Chapter 2 (part 3)
Bayesian Decision Theory

- Discriminant Functions for the Normal Density
- *Minimum Distance Classifier*
- Error Probabilities and Integrals
- Signal Detection Theory and Operating Characteristics
- Bayes Decision Theory – Discrete Features
- Independent Binary Features
- Missing and Noisy Features

Discriminant Functions for the Normal Density

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

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

- Case of multivariate normal $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$.

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

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

Eq 47

Case 1: $\Sigma_i = \sigma^2.\mathbf{I}$ (**I** stands for the identity matrix)

The features are statistically independent, and each feature has the same variance, σ^2.

Geometrically, this corresponds to the situation in which the samples fall in equal-size hyperspherical clusters, the cluster for the *i*th class being centered about the mean vector $\boldsymbol{\mu}_i$. So we have

$$|\boldsymbol{\Sigma}_i| = \sigma^{2d} \qquad\qquad \boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\mathbf{I}$$

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

Where the *Euclidean norm* is $\|\mathbf{x}-\boldsymbol{\mu}_i\|^2 = (\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)$

If **x** is equally near two different mean vectors, the optimal decision will favor the *a priori* more likely category.

Expansion of the quadratic form $(\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i)$ yields:

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

The quadratic term $\mathbf{x}^t\mathbf{x}$ is the same for all *i*, making it an ignorable additive constant.

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

where:

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

(w_{i0} is called the threshold or bias for the ith 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:

$$g_i(\mathbf{x}) = g_j(\mathbf{x})$$

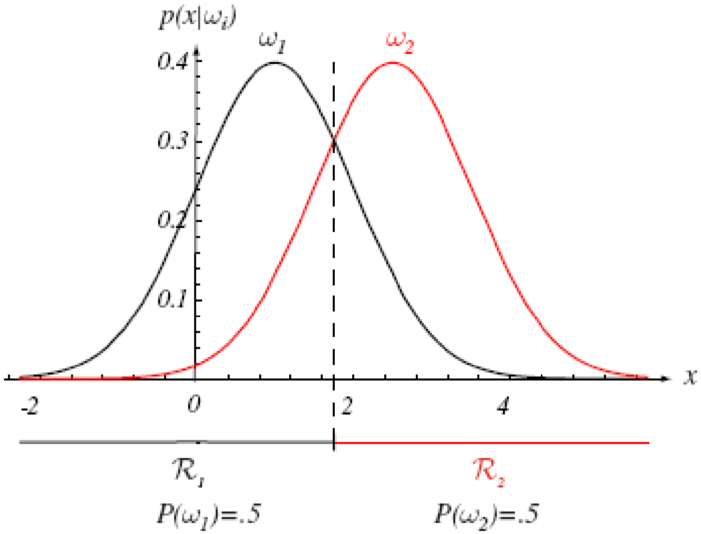
This equation can be written as

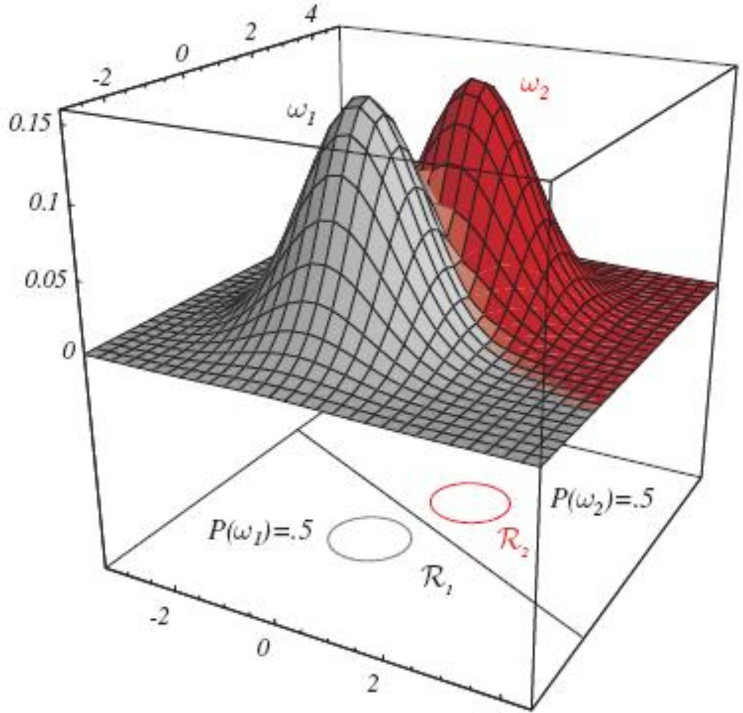
$$\boxed{\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0}$$

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

and

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





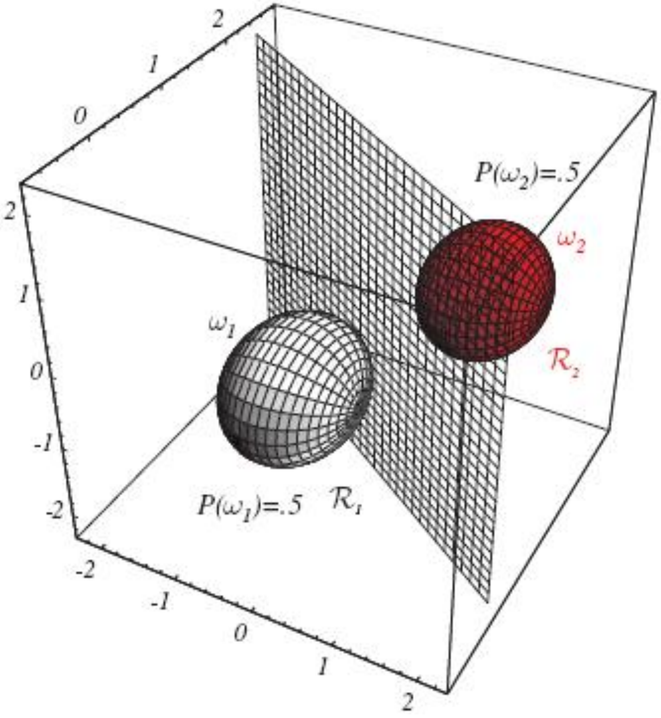


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in *d* dimensions, and the boundary is a generalized hyperplane of *d - 1* dimensions, perpendicular to the line separating the means. In these one-, two-, 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 3-dimensional case, the grid plane separates R_1 from R_2.

– The hyperplane through the point $\mathbf{x}_0$ and orthogonal to the vector $\mathbf{w}$, separating $\mathcal{R}_i$ and $\mathcal{R}_j$

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

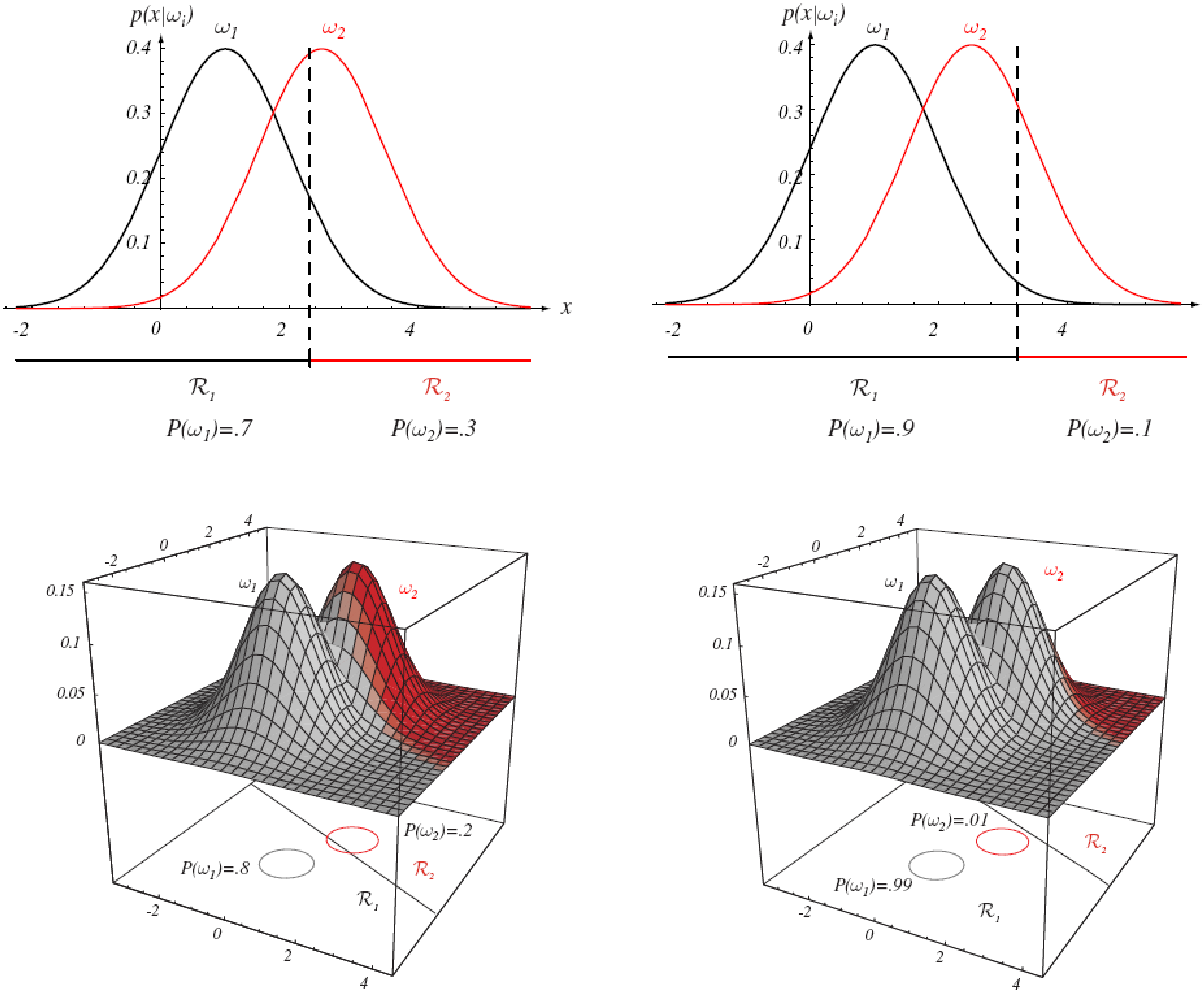
always orthogonal to the line linking the means!

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

- If $P(\omega_i) = P(\omega_j)$ the 2nd term vanishes, and thus the point $\mathbf{x}_0$ is halfway between the means, and the hyperplane is the perpendicular bisector of the line between the means (Fig. 2.11). If $P(\omega_i) \neq P(\omega_j)$, the point $\mathbf{x}_0$ shifts away from the more likely mean.

Minimum Distance Classifier

- If the prior probabilities $P(\omega_i)$ are the same for all c classes, then the $\ln P(\omega_i)$ term becomes another unimportant additive constant that can be ignored.
- The optimum decision rule can be stated very simply: to classify a feature vector **x**, measure the Euclidean distance $\mathbf{x} - \boldsymbol{\mu}_i$ from each **x** to each of the c mean vectors, and assign **x** to the category of the nearest mean. Such a classifier is called a *minimum distance classifier*. If each mean vector is thought of as being an ideal prototype or template for patterns in its class, then this is essentially a *template matching* procedure (Fig. 2.10)


$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)$=.7
$P(\omega_2)$=.3
$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)$=.9
$P(\omega_2)$=.1
0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_2)$=.2
$P(\omega_1)$=.8
$\mathcal{R}_2$
$\mathcal{R}_1$
-2
0
2
4
0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_2)$=.01
$\mathcal{R}_2$
$P(\omega_1)$=.99
$\mathcal{R}_1$
-2
0
2
4

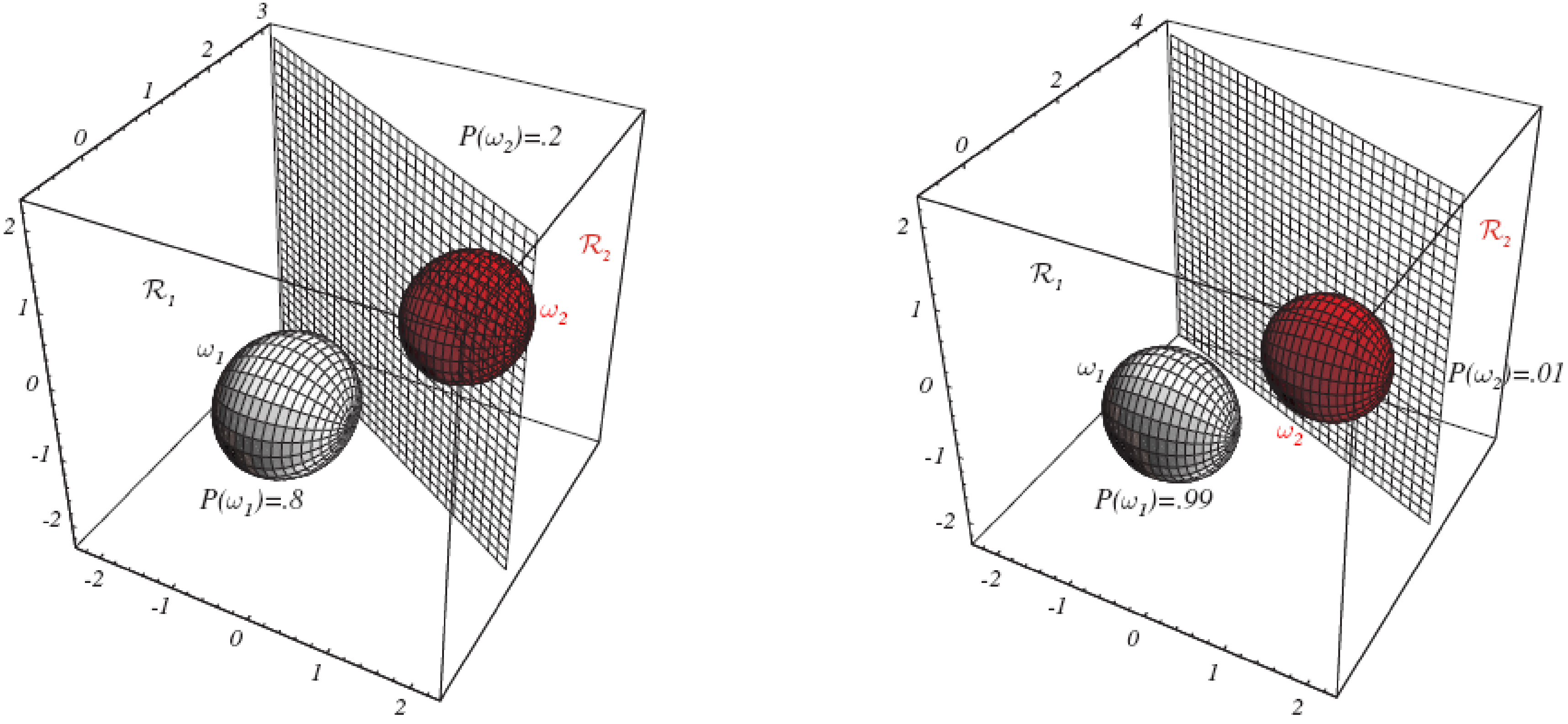


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions.

Case 2: $\mathbf{\Sigma_i = \Sigma}$ (covariance of all classes are identical but arbitrary!)

- The samples fall in hyperellipsoidal clusters of equal size and shape, the cluster for the *i*th class being centered about the mean vector $\boldsymbol{\mu}_i$.
- Since both $|\mathbf{\Sigma}_\mathrm{i}|$ and the *(d/2) ln 2π* term in Eq. 47 are independent of *i*, they can be ignored as superfluous additive constants. So the discriminant functions are

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

- If the prior probabilities *P(ω*$_i$*)* are the same for all c classes, then the *ln P(ω*$_i$*)* term can be ignored.

The optimal decision rule: To classify a feature vector **x**, measure the squared *Mahalanobis* distance $(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$ from **x** to each of the *c* mean vectors, and assign **x** to the category of the nearest mean.

Expansion of the quadratic form $(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$ and dropping $\mathbf{x}^t \boldsymbol{\Sigma}^{-1} \mathbf{x}$ term (independent of *i*) the resulting discriminant functions are again linear:

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

Where

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

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

Since the discriminants are linear, the resulting decision boundaries are again hyperplanes.

If $\mathcal{R}_i$ and $\mathcal{R}_j$ are contiguous, the boundary (Hyperplane) between them has the equation

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

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

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

The hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line between the means, because $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i\text{-}\boldsymbol{\mu}_j)$ is generally not in the direction of $\boldsymbol{\mu}_i\text{-}\boldsymbol{\mu}_j$.

However, it does intersect that line at the point $\mathbf{x}_0$ which is halfway between the means if the prior probabilities are equal.

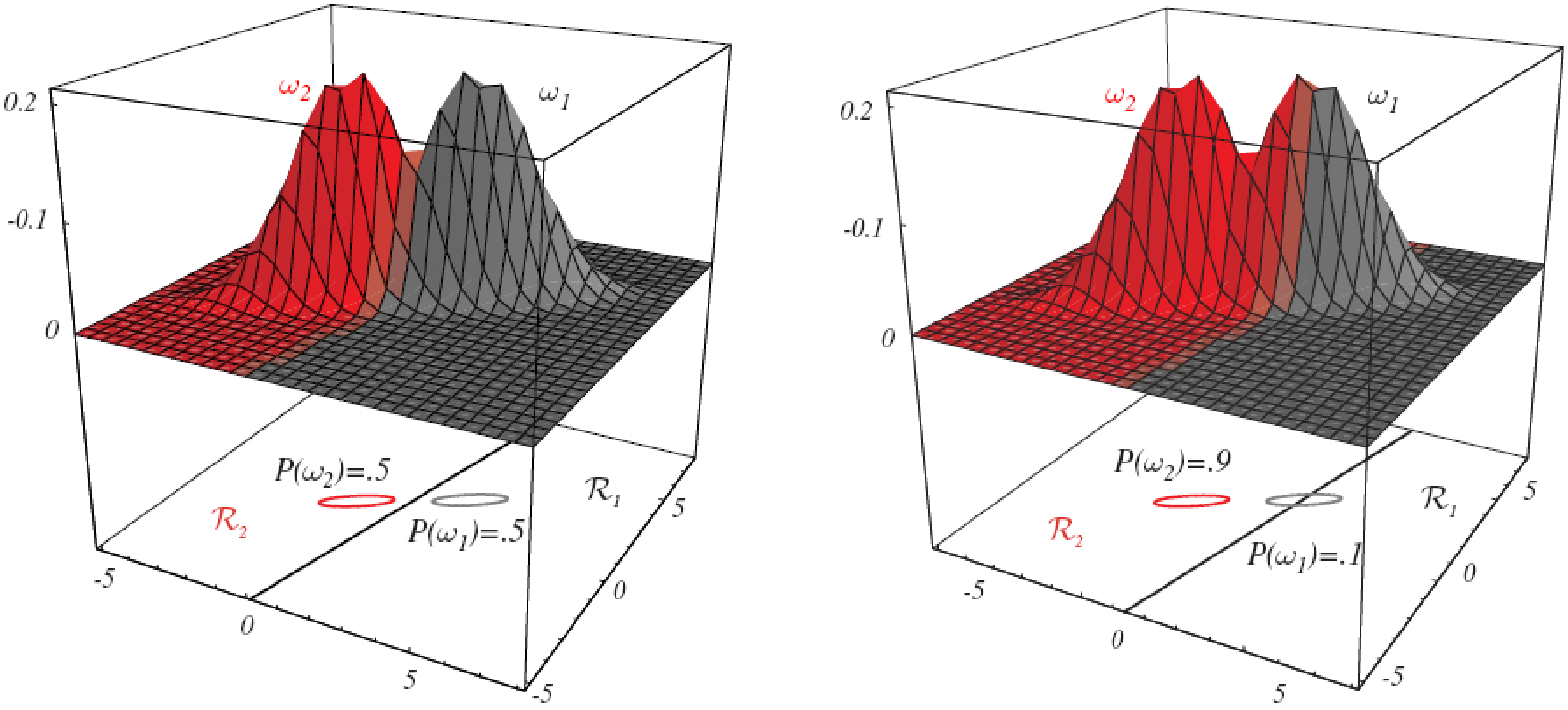


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means.

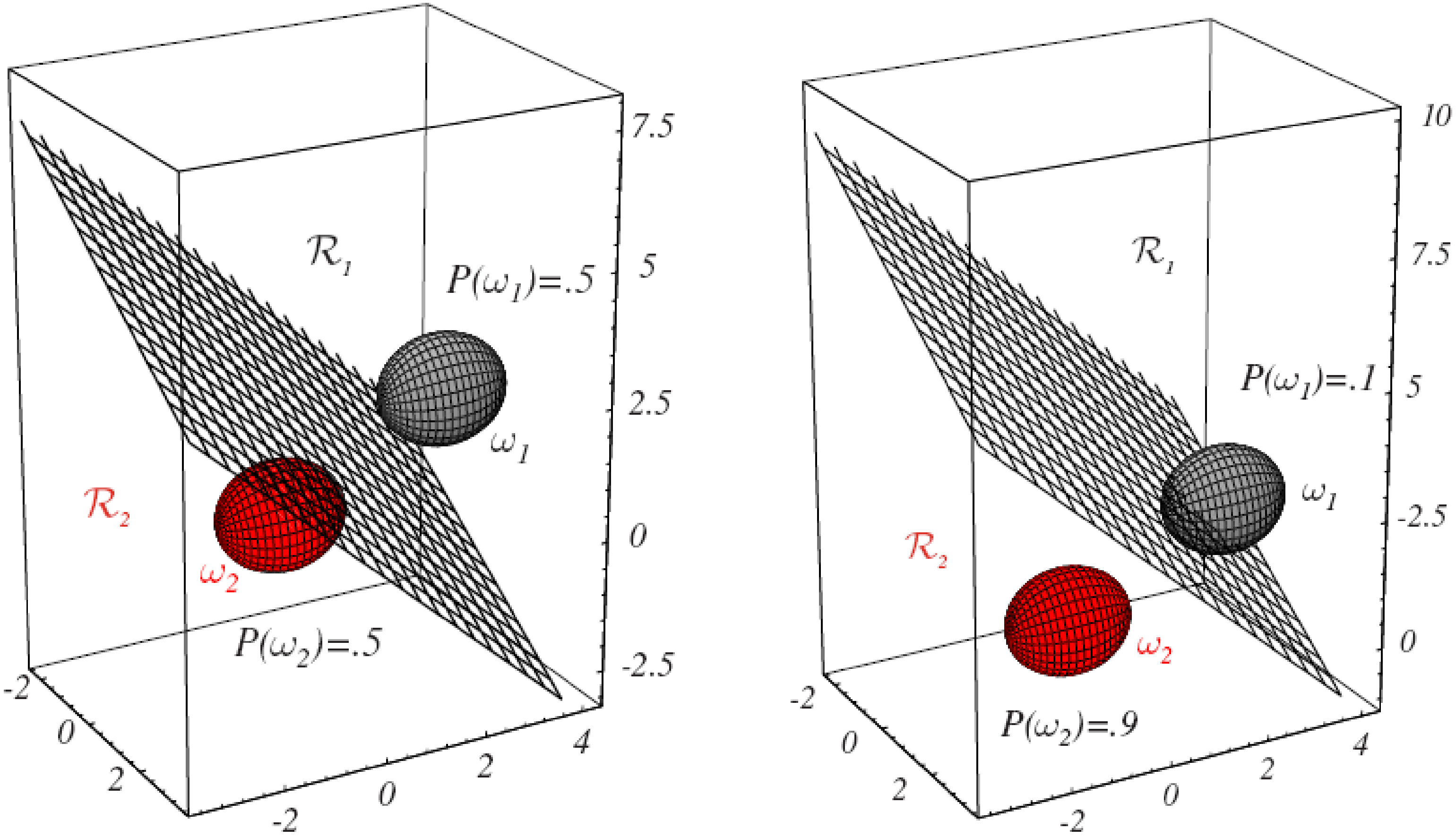


FIGURE 2.12 The same caption

Example:

In a two-class, two-dimensional classification task, the feature vectors are generated by two normal distributions sharing the same covariance matrix

$$\Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

and the mean vectors are $\boldsymbol{\mu}_{1} = [0, 0]^t$, $\boldsymbol{\mu}_{2} = [3, 3]^t$, respectively.

(a) Classify the vector $[1.0, 2.2]^t$ according to the Bayesian classifier.

It suffices to compute the Mahalanobis distance of $[1.0, 2.2]^t$ from the two mean vectors. Thus,

$$(\mathbf{x} - \boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) = [1.0, 2.2]\begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}\begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

$$(\mathbf{x} - \boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) = [-2.0, -0.8]\begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}\begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

Thus, the vector is assigned to the class with mean vector $[0, 0]^t$. Notice that the given vector $[1.0, 2.2]^t$ is closer to $[3, 3]^t$ with respect to the Euclidean distance.

Case 3: Σ_i = arbitrary

– The covariance matrices are different for each category

$$
\boxed{
\begin{aligned}
&g_i(x) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0} \\
&where: \\
&\mathbf{W}_\mathrm{i} = -\frac{1}{2}\mathbf{\Sigma}_i^{-1} \\
&\mathbf{w}_\mathrm{i} = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i \\
&w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln\left|\mathbf{\Sigma}_i\right| + \ln P(\omega_i)
\end{aligned}
}
$$

(Two category case →Hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)

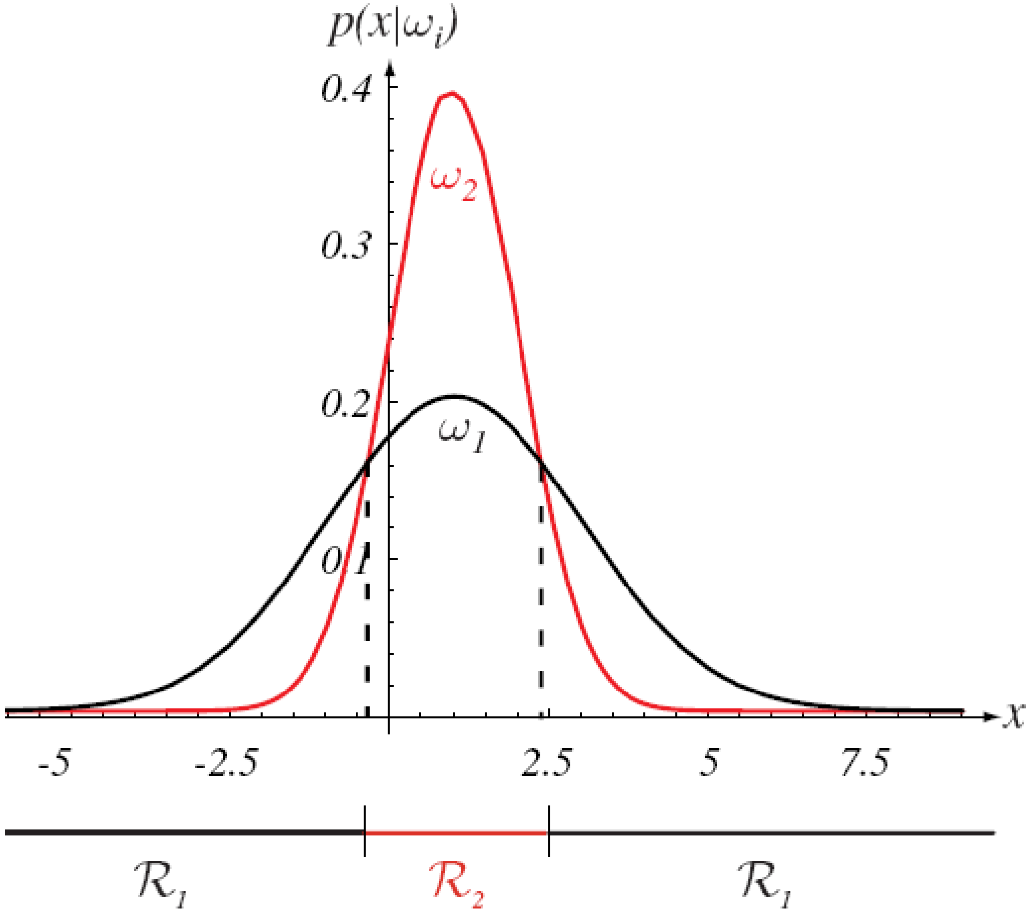
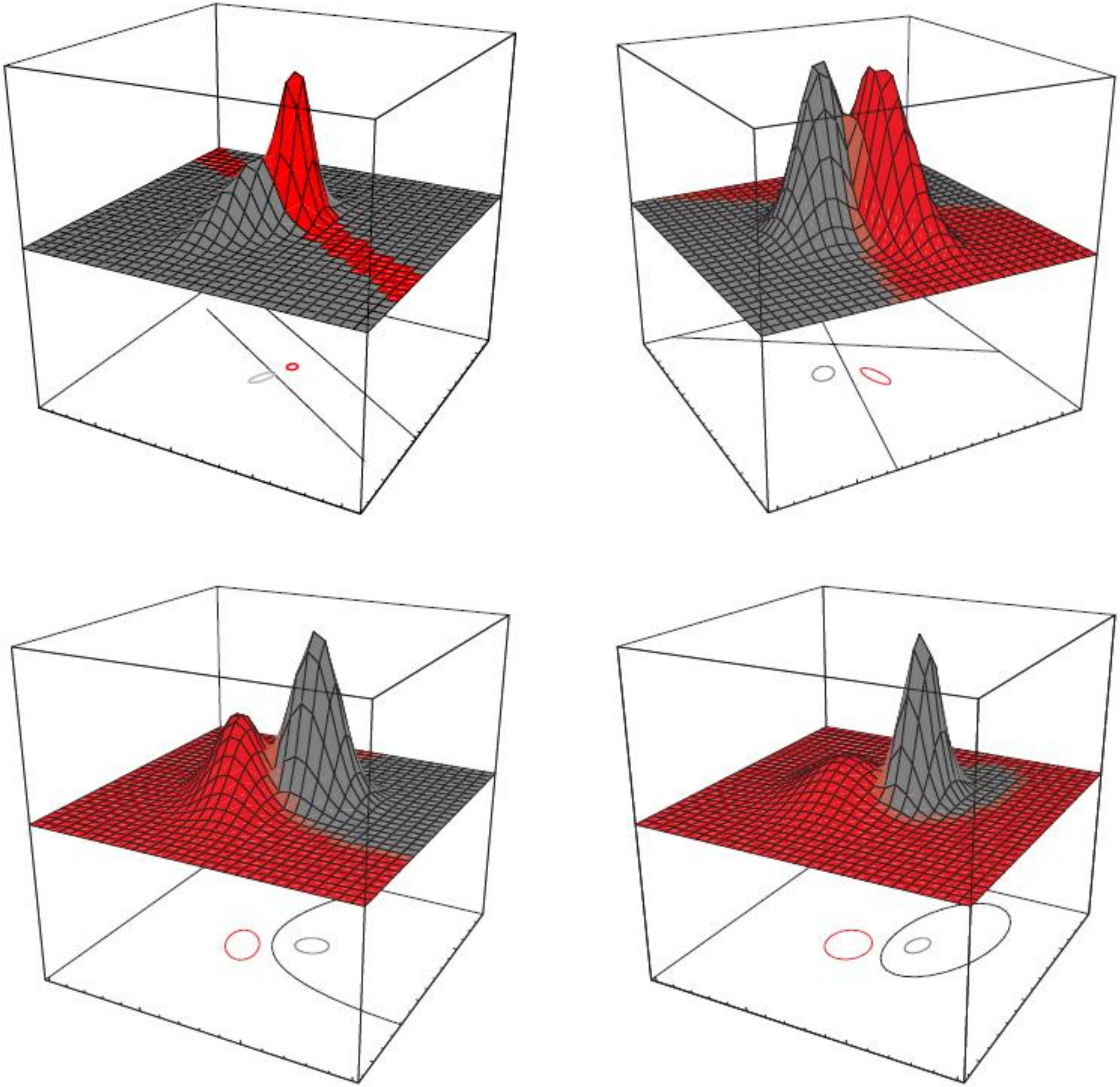


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance.



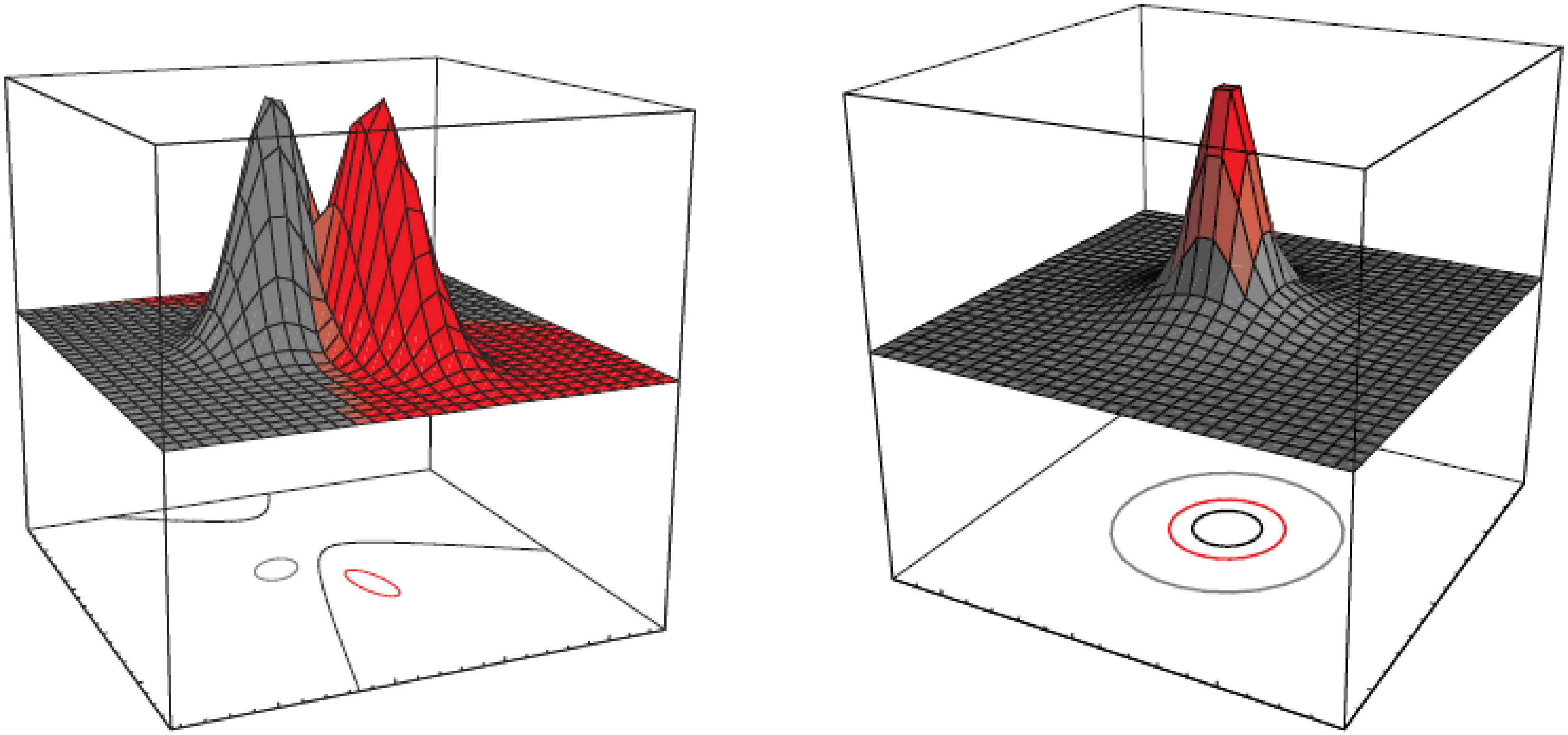
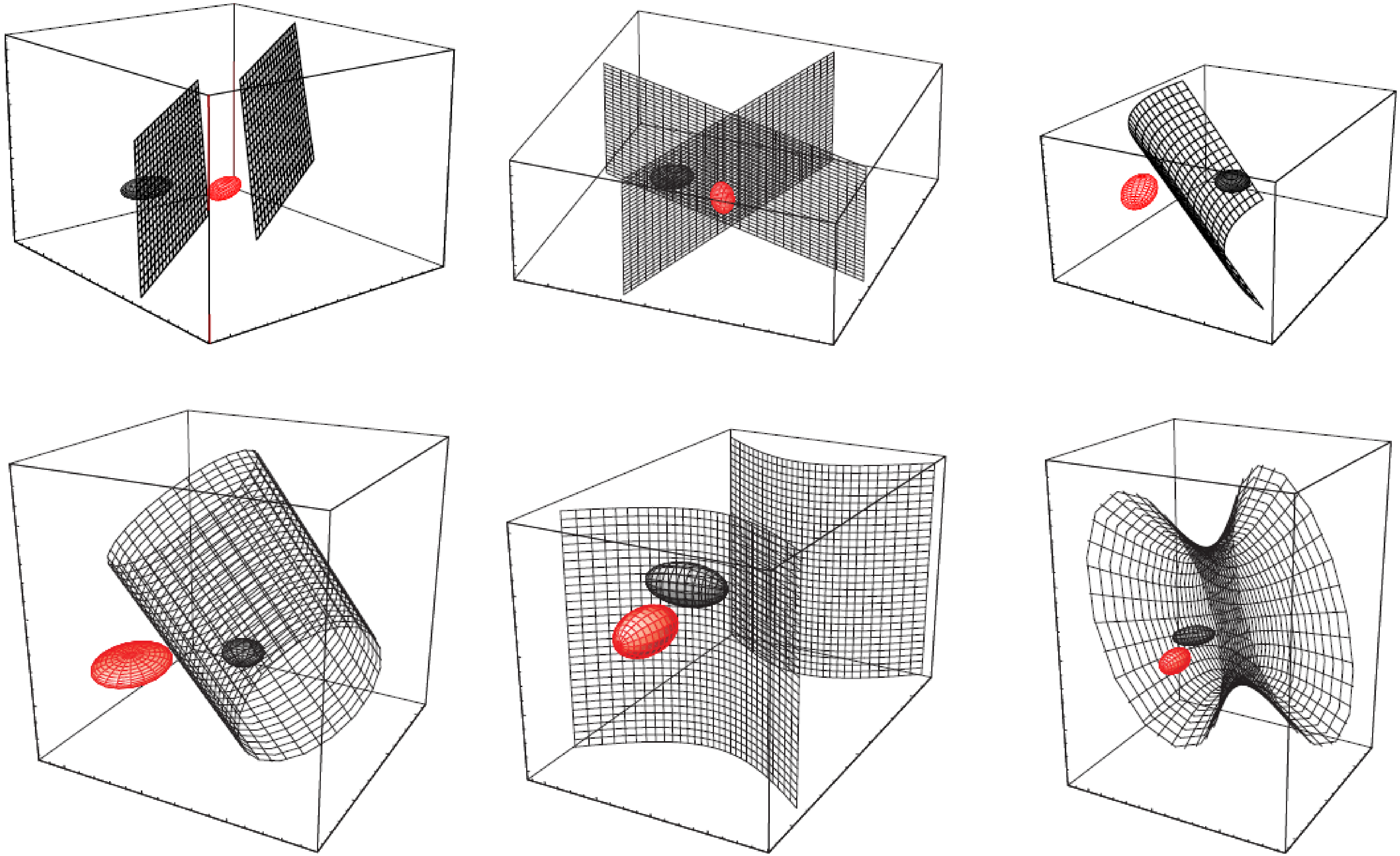


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These Variances are indicated by the contours of constant probability density.



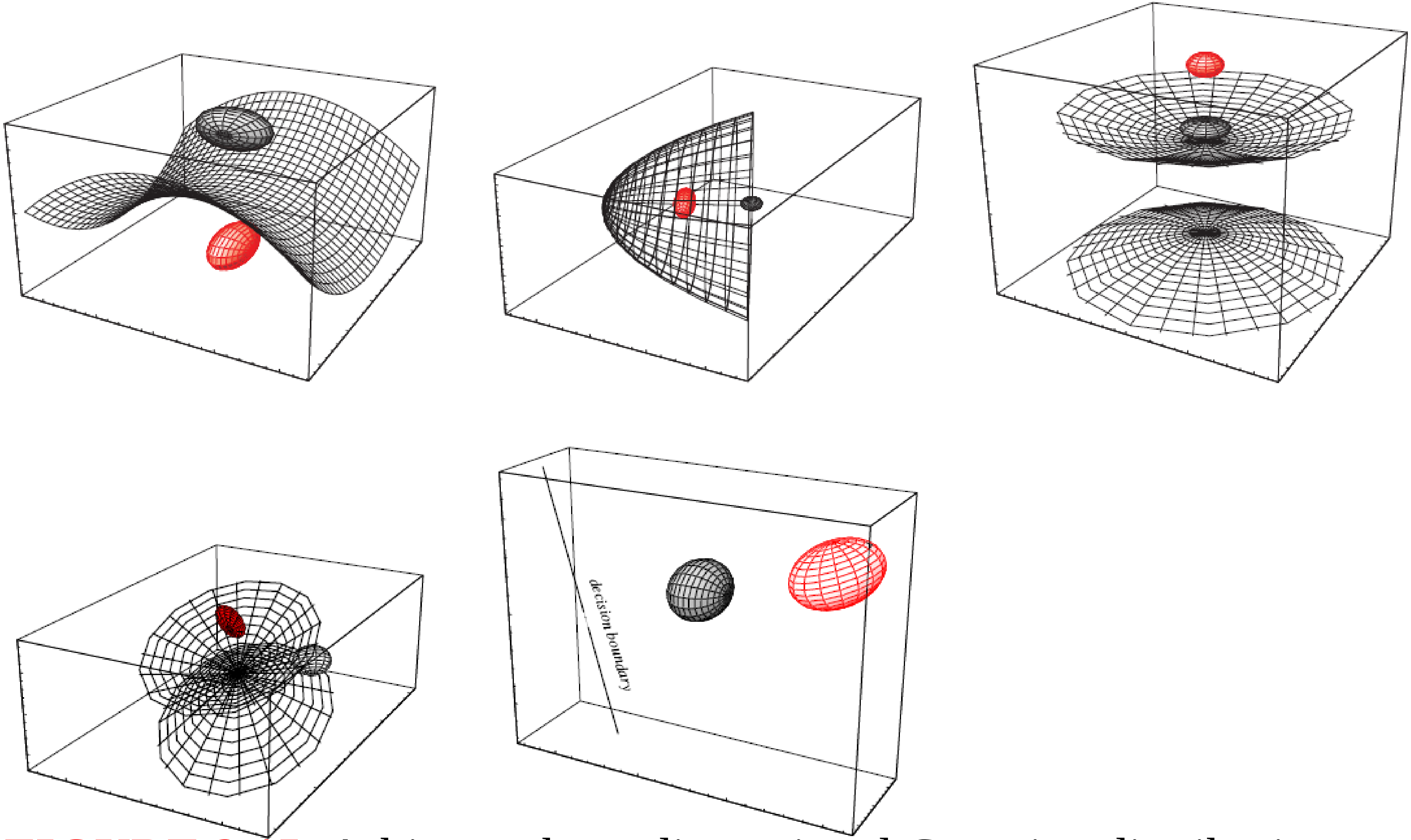


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line.

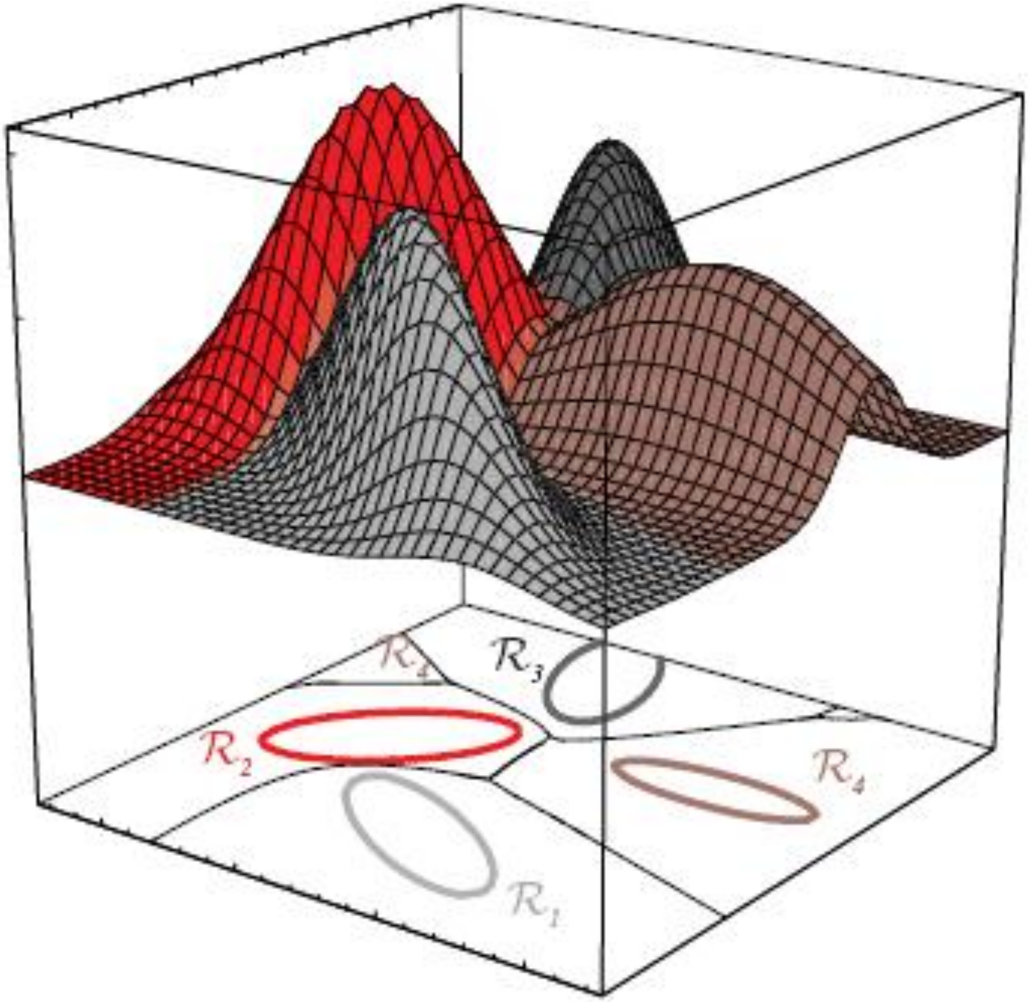
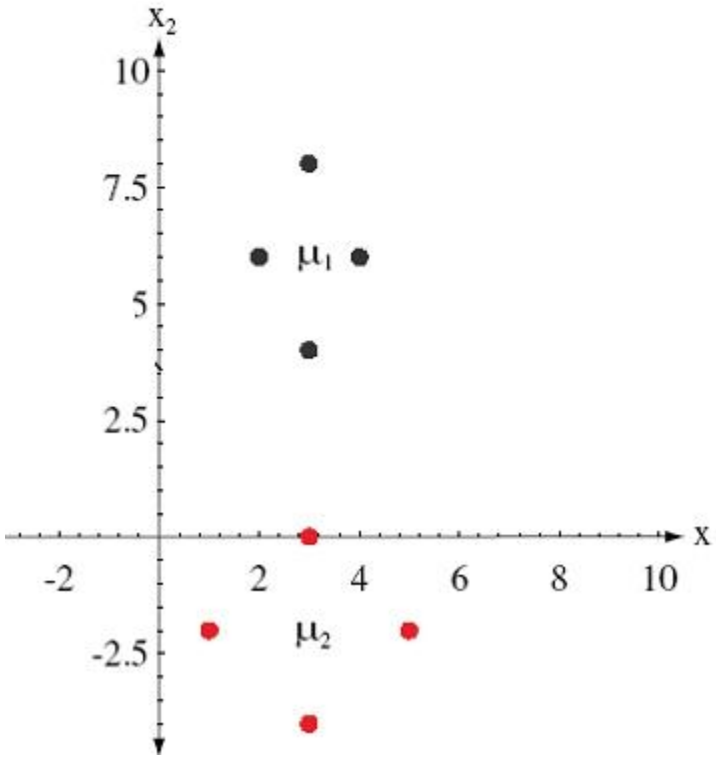


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

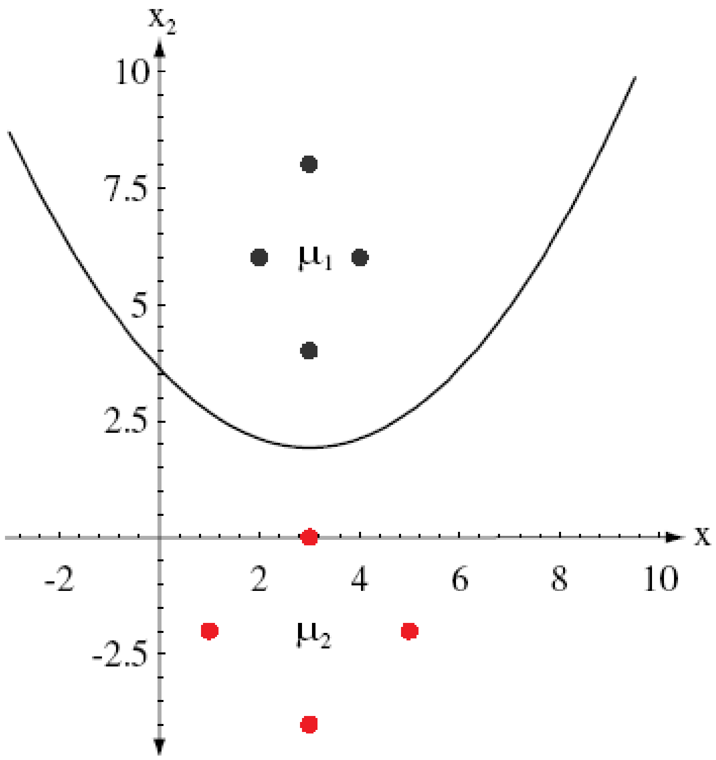
Example 1: Decision regions for two-dimensional Gaussian data



$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

Example 1: Decision regions for two-dimensional Gaussian data



$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \qquad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \quad \Sigma_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \qquad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad \Sigma_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

We assume equal prior probabilities, $P(\omega_1) = P(\omega_2) = 0.5$, and substitute these into the general form for a discriminant and setting $g_1(\mathbf{x}) = g_2(\mathbf{x})$ to obtain the decision boundary:

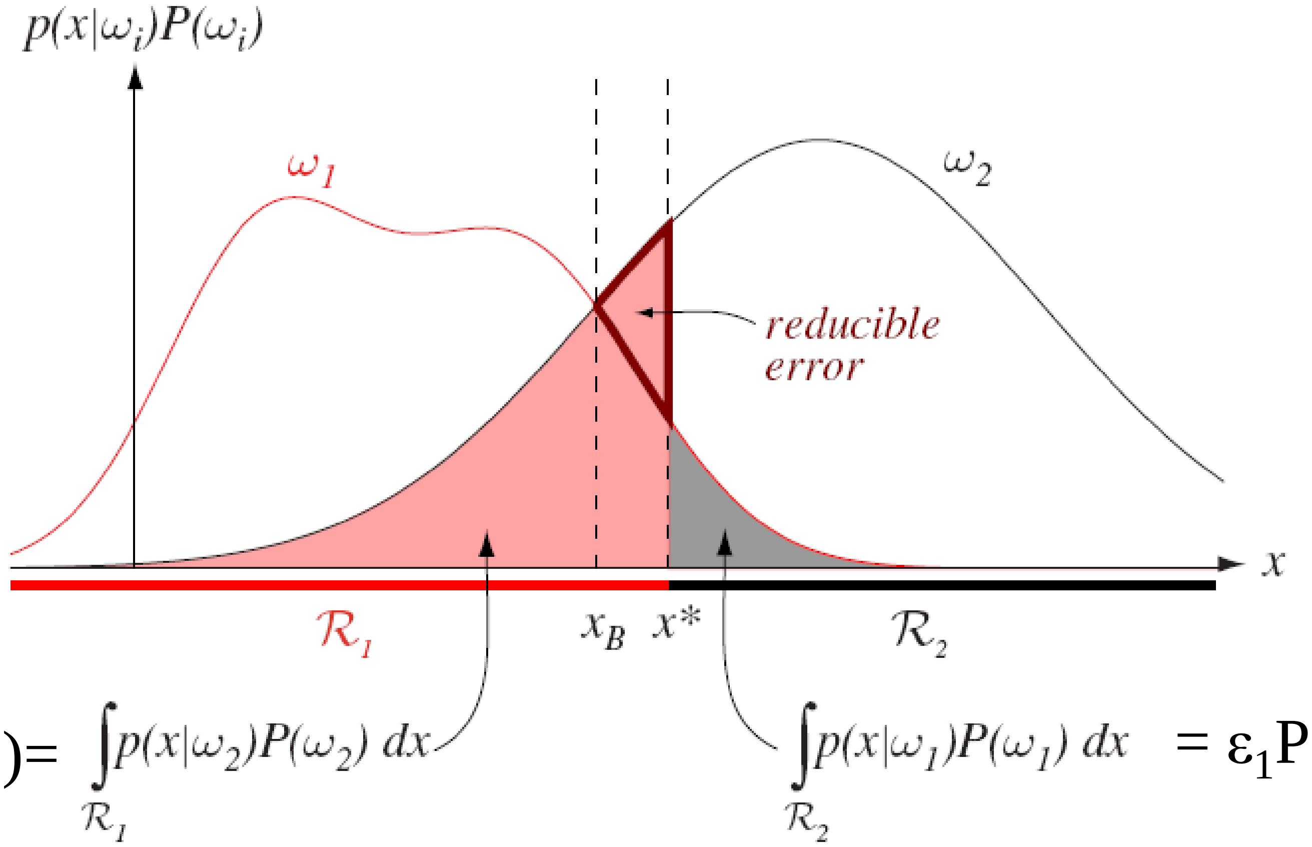
$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

This equation describes a parabola with vertex at (3, 1.83). Note that despite the fact that the variance in the data along the x_2 direction for both distributions is the same, the decision boundary does not pass through the point (3, 2), midway between the means, as we might have naively guessed. This is because for the ω_1 distribution, the probability distribution is "squeezed" in the x_1-direction more so than for the ω_2 distribution.

Error Probabilities and Integrals

- Consider first the two-category case, and suppose the dichotomizer has divided the space into two regions $\mathcal{R}_1$ and $\mathcal{R}_2$ in a possibly non-optimal way.
- Errors
  - an observation **x** falls in $\mathcal{R}_2$ and the true state of nature is ω_1 or
  - **x** falls in $\mathcal{R}_1$ and the true state of nature is ω_2.

$$\begin{aligned} P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2 | \omega_1) P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1 | \omega_2) P(\omega_2) \\ &= \int\limits_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) P(\omega_1)\ d\mathbf{x} + \int\limits_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) P(\omega_2)\ d\mathbf{x}. \end{aligned}$$



$$\varepsilon_2 P(\omega_2) = \int_{\mathcal{R}_1} p(x|\omega_2)P(\omega_2)\,dx \qquad \int_{\mathcal{R}_2} p(x|\omega_1)P(\omega_1)\,dx = \varepsilon_1 P(\omega_1)$$

FIGURE 2.17. Components of the probability of error for equal priors and (non-optimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate.

P(*correct*)?

- In the multicategory case, there are more ways to be wrong than to be right, and it is simpler to compute the probability of being correct.

$$
\begin{aligned}
P(correct) &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, \omega_i) \\
&= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i | \omega_i) P(\omega_i) \\
&= \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i) P(\omega_i) \; d\mathbf{x}.
\end{aligned}
$$

The Bayes classifier maximizes this probability.

Error Bounds for Normal Densities

- The full calculation of the error for the Gaussian case would be quite difficult, especially in high dimensions, because of the discontinuous nature of the decision regions in the above integral.
- In the two-category case the general error integral can be approximated analytically to give us an upper bound on the error.

Chernoff Bound

- To derive a bound for the error, we need the following inequality:

  $\min[a, b] \le a^{\beta} b^{1-\beta}$ for $a, b \ge 0$ and $0 \le \beta \le 1$.

- Assume $a \ge b$. Thus we need only show that $b \le a^{\beta} b^{1-\beta} = (a/b)^{\beta} b$. But this inequality is manifestly valid, since $(a/b)^{\beta} \ge 1$.

- We had

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

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

$$P(error|x) = \min\,[P(\omega_1|x), P(\omega_2|x)].$$

Thus we apply this inequality to get the bound:

$$\boxed{P(error) \le P^{\beta}(\omega_1)P^{1-\beta}(\omega_2) \int p^{\beta}(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)\ d\mathbf{x}}$$

for $0 \le \beta \le 1.$

This integral is over *all* feature space.

If the conditional probabilities are normal, this integral can be evaluated analytically, yielding:

$$\int p^{\beta}(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)\ d\mathbf{x} = e^{-k(\beta)}$$

where

$$\begin{aligned} k(\beta) \quad &= \quad \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \left[(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2\right]^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ & \qquad + \frac{1}{2}\ln\frac{(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2}{|\boldsymbol{\Sigma}_1|^{1-\beta}|\boldsymbol{\Sigma}_2|^{\beta}}. \end{aligned}$$

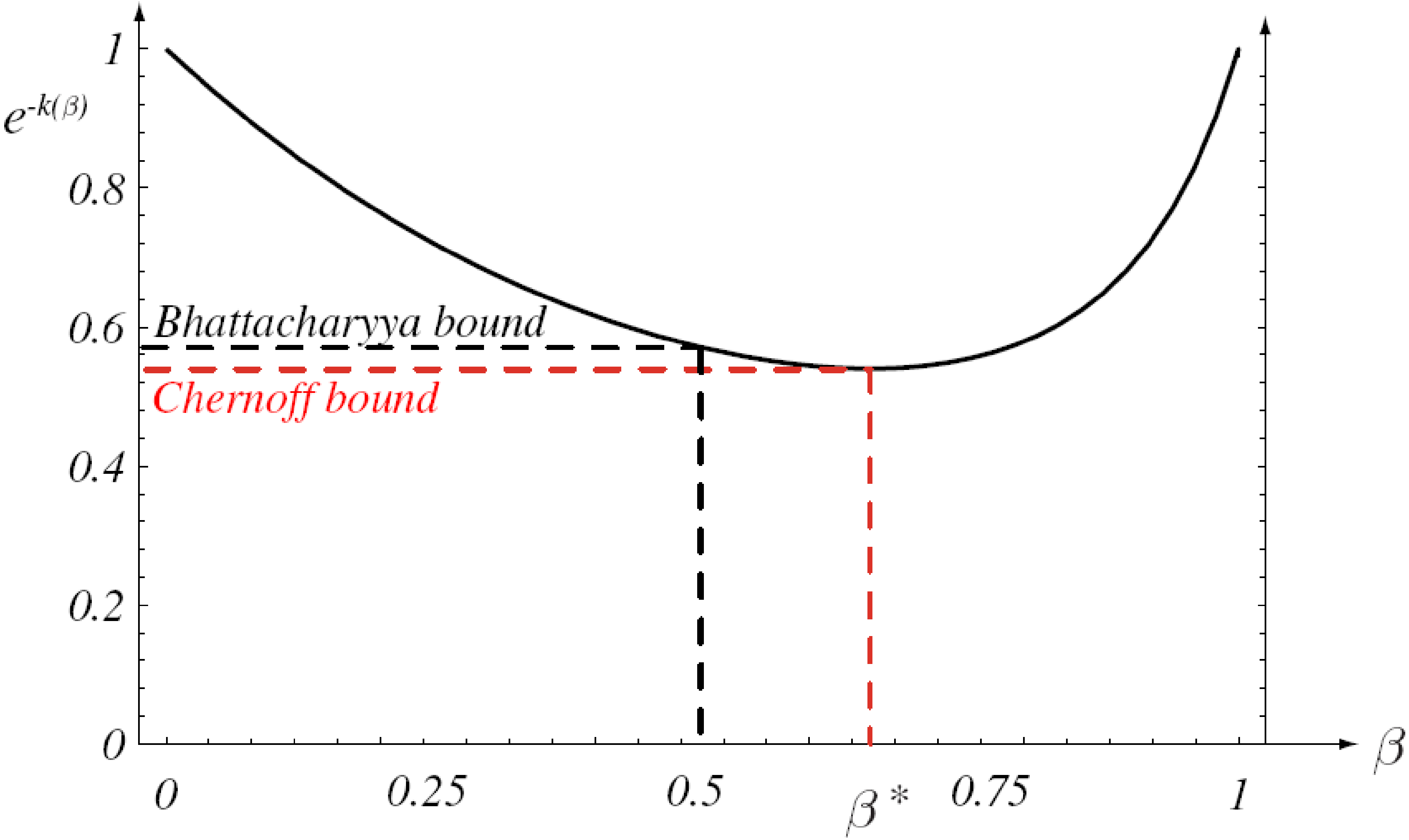


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^*= 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$).

- *k*(β) is called *Chernoff distance*. The *Chernoff bound*, on *P*(*error*) is found by analytically or numerically finding the value of β that minimizes $P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)\, e^{-k(\beta)}$ and substituting the results in Eq. *P(error)=…*

Bhattacharyya Bound

- Slightly less tight bound can be derived simply by setting the results for β = 1/2. This result is the so-called *Bhattacharyya bound* on the error. Thus,

$$\begin{aligned} P(error) &\leq \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)}\, d\mathbf{x} \\ &= \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)}, \end{aligned}$$

where

$$k(1/2) = 1/8(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2}\ln \frac{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}.$$

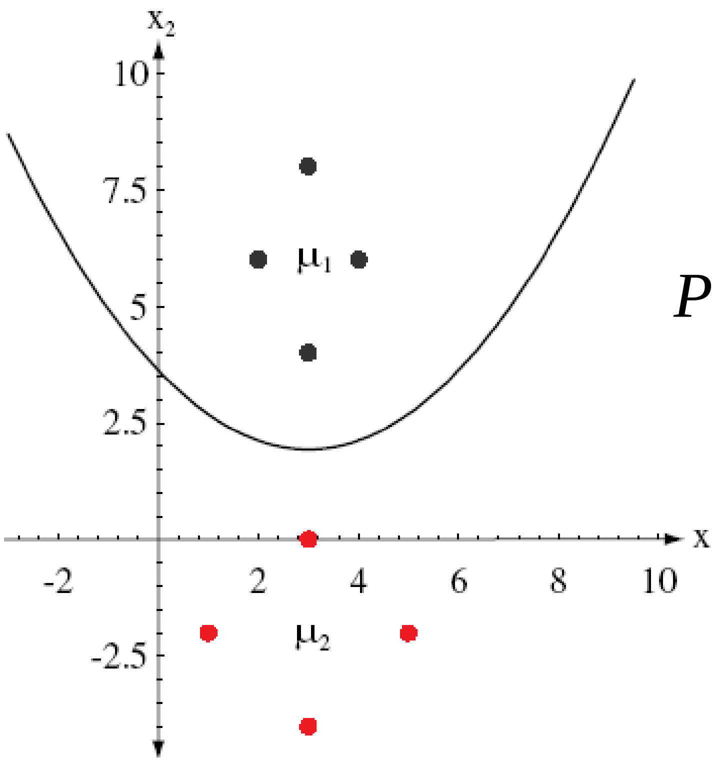
The term *k(1/2)* is called *Bhattacharyya distance*, and will be used as an important measure of the separability of two distributions.

The Chernoff and Bhattacharyya bounds may still be used even if the underlying distributions are not Gaussian. However, for distributions that deviate markedly from a Gaussian, the bounds will not be informative.

Example 2: Error bounds for Gaussian distributions.

It is a straightforward matter to calculate the Bhattacharyya bound for the two dimensional data sets of Example 1.

Example 1: Decision regions for two-dimensional Gaussian data



$$P(\omega_1) = P(\omega_2) = 0.5$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \qquad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

$$\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}$$

Calculating ….

- K(1/2)=0.41157,
- $P(error) \leq 0.008191$
- A slightly tighter bound on the error can be approximated by searching numerically for the Chernoff bound, which for this problem gives *0.008190*.
- Numerically integrating of the error rate gives an error rate of 0.0021, and thus the bounds here are not particularly tight. Such numerical integration is often impractical for Gaussians in higher than two or three dimensions.

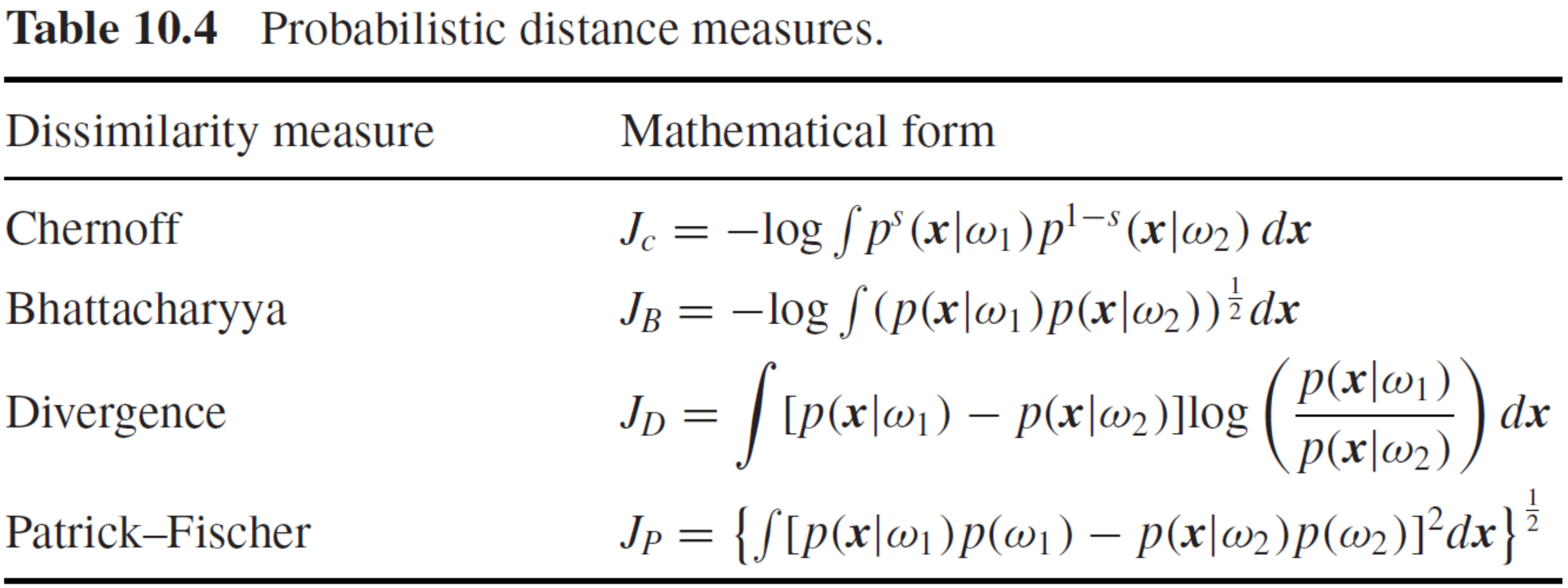


Table 10.4 Probabilistic distance measures.

Dissimilarity measure	Mathematical form
Chernoff	$J_c = -\log \int p^s(\boldsymbol{x}\|\omega_1)\,p^{1-s}(\boldsymbol{x}\|\omega_2)\,d\boldsymbol{x}$
Bhattacharyya	$J_B = -\log \int (p(\boldsymbol{x}\|\omega_1)p(\boldsymbol{x}\|\omega_2))^{\frac{1}{2}}\,d\boldsymbol{x}$
Divergence	$J_D = \int [p(\boldsymbol{x}\|\omega_1) - p(\boldsymbol{x}\|\omega_2)]\log\left(\frac{p(\boldsymbol{x}\|\omega_1)}{p(\boldsymbol{x}\|\omega_2)}\right)d\boldsymbol{x}$
Patrick–Fischer	$J_P = \left\{\int [p(\boldsymbol{x}\|\omega_1)p(\omega_1) - p(\boldsymbol{x}\|\omega_2)p(\omega_2)]^2 d\boldsymbol{x}\right\}^{\frac{1}{2}}$

Signal Detection Theory and Operating Characteristics

- Another measure of distance between two Gaussian distributions has found great use in experimental psychology, radar detection and other fields.
- Suppose we are interested in detecting a single weak pulse, such as a dim flash of light or a weak radar reflection.
- Detector detects a signal whose mean value is μ_1 when signal is absent and μ_2 when signal is present.
- The detector (classifier) employs a threshold value x^*.
- How do we compare two decision rules if they require different thresholds for optimum performance?

- Suppose we do not know μ_1, μ_2, σ nor x^* .
- We seek to find some measure of the ease of discriminating whether the pulse is present or not, in a form independent of the choice of x^*.
- Such a measure is the *discriminability*, which describes the inherent and unchangeable properties due to noise and the strength of the external signal, but not on the decision strategy (i.e., the actual choice of x^*).

Discriminability

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}.$$

A high *d' is* of course desirable.

While we do not know μ_1, μ_2, σ nor x^*, we assume here that we know the state of nature and the decision of the system. Such information allows us to find *d'*. To this end, we consider the following four probabilities:

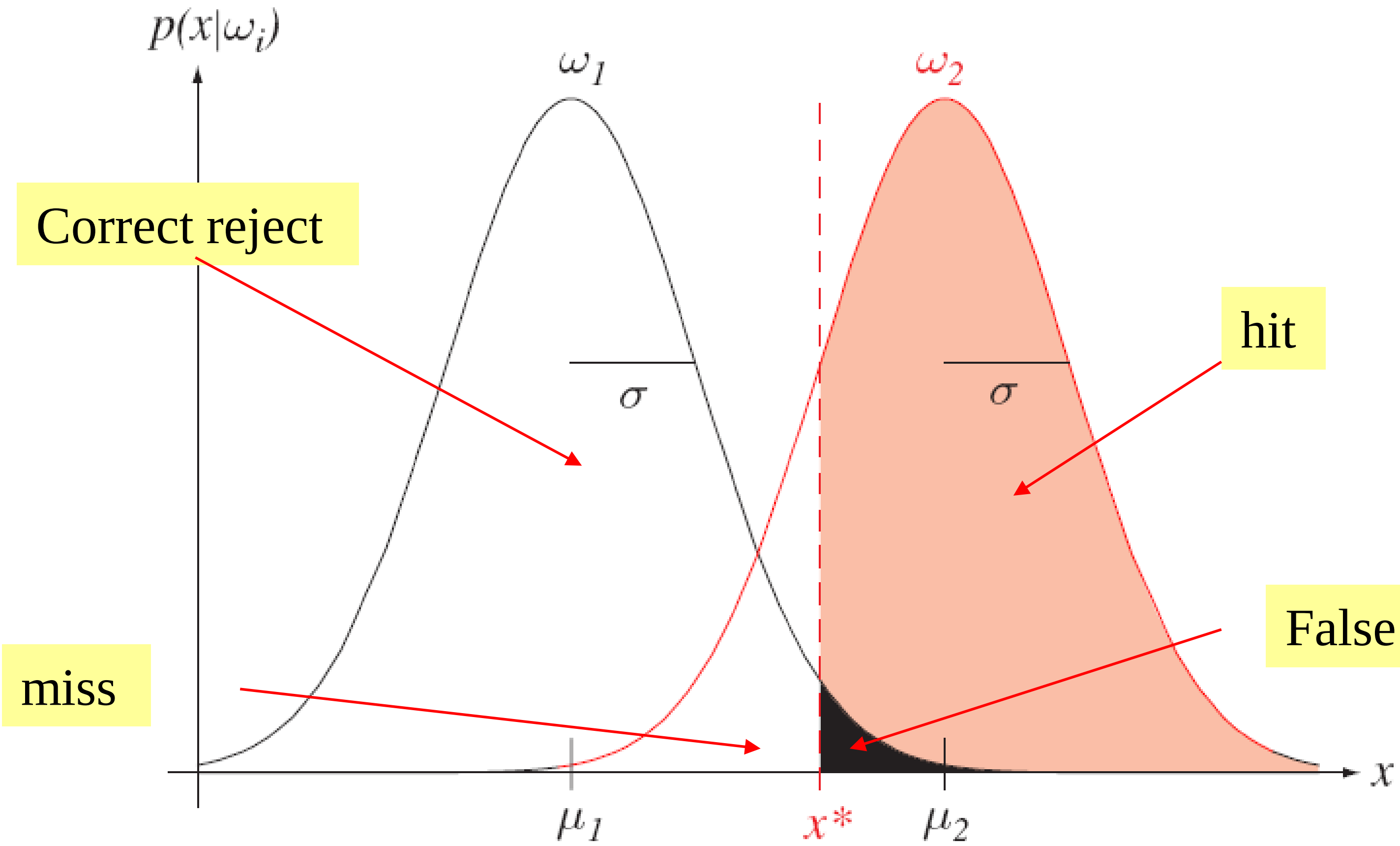
- $P(x > x^* \mid x \in \omega_2)$: *a hit* — the probability that the internal signal is above x^* given that the external signal is present. True Positive (TP)
- $P(x > x^* \mid x \in \omega_1)$: *a false alarm* — the probability that the internal signal is above x^* despite there being no external signal is present. False Positive (FP) – type I error.
- $P(x < x^* \mid x \in \omega_2)$: *a miss* — the probability that the internal signal is below x^* given that the external signal is present. False Negative (FN)- type II error.
- $P(x < x^* \mid x \in \omega_1)$: *a correct rejection* — the probability that the internal signal is below x^* given that the external signal is not present. True Negative (TN).

Terminology and derivations from a confusion matrix

- True Positive Rate (TPR) eqv. with hit rate, recall, sensitivity $TPR = TP / P = TP / (TP + FN)$
- False Positive Rate (FPR) eqv. with false alarm rate, fall-out $FPR = FP / N = FP / (FP + TN)$
- Accuracy (ACC) $ACC = (TP + TN) / (P + N)$
- Specificity (SPC) $SPC = TN / (FP + TN) = 1 - FPR$
- Positive Predictive Value (PPV) eqv. with precision $PPV = TP / (TP + FP)$
- Negative Predictive Value (NPV) $NPV = TN / (TN + FN)$
- False Discovery Rate (FDR) $FDR = FP / (FP + TP)$

Matthews Correlation Coefficient (MCC)

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$



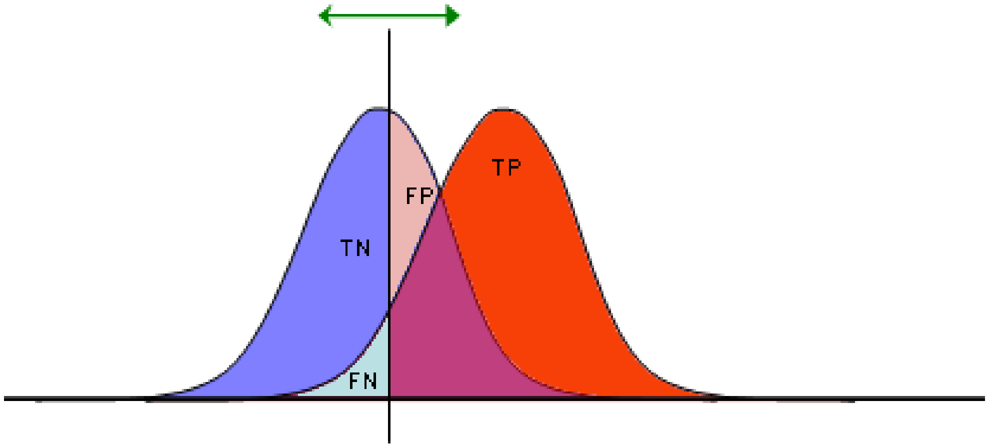
Correct reject

hit

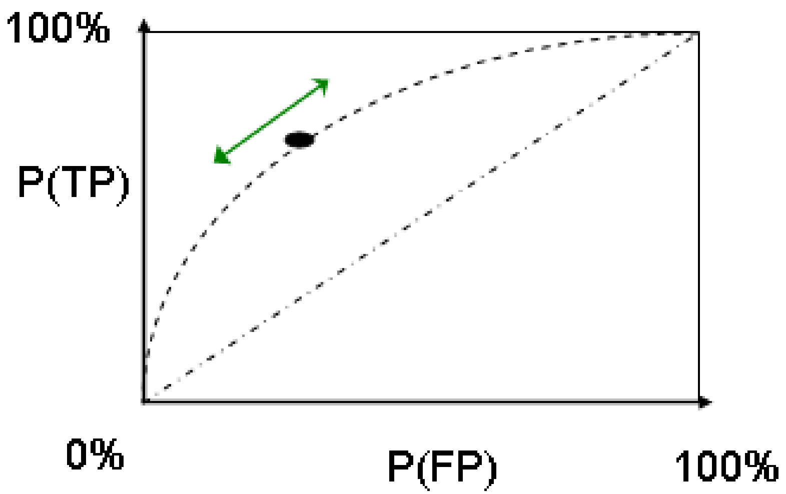
False alarm

miss

Figure 2.19: During any instant when no external pulse is present, the probability density for an internal signal is normal, i.e., $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine the probability of a hit (the red area under the ω_2 curve, above x^*) and of a false alarm (the black area under the ω_1 curve, above x^*).



TP	FP
FN	TN
1	1



Receiver Operating Characteristic (ROC)

- If we have a large number of trials (and we can assume x^* is fixed, albeit at an unknown value), we can determine these probabilities experimentally, in particular the hit and false alarm rates.
- If the densities are fixed but the threshold x^* is changed, then our hit and false alarm rates will also change.
- Thus we see that for a given discriminability *d'*, our point will move along a smooth curve — a receiver operating characteristic or ROC curve.

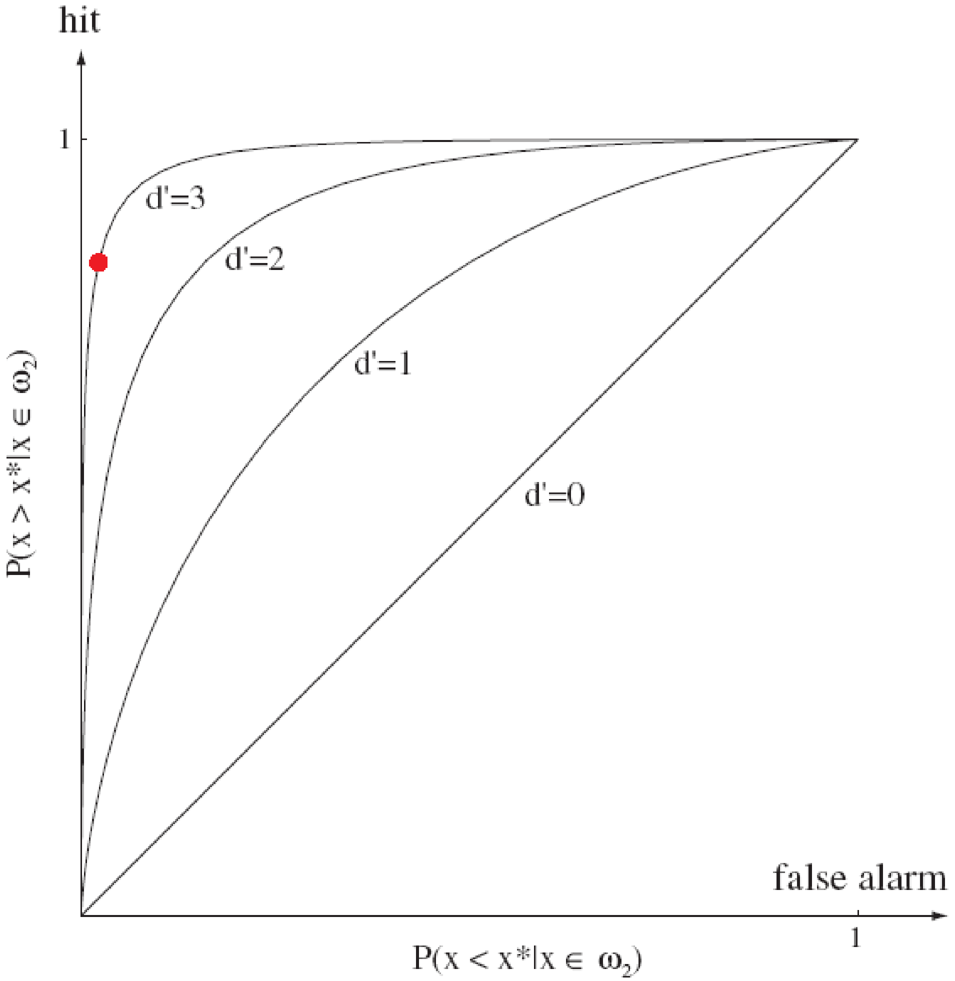


FIGURE 2.20. In a receiver operating characteristic (ROC) curve, the abscissa is the probability of false alarm, $P(x > x^* |x \in \omega_1)$, and the ordinate is the probability of hit, $P(x > x^* |x \in \omega_2)$. From the measured hit and false alarm rates (here corresponding to x^* in Fig. 2.19 and shown as the red dot), we can deduce that $d' = 3$.

Bayes Decision Theory – Discrete Features

- Suppose components of **x** are binary or integer valued. So, **x** can take only one of *m* discrete values

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

The probability density function p(**x**|ω_j) becomes singular; So:

$$\int p(\mathbf{x}|\omega_j)\, d\mathbf{x} \Longrightarrow \sum_{\mathbf{x}} P(\mathbf{x}|\omega_j)$$

Bayes’ formula:

$$\boxed{P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})},} \qquad P(\mathbf{x}) = \sum_{j=1}^{c} P(\mathbf{x}|\omega_j)P(\omega_j)$$

- The definition of the conditional risk *R(*α|*x)* is unchanged, and the fundamental Bayes decision rule remains the same: To minimize the overall risk:

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

- The basic rule to minimize the error-rate by maximizing the *posterior* probability is also unchanged as are the discriminant functions of Eqs. 25 – 27, given the obvious replacement of densities *p*(·) by probabilities *P*(·).

Independent Binary Features

- Case of independent binary features in 2 category problem

  Let $\mathbf{x} = (x_1, x_2, \ldots, x_d)^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = Prob(x_i=1|\omega_1)$$

$$q_i = Prob(x_i=1|\omega_2)$$

- By assuming conditional independence:

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

the Bernoulli distribution

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

Then the likelihood ratio is given by

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

We had
$$\left\{ \begin{aligned} g(\mathbf{x}) &= P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \\ \\ g(\mathbf{x}) &= \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}. \end{aligned} \right.$$

⇒

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

- The discriminant function in this case is:

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

where:

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

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**) ≤ 0

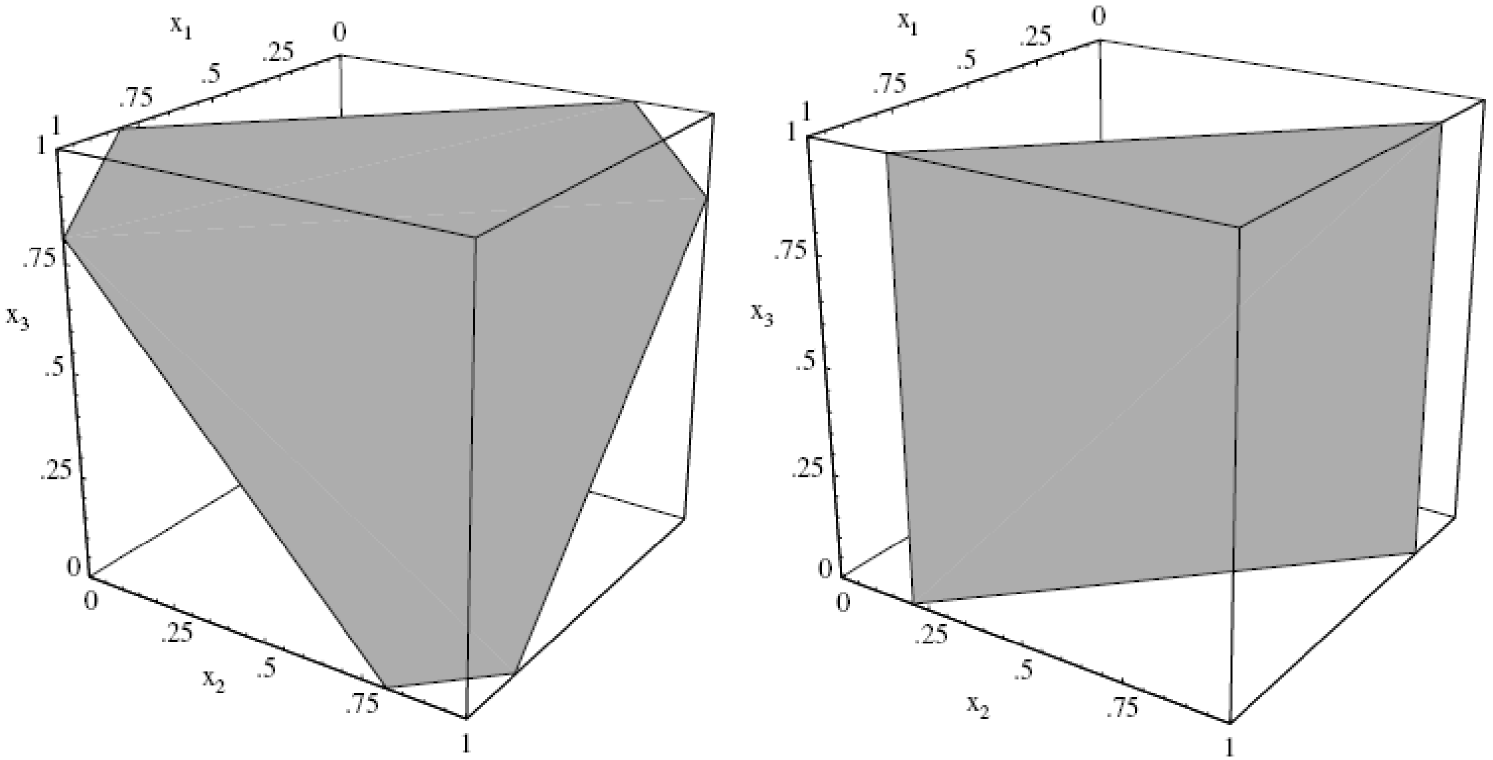
Example 3: Bayesian decisions for three-dimensional binary features

- Suppose two categories consist of independent binary features in three dimensions with known feature probabilities. Let us construct the Bayesian decision boundary if $P(\omega_1) = P(\omega_2) = 0.5$ and the individual components obey:

$$\begin{cases} p_i = 0.8 \\ q_i = 0.5 \end{cases} \qquad i = 1, 2, 3.$$

$$w_i = \ln \frac{.8(1 - .5)}{.5(1 - .8)} = 1.3863$$

$$w_0 = \sum_{i=1}^{3} \ln \frac{1-.8}{1-.5} + \ln \frac{.5}{.5} = 1.2.$$



The decision boundary for the Example involving three-dimensional binary features. On the left we show the case $p_i = 0.8$ and q_i =0.5. On the right we use the same values except $p_3 = q_3$, which leads to $w_3 = 0$ and a decision surface parallel to the x_3 axis.

- The surface g(**x**) = 0, is shown on the left of the figure, the boundary places points with two or more “yes” answers into category ω_1, since that category has a higher probability of having any feature have value 1.
- If we have:

$$\begin{cases} p_1 = p_2 = 0.8, \quad p_3 = 0.5 \\ q_1 = q_2 = q_3 = 0.5 \end{cases}$$

- In this case feature x_3 gives us no predictive information about the categories, and hence the decision boundary is parallel to the x_3 axis.

Missing and Noisy Features

- Suppose we develop a Bayes classifier using uncorrupted data, but our input (test) data are then corrupted in particular known ways. How can we classify such corrupted inputs to obtain a minimum error now?
- **Missing Features**
- Let $\mathbf{x} = [\mathbf{x}_g, \mathbf{x}_b]$, where $\mathbf{x}_g$ represents the known or "good" features and $\mathbf{x}_b$ represents the "bad" ones, i.e., either unknown or missing.
- We seek the Bayes rule given the good features, and for that the posterior probabilities are needed.

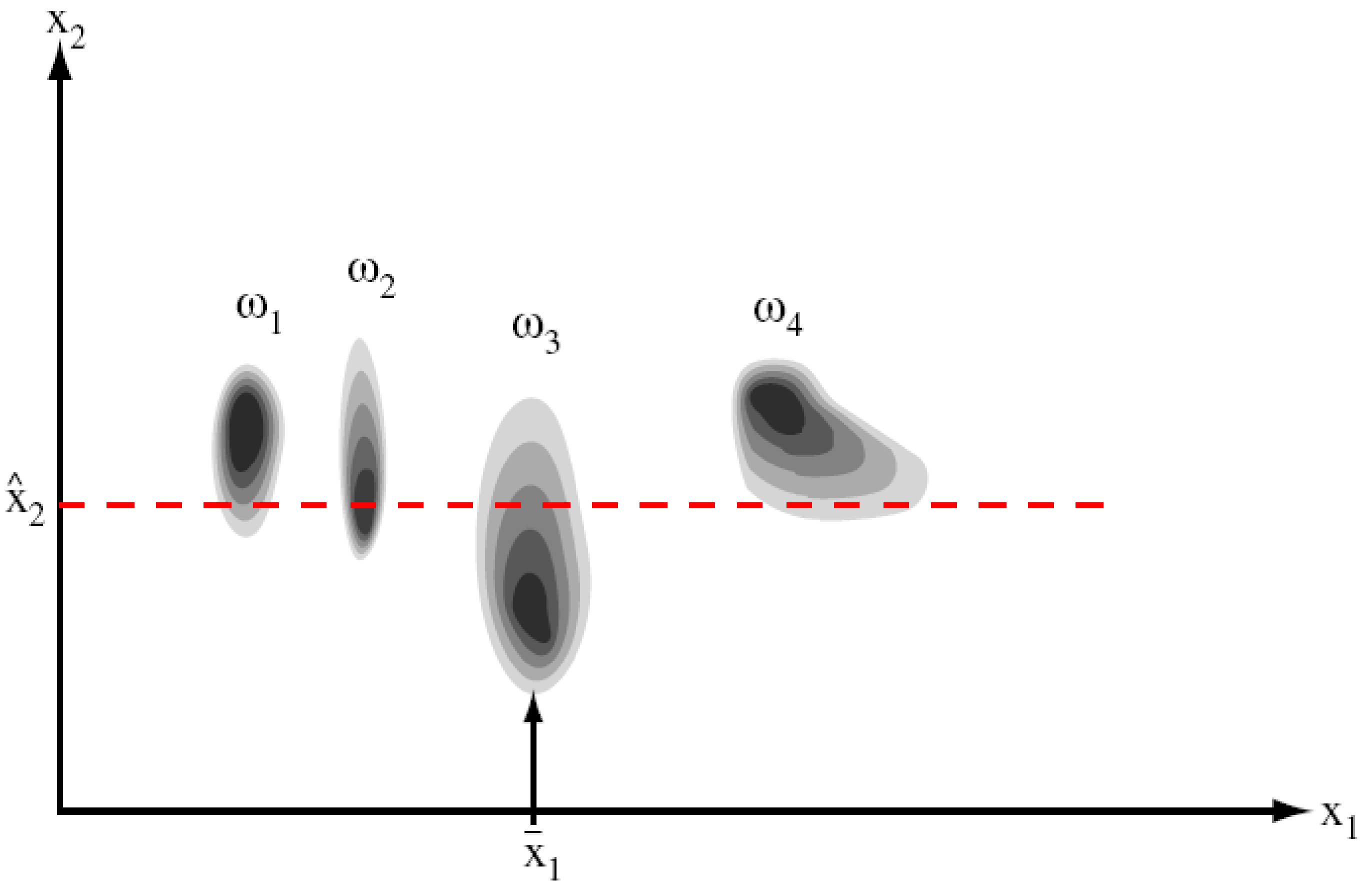


Figure 2.22: Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, x_1) and the other is measured to have value $\hat{x}_2$ (red dashed line), we want our classifier to classify the pattern as category ω_2, because $p(\hat{x}_2 \mid \omega_2)$ is the largest of the four likelihoods.

the posteriors are

$$
\begin{aligned}
P(\omega_i|\mathbf{x}_g) &= \frac{p(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b)\ d\mathbf{x}_b}{p(\mathbf{x}_g)} \\
&= \frac{\int P(\omega_i|\mathbf{x}_g, \mathbf{x}_b) p(\mathbf{x}_g, \mathbf{x}_b)\ d\mathbf{x}_b}{p(\mathbf{x}_g)} \\
&= \frac{\int g_i(\mathbf{x}) p(\mathbf{x})\ d\mathbf{x}_b}{\int p(\mathbf{x})\ d\mathbf{x}_b},
\end{aligned}
$$

where $g_i(\mathbf{x}) = g_i(\mathbf{x}_g, \mathbf{x}_b) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_b)$

is one form of our discriminant function.

We refer to $\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b)\ d\mathbf{x}_b$ as a *marginal distribution*;

we say the full joint distribution is marginalized over the variable $\mathbf{x}_b$.

Finally we use the Bayes decision rule on the resulting posterior probabilities, i.e., choose ω_i if $P(\omega_i|\mathbf{x}_g) > P(\omega_j\,|\mathbf{x}_g)$ for all i and j.

Noisy Features

A particular feature has been corrupted by statistically independent noise.

We assume we have uncorrupted (good) features $\mathbf{x}_g$, as before, and a *noise model*, expressed as $p(\mathbf{x}_b|\mathbf{x}_t)$. Here we let $\mathbf{x}_t$ denote the true value of the observed $\mathbf{x}_b$ features, i.e., without the noise present; that is, the $\mathbf{x}_b$ are observed instead of the true $\mathbf{x}_t$. We assume that if $\mathbf{x}_t$ were known, $\mathbf{x}_b$ would be independent of ω_i and $\mathbf{x}_g$. From such an assumption we get:

$$P(\omega_i|\mathbf{x}_g, \mathbf{x}_b) = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)\ d\mathbf{x}_t}{p(\mathbf{x}_g, \mathbf{x}_b)}.$$

$$p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)$$

By our independence assumption, if we know $\mathbf{x}_t$, then $\mathbf{x}_b$ does not provide any additional information about ω_i.

We have $P(\omega_i|\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_t).$

$p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = p(\mathbf{x}_b|\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)$,
and $p(\mathbf{x}_b|\mathbf{x}_g, \mathbf{x}_t) = p(\mathbf{x}_b|\mathbf{x}_t)$

$$\begin{aligned} P(\omega_i|\mathbf{x}_g, \mathbf{x}_b) &= \frac{\int P(\omega_i|\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t}{\int p(\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t} \\ &= \frac{\int g_i(\mathbf{x})p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t}{\int p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)\ d\mathbf{x}_t}, \end{aligned}$$

We use this as discriminant functions for classification in the manner dictated by Bayes.

* Compound Bayesian Decision Theory and Context

- In the fish problem our original assumption was that the sequence of types of fish was so unpredictable that the state of nature looked like a random variable. Now we consider the possibility that the consecutive states of nature might not be statistically independent. → performance improvement.
- The way in which we exploit such *context* information is somewhat different when we can wait for *n* fish to emerge and then make all *n* decisions jointly than when we must decide as each fish emerges. The first problem is a *compound* decision problem, and the second is a *sequential compound* decision problem.

- let $\boldsymbol{\omega} = (\omega(1), ..., \omega(n))^t$ be a vector denoting the n states of nature, with $\omega(i)$ taking on one of the c values $\omega_1, ..., \omega_c$.
- Let $P(\boldsymbol{\omega})$ be the prior probability for the n states of nature.
- Let $\boldsymbol{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ be a matrix giving the n observed feature vectors, with $\mathbf{x}_i$ being the feature vector obtained when the state of nature was $\omega(i)$.
- Finally, let $p(\boldsymbol{X}|\boldsymbol{\omega})$ be the conditional probability density function for $\boldsymbol{X}$ given the true set of states of nature $\boldsymbol{\omega}$. The posterior prob. of $\boldsymbol{\omega}$ is given by

$$P(\boldsymbol{\omega}|X) = \frac{p(X|\boldsymbol{\omega})P(\boldsymbol{\omega})}{p(X)} = \frac{p(X|\boldsymbol{\omega})P(\boldsymbol{\omega})}{\sum_{\boldsymbol{\omega}} p(X|\boldsymbol{\omega})P(\boldsymbol{\omega})}.$$

- One can define a loss matrix for the compound decision problem and seek a decision rule that minimizes the compound risk.
- In practice the computation of $P(\boldsymbol{\omega}|\boldsymbol{X})$ can easily prove to be an enormous task. There are c^n possible values of $\boldsymbol{\omega}$ to consider.
- If the distribution of the feature vector $\mathbf{x}_i$ depends only on the corresponding state of nature $\omega(i)$, not on the values of the other feature vectors or the other states of nature. Then the joint density $p(\boldsymbol{X}|\boldsymbol{\omega})$ is merely the product of the component densities $p(\mathbf{x}_i|\omega(i))$:

$$p(X|\boldsymbol{\omega}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\omega(i)).$$

Homework#1

- Textbook: Chapter 2
- 2.1, 2.2, 2.3, 2.4
- 2.5, 2.6, 2.23
- 2.13, 2.14, 2.24, 2.31

Computer Assignment #1

Computer Exercise 2.1, 2.2