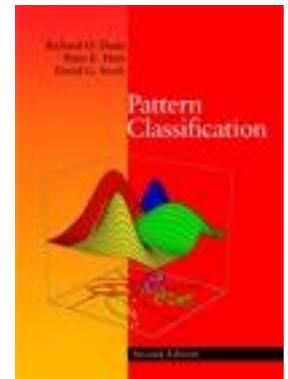
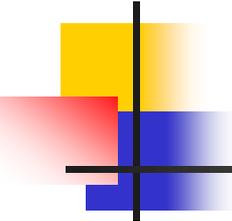


# Chapter 2:

## Bayesian Decision Theory (Part 2)

- Minimum-Error-Rate Classification
- Classifiers, Discriminant Functions and Decision Surfaces
- The Normal Density
- Entropy and Information





# Minimum-Error-Rate Classification

---

- Actions are decisions on classes  
If action  $\alpha_i$  is taken and the true state of nature is  $\omega_j$  then:  
the decision is correct if  $i = j$  and is error if  $i \neq j$
- Seek a decision rule that minimizes the *probability of error* which is the *error rate*

- Introduction of the symmetrical or zero-one loss function:

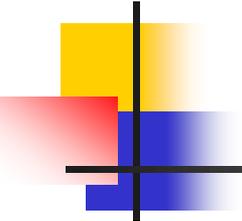
$$\lambda(\alpha_i | \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \dots, c$$

Therefore, the conditional risk is:

$$R(\alpha_i | \mathbf{x}) = \sum_{j=1}^{j=c} \lambda(\alpha_i | \omega_j) P(\omega_j | \mathbf{x}) \quad (13)$$

$$= \sum_{j \neq i} P(\omega_j | \mathbf{x}) = 1 - P(\omega_i | \mathbf{x})$$

*"The risk corresponding to this loss function is the average probability error"*

- 
- 
- Minimize the risk requires maximize  $P(\omega_i / \mathbf{x})$   
(since  $R(\alpha_i / \mathbf{x}) = 1 - P(\omega_i / \mathbf{x})$ )
  - For Minimum error rate
    - Decide  $\omega_i$  if  $P(\omega_i / \mathbf{x}) > P(\omega_j / \mathbf{x}) \forall j \neq i$

- Regions of decision and zero-one loss function:

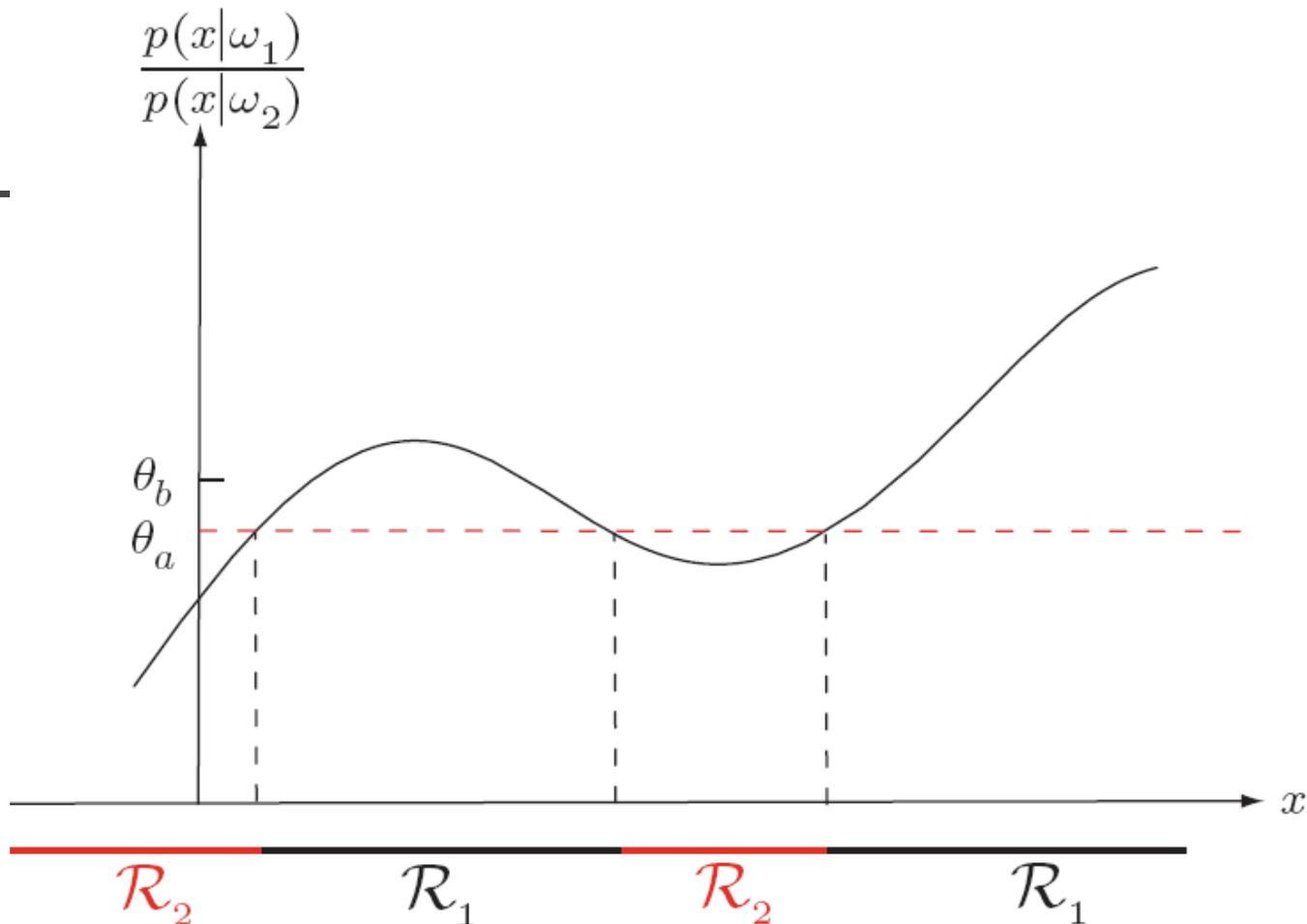
Let  $\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda$  then decide  $\omega_1$  if :  $\frac{p(x | \omega_1)}{p(x | \omega_2)} > \theta_\lambda$

- If  $\lambda$  is the zero-one loss function which means:

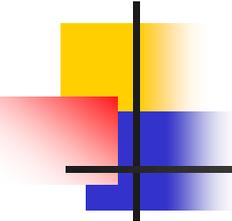
$$\lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\text{then } \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

$$\text{if } \lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix} \text{ then } \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$



**Figure 2.3:** The likelihood ratio  $p(x|\omega_1)/p(x|\omega_2)$  for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold  $\theta_a$ . If our loss function penalizes miscategorizing  $\omega_2$  as  $\omega_1$  patterns more than the converse, (i.e.,  $\lambda_{12} > \lambda_{21}$ ), we get the larger threshold  $\theta_b$ , and hence  $\mathcal{R}_1$  becomes smaller.



# Minimax Criterion

---

- We may want to design our classifier to perform well over a *range* of prior probabilities.
- A reasonable approach is then to design our classifier so that the *worst* overall risk for any value of the priors is as small as possible — that is, **minimize the maximum possible overall risk**.
- Let  $R_1$  denote that (as yet unknown) region in feature space where the classifier decides  $\omega_1$  and likewise for  $R_2$  and  $\omega_2$

The overall risk in terms of conditional risks is (see Eqs 12,13):

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x}) d\mathbf{x}, \quad \longrightarrow$$

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

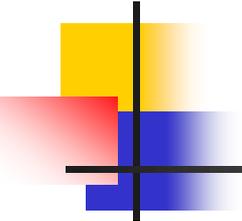
We use

$$P(\omega_2) = 1 - P(\omega_1) \text{ and that } \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1) d\mathbf{x} = 1 - \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) d\mathbf{x}$$

So we have:

$$\begin{aligned}
 R(P(\omega_1)) &= \overbrace{\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) d\mathbf{x}}^{= R_{mm}, \text{ minimax risk}} \\
 &+ P(\omega_1) \underbrace{\left[ (\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) d\mathbf{x} \right]}_{= 0 \text{ for minimax solution}}.
 \end{aligned} \tag{22}$$

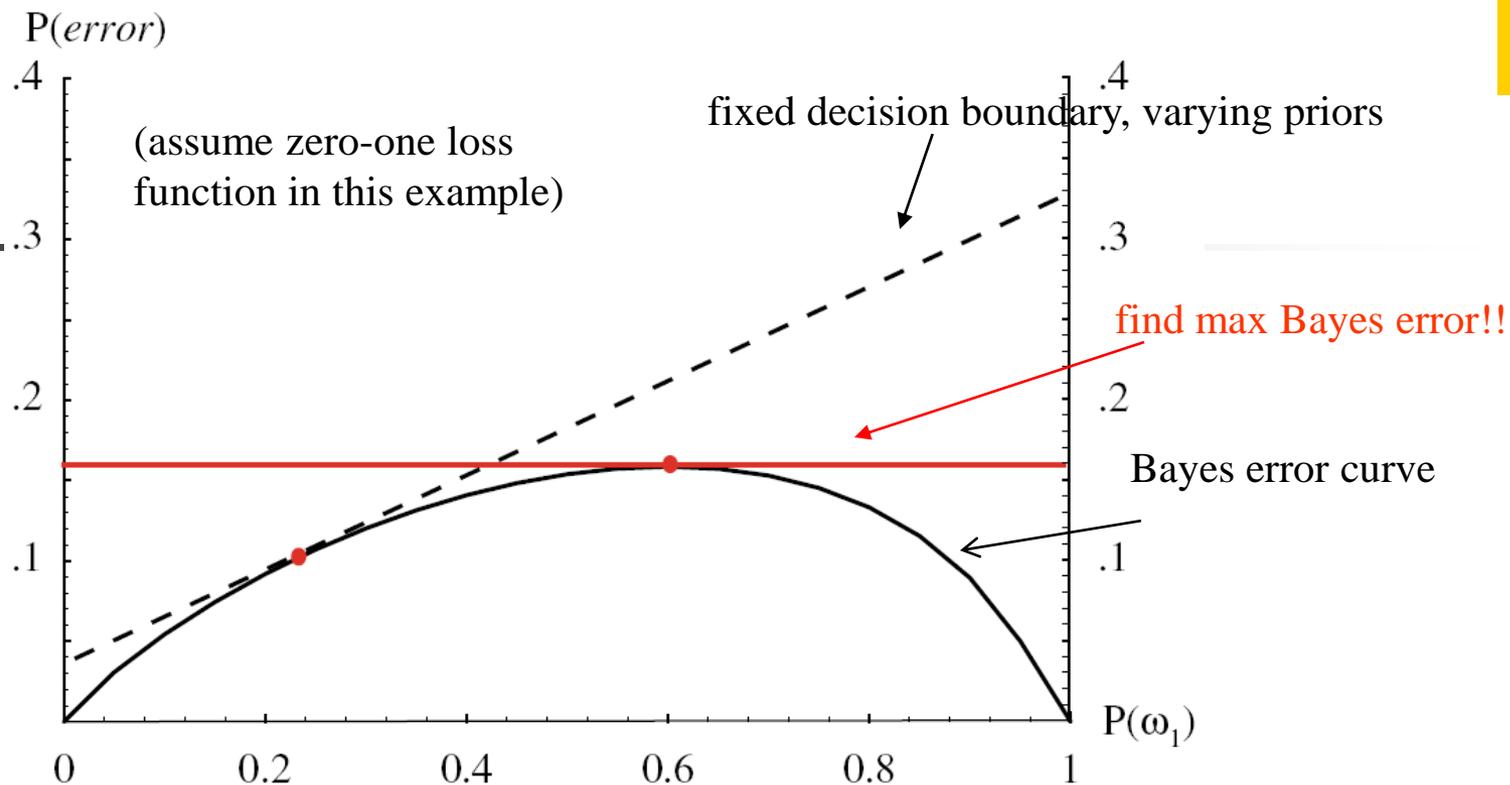
Note that the overall risk is linear in  $P(\omega_1)$ . If we can find a boundary such that the constant of proportionality is 0, then the risk is independent of priors. This is the *minimax solution*, and the *minimax risk*,  $R_{mm}$ , can be read from Eq. 22:



---

$$\begin{aligned} R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2) d\mathbf{x} \\ &= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1) d\mathbf{x}. \end{aligned}$$

The value of the minimax risk,  $R_{mm}$ , is hence equal to the worst Bayes risk.



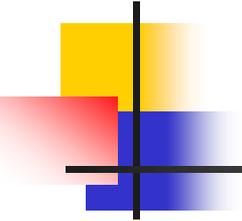
**Figure 2.4:** The curve at the bottom shows the minimum (Bayes) error as a function of prior probability  $P(\omega_1)$  in a two-category classification problem of fixed distributions. For each value of the priors (e.g.,  $P(\omega_1) = 0.25$ ) there is a corresponding optimal decision boundary and associated Bayes error rate. For any (fixed) such boundary, if the priors are then changed, the probability of error 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 our decision boundary for the maximum Bayes error (here  $P(\omega_1) = 0.6$ ), and thus the error will not change as a function of prior, as shown by the solid red horizontal line.

# Neyman-Pearson Criterion

- An alternative to the Bayes decision rules for a two-class problem is the Neyman–Pearson test.
- We may classify a pattern of class  $\omega_1$  as belonging to class  $\omega_2$  or a pattern from class  $\omega_2$  as belonging to class  $\omega_1$ .
- Let the probability of these two errors be  $\epsilon_1$  and  $\epsilon_2$  respectively, so that

$$\epsilon_1 = \int_{\Omega_2} p(\mathbf{x}|\omega_1) d\mathbf{x} = \text{error probability of Type I}$$

and


$$\epsilon_2 = \int_{\Omega_1} p(\mathbf{x}|\omega_2) d\mathbf{x} = \text{error probability of Type II}$$

- The Neyman–Pearson decision rule is to minimise the error  $\epsilon_1$  subject to  $\epsilon_2$  being equal to a constant,  $\epsilon_0$ , say.
- If class  $\omega_1$  is termed the positive class and class  $\omega_2$  the negative class, then  $\epsilon_1$  is referred to as the *false negative rate*, the proportion of positive samples incorrectly assigned to the negative class;  $\epsilon_2$  is the *false positive rate*, the proportion of negative samples classified as positive.

We seek the minimum of

$$r = \int_{\Omega_2} p(\mathbf{x}|\omega_1) d\mathbf{x} + \mu \left\{ \int_{\Omega_1} p(\mathbf{x}|\omega_2) d\mathbf{x} - \epsilon_0 \right\}$$

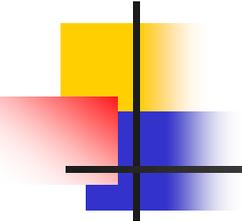
where  $\mu$  is a Lagrange multiplier and  $\epsilon_0$  is the specified false alarm rate. Then we have:

$$r = (1 - \mu\epsilon_0) + \int_{\Omega_1} \{\mu p(\mathbf{x}|\omega_2) - p(\mathbf{x}|\omega_1)\} d\mathbf{x}$$

This will be minimized if we choose  $\Omega_1$  such that the integrand is negative, i.e.

$$\text{if } \mu p(\mathbf{x}|\omega_2) - p(\mathbf{x}|\omega_1) < 0, \quad \text{then } \mathbf{x} \in \Omega_1$$

$$\text{or} \quad \text{if } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \mu, \quad \text{then } \mathbf{x} \in \Omega_1$$



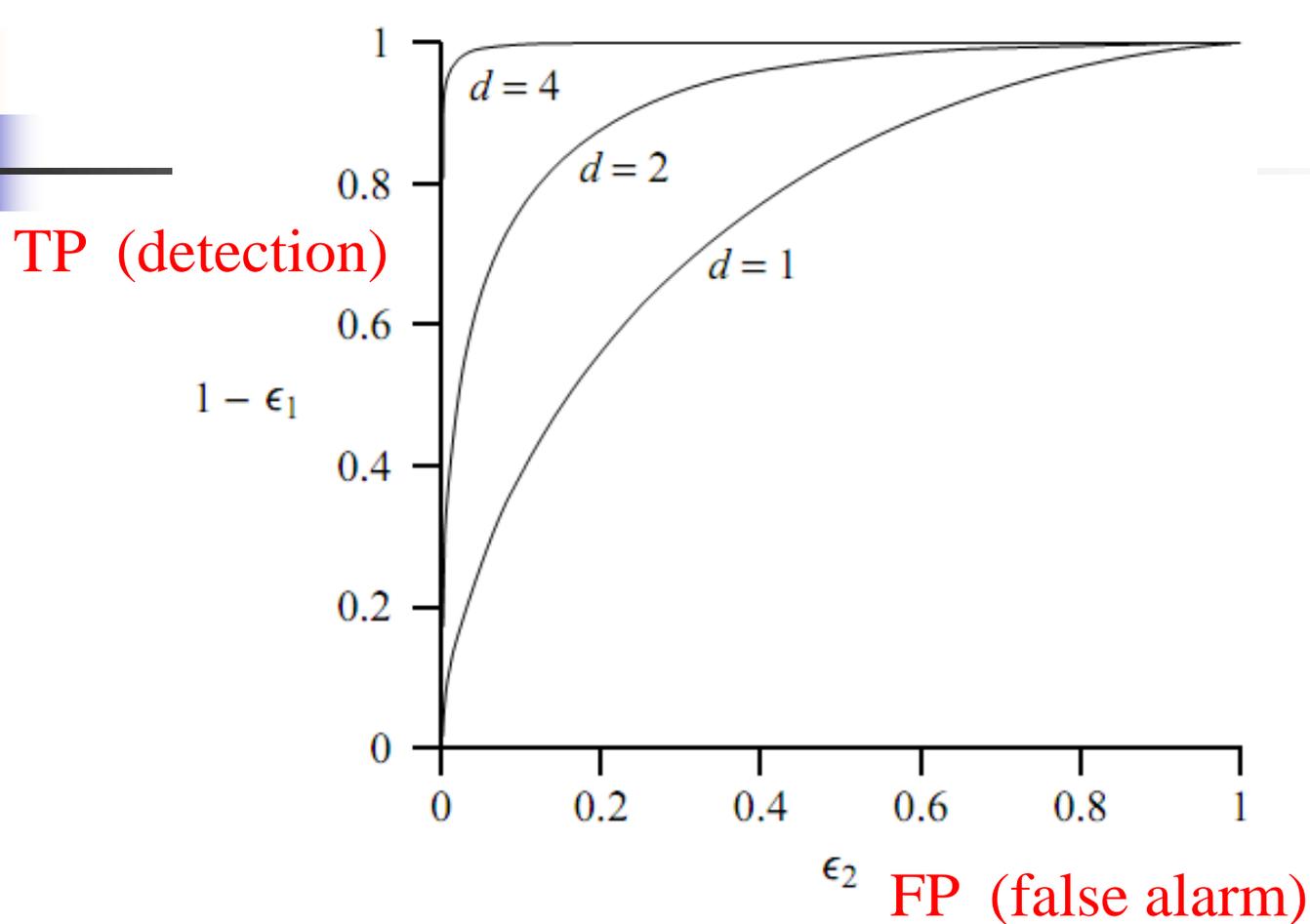
---

Thus the decision rule depends only on the within-class distributions and ignores the *a priori probabilities*.

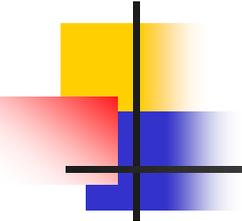
The threshold  $\mu$  is chosen so that 
$$\int_{\Omega_1} p(\mathbf{x}|\omega_2) d\mathbf{x} = \epsilon_0,$$

the specified false alarm rate. However, in general  $\mu$  cannot be determined analytically and requires numerical calculation.

Often, the performance of the decision rule is summarized in a receiver operating characteristic (ROC) curve, which plots the true positive against the false positive i.e.  $1 - \epsilon_1$  vs  $\epsilon_2$  as the threshold  $\mu$  is varied..



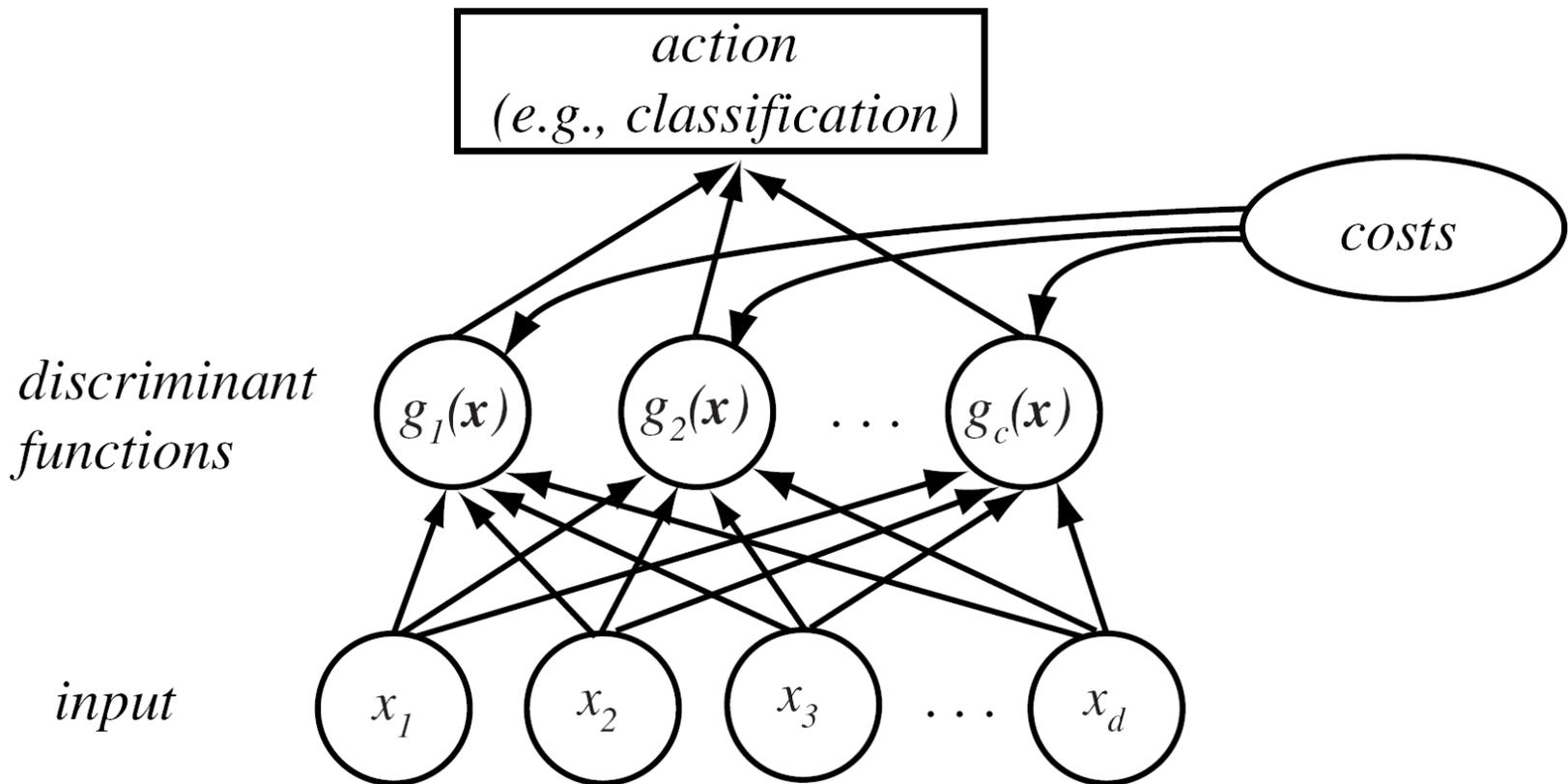
ROC curve for the univariate case of two normally distributed classes of unit variance and means separated by a distance,  $d$ .



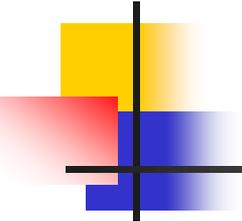
# Classifiers, Discriminant Functions and Decision Surfaces

---

- The multi-category case
  - Set of discriminant functions  $g_i(\mathbf{x})$ ,  $i = 1, \dots, c$
  - The classifier assigns a feature vector  $\mathbf{x}$  to class  $\omega_i$   
if:  $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$



**FIGURE 2.5.** The functional structure of a general statistical pattern classifier which includes  $d$  inputs and  $c$  discriminant functions  $g_i(\mathbf{x})$ . A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident.

- 
- Let  $g_i(\mathbf{x}) = -R(\alpha_i / \mathbf{x})$

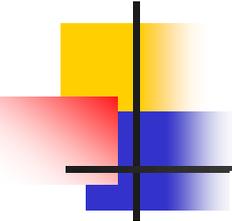
(max. discriminant corresponds to min. risk!)

- For the minimum error rate, we take

$$g_i(\mathbf{x}) = P(\omega_i / \mathbf{x})$$

(max. discrimination corresponds to max. posterior!)

- We may 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.

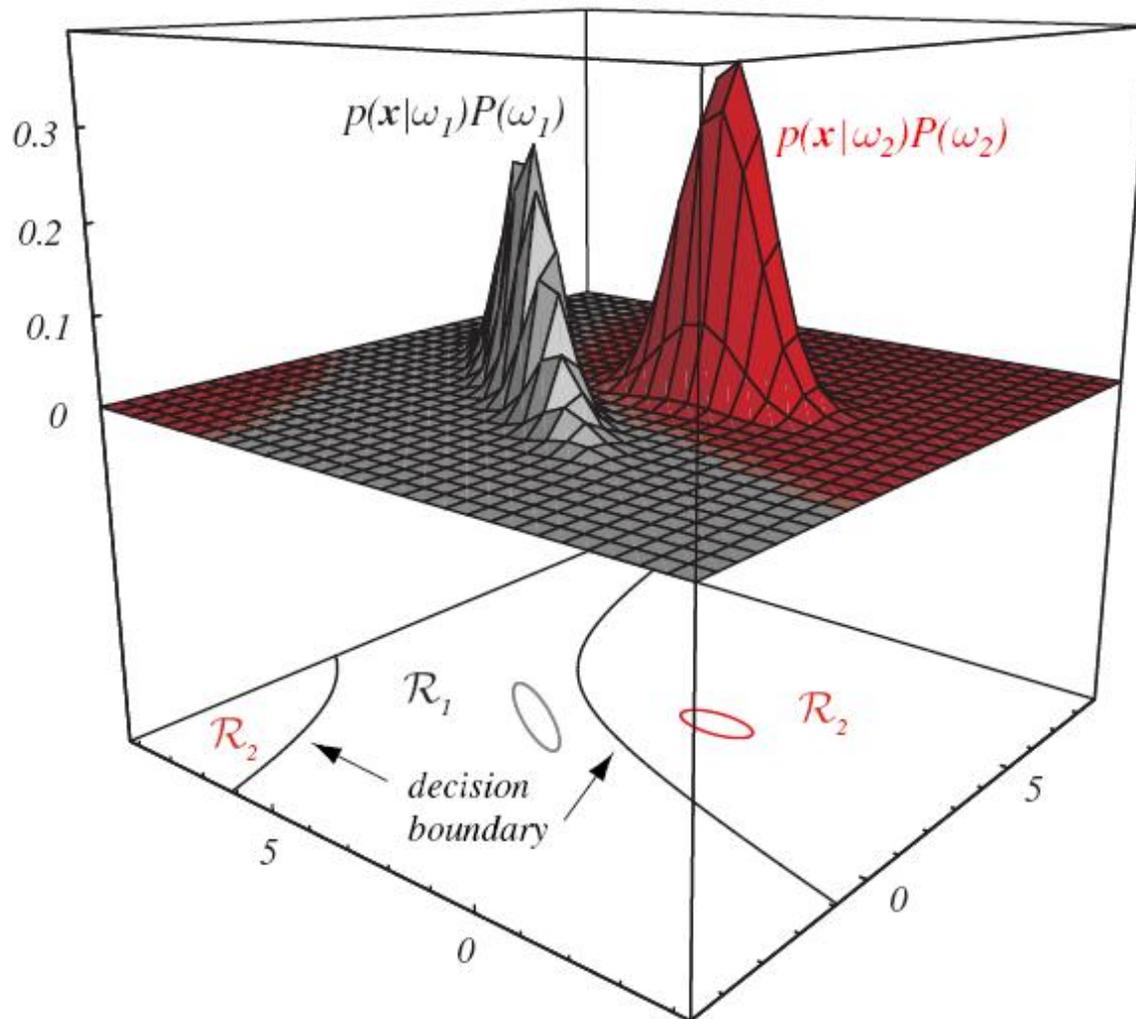
- 
- $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)}$
  - $g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$
  - $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i),$

Some of the above choices can be much simpler to understand or to compute than others.

- Feature space divided into  $c$  **decision regions**

if  $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall j \neq i$  then  $\mathbf{x}$  is in  $R_i$

( $R_i$  means assign  $\mathbf{x}$  to  $\omega_i$ ) (It does not depend on the form of discriminant functions)



**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  $\mathcal{R}_2$  is not simply connected.

## The two-category case

- A classifier is a “dichotomizer” that has two discriminant functions  $g_1$  and  $g_2$

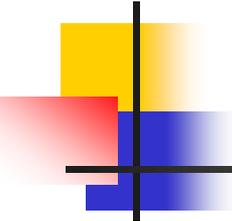
Let  $g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x})$

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

- Of the various forms in which the minimum-error-rate discriminant function can be written, the following two are particularly convenient:

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

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



# The Normal Density

---

The structure of a Bayes classifier is determined by the conditional densities  $p(\mathbf{x}/\omega_i)$  as well as by the prior probabilities.

The multivariate normal or Gaussian density has received more attention than the others, because it is **analytically tractable**.

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

$$\mathcal{E}[f(x)] \equiv \int_{-\infty}^{\infty} f(x)p(x)dx.$$

If we have samples in a set  $D$  from a discrete distribution:

$$\mathcal{E}[f(x)] = \sum_{x \in D} f(x)P(x), \text{ where } P(x) \text{ is the probability mass at } x.$$

# The Normal Density

## ■ Univariate density

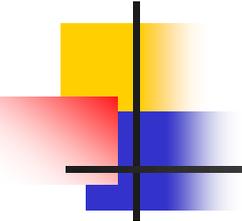
- Density which is analytically tractable
- Continuous density
- A lot of processes are asymptotically Gaussian
- Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

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

Where:

$\mu$  = mean (or expected value) of  $x$

$\sigma^2$  = expected squared deviation or variance

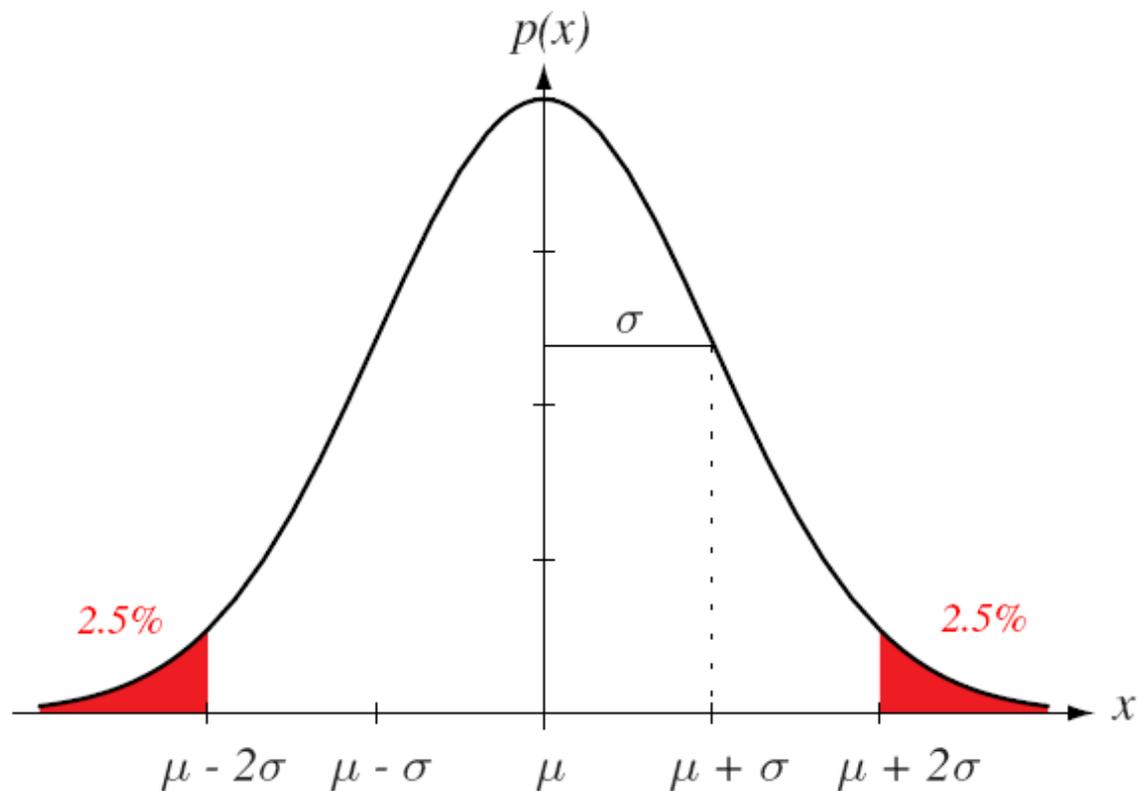


---

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x) dx,$$

$$\sigma^2 \equiv \mathcal{E}[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx.$$

The univariate normal density is completely specified by two parameters: its mean  $\mu$  and variance  $\sigma^2$ .



**FIGURE 2.7.** A univariate normal distribution has roughly 95% of its area in the range  $|x - \mu| \leq 2\sigma$ , as shown. The peak of the distribution has value  $p(\mu) = 1/\sqrt{2\pi} \sigma$

# Entropy and information

- Assume we have a discrete set of symbols  $\{v_1 v_2 \dots v_m\}$  with associated probabilities  $P_i$ . The entropy of the discrete distribution — a measure of the randomness or unpredictability of a sequence of symbols drawn from it — is

$$H = - \sum_{i=1}^m P_i \log_2 P_i,$$

where since we use the logarithm base 2 entropy is measured in *bits*.

- For a given number of symbols  $m$ , the uniform distribution in which each symbol is equally likely, is the maximum entropy distribution (and  $H = \log_2 m$  bits) — we have the maximum uncertainty about the identity of each symbol that will be chosen.

## Entropy:

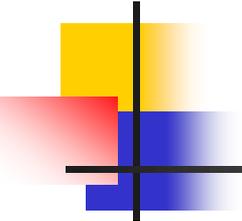
The entropy is a non negative quantity that describes the fundamental uncertainty in the values of points selected randomly from a distribution.

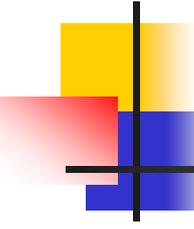
There is a deep relationship between the normal distribution and *entropy* and is given by.

$$H(p(x)) = - \int p(x) \ln p(x) dx$$

and measured in *nats*.

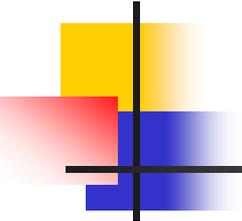
If a  $\log_2$  is used instead, the unit is the *bit*.

- 
- 
- It can be shown that the normal distribution has the maximum entropy of all distributions having a given **mean** and **variance**.
  - As stated by the *Central Limit Theorem*, the aggregate effect of a large number of small, independent random disturbances will lead to a Gaussian distribution.



# The Kullback-Leibler Divergence or the relative entropy

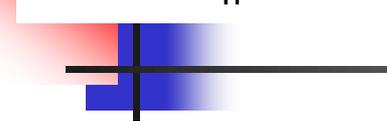
Consider some unknown distribution  $p(x)$ , and suppose that we have modelled this using an approximating distribution  $q(x)$ . If we use  $q(x)$  to construct a coding scheme for the purpose of transmitting values of  $x$  to a receiver, then the average additional amount of information (in nats) required to specify the value of  $x$  (assuming we choose an efficient coding scheme) as a result of using  $q(x)$  instead of the true distribution  $p(x)$  is given by:


$$\begin{aligned}\text{KL}(p\|q) &= - \int p(\mathbf{x}) \ln q(\mathbf{x}) \, d\mathbf{x} - \left( - \int p(\mathbf{x}) \ln p(\mathbf{x}) \, d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \, d\mathbf{x}\end{aligned}$$

$$\text{KL}(p\|q) \neq \text{KL}(q\|p)$$

The KL distance is closely related to the *mutual information* measure,  $I$ , between  $l$  scalar random variables,  $x_i$ ,  $i=1, 2, \dots, l$ . Indeed, let us compute the KL distance between the joint pdf  $p(\mathbf{x})$  and the pdf resulting from the product of the corresponding marginal<sup>†</sup> probability densities, that is,

<sup>†</sup> In the study of several random variables, the statistics of each are called marginal.

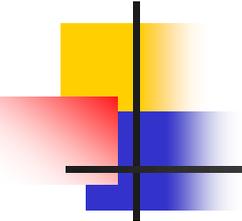

$$KL(p \parallel q) = L = - \int p(\mathbf{x}) \ln \frac{\prod_{i=1}^l p_i(x_i)}{p(\mathbf{x})} d\mathbf{x}$$

$$= \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} - \sum_{i=1}^l \int p(\mathbf{x}) \ln p_i(x_i) d\mathbf{x}$$

$$= -H(\mathbf{x}) - \sum_{i=1}^l \int p(\mathbf{x}) \ln p_i(x_i) d\mathbf{x}$$

Carrying out the integrations on the right-hand side it is straightforward to see the KL distance is equal to the mutual information,  $I$ , defined as:

$$I(x_1, x_2, \dots, x_l) = -H(\mathbf{x}) + \sum_{i=1}^l H(x_i)$$



---

where  $H(x_i)$  is the associated entropy of  $x_i$ , defined as

$$H(x_i) = - \int p_i(x_i) \ln p_i(x_i) dx_i$$

It is now easy to see that if the variables  $x_i$ ,  $i=1, 2, \dots, l$ , are statistically independent their mutual information  $I$  is zero. Indeed, in this case  $\prod_{i=1}^l p_i(x_i) = p(\mathbf{x})$ , hence  $L = I(x_1, x_2, \dots, x_l) = 0$ .

## ■ Multivariate density

- Multivariate normal density in  $d$  dimensions is:

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

where:

$\mathbf{x} = (x_1, x_2, \dots, x_d)^t$  (t stands for the transpose vector form)

$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_d)^t$  mean vector

$\boldsymbol{\Sigma} = d \times d$  covariance matrix

$|\boldsymbol{\Sigma}|$  and  $\boldsymbol{\Sigma}^{-1}$  are determinant and inverse respectively

## The *inner (dot) product*

$$\mathbf{a}^t \mathbf{b} = \sum_{i=1}^d a_i b_i,$$

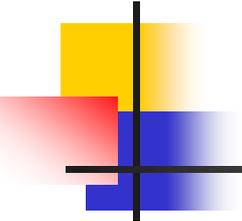
Abbreviation  $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

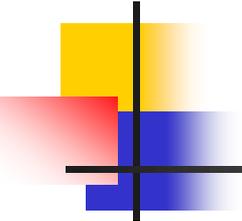
$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

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

$$\mu_i = \mathcal{E}[x_i]$$

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

- 
- The covariance matrix  $\Sigma$  is always **symmetric** and **positive semidefinite**.
  - (A matrix  $\mathbf{A}$  is pos. semidefinite if:  $\mathbf{z}^t \mathbf{A} \mathbf{z} \geq 0$  for any  $\mathbf{z}$ .)
  - In the case in which  $\Sigma$  is positive definite, the determinant of  $\Sigma$  is strictly positive.
  - The diagonal elements  $\sigma_{ii}$  are the variances of the respective  $x_i$  (i.e.,  $\sigma^2_i$ ), and the off-diagonal elements  $\sigma_{ij}$  are the *covariances* of  $x_i$  and  $x_j$ .
  - If  $x_i$  and  $x_j$  are *statistically independent*,  $\sigma_{ij} = 0$ .

- 
- Linear combinations of jointly normally distributed random variables, independent or not, are normally distributed.
  - If  $\mathbf{A}$  is a  $d$ -by- $k$  matrix and  $\mathbf{y} = \mathbf{A}^t \mathbf{x}$  is a  $k$ -component vector, then  $p(\mathbf{y}) \sim N(\mathbf{A}^t \boldsymbol{\mu}, \mathbf{A}^t \boldsymbol{\Sigma} \mathbf{A})$
  - 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  $\mathbf{x}$  onto  $\mathbf{a}$ .

# .... Diagonalization

(refer to Ch06-LinearTransformations.ppt)

Perform a change of basis (similarity transformation) using the eigenvectors as the basis vectors. If the eigenvalues are distinct, the new matrix will be diagonal.

$$\mathbf{B} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_n \end{bmatrix} \quad \begin{array}{l} \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\} \\ \{\lambda_1, \lambda_2, \dots, \lambda_n\} \end{array} \quad \begin{array}{l} \text{Eigenvectors} \\ \text{Eigenvalues} \end{array}$$

$$[\mathbf{B}^{-1} \mathbf{A} \mathbf{B}] = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix}$$

$\Phi = [\phi_1 \quad \dots \quad \phi_n]$   $d \times d$  matrix consisting of  $d$  eigenvectors

$\mathbf{y} = \Phi^t \mathbf{x}$  use  $\Phi$  as the transformation matrix  $\mathbf{A}$

$\Sigma_{\mathbf{y}} = \Phi^t \Sigma_{\mathbf{x}} \Phi = \Lambda$  covariance matrix of transformed vector

note:  $(\Phi^t)^t = \Phi$  and  $\Phi^{-1} = \Phi^t$

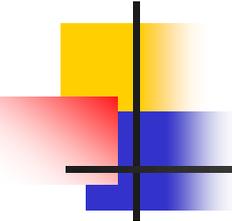
## *Whitening Transformation*

$$\mathbf{y} = \Lambda^{-1/2} \Phi^t \mathbf{x} = (\Phi \Lambda^{-1/2})^t \mathbf{x}$$

use  $\Phi \Lambda^{-1/2}$  as the transformation matrix  $\mathbf{A}$

$$\Sigma_{\mathbf{y}} = \Lambda^{-1/2} \Phi^t \Sigma_{\mathbf{x}} \Phi \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = \mathbf{I}$$

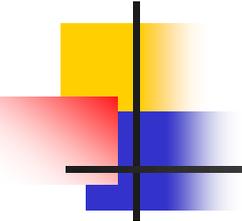
covariance matrix of transformed vector is identity matrix



# Properties

---

- If we define  $\Phi$  to be the matrix whose columns are the orthonormal eigenvectors of  $\Sigma$ , and  $\Lambda$  the diagonal matrix of the corresponding eigenvalues, then the transformation  $\mathbf{A}_w = \Phi\Lambda^{-1/2}$  applied to the coordinates insures that the transformed distribution has covariance matrix equal to the identity matrix.
- In signal processing, the transform  $\mathbf{A}_w$  is called a *whitening* transformation, since it makes the spectrum of eigenvectors of the transformed distribution uniform.

- 
- The multivariate normal density is completely specified by  $d + d(d + 1)/2$  parameters — the elements of the mean vector  $\boldsymbol{\mu}$  and the independent elements of the covariance matrix  $\boldsymbol{\Sigma}$ .
  - Samples drawn from a normal population tend to fall in a single cloud or cluster (Fig. 2.9); the center of the cluster is determined by the mean vector, and the shape of the cluster is determined by the covariance matrix.

- Whitening transformations are not orthonormal transformations because

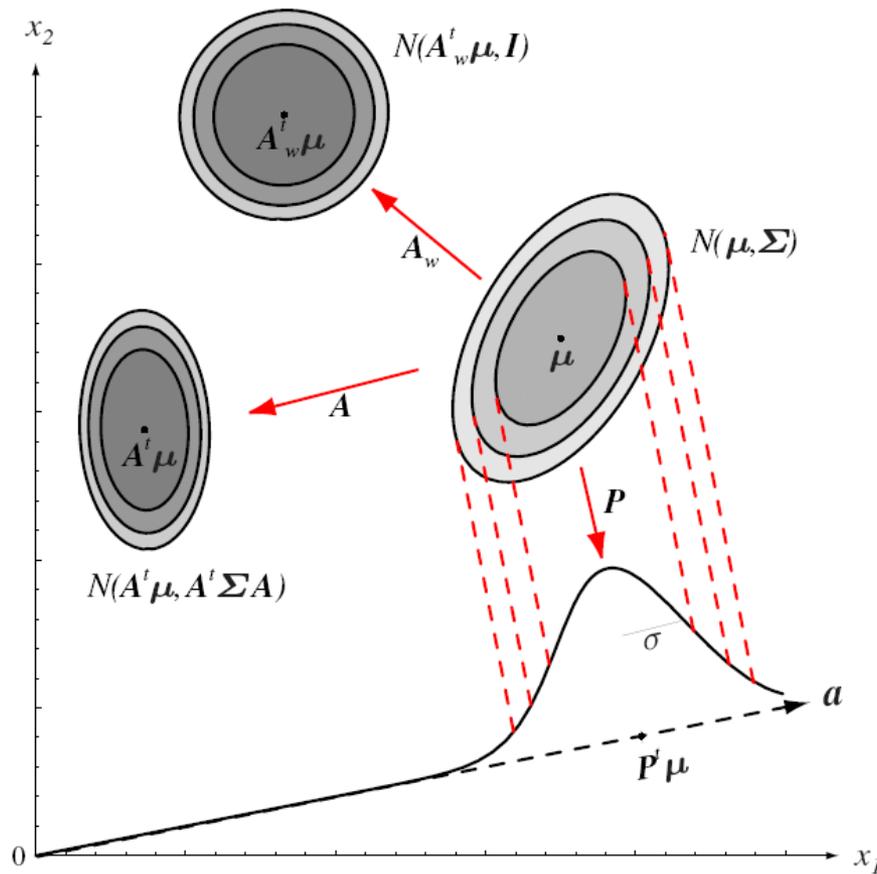
$$\left(\Phi\Lambda^{-1/2}\right)^t \Phi\Lambda^{-1/2} = \Lambda^{-1/2}\Phi^t\Phi\Lambda^{-1/2} = \Lambda^{-1/2}\Lambda^{-1/2} = \Lambda^{-1} \neq \mathbf{I}$$

Therefore, Euclidean distances are not preserved:

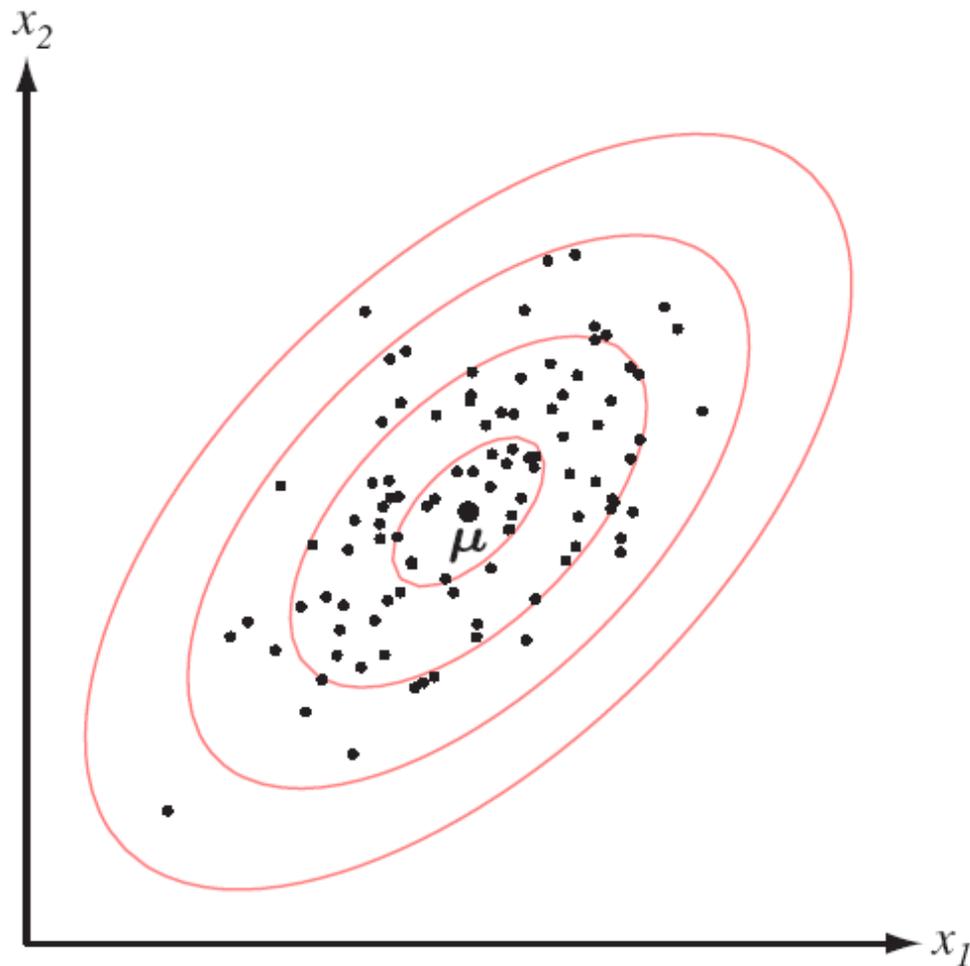
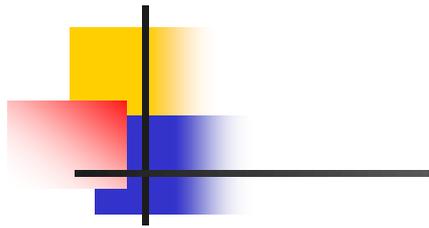
$$\|\mathbf{y}\|^2 = \mathbf{y}^t\mathbf{y} = \left(\mathbf{x}^t\Phi\Lambda^{-1}\Phi^t\mathbf{x}\right) = \mathbf{x}^t\Sigma_{\mathbf{x}}^{-1}\mathbf{x} \neq \|\mathbf{x}\|^2$$

- After a whitening transformation, the covariance matrix is invariant under any orthonormal transformation, because

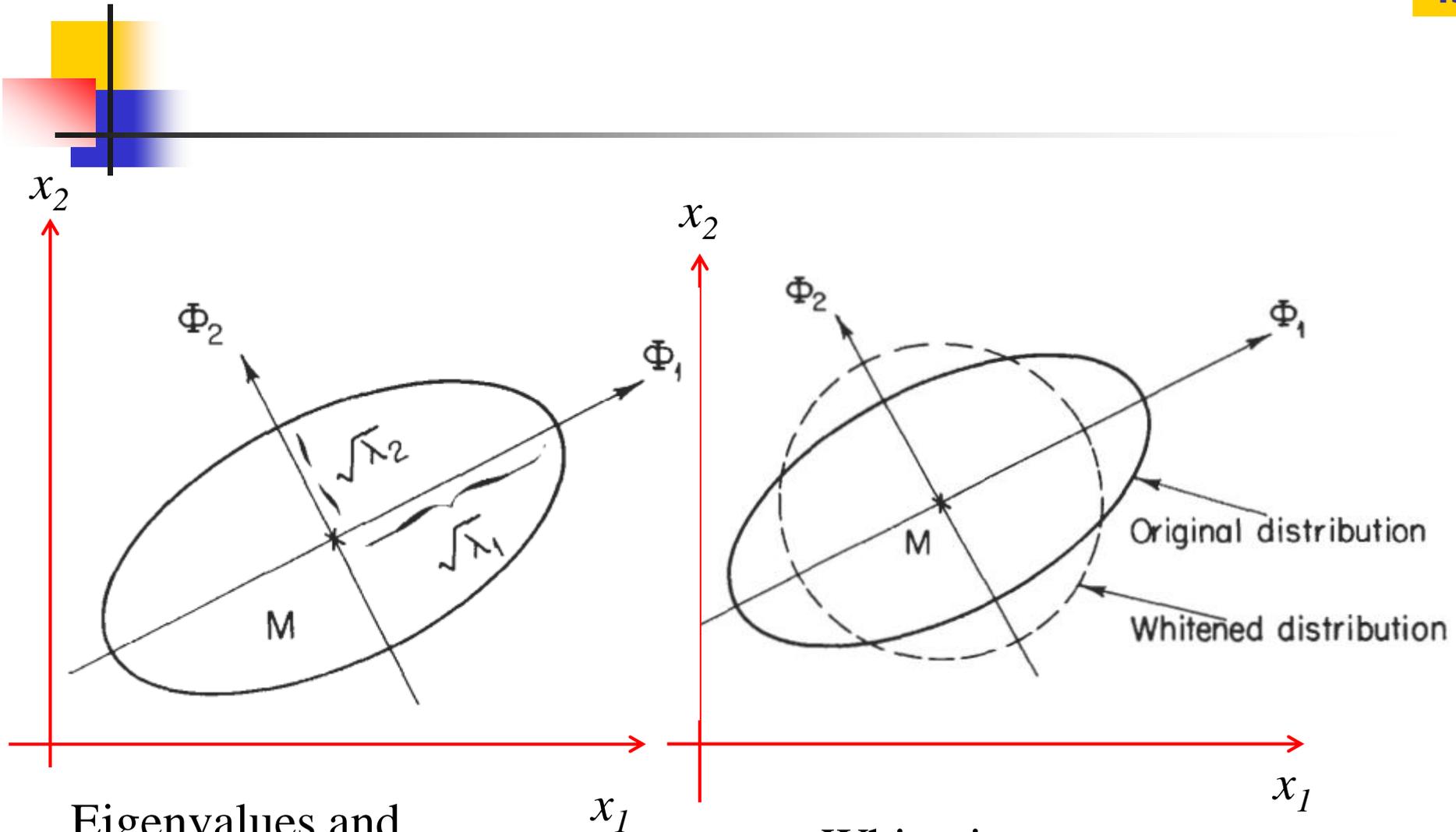
$$\Psi^t\mathbf{I}\Psi = \Psi^t\Psi = \mathbf{I}$$



**FIGURE 2.8.** The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation,  $\mathbf{A}$ , takes the source distribution into distribution  $N(\mathbf{A}^t \mu, \mathbf{A}^t \Sigma \mathbf{A})$ . Another linear transformation—a projection  $\mathbf{P}$  onto a line defined by vector  $\mathbf{a}$ —leads to  $N(\mu, \sigma^2)$  measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original  $x_1$ - $x_2$  space. A whitening transform,  $\mathbf{A}_w$ , leads to a circularly symmetric Gaussian, here shown displaced.

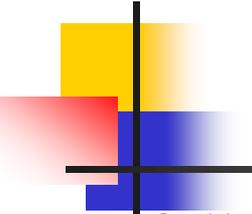


**FIGURE 2.9.** Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean. The ellipses show lines of equal probability density of the Gaussian.



Eigenvalues and  
eigenvectors of a  
distribution.

Whitening process



It follows from Multivariate normal density function that the loci of points of constant density are hyperellipsoids for which the quadratic form  $(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$  is constant. The principal axes of these hyperellipsoids are given by the eigenvectors of  $\boldsymbol{\Sigma}$  (described by  $\boldsymbol{\Phi}$ ); the eigenvalues (described by  $\boldsymbol{\Lambda}$ ) determine the lengths of these axes.

*Mahalanobis distance* from  $\mathbf{x}$  to  $\boldsymbol{\mu}$

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

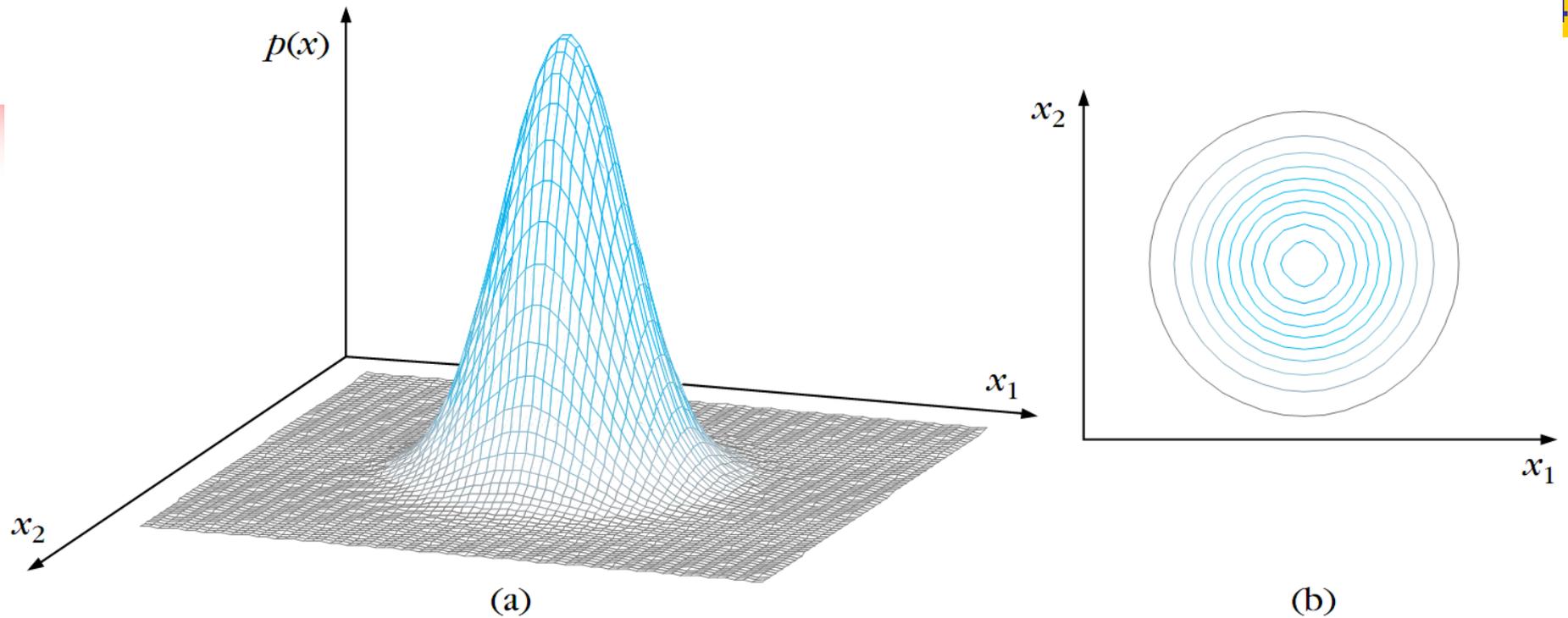
Thus, the contours of constant density are hyperellipsoids of constant Mahalanobis distance to  $\boldsymbol{\mu}$  and the volume of these hyperellipsoids measures the scatter of the samples about the mean.

- The volume of the hyperellipsoid corresponding to a Mahalanobis distance  $r$  is given by

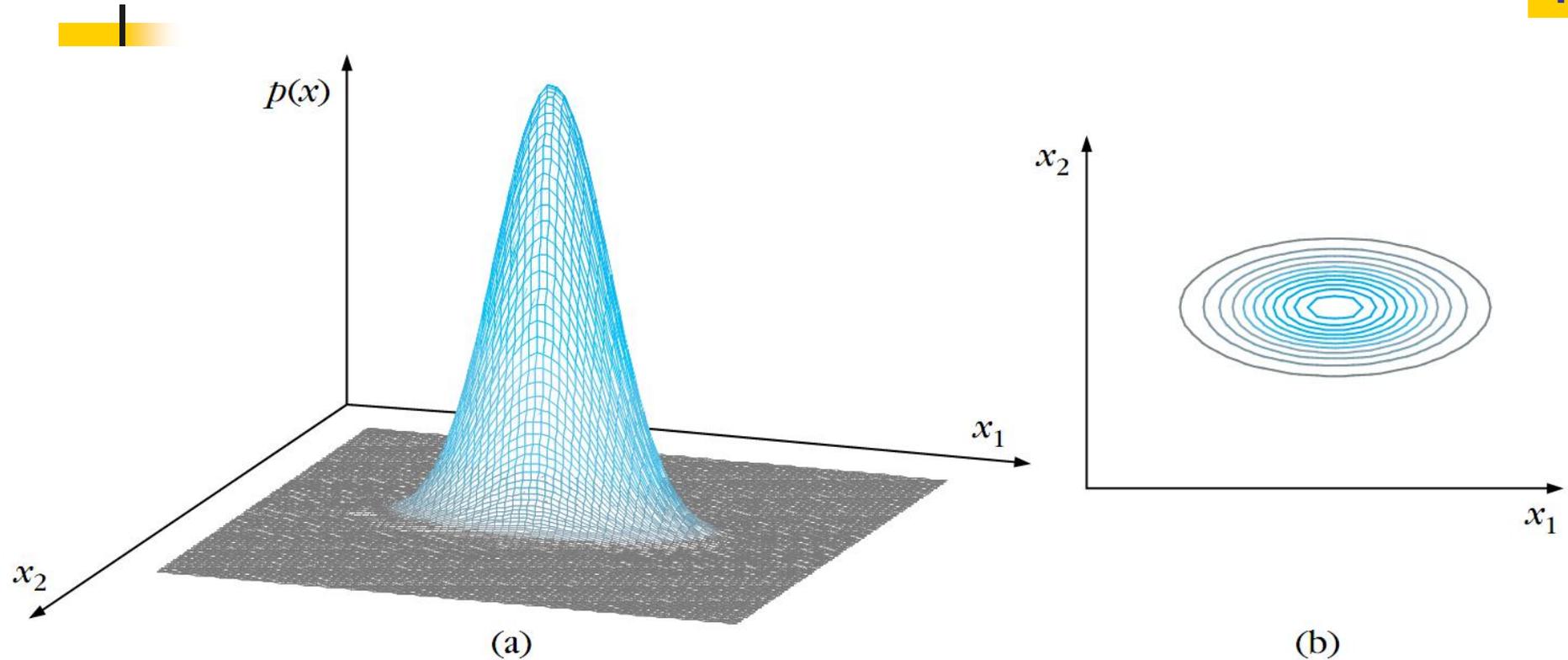
$$V = V_d |\Sigma|^{1/2} r^d,$$

where  $V_d$  is the volume of a  $d$ -dimensional unit hypersphere:

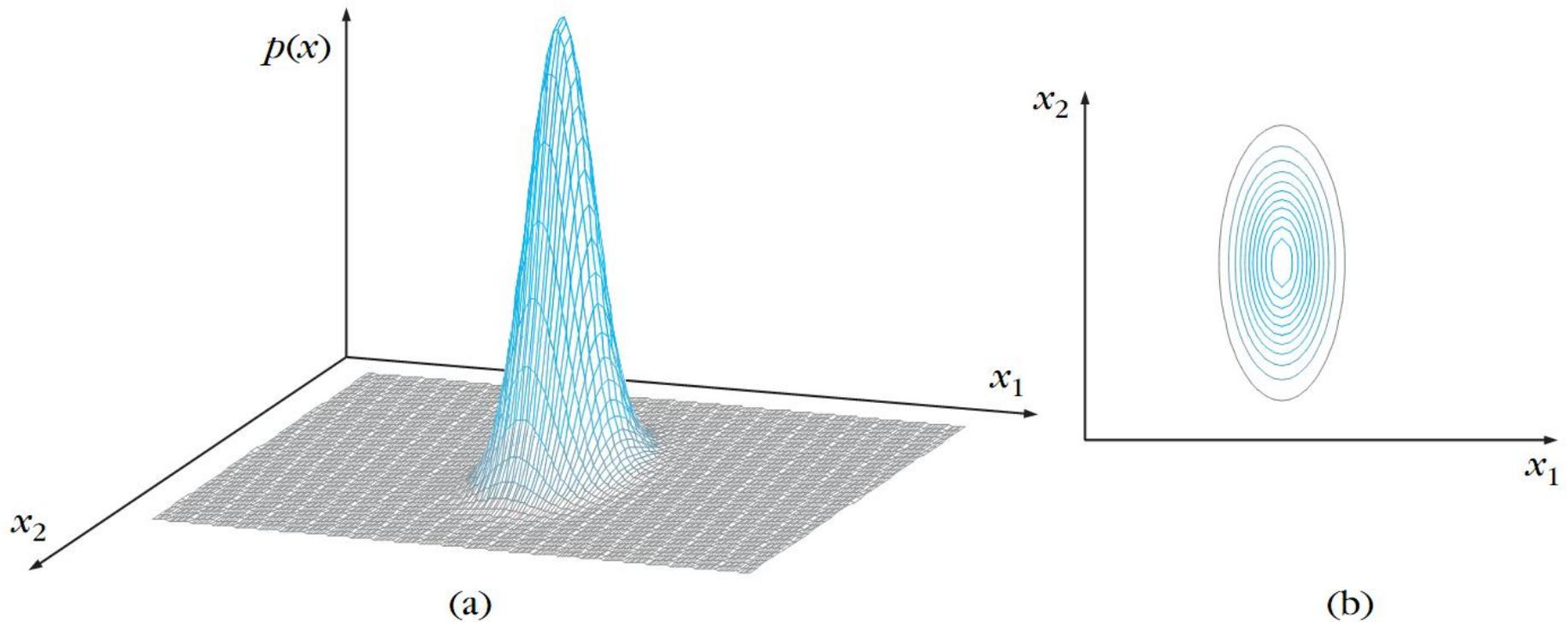
$$V_d = \begin{cases} \pi^{d/2} / (d/2)! & d \text{ even} \\ 2^d \pi^{(d-1)/2} \left(\frac{d-1}{2}\right)! / (d)! & d \text{ odd.} \end{cases}$$



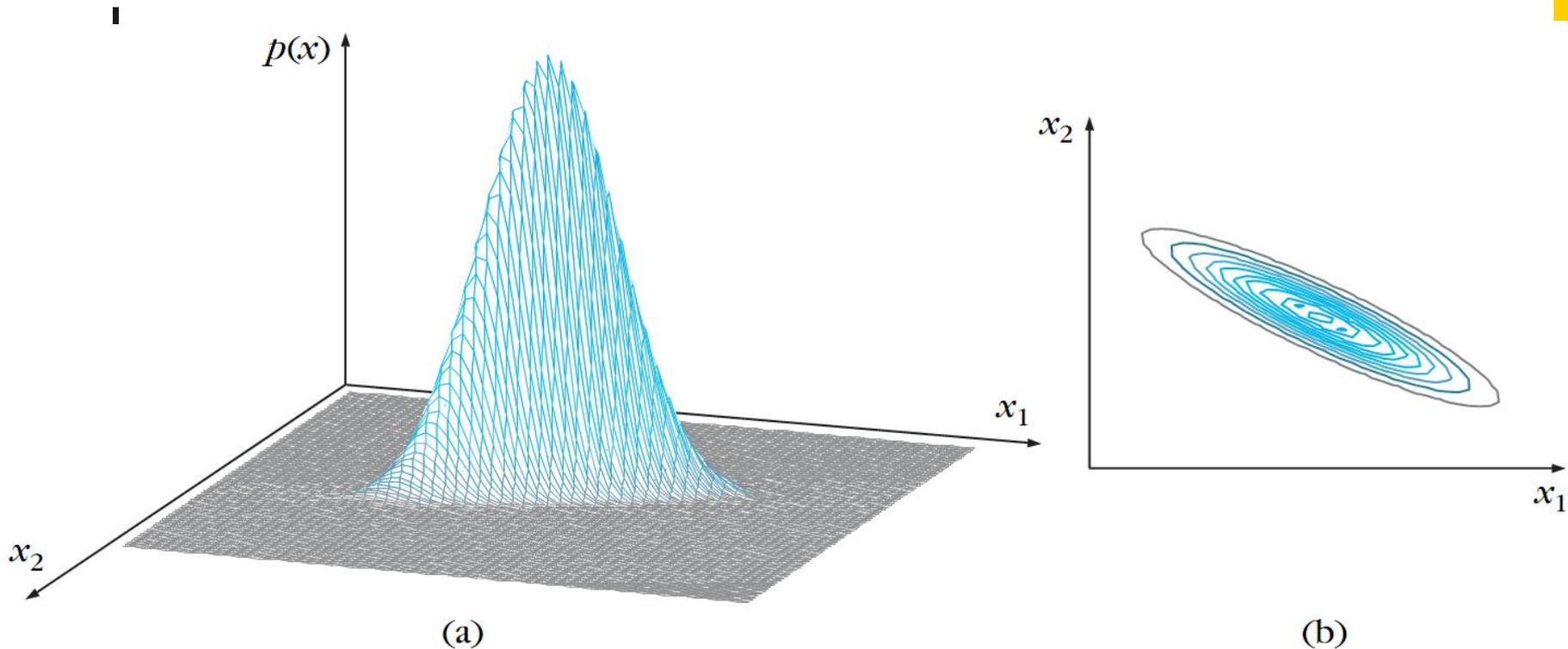
- (a) The graph of a two-dimensional Gaussian pdf and  
(b) the corresponding iso-value curves for a diagonal  $\Sigma$  with  $\sigma_1^2 = \sigma_2^2$ .  
The graph has a spherical symmetry showing no preference in any direction.



- (a) The graph of a two-dimensional Gaussian pdf and  
(b) the corresponding iso-value curves for a diagonal  $\Sigma$  with  $\sigma_1^2 \gg \sigma_2^2$ .  
The graph is elongated along the  $x_1$  direction.



- (a) The graph of a two-dimensional Gaussian pdf and  
(b) the corresponding iso-value curves for a diagonal  $\Sigma$  with  $\sigma_1^2 \ll \sigma_2^2$ .  
The graph is elongated along the  $x_2$  direction.



- (a) The graph of a two-dimensional Gaussian pdf and  
(b) the corresponding iso-value curves for a case of a nondiagonal  $\Sigma$ . Playing with the values of the elements of  $\Sigma$  one can achieve different shapes and orientations

# Sample generation

- To generate samples which are to be normally distributed according to a given expected vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .
- From the given  $\boldsymbol{\Sigma}$ , find the whitening transformation of  $\mathbf{y} = \boldsymbol{\Lambda}^{-1/2} \boldsymbol{\Phi}^t \mathbf{x}$ . In the transformed space,  $\boldsymbol{\Sigma}_y = \mathbf{I}$ .
- Generate  $N$  independent, normally distributed numbers for each  $y_i$  ( $i=1, \dots, n$ ) with zero expected value and unit variance. Then, form  $N$  vectors  $y_1, y_2, \dots, y_N$ .
- Transform back the generated samples to the  $\mathbf{x}$ -space by
$$\mathbf{x}_k = \boldsymbol{\Phi} \boldsymbol{\Lambda}^{1/2} \mathbf{y}_k \quad (k = 1, \dots, N).$$
- Add  $\boldsymbol{\mu}$  to the samples in the  $\mathbf{x}$ -space as  $\mathbf{x}_k + \boldsymbol{\mu}$  ( $k=1, \dots, N$ ).