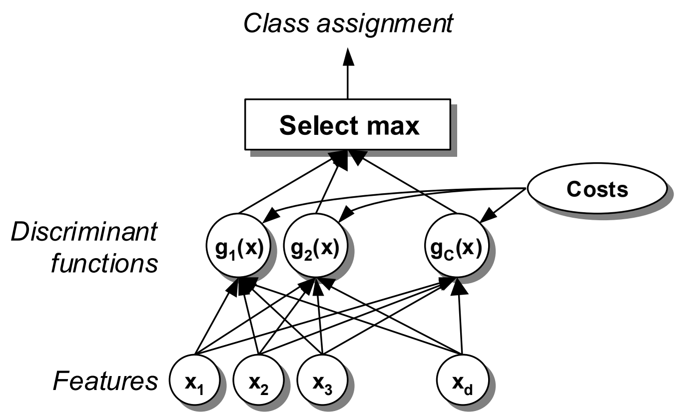
LECTURE 5: Quadratic classifiers

- **Bayes classifiers for Normally distributed classes**
  - Case 1: $\Sigma_i=\sigma^2 I$
  - Case 2: $\Sigma_i=\Sigma$ (Σ diagonal)
  - Case 3: $\Sigma_i=\Sigma$ (Σ non-diagonal)
  - Case 4: $\Sigma_i=\sigma_i^2 I$
  - Case 5: $\Sigma_i\neq\Sigma_j$ general case
- **Numerical example**
- **Linear and quadratic classifiers: conclusions**

Bayes classifiers for Normally distributed classes

- **On Lecture 4 we showed that the decision rule (MAP) that minimized the probability of error could be formulated in terms of a family of discriminant functions**

$$\text{choose } \omega_i \text{ if } g_i(x) > g_j(x)\ \forall j \neq i$$
$$\text{where } g_i(x) = P(\omega_i \mid x)$$



  - As we will show, for classes that are normally distributed, this family of discriminant functions can be reduced to very simple expressions

- **General expression for Gaussian densities**
  - The multivariate Normal density function was defined as

$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

  - Using Bayes rule, the MAP discriminant function becomes

$$g_i(x) = P(\omega_i \mid x) = \frac{P(x \mid \omega_i)P(\omega_i)}{P(x)} = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] P(\omega_i)\frac{1}{P(x)}$$

  - Eliminating constant terms

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

  - We take natural logs since the logarithm is a monotonically increasing function

$$\boxed{g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)}$$

    - This expression is called a **quadratic discriminant function**

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

- **This situation occurs when the features are statistically independent with the same variance for all classes***
  - In this case, the quadratic discriminant function becomes

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\left(\sigma^2 I\right)^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\sigma^2 I\right|\right) + \log(P(\omega_i)) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) - \frac{1}{2}N\log\left(\sigma^2\right) + \log(P(\omega_i)) \overset{\text{dropping the second term}}{=}$$

$$= -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \log(P(\omega_i))$$

  - Expanding this expression

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \log(P(\omega_i)) = -\frac{1}{2\sigma^2}\left(x^Tx - 2\mu_i^Tx + \mu_i^T\mu_i\right) + \log(P(\omega_i))$$

  - Eliminating the term x^Tx, which is constant for all classes

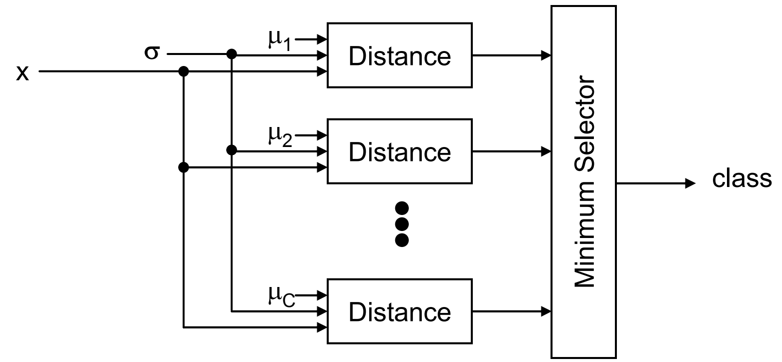
$$g_i(x) = -\frac{1}{2\sigma^2}\left(-2\mu_i^Tx + \mu_i^T\mu_i\right) + \log(P(\omega_i)) = w_i^Tx + w_{i0}$$

$$\text{where} \begin{cases} w_i = \dfrac{\mu_i}{\sigma^2} \\ w_{i0} = -\dfrac{1}{2\sigma^2}\mu_i^T\mu_i + \log(P(\omega_i)) \end{cases}$$

    - Since the discriminant is linear, the decision boundaries $g_i(x)=g_j(x)$, will be hyper-planes
  - If we assume equal priors

$$\boxed{g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)}$$

    - This is called a **minimum-distance** or **nearest mean** classifier
    - The loci of constant probability for each class are hyper-spheres
    - For unit variance ($\sigma^2=1$), the distance becomes the Euclidean distance



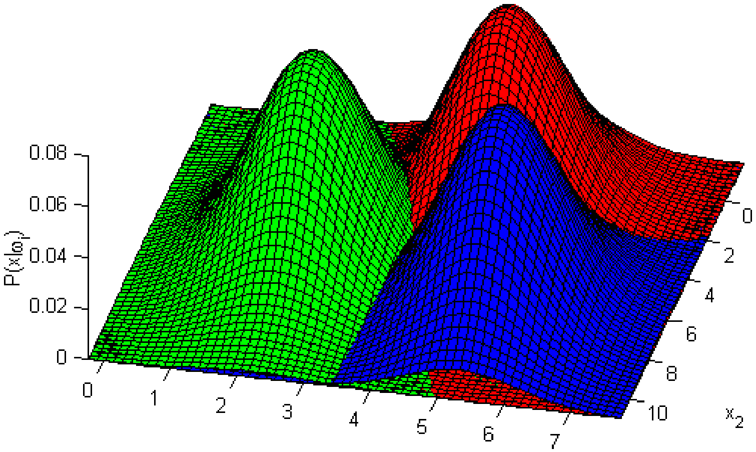
From [Schalkoff, 1992]

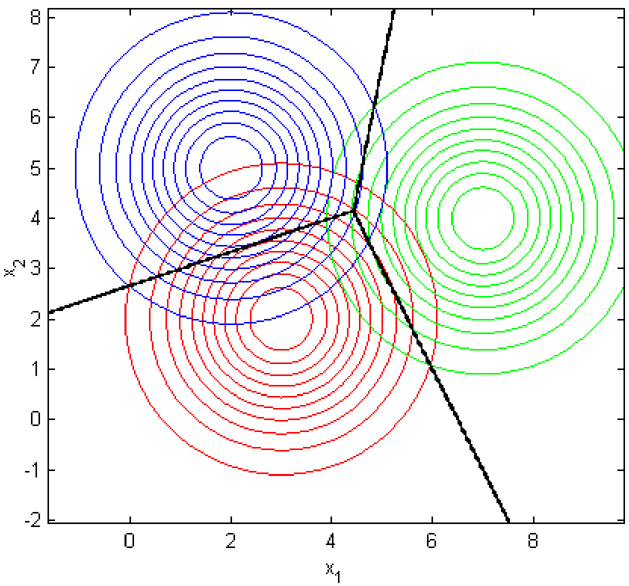
Case 1: $\Sigma_i=\sigma^2 I$, example

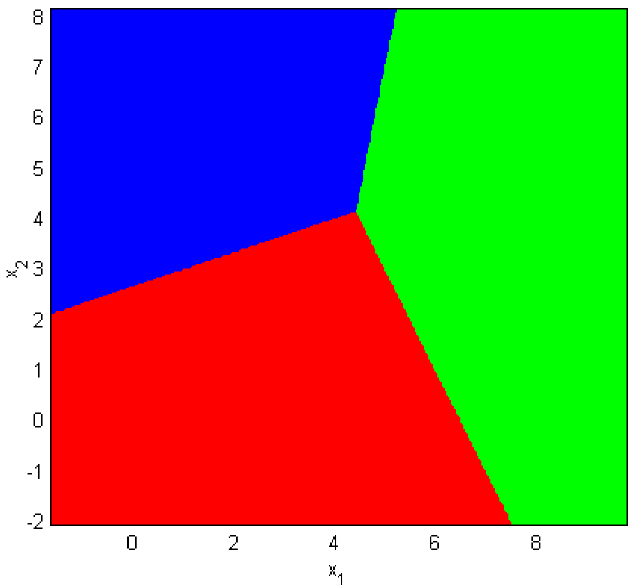
- **To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors**

$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [7 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$







Case 2: $\Sigma_i=\Sigma$ (Σ diagonal)

- **The classes still have the same covariance matrix, but the features are allowed to have different variances**
  - In this case, the quadratic discriminant function becomes

$$
\begin{aligned}
g_i(x) &= -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\Sigma_i\right|\right) + \log\left(P(\omega_i)\right) = \\
&= -\frac{1}{2}(x-\mu_i)^T \begin{bmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_N^{-2} \end{bmatrix} (x-\mu_i) - \frac{1}{2}\log\left(\begin{vmatrix} \sigma_1^{2} & & \\ & \ddots & \\ & & \sigma_N^{2} \end{vmatrix}\right) + \log\left(P(\omega_i)\right) = \\
&= -\frac{1}{2}\sum_{k=1}^{N}\frac{(x[k]-\mu_i[k])^2}{\sigma_k^2} - \frac{1}{2}\log\prod_{k=1}^{N}\sigma_k^2 + \log\left(P(\omega_i)\right) = \\
&= -\frac{1}{2}\sum_{k=1}^{N}\frac{x[k]^2 - 2x[k]\mu_i[k] + \mu_i[k]^2}{\sigma_k^2} - \frac{1}{2}\log\prod_{k=1}^{N}\sigma_k^2 + \log\left(P(\omega_i)\right)
\end{aligned}
$$

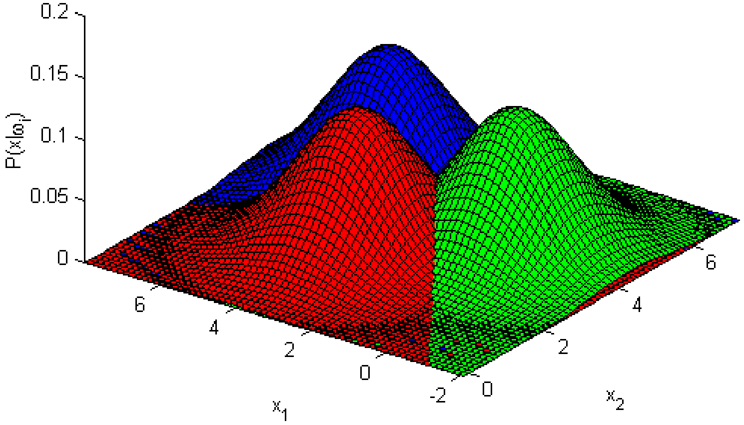
  - Eliminating the term $x[k]^2$, which is constant for all classes

$$
\boxed{g_i(x) = -\frac{1}{2}\sum_{k=1}^{N}\frac{2x[k]\mu_i[k] + \mu_i[k]^2}{\sigma_k^2} - \frac{1}{2}\log\prod_{k=1}^{N}\sigma_k^2 + \log\left(P(\omega_i)\right)}
$$

  - This discriminant is linear, so the decision boundaries $g_i(x)=g_j(x)$, will also be hyper-planes
  - The loci of constant probability are hyper-ellipses aligned with the feature axes
  - Note that the only difference with the previous classifier is that the distance of each axis is normalized by the variance of the axis

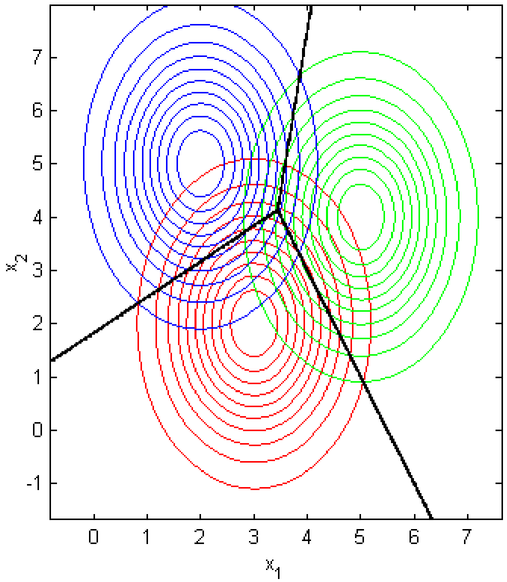
Case 2: $\Sigma_i = \Sigma$ (Σ diagonal), *example*

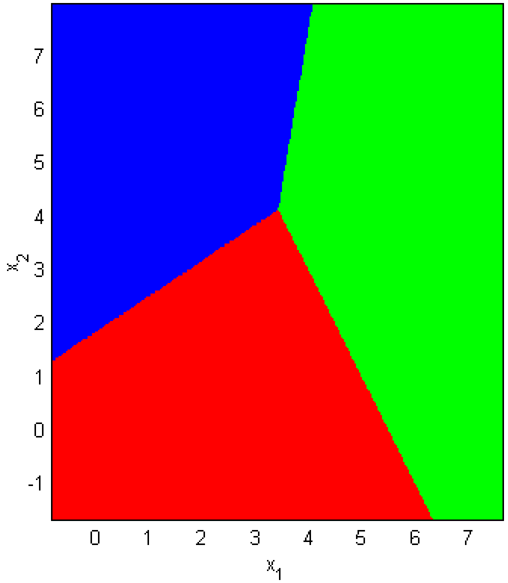
- **To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors**



$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$





Case 3: $\Sigma_i=\Sigma$ (Σ non-diagonal)

- **In this case, all the classes have the same covariance matrix, but this is no longer diagonal**
- **The quadratic discriminant becomes**

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\,\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\Sigma_i\right|\right) + \log\left(P(\omega_i)\right) =$$
$$= -\frac{1}{2}(x-\mu_i)^T\,\Sigma^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\Sigma\right|\right) + \log\left(P(\omega_i)\right)$$

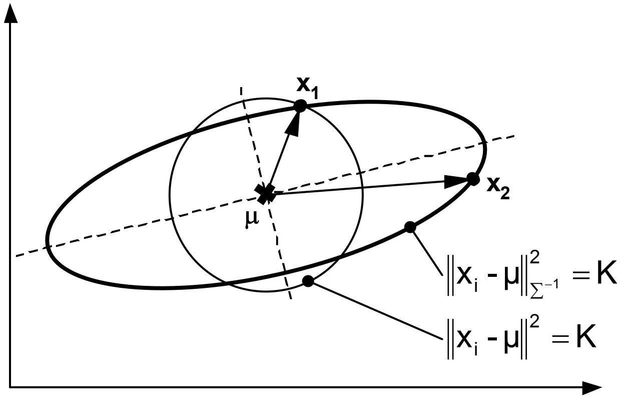
- **Eliminating the term log|Σ|, which is constant for all classes**

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\,\Sigma_i^{-1}(x-\mu_i) + \log\left(P(\omega_i)\right)$$

  - The quadratic term is called the Mahalanobis distance, a very important distance in Statistical PR

Mahalanobis Distance

$$\left\|x-y\right\|^2_{\Sigma^{-1}} = (x-y)^T\,\Sigma^{-1}(x-y)$$



- **The Mahalanobis distance is a vector distance that uses a Σ^{-1} norm**
  - Σ^{-1} can be thought of as a stretching factor on the space
  - Note that for an identity covariance matrix (Σ=I), the Mahalanobis distance becomes the familiar Euclidean distance

Case 3: $\Sigma_i=\Sigma$ (Σ non-diagonal)

- Expansion of the quadratic term in the discriminant yields

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)+\log(P(\omega_i)) = -\frac{1}{2}\left(x^T\Sigma^{-1}x-2\mu_i^T\Sigma^{-1}x+\mu_i^T\Sigma^{-1}\mu_i\right)+\log(P(\omega_i))$$

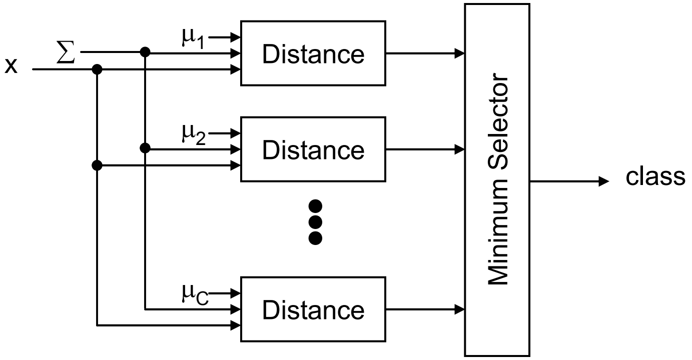
- Removing the term $x^T\Sigma^{-1}x$, which is constant for all classes

$$g_i(x) = -\frac{1}{2}\left(-2\mu_i^T\Sigma^{-1}x+\mu_i^T\Sigma^{-1}\mu_i\right)+\log(P(\omega_i))$$

- Reorganizing terms we obtain

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

$$\text{where} \begin{cases} w_i = \Sigma^{-1}\mu_i \\ w_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \log P(\omega_i) \end{cases}$$



- This discriminant is linear, so the decision boundaries will also be hyper-planes
- The constant probability loci are hyper-ellipses aligned with the eigenvectors of Σ
- If we can assume equal priors

$$\boxed{g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)}$$

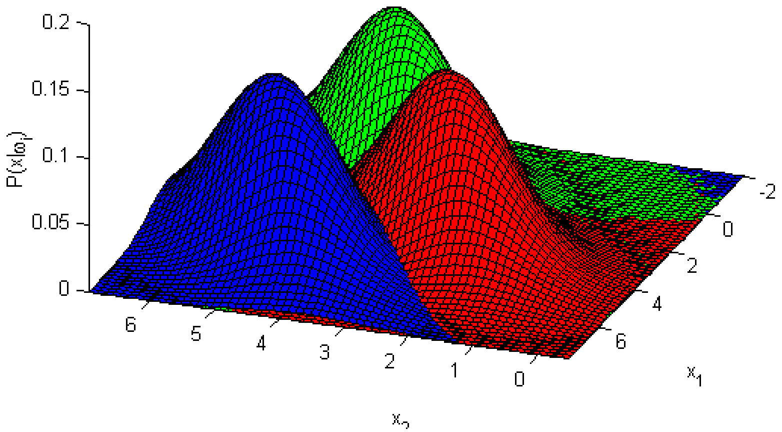
- The classifier becomes a minimum (Mahalanobis) distance classifier

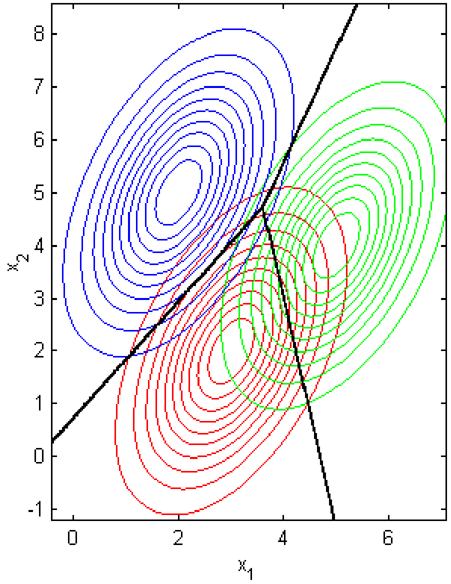
Case 3: $\Sigma_i=\Sigma$ (Σ non-diagonal), example

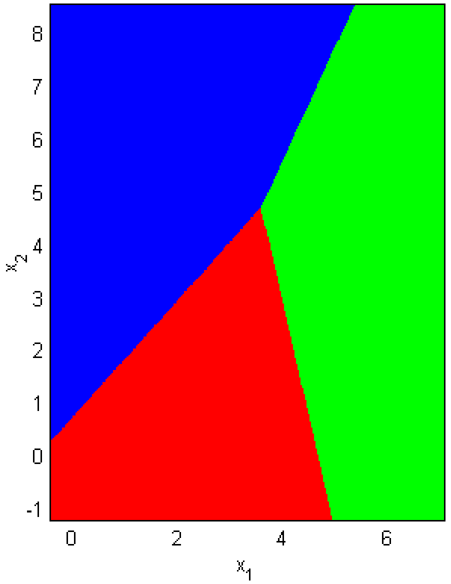
- **To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors**

$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix}$$







Case 4: $\Sigma_i = \sigma_i^2 I$

- **In this case, each class has a different covariance matrix, which is proportional to the identity matrix**
  - The quadratic discriminant becomes

$$
\begin{aligned}
g_i(x) &= -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\Sigma_i\right|\right) + \log\left(P(\omega_i)\right) = \\
&= -\frac{1}{2}(x-\mu_i)^T \sigma_i^{-2}(x-\mu_i) - \frac{1}{2}N\log\left(\sigma_i^2\right) + \log\left(P(\omega_i)\right)
\end{aligned}
$$

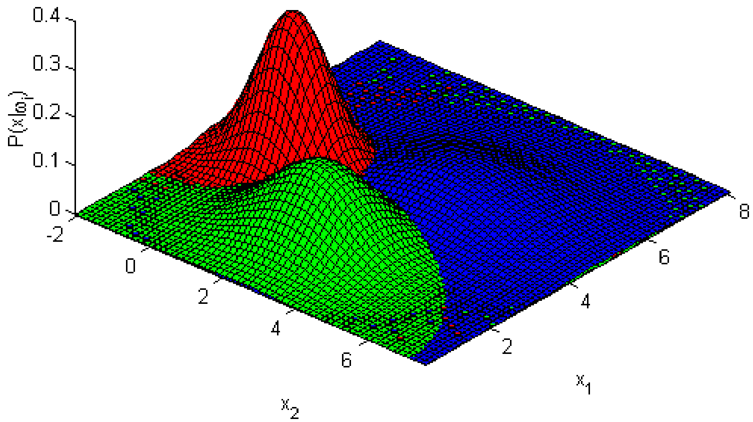
- **This expression cannot be reduced further so**
  - The decision boundaries are quadratic: hyper-ellipses
  - The loci of constant probability are hyper-spheres aligned with the feature axis

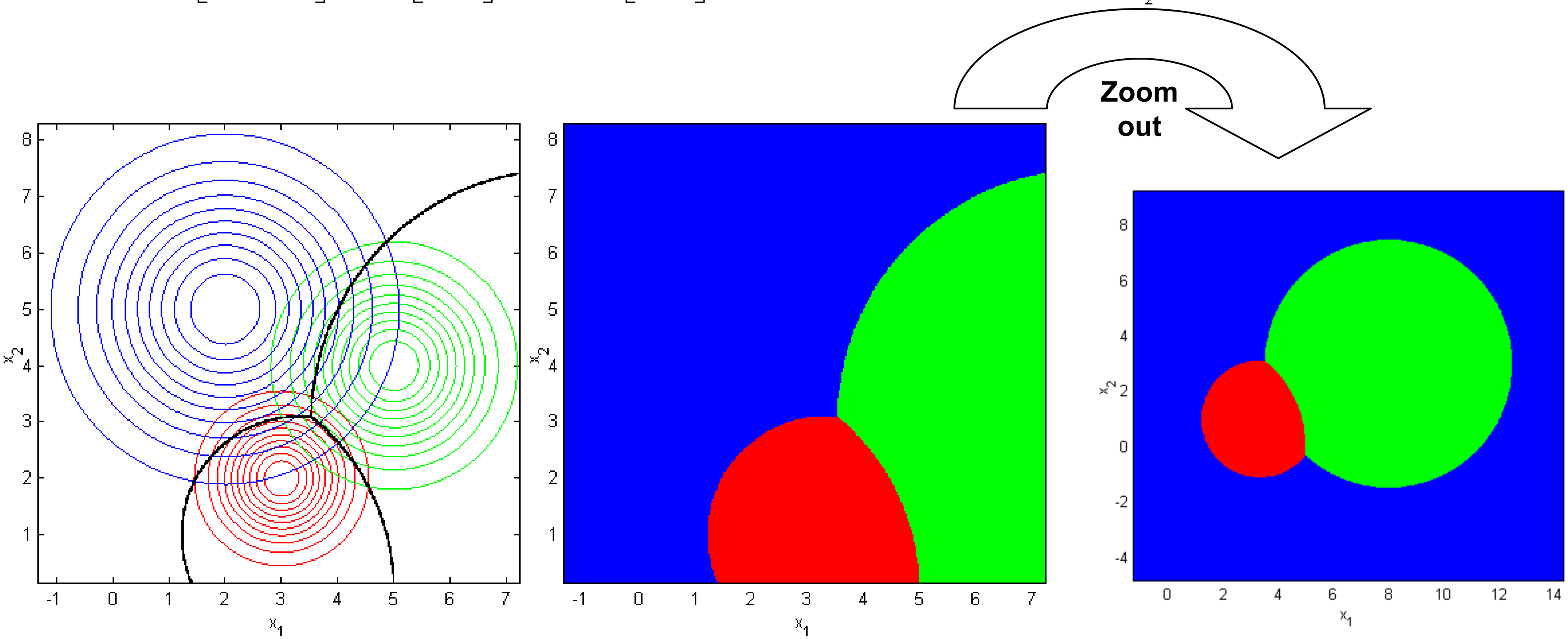
Case 4: $\Sigma_i=\sigma_i^2 I$, example

- To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors

$$\mu_1=\begin{bmatrix}3 & 2\end{bmatrix}^T \qquad \mu_2=\begin{bmatrix}5 & 4\end{bmatrix}^T \qquad \mu_3=\begin{bmatrix}2 & 5\end{bmatrix}^T$$

$$\Sigma_1=\begin{bmatrix}0.5 & 0\\ 0 & 0.5\end{bmatrix} \quad \Sigma_2=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} \quad \Sigma_3=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}$$





Case 5: $\Sigma_i \neq \Sigma_j$ general case

- **We already derived the expression for the general case at the beginning of this discussion**

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

  - Reorganizing terms in a quadratic form yields

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

$$\text{where } \begin{cases} W_i = -\frac{1}{2}\Sigma_i^{-1} \\ w_i = \Sigma_i^{-1}\mu_i \\ w_{i0} = -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i)) \end{cases}$$

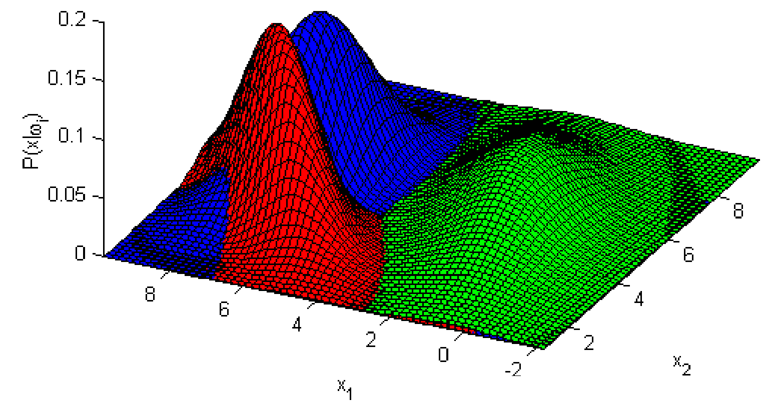
  - The loci of constant probability for each class are hyper-ellipses, oriented with the eigenvectors of Σ_i for that class
  - The decision boundaries are again quadratic: hyper-ellipses or hyper-parabolloids
  - Notice that the quadratic expression in the discriminant is proportional to the Mahalanobis distance using the class-conditional covariance Σ_i

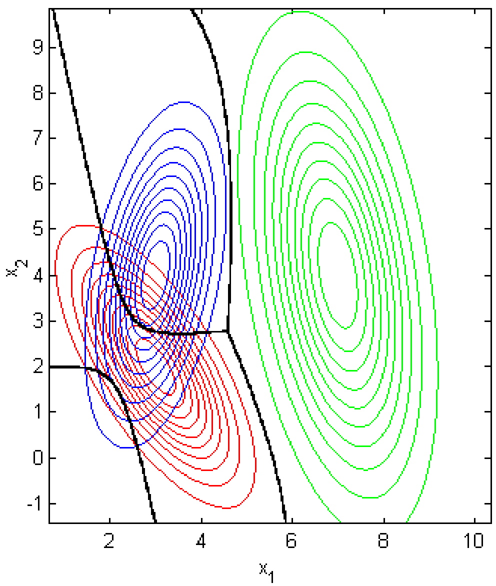
Case 5: $\Sigma_i \neq \Sigma_j$ general case, example

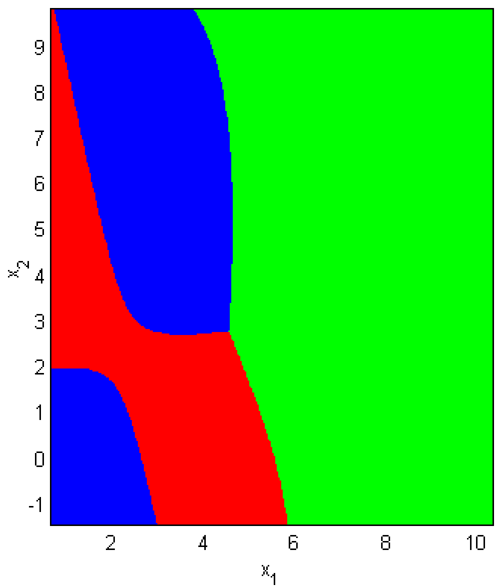
- **To illustrate the previous result, we will compute the decision boundaries for a 3-class, 2-dimensional problem with the following class mean vectors and covariance matrices and equal priors**

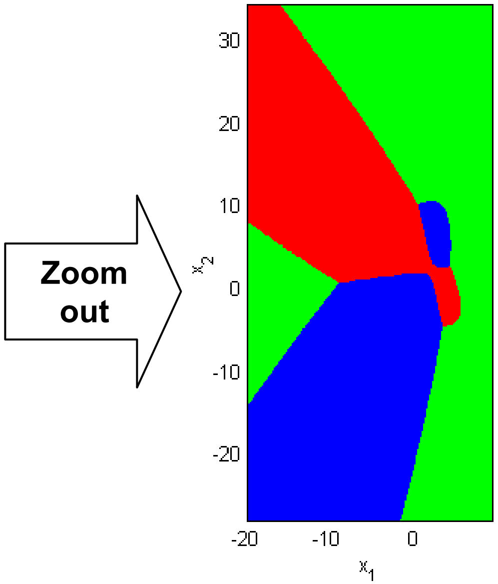
$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$









Numerical example

- **Derive a linear discriminant function for the two-class 3D classification problem defined by the following Gaussian Likelihoods**

$$\mu_1 = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T;\ \mu_2 = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T;\ \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix};\ p(\omega_2) = 2p(\omega_1)$$

- **Solution**

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \log P(\omega_i) = -\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z-\mu_z \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z--\mu_z \end{bmatrix} + \log P(\omega_i)$$

$$g_1(x) = -\frac{1}{2}\begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix} + \log\frac{1}{3};\quad g_2(x) = -\frac{1}{2}\begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix} + \log\frac{2}{3}$$

$$g_1(x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} g_2(x) \Rightarrow -2\left(x^2+y^2+z^2\right) + \log\frac{1}{3} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} -2\left((x-1)^2+(y-1)^2+(z-1)^2\right) + \log\frac{2}{3}$$

$$\boxed{x+y+z \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \frac{6-\log 2}{4} = 1.32}$$

- **Classify the test example x_u=[0.1 0.7 0.8]^T**

$$0.1+0.7+0.8 = 1.6 \underset{\omega_1}{\overset{\omega_2}{\gtrless}} 1.32 \Rightarrow x_u \in \omega_2$$

Conclusions

- **From the previous examples we can extract the following conclusions**
  - The Bayes classifier for normally distributed classes (general case) is a quadratic classifier
  - The Bayes classifier for normally distributed classes with equal covariance matrices is a linear classifier
  - The minimum Mahalanobis distance classifier is Bayes-optimal for
    - normally distributed classes and
    - equal covariance matrices and
    - equal priors
  - The minimum Euclidean distance classifier is Bayes-optimal for
    - normally distributed classes and
    - equal covariance matrices proportional to the identity matrix and
    - equal priors
  - Both Euclidean and Mahalanobis distance classifiers are linear classifiers
- **The goal of this discussion was to show that some of the most popular classifiers can be derived from decision-theoretic principles and some simplifying assumptions**
  - It is important to realize that using a specific (Euclidean or Mahalanobis) minimum distance classifier implicitly corresponds to certain statistical assumptions
  - The question whether these assumptions hold or don't can rarely be answered in practice; in most cases we are limited to posing and answering the question "*does this classifier solve our problem or not?*"