

# Bayesian Decision Theory



Dr. Pritam Anand.  
Assistant Professor,  
DA-IICT, Gandhinagar.

# An Introduction

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

# Classification Problem



Length ( cm )	Height (cm)	Number of fins	Weight (Kg)	Color	Fish type
17.8	22.9	8	5.1`	Orange	Salman
14.8	20.5	7	4.9	Black	Sea bass
16.34	12.76	6	6.6	Grey	Salman
10.34	8.76	3	3.8	Grey	Salman
---	-----	-----	-----	-----	-----
11.30	17.76	6	9.8	Orange	Sea Bass

12 cm  
=

$$P(X=12\text{cm} / \omega_1) P(\omega_1) - P(\omega_1 / X)$$

$$P(X=12\text{cm} / \omega_2) P(\omega_2) - P(\omega_2 / X)$$

# Prior Probability

- More generally, we assume that there is some a priori probability (or simply prior)  $P(\omega_1)$  that the next fish is sea bass, and prior 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.
- These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears.

# Prior Probability

- Suppose for a moment that we were forced to make a decision about the type of fish that will appear next without being allowed to see it.
- If a decision must be made with so little information, it seems logical to use the following decision rule: Decide  $\omega_1$  if  $P(\omega_1) > P(\omega_2)$ ; otherwise decide  $\omega_2$ .

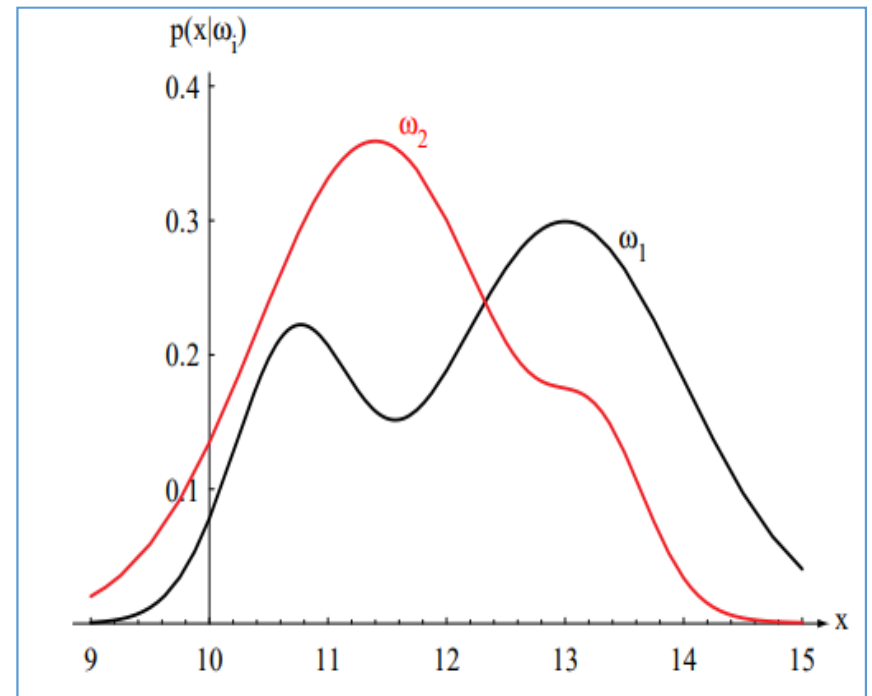
# Improving the Decision rule

- In most circumstances we are not asked to make decisions with so little information.
- In our example, we might for instance use a lightness measurement  $x$  to improve our classifier. Different fish will yield different lightness readings and we express this variability in probabilistic term using  $p(x|\omega_1)$  and  $p(x|\omega_2)$ .

.

# Improving the Decision rule

- In most circumstances we are not asked to make decisions with so little information.
- In our example, we might for instance use a lightness measurement  $x$  to improve our classifier. Different fish will yield different lightness readings and we express this variability in probabilistic term using  $p(x|\omega_1)$  and  $p(x|\omega_2)$ .



# Posterior Likelihood

- Suppose that we know both the prior probabilities  $P(\omega_j)$  and the conditional densities  $p(x|\omega_j)$ .
- We note first that the (joint) probability density of finding a pattern that is in category  $\omega_j$  and has feature value  $x$  can be written two ways:

$$p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j).$$



# Posterior Likelihood

- Suppose that we know both the prior probabilities  $P(\omega_j)$  and the conditional densities  $p(x|\omega_j)$ .
- We note first that the (joint) probability density of finding a pattern that is in category  $\omega_j$  and has feature value  $x$  can be written two ways:

$$p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j).$$

$$P(\omega_1/x) = \frac{P(x/\omega_1)P(\omega_1)}{P(x)}$$

$$\begin{aligned} P(\omega_1/x) &= \frac{P(\omega_1, x)}{P(x)} \\ &= \frac{P(x/\omega_1)P(\omega_1)}{P(x)} \end{aligned}$$

# Bayes' Formula

$$P(\omega_1 | x) = \frac{P(x | \omega_1) P(\omega_1)}{P(x)}$$

$$P(\omega_2 | x) = \frac{P(x | \omega_2) P(\omega_2)}{P(x)}$$

likelihood

posterior

prior

evidence

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

where in this case of two categories

$$P(\omega_1 | x = 14.5)$$

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

$$P(x | \omega_1)$$

$$P(x = 14.5 | \omega_1)$$

# Posterior Probability

- We call  $p(x | \omega_j)$  the likelihood of  $\omega_j$  with respect to  $x$  .
- Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability.
- The evidence factor,  $p(x)$ , can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must.



# Posterior Probability

$$P(\omega_1 / x) \geq P(\omega_2 / x)$$

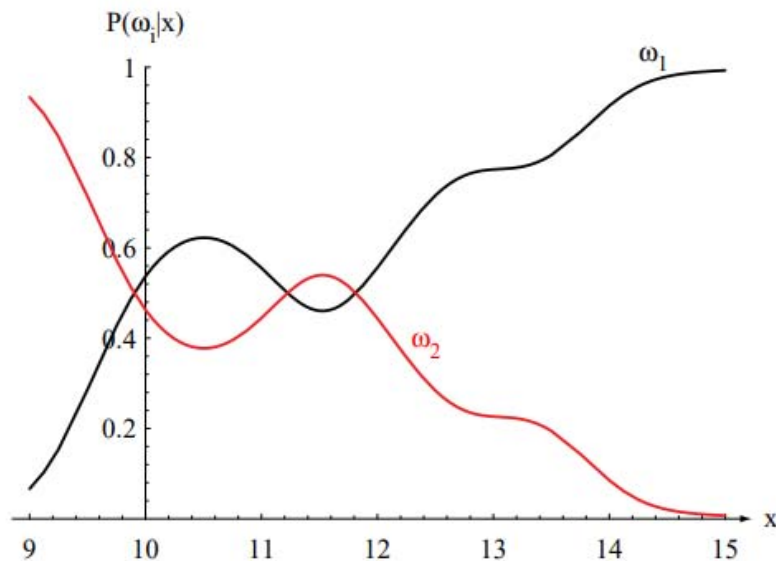
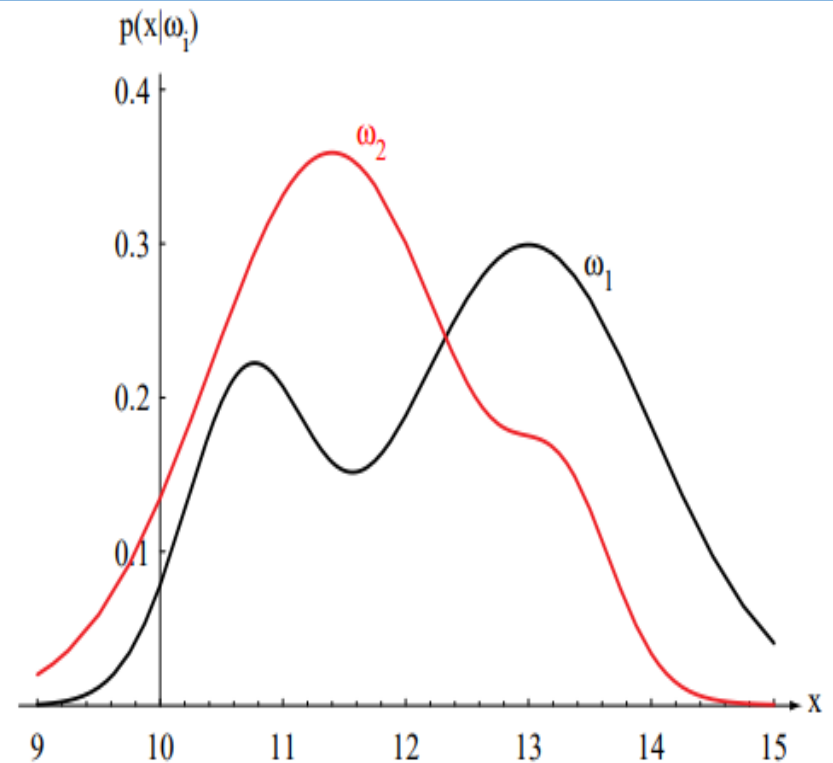
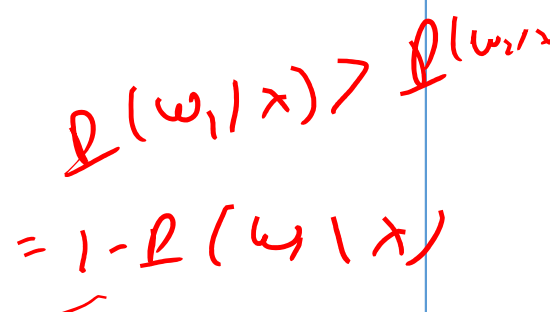


Figure 2.2: Posterior probabilities for the particular priors  $P(\omega_1) = 2/3$  and  $P(\omega_2) = 1/3$  for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value  $x = 14$ , the probability it is in category  $\omega_2$  is roughly 0.08, and that it is in  $\omega_1$  is 0.92. At every  $x$ , the posteriors sum to 1.0.



# Decision Rule

- 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  $\omega_1$ , otherwise, we choose  $\omega_2$ .
- Whenever we observe a particular  $x$ ,  
 $P(\text{error}|x) = P(\omega_1|x)$  if we decide  $\omega_2$ .  
 $P(\omega_2|x)$  if we decide  $\omega_1$ .  

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

# Decision Rule

- Decide  $\omega_1$  if  $P(\omega_1|x) > P(\omega_2|x)$ ,  
decide  $\omega_2$  , otherwise.

and under this rule

$$P(\text{error}|x) = \min [P(\omega_1|x), P(\omega_2|x)]$$

# Decision Rule

- Note that the evidence,  $p(x)$  is unimportant as far as making a decision is concerned.
- Its presence in Eq. 1 assures us that  $P(\omega_1|x) + P(\omega_2|x) = 1$ . By eliminating this scale factor, we obtain the following completely equivalent decision rule

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





# Discriminant Functions

$$g_1(x) = \log(p(x|\omega_1) P(\omega_1))$$
$$g_2(x) = \log(p(x|\omega_2) P(\omega_2))$$

- Define a set of discriminant functions for each class

$$g_i(x), i = 1, \dots, c.$$

- The classifier is said to assign a feature vector  $x$  to class  $\omega_i$  if  $g_i(x) > g_j(x)$  for all  $j \neq i$

$$g_i(x) = p(\omega_i | x) \text{ or}$$

$$g_i(x) = p(x | \omega_i) P(\omega_i) \text{ or}$$

$$g_i(x) = \ln p(x | \omega_i) + \ln P(\omega_i),$$

# Discriminant Functions

- The effect of any decision rule is to divide the feature decision space into  $c$  decision regions,  $R_1, \dots, R_c$ .
- If  $g_i(x) > g_j(x)$  for all  $j \neq i$ , then  $x$  is in region  $R_i$ , and the decision rule calls for us to assign  $x$  to  $\omega_i$ .

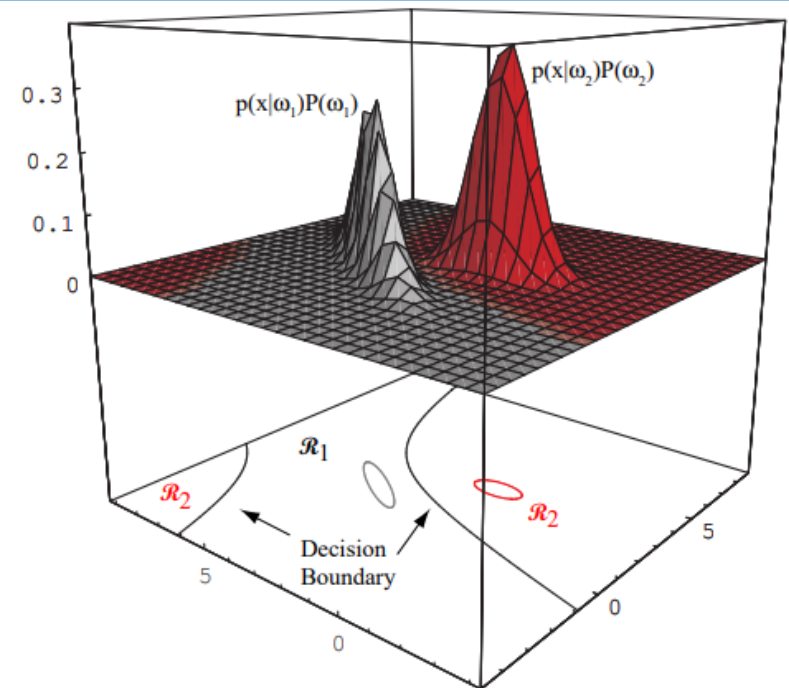


Figure 2.6: In this two-dimensional two-category classifier, the probability densities are Gaussian (with  $1/e$  ellipses shown), the decision boundary consists of two hyperbolas, and thus the decision region  $R_2$  is not simply connected.

# The Two-Category Case

- Instead of using two dichotomizer discriminant functions  $g_1$  and  $g_2$  and assigning  $x$  to  $\omega_1$  if  $g_1(x) > g_2(x)$ ,

- It is more common to define a single discriminant function

$$g(x) = g_1(x) - g_2(x)$$

and to use the following decision rule:

$$g(x) > 0$$

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

$$\Rightarrow g_1(x) > g_2(x) \Rightarrow \text{Class } \omega_1$$

# The Two-Category Case

- Decide  $\omega_1$  if  $g(x) > 0$ ; otherwise decide  $\omega_2$ .

- $g(x) = P(\omega_1 | x) - P(\omega_2 | x)$   
or

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

$$\log\left(\frac{p(x|\omega_1)}{p(x|\omega_2)}\right)$$

$$g_1(x) = \log p(x|\omega_1) + \log p(\omega_1)$$
$$g_2(x) = \log p(x|\omega_2) + \log p(\omega_2)$$

$$g(x) = g_1(x) - g_2(x)$$
$$= \log p(x|\omega_1) - \log p(x|\omega_2) + \log p(\omega_1) - \log p(\omega_2)$$

# Discriminant Functions for the Normal Density

- we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i).$$

$$g_1(x) = \log p(x|\omega_1) + \log P(\omega_1)$$

$$g_2(x) = \log p(x|\omega_2) + \log P(\omega_2)$$

- This expression can be readily evaluated if the densities  $p(x|\omega_i)$  are multivariate normal, i.e.,

$$\text{if } p(x|\omega_i) \sim N(\mu_i, \Sigma_i).$$

In this case, we have

$$g(x) = g_1(x) - g_2(x)$$

$g(x) > 0 \rightarrow \text{decide } 1$   
Otherwise - decide 2 or other

$$p(x|\omega_1) \sim N(\mu_1, \Sigma_1)$$

$$p(x|\omega_2) \sim N(\mu_2, \Sigma_2)$$

# Discriminant Functions for the Normal Density

- we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(x) = \ln p(x | \omega_i) + \ln P(\omega_i).$$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$

- This expression can be readily evaluated if the densities  $p(x | \omega_i)$  are multivariate normal, i.e.,

$$\text{if } p(x | \omega_i) \sim N(\mu_i, \Sigma_i).$$

$$p(x | \omega_i)$$

In this case, we have

$$g_i(x) = \left[ \frac{-1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \right] - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

$$p(x/w_1) \sim N(\mu_1, \Sigma_1)$$

$$g_1(x) = \log p(x/w_1) + \log p(w_1)$$

$$\log \left( \frac{1}{(\sqrt{2\pi})^d |\Sigma_1|^{1/2}} e^{-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)} \right) + \log p(w_1)$$

$$= \left( -\frac{d}{2} \log \sqrt{2\pi} + \frac{1}{2} \log |\Sigma_1| \right) - \frac{1}{2} (x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1) + \log p(w_1)$$

$$\Sigma_1 = \Sigma_2 = \sigma^2 I$$

$$-\frac{1}{2} (x-\mu_1)^T \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}^{-1} (x-\mu_1)$$



# Discriminant Functions for the Normal Density

- we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(x) = \ln p(x | \omega_i) + \ln P(\omega_i).$$

- This expression can be readily evaluated if the densities  $p(x | \omega_i)$  are multivariate normal, i.e.,

$$\text{if } p(x | \omega_i) \sim N(\mu_i, \Sigma_i).$$

In this case, we have

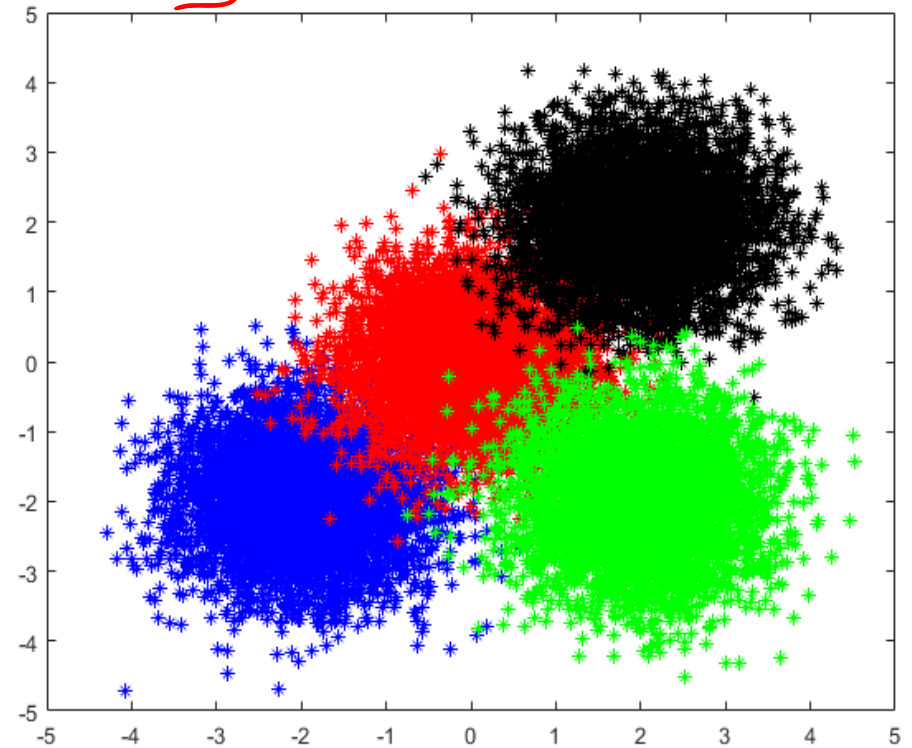
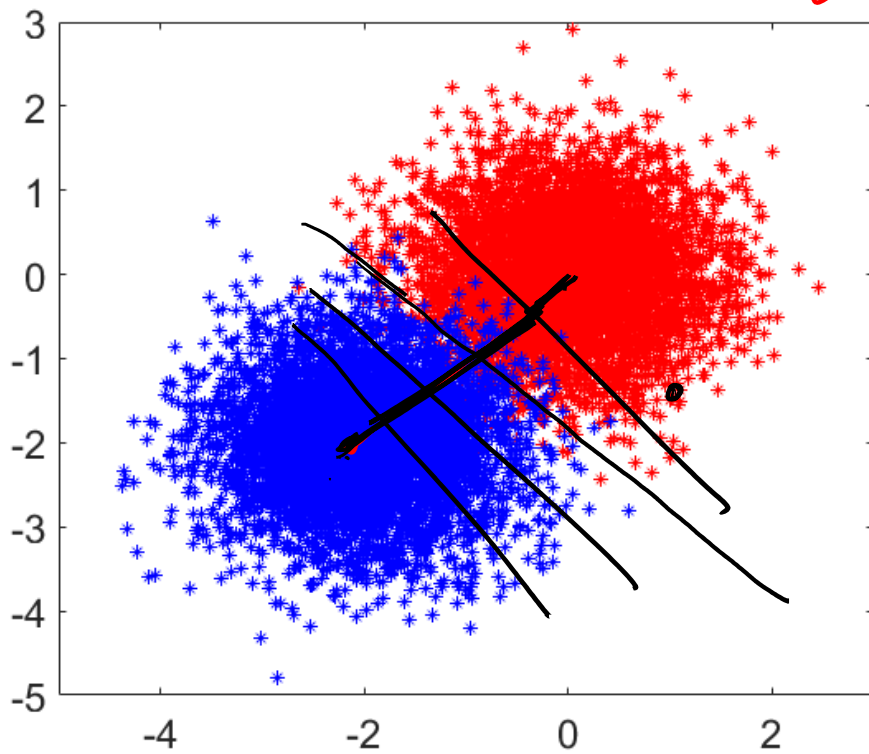
$$g_i(x) = \frac{-1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

Case 1 :  $\Sigma_i = \sigma^2 I$

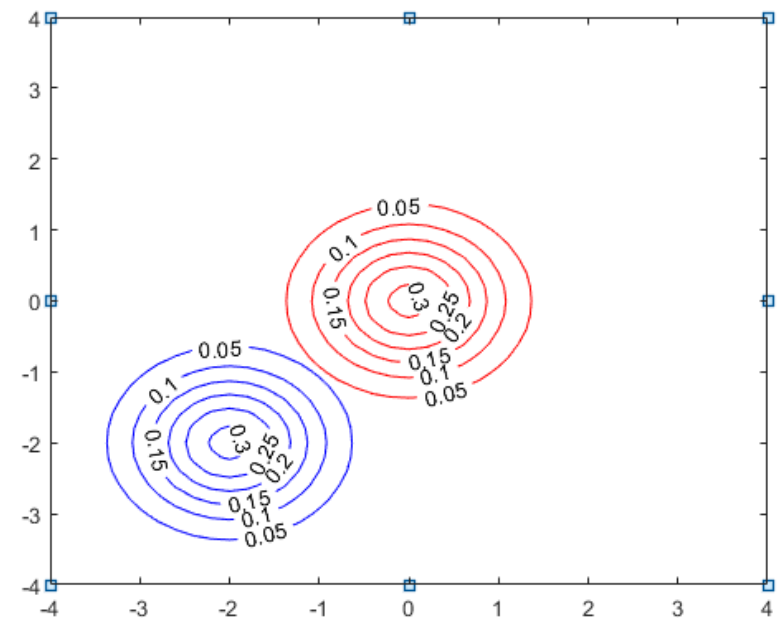
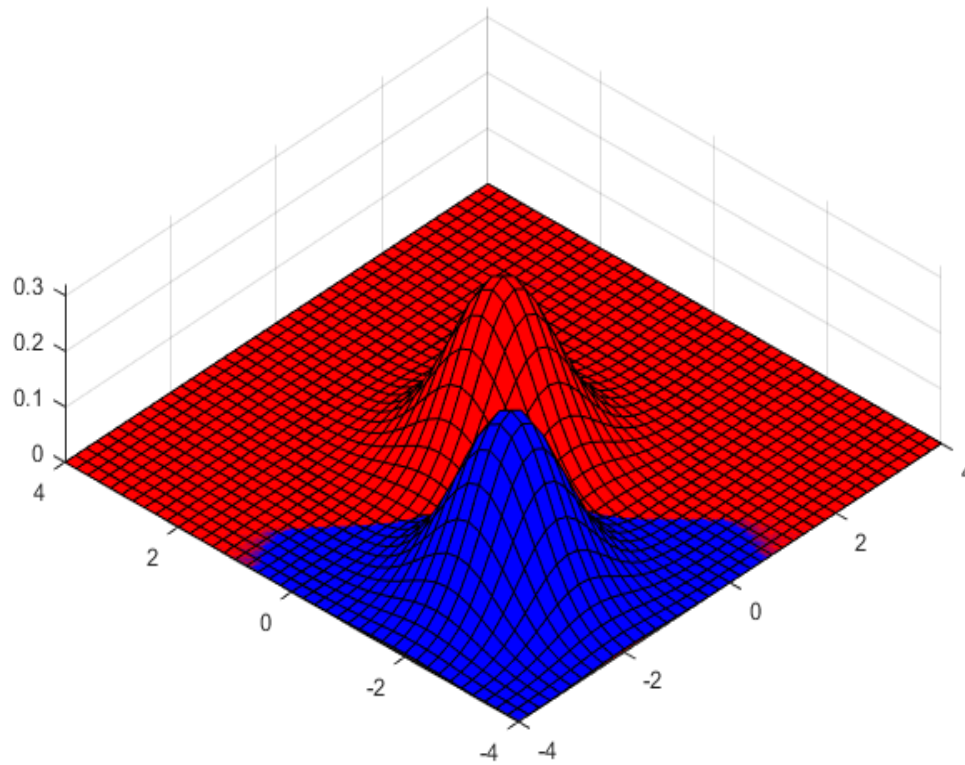
$$\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix}$$

$$\begin{bmatrix} \sigma_1^2 \\ \sigma_2^2 \end{bmatrix}$$

$$\Sigma_1 = \Sigma_2 = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$



Case 1 :  $\Sigma_i = \sigma^2 I$



Case 1 :  $\Sigma_i = \sigma^2 I$   ~~$g_1(x) = -\frac{1}{2} \|x - \mu_1\|^2$~~   ~~$+ \ln P(w_1)$~~   
 $g_2(x) = -\frac{1}{2} \|x - \mu_2\|^2$

- $g_i(x) = -\frac{(x - \mu_i)^T (x - \mu_i)}{2\sigma^2} + \ln P(w_i)$

$$g_i(x) = -\frac{(x^T x - 2\mu_i^T x + \mu_i^T \mu_i)}{2\sigma^2} + \ln P(w_i)$$

However, the quadratic term  $x^T x$  is the same for all  $i$ , making it an ignorable additive constant.

$$g_1(x) = \log p(x|w_1) + \log p(w_1)$$

$$g_2(x) = \log p(x|w_2) + \log p(w_2)$$

~~x~~  $x|w_1 \sim N(\mu_1, \sigma^2 I)$   
 $x|w_2 \sim N(\mu_2, \sigma^2 I)$

$$\log \left( \frac{1}{(\sqrt{2\pi})^d |\Sigma|^{1/2}} \exp \frac{-1}{2} \frac{(x-\mu_1)^T (x-\mu_1)}{\sigma^2} \right) + \log p(w_1)$$

$$= \frac{-d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \frac{(x-\mu_1)^T (x-\mu_1)}{\sigma^2} + \log p(w_1)$$

$$g_2(x) = \frac{-d}{2} \log 2\pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \frac{(x-\mu_2)^T (x-\mu_2)}{\sigma^2} + \log p(w_2)$$

$$g_1(x) = \frac{-1}{2} \frac{(x-\mu_1)^T (x-\mu_1)}{\sigma^2} + \log p(w_1)$$

Case 1 :  $\Sigma_i = \sigma^2 I$

$$g_1(x) = \frac{2\mu_1^T x}{2\sigma^2} - \frac{\mu_1^T \mu_1}{2\sigma^2} + \log P(w_1)$$

$$g_i(x) = - \frac{(x^T x - 2\mu_i^T x + \mu_i^T \mu_i)}{2\sigma^2} + \ln P(w_i)$$

However, the quadratic term  $x^T x$  is the same for all  $i$ , making it an ignorable additive constant. Thus, we obtain the equivalent linear discriminant functions

$$g_i(x) = \frac{1}{\sigma^2} \mu_i^T x + \left[ \frac{-1}{2\sigma^2} \mu_i^T \mu_i + \ln P(w_i) \right] - w_0$$

$$g_i(x) = w_1^T x + w_0$$

$$w_1 = \frac{1}{\sigma^2} \mu_i$$

Case 1 :  $\Sigma_i = \sigma^2 I$  , for two category case

$$g_1(x) = \frac{1}{\sigma^2} \mu_1^T x + \frac{-1}{2\sigma^2} \mu_1^T \mu_1 + \ln P(w_1)$$

$$g_2(x) = \frac{1}{\sigma^2} \mu_2^T x + \frac{-1}{2\sigma^2} \mu_2^T \mu_2 + \ln P(w_2)$$

$$g(x) = g_1(x) - g_2(x) = 0$$

$$\frac{1}{\sigma^2} (\mu_1 - \mu_2)^T x + \frac{-1}{2\sigma^2} (\mu_1^T \mu_1 - \mu_2^T \mu_2) + \frac{\ln P(w_1)}{\ln P(w_2)} = 0$$

$$(\mu_1 - \mu_2)^T x + \frac{-1}{2} (\mu_1 - \mu_2)^T (\mu_1 + \mu_2) + \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)^T (\mu_1 - \mu_2)}{||(\mu_1 - \mu_2)||^2} = 0$$

$$\log p(w_1) - \log p(w_2)$$

Case 1 :  $\Sigma_i = \sigma^2 I$ , for two category case  $\log \frac{p(w_1)}{p(w_2)}$

$$(\mu_1 - \mu_2)^T x + \frac{-1}{2} (\mu_1 - \mu_2)^T (\mu_1 + \mu_2) + \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)^T (\mu_1 - \mu_2)}{\|(\mu_1 - \mu_2)\|^2} = 0$$

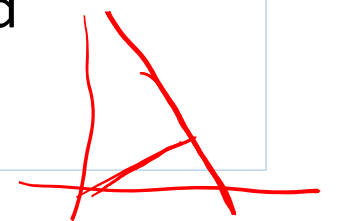
$$(\mu_1 - \mu_2)^T \left( x - \left( \frac{1}{2} (\mu_1 + \mu_2) - \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)}{\|(\mu_1 - \mu_2)\|^2} \right) \right) = 0$$

$$w^T (x - x_0) = 0$$

$$3x_1 + 3x_2 = 5$$

$$\frac{1}{\sqrt{18}} [3 \ 3] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \frac{-5}{\sqrt{18}} = 0$$

This equation defines a hyperplane through the point  $x_0$  and orthogonal to the vector  $w$ .





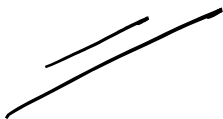
Case 1 :  $\Sigma_i = \sigma^2 I$  , for two category case

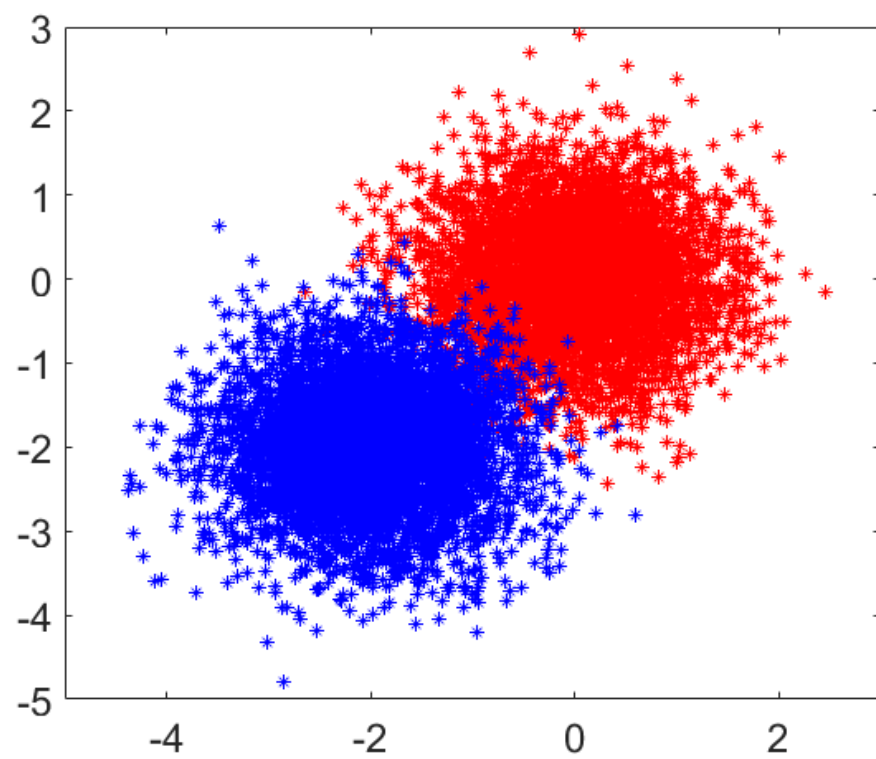
$$(\mu_1 - \mu_2)^T \left( X - \left( \frac{1}{2} (\mu_1 + \mu_2) - \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)}{\|(\mu_1 - \mu_2)\|^2} \right) \right) = 0$$

$$w^T (x - x_0) = 0$$

This equation defines a hyperplane through the point  $x_0$  and orthogonal to the vector  $w$ .

Since  $w = \mu_1 - \mu_2$  , the hyperplane separating  $R_1$  and  $R_2$  is orthogonal to the line linking the means. 

If  $P(w_1) = P(w_2)$  , the second term on the right of Equation vanishes, and thus the point  $x_0$  is halfway between the means, and the hyperplane is the perpendicular bisector of the line between the means . 



Case 1 :  $\Sigma_i = \sigma^2 I$  , for two category case

$$(\mu_1 - \mu_2)^T \left( X - \left( \frac{1}{2}(\mu_1 + \mu_2) - \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)}{\|(\mu_1 - \mu_2)\|^2} \right) \right) = 0$$

$$w^T(x - x_0) = 0$$

Since  $w = \mu_1 - \mu_2$  , the hyperplane separating  $R_1$  and  $R_2$  is orthogonal to the line linking the means.

Case 1 :  $\Sigma_i = \sigma^2 I$  , for two category case

$$(\mu_1 - \mu_2)^T \left( X - \left( \frac{1}{2} (\mu_1 + \mu_2) - \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)}{\|(\mu_1 - \mu_2)\|^2} \right) \right) = 0$$

$$w^T (x - x_0) = 0$$

This equation defines a hyperplane through the point  $x_0$  and orthogonal to the vector  $w$ .

Since  $w = \mu_1 - \mu_2$  , the hyperplane separating  $R_1$  and  $R_2$  is orthogonal to the line linking the means.

If  $P(\omega_1) \neq P(\omega_2)$  , the second term on the right of Equation vanishes, and thus the point  $x_0$  is halfway between the means, and the hyperplane is the perpendicular bisector of the line between the means .

## Case 1 : $\Sigma_i = \sigma^2 I$ , for two category case

$$(\mu_1 - \mu_2)^T \left( x - \left( \frac{1}{2}(\mu_1 + \mu_2) - \frac{\sigma^2 \ln P(w_1)}{\ln P(w_2)} \frac{(\mu_1 - \mu_2)}{\|(\mu_1 - \mu_2)\|^2} \right) \right) = 0$$
$$w^T(x - x_0) = 0$$

This equation defines a hyperplane through the point  $x_0$  and orthogonal to the vector  $w$ .

Since  $w = \mu_1 - \mu_2$  , the hyperplane separating  $R_1$  and  $R_2$  is orthogonal to the line linking the means.

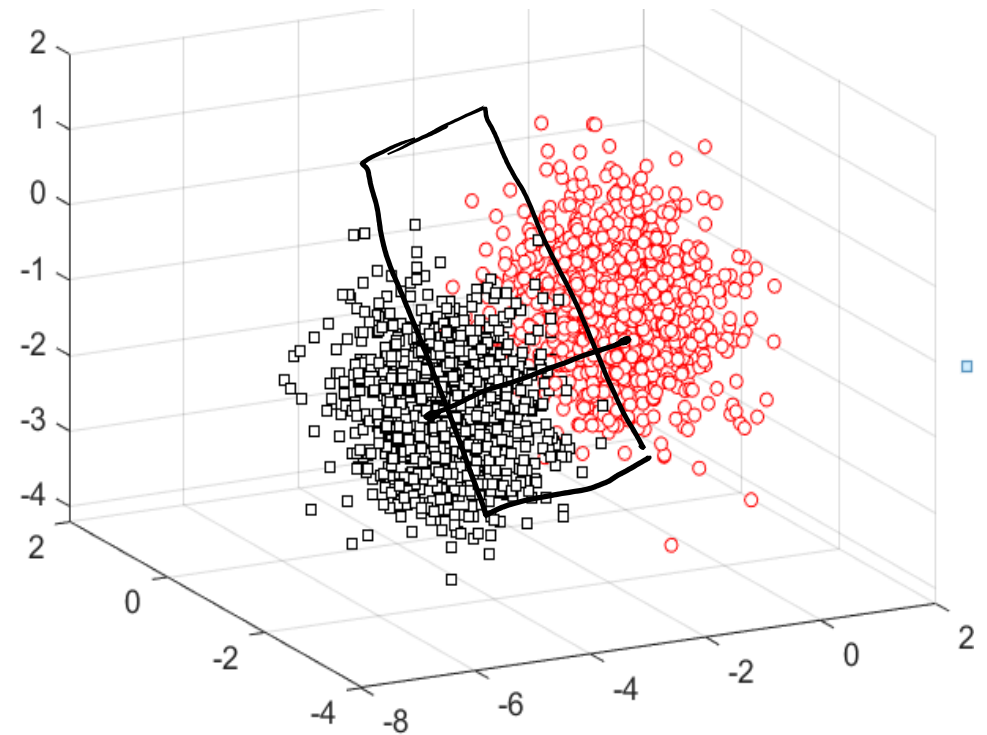
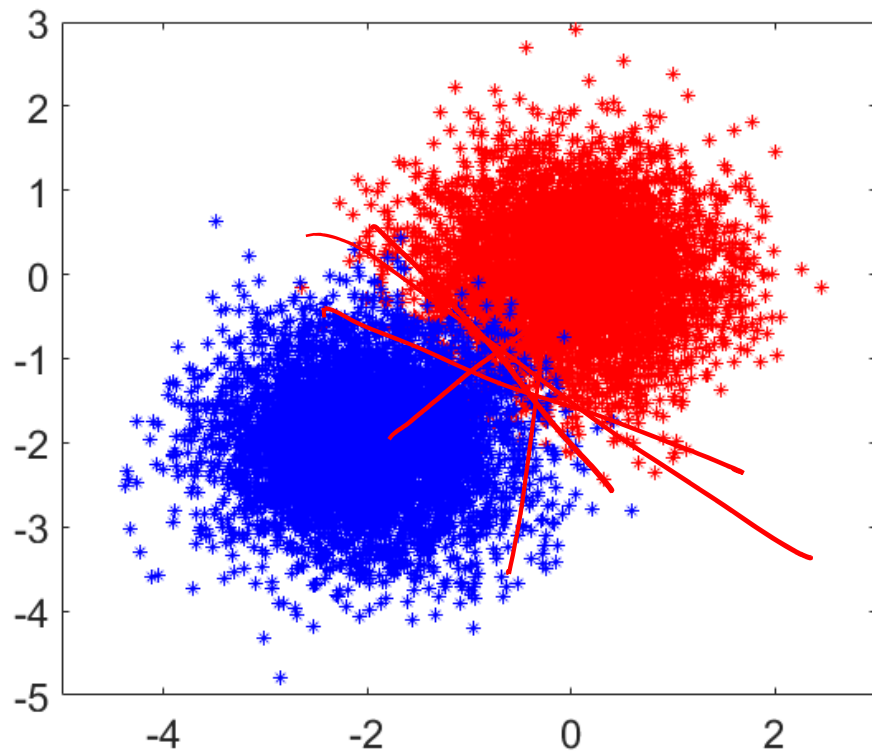
If  $P(\omega_1) \neq P(\omega_2)$  , the second term on the right of Equation vanishes, and thus the point  $x_0$  is halfway between the means, and the hyperplane is the perpendicular bisector of the line between the means .

Note, however, that if the variance  $\sigma^2$  is small relative to the squared distance  $\|\mu_1 - \mu_2\|^2$  , then the position of the decision boundary is relatively insensitive to the exact values of the prior probabilities.

## Case 1 : $\Sigma_i = \sigma^2 I$

- 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.
- When this happens, the optimum decision rule can be stated very simply:  
to classify a feature vector  $x$ , measure the Euclidean distance  $||x - \mu_i||$  from each  $x$  to each of the  $c$  mean vectors, and assign  $x$  to the category of the nearest mean.
- Such a classifier is minimum called a ~~minimum distance classifier~~. If each mean vector is thought of as being an distance classifier ideal prototype or template for patterns in its class, then this is essentially a template matching procedure .

Case 1:  $\Sigma_i = \sigma^2$



Case 1:  $\Sigma_i = \Sigma$

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- Another simple case arises when the covariance matrices for all of the classes are identical but otherwise arbitrary. Geometrically, this corresponds to the situation in which the samples fall in hyperellipsoidal clusters of equal size and shape, the cluster for the  $i^{\text{th}}$  class being centered about the mean vector  $\mu_i$ .

$$g_i(x) = \underbrace{-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i|}_{\text{independent of } i} + \ln P(\omega_i) \quad \Sigma_i = \Sigma$$

Since both  $|\Sigma_i|$  and the  $(d/2) \ln 2\pi$  term in above equations are independent of  $i$ , they can be ignored as superfluous additive constants. This simplification leads to the discriminant functions

$$g_i(x) = \underbrace{-\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i)}_{\text{discriminant function}} + \ln P(\omega_i)$$

$$g_1(x) = -\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)$$

$$g_2(x) = -\frac{1}{2}(x - \mu_2)^T \Sigma^{-1}(x - \mu_2)$$



## Case 1 : $\Sigma_i = \Sigma$

- If the prior probabilities  $P(\omega_i)$  are the same for all  $c$  classes, then the  $\ln P(\omega_i)$  term can be ignored.  
In this case, the optimal decision rule can once again be stated very simply:  
to classify a feature vector  $x$ ,  
measure the squared Mahalanobis distance =  $\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i)$   
from  $x$  to each of the  $c$  mean vectors,  
and assign  $x$  to the category of the nearest mean.
- As before, unequal prior probabilities bias the decision in favor of the a priori more likely category.

## Case 1 : $\Sigma_i = \Sigma$

- If the prior probabilities  $P(\omega_i)$  are the same for all  $c$  classes, then the  $\ln P(\omega_i)$  term can be ignored.  
In this case, the optimal decision rule can once again be stated very simply:  
to classify a feature vector  $x$ ,  
measure the squared Mahalanobis distance =  $\frac{1}{2}(x - \mu_i)^T \Sigma^{-1}(x - \mu_i)$   
from  $x$  to each of the  $c$  mean vectors,  
and assign  $x$  to the category of the nearest mean.
- As before, unequal prior probabilities bias the decision in favor of the a priori more likely category.

$$g(x) = g_1(x) - g_2(x) = 0$$

$$a^T b = b^T a$$

$$\frac{-1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) + \left( \log p(\mu_1) - \left( \frac{-1}{2} (x - \mu_2)^T \Sigma^{-1} (x - \mu_2) \right) + \log p(\mu_2) \right)$$

$$= \frac{-1}{2} \cancel{x^T \Sigma^{-1} x} = \mu_1^T \Sigma^{-1} \mu_1 - 2 \underbrace{(x^T \Sigma^{-1} \mu_1)} + \log p(\mu_1) =$$

$$- \left( \frac{-1}{2} \cancel{x^T \Sigma^{-1} x} - \mu_2^T \Sigma^{-1} \mu_2 - 2 \underbrace{x^T \Sigma^{-1} \mu_2} + \log p(\mu_2) \right) = 0$$

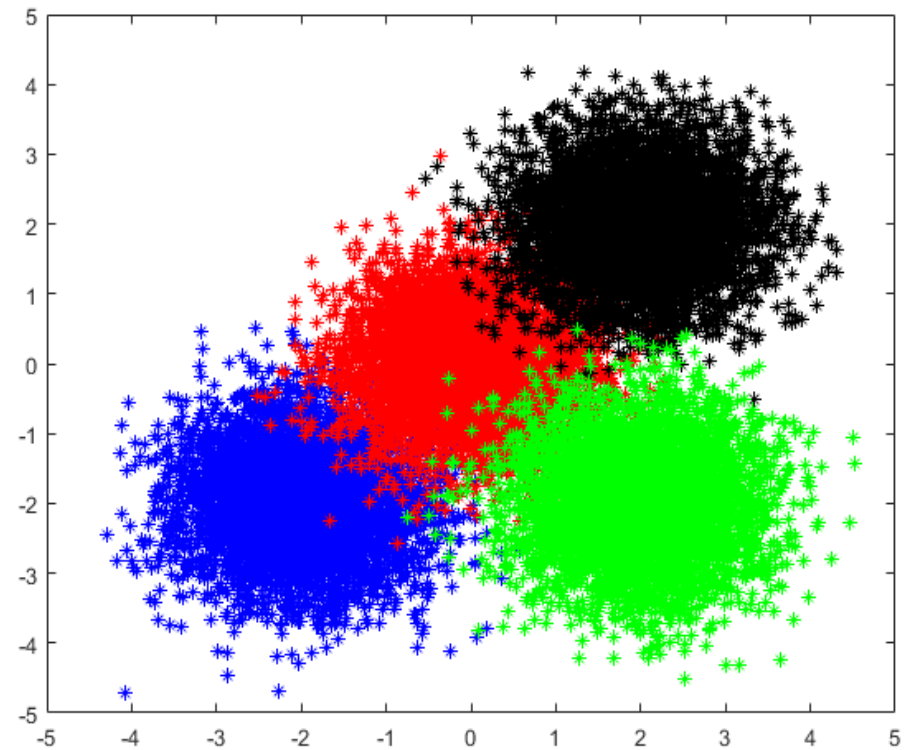
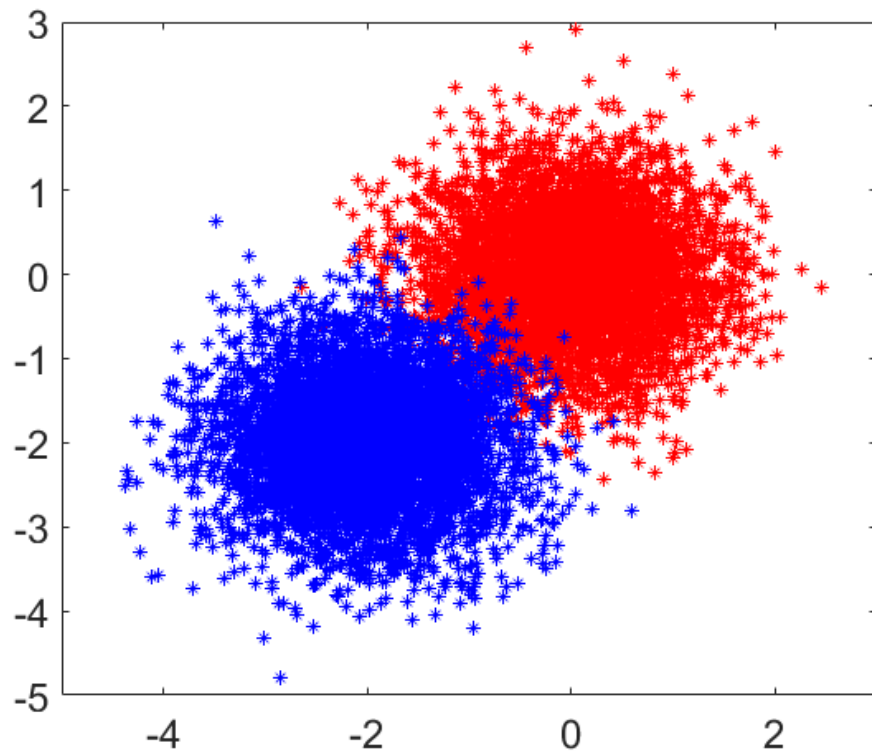
$$= \Sigma^{-1} (\mu_2 - \mu_1)^T x + \frac{1}{2} (\mu_2^T \Sigma^{-1} \mu_2 + \mu_1^T \Sigma^{-1} \mu_1) + \log \left( \frac{p(\mu_1)}{p(\mu_2)} \right) \\ \Sigma^{-1} (\mu_1 - \mu_2)^T x + \frac{1}{2} \underbrace{(\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)}_{\mu_1 - \mu_2} - \log \frac{p(\mu_1)}{p(\mu_2)} = 0$$

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

$$\begin{aligned} & \Sigma^{-1}(\mu_1 - \mu_2)^T x + \frac{1}{2}(\mu_1 + \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2) - \log\left(\frac{P(\omega_1)}{P(\omega_2)}\right) = 0 \\ & \Sigma^{-1}(\mu_1 - \mu_2)^T \left[ x - \left( \frac{1}{2}(\mu_1 + \mu_2) - \log\left(\frac{P(\omega_1)}{P(\omega_2)}\right) \frac{(\mu_1 - \mu_2)}{(\mu_1 - \mu_2)^T \Sigma^{-1}(\mu_1 - \mu_2)} \right) \right] \\ & \qquad \qquad \qquad = 0 \end{aligned}$$

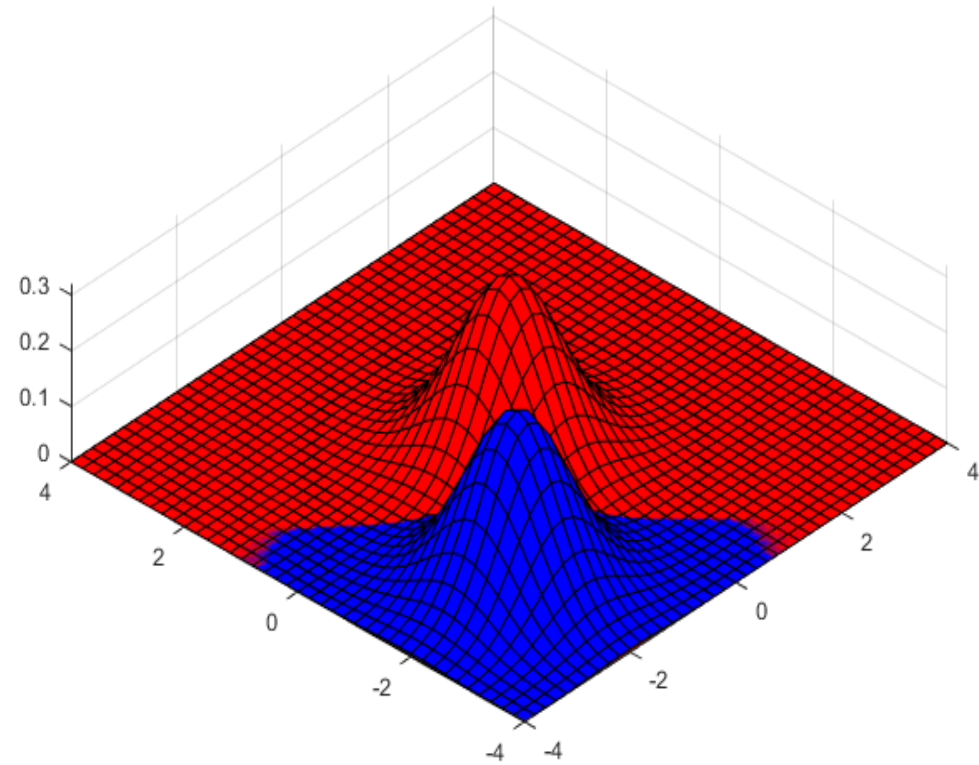
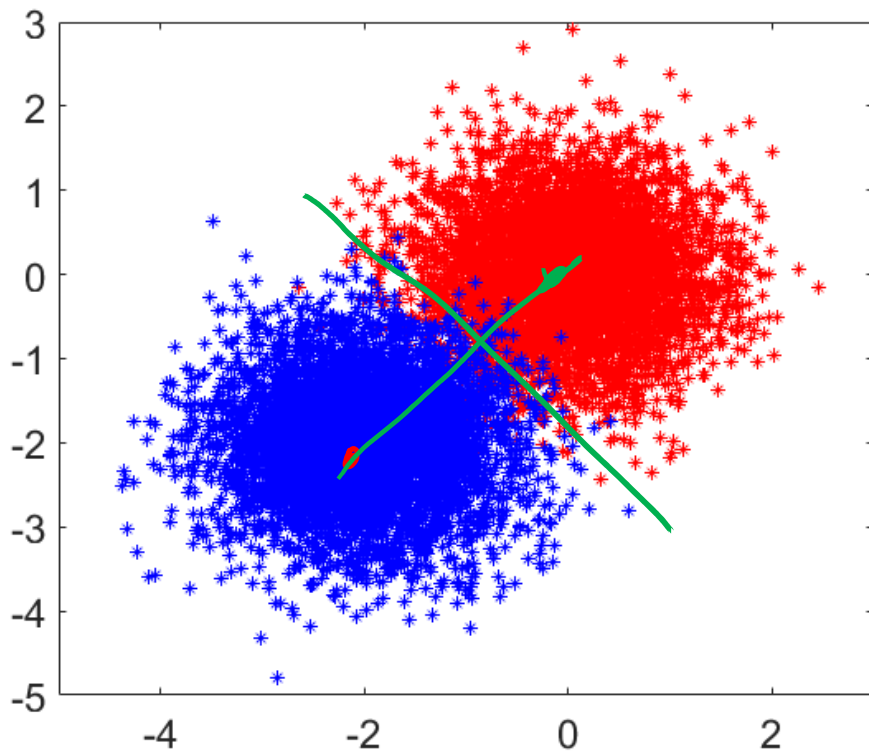


Case 1:  $\Sigma_i = \sigma^2$

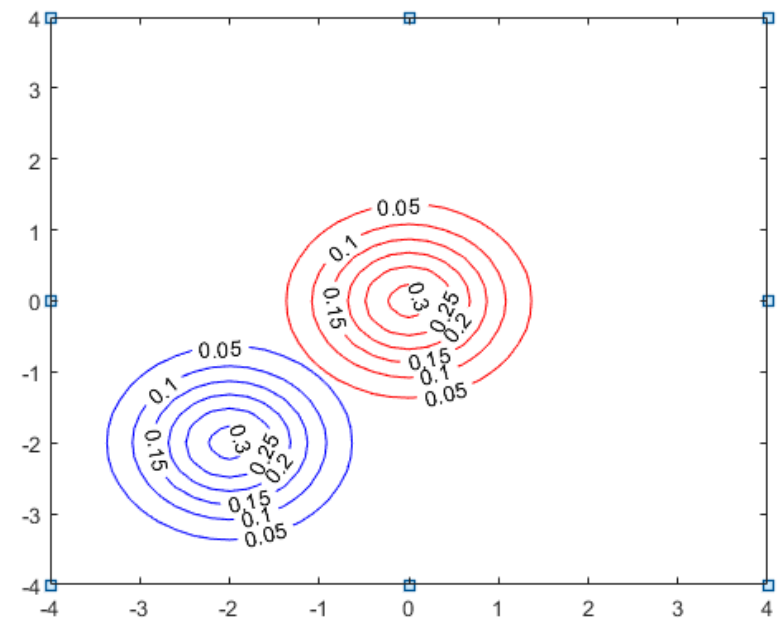
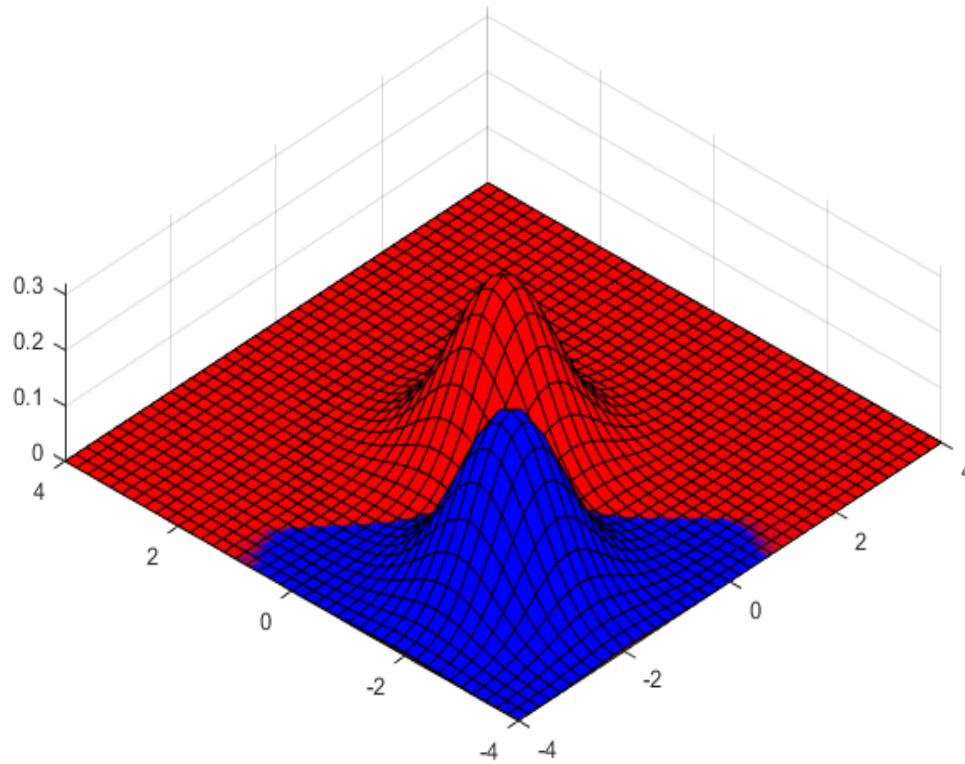


Case 1:  $\Sigma_i = \sigma^2$

$$\sum (y_1 - y_2)$$

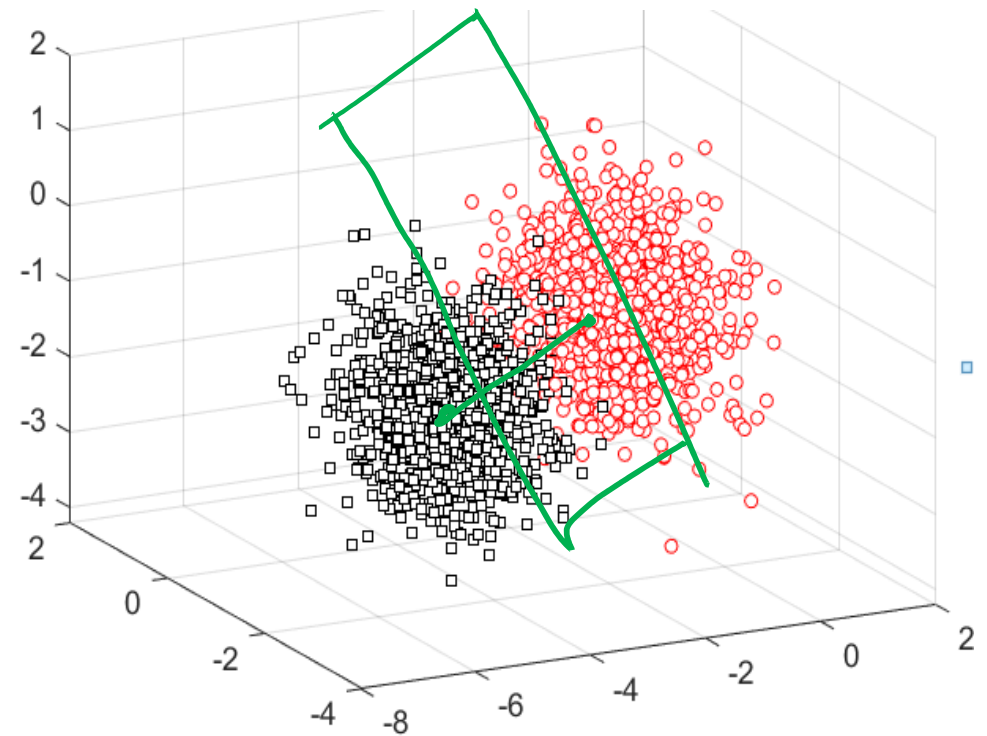
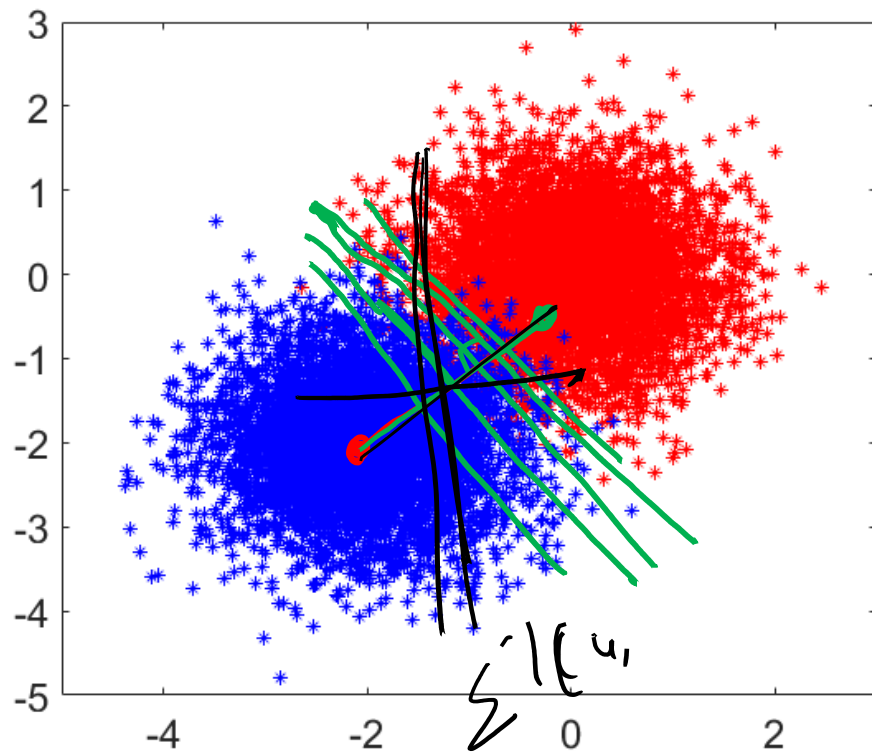


Case 1:  $\Sigma_i = \sigma^2$





Case 1:  $\Sigma_i = \sigma^2$

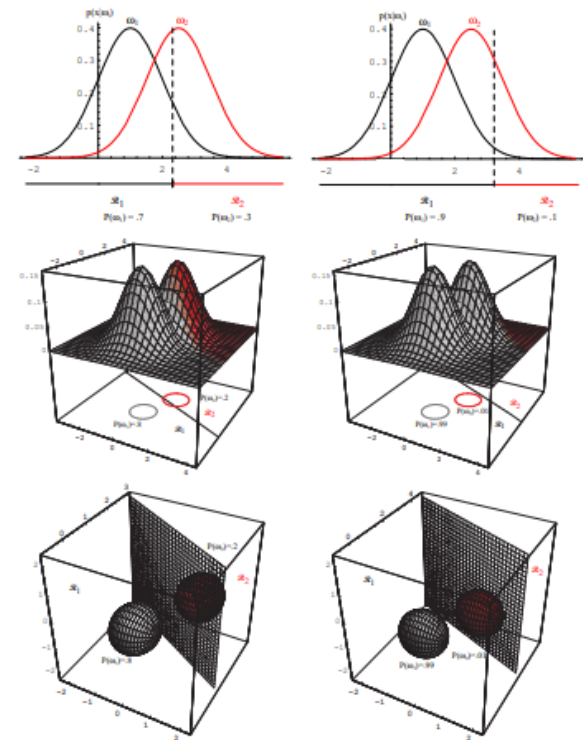


Case 1:  $\Sigma_i = \sigma^2$

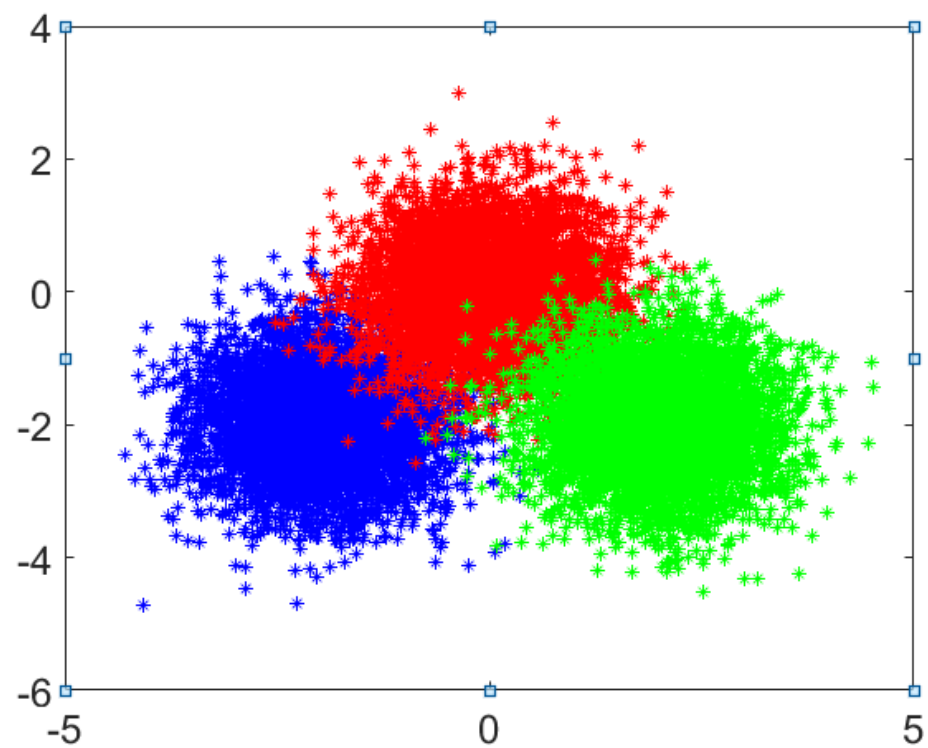
Case 1:  $\Sigma_i = \sigma^2$

Case 1:  $\Sigma_i = \sigma^2$

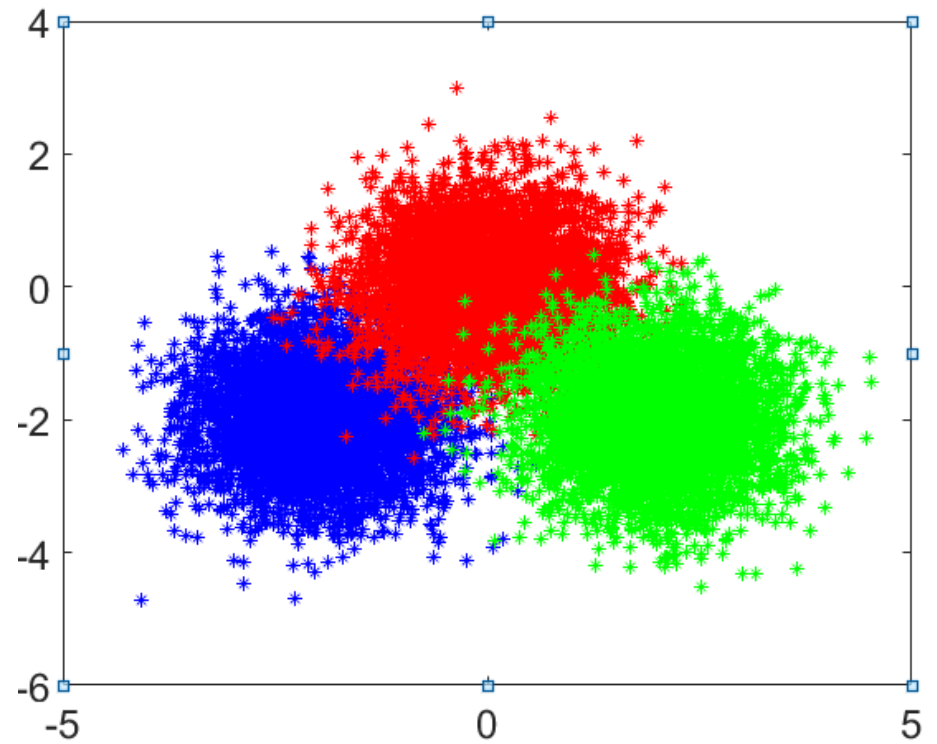
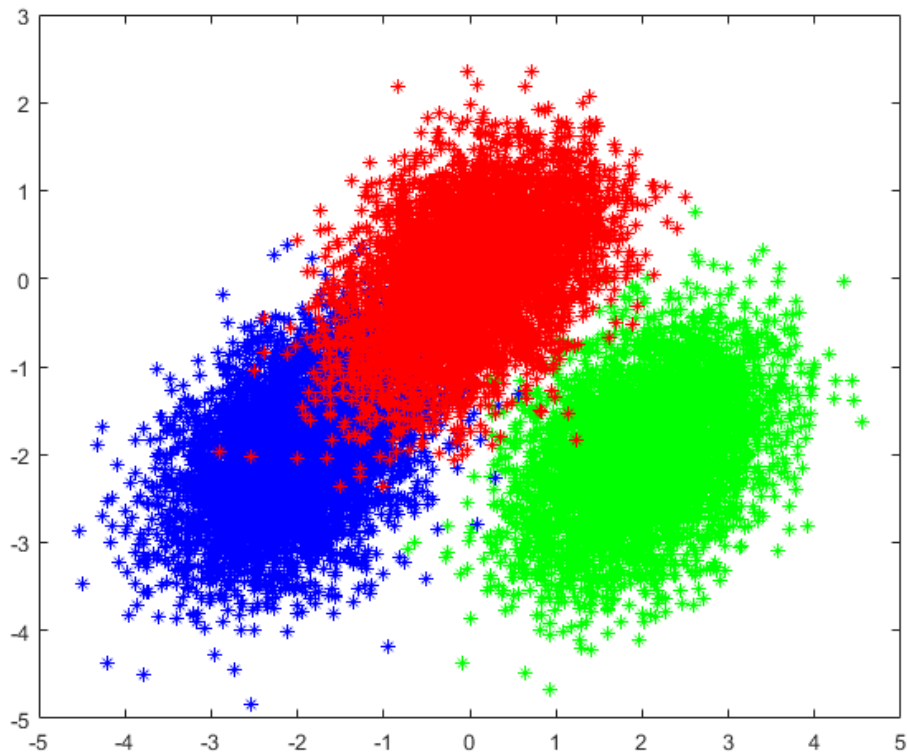
Case 1:  $\Sigma_i = \sigma^2$



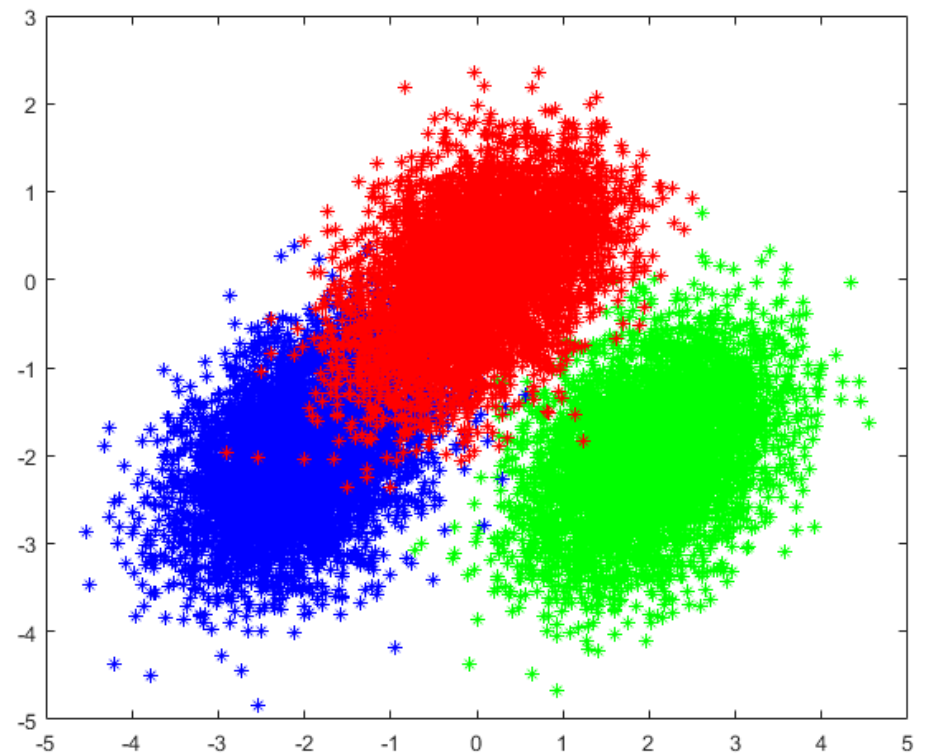
Case 1:  $\Sigma_i = \sigma^2$



Case 2:  $\Sigma_i = \Sigma$



Case 2:  $\Sigma_i = \Sigma$

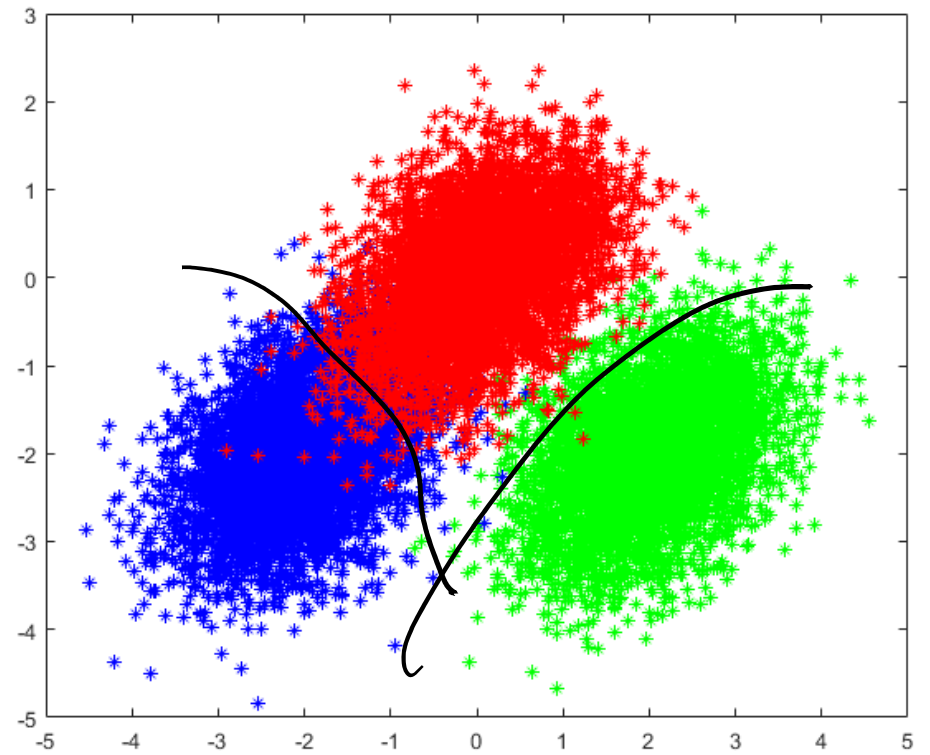




Case 2:  $\Sigma_i = \Sigma$

Case 2:  $\Sigma_i = \Sigma$

Case 2:  $\Sigma_i = \Sigma$



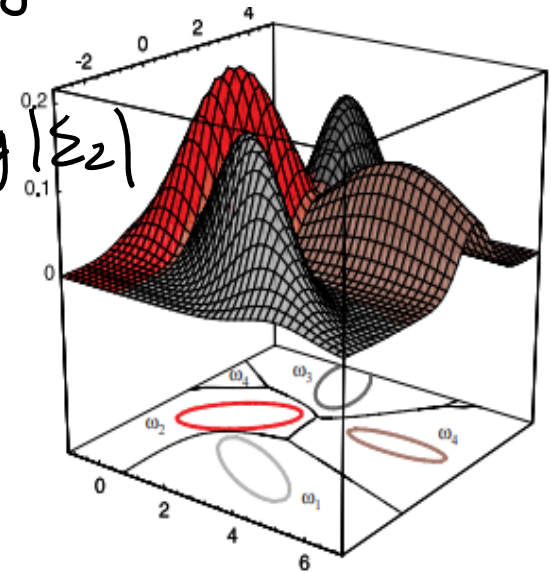


Case 3 :  $\Sigma_i$  are arbitrary

$$g_1(x) = \frac{-1}{2} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + \log p(\omega_1) - \frac{1}{2} \log |\Sigma_1|$$

$$g_2(x) = \frac{-1}{2} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) + \log p(\omega_2) - \frac{1}{2} \log |\Sigma_2|$$

$$g(x) = g_1(x) - g_2(x)$$



Case 3 :  $\Sigma_i$  are arbitrary

Case 3 :  $\Sigma_i$  are arbitrary

## Case 3 : $\Sigma_i$ are arbitrary

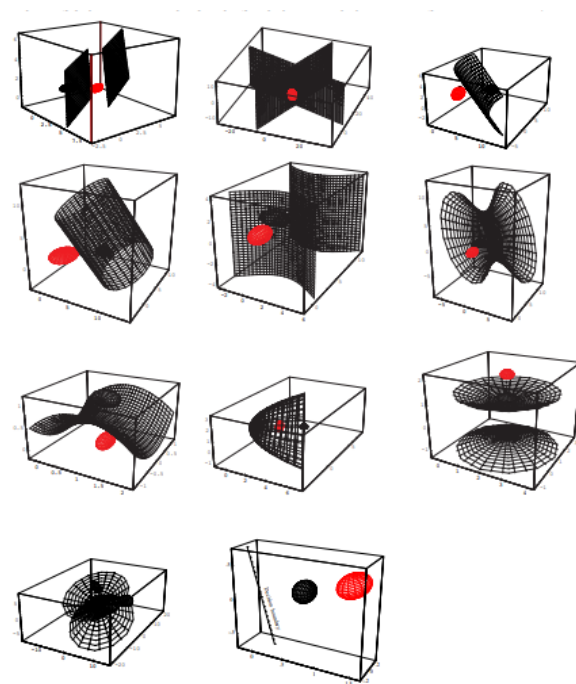


Figure 2.15: Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line.



