5. Bayesian Decision Theory

See: Duda and Hart Chapter 2.

5.1 The Bayes Classifier

Classifying Fish

- Simple model:
  - No posterior knowledge (i.e. no measurements)
  - Two classes

    ω_1 =“sea bass”

    ω_2=“salmon”
  - Given: $P(\omega_1)$ and $P(\omega_2)$
  - Goal:
    - Minimize the number of fish that get the wrong label

How would you set up a decision rule?

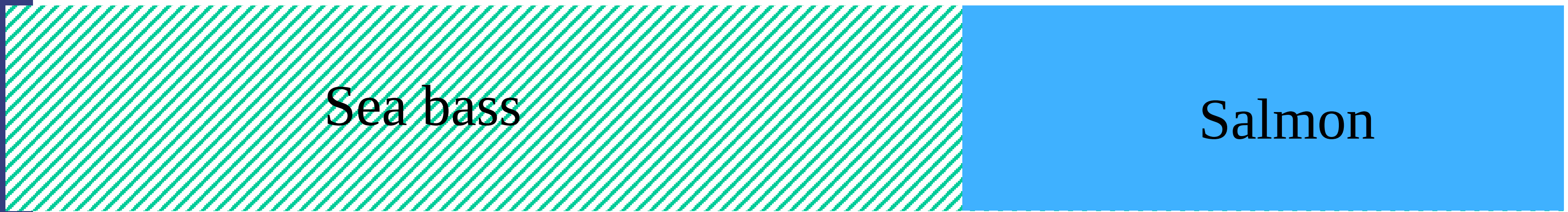
Classifying Fish

Sea bass	Salmon
$P(\omega_1)$	$P(\omega_2)$

Classify every fish as

Classifying Fish

Incorrectly classified

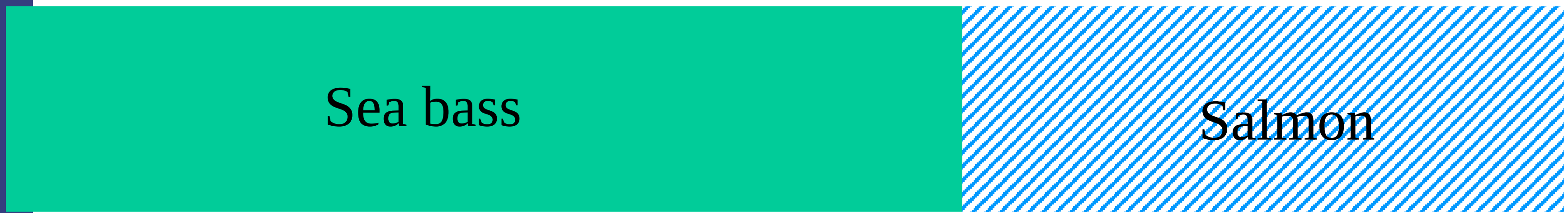


$P(\omega_1)$ $P(\omega_2)$

Classify every fish as salmon

Classifying Fish

Incorrectly classified



$P(\omega_1)$ $P(\omega_2)$

Classify every fish as “see bass”

Smaller number of fish with wrong label

Generalization

- Minimize number of wrong labels

  $\mapsto$pick class with highest probability

Formal notation:

$$\overline{\omega_i} = \underset{\omega_k}{\arg\max}\, P(\omega_k)$$

Available Measurements x

- Feature vector x from measurement
- Probabilities depend on x

$$P(\omega_k \mid x)$$

- Definition conditional probability:

$$P(\omega_k \mid x) = \frac{P(\omega_k, x)}{P(x)}$$

Bayes Decision Rule: Draft Version

- Bayes decision rule

$$\overline{\omega_i} = \arg\max_{\omega_k} P(\omega_k \mid x)$$

Ugly: usually x is measured for a given class ω_k

Rewrite Bayes Decision Rule

$$\overline{\omega_i} = \arg\max_{\omega_k} P(\omega_k \mid x)$$

Use definition of cond. probability

$$= \arg\max_{\omega_k} \frac{P(x \mid \omega_k)P(\omega_k)}{P(x)}$$

$$P(\omega_k \mid x) = \frac{P(\omega_k, x)}{P(x)}$$

$$= \frac{P(x \mid \omega_k)P(\omega_k)}{P(x)}$$

$$= \arg\max_{\omega_k} P(x \mid \omega_k)P(\omega_k)$$

P(x) does not affect decision

Bayes Decision Rule

$$\overline{\omega_i} = \arg\max_{\omega_k} P(x \mid \omega_k) P(\omega_k)$$

Terminology

Prior: $P(\omega_k)$

Posterior: $P(\omega_k \mid x)$

Cost of Making Errors

- The fish is a “salmon”
- You classify it as a “sea bass”
- You sell it as a “sea bass”

$\mapsto$ angry customer

Cost Making Errors

- The fish is a “sea bass”
- You classify it as a “salmon”
- You sell it as a “salmon”

$\mapsto$ lost revenue

Loss Function

		Fish is a	
		Sea bass	Salmon
Sold as	Sea bass	0$	2$
	Salmon	1$	0$

Loss Function and Conditional Risk

- True classes $\{\omega_1, \omega_2, \ldots, \omega_c\}$
- Actions taken $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$
- Loss function $\lambda(\alpha_i \mid \omega_j)$
- Conditional risk

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

How to include p(x)
to estimate overall loss/risk?

Overall Risk

- Decision rule: map feature vector to action
  - $x \mapsto \alpha$
- Goal:

  Determine decision rule that minimizes overall risk:

$$R = \int R(\alpha(x) \mid x)\, p(x)\, dx$$

$\mapsto$ to minimize R, pick the action that minimizes the conditional risk for a specific x

Example: two-class problem (1)

- Classes: ω_1, ω_2
- Actions: $\alpha_{1,} \alpha_2$
- For simplicity: loss: $\lambda_{ij}=\lambda(\alpha_i| \omega_j)$
- Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x)$$

Example: two-class problem (2)

- Example actions
  - α_1: decide that the class is ω_1
  - α_2: decide that the class is ω_2
- decide that the class is ω_1 if:

$$R(\alpha_1 \mid x) < R(\alpha_2 \mid x) \quad \Rightarrow$$

$$\lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x) < \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x) \quad \Rightarrow$$

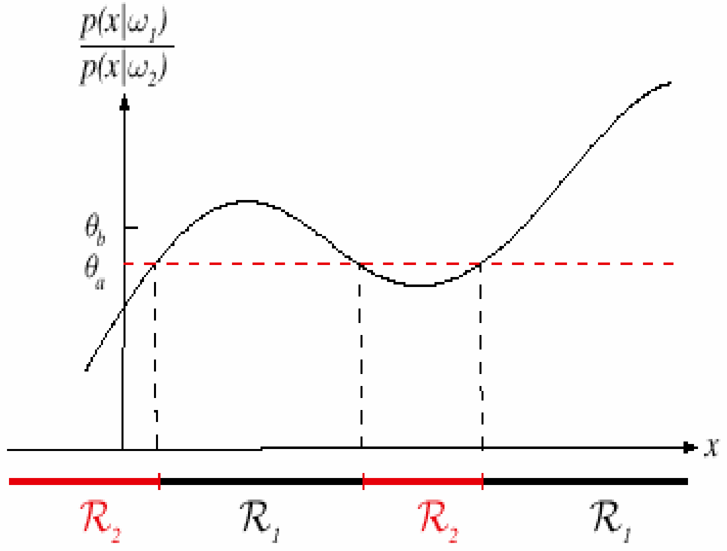
$$(\lambda_{12} - \lambda_{22}) P(\omega_2 \mid x) < (\lambda_{21} - \lambda_{11}) P(\omega_1 \mid x)$$

Replaces Bayes decision rule

Example: two-class problem (3)

- Rephrase:

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



$\mapsto$ tune threshold θ to tune overall risk (loss)

Minimum Error Rate Classification

General case difficult to handle
Important special case: minimze the number of errors

Actions:

α_i: decide that the class is ω_i

"Zero-one-loss"-function

$$\lambda(a_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \ldots, c$$

Conditional Risk for zero-one Loss Function

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

How can you simplify this?

$$= \sum_{j=1, i \neq j}^{c} P(\omega_j \mid x)$$

Def. of zero-one loss function

$$= 1 - P(\omega_i \mid x)$$

Normalization of probability

Minimum Error Rate/ Bayes Decision Rule

- Pick i that minimizes risk:

$$R(\alpha_i \mid x) = 1 - P(\omega_i \mid x)$$

$\mapsto$ pick i that maximizes conditional probability

$$P(\omega_i \mid x)$$

$\mapsto$ Bayes decision rule

Example: two-class problem (3)

- Minimum error rate applied to example
- Action α_1: decide that the class is ω_1
- Take this action if

$$(\lambda_{12} - \lambda_{22})P(\omega_2 \mid x) < (\lambda_{21} - \lambda_{11})P(\omega_1 \mid x) \quad \Rightarrow$$
$$P(\omega_2 \mid x) < P(\omega_1 \mid x)$$

$\mapsto$ Recover Bayes Decision Rule

Summary 5.1. The Bayes Classifier

- Bayes classifier

$$\overline{\omega_i} = \arg\max_{\omega_k} P(x \mid \omega_k) P(\omega_k)$$

- Minimizes number of classification errors
- Generalization: minimize loss (“risk”)

5.2 Normal Distributions

Motivation

- Try to describe probability of (multidimensional) continuous data
- Normal distribution very often found in nature

Some Definitions from Statistics

- Probability density

$$p(x)$$

- Expectation value

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)\,p(x)$$

Basic Expectation Values

- Normalization $$1 = \mathcal{E}[1] = \int_{-\infty}^{\infty} p(x)dx$$
- Mean $$\mu = \mathcal{E}[x] = \int_{-\infty}^{\infty} x\ p(x)dx$$
- Variance $$\sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2\ p(x)dx$$

  (σ is called standard deviation)
- Entropy $$\mathrm{H} = \mathcal{E}[-\ln p(x)] = \int_{-\infty}^{\infty} [-\ln p(x)]p(x)dx = -\int_{-\infty}^{\infty} p(x)\ln p(x)\ dx$$

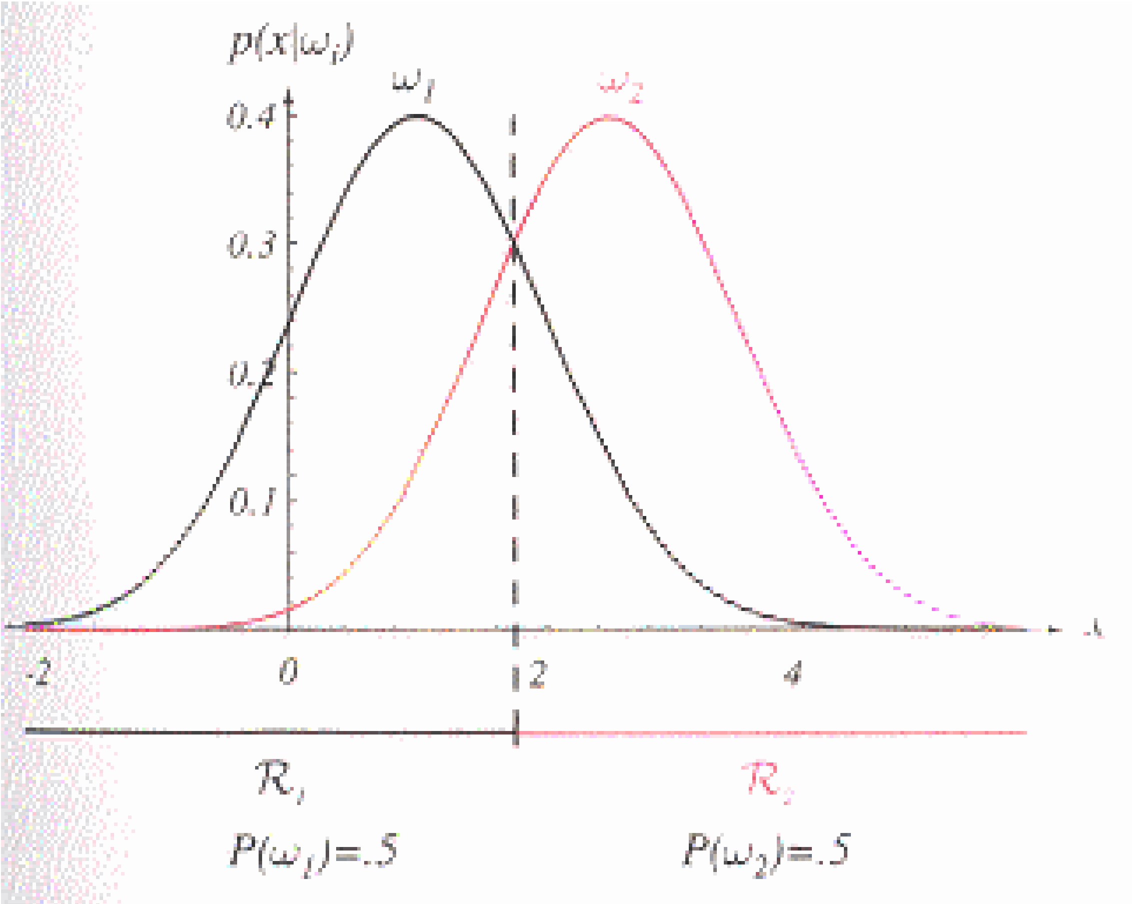
One Dimensional Gaussian Density (Univariate Density)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

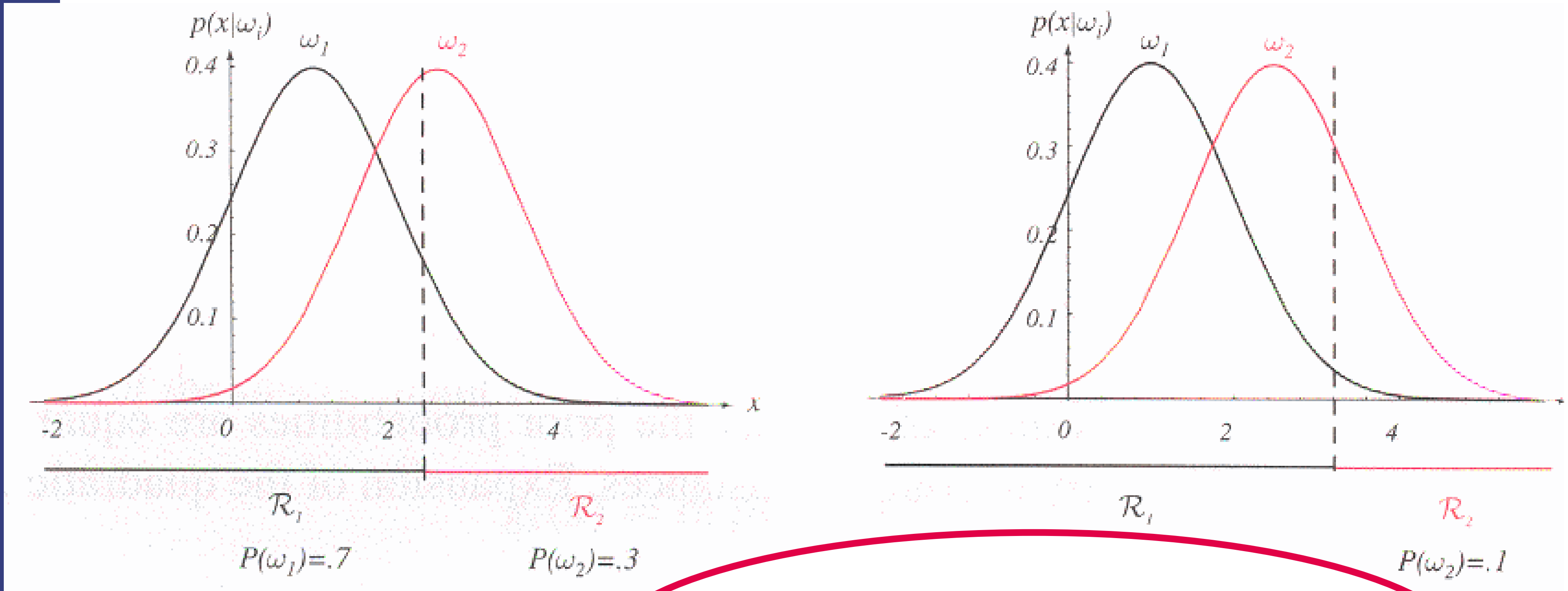
Terminology: “Normal Distribution” is a different term for Gaussian Densities

$\mapsto$ maple

Decision Surface in 1 Dimension (identical variance, prior)

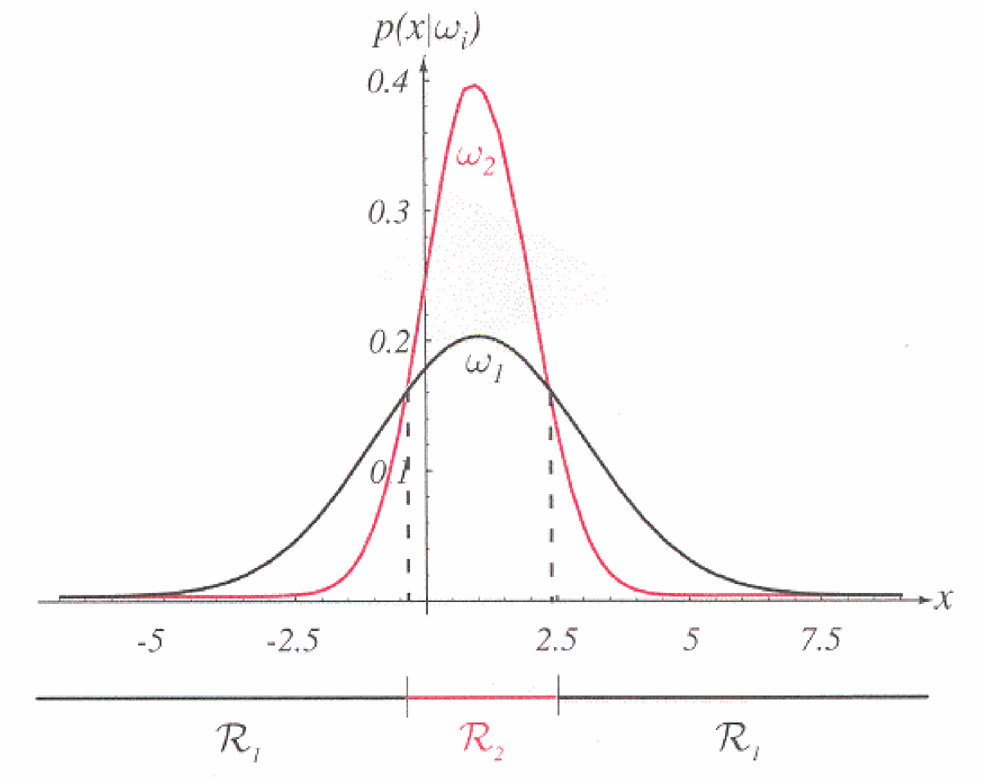


Decision Surface in 1 Dimension: Changing the Prior



Can you think of a case with two decision boundaries?

General Case: 1 Dimension



Excursion: Reminder of Linear Algebra

Vector (column-vector)

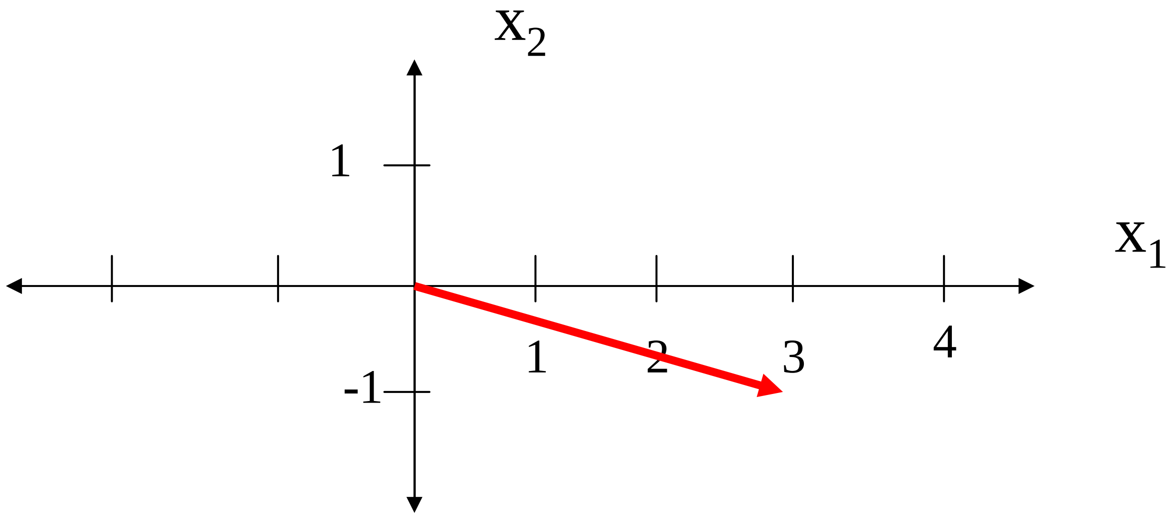
- A vector is a kind of multidimensional arrow
- In general

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{pmatrix}$$

is called an N-dimensional vector

Vector: Example

- Example $\vec{x} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$ is a 2-dimensional vector



Matrix

- A matrix is a rectangular scheme of numbers
- In general

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{M1} & \dots & \dots & a_{MN} \end{pmatrix}$$

is called an MxN matrix

•Very often M=N

Matrix: Example

- A 2x2 matrix

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}$$

Multiplication of a Matrix with a Vector

- Definition:

$$A\vec{x} = \vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}$$

with $$y_i = \sum_{j=1}^{N} a_{ij} x_j$$

Example

Given $A = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}$ $\vec{x} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$

Multiplication

$$\vec{y} = A\vec{x} = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix}\begin{pmatrix} 3 \\ -1 \end{pmatrix} = \begin{pmatrix} 5 \\ 7 \end{pmatrix}$$

Transposed of a vector: row vector

Given a column vector

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \cdots \\ x_N \end{pmatrix}$$

we can create a row vector by transposing the column vector:

$$\vec{x}^{\,t} = \begin{pmatrix} x_1 & x_2 & \ldots & x_N \end{pmatrix}$$

Multiplication of a Row Vector with a Matrix

- Definition:

$$\vec{x}^{\,t} A = \vec{y}^{\,t} = \begin{pmatrix} y_1 & y_2 & \dots & y_N \end{pmatrix}$$

with $$y_i = \sum_{j=1}^{N} x_j a_{ji}$$

Example

Given

$$A = \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} \qquad \vec{x} = \begin{pmatrix} 3 \\ -1 \end{pmatrix}$$

Multiplication

$$\vec{y}^{\,t} = \vec{x}^{\,t} A = \begin{pmatrix} 3 & -1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 1 \end{pmatrix}$$

Inner Product of Vectors

- Definition: $\vec{a}^{\,t}\vec{b} = \sum_{i=1}^{N} a_i b_i$

•Example:

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{pmatrix} \qquad \vec{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_N \end{pmatrix}$$

Hence: $(\vec{x}-\vec{\mu})^t(\vec{x}-\vec{\mu}) = \sum_{i=1}^{N}(x_i-\mu_i)^2$ (Euclidian Distance)

Multiplication of two Matrices

$$C = AB = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{N1} & \dots & \dots & a_{NN} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1N} \\ b_{21} & b_{22} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ b_{N1} & \dots & \dots & b_{NN} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1N} \\ c_{21} & c_{22} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ c_{N1} & \dots & \dots & c_{NN} \end{pmatrix}$$

with $$c_{ij} = \sum_{k=1}^{N} a_{ik} b_{ki}$$

Note: in general $AB \neq BA$

Unit Matrix/Inverse of a Matrix/Determinant

Unit Matrix

$$1 = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & 0 \\ 0 & \ldots & 0 & 1 \end{pmatrix}$$

Inverse: matrix A^{-1} that satisfies

$$AA^{-1} = 1$$

Determinant |A|: kind of an absolute value for matrices

More Examples

-> maple

Multi Dimensional Gaussian Density (Multivariate Density)

$\mapsto$ blackboard

5.3.1 Discriminant Functions for Normal Distributions

Decision Boundaries in 2d

-> maple script

Discriminant Functions for the Normal Density

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

$$g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$$

- Case of multivariate normal

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

Decision Surface in 2+3 Dimensions for identical and uniform variance ($\Sigma_i = \sigma^2 I$)

Can drop covariance term from all discriminant functions

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

where :

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

(ω_{i0} is called the threshold for the ith category!)

Terminology "linear machine"

- A classifier that uses linear discriminant functions is called "a linear machine"

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

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

Decision Boundary

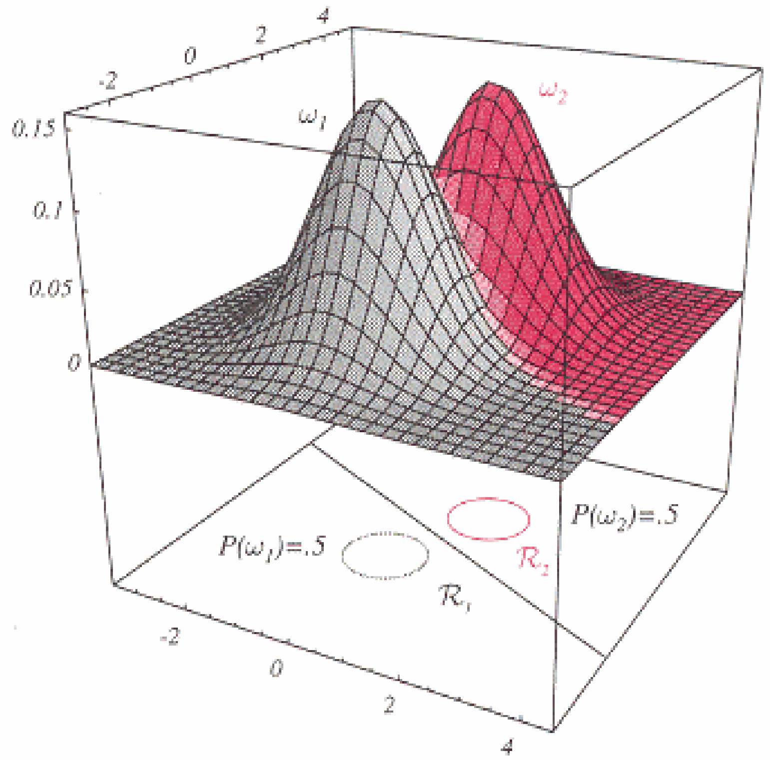
- The hyperplane separating R_i and R_j

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

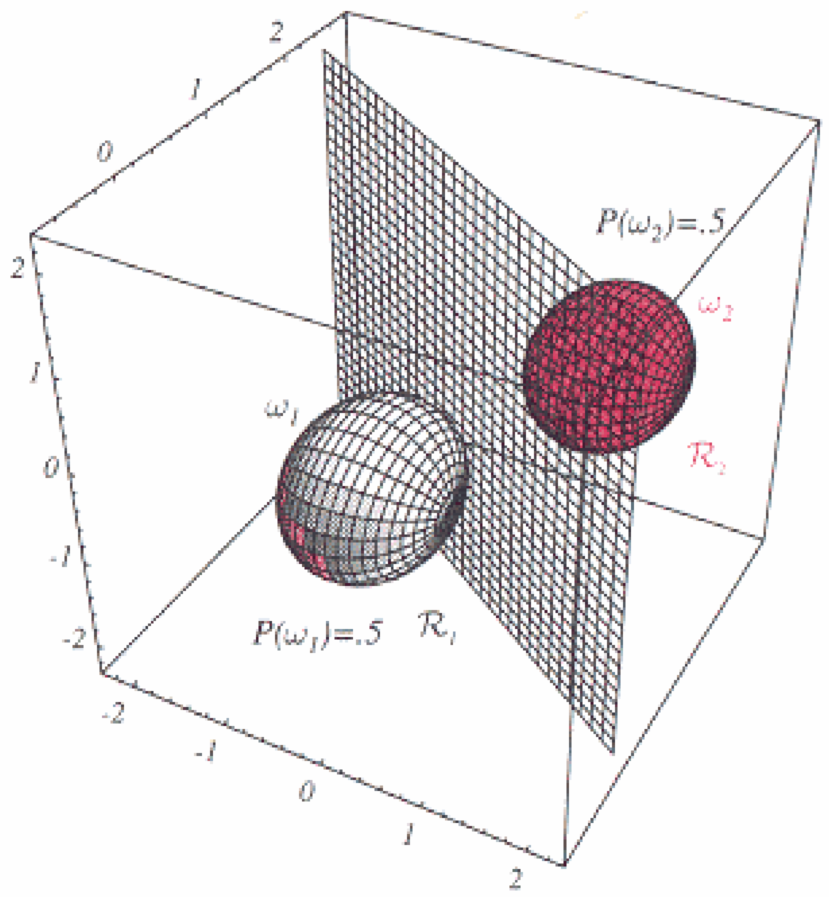
always orthogonal to the line linking the means!

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

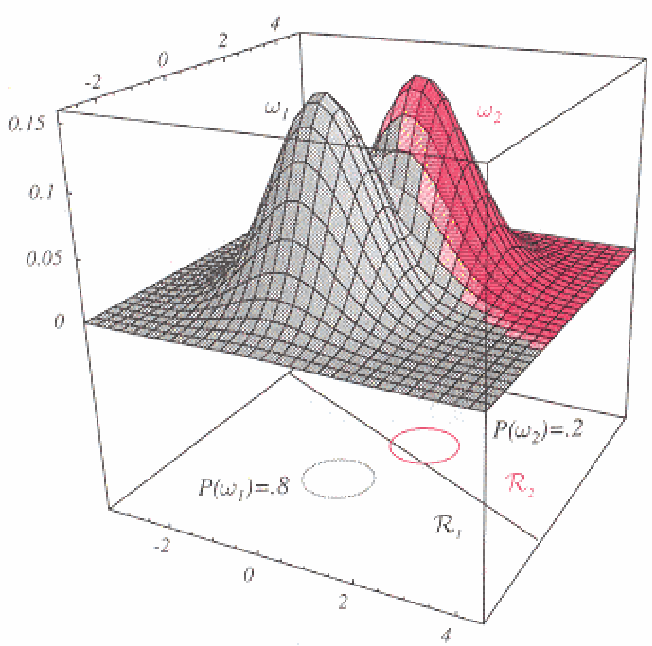
Decision Surface in 2 Dimensions (identical and uniform variance, identical prior)

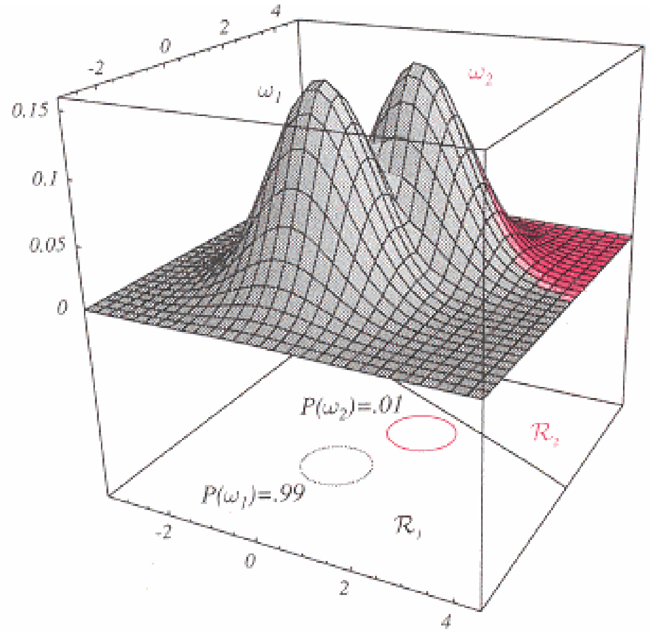


Decision Surface in 3 Dimensions (identical and uniform variance, identical prior)

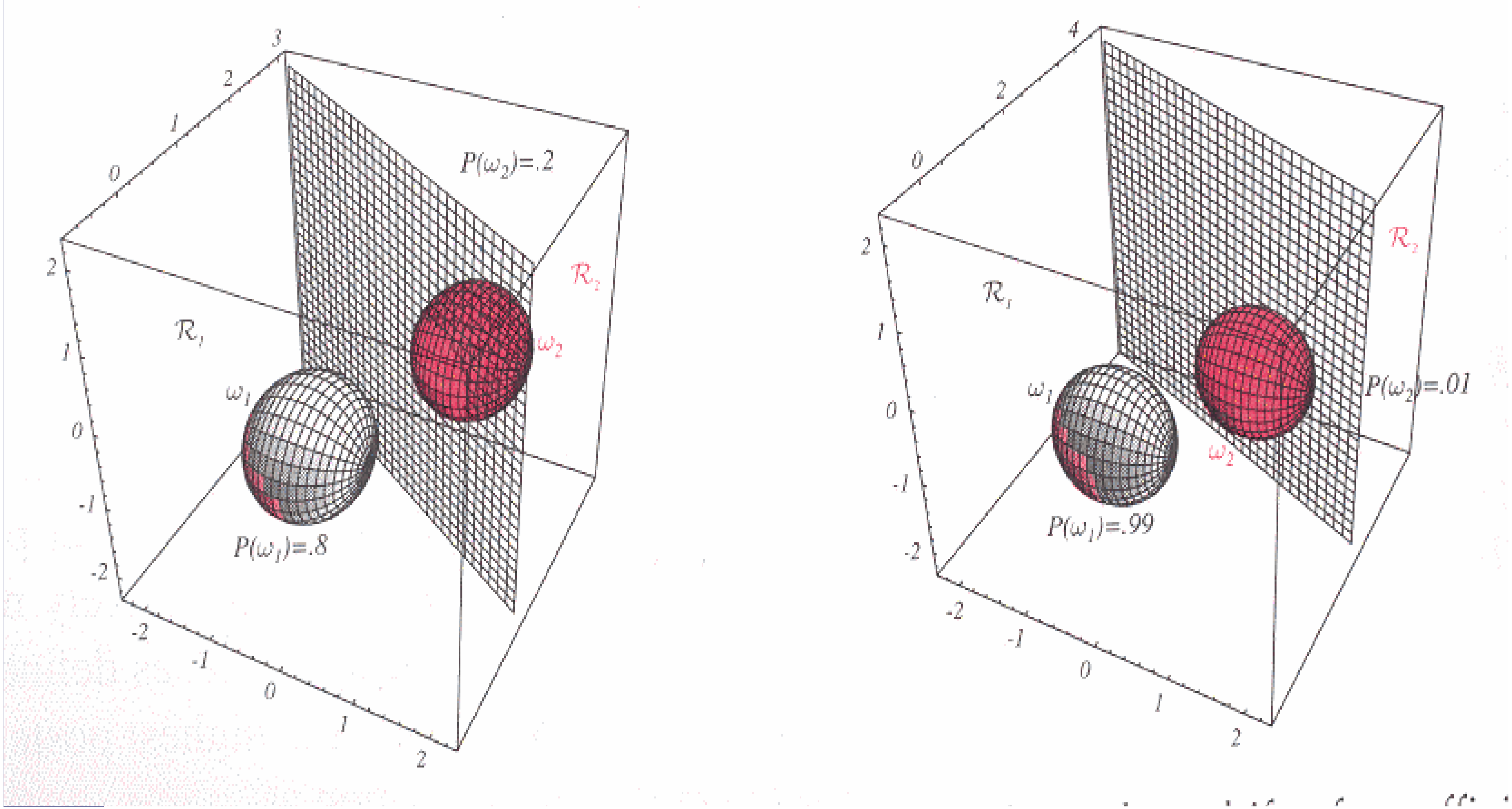


Decision Surface in 2 Dimensions: Changing the Prior





Decision Surface in 3 Dimensions: Changing the Prior



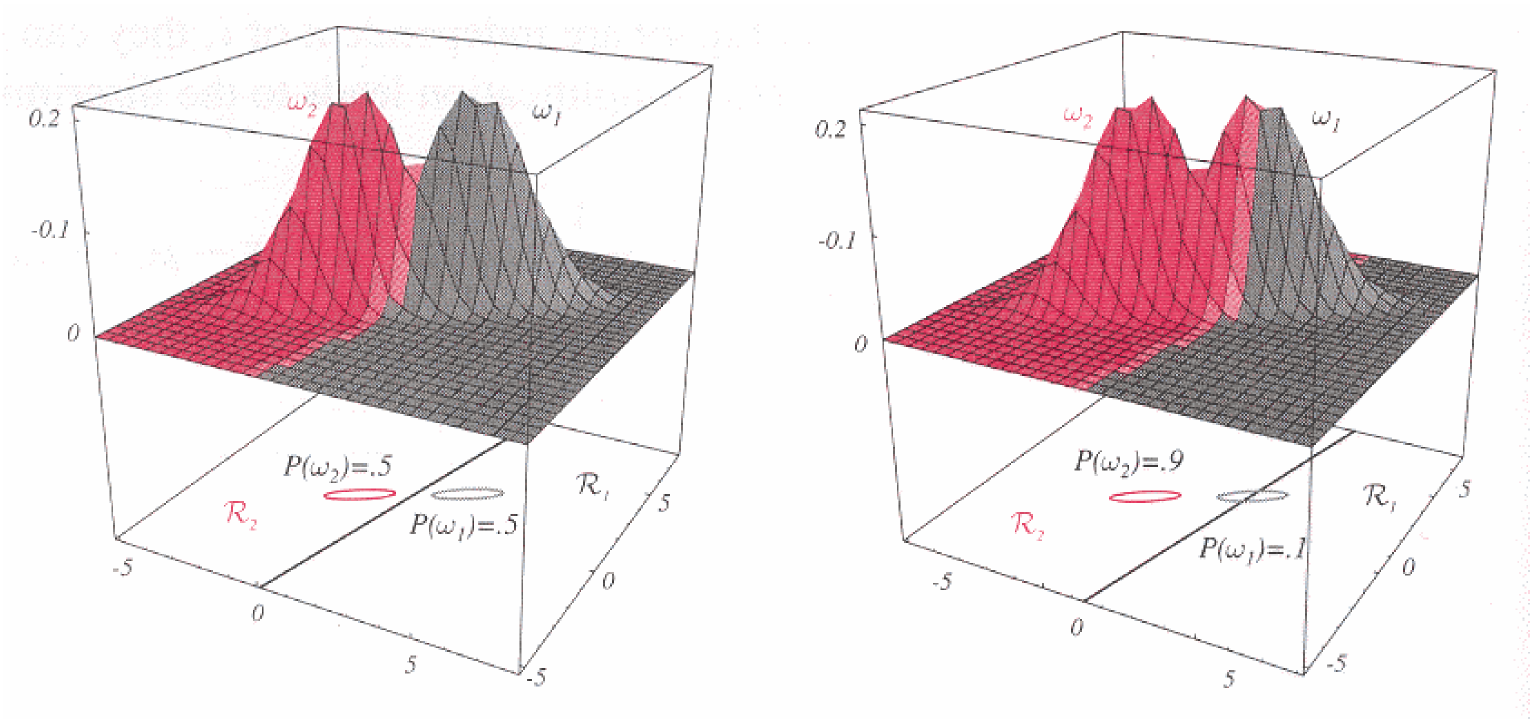
Covariance of all classes are identical but arbitrary! ($\Sigma_i = \Sigma$)

- Hyperplane separating R_i and R_j

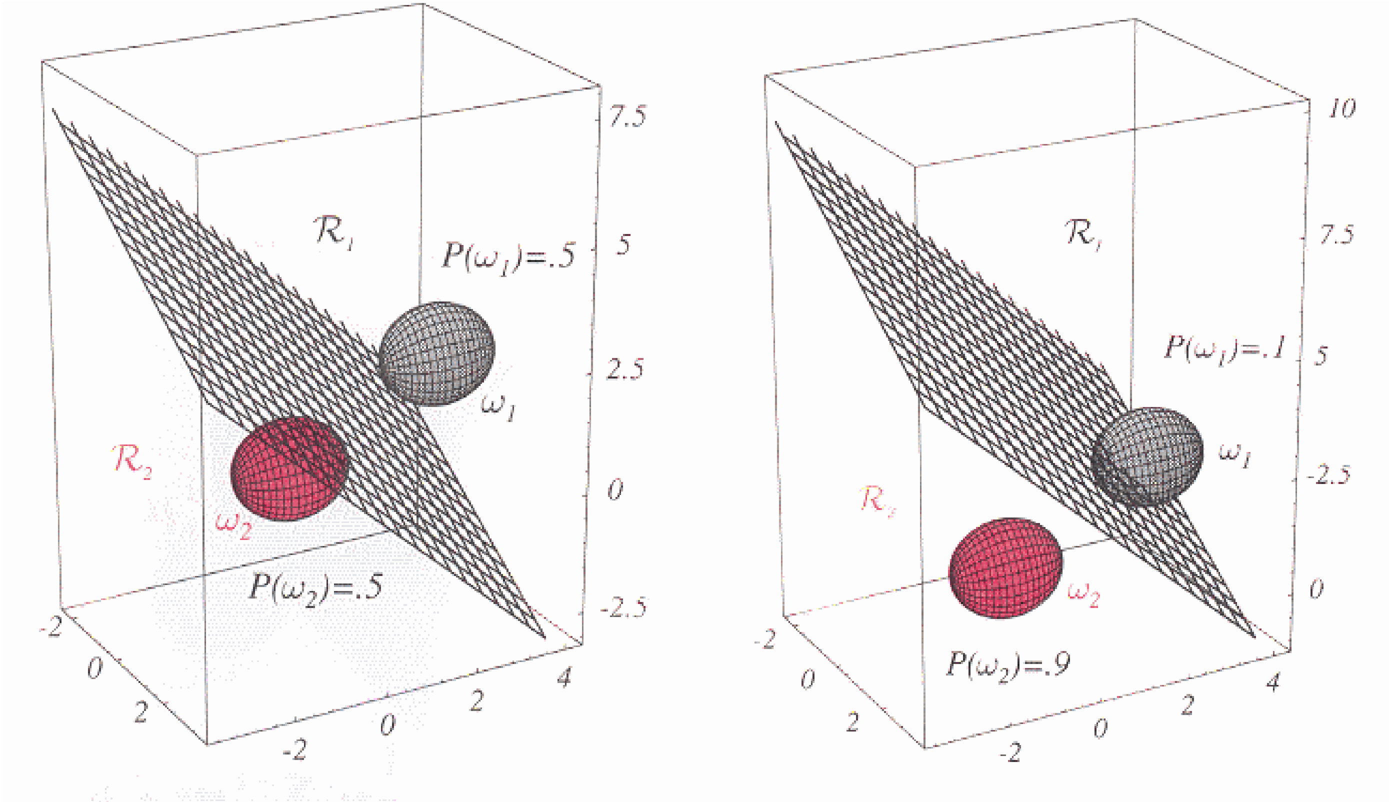
$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln\left[P(\omega_i)/P(\omega_j)\right]}{(\mu_i - \mu_j)^t \Sigma^{-1}(\mu_i - \mu_j)}.(\mu_i - \mu_j)$$

(the hyperplane separating R_i and R_j is generally not orthogonal to the line between the means!)

Arbitrary Variance; Identical for all Gaussians



Arbitrary Variance; Identical for all Gaussians



Covariance matrices are different for each category (Σ_i = arbitrary)

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

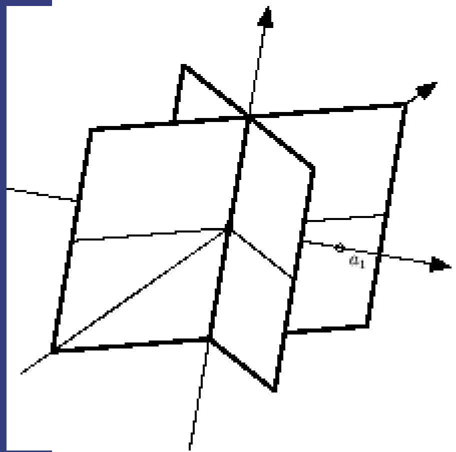
where :

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

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

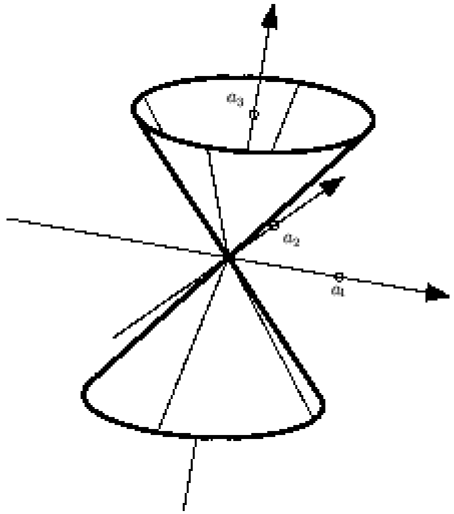
$$w_{i0} = -\frac{1}{2}\mu_i^t \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

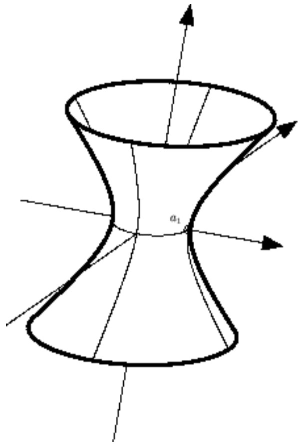
Quadrics in 3 Dimensions



Intersecting planes

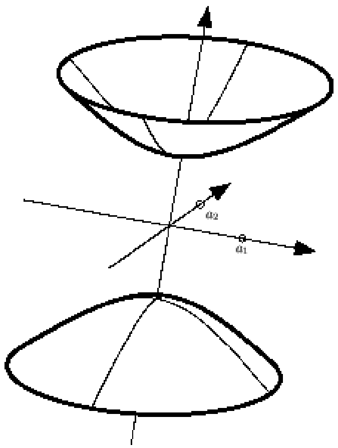
Cones



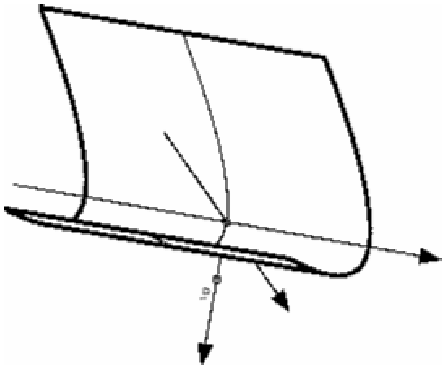


hyperboloid of 1 sheet

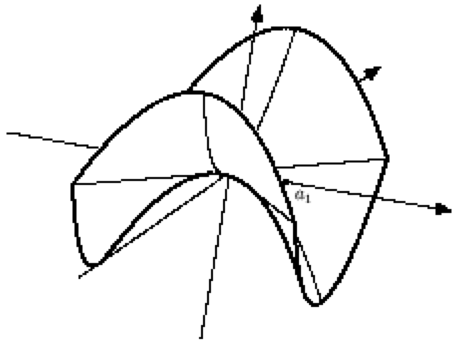
hyperboloid of 2 sheets



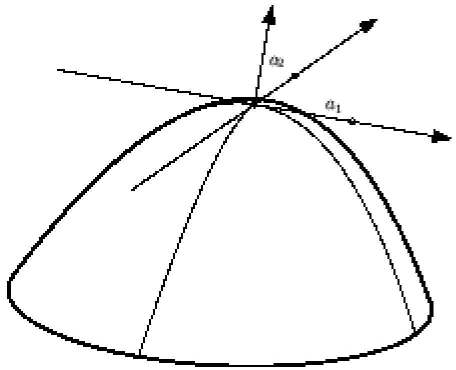
Quadrics in 3 Dimensions



parabolic cylinder

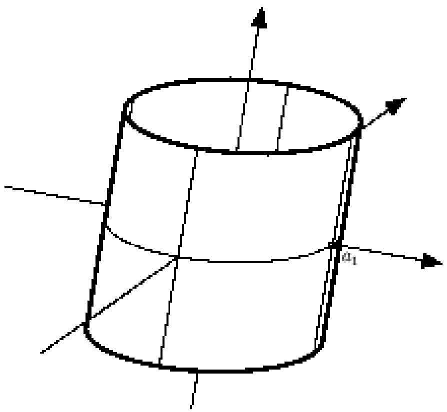


hyperbolic paraboloid

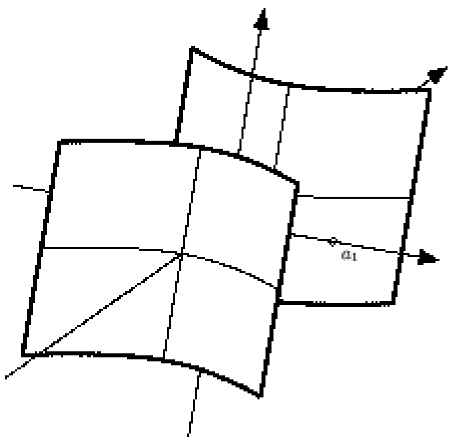


elliptic paraboloid

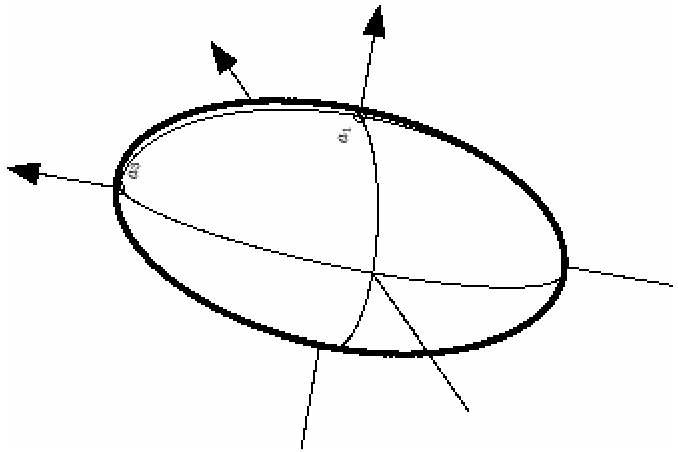
elliptic cylinder



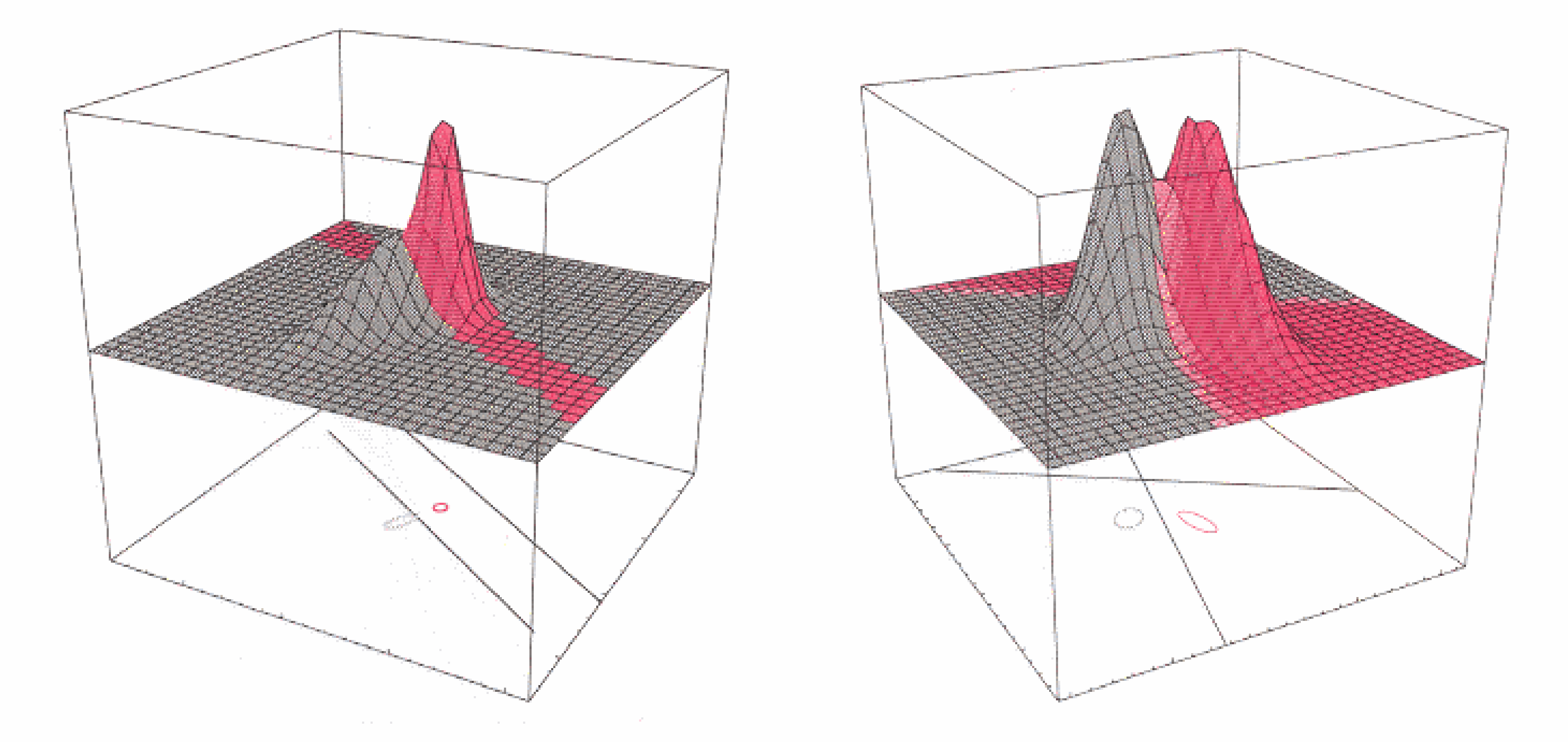
hyperbolic cylinder



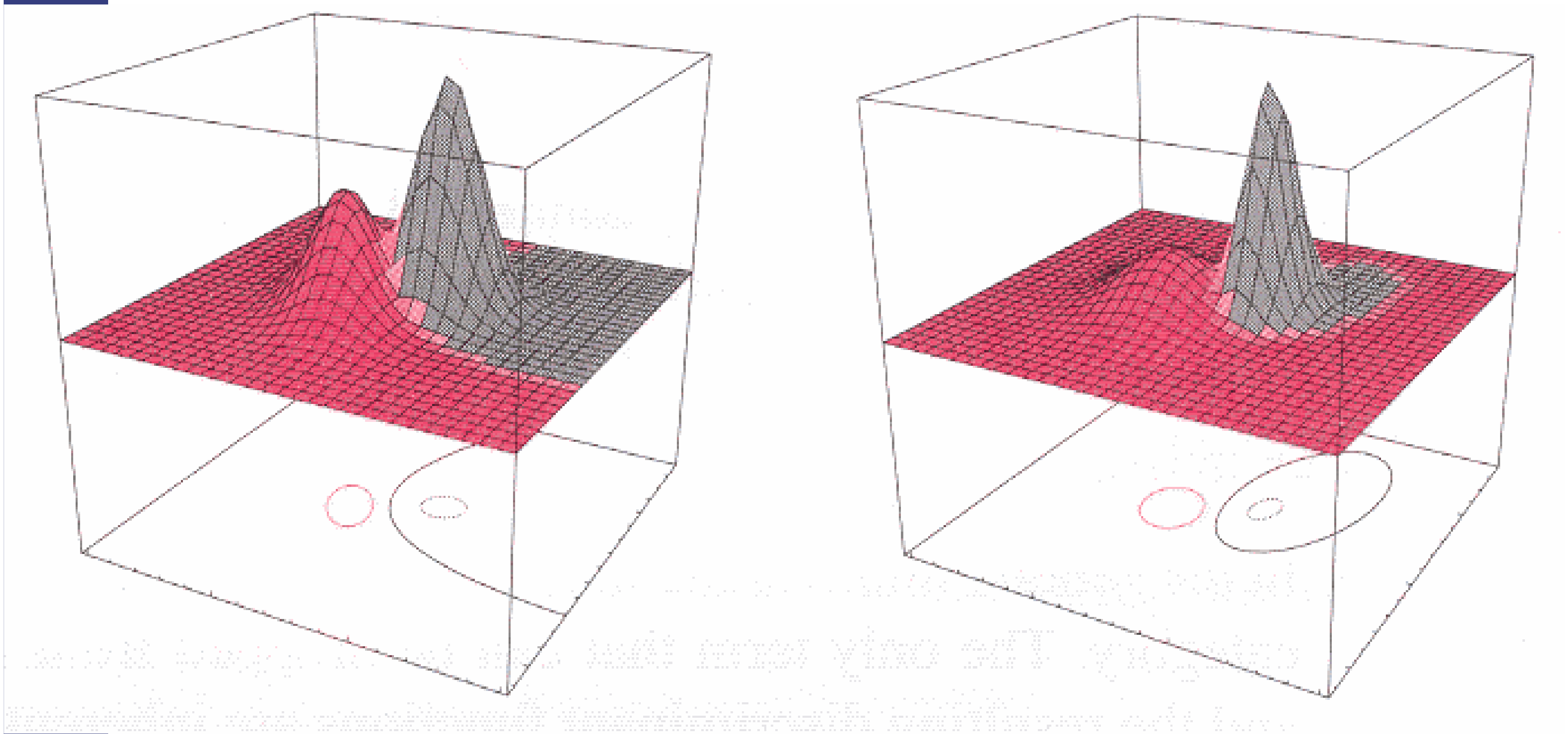
ellipsoid



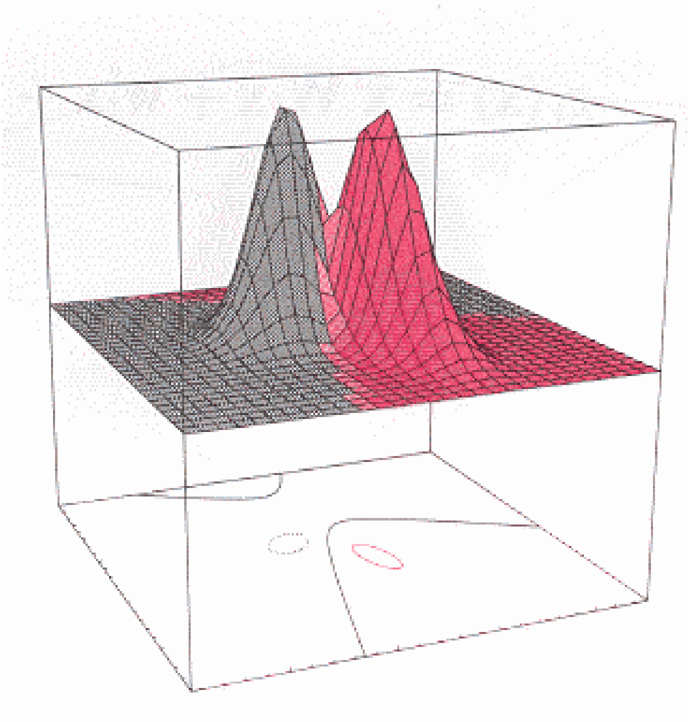
General Case: 2 Dimensions

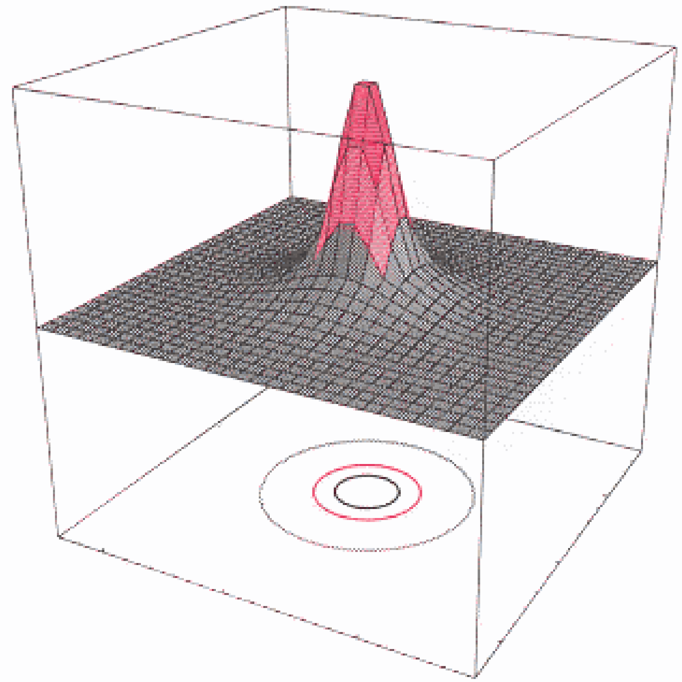


General Case: 2 Dimensions

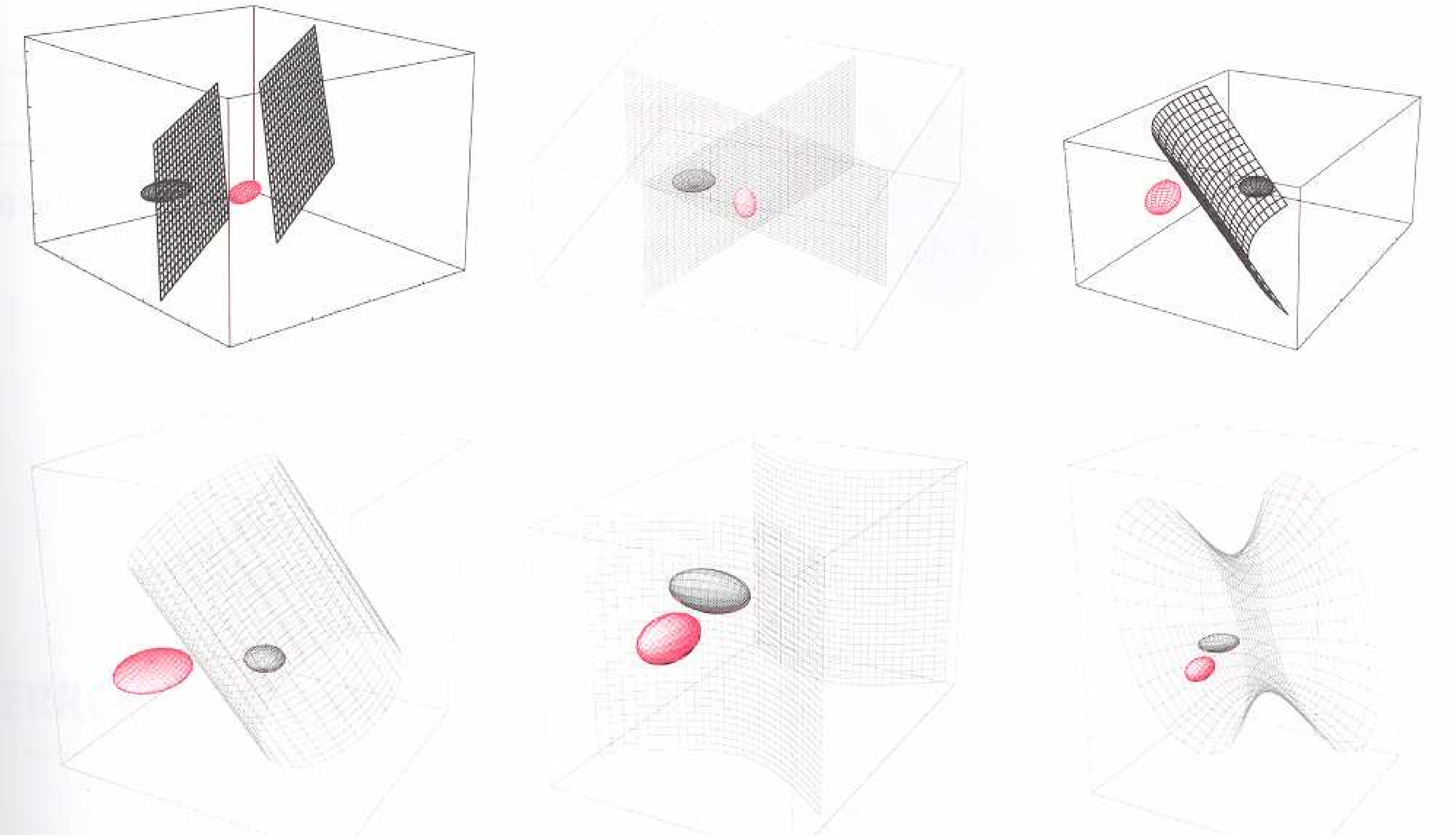


General Case: 2 Dimensions

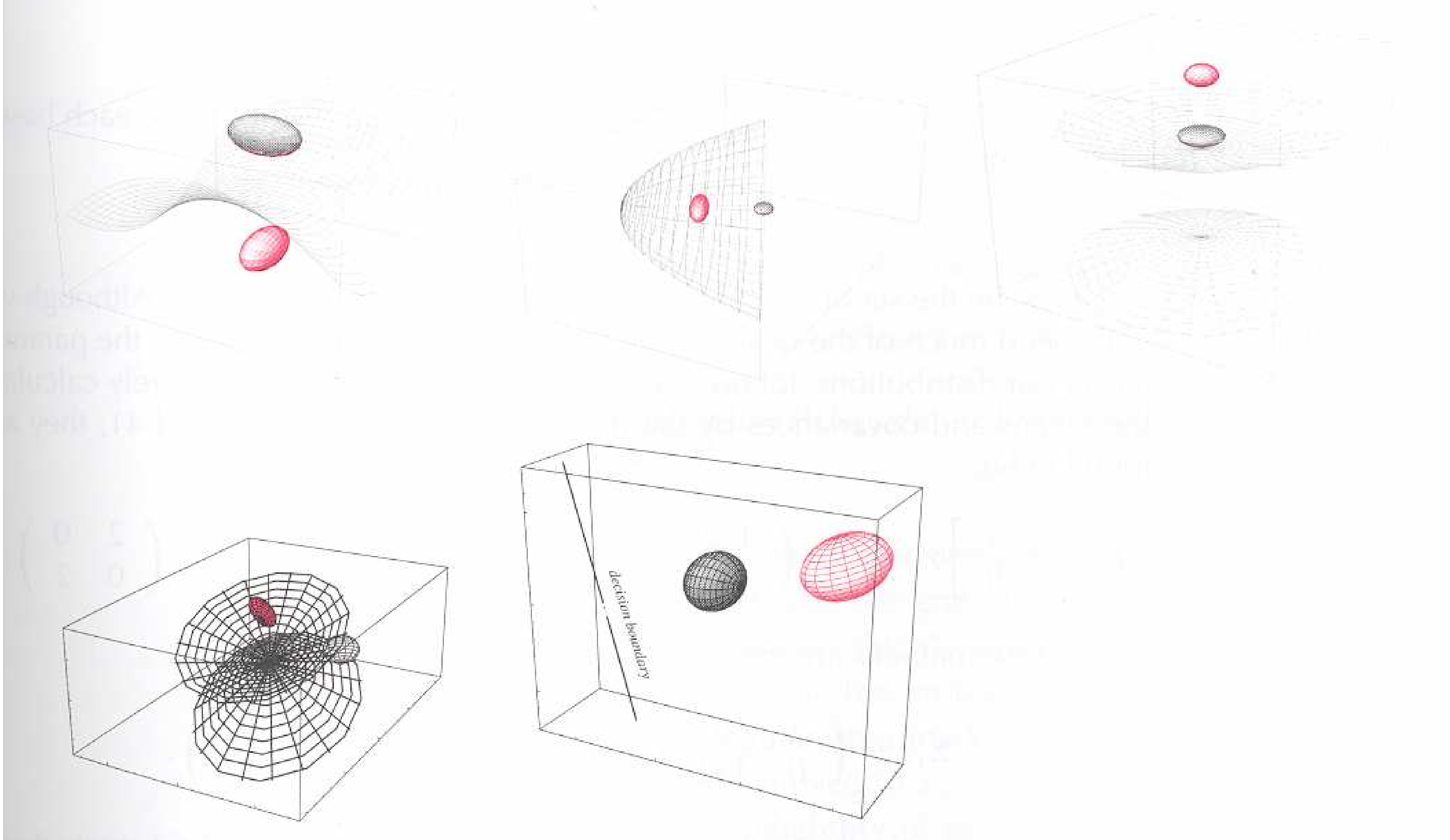




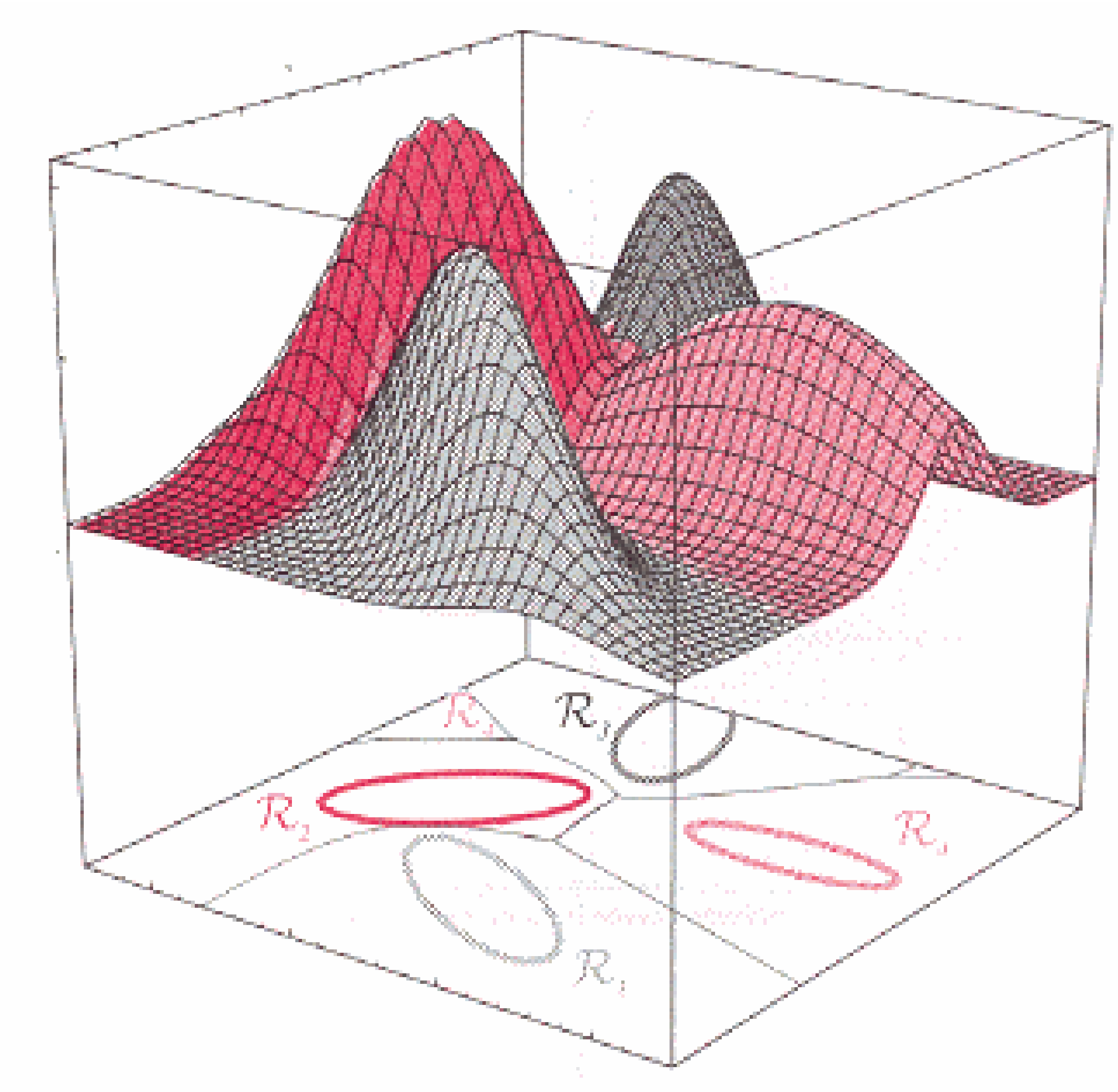
General Case: 3 Dimensions



General Case: 3 Dimensions



General Case: 2 Dimensions; many Classes



Summary

- Decision boundaries for normal distributions:
  - Lines
  - Planes
  - Other quadrics