Chapter 2
BAYESIAN DECISION THEORY

Wei-Yang Lin

Department of Computer Science

& Information Engineering

mailto:wylin@cs.ccu.edu.tw

2.1 **Introduction**

2.2 Bayesian Decision Theory

2.3 Minimum-Error-Rate Classification

2.4 Classifiers, Discriminant Functions, and Decision Surfaces

2.5 The Normal Density

2.6 Discriminant Functions for the Normal Density

2.8.3 Signal Detection Theory and Operating Characteristics

2.1 Introduction

- Bayesian decision theory is a fundamental approach to the problem of classification.
- This approach is based on quantifying the tradeoffs between various decisions and the costs that accompany such decisions.
- It makes the assumption that all the relevant probability values are known.

- While we will give a quite general development of Bayesian decision theory, we begin our discussion with a specific example.

- We let ω denote the **state of nature**, with $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon. discrete

- Because the state of nature is unpredictable, we consider ω to be a random variable.

- If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon.

大寫

- More generally, we assume that there is some **prior probability** $P(\omega_1)$ that the next fish is sea bass, and some prior probability $P(\omega_2)$ that it is salmon.
- If we assume there are no other types of fish relevant here, then $P(\omega_1)$ and $P(\omega_2)$ sum to one.

- Suppose that we were forced to make a decision about the type of fish that will appear next.
- If a decision must be made with so little information, it seems logical to use the following **decision rule**:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1) > P(\omega_2)\text{; otherwise decide } \omega_2.$$

- In most circumstance we are not asked to make decisions with so little information.

- In our example, we might use a lightness measurement x to improve our classifier.

- We consider x to be a continuous random variable whose distribution depends on the state of nature and is expressed as $p(x|\omega)$[a].

特徵

discrete

continue

[a] We use an uppercase $P(\cdot)$ to denote a probability mass function and use a lowercase $p(\cdot)$ to denote a probability density function

- $p(x|\omega)$, the probability density function for x given that the sate of nature is ω, is called the **class-conditional probability density**.

- The difference between $p(x|\omega_1)$ and $p(x|\omega_2)$ describes the difference in lightness between populations of sea bass and salmon (Figure 1).

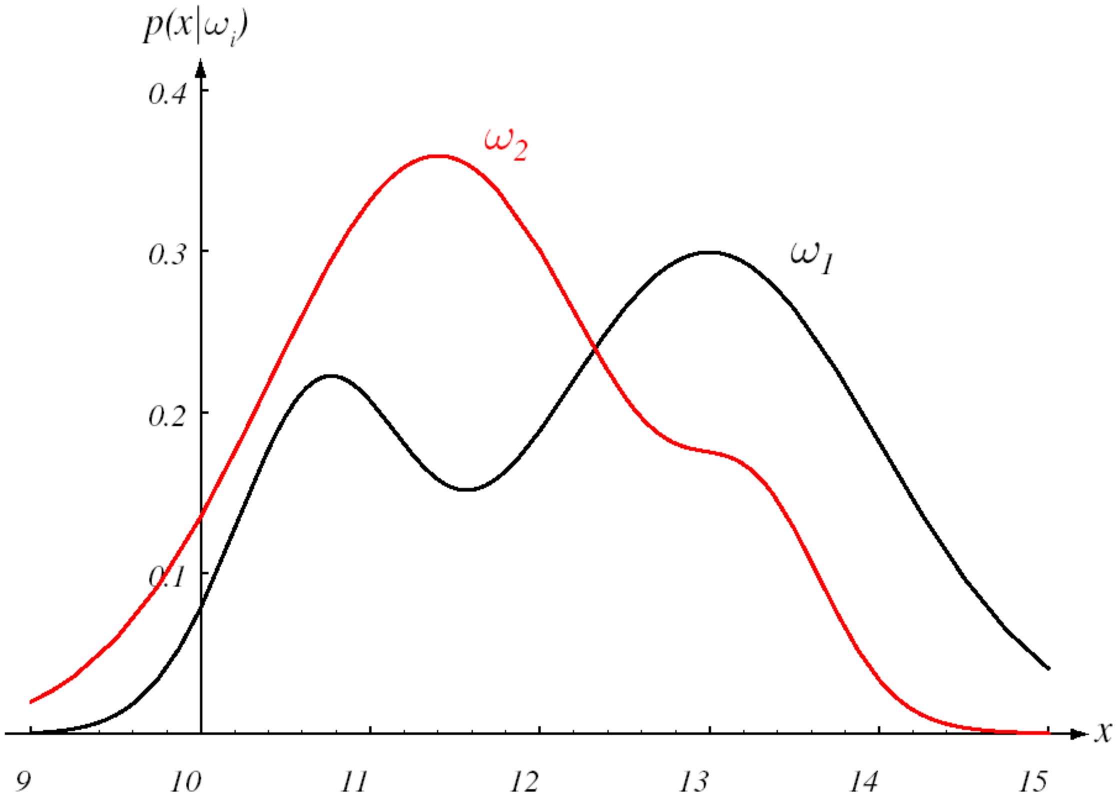


Figure 1: Class-conditional density functions show the distribution of x given the pattern is in category ω_i.

- Suppose that we know both the prior probabilities $P(\omega_j)$ and the conditional densities $p(x|\omega_j)$ for $j = 1, 2$.

- How does this change our attitude concerning the true state of nature?

- We note first that the joint probability density of a pattern that is in category ω_j and has feature value x can be written in two ways:

$$p(\omega_j, x) = \underline{P}(\overset{\text{discrete}}{\omega_j}|x)p(x) = \underline{p}(\overset{\text{continue}}{x}|\omega_j)P(\omega_j) \tag{1}$$

- Rearrange Equation (1) leads us to the answer to our question, which is called **Bayes formula**:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} \tag{2}$$

where in this case of two categories

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j) \tag{3}$$

- Bayes formula can be expressed formally in English by saying that

$$posterior = \frac{likelihood \times prior}{evidence} \tag{4}$$

- Bayes formula shows that by observing the value of x we can convert the prior probability $P(\omega_j)$ to the **a posteriori probability** (or **posterior**) $P(\omega_j|x)$, the probability of the state of nature being ω_j given that feature value x has been measured.

- We call $p(x|\omega_j)$ the **likelihood** of ω_j with respect to x, a term chosen to indicate that the category ω_j for which $p(x|\omega_j)$ is large is more likely to be the true category.
- Note that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability.

- The **evidence**, $p(x)$, can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one.
- The variation of $P(\omega_j|x)$ with x is illustrated in Figure 2.

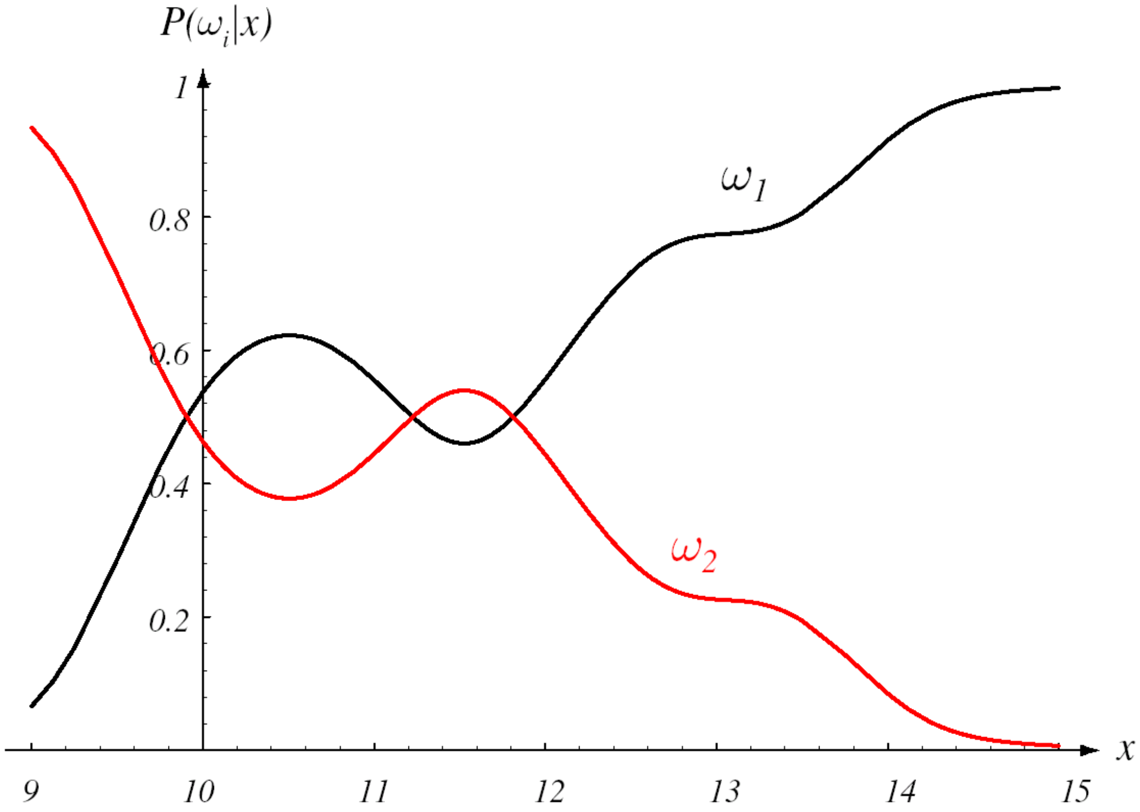
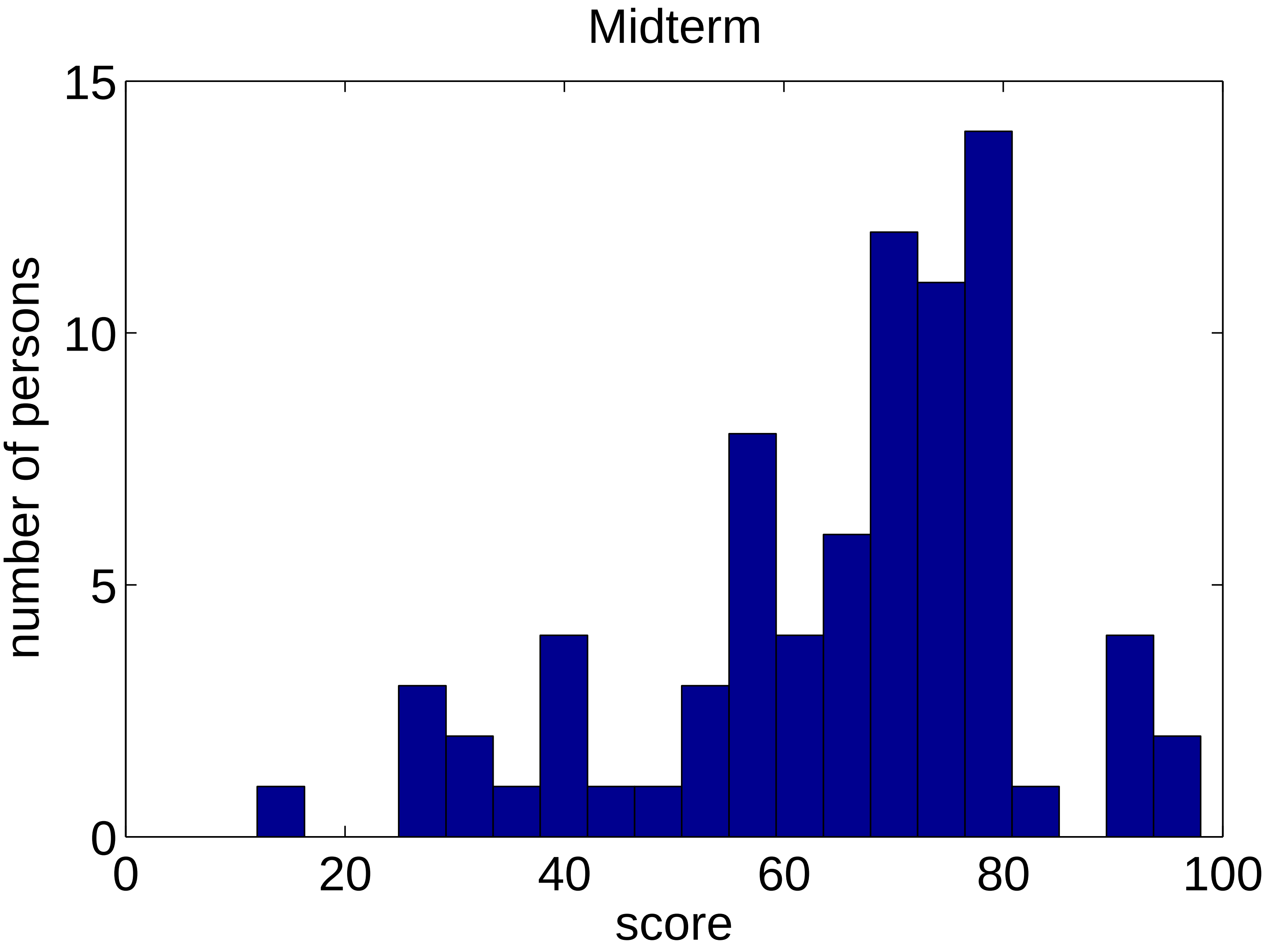
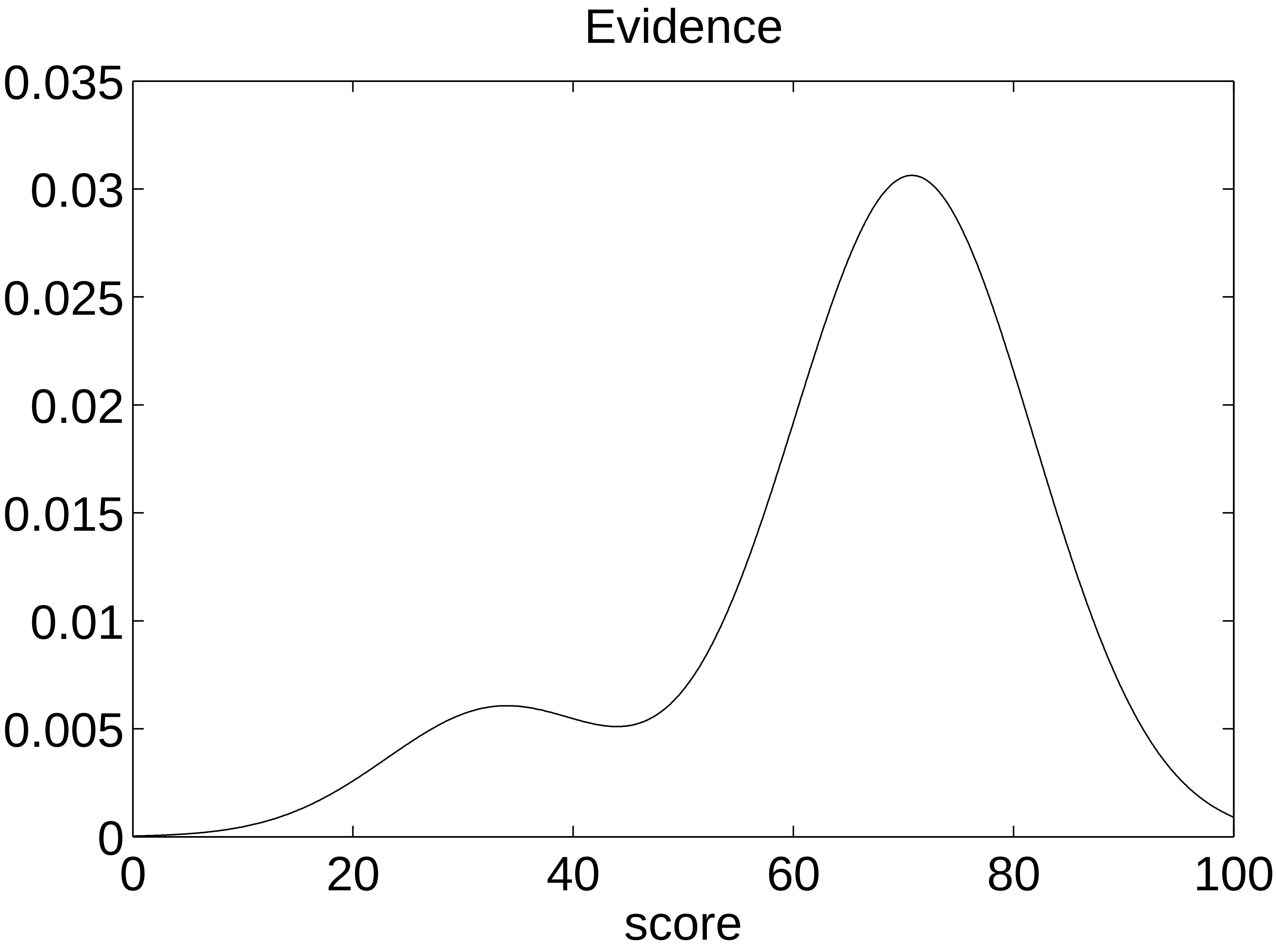
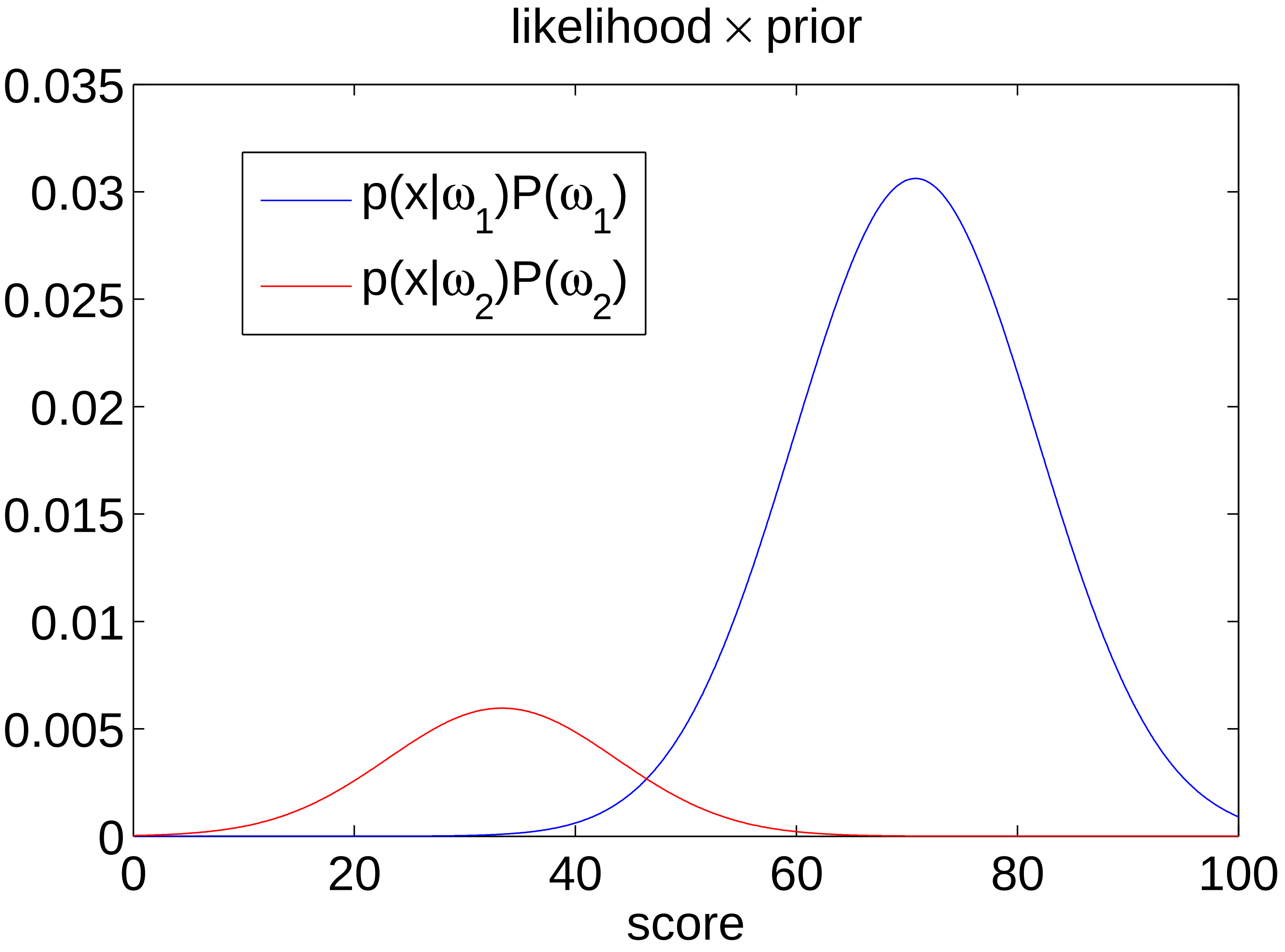


Figure 2: Posterior probabilities for $P(\omega_1) = 2/3$, $P(\omega_2) = 1/3$, and the class-conditional densities shown in Figure 1.


Midterm
number of persons
score
0
5
10
15
0
20
40
60
80
100


Evidence
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
score


likelihood × prior
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
20
40
60
80
100
score
$p(x|\omega_1)P(\omega_1)$
$p(x|\omega_2)P(\omega_2)$

- If we have an observation x for which $P(\omega_1|x)$ is greater than $P(\omega_2|x)$, we would naturally be inclined to decide that the true state of nature is ω_1.
- Conversely, if $P(\omega_2|x)$ is greater than $P(\omega_1|x)$, we would naturally be inclined to choose ω_2.
- To justify this decision procedure, let us calculate the probability of error whenever we make a decision.

- Whenever we observe a particular x, the probability of error is

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1. \end{cases} \tag{5}$$

- Clearly, for a given x we can minimize the probability of error by deciding ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and ω_2 otherwise.

- But, will this rule minimize the average probability of error?

- Yes, because the average probability of error is given by

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error|x)p(x)dx \qquad (6)$$

- If for every x we ensure that $P(error|x)$ is as small as possible, then the integral must be as small as possible.

- Thus, we have justified the following **Bayes decision rule** for minimizing the probability of error:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) > P(\omega_2|x); \text{ otherwise decide } \omega_2.$$

- This form of decision rule emphaiszes the importance of the posterior probabilities.
- By using Eq. 2, we can express the rule in terms of the conditional densities and prior probabilities.

  Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

- Note that the evidence $p(x)$ in Eq. 2 is unimportant as far as making a decision is concerned.

- If for some x we have $p(x|\omega_1) = p(x|\omega_2)$, then that particular observation gives us no information about the state of nature.
  - In this case, the decision hinges entirely on the prior probabilities.
- On the other hand, if $p(\omega_1) = p(\omega_2)$, then the state of natures are equally probable.
  - In this case, the decision is based entirely on the likelihoods.
- In general, both of these factors are important and the Bayes decision rule combines them to achieve the minimum error rate.

2.1 Introduction

2.2 **Bayesian Decision Theory**

2.3 Minimum-Error-Rate Classification

2.4 Classifiers, Discriminant Functions, and Decision Surfaces

2.5 The Normal Density

2.6 Discriminant Functions for the Normal Density

2.8.3 Signal Detection Theory and Operating Characteristics

2.2 Bayesian Decision Theory

- We shall now generalize the ideas just considered in four ways:
  - By allowing the use of more than one feature
  - By allowing more than two states of nature
  - By allowing actions other than merely deciding the state of nature
  - By introducing a loss function

- Allowing the use of more than feature merely requires replacing the scalar x by the **feature vector** $\mathbf{x} \in \mathbb{R}^d$.
- Allowing more than two states of nature provides us with a useful generalization.
- Allowing actions other than classification primarily allows the possibility of rejection.
- The **loss function** states exactly how costly each action is.

- Let $\{\omega_1, \ldots, \omega_c\}$ be the set of c states of nature and let $\{\alpha_1, \ldots, \alpha_a\}$ be the set of a possible actions.
- The loss function $\lambda(\alpha_i|\omega_j)$ describes the loss incurred for taking action α_i when the state of nature is ω_j.
- Let $\mathbf{x}$ be a d-dimensional random vector and let $p(\mathbf{x}|\omega_j)$ be the state-conditional probability density function for $\mathbf{x}$.
- As before, $P(\omega_j)$ describes the prior probability that the state of nature is ω_j.

- The posterior probability $P(\omega_j|\mathbf{x})$ can be computed from $P(\mathbf{x}|\omega_j)$ by Bayes formula:

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})} \tag{7}$$

where the evidence is

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j) \tag{8}$$

- Suppose that we observe a particular $\mathbf{x}$ and that we contemplate taking action α_i.

- If the true state of nature is ω_j, we will incur the loss $\lambda(\alpha_i|\omega_j)$.

- The expected loss associated with taking action α_i is

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \tag{9}$$

- An expected loss is called a **risk**, and $R(\alpha_i|\mathbf{x})$ is called the **conditional risk**.

- We shall now show that the **Bayes decision rule** actually provides the optimal performance.

- Stated formally, our problem is to find a decision rule that minimizes overall risk.

- To be more specific, for every $\mathbf{x}$ the **decision rule** $\alpha(\mathbf{x})$ assumes one of the values $\alpha_1, \ldots, \alpha_a$.

- The overall risk R is the expected loss associated with a given decision rule.

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{10}$$

where $d\mathbf{x}$ is the notation for a d-dimensional volume element.

- Clearly, if $\alpha(\mathbf{x})$ is chosen so that $R(\alpha(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$, then the overall risk will be minimized.

- This justifies the Bayes decision rule: Compute the conditional risk $R(\alpha_i|\mathbf{x})$ for $i = 1, \ldots, a$ and then select the action with minimal conditional risk.

2.2.1 Two-Category Classification

- Let us consider the special case of two-category classification problems.
- Here, action α_1 corresponds to deciding that the true state of nature is ω_1.
- Similarly, action α_2 corresponds to deciding ω_2.
- For notational simplicity, let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be the loss incurred for deciding ω_i when the true state of nature is ω_j.

- The conditional risks are given by

$$
\begin{aligned}
R(\alpha_1|\mathbf{x}) &= \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \qquad (11)\\
R(\alpha_2|\mathbf{x}) &= \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}) \qquad (12)
\end{aligned}
$$

- Thus, the resulting decision rule is to decide ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$.

- We can obtain an equivalent rule as follows:

$$\text{Decide } \omega_1 \text{ if } (\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}) \tag{13}$$

- Ordinarily, the loss incurred for making an error is greater than the loss incurred for being correct, i.e. $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are positive.

- Under this assumption, another alternative is to decide ω_1 if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \cdot \frac{P(\omega_2)}{P(\omega_1)} \tag{14}$$

- In other words, decide ω_1 if the **likelihood ratio** $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$ exceeds a threshold value that is independent of the observation $\mathbf{x}$.

2.1 Introduction

2.2 Bayesian Decision Theory

2.3 **Minimum-Error-Rate Classification**

2.4 Classifiers, Discriminant Functions, and Decision Surfaces

2.5 The Normal Density

2.6 Discriminant Functions for the Normal Density

2.8.3 Signal Detection Theory and Operating Characteristics

2.3 Minimum-Error-Rate Classification

- In classification problems, the action α_i is usually interpreted as the decision that the true state of nature is ω_i.
- If a decision α_i is taken and the true state of nature is ω_j, then the decision is wrong.
- It is natural to seek a decision rule that minimizes the probability of error.

- The loss function for this case is the so-called **zero-one loss function**.

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \cdots, c \tag{15}$$

- Thus, no loss is assigned to a correct decision and a unit loss is assigned to any error.

- The conditional risk is

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \\ &= \sum_{j\neq i} P(\omega_j|\mathbf{x}) \qquad \text{(P.14, sum=1)} \\ &= 1 - P(\omega_i|\mathbf{x}) \qquad (16) \end{aligned}$$

- Thus, to minimize the average probability of error, we should select the i that maximizes the posterior probability $P(\omega_i|\mathbf{x})$.

- In other words, for **minimum error rate**:

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \qquad \text{for all } j \neq i$$

- The region where we decide ω_i is denoted $\mathcal{R}_i$; such a region need not to be simply connected (Figure 3).

- If we employ a zero-one loss, the decision regions are determined by θ_a.

- If we penalizes miscategorizing ω_2 as ω_1 more than the converse, we get the larger θ_b. Hence, $\mathcal{R}_1$ becomes smaller.

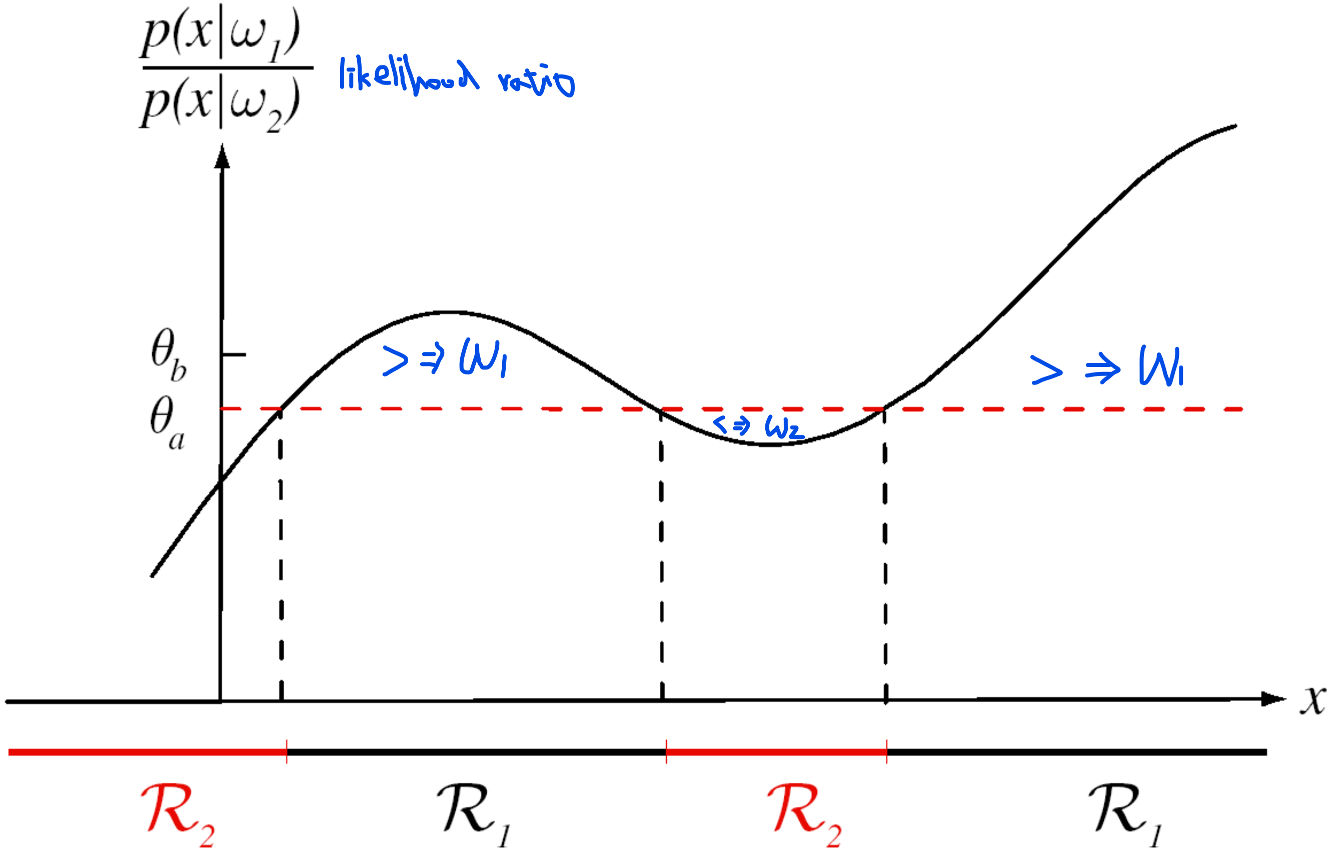


Figure 3: The likelihood ratio for the conditional-densities shown in Figure 1.

2.3.1 Minimax Criterion

- Sometimes we might expect a classifier to perform well over a range of prior probabilities.
- For instance, in the fish categorization problem, the prior probabilities might vary widely and in an unpredictable way.
- Or, we want to use a classifier in different plants where we do not know the prior probabilities.
- A reasonable approach is to design a classifier so that the worst overall risk for is as small as possible.

- Let $\mathcal{R}_1$ denote that region where a classifier decides ω_1 and likewise for $\mathcal{R}_2$ and ω_2.

- Then, we write the overall risk Eq. (10) in terms of conditional risk:

$$\begin{aligned} R &= \int_{\mathcal{R}_1} [\lambda_{11} P(\omega_1) p(\mathbf{x}|\omega_1) + \lambda_{12} P(\omega_2) p(\mathbf{x}|\omega_2)] d\mathbf{x} \\ &+ \int_{\mathcal{R}_2} [\lambda_{21} P(\omega_1) p(\mathbf{x}|\omega_1) + \lambda_{22} P(\omega_2) p(\mathbf{x}|\omega_2)] d\mathbf{x} \qquad (17) \end{aligned}$$

- We use the fact that $P(\omega_2) = 1 - P(\omega_1)$ and that $\displaystyle\int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1)d\mathbf{x} = 1 - \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)d\mathbf{x}$ to write the overall risk as:

$$\begin{aligned} R(P(\omega_1)) &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)d\mathbf{x} \\ &+ P(\omega_1) \Bigg[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)d\mathbf{x} \\ &- (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)d\mathbf{x} \Bigg] \end{aligned} \tag{18}$$

- The Eq. (18) shows that the overall risk is linear in $P(\omega_1)$.
- If we can find a decision rule such that the second term in Eq. (18) is zero, then the risk is independent of priors.
- The resulting **minimax risk**, R_{mm}, is:

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(\mathbf{x}|\omega_2) d\mathbf{x} \\ &= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{R_2} p(\mathbf{x}|\omega_1) d\mathbf{x} \qquad (19) \end{aligned}$$

- Figure 4 illustrates the approach.
- The curve at the bottom shows the overall error rate as a function of $P(\omega_1)$.
- For each value (e.g., $P(\omega_1) = 0.25$), there is a corresponding decision rule and associated overall error rate.
- For a fixed decision rule, if the priors are changed, the overall error rate will change as a linear function of $P(\omega_1)$ (shown by the dashed line).
- The maximum such error will occur at an extreme value of the prior, here at $P(\omega_1) = 1$.

- To minimize the maximum of such error, we should design a classifier for the maximum overall error rate, and thus the overall error will not change as a function of prior.
- Briefly stated, we search for the prior for which the Bayes risk is maximum, and the corresponding decision rule then gives the minimax solution.
- The value of the minimax risk, R_{mm}, is hence equal to the worst Bayes risk.

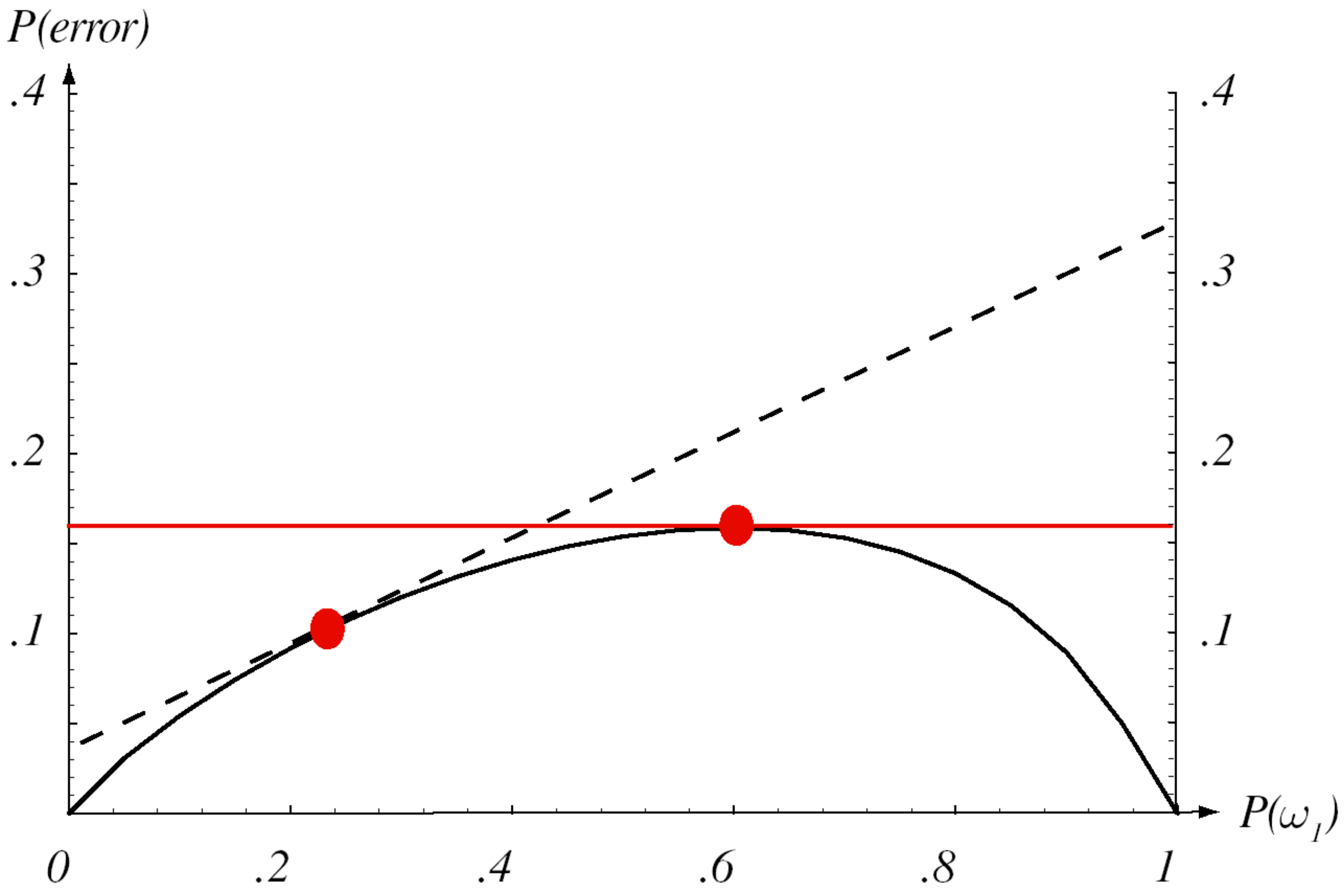


Figure 4: The overall error will not change as a function of prior, as shown by the red line.

- The minimax criterion finds greater use in game theory than it does in typical pattern recognition.
- In game theory, you have a hostile opponent who can be expected to take an action maximally detrimental to you.
- Thus, it makes great sense for you to take an action so that the worst case cost is minimized.

Example 1

Assume $p(x|\omega_1) \sim N(0, 1)$ and $p(x|\omega_2) \sim N(2, 1)$, under a zero-one loss. Please specify the decision regions based on minimax criterion. What is the resulting overall risk?

Answers:

(a) $\int_{R_2} p(x|w_1)dx = \int_{R_1} p(x|w_2)dx$

$$\Rightarrow \int_{\theta}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[\frac{-1}{2}x^2]dx = \int_{-\infty}^{\theta} \frac{1}{\sqrt{2\pi}} \exp[\frac{-1}{2}(x-2)^2]dx$$

$$\Rightarrow \Phi(-\theta) = \Phi(\frac{\theta - 2}{1})$$

$$\Rightarrow -\theta = \theta - 2$$

$$\Rightarrow \theta = 1$$

(b) $$R = \int_{R_1} p(x|w_2)dx$$

$$= \int_{-\infty}^{1} p(x|w_2)dx$$

$$= \Phi(\frac{1-2}{1})$$

$$= \Phi(-1)$$

$$= 0.1587$$

Example 2

Consider two one-dimensional normal distributions $p(x|\omega_1) \sim N(-1,1)$ and $p(x|\omega_2) \sim N(1,1)$ with zero-one loss function.

(a) For each value of $P(\omega_1)$, there is a corresponding Bayes decision rule. The resulting overall risk is called the Bayes risk. Plot Bayes risk as a function of $P(\omega_1)$.

(b) Please state the Bayes decision rule for $P(\omega_1) = 0.25$. For this decision rule, if the priors are then changed, the overall risk will change as well. For this fixed classifier, plot the over risk as a function of $P(\omega_1)$.

(c) Please state the minimax decision rule and calculate the minimax risk. For the minimax classifier (fixed), plot the overall risk as a function of $P(\omega_1)$.

Answers:

(a)

$$\begin{aligned}
\frac{p(\theta|\omega_1)}{p(\theta|\omega_2)} &= \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\cdot\frac{P(\omega_2)}{P(\omega_1)} \\
\frac{\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-(\theta+1)^2}{2}\right]}{\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-(\theta-1)^2}{2}\right]} &= \frac{(1-0)}{(1-0)}\cdot\frac{1-P(\omega_1)}{P(\omega_1)} \\
\exp\left[-2\theta\right] &= \frac{1-P(\omega_1)}{P(\omega_1)} \\
\theta &= \frac{-1}{2}\cdot\ln\frac{1-P(\omega_1)}{P(\omega_1)}
\end{aligned}$$

(a)

$$
\begin{aligned}
R(P(\omega_1)) &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(x|\omega_2)dx \\
&+ P(\omega_1) \left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(x|\omega_1)dx \right. \\
&- \left. (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(x|\omega_2)dx \right] \\
&= \int_{-\infty}^{\theta} p(x|\omega_2)dx + P(\omega_1) \left[\int_{\theta}^{\infty} p(x|\omega_1)dx \right. \\
&- \left. \int_{-\infty}^{\theta} p(x|\omega_2)dx \right]
\end{aligned}
$$

(a)

$$\begin{aligned} R(P(\omega_1)) &= \Phi(\frac{\theta-1}{1}) + P(\omega_1)[1-\Phi(\frac{\theta+1}{1}) - \Phi(\frac{\theta-1}{1})] \\ &= \Phi(\theta-1) + P(\omega_1)[1-\Phi(\theta+1)-\Phi(\theta-1)] \end{aligned}$$

- Note that the $X \sim N(m, \sigma^2)$ cdf can always be expressed using the standard normal cdf (See page 187 in [1]).

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\big[\frac{-1}{2}\big(\frac{\tau - m}{\sigma}\big)^2\big] d\tau = \Phi(\frac{x-m}{\sigma}).$$

(b)

$$\begin{aligned} \theta &= \frac{-1}{2}\cdot\ln\frac{1-0.25}{0.25} = -0.5493 \\ R(P(\omega_1)) &= \Phi(\theta-1) + P(\omega_1)[1-\Phi(\theta+1)-\Phi(\theta-1)] \end{aligned}$$

So, we decide ω_1 if $x < -0.5493$. Otherwise, decide ω_2.

(c) $\int_{R_2} p(x|w_1)dx = \int_{R_1} p(x|w_2)dx$

$$\Rightarrow \int_{\theta}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[\frac{-1}{2}(x+1)^2]dx = \int_{-\infty}^{\theta} \frac{1}{\sqrt{2\pi}} \exp[\frac{-1}{2}(x-1)^2]dx$$

$$\Rightarrow 1 - \Phi(\frac{\theta+1}{1}) = \Phi(\frac{\theta-1}{1})$$

$$\Rightarrow \Phi(-\theta-1) = \Phi(\theta-1)$$

$$\Rightarrow \theta = 0$$

So, we decide ω_1 if $x < 0$. Otherwise, decide ω_2.

(c)

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} p(x|\omega_2)dx \\ &= \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp[\frac{-1}{2}(x-1)^2]dx \\ &= \Phi(0-1) = 0.1587 \end{aligned} \tag{20}$$

2.3.2 Neyman-Pearson Criterion

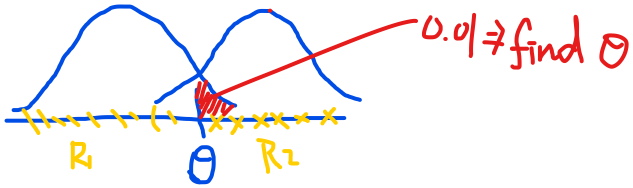
- In some problems, we might wish to minimize the total risk subject to a constraint $\int R(\alpha_i|\mathbf{x})d\mathbf{x} < constant$ for some particular i.
- Such a constraint might arise when there is a fixed resource that accompanies one particular action α_i.

- For instance, in an international airport, there might be a policy that we must **NOT** misclassify more than 0.1% of travelers as terrorists. We might then seek a decision rule that maximizes the chance of detecting a terrorist correctly subject to this constraint.

- We generally satisfy such a **Neyman-Pearson criterion** by adjusting decision boundaries numerically.

- For Gaussian and some other distributions, Neyman-Pearson solutions can be found analytically (Problem 6 and 7).

Example 3



單变

Consider the Neyman-Person criterion for two univariate normal distributions: $p(x|\omega_1) \sim N(-2, 2)$, $p(x|\omega_2) \sim N(2, 2)$ and $P(\omega_1) = P(\omega_2)$.

mean σ^2, variance

1. Suppose the maximum acceptable error rate for classifying a pattern that is actually in ω_1 as if it were in ω_2 is 0.01. What is the resulting single-point decision boundary?

2. For this boundary, what is the error rate for classifying ω_2 as ω_1?

3. What is the overall error rate? Also, compare your result to the Bayes error rate.

$$\frac{P(x,W_1)}{P(x,W_2)} = \frac{\overset{1}{\lambda_{12}} - \overset{0}{\lambda_{22}}}{\overset{1}{\lambda_{21}} - \overset{0}{\lambda_{11}}} \, \frac{\cancel{P(W_2)}}{\cancel{P(W_1)}}$$

Answers:

(a) $E_1 = 1 - \int_{-\infty}^{\theta} p(\mathbf{x}|w_1)d\mathbf{x}$

$$\Rightarrow E_1 = 1 - \overbrace{\Phi(\frac{\theta - \mu_1}{\sigma_1})}^{\text{standard normal distribution}}$$

$$\Rightarrow E_1 = 0.01 = 1 - \Phi(\frac{\theta + 2}{\sqrt{2}})$$

$$\Rightarrow \theta = \Phi^{-1}(0.99) \times \sqrt{2} - 2 = 2.33 \times \sqrt{2} - 2 = 1.295$$

(b) $E_2 = \int_{-\infty}^{\theta} p(\mathbf{x}|w_2) d\mathbf{x}$

$$= \Phi(\frac{\theta - \mu_2}{\sigma_2})$$

$$= \Phi(\frac{1.295 - 2}{\sqrt{2}})$$

$$= \Phi(-0.5)$$

$$= 0.3085$$

$$\Rightarrow \text{error rate} = E_2 \times P(w_2) = 0.3085 \times \frac{1}{2} = 0.15425$$

(c) Overall error rate of NP $= E_1 \times P(w_1) + E_2 \times P(w_2) = 0.15925$

For Bayes, $\theta = \frac{1}{2}(\mu_1 + \mu_2) = 0$

$$\Rightarrow E_1 = 1 - \Phi(\frac{0-(-2)}{\sqrt{2}}) = 0.0793$$

$$\Rightarrow E_2 = \Phi(\frac{0-2}{\sqrt{2}}) = 0.0793$$

Overall error rate of Bayes $= E_1 \times P(w_1) + E_2 \times P(w_2) = 0.0793$

2.1 Introduction

2.2 Bayesian Decision Theory

2.3 Minimum-Error-Rate Classification

2.4 **Classifiers, Discriminant Functions, and Decision Surfaces**

2.5 The Normal Density

2.6 Discriminant Functions for the Normal Density

2.8.3 Signal Detection Theory and Operating Characteristics

2.4.1 The Multicategory Case

- There are many different ways to represent pattern classifiers.
- One of the most useful is in terms of a set of discriminant functions $g_i(\mathbf{x}), i = 1, \ldots, c$.
- The classifier is said to assign a feature vector $\mathbf{x}$ to class ω_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \text{for all } j \neq i \tag{21}$$

- A Bayes classifier is naturally represented in this way.
- For the general case, we can let $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$ because the maximum discriminant function corresponds to the minimum conditional risk.
- For the minimum-error-rate case, we can simplify things by taking $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$ so that the maximum discriminant function corresponds to the maximum posterior probability.

- Clearly, the choice of discriminant functions is not unique.
- In general, if we replace every $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$, where $f(\cdot)$ is a monotonically increasing function, the resulting classification is unchanged.
- This observation can lead to analytical and computational simplifications.

- For instance, any of the following choices gives identical classification results, but some can be much simpler to compute than others:

$$\begin{aligned} g_i(\mathbf{x}) &= P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)} && (22) \\ g_i(\mathbf{x}) &= p(\mathbf{x}|\omega_i)P(\omega_i) && (23) \\ g_i(\mathbf{x}) &= \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) && (24) \end{aligned}$$

  where ln denotes natural logarithm.

- Even though the discriminant functions can be written in a variety of forms, the decision rules are equivalent.

- The effect of any decision rule is to divide the feature space into c **decision regions**, $\mathcal{R}_1, \ldots, \mathcal{R}_c$.

- If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$, then the decision rule calls for us to assign $\mathbf{x}$ to ω_i, i.e., $\mathbf{x}$ is in $\mathcal{R}_i$.

- The decision regions are separated by **decision boundaries** (Figure 5).

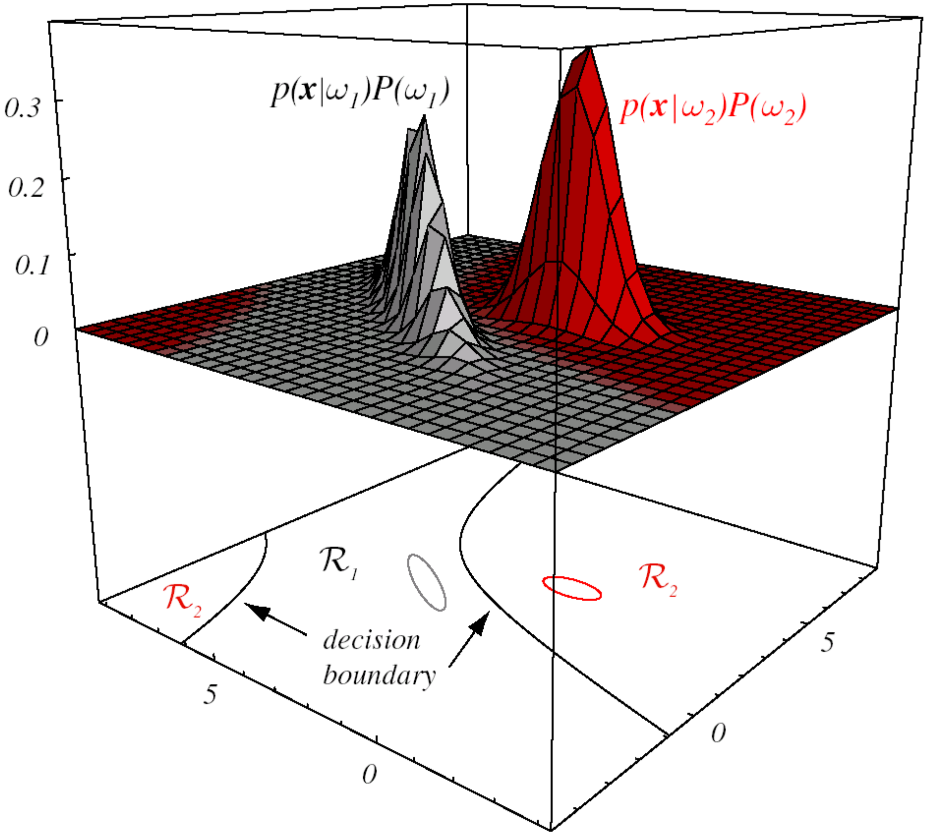


Figure 5: The decision boundaries consist of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected.

2.4.2 The Two-Category Case

- While the two-category problem is just a special case, it has traditionally received much more attentions.
- Instead of using two discriminant functions, it is more common to define a single discriminant function

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) \tag{25}$$

and the decision rule is to

$$\text{Decide } \omega_1 \text{ if } g(\mathbf{x}) > 0\text{; otherwise decide } \omega_2.$$

- Of the various forms in which the minimum-error-rate discriminant function can be written, the following two (derived from Eqs. (22) and (24)) are particularly convenient:

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \tag{26}$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \tag{27}$$

Example 4

Consider a two-class classification problem. Given the class-conditional probability density functions

$$
\begin{aligned}
p(x|\omega_1) &\sim N(0,1) \\
p(x|\omega_2) &\sim N(0,2)
\end{aligned}
$$

with $P(\omega_1) = 0.3$ and $P(\omega_2) = 0.7$. Please specify the decision regions based on Bayesian decision theory.

Answers:

$$\frac{p(x|w_1)P(w_1)}{p(x|w_2)P(w_2)} = \frac{\sigma_2 P(w_1)e^{-\frac{1}{2}(\frac{x}{\sigma_1})^2}}{\sigma_1 P(w_2)e^{-\frac{1}{2}(\frac{x}{\sigma_2})^2}}$$

By taking logarithm, we have

$$\ln\{\frac{\sigma_2 P(w_1)}{\sigma_1 P(w_2)}\} - \frac{1}{2}x^2(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}) = -0.5007 - \frac{x^2}{4} < 0$$

$\Rightarrow$ We always decide w_2

$\Rightarrow R_2 = [-\infty, \infty],\ R_1 = \{\emptyset\}$

2.1 Introduction

2.2 Bayesian Decision Theory

2.3 Minimum-Error-Rate Classification

2.4 Classifiers, Discriminant Functions, and Decision Surfaces

2.5 **The Normal Density**

2.6 Discriminant Functions for the Normal Density

2.8.3 Signal Detection Theory and Operating Characteristics

2.5 The Normal Density

- Of the various density functions, none has received more attention than the multivariate normal or Gaussian density.
- To a large extent, this attention is due to its analytical tractability.
- In this section, we provide a brief exposition of the multivariate normal density.

- Recall the definition of the **expected value** of a scalar function $f(x)$, defined for some density $p(x)$:

$$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx \tag{28}$$

- If the values of x are restricted to points in a discrete set $\mathcal{D}$, we must sum over all samples as

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{D}} f(x)P(x) \tag{29}$$

where $P(x)$ is a probability mass function.

2.5.1 Univariate Density

- We begin with the univariate normal density,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \frac{(x-m)^2}{\sigma^2} \right] \tag{30}$$

- The expected value of x is

$$m = \mathcal{E}[x] = \int_{-\infty}^{\infty} x p(x) dx, \tag{31}$$

- The **variance** of x is

$$\sigma^2 = \mathcal{E}[(x-m)^2] = \int_{-\infty}^{\infty} (x-m)^2 p(x) dx, \tag{32}$$

- For simplicity, we often abbreviate Eq. (30) by writing $p(x) \sim N(m, \sigma^2)$ to say that x is distributed normally with mean m and variance σ^2.

- Samples from a normal distribution tend to cluster about the mean, with a spread related to the standard deviation σ (Figure 6).

- Moreover, as stated by the **Central Limit Theorem**, the sum of a large number of independent random variables will lead to a Gaussian distribution (computer exercise 5).

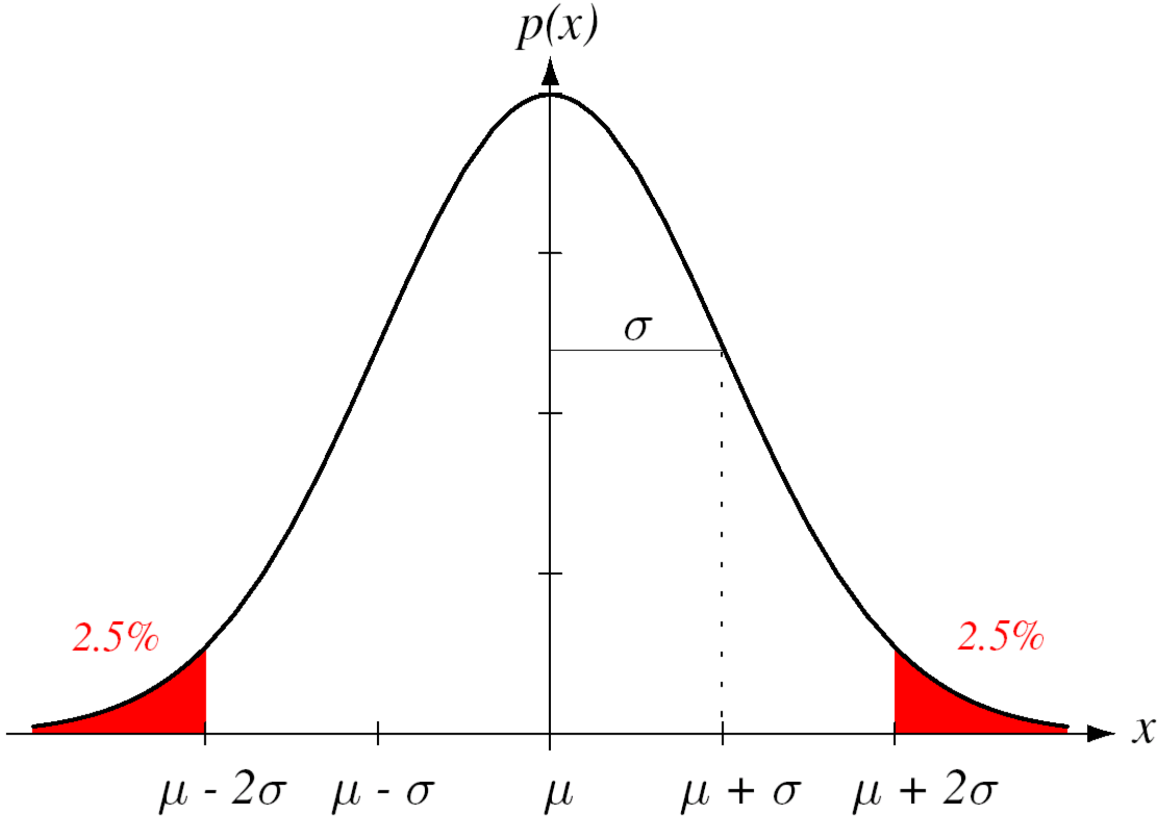


Figure 6: A univariate normal distribution has roughly 95% of its area in the range $|x - m| \leq 2\sigma$, as shown.

2.5.2 Multivariate Density

- The multivariate normal density is

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \qquad (33)$$

where $\mathbf{x}$ is a d-component column vector, $\boldsymbol{\mu}$ is a d-component mean vector, $\mathbf{\Sigma}$ is a d-by-d **covariance matrix**, and $|\mathbf{\Sigma}|$ and $\mathbf{\Sigma}^{-1}$ are its determinant and inverse, respectively. Also, we let $(\mathbf{x}-\boldsymbol{\mu})^t$ denote the transpose of $\mathbf{x}-\boldsymbol{\mu}$.

- Formally, we have

$$\boldsymbol{\mu} = \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}, \tag{34}$$

and

$$\boldsymbol{\Sigma} = \mathcal{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}, \tag{35}$$

where the expected value of a vector or a matrix is found by taking the expected values of its components.

- In other words, if x_i is the ith component of $\mathbf{x}$, m_i the ith component of $\boldsymbol{\mu}$, and σ_{ij} the ijth component of $\boldsymbol{\Sigma}$, then

$$m_i = \mathcal{E}[x_i] \tag{36}$$

and

$$\sigma_{ij} = \mathcal{E}[(x_i - m_i)(x_j - m_j)]. \tag{37}$$

- The diagonal elements σ_{ii} are the variances of the respective x_i and off-diagonal elements are the **covariances** of x_i and x_j.
- If x_i and x_j are **statistically independent**, then $\sigma_{ij} = 0$.

- Linear combination of jointly normally distributed random variables are normally distributed.

- In particular, if $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{A}$ is a d-by-k matrix and $\mathbf{y} = \mathbf{A}^t\mathbf{x}$, then $p(\mathbf{y}) \sim N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$, as illustrated in Figure 7.

- In the special case where $k = 1$ and $\mathbf{A}$ is a unit-length vector $\mathbf{a}$, $y = \mathbf{a}^t\mathbf{x}$ is a scalar that represents the projection of $\mathbf{x}$ onto a line in the direction of $\mathbf{a}$; in that case $\mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a}$ is the variance of the projection of $\mathbf{x}$.

- It is sometimes convenient to perform a coordinate transformation that converts an arbitrary multivariate normal distribution into a spherical one.

- If we define $\mathbf{\Phi}$ to be the matrix whose columns are the eigenvectors of $\mathbf{\Sigma}$, and $\mathbf{\Lambda}$ the diagonal matrix of the corresponding eigenvalues, then the **whitening transformation**

$$\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2} \tag{38}$$

  ensures that the transformed distribution has covariance matrix equal to the identity matrix.

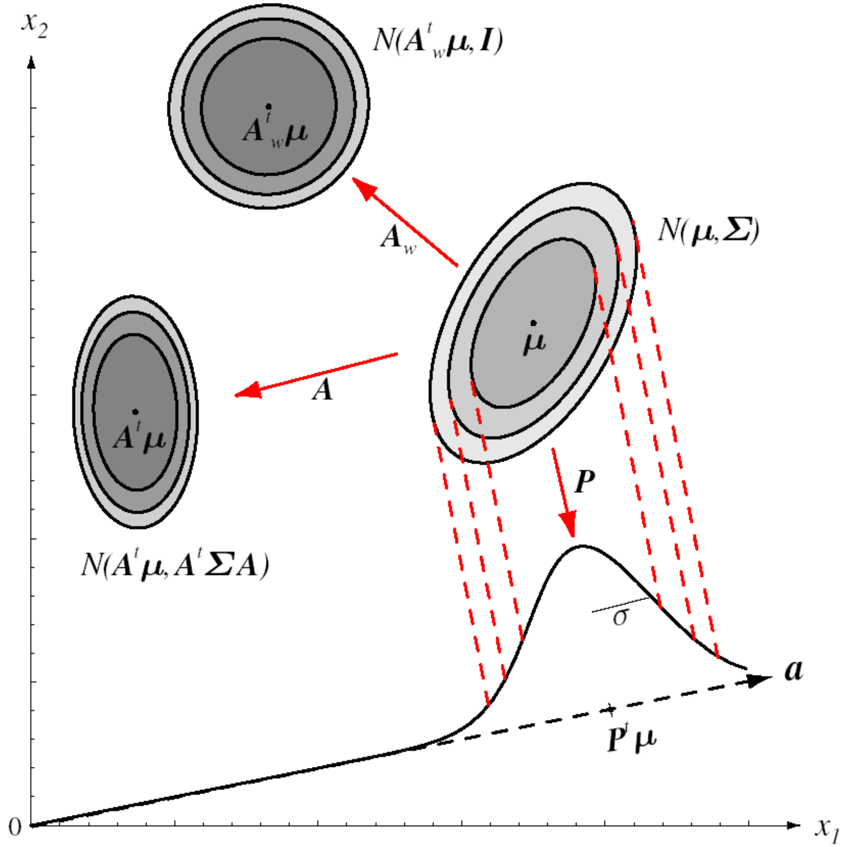

x_2
$N(\mathbf{A}_w^t\boldsymbol{\mu}, \mathbf{I})$
$\mathbf{A}_w^t\boldsymbol{\mu}$
$\mathbf{A}_w$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\boldsymbol{\mu}$
$\mathbf{A}$
$\mathbf{A}^t\boldsymbol{\mu}$
$\mathbf{P}$
$N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$
σ
$\mathbf{a}$
$\mathbf{P}^t\boldsymbol{\mu}$
0
x_1

Figure 7: A whitening transformation convert a normal distribution into a circularly symmetric Gaussian.

- Samples drawn from a normal population tend to fall in a single cloud (Figure 8).
- The center of the cloud is determined by the mean vector.
- The shape of the cloud is determined by the covariance matrix.
- The loci of points of constant density are hyperellipsoids for which $(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is constant.

- The principal axes of these hyperellipsoids are given by the eigenvectors of $\boldsymbol{\Sigma}$; the eigenvalues determine the lengths of these axes.

- The quantity

$$r^2 = (\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{39}$$

  is called the squared **Mahalanobis distance** from $\mathbf{x}$ to $\boldsymbol{\mu}$.

- In Figure 8, an ellipse consists of points with an equal Mahalanobis distance from the mean $\boldsymbol{\mu}$.

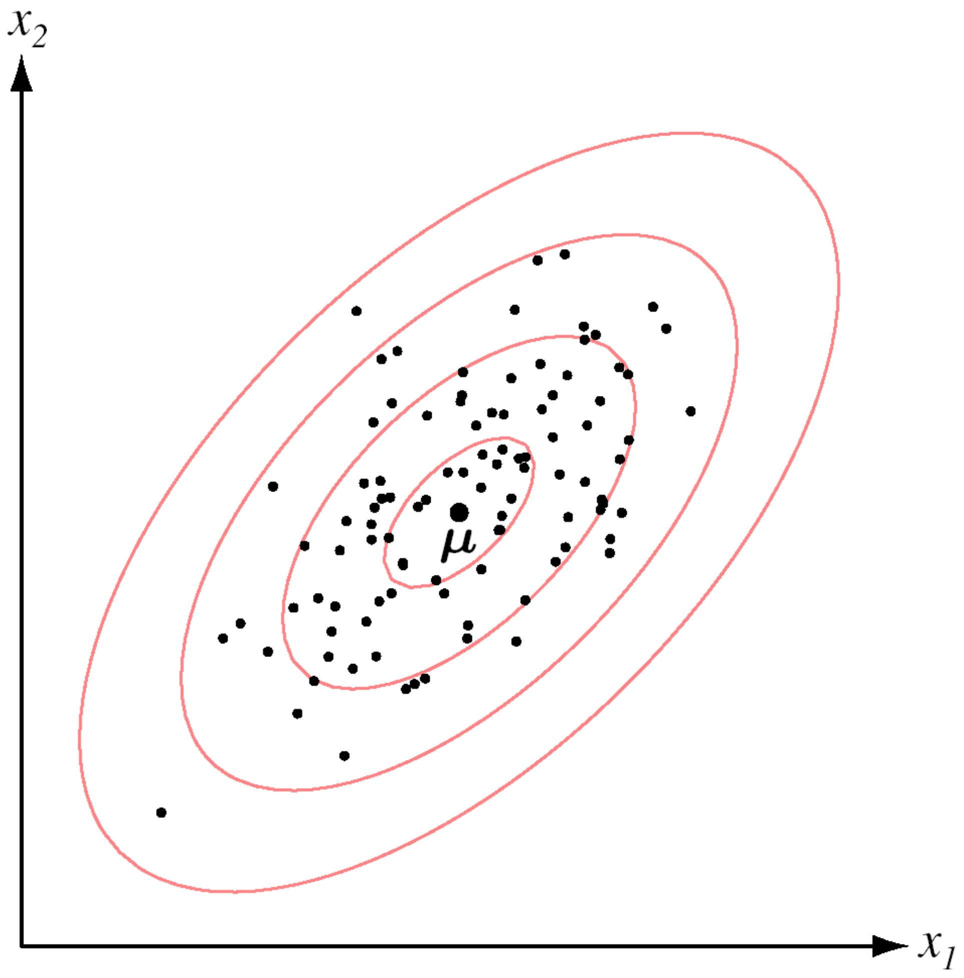


Figure 8: Samples drawn from a two-dimensional Gaussian density.

2.1 Introduction

2.2 Bayesian Decision Theory

2.3 Minimum-Error-Rate Classification

2.4 Classifiers, Discriminant Functions, and Decision Surfaces

2.5 The Normal Density

2.6 **Discriminant Functions for the Normal Density**

2.8.3 Signal Detection Theory and Operating Characteristics

2.6 Discriminant Functions for the Normal Density

- In section 2.4.1, the minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \tag{40}$$

- This expression can be evaluated if the densities are multivariate normal. Thus, we have

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \tag{41}$$

Let us examine this discriminant function for a number of special cases.

2.6.1 Case 1: $\boldsymbol{\Sigma}_i = \sigma^2\mathbf{I}$

- In this case, the covariance matrix is diagonal, being σ^2 times the identity matrix.
- Geometrically, this corresponds to equal-size hyperspherical clusters.

$$|\boldsymbol{\Sigma}_i| = \sigma^{2d} \quad \text{and} \quad \boldsymbol{\Sigma}_i^{-1} = \frac{1}{\sigma^2}\mathbf{I}. \tag{42}$$

- Both $|\boldsymbol{\Sigma}_i|$ and $(d/2)\ln 2\pi$ in Eq. (41) are constants that can be ignored.

- Thus, we obtain the simplified discriminant functions

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 + \ln P(\omega_i) \tag{43}$$

where $\|\cdot\|$ denotes the Euclidean norm, that is

$$\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i) \tag{44}$$

- Furthermore, it is actually not necessary to compute norms.
- Note that the expansion of $(\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i)$ yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \qquad (45)$$

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

- Thus, we obtain the equivalent **linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0} \tag{46}$$

where

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i \tag{47}$$

and

$$w_{i0} = \frac{-1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i) \tag{48}$$

- We call w_{i0} the **threshold** or **bias** for ith category.

- The decision boundaries are pieces of hyperplanes defined by the equations $g_i(\mathbf{x}) = g_j(\mathbf{x})$ for two categories with the highest posterior probabilities.

- For this particular case, the equation can be written as

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \tag{49}$$

where

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \tag{50}$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{51}$$

- These equations define a hyperplane through the point $\mathbf{x}_0$ and orthogonal to the vector $\mathbf{w}$.

- If $P(\omega_i) = P(\omega_j)$, the second term on the right of Eq. (51) vanishes. Thus, the point $\mathbf{x}_0$ is halfway between the means (Figure 9, 11, and 13).

- If $P(\omega_i) \neq P(\omega_j)$, the point $\mathbf{x}_0$ shifts away from the more likely mean (Figure 10, 12, and 14).

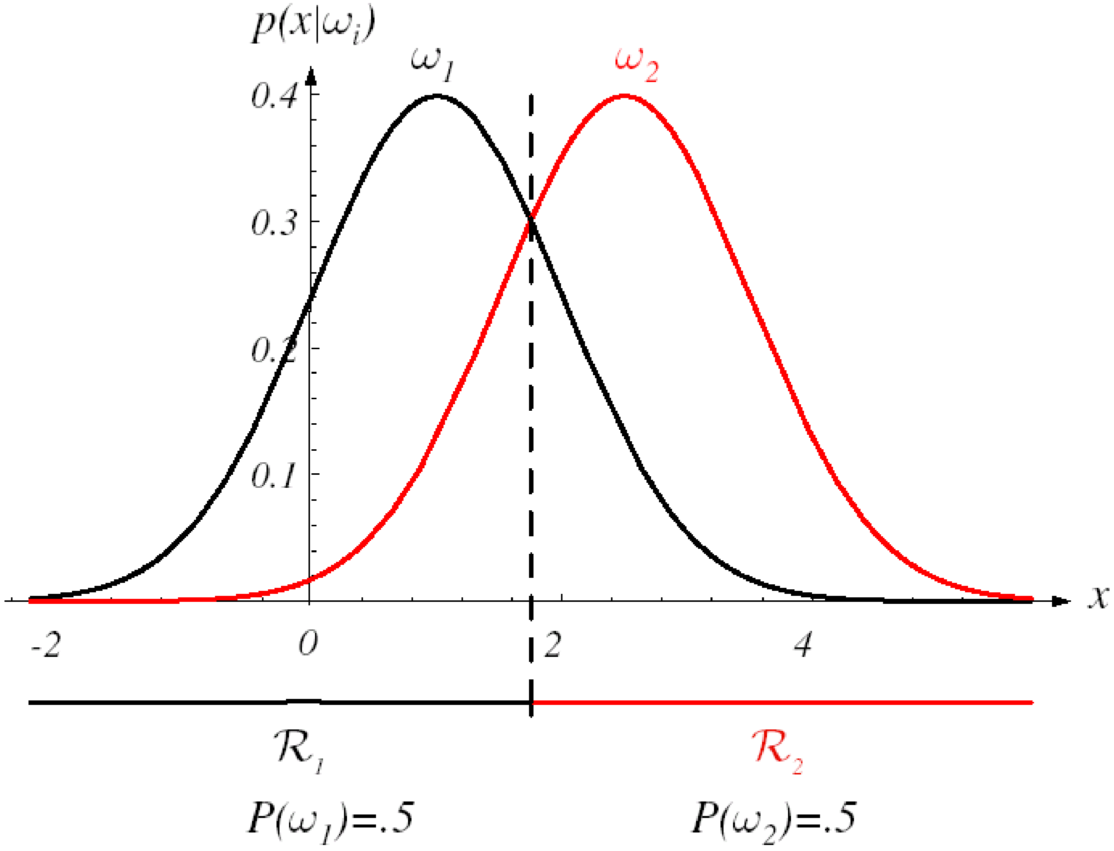


Figure 9: An one-dimensional example

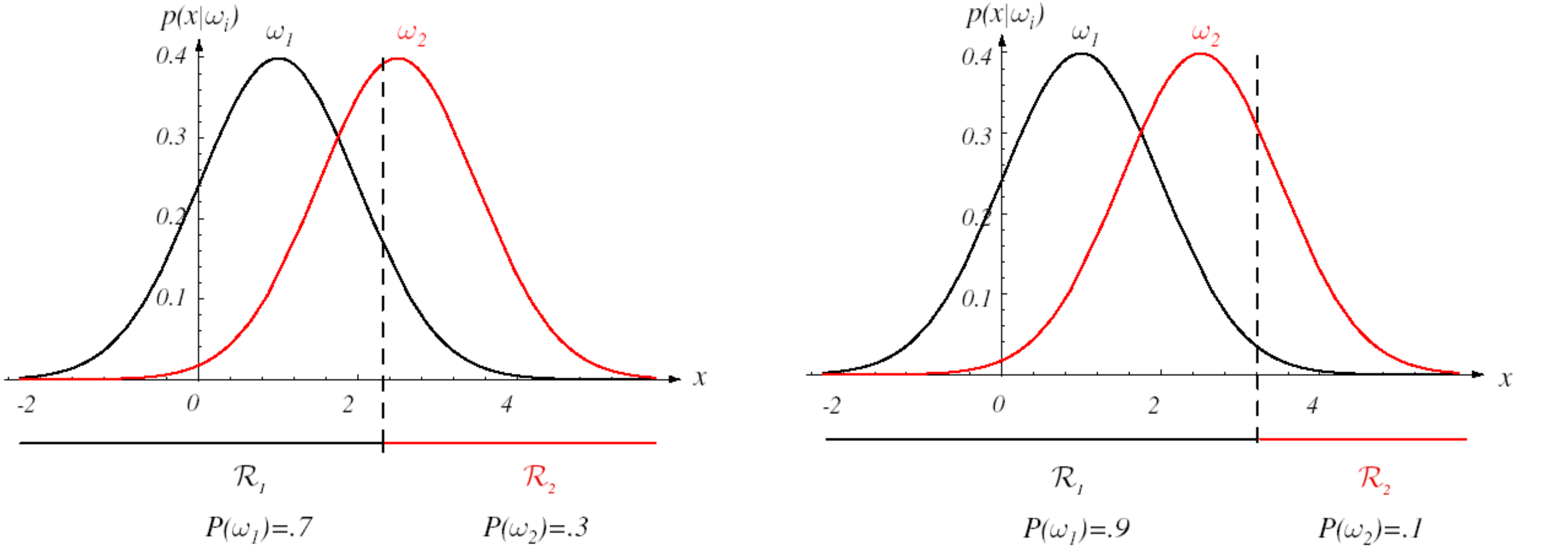


Figure 10: As priors are changed, the decision boundary shifts.

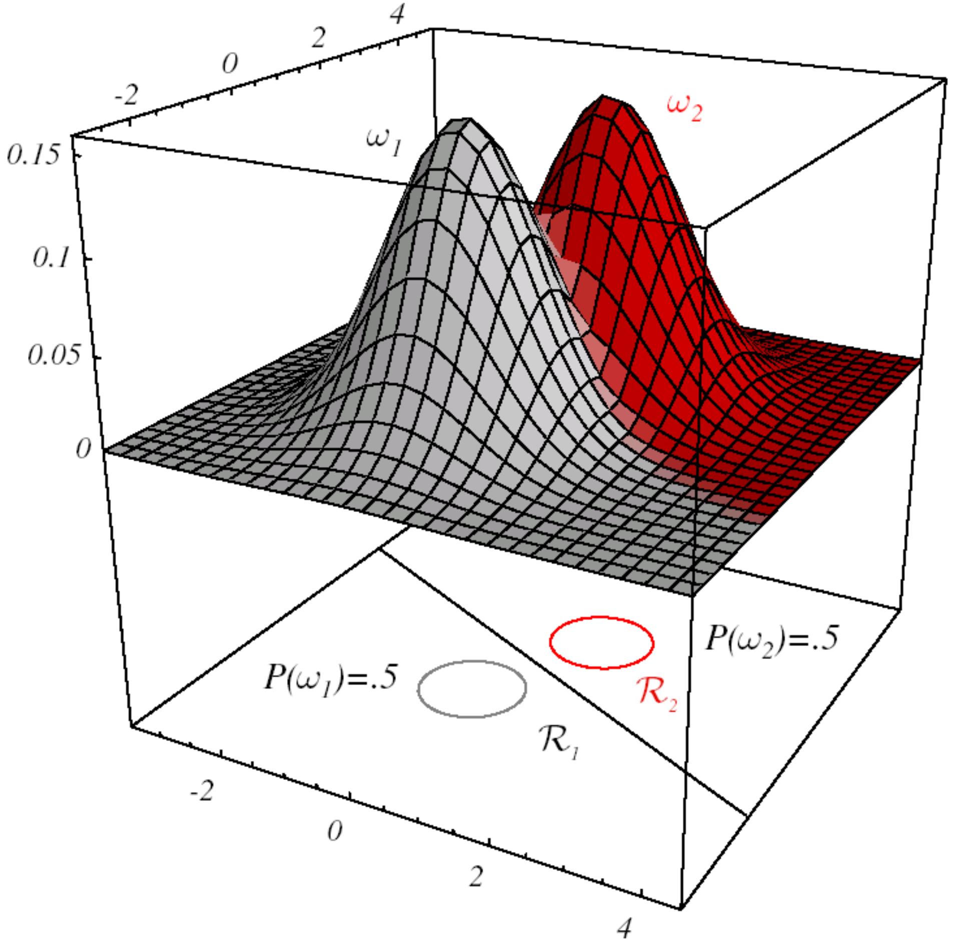


Figure 11: A two-dimensional example

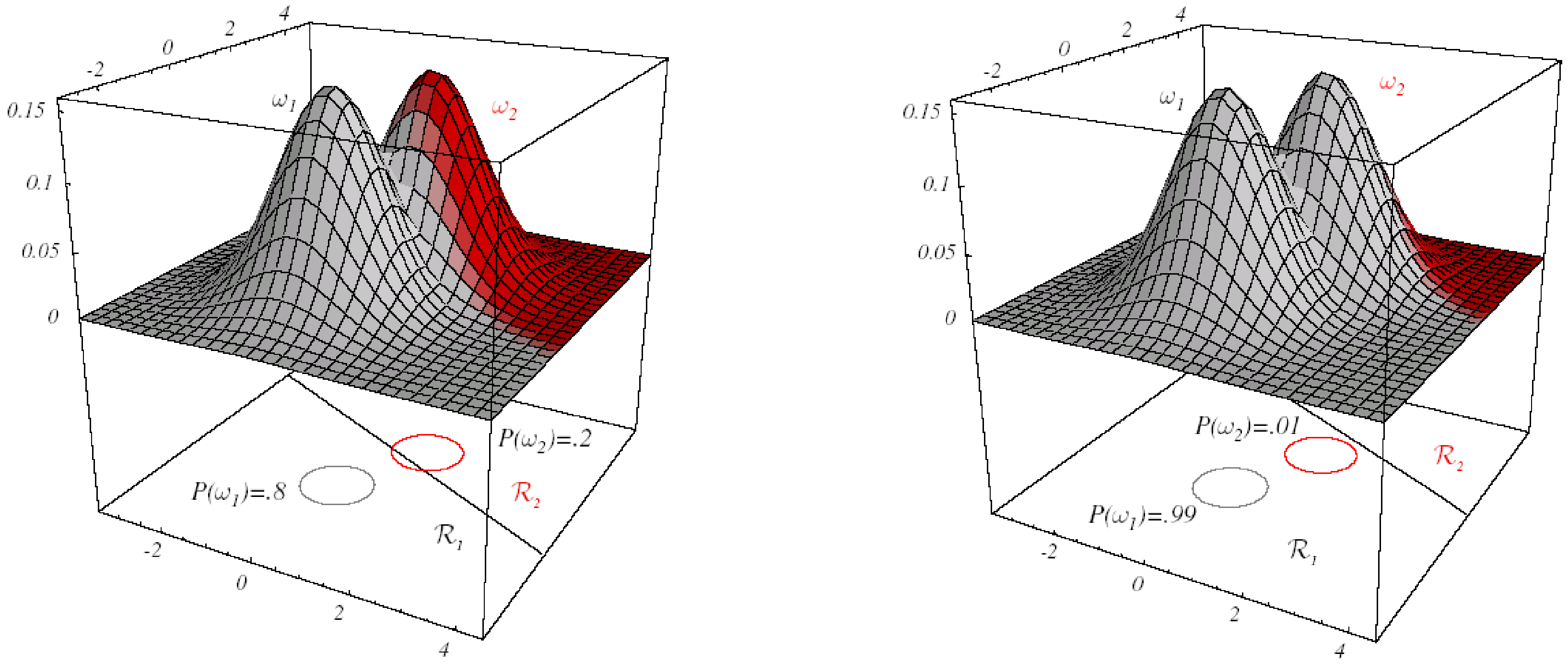


Figure 12: As priors are changed, the decision boundary shifts.

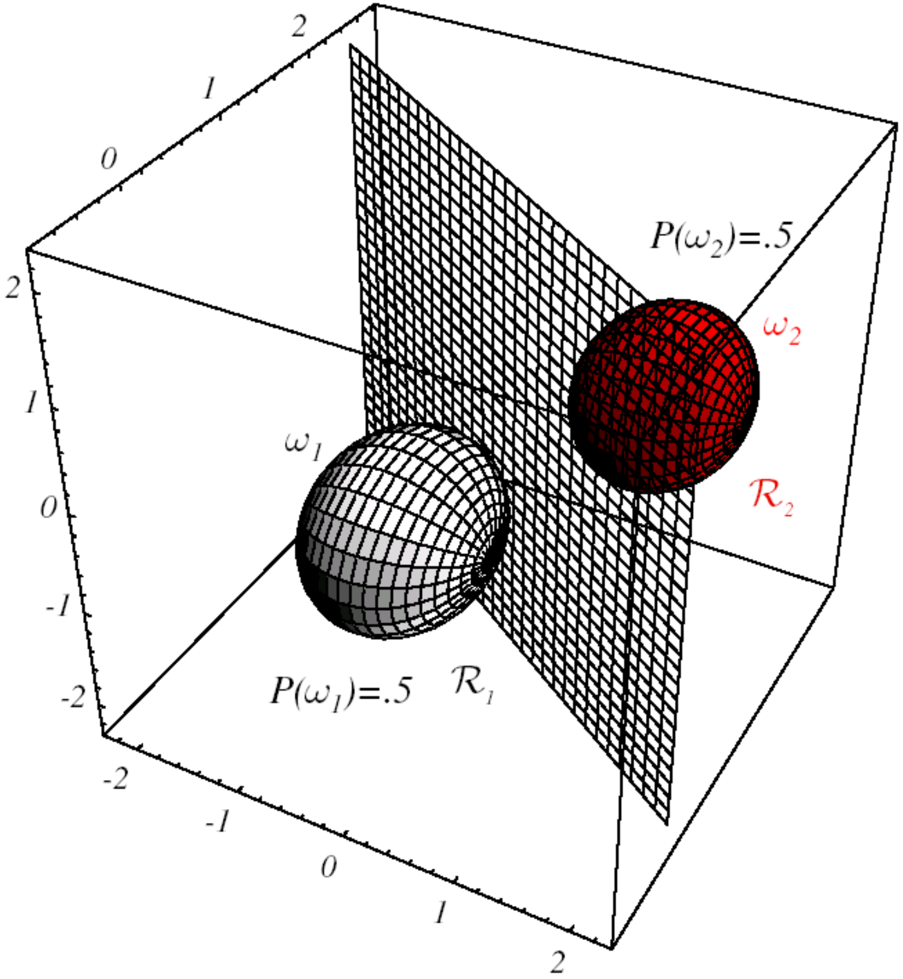


Figure 13: A three-dimensional example

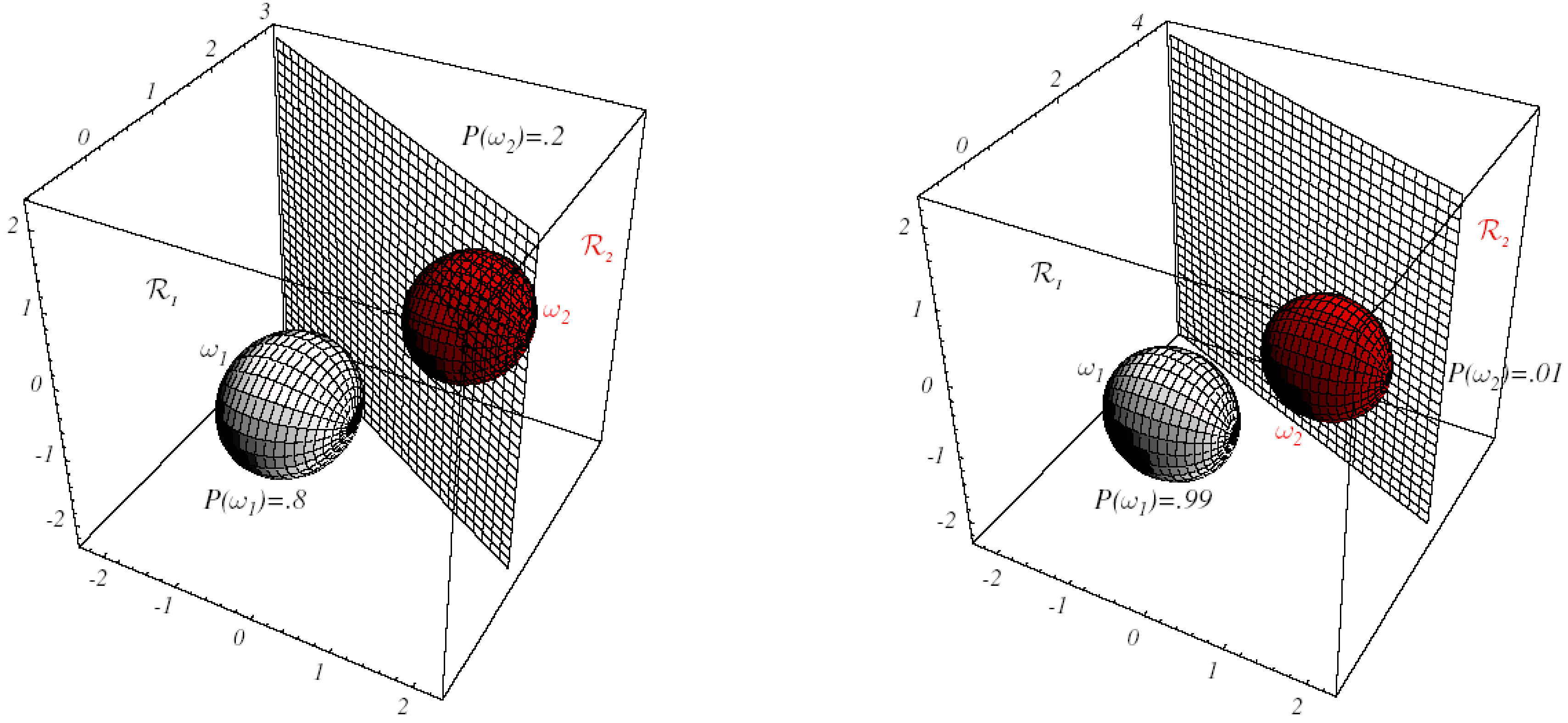


Figure 14: As priors are changed, the decision boundary shifts.

- If the priors are the same for all categories, the $\ln P(\omega_i)$ becomes unimportant terms.
- When this happens, the optimal decision rule can be stated very easily:

  > Compute the squared Euclidean distances $(\mathbf{x} - \boldsymbol{\mu}_i)^t(\mathbf{x} - \boldsymbol{\mu}_i)$ and assign a feature vector $\mathbf{x}$ to the category of the nearest mean.

- Such a classifier is called a **minimum-distance classifier**.

2.6.2 Case 2: $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

- Another simple case arises when the covariance matrices for all categories are identical.
- Geometrically, this corresponds to the situation in which all samples fall in hyperellipsoidal clusters of equal size and shape.
- The terms, $|\boldsymbol{\Sigma}_i|$ and $(d/2)\ln 2\pi$, in Eq. 41 can be ignored because they are independent of i.

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i) \qquad (52)$$

- If the priors are the same for all categories, the term $\ln P(\omega_i)$ can also be ignored.

- When this happens, the optimal decision rule can once again be stated very easily:

  > Compute the squared Mahalanobis distances $(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)$ and assign a feature vector $\mathbf{x}$ to the category of the nearest mean.

- Expansion of the form $(\mathbf{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)$ results in a term $\mathbf{x}^t\boldsymbol{\Sigma}_i^{-1}\mathbf{x}$ which is independent of i.
- After this term is dropped, the resulting discriminant functions are again linear:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0} \tag{53}$$

where

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \tag{54}$$

and

$$w_{i0} = \frac{-1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i) \tag{55}$$

- Because the discriminants are linear, the resulting decision boundaries are again hyperplanes (Figure 15 and 16).

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

$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_0) = 0 \tag{56}$$

where

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{57}$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^t\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \tag{58}$$

- Because $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ is generally not in the direction of $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, the decision hyperplane is generally not orthogonal to the line connecting the means.

- If the priors are equal, $\mathbf{x}_0$ is halfway between the means.

- If $P(\omega_i) \neq P(\omega_j)$, the decision hyperplane is shifted away from the more likely mean (Figure 15 and 16).

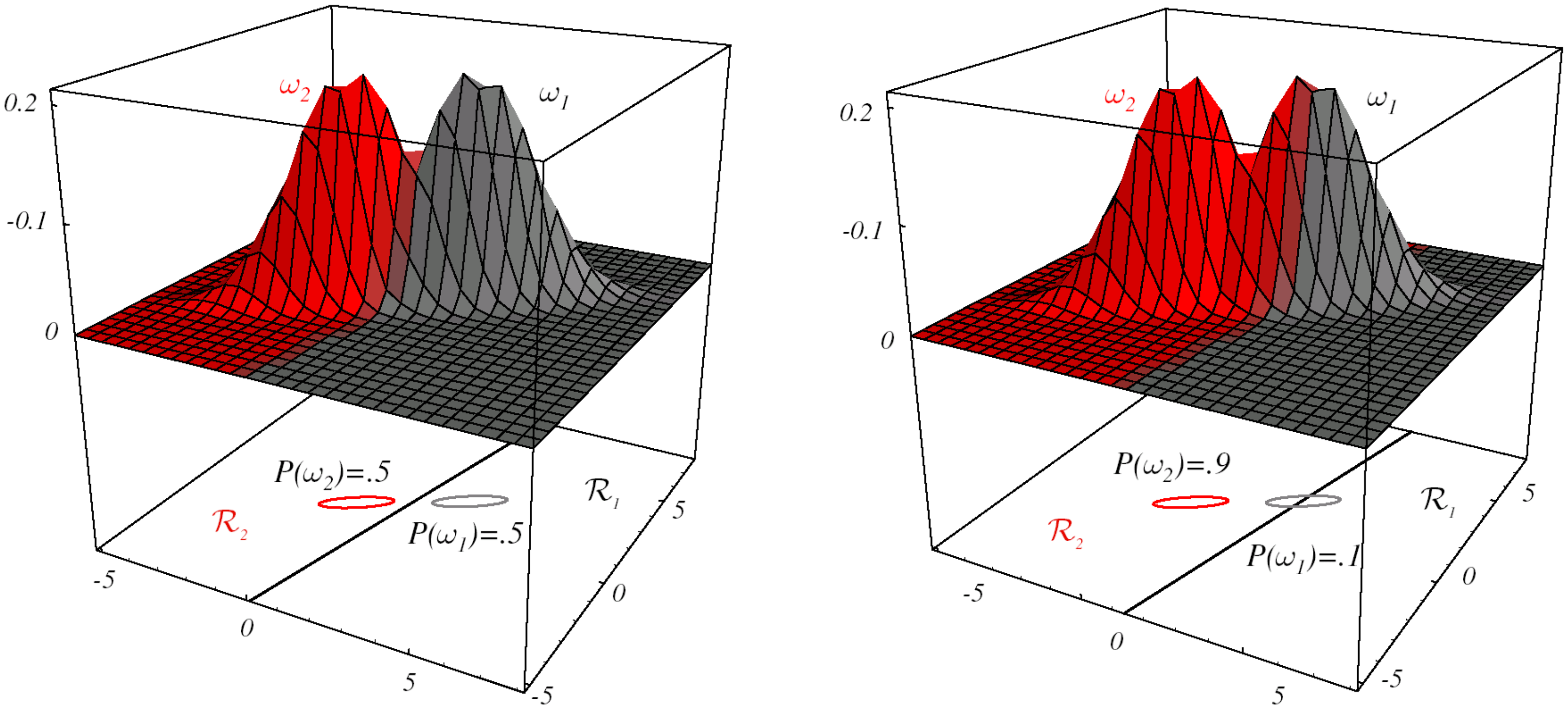


Figure 15: For equal but asymmetric Gaussian distributions, the decision hyperplanes are not perpendicular to the line connecting the means in general.

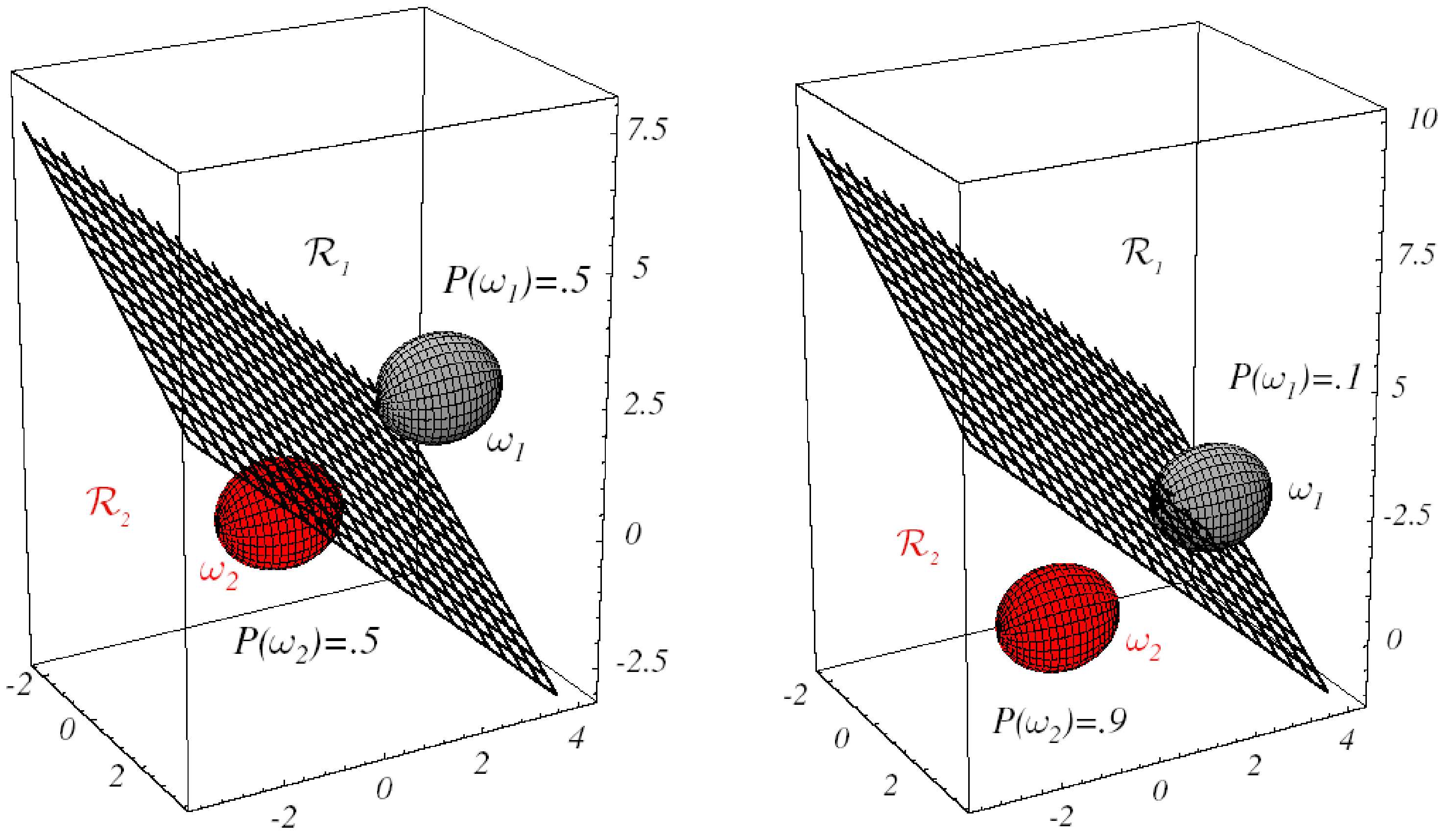


Figure 16: For equal but asymmetric Gaussian distributions, the decision hyperplanes are not perpendicular to the line connecting the means in general.

2.6.3 Case 3: $\boldsymbol{\Sigma}_i$ = arbitrary

- The only term that can be ignored from Eq. 41 is the $(d/2)\ln 2\pi$, and the resulting discriminants are inherently quadratic:

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0} \tag{59}$$

where

$$\mathbf{W}_i = \frac{-1}{2}\boldsymbol{\Sigma}_i^{-1} \tag{60}$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \tag{61}$$

and

$$w_{i0} = \frac{-1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \tag{62}$$

- In the two-category case, the decision boundaries are **hyperquadrics**.

- They can assume any of the general forms: hyperplanes, pair of hyperplanes, hyperellipsoids, and hyperhyperboloids of various types.

- Even in one dimension, for arbitrary variance the decision regions need not to be simply connected.

- Two-dimensional examples are shown in Figs 18, 19, and 20.

- Three-dimensional examples are shown in Figs 21, 22, and 23.

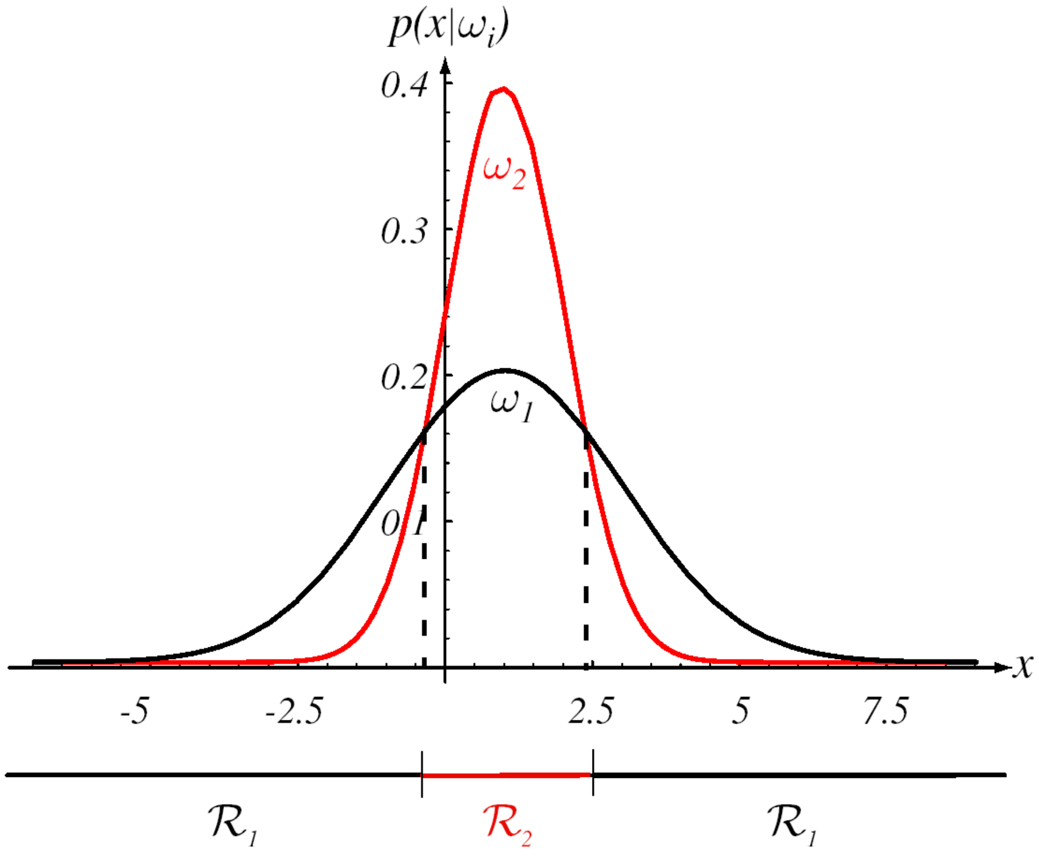


Figure 17: Non-simply connected decision regions can arise for Gaussian having unequal variance.

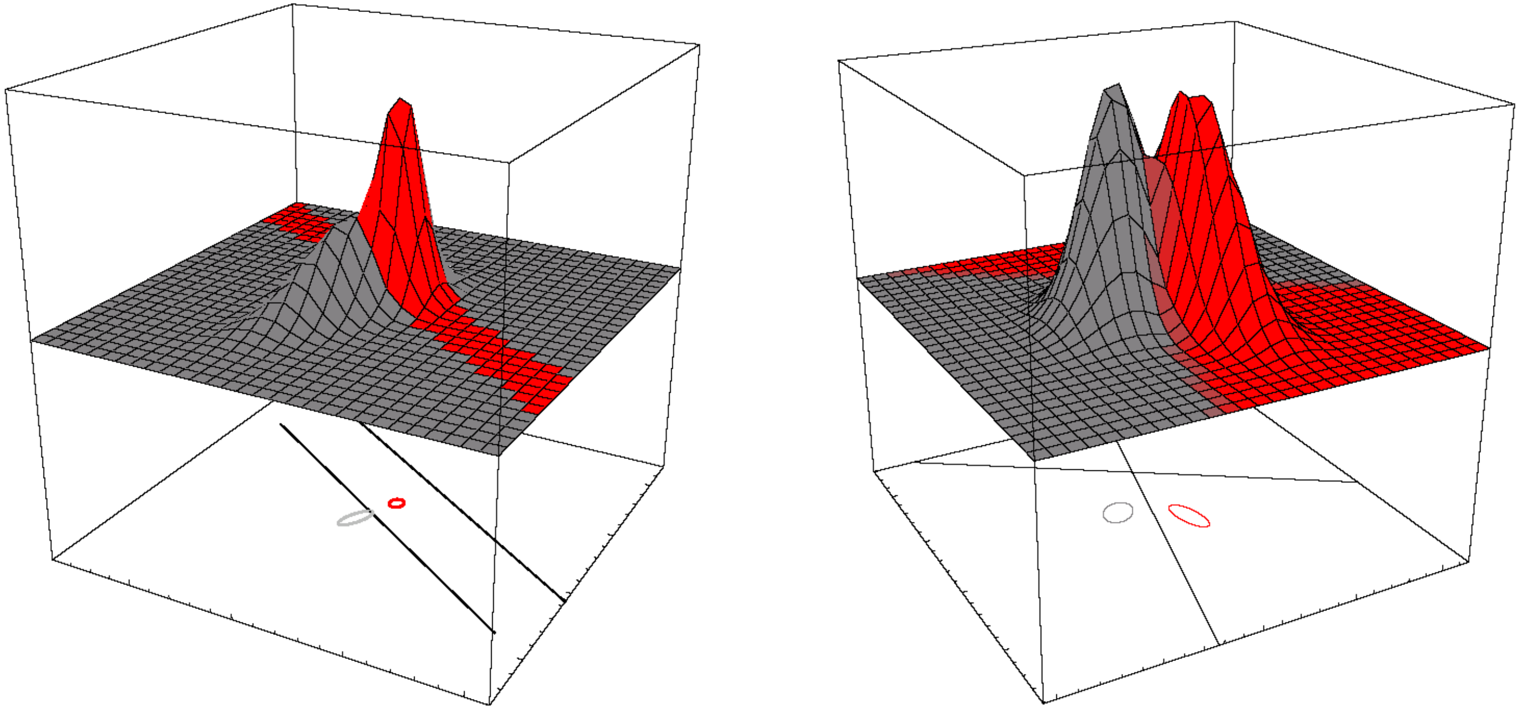


Figure 18: Arbitrary Gaussian distributions lead decision boundaries that are generally hyperquadratics.

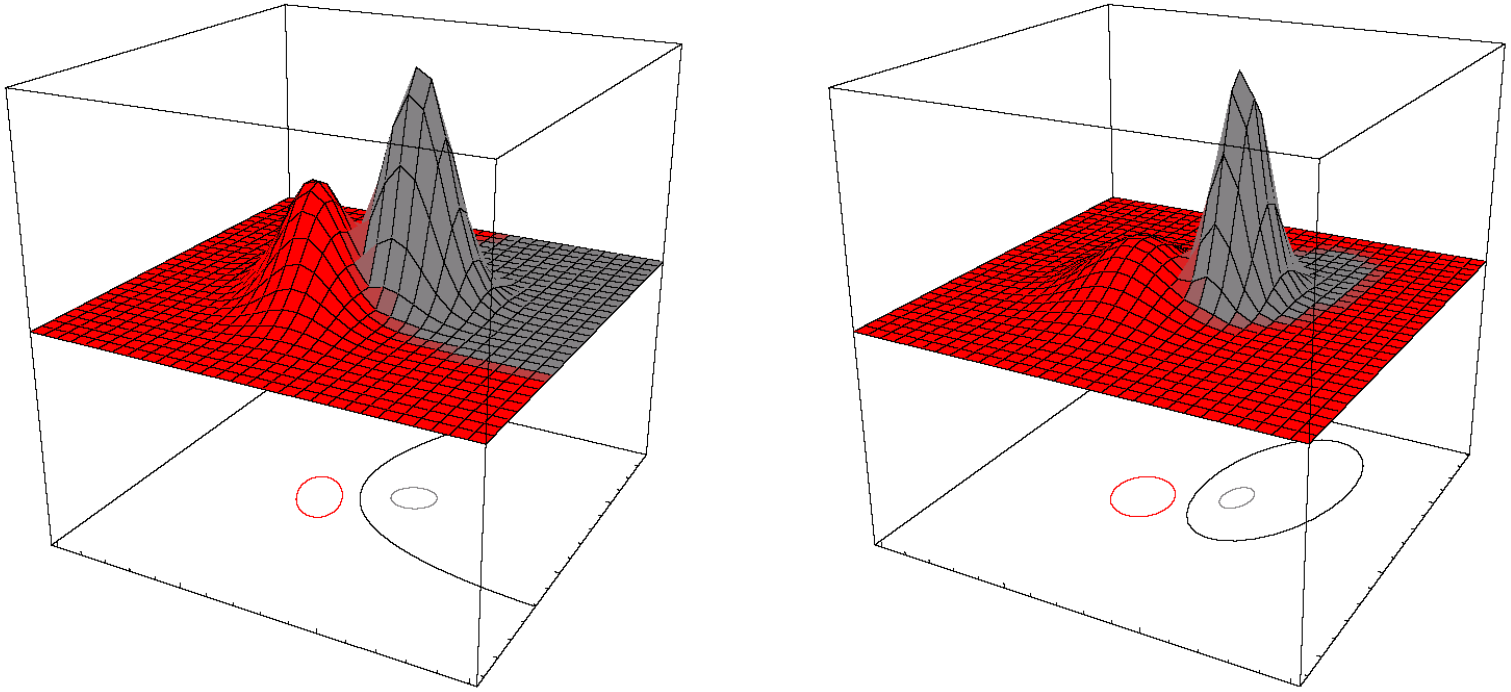


Figure 19: Arbitrary Gaussian distributions lead decision boundaries that are generally hyperquadratics.

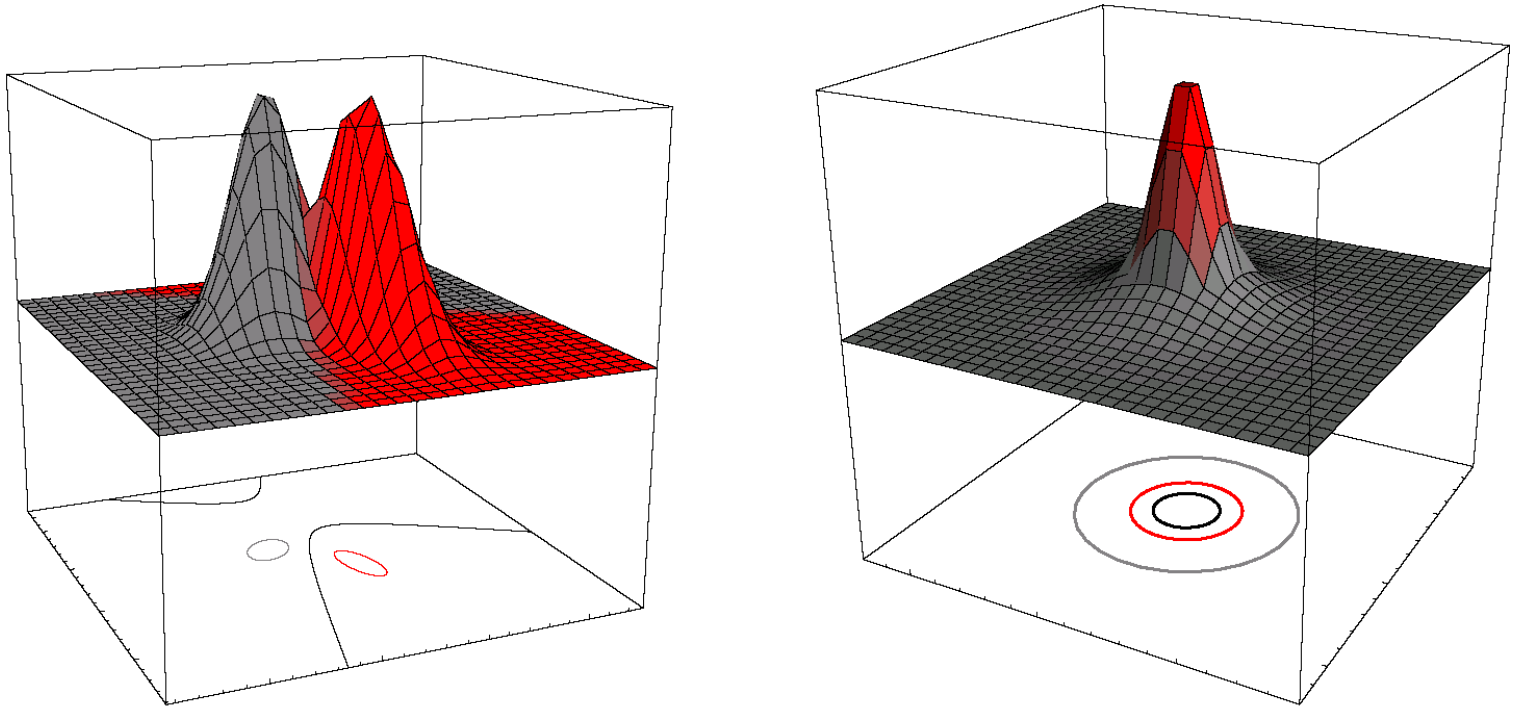


Figure 20: Arbitrary Gaussian distributions lead decision boundaries that are generally hyperquadratics.

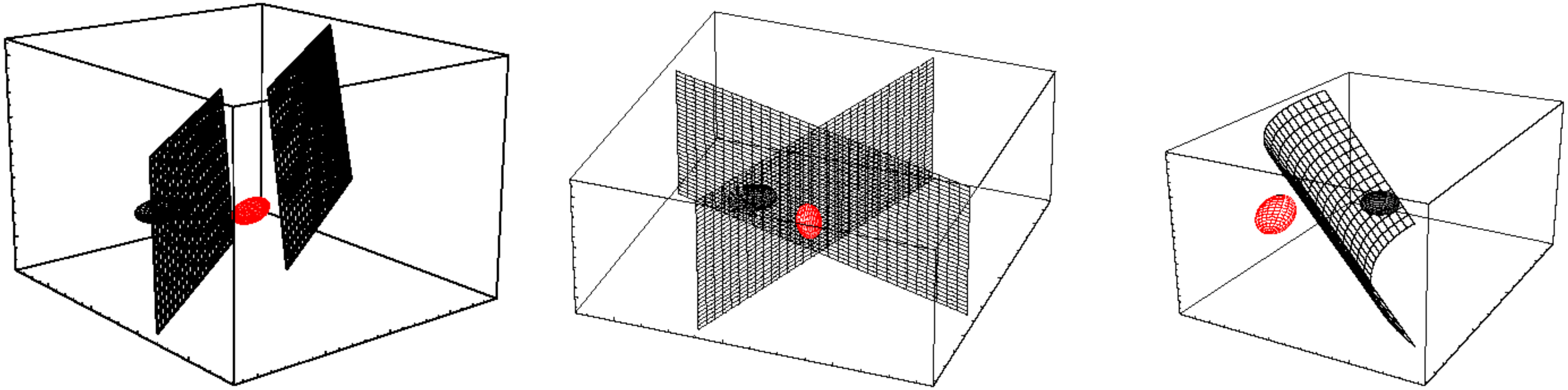


Figure 21: Arbitrary three-dimensional Gaussian distributions yield decision boundaries that are two-dimensional hyperquadratics.

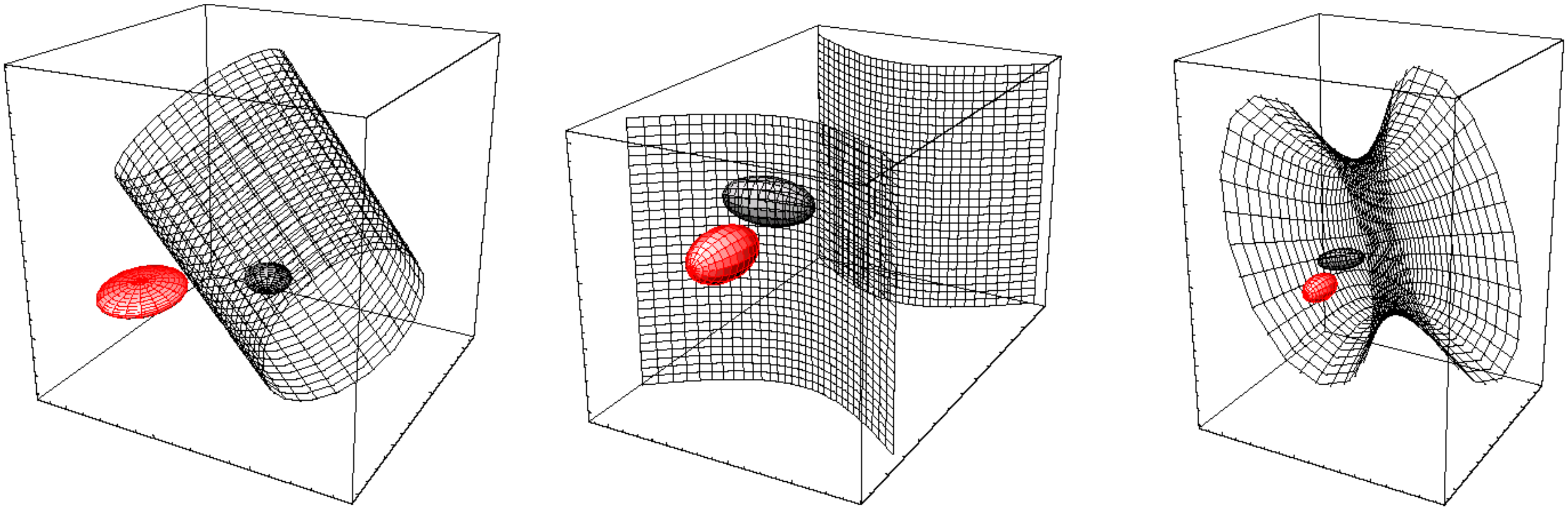


Figure 22: Arbitrary three-dimensional Gaussian distributions yield decision boundaries that are two-dimensional hyperquadratics.

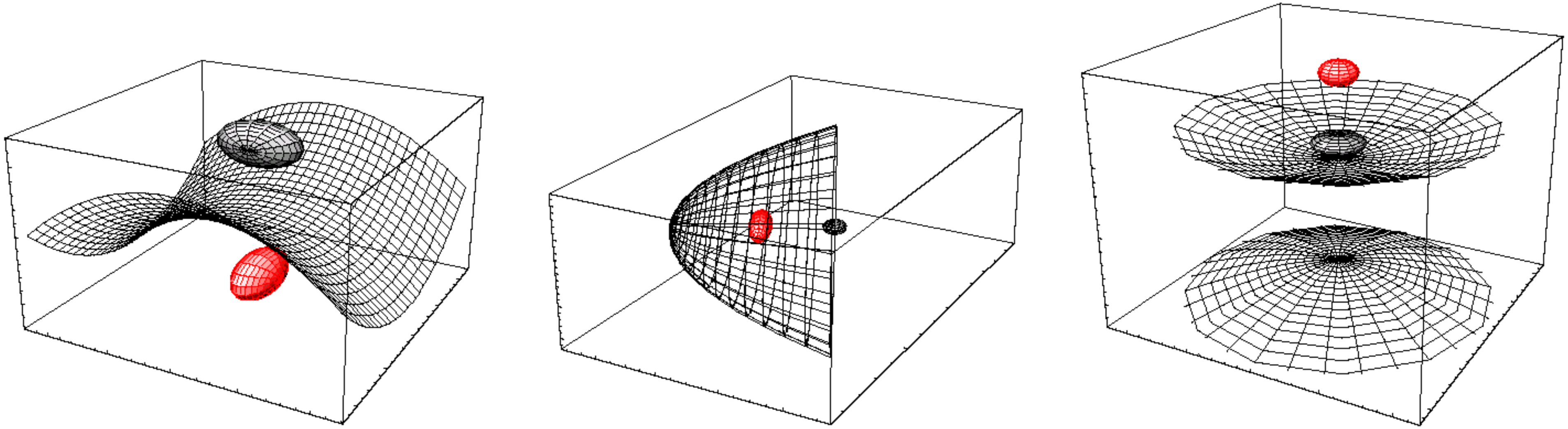


Figure 23: Arbitrary three-dimensional Gaussian distributions yield decision boundaries that are two-dimensional hyperquadratics.

- The extension to more than two categories is straightforward.
- Figure 24 shows the decision boundaries for a four-category made up of Gaussian distributions.
- Of course, if the distributions are more complicated, the decision regions can be more complex.

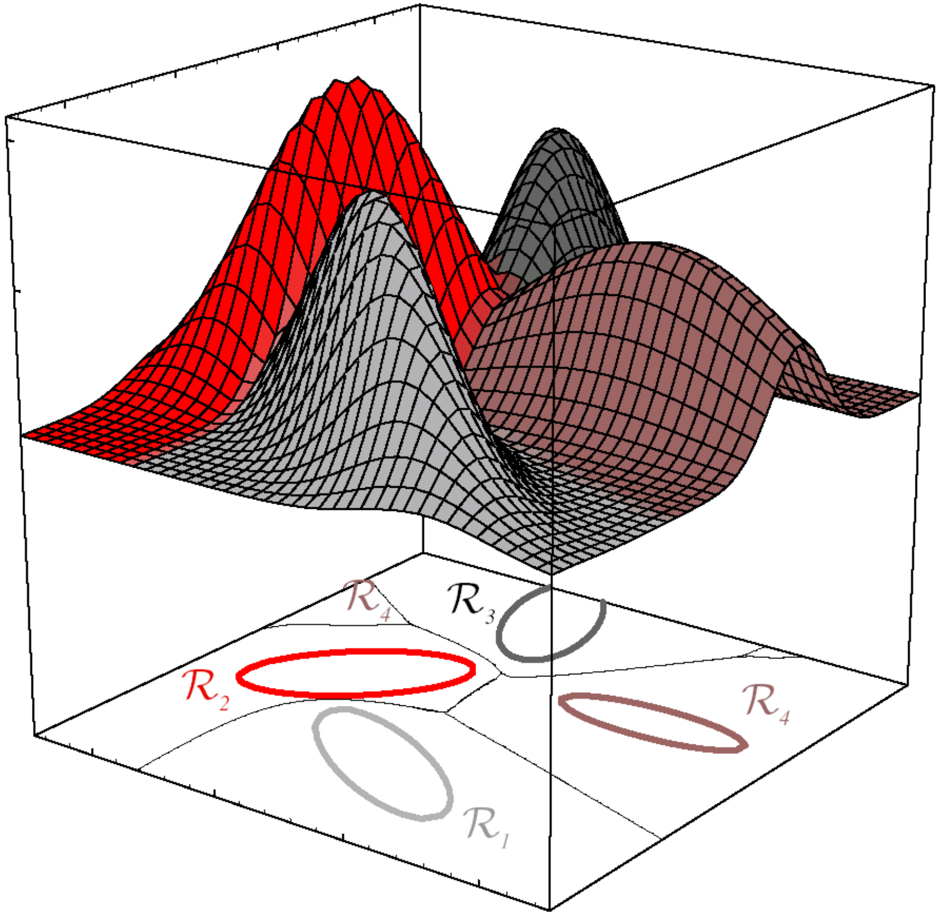


Figure 24: The decision regions for four normal distributions.

Example 5

Decision Regions for Two-Dimensional Gaussian Data

To illustrate these ideas, we calculate the decision boundary for the data in the Figure 25.

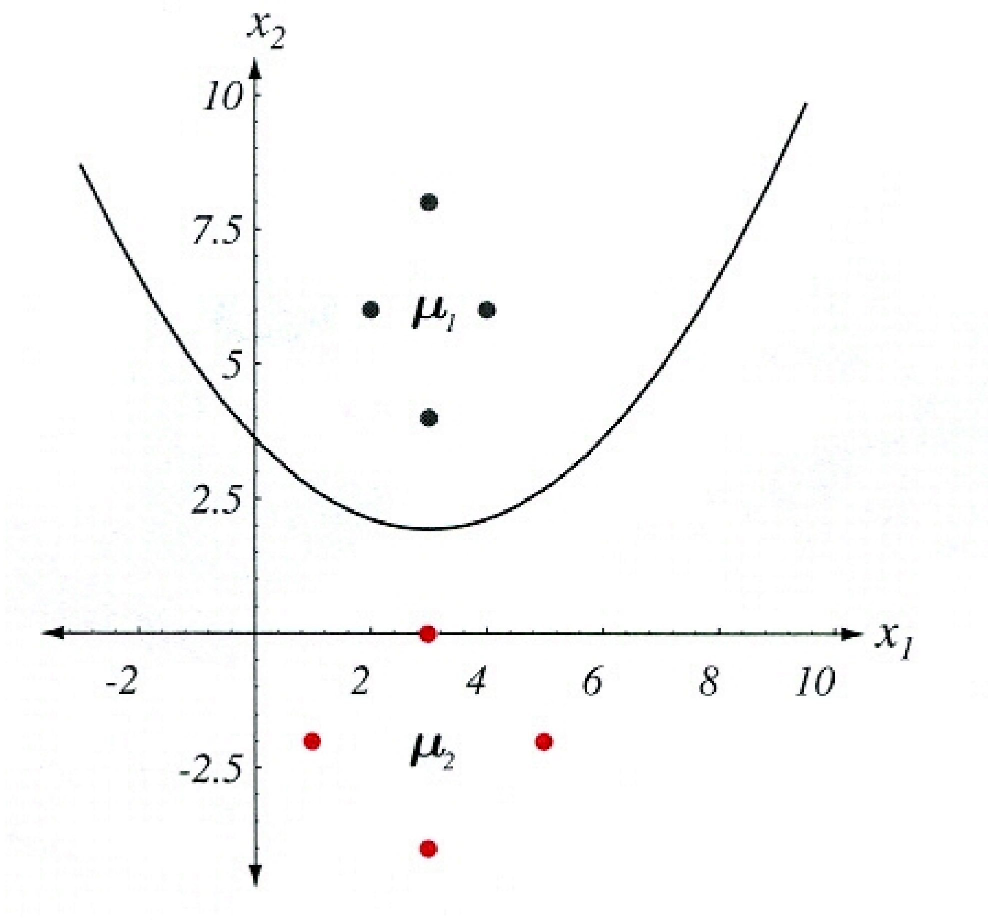


Figure 25: Two-category data and the computed decision boundary

- Let ω_1 be the set of black points, and ω_2 the red points.
- Although we will spend much of the next chapter on the parameter estimation, for now we simply calculate the means and covariances by

$$
\begin{aligned}
\boldsymbol{\mu}_1 &= \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix} \\
\boldsymbol{\mu}_2 &= \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}
\end{aligned}
$$

- Thus, the inverse matrices are

$$\mathbf{\Sigma}_1^{-1} = \begin{bmatrix} 2 & 0 \\ 0 & 1/2 \end{bmatrix} \quad \mathbf{\Sigma}_2^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

- We assume equal priors and substitute these into the Eqs (59) -(62).

- By setting $g_1(\mathbf{x}) = g_2(\mathbf{x})$, we obtain the following decision boundary.

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2 \tag{63}$$

Some observations

- The decision boundary is a parabola with vertex at $\begin{bmatrix} 3 & 1.83 \end{bmatrix}^t$.
- The decision boundary does not pass through the point $\begin{bmatrix} 3 & 2 \end{bmatrix}^t$, the midway between two means, as we might have naively guessed.
- This is because for ω_1, the distribution is squeezed more in the x_1 direction. Hence, the ω_1 distribution is increased along the x_2 direction.
- The resulting decision boundary lies slightly lower than the midway point.

2.1 Introduction

2.2 Bayesian Decision Theory

2.3 Minimum-Error-Rate Classification

2.4 Classifiers, Discriminant Functions, and Decision Surfaces

2.5 The Normal Density

2.6 Discriminant Functions for the Normal Density

2.8.3 **Signal Detection Theory and Operating Characteristics**

2.8.3 Signal Detection Theory and Operating Characteristics

- Suppose we are interested in detecting a weak signal, such as a radar reflection.
- Because of random noise, the observation x is a random variable.
- Our model is that the observation x has mean μ_2 when the signal of interest is present, and mean μ_1 when it is not present (Figure 26).

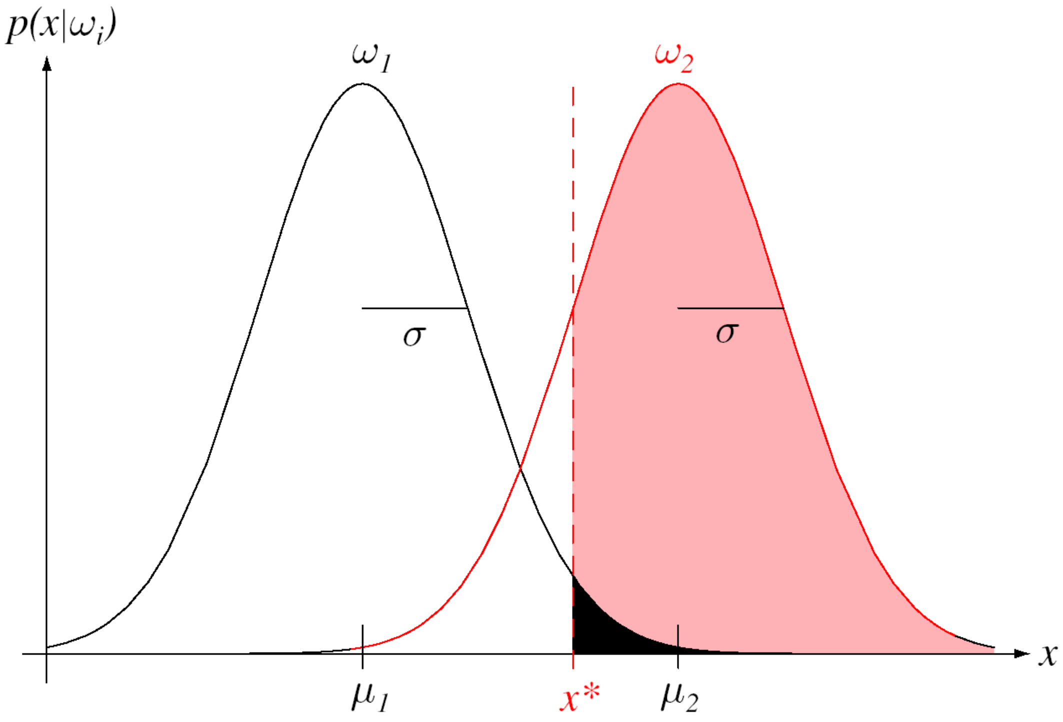


Figure 26: When the signal of interest is not present, the density is $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. The decision threshold x^* will determine the probabilities of hit and of false alarm.

- A detector (classifier) employs a threshold values x^* for determining whether the signal of interest is present.
- Sometimes, one might seek to find some measure of the ease of detection (making decision), in a form independent of the choice of x^*.
- Such a measure is called **discriminability** , which completely depends on the underlying distributions, but not on a decision strategy.
- A popular choice for discriminability is

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma} \tag{64}$$

A high d' is of course desirable.

Let us consider the following four probabilities:

- $P(x > x^*|x \in \omega_2)$: a **hit** – the probability that an observation x is above x^* given that the signal of interest is present.
- $P(x > x^*|x \in \omega_1)$: a **false alarm** – the probability that an observation x is above x^* given that the signal of interest is not present.
- $P(x < x^*|x \in \omega_2)$: a **miss** – the probability that an observation x is below x^* given that the signal of interest is present.
- $P(x < x^*|x \in \omega_1)$: a **correct rejection** – the probability that an observation x is below x^* given that the signal of interest is not present.

- If the densities are fixed but the threshold x^* is changed, then our hit and false alarm rates will also change.

- The pair of hit and false alarm rates will move along a curve – a **receiver operating characteristic** or **ROC** curve (Figure 27).

- It is a simple matter to generalize the above discussion to multidimensional distributions.

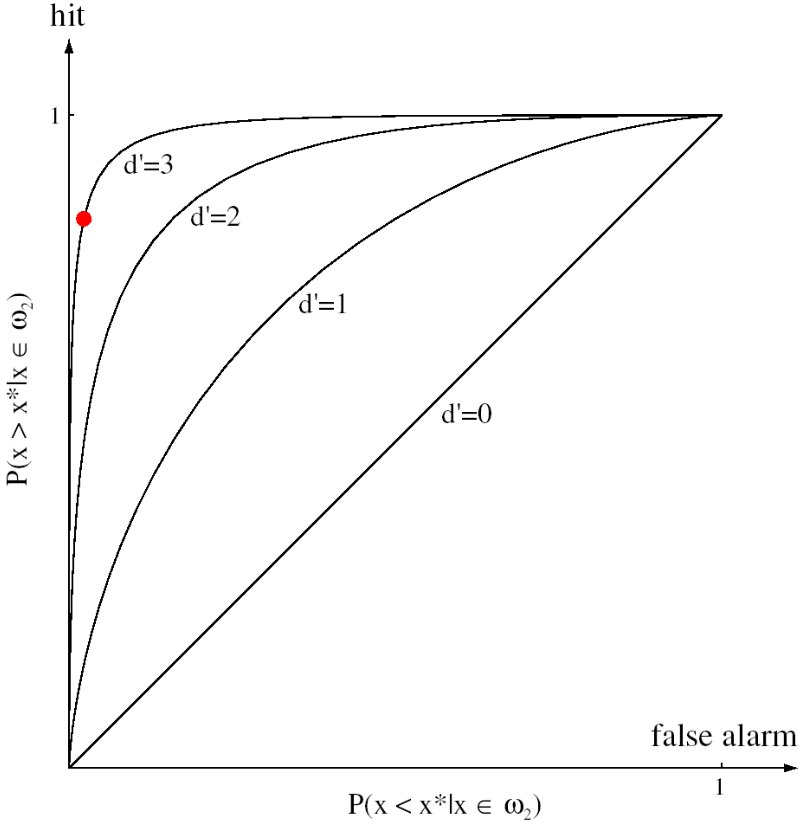


Figure 27: In an ROC curve, the abscissa is the probability of false alarm and the ordinate is the probability of hit.

Summary

- For practical applications, the main problem in applying Bayesian decision theory is that the conditional densities are not known.
- In some cases, we may know the form these densities, but we may not know the parameter values.
- For instance, the densities are known to be, or assumed to be, multivariate normal, but the values of means and covariance matrices are not known.
- Most of the following chapter will be devoted to various approaches for dealing with such problems.

References

[1] J. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*. Cambridge University Press, 2006.