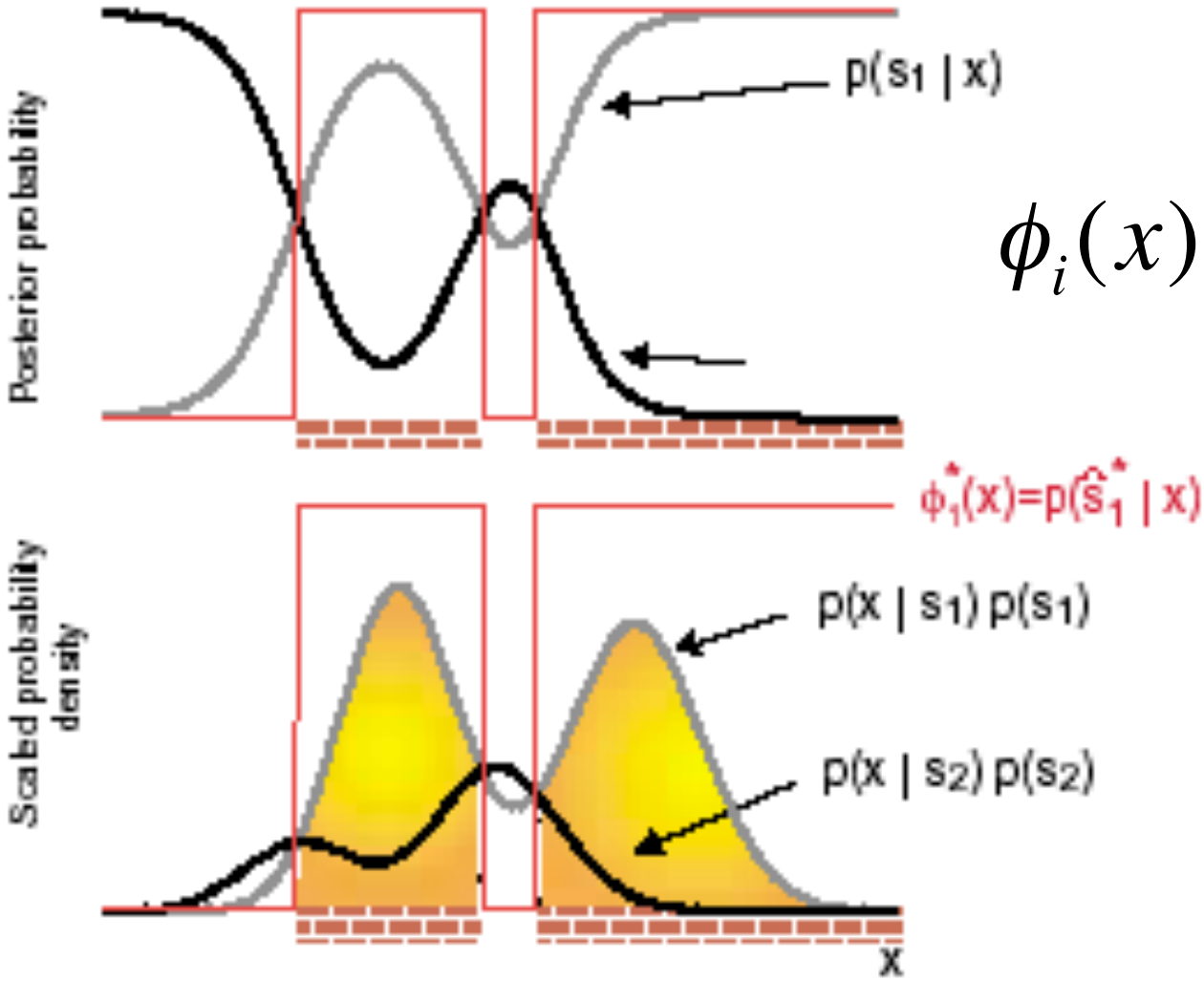
Errors



- Pick ω_1 when ω_0 is true.
  - False Alarm (rate)
  - False Positive (rate)
  - Type I error
- Pick ω_0 when ω_1 is true.
  - Miss (rate)
  - False Negative (rate)
  - Type II error

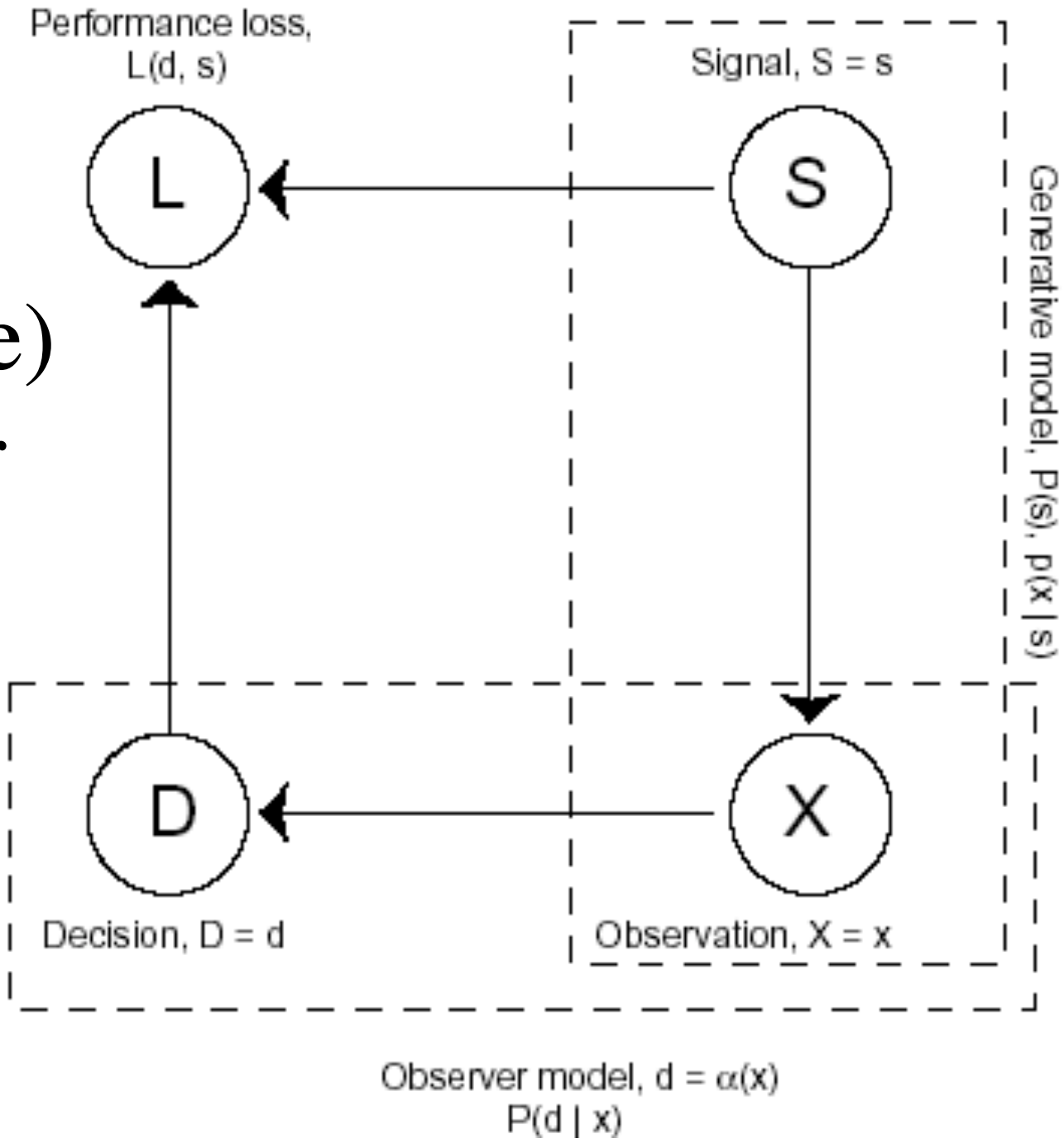
Indicator Variables

$$\phi_i(x) = \begin{cases} 1 & \text{if } g_i(x) > g_j(x) \quad \forall i \neq j \\ 0 & \text{otherwise} \end{cases}$$



$$\phi_i(x) = p(\alpha_1 \mid x)$$

Graphical (Generative) Model for Decision Theory



Computing Expected Success

Choose :

$$p(\alpha_i \mid x) = \phi_i(x) = \mathbf{1}\left(g_i(x) > g_j(x)\ \forall\ i \neq j\right)$$

- Decision Rule

$$P(\alpha_i \mid \omega_i) = \int_{-\infty}^{\infty} p(\alpha_i \mid x) p(x \mid \omega_i) dx$$

- Expected success

$$\Re_i = \left\{x : \phi_i(x) > 0\right\}$$

$$\Re_i = \left\{g_i(x) > g_j(x)\ \forall\ i \neq j\right\}$$

- Decision Region

$$P(\alpha_1 \mid \omega_1) = \int_{\Re_1} p(x \mid \omega_1) dx$$

- Expected success

ROC curves

- For two-state problems, the Bayes decision rule is: $\log \frac{P(x|\omega_1)}{P(x|\omega_2)} > T$

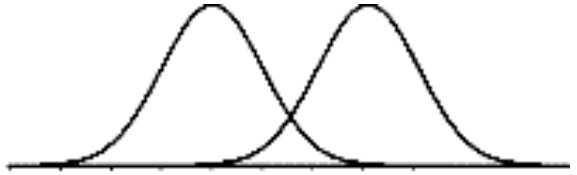
where T depends on the priors and the loss function.

- The observer may use the correct log-likelihood ratio, but have the wrong threshold.
- E.g. the observer's loss function choice may incorrectly penalize false negatives (trigger-shy) or false positives (trigger-happy).
- The ROC curve plots the proportion of correct responses (hits) against the false positives as the threshold T changes.
- Requires altering the loss function of observers by rewards (chocolate) and penalties (electric shocks).
- The ROC curve gives information which is independent of the observer's loss function.

ROC curves

Vary Criterion

d' = 1 (lots of overlap)



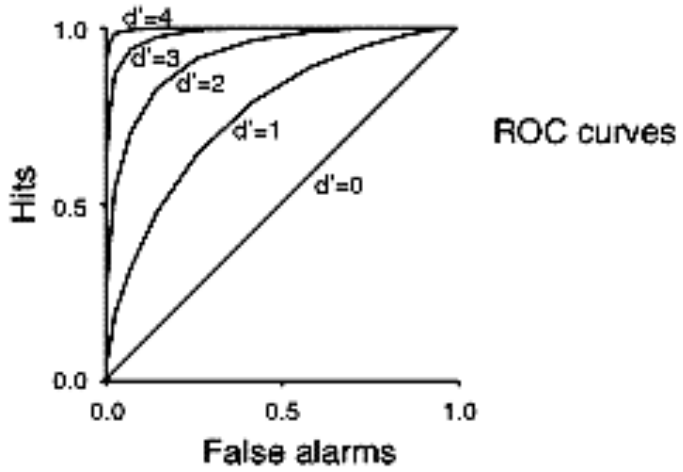
d' = 3 (not much overlap)

d' = 1

Hits = 97.5%
False alarms = 84%

Hits = 84%
False alarms = 50%

Hits = 50%
False alarms = 16%



ROC curves

The Normal Density

- Univariate density
  - Density which is analytically tractable
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

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

Where:

μ = mean (or expected value) of x

σ^2 = expected squared deviation or variance

Learning Classifiers: Parametric Approach

- Model class conditional densities using a formula with unknown parameters
- Learn the parameters from data
- Apply Bayesian decision theory to do subsequent classification.

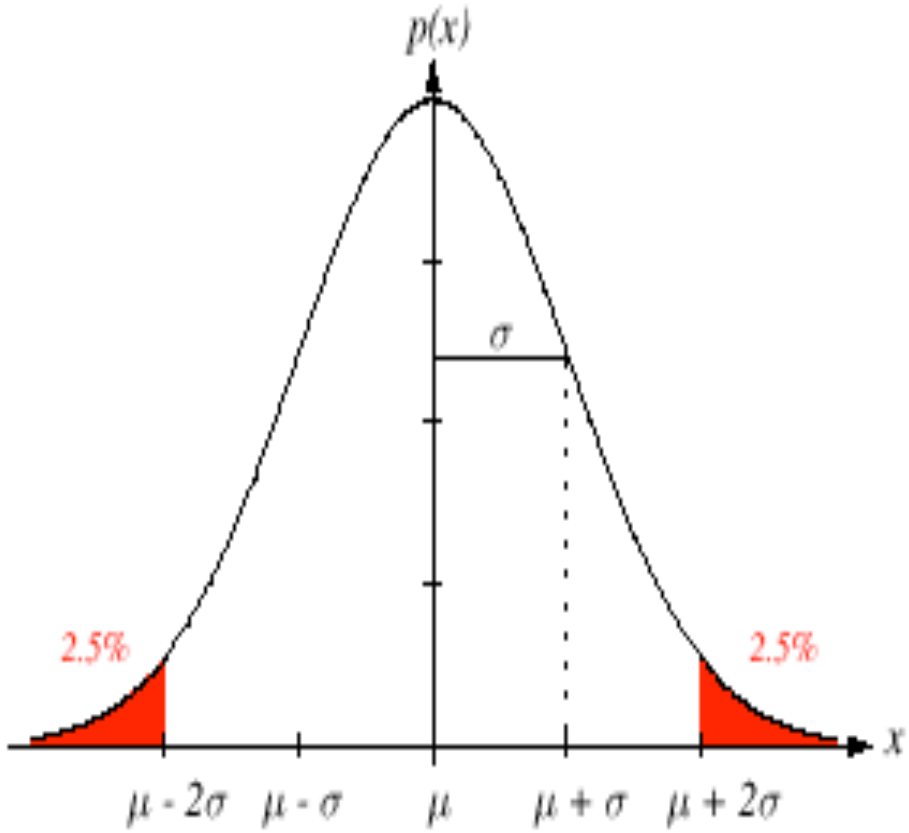


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- Multivariate density
  - Multivariate normal density in d dimensions is:

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

where:

$x = (x_1, x_2, \ldots, x_d)^t$ (t stands for the transpose vector form)

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

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

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

Special case $\Sigma_i=\Sigma$

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

$$g_i(x) - g_j(x) > 0$$

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

$$\left(-\frac{1}{2}(x-\mu_j)^t \Sigma^{-1}(x-\mu_j) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| + \ln P(\omega_j)\right)$$

$$\text{Now} \quad (x-\mu_i)^t \Sigma^{-1}(x-\mu_i) = x^t\Sigma^{-1}x - 2\mu_i^t\Sigma^{-1}x + \mu_i^t\Sigma^{-1}\mu_i$$

$$= \mu_i^t\Sigma^{-1}x - \mu_j^t\Sigma^{-1}x - \frac{1}{2}\mu_i^t\Sigma^{-1}\mu_i + \frac{1}{2}\mu_j^t\Sigma^{-1}\mu_j + \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$= w_i^t x - w_j^t x + w_{i0} - w_{j0}$$

- Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)

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

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

(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

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

$$g_i(x) - g_j(x)$$

$$= w_i^t x + w_{i0} - w_j^t x + w_{j0}$$

$$= \left(w_i^t - w_j^t\right)x + \left(w_{i0} - w_{j0}\right)$$

$$= \left(w_i^t - w_j^t\right)(x - x_0)$$

For Identity covariance case :

$$= \frac{1}{\sigma^2}\left(\mu_i^t - \mu_j^t\right)(x - x_0)$$

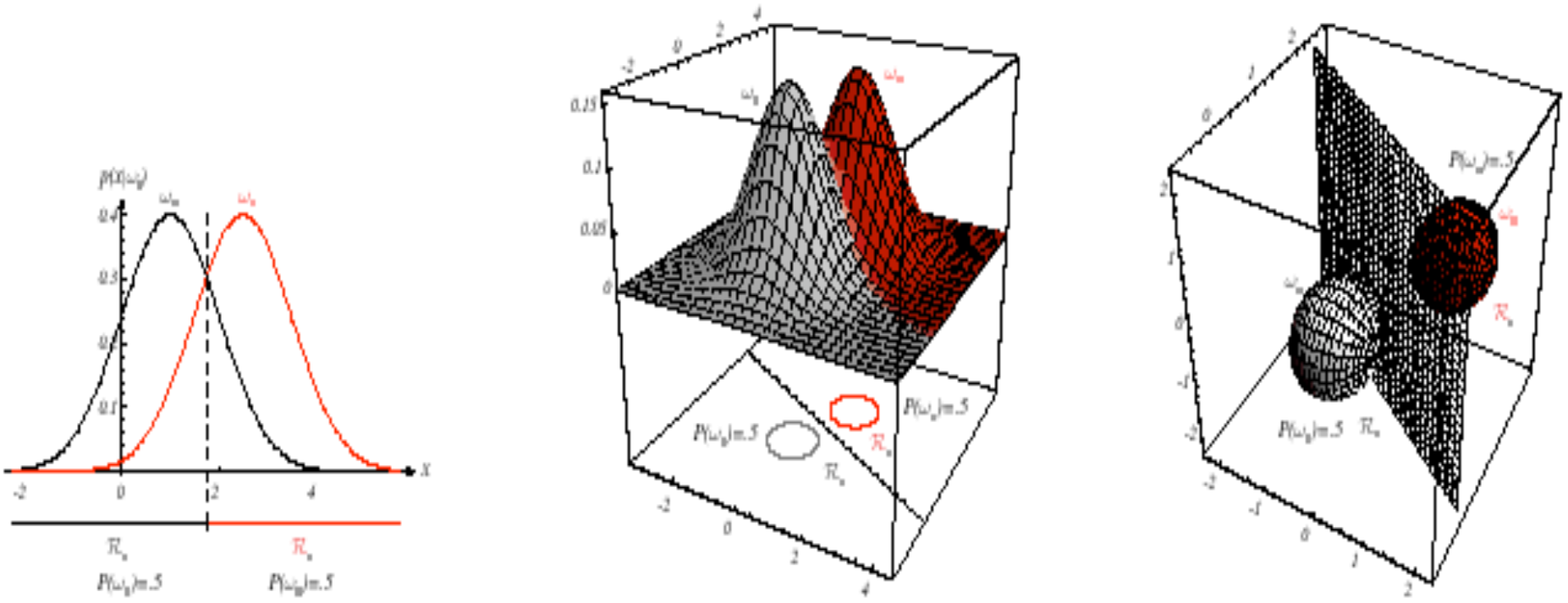


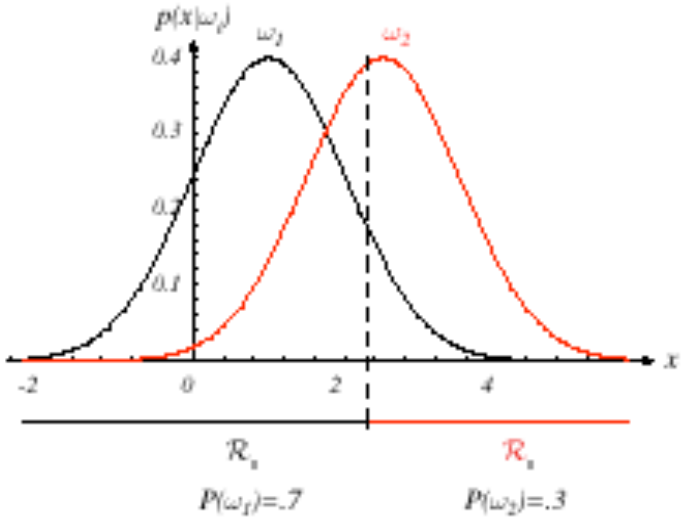
FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-,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

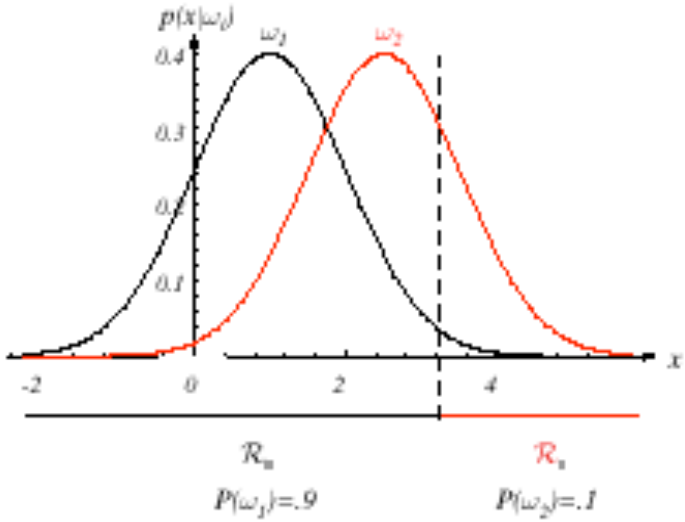
– The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$

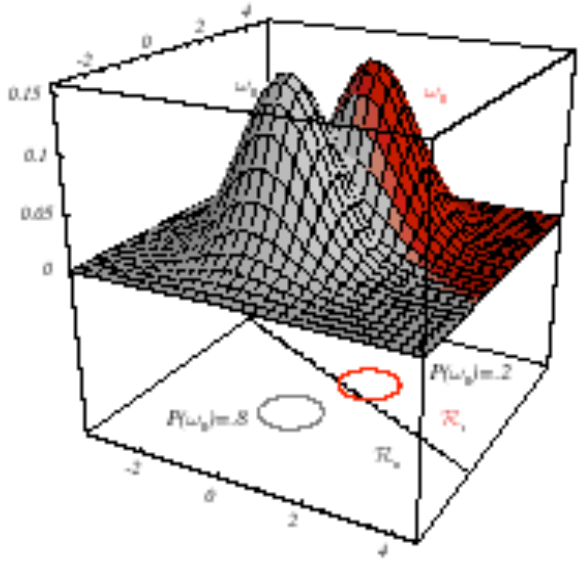
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

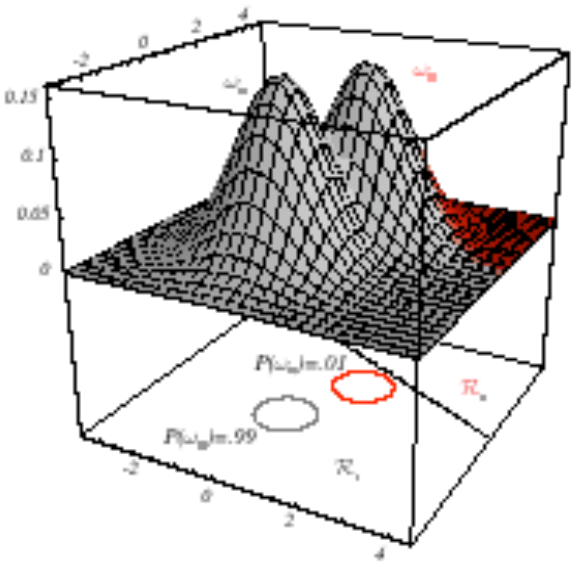
always orthogonal to the line linking the means!

$$if \;\; P(\omega_i) = P(\omega_j) \;\; then \;\; x_0 = \frac{1}{2}(\mu_i + \mu_j)$$


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


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


0.15
0.1
0.05
0
P(ω2)=.2
P(ω1)=.8


0.15
0.1
0.05
0
P(ω2)=.01
P(ω1)=.99

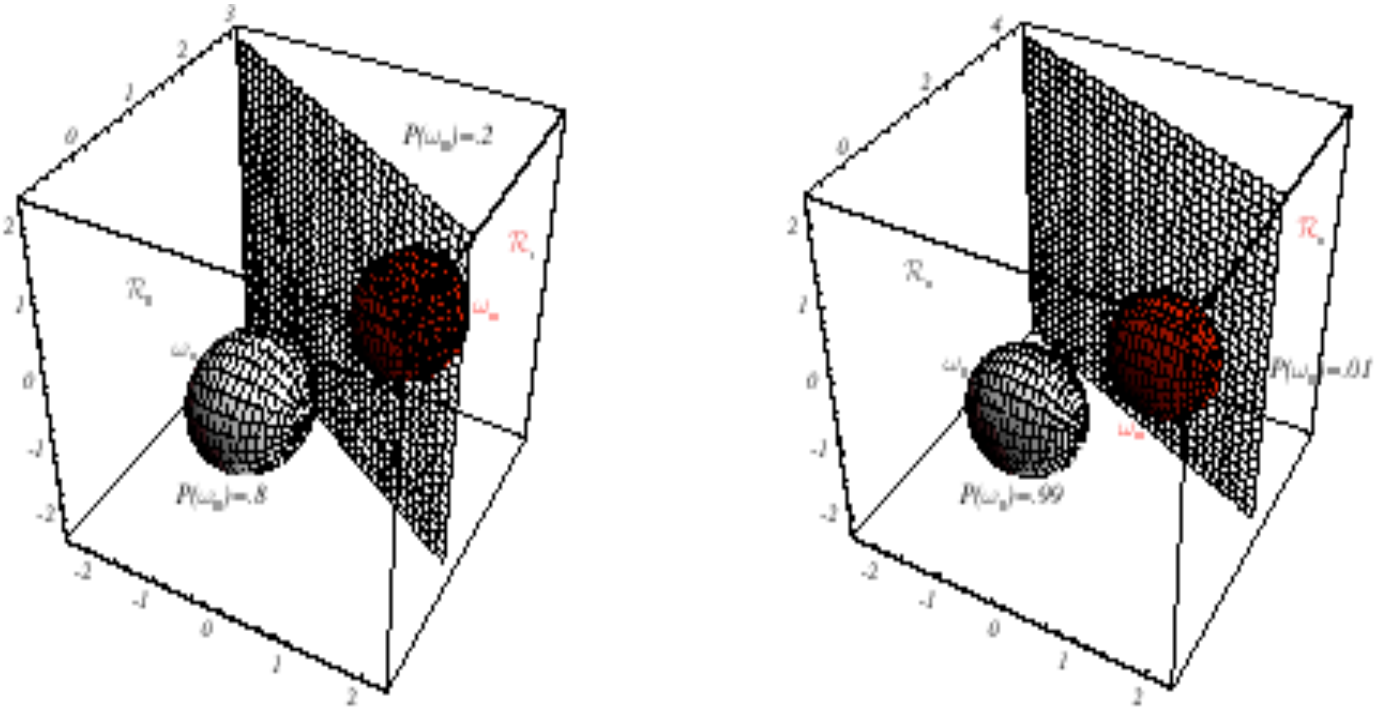


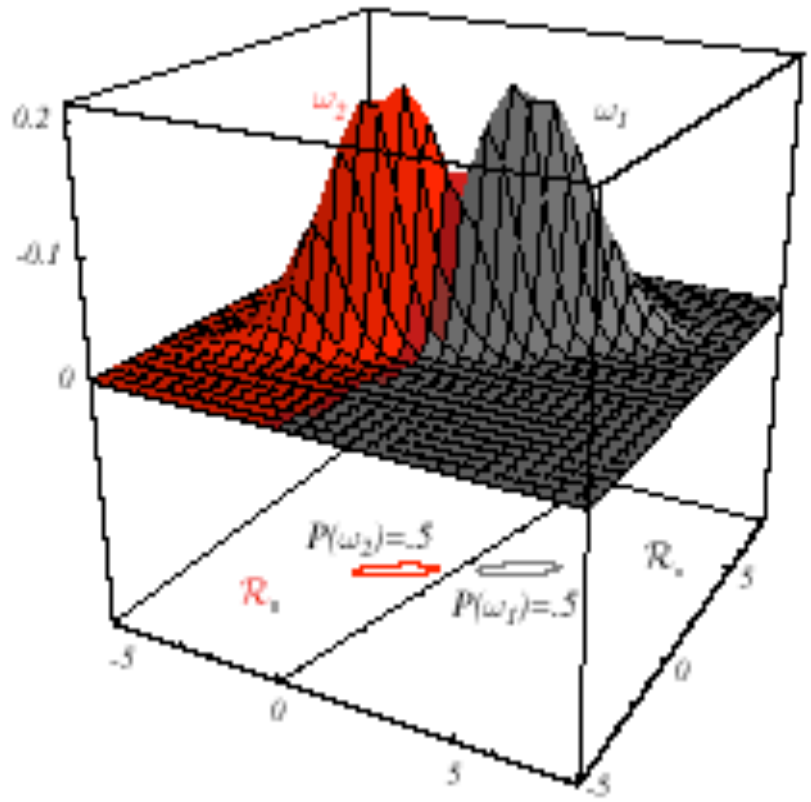
FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*.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

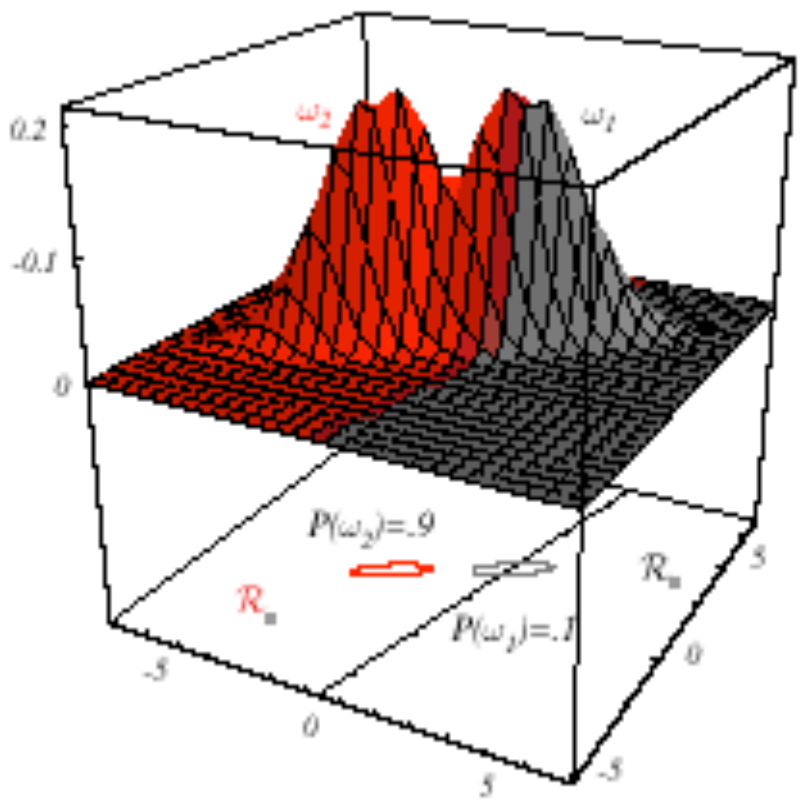
- Case $\Sigma_i = \Sigma$ (covariance of all classes are identical but arbitrary!)

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

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

(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
ω_2
ω_1
$P(\omega_2)=.5$
$P(\omega_1)=.5$
$\mathcal{R}_2$
$\mathcal{R}_1$
-5
0
5
-5
0
5


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)=.9$
$P(\omega_1)=.1$
$\mathcal{R}_2$
$\mathcal{R}_1$
-5
0
5
-5
0
5

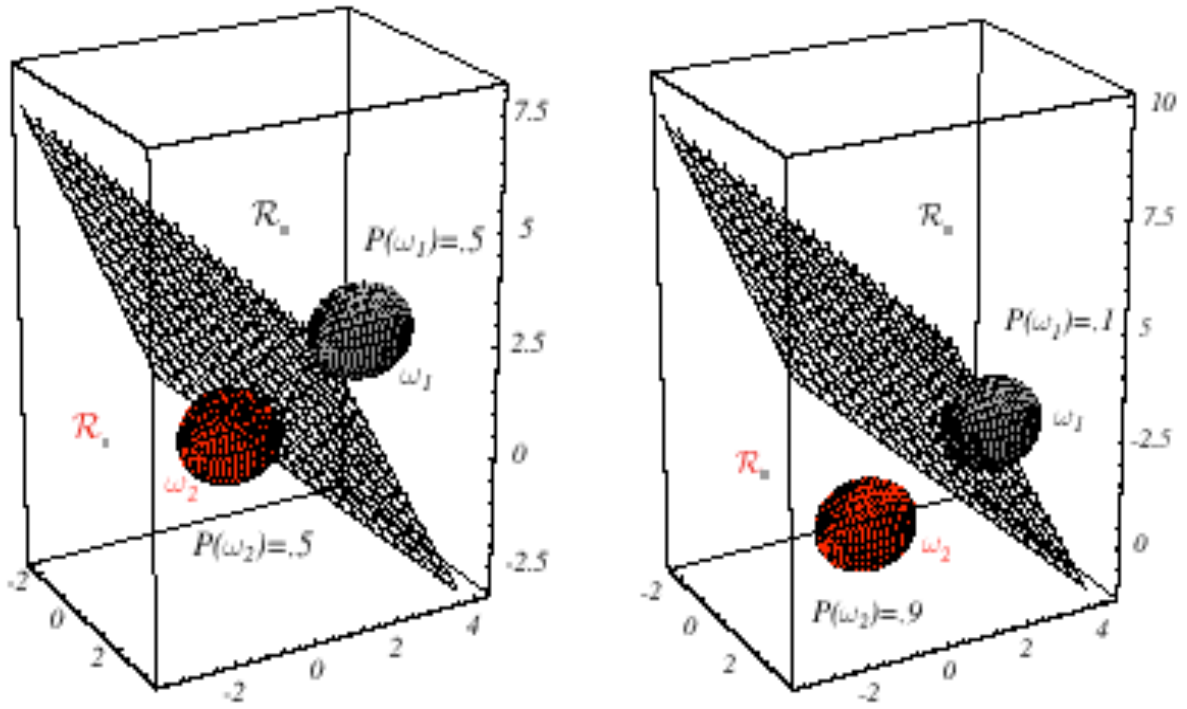


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means. From: Richard O. Duda,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

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

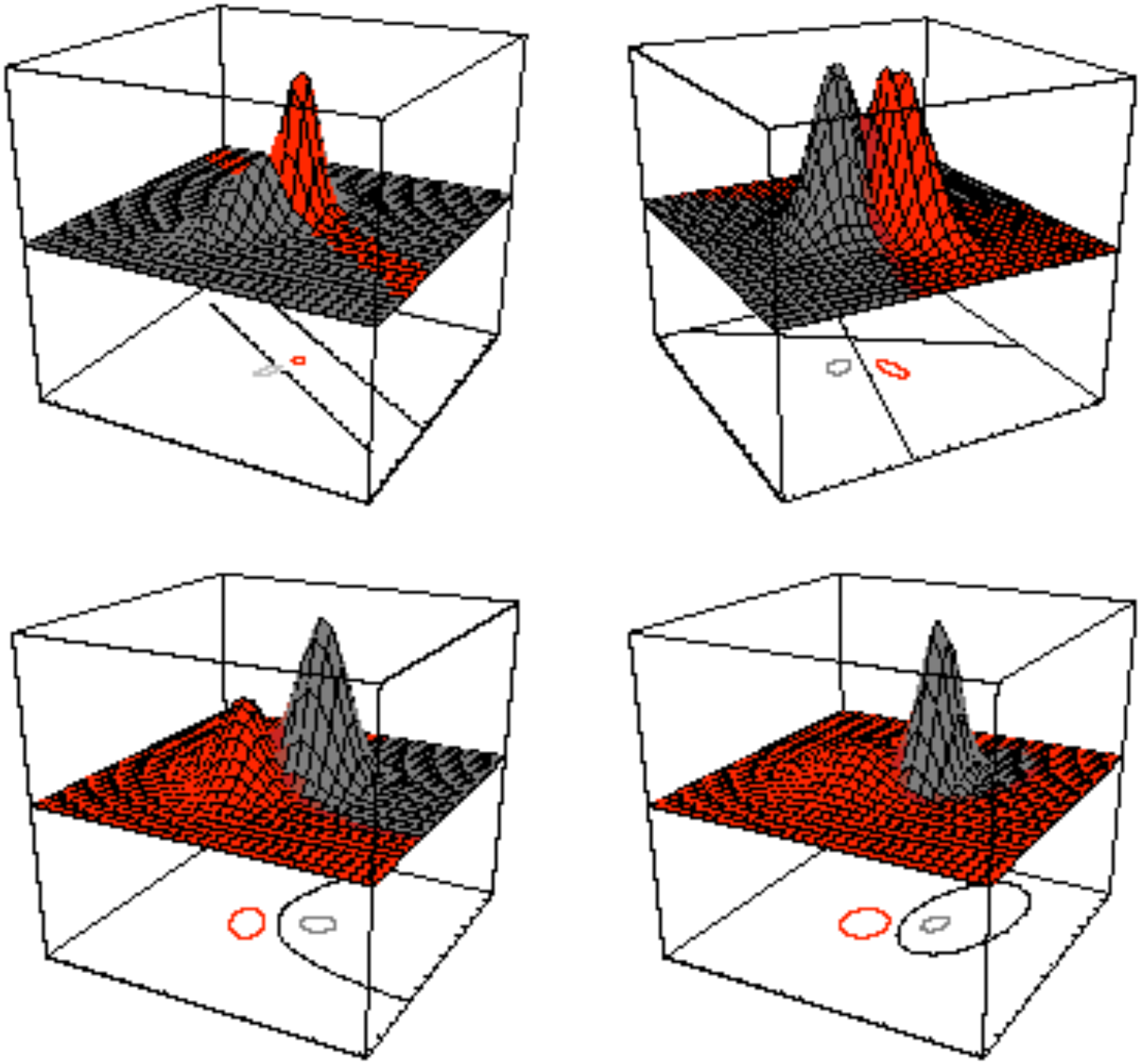
where :

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

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

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

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



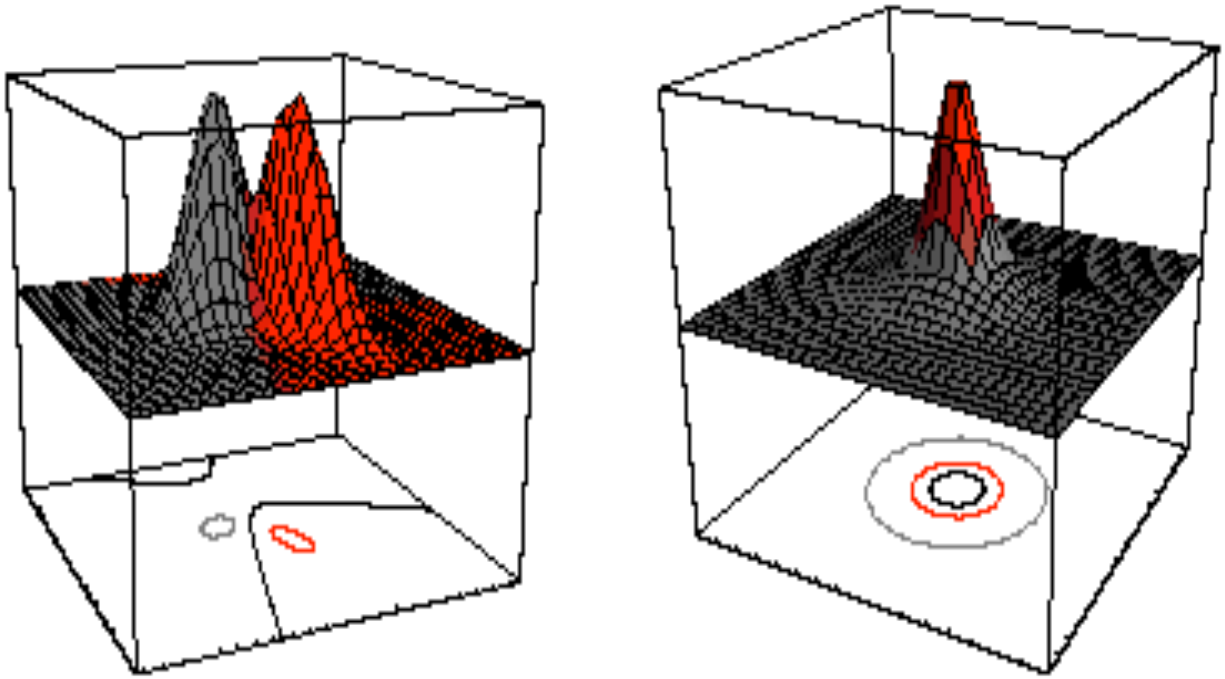


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability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

Bayes Decision Theory – Discrete Features

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

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

- 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:

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

- 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 :

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

decide ω_1 ***if*** $g(x) > 0$ ***and*** ω_2 ***if*** $g(x) \leq 0$