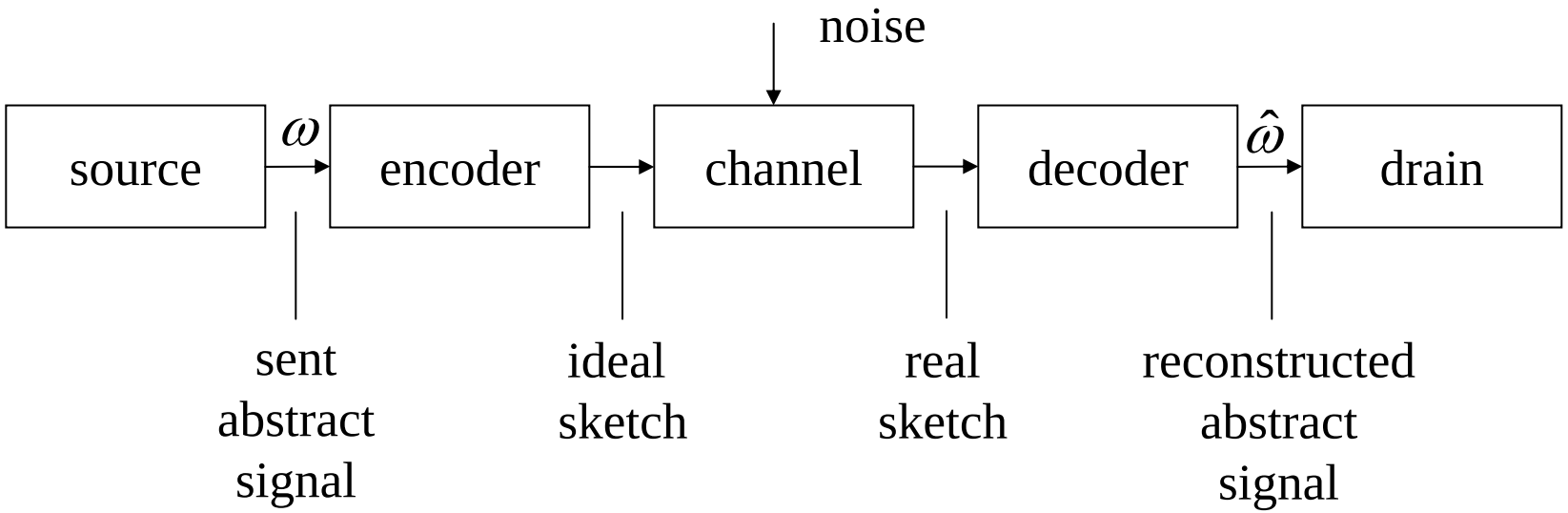
# Chapter 7

## Bayes or optimal classifier

# Designing an optimal classifier

- Last link in the pattern recognition chain
- The classifier has to assign a feature vector to a category in an optimal way
- The design is based on statistic decision theory
- Description of the recognition system analogous to a news transmission system:



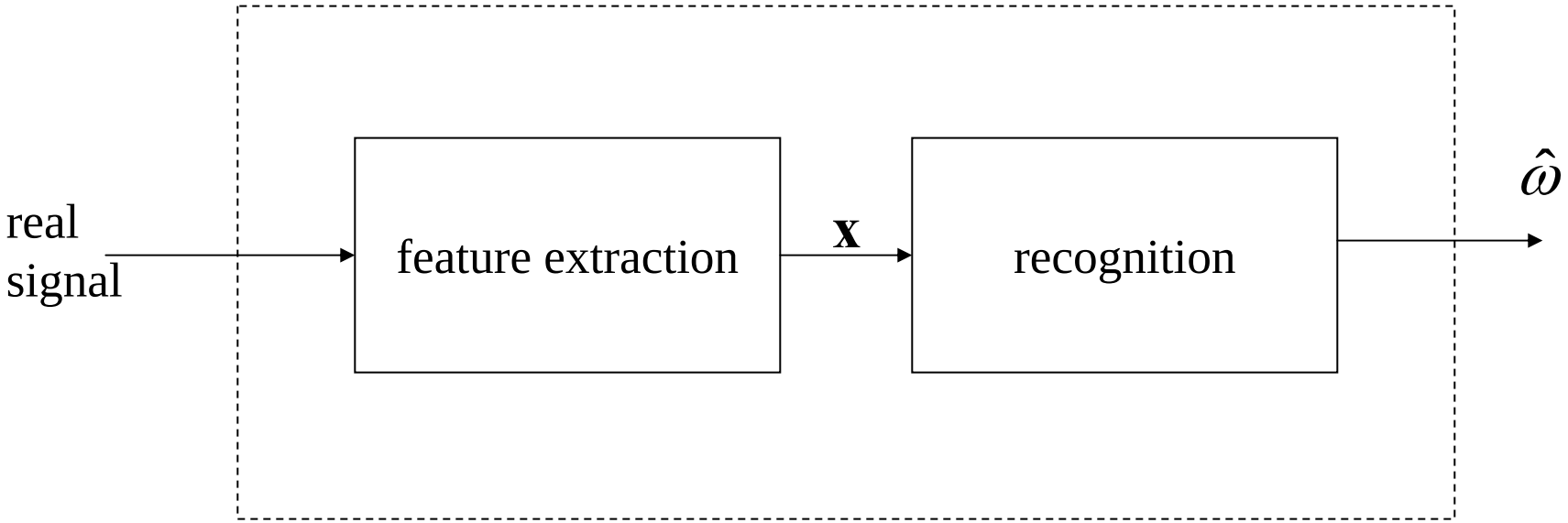
# About the single components

Coder: from an abstract source signal *K* results a character: e.g. OCR-B

Noise: includes all variations like e.g. actual printing or writing, possible fouling, scanner error etc.

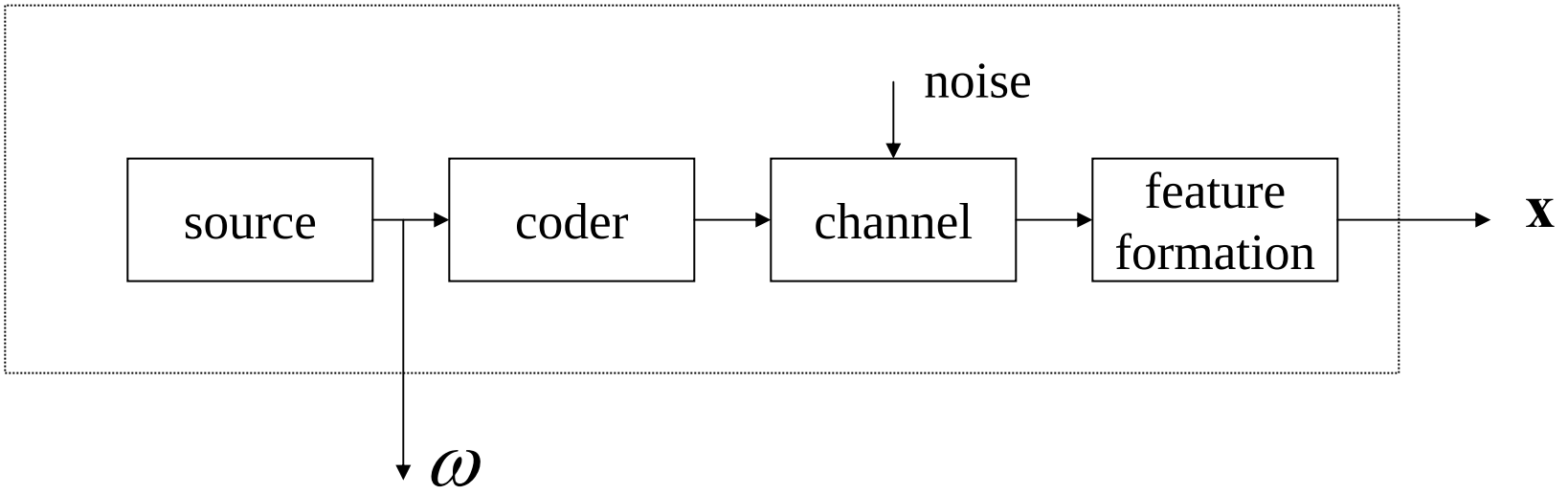
Decoder: reading machine, reconstructs the signal sent, passes decision to drain

# Structure of a decoder



# Stochastic model

The process that generates signals produces pairs of varietes, that are connected to each other (ω,**x**)




The models‘ statistic properties are described completely by the **joint probabilty distribution**:

$$p(\omega, \mathbf{x}) = p(\mathbf{x}, \omega) \qquad \omega \in \{\omega_i\} \quad i = 1, 2, \ldots K$$

ω discret, x continuous

# Optimality criterion

Sought is a classifier, that classifies in the “best possible way” according to a given performance index (optimal classifier)

Choosing ***minimization of wrong classifications*** for plenty of tests, results in a classifier, that maximizes the A-posteriori-probablity (maximum-A-Posteriori-classifier):

$$\max_{K}\{P(\omega_k \mid \mathbf{x})\} \qquad \text{MAP or Bayes classifier}$$

The optimal decision is based on the **a-posteriori-** or ***inference probability*** $P(\omega_k|\mathbf{x})$, which is the ***conditional probability***, for the observed value $\mathbf{x}$ to originate in $\omega_k$ .

Using Bayes‘ theorem this can be transformed as follows:

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

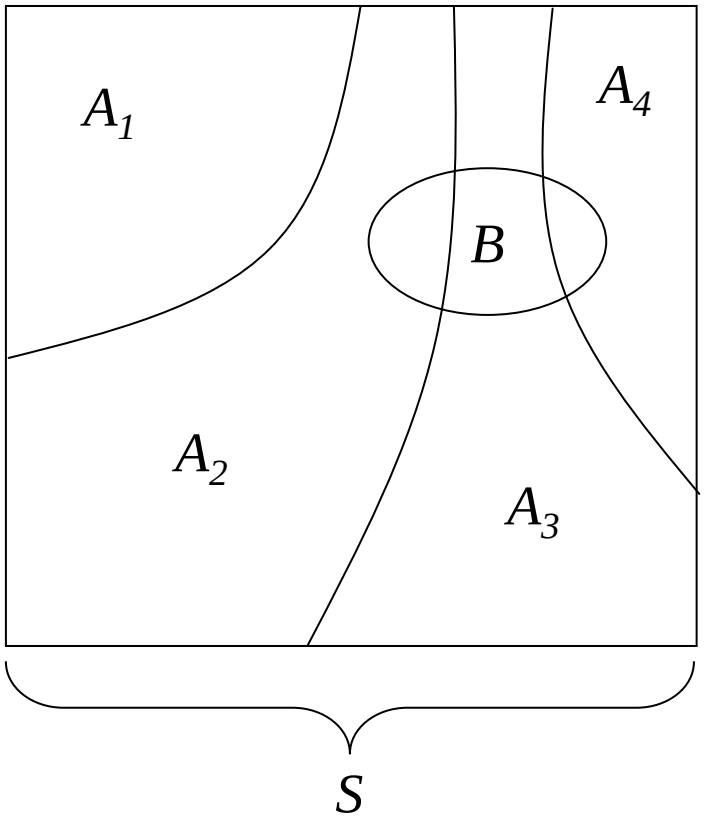
with marginal distribution:

$$p(\mathbf{x}) = \sum_K p(\mathbf{x}, \omega_k) = \sum_K p(\mathbf{x} \mid \omega_k) P(\omega_k)$$

# Bayes‘ theorem

Let $A_i$ be disjoint (that exclude each other) events

$$S = \bigcup_{i=1}^{n} A_i \quad \text{sample space}$$




Let $B$ be an arbitrary event. Then:

$$P(A_i \mid B) = \frac{P(A_i, B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j)P(A_j)}$$

$$\Rightarrow P(A_i \mid B)P(B) = P(A_i, B) = P(B \mid A_i)P(A_i)$$

# Example

The application of the theorem is demonstrated with an example. The event “A cars’ tires squeak” B occurs with probability P(B)=0,05 ; the hyothesis A ,“The cars’ tires are poorly adjusted.” with probablity P(A)=0,02.

Furthermore we suppose, that poorly adjusted tires sometimes, not always, cause the tires to squeak. The conditional probability for that is P(B|A)=0,7. In case we observe squeaking tires, we can calculate the probablity of poorly adjusted tires using Bayes’ theorem:

$$P(A \mid B) = \frac{P(B \mid A) * P(A)}{P(B)} = \frac{0,7 * 0,02}{0,05} = 0,28$$

Since we observed event B, the probability of hypothesis A increased from 0,02 to 0,28. Calculation of P(A|B) based on P(A) can be viewed as a re-evaluation of hypothesis A in case event B occurs.
This makes the theorem very valuable. It can be used to calculate the propagation of incertainty. Its disadvantage lies in the great amount of data, because probability and conditional probabiliy have to be stored for every single event and every single hypothesis. Also, that data is difficult to obtain and mostly even cannot be obtained with mathmatical precision. [Gottlob1990]

# Optimal classifier

$$\max_k \{P(\omega_k \mid \mathbf{x})\}$$

$$\sim \max_k \{p(\mathbf{x} \mid \omega_k) P(\omega_k)\}$$

Bayes or MAP-classifier

The MAP criterion can be attributed to class-specific distribution densities, which can be measured.

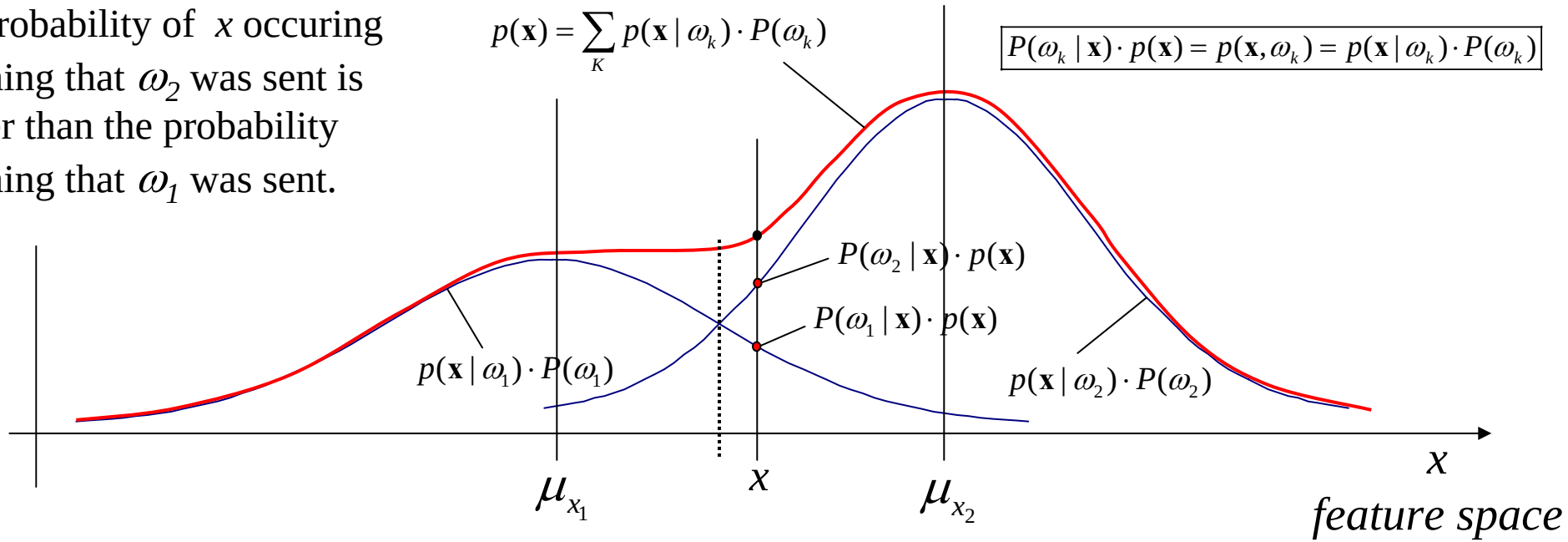
$$\max_K \{p(\mathbf{x} \mid \omega_k)\}$$

Maximum-Likelihood classifier

(p(x|ωk) cond. probabilty of x wrt. $\omega_k$)

# Two-class problem with Gaussian distribution densities and a scalar feature $x$

The probability of $x$ occuring assuming that $\omega_2$ was sent is greater than the probability assuming that $\omega_1$ was sent.




$$p(\mathbf{x} \mid \omega_i) P(\omega_i) \gtrless^{?} p(\mathbf{x} \mid \omega_j) P(\omega_j)$$

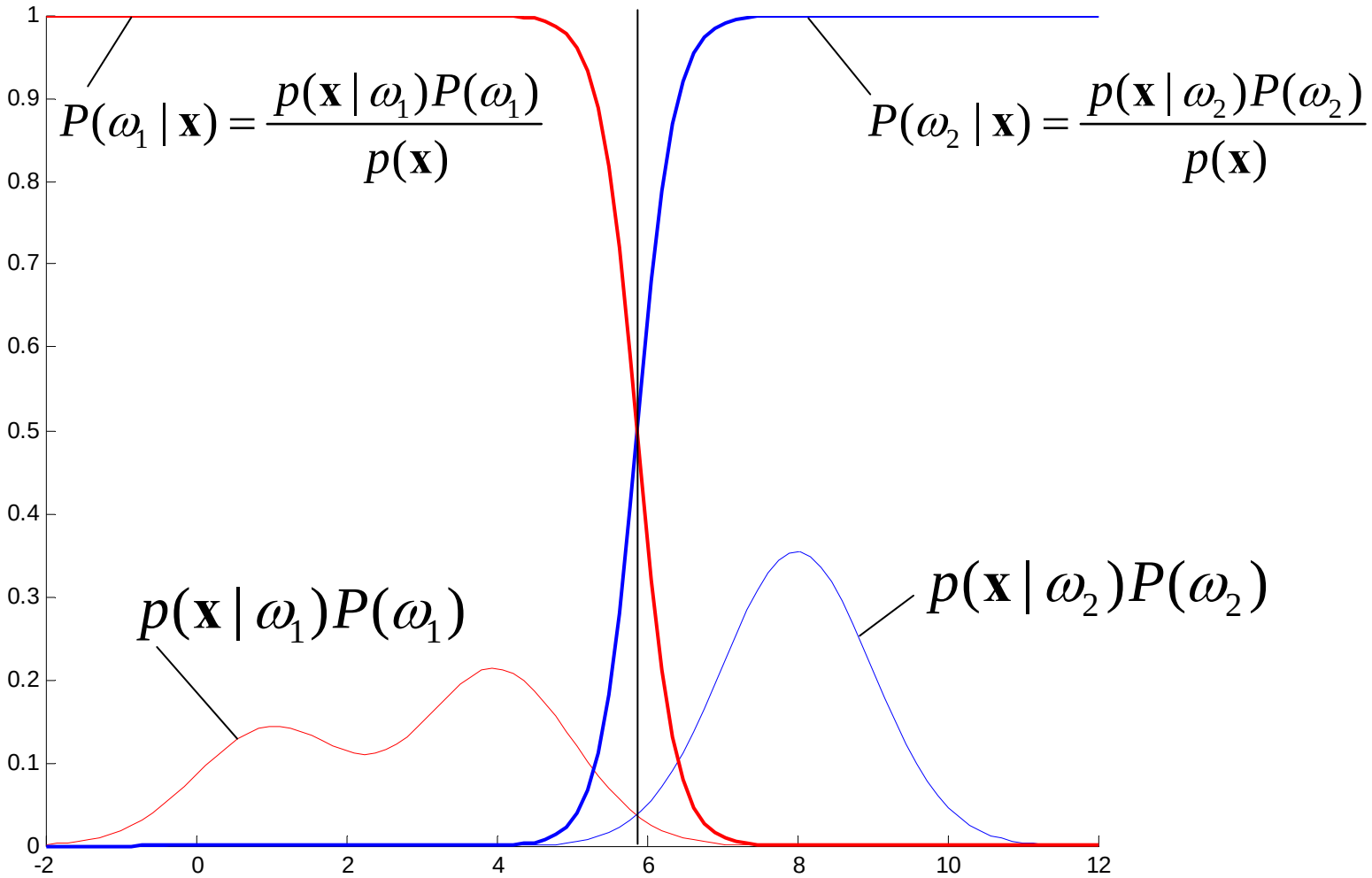
$p(\mathbf{x}|\omega_k)$ *class specific distribution density* for feature vectors $\mathbf{x}$, to be assigned to class $k$ .

A-priori probability for frequency of source signals: $P(\omega_k)$
(source statistics, prob. of occurence for events $\omega_k$, e.g. letters of a language)

Probability for x being measured results from overlapping effects of $\omega_k$ being sent:

$$p(\mathbf{x}) = p(\mathbf{x} \mid \omega_1) P(\omega_1) + p(\mathbf{x} \mid \omega_2) P(\omega_2) + \cdots = \sum_K p(\mathbf{x} \mid \omega_k) P(\omega_k)$$

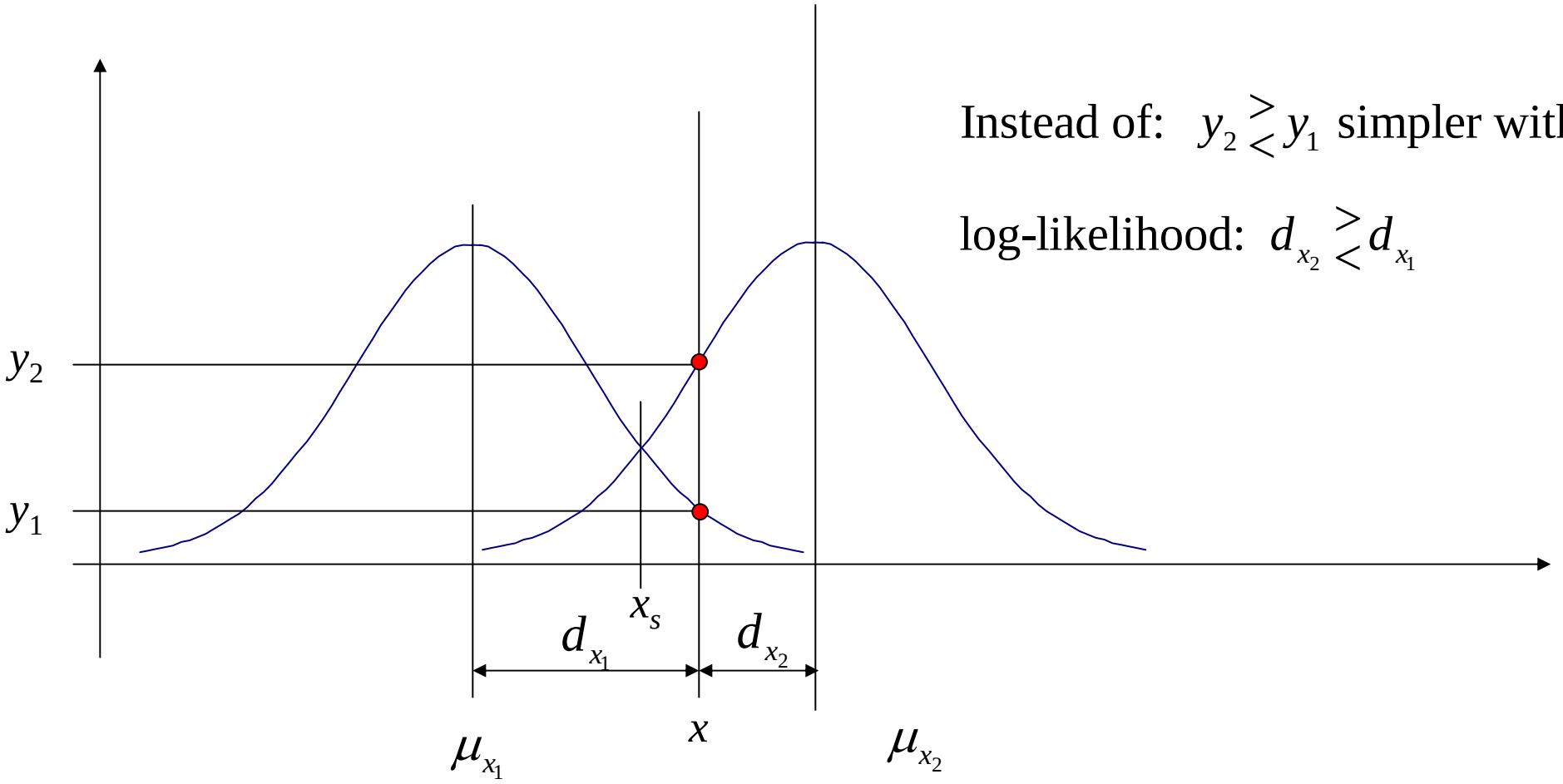
# MAP decision




$$p(\mathbf{x}) = \sum_K p(\mathbf{x}, \omega_k) = \sum_K p(\mathbf{x} \mid \omega_k)P(\omega_k)$$

$$p(\mathbf{x} \mid \omega_1)P(\omega_1) = 0,2\,p(x-1) + 0,3\,p(x-4)$$
$$p(\mathbf{x} \mid \omega_2)P(\omega_2) = 0,5\,p(x-8)$$

# Decision using Log-Likelihood

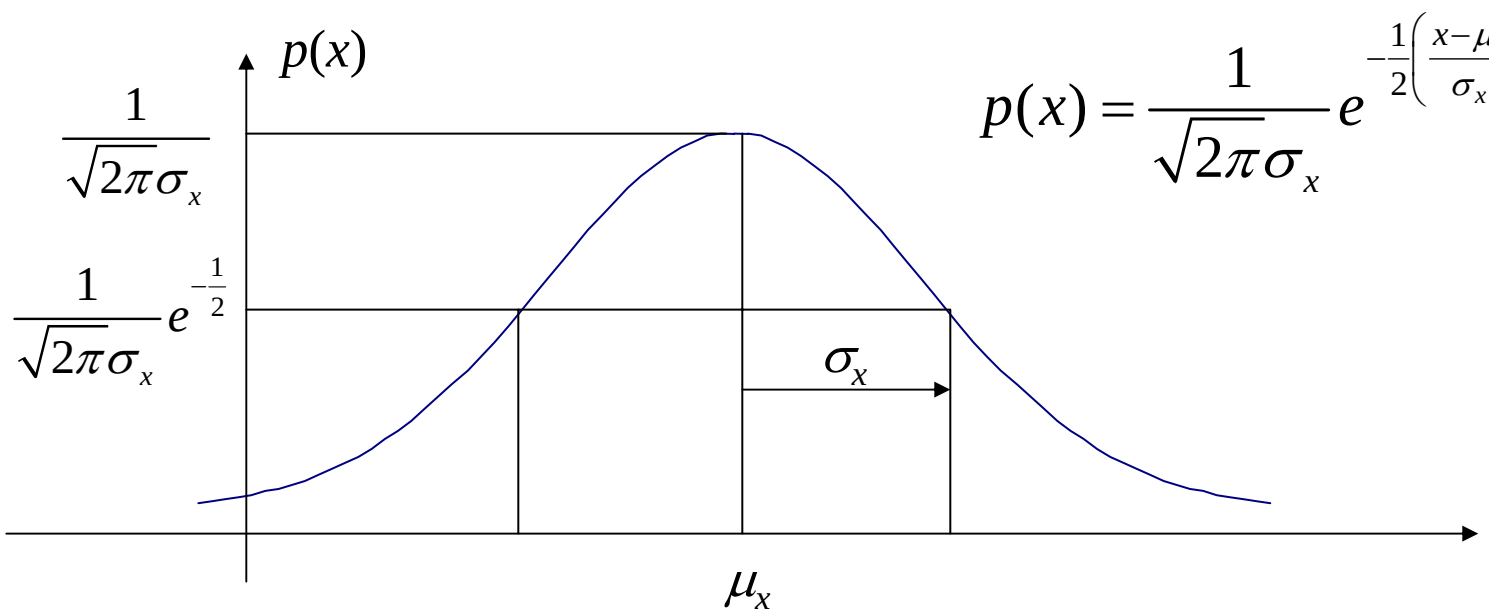



Instead of: $y_2 \gtrless y_1$ simpler with

log-likelihood: $d_{x_2} \gtrless d_{x_1}$

# Normally distributed class specific features $p(\mathbf{x}|\omega_k)$

One-dimensional case:




Expected value of $x$: $\mu_x = E\{x\} = \int_{x=-\infty}^{x=+\infty} x \cdot p(x)dx$

Variance: $\mathrm{var}(x) = \sigma_x^2 = E\{(x-\mu_x)^2\} = \int_{x=-\infty}^{x=+\infty} (x-\mu_x)^2 \cdot p(x)dx$

Standard deviation: $\sigma_x = \sqrt{\mathrm{var}(x)}$

# *N*-dimensional normal distribution

Expected value: $\mu_x = E(\mathbf{x})$ (vector)

Instead of $\sigma^2$ auto-covariance matrix:

$$\mathbf{K} = \mathbf{C}_{\mathbf{xx}} = E\{(\mathbf{x} - \overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}})^T\} = \mathbf{R}_{\mathbf{xx}} - \overline{\mathbf{x}}\overline{\mathbf{x}}^T$$

N-dimensional normal distribution: $p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{K})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})^T \mathbf{K}^{-1} (\mathbf{x}-\boldsymbol{\mu}_\mathbf{x})}$

$$\mathbf{K} = \begin{bmatrix} K_{1,1} & K_{1,2} & \cdots & K_{1,N} \\ K_{2,1} & K_{2,2} & \cdots & K_{2,N} \\ \vdots & \vdots & \vdots & \vdots \\ K_{N,1} & K_{N,2} & \cdots & K_{N,N} \end{bmatrix}$$

$$K_{m,n} = E\{(x_m - \mu_{x_m})(x_n - \mu_{x_n})\}$$

$$K_{n,n} = E\{(x_n - \mu_{x_n})^2\}$$

$\mathbf{K}$: a) symmetrical

b) positive semidefinite

# *N*-dimensional normal distribution

From the positive semi-definiteness $\mathbf{a}^T\mathbf{K}\mathbf{a} \geq 0$ for arbitrary $\mathbf{a} \neq 0$ follows:

If one or more components are linear combinations of other components, **K** is semi-definite, otherwise positive definite (which we assume generally).

If **K** pos. definite, then also $\mathbf{K}^{-1}$ => $\det(\mathbf{K})>0$ and $\det(\mathbf{K}^{-1})>0$ .

Loci of constant probability densities:

$$Q = (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^T \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}) = const.$$

In general, this quadratic form results in conic sections and for positive definite $\mathbf{K}^{-1}$ result *N*-dimensional ellipsoids.

# $N$=2: ellipses

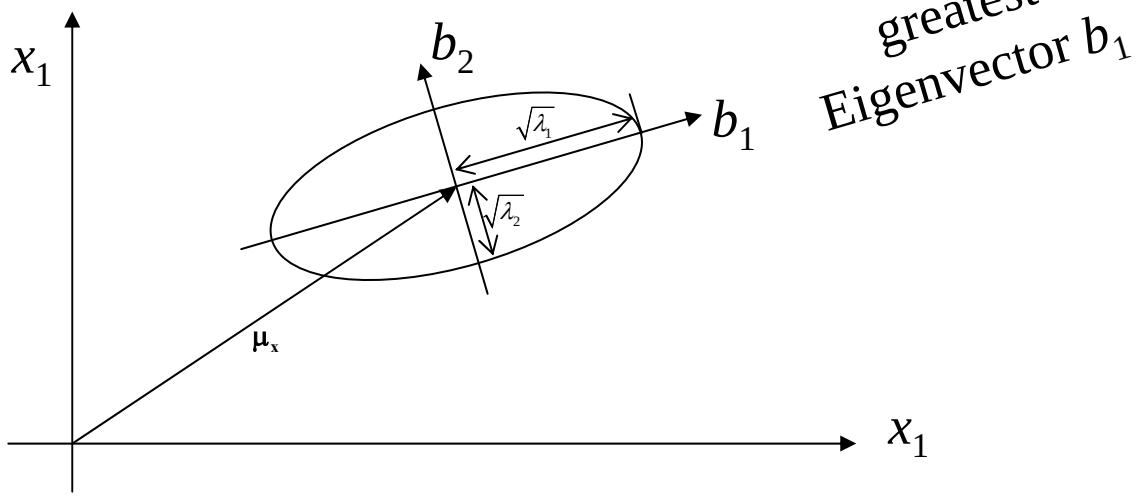
From Eigenvalue equation:

$$\mathbf{Kb} = \lambda\mathbf{b} \Rightarrow [\mathbf{K} - \lambda\mathbf{I}]\mathbf{b} = 0$$

result

Eigenvalues: $\lambda_1$, $\lambda_2$

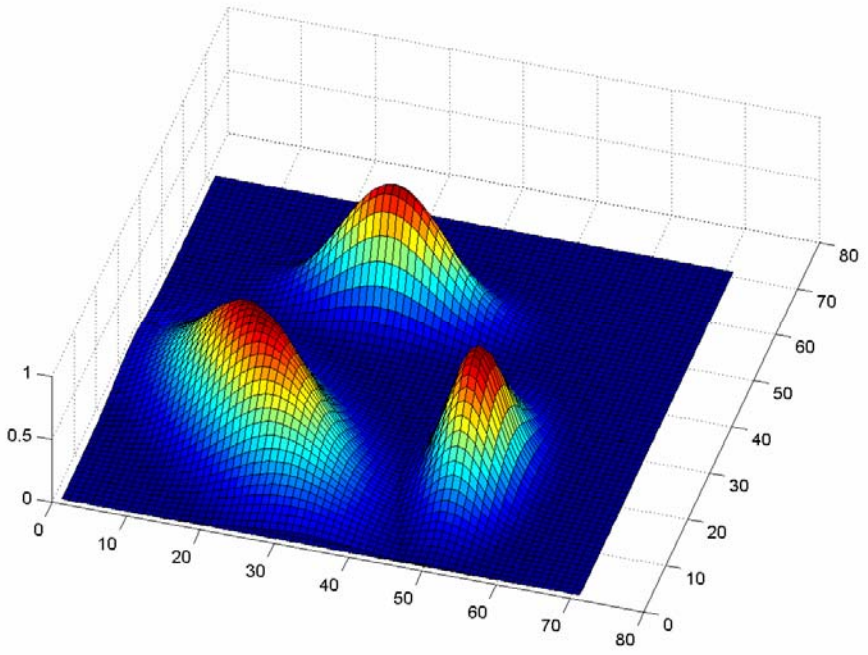
and Eigenvectors: $b_1, b_2$

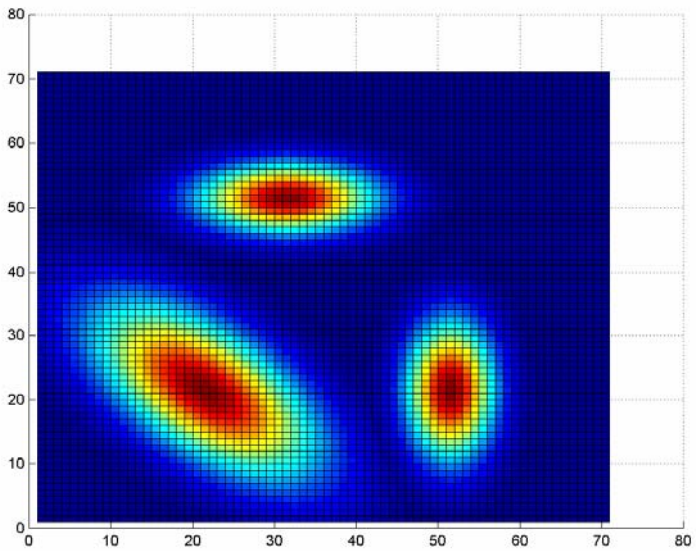


# The Bayes classifier

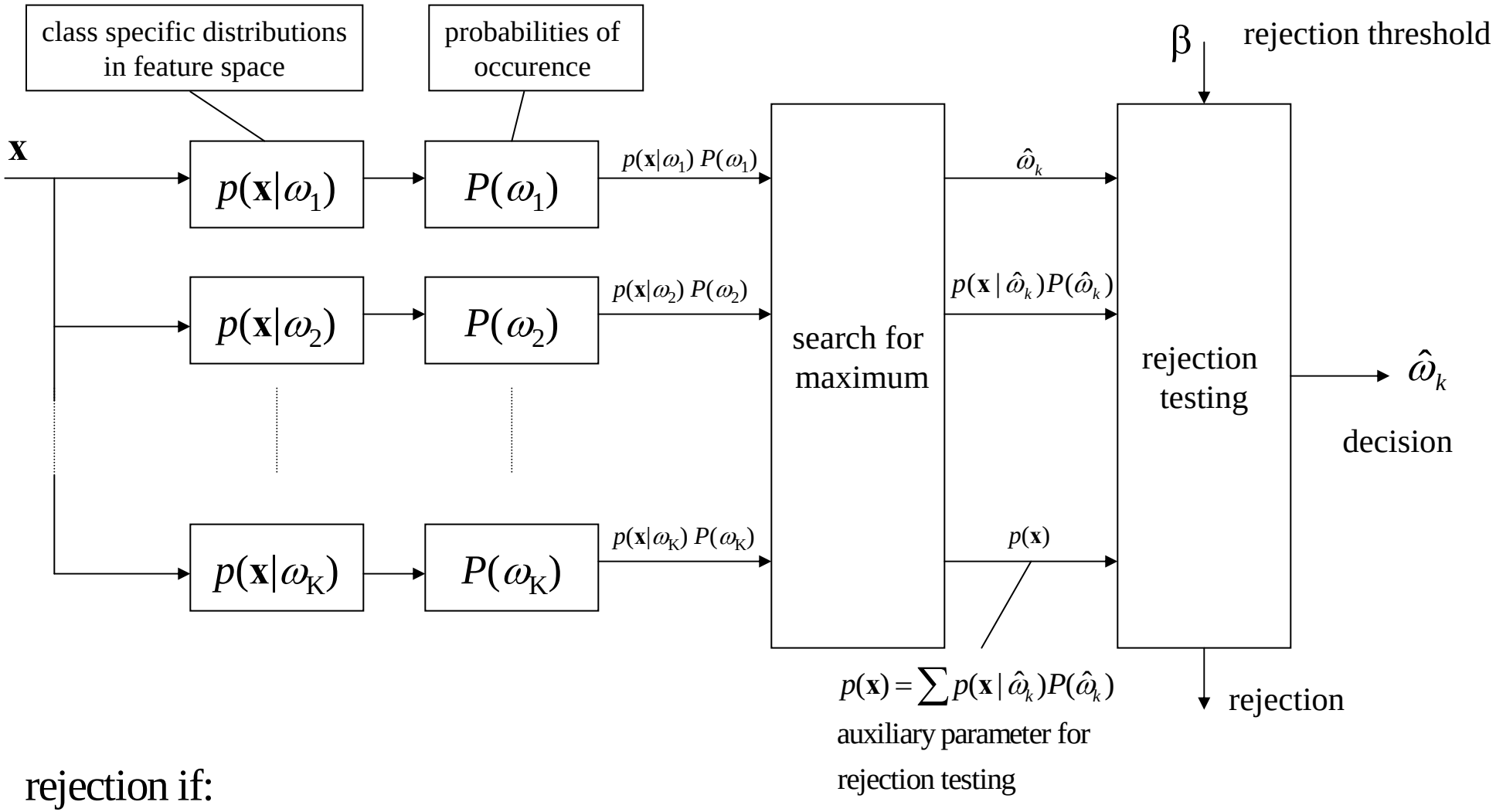
$$\max_{\omega_i} P(\omega_i \mid x)$$

Assumption: class specific Gaussian distributions



# Optimal recognition system




rejection if:

$p(\mathbf{x}\,|\,\hat{\omega}_k)P(\hat{\omega}_k) < \beta\, p(\mathbf{x})$

i.e. $P(\hat{\omega}_k\,|\,\mathbf{x}) < \beta$

If probability is too small

➔ rejection (otherwise decision very uncertain)

# Note: positive definiteness of covariance matrix **K**

The observations of random processes are expected to be independent.

$$\text{Assumpt.:}\quad Q = \mathbf{z}^T \mathbf{K}\mathbf{z} > 0 \quad \text{for} \quad \forall \mathbf{z} \neq \mathbf{0}$$

$$Q = \mathbf{z}^T E\{(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\}\mathbf{z}$$

$$= E\{\mathbf{z}^T(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{z}\} = E\{[\underbrace{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{z}}_{=w\ \text{(scalar)}}]^2\}$$

$$= E\{w^2\} > 0 \quad \text{for} \quad w \neq 0$$

In the singular case Q=0 the random process is $(\mathbf{x}\text{-}\boldsymbol{\mu}) \perp \mathbf{z}$ , i.e. only a linear subspace of the *N*-dimensional observed space $\mathbb{R}^N$ is taken. This is the case if the random variables do not span the complete space, i.e. one vector is linearly dependent on other vectors (e.g. if the 3-dimensional observations always lie within a plane).

For single vectors orthogonality $(\mathbf{x}\text{-}\boldsymbol{\mu}) \perp \mathbf{z}$ may be given, but not for the whole ensemble, so that E{...}=0.

# Consequences of **K**‘s positive definiteness

- **K** is regular and there exists $\mathbf{K}^{-1}$
- $\det(\mathbf{K})>0$
- $\mathbf{K}^{-1}$ is also positive definite
- $\det(\mathbf{K}^{-1})>0$
- The Eigenvalues of **K** are positive

# Case 1: class-wise arbitrary normally distributed features

This assumption specifies the MAP criterion even more:

$$p(\mathbf{x}, \omega_k) = p(\mathbf{x} \mid \omega_k) \cdot P(\omega_k)$$

The signal generating process can be broken down into $K$ independent subprocesses $\{p(\mathbf{x}|\omega_k)\}$ with parameters:

$\boldsymbol{\mu}_{x_k} = E\{\mathbf{x} \mid \omega_k\}$ class specific expected value

$\mathbf{K}_k = E\{(\mathbf{x} - \boldsymbol{\mu}_{x_k})(\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mid \omega_k\}$ class specific covariance matrix

Calculating the k-th decision function of the MAP criterion, results in:

$$D_k(\mathbf{x}) = p(\mathbf{x} \mid \omega_k) \cdot P(\omega_k) = \frac{P(\omega_k)}{\sqrt{(2\pi)^N \det(\mathbf{K}_k)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{x_k})^T \mathbf{K}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_{x_k})}$$

Using a monotonous map ln(...), which does not change proportions, results in:

$$D'_k(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2}\ln(\det(\mathbf{K}_k)) - \frac{1}{2}[(\mathbf{x}-\boldsymbol{\mu}_{x_k})^T \mathbf{K}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_{x_k})]$$

with concluding maximum comparison.

The borders of the classes are:

$$D_i'(\mathbf{x}) \stackrel{!}{=} D_j'(\mathbf{x})$$

which results in the interface $g_{ij}(\mathbf{x}) = 0$, with:

$$g_{ij}(\mathbf{x}) = \ln \frac{\det \mathbf{K}_i}{\det \mathbf{K}_j} - 2 \ln \frac{P(\omega_i)}{P(\omega_j)}$$
$$+ (\mathbf{x} - \boldsymbol{\mu}_{x_i})^T \mathbf{K}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_{x_i}) - (\mathbf{x} - \boldsymbol{\mu}_{x_j})^T \mathbf{K}_j^{-1} (\mathbf{x} - \boldsymbol{\mu}_{x_j})$$

From the difference of two square forms results from a common square form:

$$\boxed{g_{ij}(\mathbf{x}) = g_0 + (\mathbf{x} - \mathbf{x}_0)^T \mathbf{M}^{-1} (\mathbf{x} - \mathbf{x}_0)}$$

with:

$$g_0 = \ln\frac{\det \mathbf{K}_i}{\det \mathbf{K}_j} - 2\ln\frac{P(\omega_i)}{P(\omega_j)} + \boldsymbol{\mu}_{x_i}^T \mathbf{K}_i^{-1}\boldsymbol{\mu}_{x_i}$$

$$- \boldsymbol{\mu}_{x_j}^T \mathbf{K}_j^{-1}\boldsymbol{\mu}_{x_j} + \mathbf{x}_0^T \mathbf{M}^{-1}\mathbf{x}_0$$

$$\mathbf{x}_0 = \mathbf{M}[\mathbf{K}_i^{-1}\boldsymbol{\mu}_{x_i} - \mathbf{K}_j^{-1}\boldsymbol{\mu}_{x_j}]$$

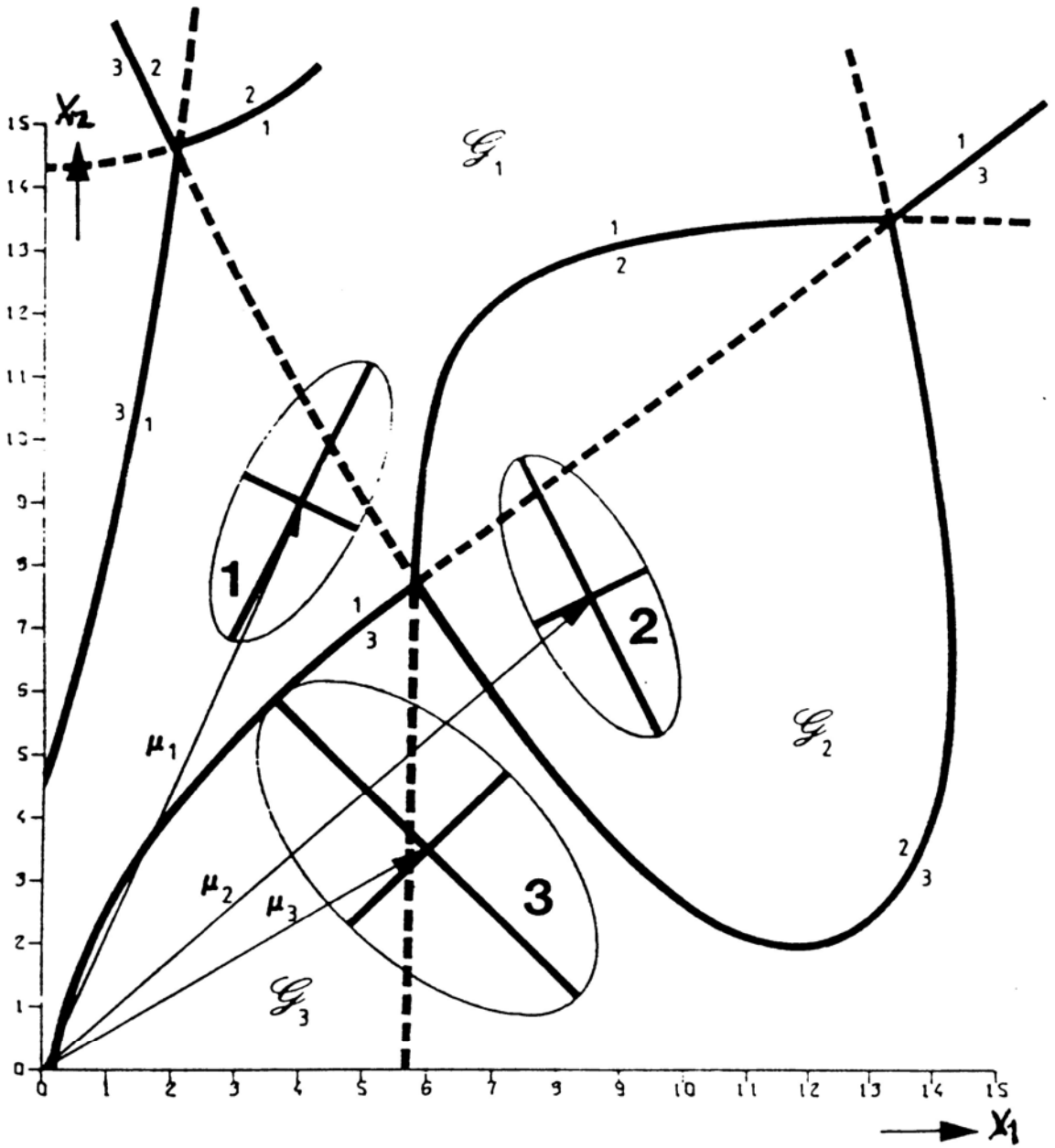
$$\mathbf{M} = [\mathbf{K}_i^{-1} - \mathbf{K}_j^{-1}]^{-1} = \mathbf{K}_i[\mathbf{K}_j - \mathbf{K}_i]^{-1}\mathbf{K}_j$$

$$= \mathbf{K}_j[\mathbf{K}_j - \mathbf{K}_i]^{-1}\mathbf{K}_i$$

The matrix $\mathbf{M}^{-1}$ that characterizes the square form is now not neccessarly positive definite => the interfaces of the regions are *general* conic sections (for N=2: ellipses, parables, hyperbolas, lines)

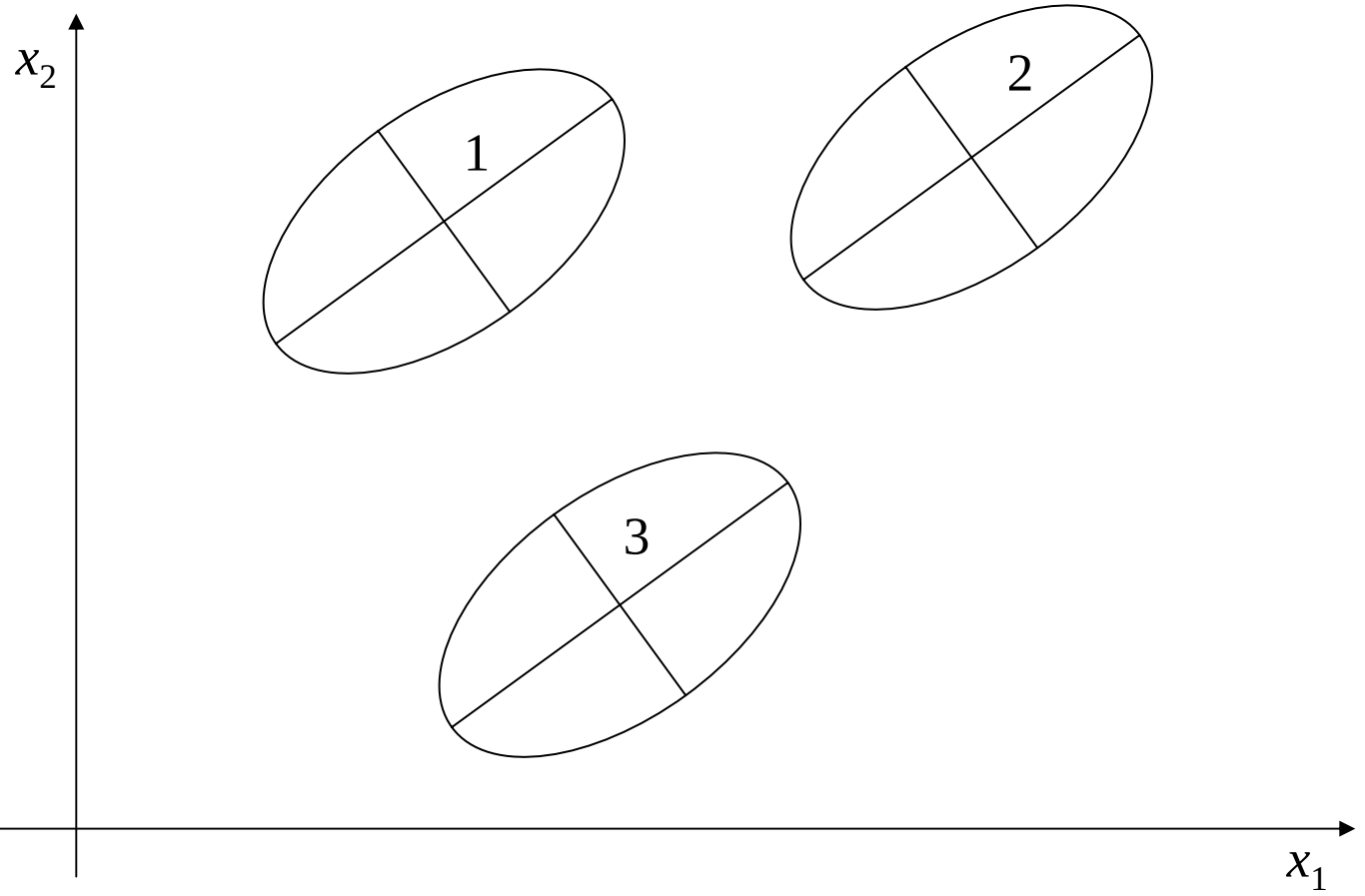
The discrimination functions $D'_k(\mathbf{x})$ are with regard to the feature space quadratic functions or polynoms of degree 2 (***quadratic or polynomial classifier***)

# class-wise normally distributed features




(taken from J. Schürmann: „Polynomklassifikatoren für die Zeichenerkennung“, Oldenbourg Verlag)

# Case 2: class-wise normally distributed features with *identical* covariance matrices **K**



# Case 2: class-wise normally distributed features with identical covariance matrices **K**

decision functions:

$$D_k'(\mathbf{x}) = \ln P(\omega_k) - \frac{1}{2}\ln(\det \mathbf{K}) - \frac{1}{2}[(\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mathbf{K}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{x_k})]$$

$$D_k'' = -2D_k' - \ln(\det \mathbf{K})$$

$$\Rightarrow D_k'' = -2\ln P(\omega_k) + (\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mathbf{K}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{x_k})$$

For equal a-priori probabilities $P(\omega_k) = 1/K$ follows:

$$D_k''' = D_k'' - 2\ln K$$

$$\Rightarrow \boxed{D_k''' = d_M^2 = (\mathbf{x} - \boldsymbol{\mu}_{x_k})^T \mathbf{K}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{x_k})}$$

***Mahalanobis-distance classifier***

This is a general weighted square metrics.

# Expressing as *linear* classifier

The decision functions still contains a square term, which is identical for every class and thus can be eliminated. Thus the classifier can be expressed linearly.

Alternatively results:

$$D_k'' = D_k' + \frac{1}{2}[\ln(\det \mathbf{K}) + \mathbf{x}^T \mathbf{K}^{-1} \mathbf{x}]$$

$$\Rightarrow D_k'' = \ln P(\omega_k) - \frac{1}{2} \boldsymbol{\mu}_{x_k}^T \mathbf{K}^{-1} \boldsymbol{\mu}_{x_k} + \boldsymbol{\mu}_{x_k}^T \mathbf{K}^{-1} \mathbf{x}$$

This term is linear in $\mathbf{x}$ !

$$\Rightarrow \boxed{D_k''(\mathbf{x}) = a_{0k} + \mathbf{a}_k^T \mathbf{x} = a_{0k} + \langle \mathbf{a}_k, \mathbf{x} \rangle}$$

***hyperplane as separation plane!***

with:

$$a_{0k} = \ln P(\omega_k) - \frac{1}{2} \boldsymbol{\mu}_{x_k}^T \mathbf{K}^{-1} \boldsymbol{\mu}_{x_k}$$

$$\mathbf{a}_k = \mathbf{K}^{-1} \boldsymbol{\mu}_{x_k}$$

# case 3: class-wise normally distributed features with unit matrix as covariance matrix $\mathbf{K}=\sigma^2\mathbf{I}$

(sterical invariant relations, hyperspheres)

decision function:

$$D'_k(\mathbf{x}) = \ln P(\omega_k) - N \ln \sigma - \frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\mu}_{x_k})^T(\mathbf{x}-\boldsymbol{\mu}_{x_k})$$

For constant A-priori prob. follows:

$$\Rightarrow \boxed{D''_k = \left\|\mathbf{x}-\boldsymbol{\mu}_{x_k}\right\|^2}$$

***Euclidian metrics***
***Minimal-distance-classifier***

Also as linear classifier:

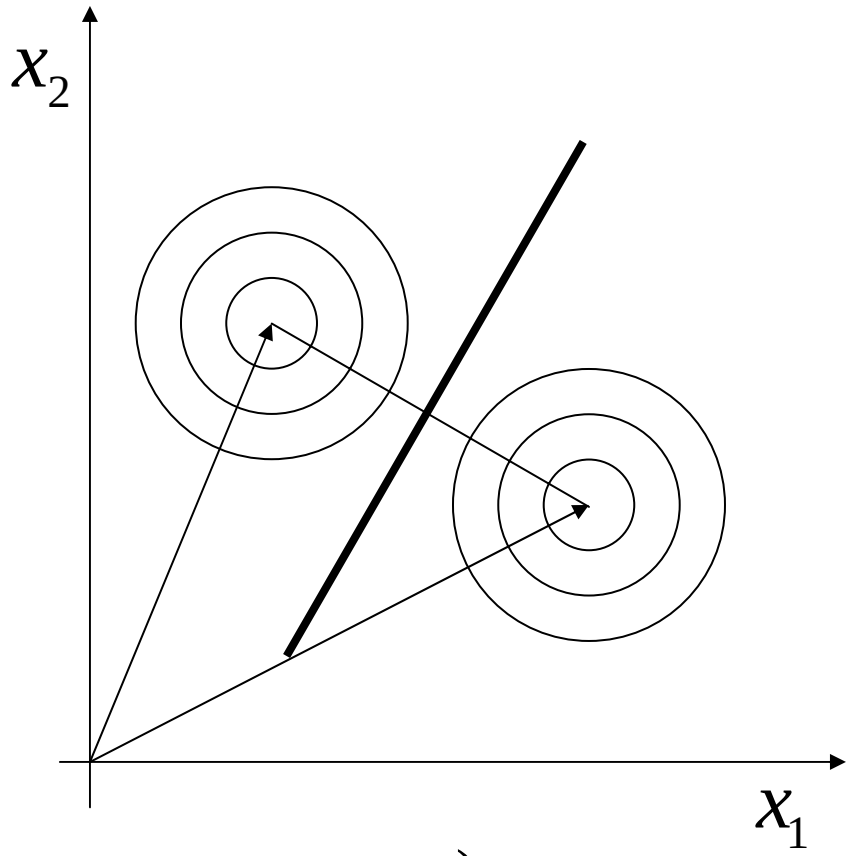
$$D'''_k = \frac{1}{2}(\|\mathbf{x}\|^2 - D''_k)$$

$$\Rightarrow \boxed{D'''_k(\mathbf{x}) = a_{0k} + \mathbf{a}_k^T\mathbf{x}}$$

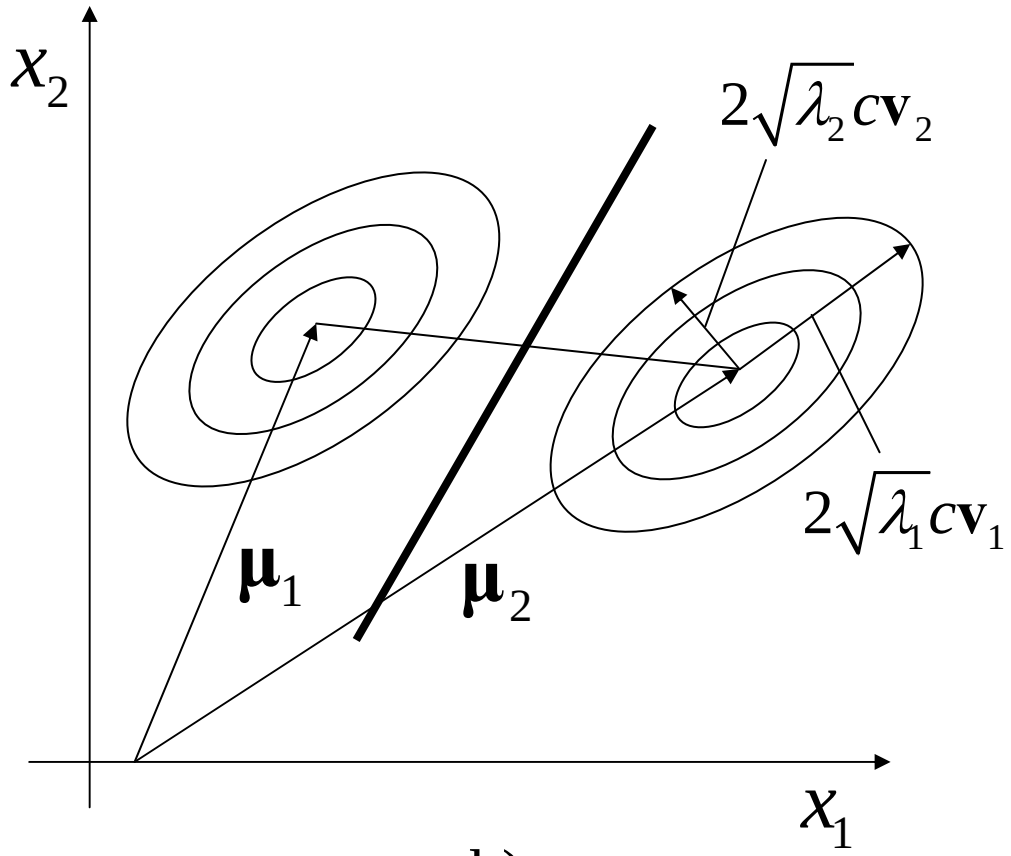
with:

$$D'''_k = \frac{1}{2}(\|\mathbf{x}\|^2 - (\|\mathbf{x}\|^2 - \|\boldsymbol{\mu}_{x_k}\|^2 - 2<\mathbf{x},\boldsymbol{\mu}_{x_k}>)) = \underbrace{-\frac{1}{2}\|\boldsymbol{\mu}_{x_k}\|^2}_{a_{0k}} + <\underbrace{\boldsymbol{\mu}_{x_k}}_{\mathbf{a}_k},\mathbf{x}>$$

# Curves of constant a) Euclidian and b) Mahalanobis-distance $d_M$ to expected value of respective class




a)




b)

# Decision limits of Bayes classifier for normally distributed sample classes

- Case 1: matlab-Bayes-Fall1.bat
- Case 2: matlab-Bayes-Fall2.bat
- Case 3: matlab-Bayes-Fall3.bat

# Transformation of the Mahalanobis metrics to sterical invariant measures

The covariance matrix can be diagonalized with KLT:

$$\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N) = \mathbf{A}^T \mathbf{K} \mathbf{A}$$

resp:

$$\mathbf{K} = \mathbf{A} \mathbf{\Lambda} \mathbf{A}^T$$

The unit matrix holds: $\mathbf{A}^T = \mathbf{A}^{-1}$

The Eigenvalues and the Eigenvectors of K define the diagonal matrix, and the Eigenvectors define the tansformation matrix:

$$\mathbf{A} = \left[ \mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_N \right]$$

Curves of constant Mahalanobis distance yield to:

$$(\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{A} \mathbf{\Lambda}^{-1} \mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu}_i) = c^2$$

# Transformation to sterical invariant measures

Introducing a coordinate transformation: $\mathbf{x}' = \mathbf{A}^T \mathbf{x}$

The original coordinates are being projected onto the Eigenvectors and thus the following curves with constant Mahalanobis-distance result:

$$\frac{(x_1' - \mu_{i1}')^2}{\lambda_1} + \cdots + \frac{(x_N' - \mu_{iN}')^2}{\lambda_N} = c^2$$

This is a hyperellipsoid in the new coordinate system.

With $x_k'' = x_k' / \lambda_k$ and $\mu_{ik}'' = \mu_{ik}' / \lambda_k$

sterical invariant (Euclidian) measures result:

$(x_1'' - \mu_{i1}'')^2 + \cdots + (x_N'' - \mu_{iN}'')^2 = c^2$ (spheres)

# Example: Two class problem of dimension 2

Let the covariance matrix and the expected values:

$$\mathbf{K} = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix} \quad \boldsymbol{\mu}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^T \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 3 & 3 \end{bmatrix}^T$$

Classify observation $\mathbf{x} = \begin{bmatrix} 1.0 & 2.2 \end{bmatrix}^T$ using Bayes.

Classification is done by calculating the Mahalanobis distance to both expected values:

$$d_M^2(\boldsymbol{\mu}_1, \mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_1)^T \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) = \begin{bmatrix} 1.0 & 2.2 \end{bmatrix} \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

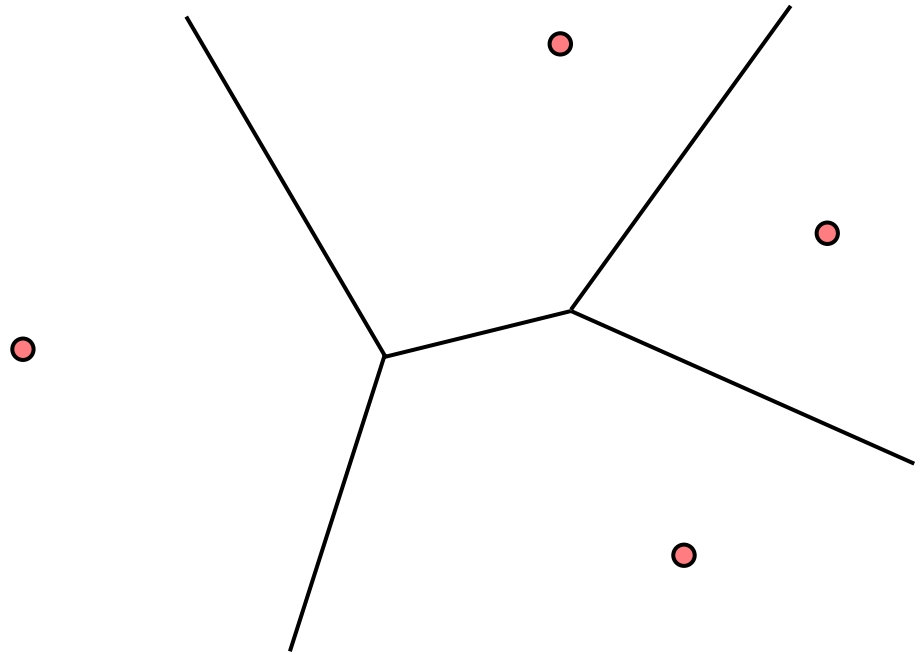
analogous:

$$d_M^2(\boldsymbol{\mu}_2, \mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu}_2)^T \mathbf{K}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) = \begin{bmatrix} -2.0 & -0.8 \end{bmatrix} \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

i.e. the obervation is being classified as class 1. Note, that the observation regarding the Euclidian distance is closer to classe 2!!

$$\left\| \mathbf{x} - \boldsymbol{\mu}_1 \right\|^2 = 5.84 \qquad \left\| \mathbf{x} - \boldsymbol{\mu}_2 \right\|^2 = 4.64$$

# The Voronoi diagramm in the two-dimensional space for the Euclidian distance



Decomposing the plane into regions $R_i$ for a set of points. Each region contains exactly the points, that are closer to the particular points than any other point:

$$R_i = \{\mathbf{x} : d(\mathbf{x}, \mathbf{x}_i) < d(\mathbf{x}, \mathbf{x}_j) \quad \text{for} \quad i \neq j\}$$

With cumulative distribution function $F(x_0) = P(x \le x_0)$ holds:

$$F(x) = \int_{-\infty}^{x} g(u)du = \frac{1}{2}(1 + \text{erf}\left(\frac{x-\mu_x}{\sqrt{2}\sigma}\right))$$

and with Gaussian error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_{0}^{x} e^{-t^2} dt$$

resp. complementary error function

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_{x}^{\infty} e^{-t^2} dt = 1 - \text{erf}(x)$$

the prob. of an error results as:

$$P(E) = \int_{x=d/2}^{\infty} g(u)du = 1 - F(x = d/2)$$

$$= \frac{1}{2}(1 - \text{erf}\left(\frac{x-\mu_x}{\sqrt{2}\sigma}\right))\Big|_{\substack{x=d/2 \\ \mu_x=0}} = \frac{1}{2}\text{erfc}\left(\frac{x-\mu_x}{\sqrt{2}\sigma}\right)\Big|_{\substack{x=d/2 \\ \mu_x=0}}$$

$$\boxed{P(E) = \frac{1}{2}\text{erfc}\left(\frac{d/2}{\sqrt{2}\sigma}\right)}$$

The prob. for a false classification decreases with growing class distance and increases with growing statistical spread of features.

The complete probability for an error in a $N$-dimensional feature space calculates from the minimum distance $d_{\min}$ to (Forney):

$$P(E) = const \cdot \frac{1}{2} \operatorname{erfc}\left( \frac{d_{\min} / 2}{\sqrt{2}\sigma} \right)$$

(without proof)