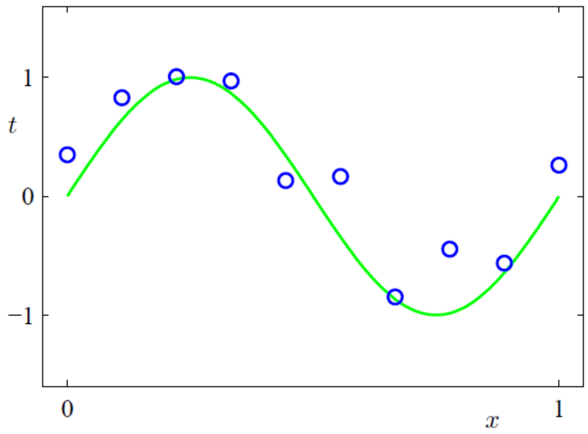
# 02457 Non-linear signal processing

## 2017 - Lecture 3

**Lars Kai Hansen**
**Technical University of Denmark**
DTU Compute, Lyngby, Denmark

# Outline lecture 3

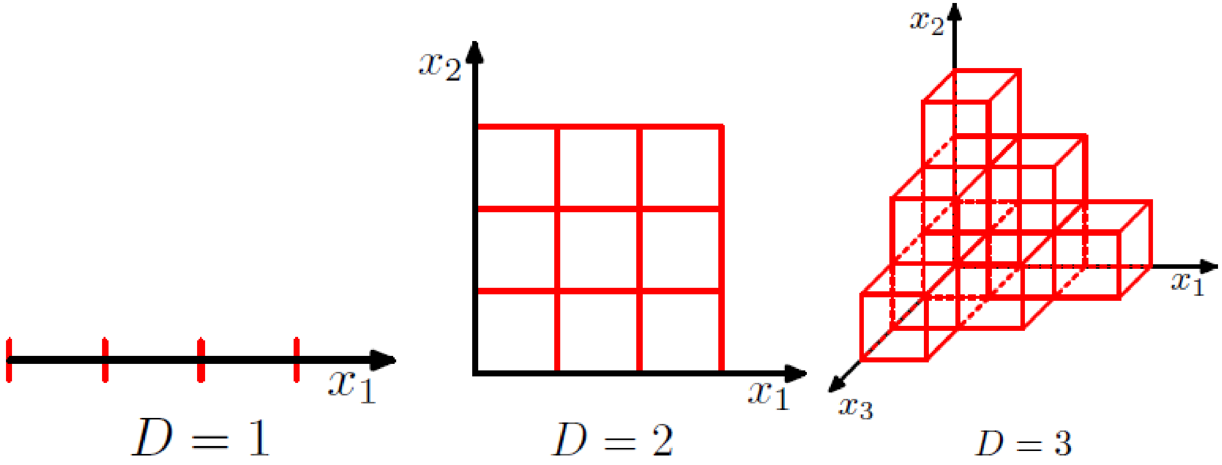


- Multivariate normal distribution
- Matrices, eigenvalues, eigenvectors
- Principal components
- Principal components of functional brain images

- Learning problem
- The likelihood function
- Least squares
- Linear regression
- Linear discriminants

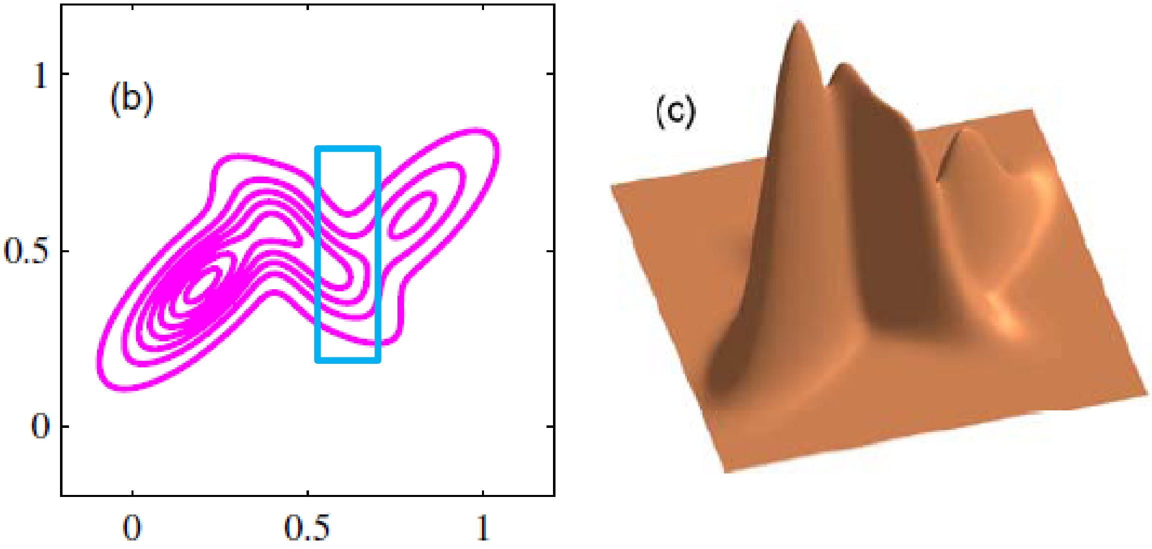
# Features and the curse of dimensionality

- To specify a map (e.g. a discriminant function) on a $d$-dimensional space by dividing the relevant parts of the this space into $L$ cells pr. dimension requires $L^d$ cells.



# Multivariate processes

$$P(x_j \in [a_j, b_j] | j = 1, .., d) = \int_{a_1}^{b_1} ... \int_{a_d}^{b_d} p(\mathbf{x}) d\mathbf{x}$$



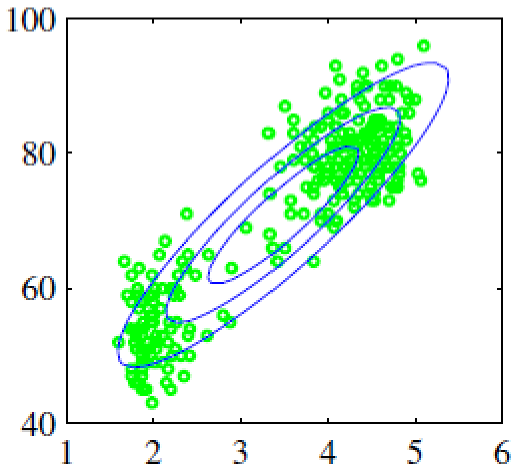
# Dependency

$$\mathbb{E}_{\mathbf{x}}\{f(\mathbf{x})\} = \int_{\text{Domain of } \mathbf{x}} f(\mathbf{x})p(\mathbf{x})d\mathbf{x} \qquad \mu_j = \mathbb{E}_{\mathbf{x}}\{x_j\}$$

$$\boxed{\boldsymbol{\Sigma}_{j,k} = \mathrm{cov}(x_j, x_k) = \mathbb{E}_{\mathbf{x}}\{(x_j - \mu_j)(x_k - \mu_k)\}}$$

$$\boldsymbol{\Sigma} = \int_{\text{Domain of } \mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x})d\mathbf{x}$$

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$



# Multivariate normal distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

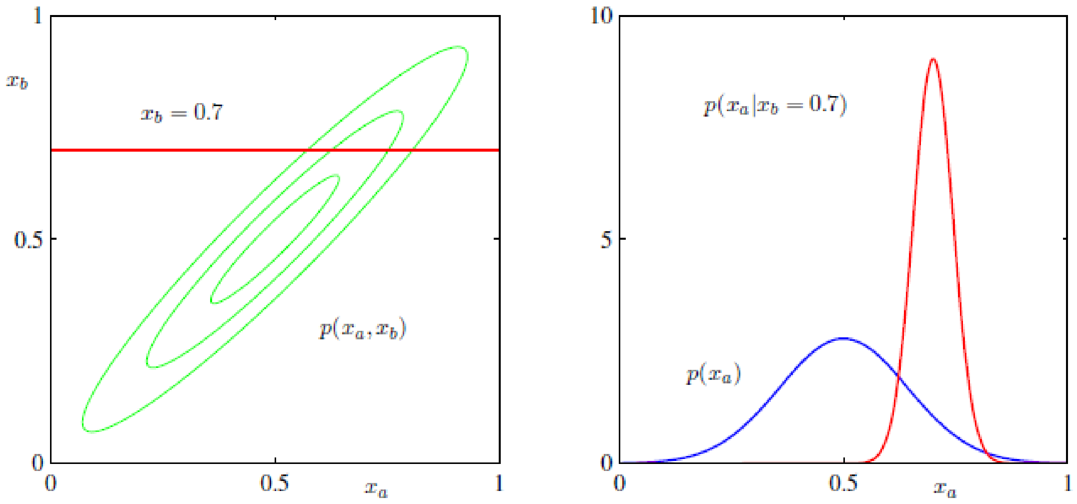


**Figure 2.9** The plot on the left shows the contours of a Gaussian distribution $p(x_a, x_b)$ over two variables, and the plot on the right shows the marginal distribution $p(x_a)$ (blue curve) and the conditional distribution $p(x_a|x_b)$ for $x_b = 0.7$ (red curve).

# Matrices recap

Appendix C. Properties of Matrices

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}} \qquad \mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \qquad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}.$$

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA}).$$

$$|\mathbf{A}| = \sum (\pm 1) A_{1i_1} A_{2i_2} \cdots A_{Ni_N} \qquad |\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

$$|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}| \qquad |\mathbf{A}^{-1}| = \frac{1}{|\mathbf{A}|}$$

# Real symmetric matrices recap

Appendix C. Properties of Matrices

$$\mathbf{A}\mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad |\mathbf{A} - \lambda_i \mathbf{I}| = 0 \qquad \mathbf{A} = \sum_{i=1}^{M} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

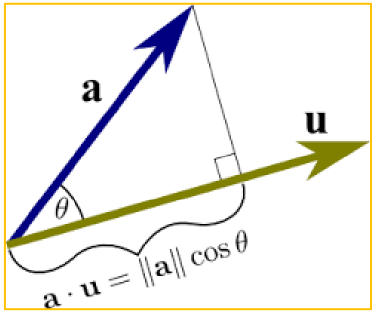
$$\mathbf{u}_i^{\mathrm{T}} \mathbf{u}_j = I_{ij} \qquad \mathbf{A}^{-1} = \sum_{i=1}^{M} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}.$$

We can take the eigenvectors $\mathbf{u}_i$ to be the columns of an $M \times M$ matrix $\mathbf{U}$, which from orthonormality satisfies

$$\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}. \tag{C.37}$$

$$\mathbf{A}\mathbf{U} = \mathbf{U}\boldsymbol{\Lambda} \qquad \mathbf{U}^{\mathrm{T}}\mathbf{A}\mathbf{U} = \boldsymbol{\Lambda}$$

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}}$$



# Analysis of the covariance matrix

**Figure 2.7** The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues $\lambda_i$.

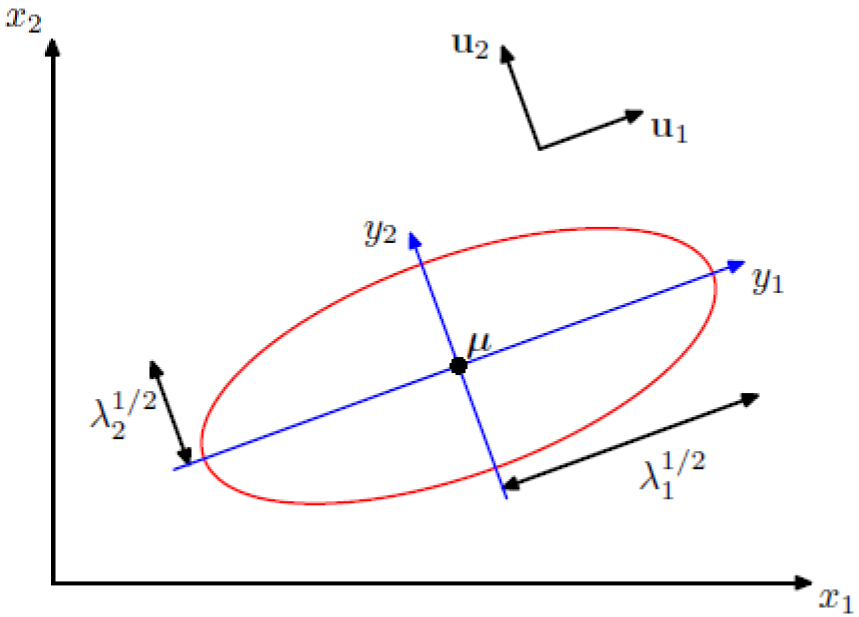
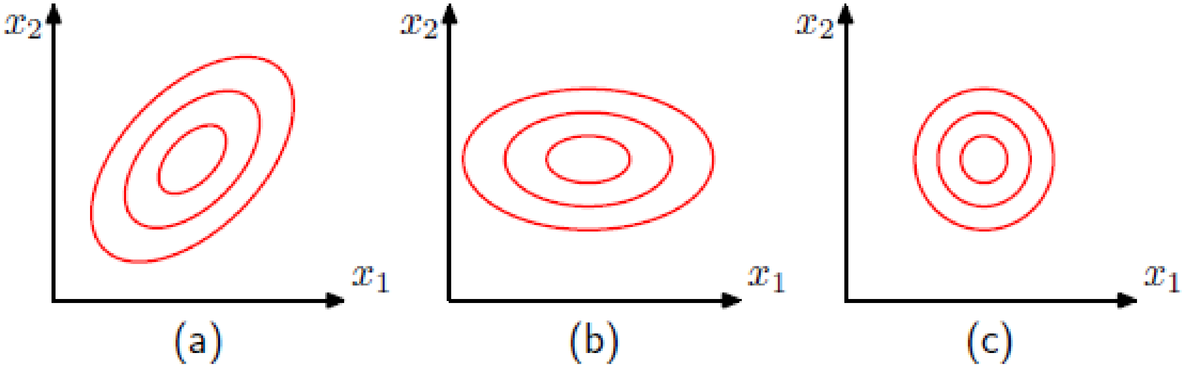


**Figure 2.8** Contours of constant probability density for a Gaussian distribution in two dimensions in which the covariance matrix is (a) of general form, (b) diagonal, in which the elliptical contours are aligned with the coordinate axes, and (c) proportional to the identity matrix, in which the contours are concentric circles.



# Expectations in the normal distribution

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}\end{aligned}$$

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

# Expectations in th normal distribution

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\mathbf{x}^{\mathrm{T}}\,\mathrm{d}\mathbf{x}$$

$$= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^{\mathrm{T}}\,\mathrm{d}\mathbf{z}$$

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}.$$

# Principal components (Bishop chapter 12)

Consider a data set of observations $\{\mathbf{x}_n\}$ where $n = 1, \ldots, N$

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n$$

Variance of projection on unit vector

$$\frac{1}{N}\sum_{n=1}^{N} \left\{\mathbf{u}_1^{\mathrm{T}}\mathbf{x}_n - \mathbf{u}_1^{\mathrm{T}}\overline{\mathbf{x}}\right\}^2 = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$$

where $\mathbf{S}$ is the data covariance matrix defined by

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \overline{\mathbf{x}})(\mathbf{x}_n - \overline{\mathbf{x}})^{\mathrm{T}}.$$

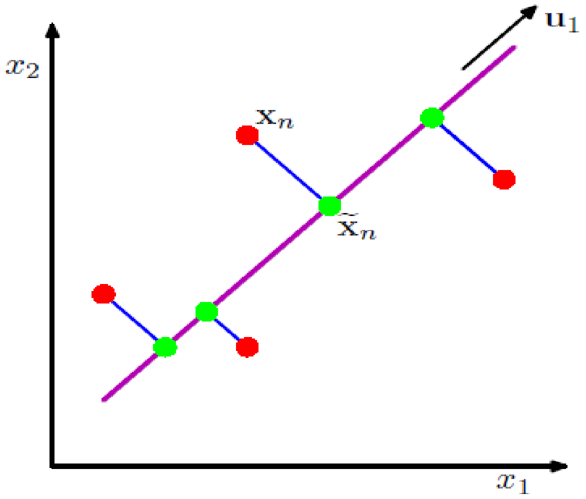
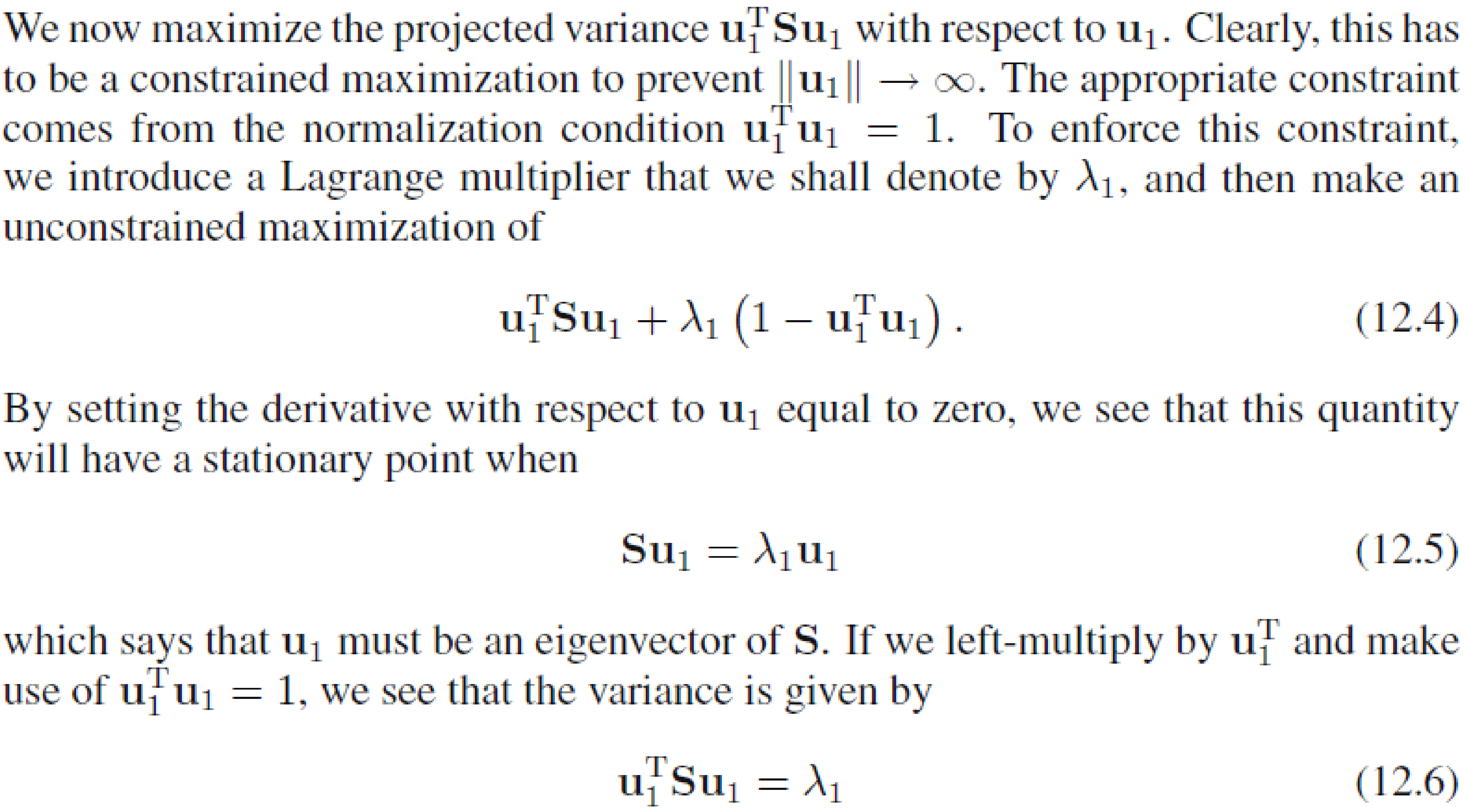


**Figure 12.2** Principal component analysis seeks a space of lower dimensionality, known as the principal subspace and denoted by the magenta line, such that the orthogonal projection of the data points (red dots) onto this subspace maximizes the variance of the projected points (green dots). An alternative definition of PCA is based on minimizing the sum-of-squares of the projection errors, indicated by the blue lines.

# Principal components



We now maximize the projected variance $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$ with respect to $\mathbf{u}_1$. Clearly, this has to be a constrained maximization to prevent $\|\mathbf{u}_1\| \to \infty$. The appropriate constraint comes from the normalization condition $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$. To enforce this constraint, we introduce a Lagrange multiplier that we shall denote by $\lambda_1$, and then make an unconstrained maximization of

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 + \lambda_1\left(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1\right). \tag{12.4}$$

By setting the derivative with respect to $\mathbf{u}_1$ equal to zero, we see that this quantity will have a stationary point when

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \tag{12.5}$$

which says that $\mathbf{u}_1$ must be an eigenvector of $\mathbf{S}$. If we left-multiply by $\mathbf{u}_1^{\mathrm{T}}$ and make use of $\mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1$, we see that the variance is given by

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1 \tag{12.6}$$

# Principal components of handwritten digits

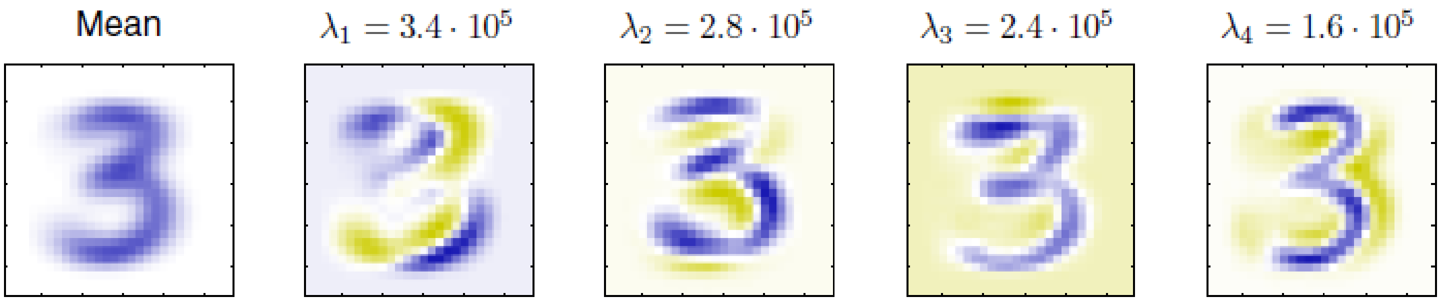


**Figure 12.3** The mean vector $\overline{\mathbf{x}}$ along with the first four PCA eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_4$ for the off-line digits data set, together with the corresponding eigenvalues.

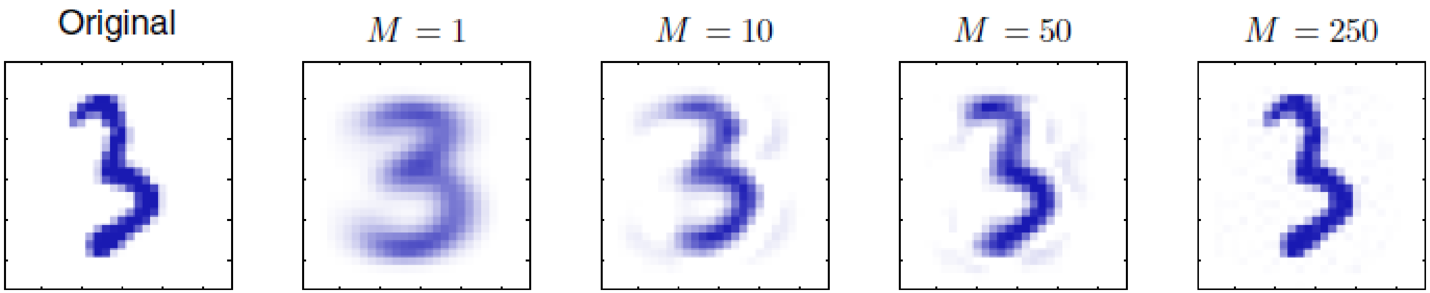
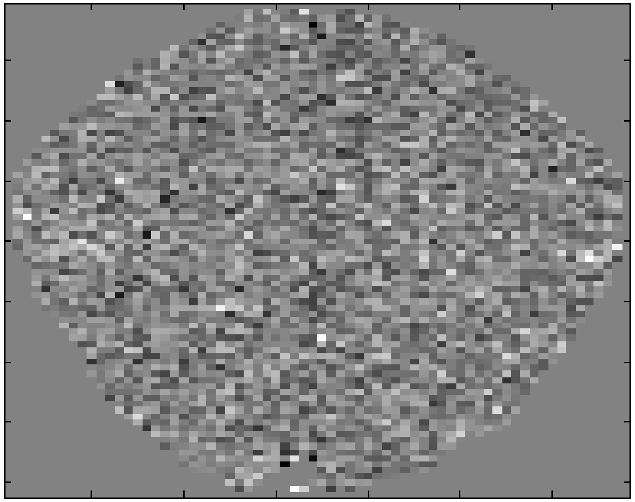


**Figure 12.5** An original example from the off-line digits data set together with its PCA reconstructions obtained by retaining $M$ principal components for various values of $M$. As $M$ increases the reconstruction becomes more accurate and would become perfect when $M = D = 28 \times 28 = 784$.

# Functional Magnetic Resonance Imaging

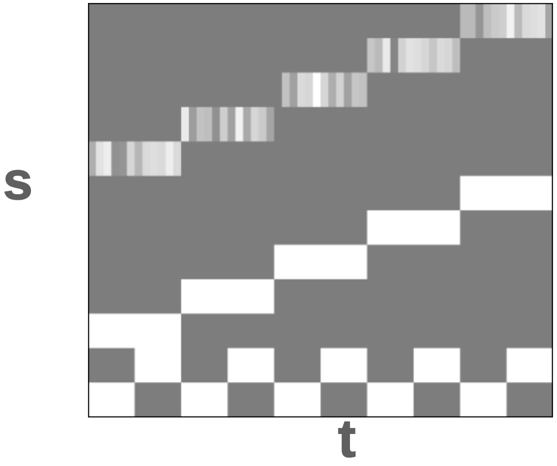
- Indirect measure of neural activity – hemodynamics
- Invented in 1992
- A cloudy window to the human brain
- Challenges:
  - Signals are multi-dimensional mixtures
  - No simple relation between measures and brain state - ”what is signal and what is noise”?



TR = 333 ms

# Multivariate neuroimaging models

Neuroimaging aims at extracting the mutual information between stimulus and response.



- Stimulus: Macroscopic variables, ”design matrix” … *s(t)*

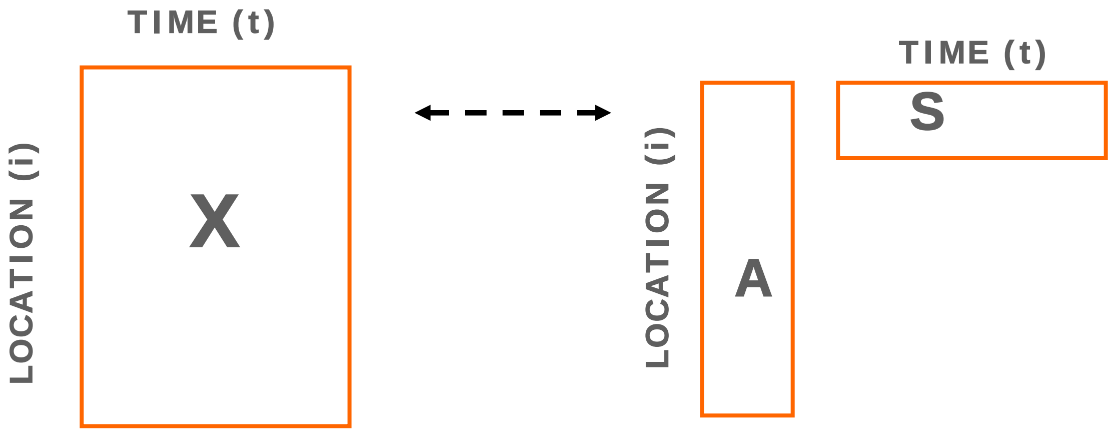
- Response: Micro/meso-scopic variables, the neuroimage … *x(t)*

- Mutual information is stored in the joint distribution … *p(x,s).*

*Often s(t) is assumed known….unsupervised methods consider s(t) or parts of s(t) ”hidden”…..*

# Factor models

Represent a datamatrix by a low-dimensional approximation



$$X(i,t) \approx \sum_{k=1}^{K} A(i,k)S(k,t)$$

# PCA of fMRI

NeuroImage **9**, 534–544 (1999)
Article ID nimg.1998.0425, available online at http://www.idealibrary.com on IDEAL®

## Generalizable Patterns in Neuroimaging: How Many Principal Components?

Lars Kai Hansen,*[,1] Jan Larsen,* Finn Årup Nielsen,* Stephen C. Strother,†,|| Egill Rostrup,‡ Robert Savoy,§
Nicholas Lange,¶ John Sidtis,|| Claus Svarer,** and Olaf B. Paulson**

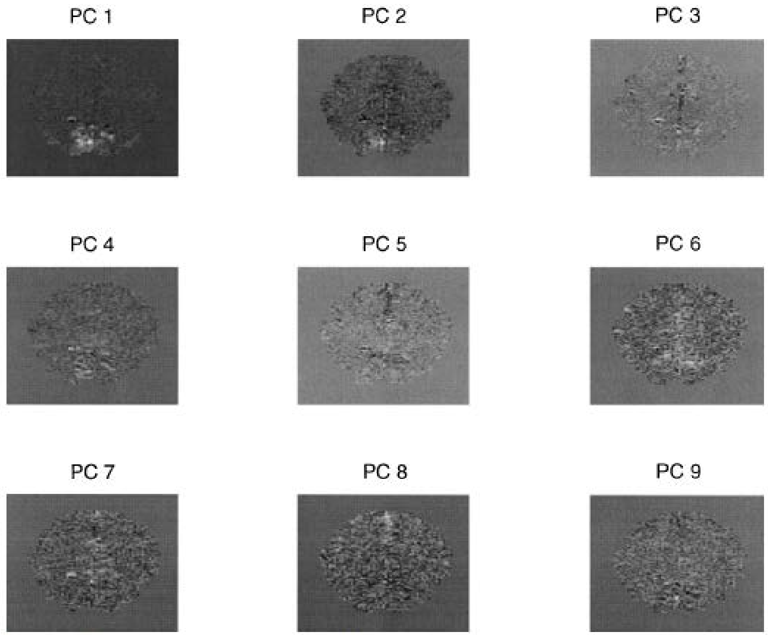


**FIG. 6.** Data set II. Covariance eigenimages corresponding to the nine most significant principal components. The eigenimage corresponding to the dominating first PC is focused in visual areas. Using the bias/variance trade-off curves in Fig. 4, we find that only the eigenimages corresponding to the first three components generalize.

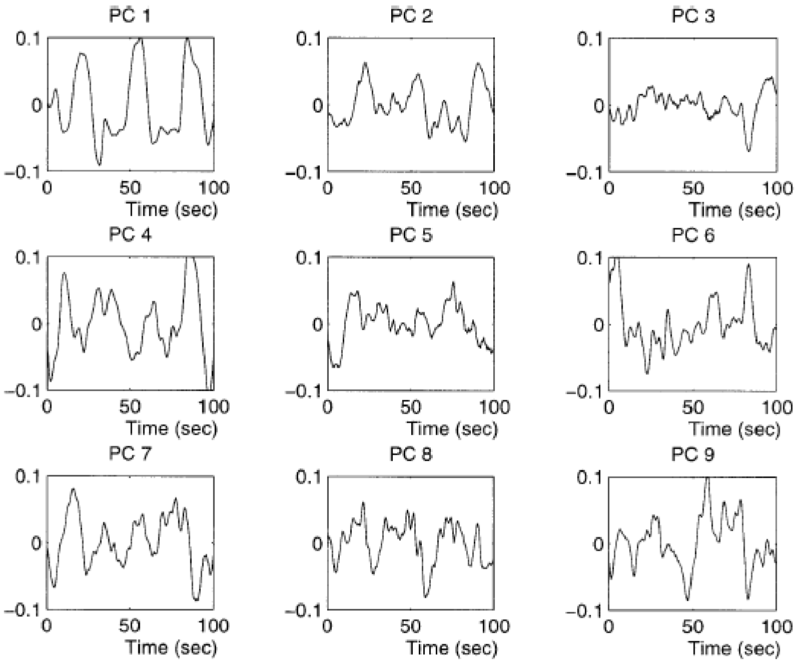
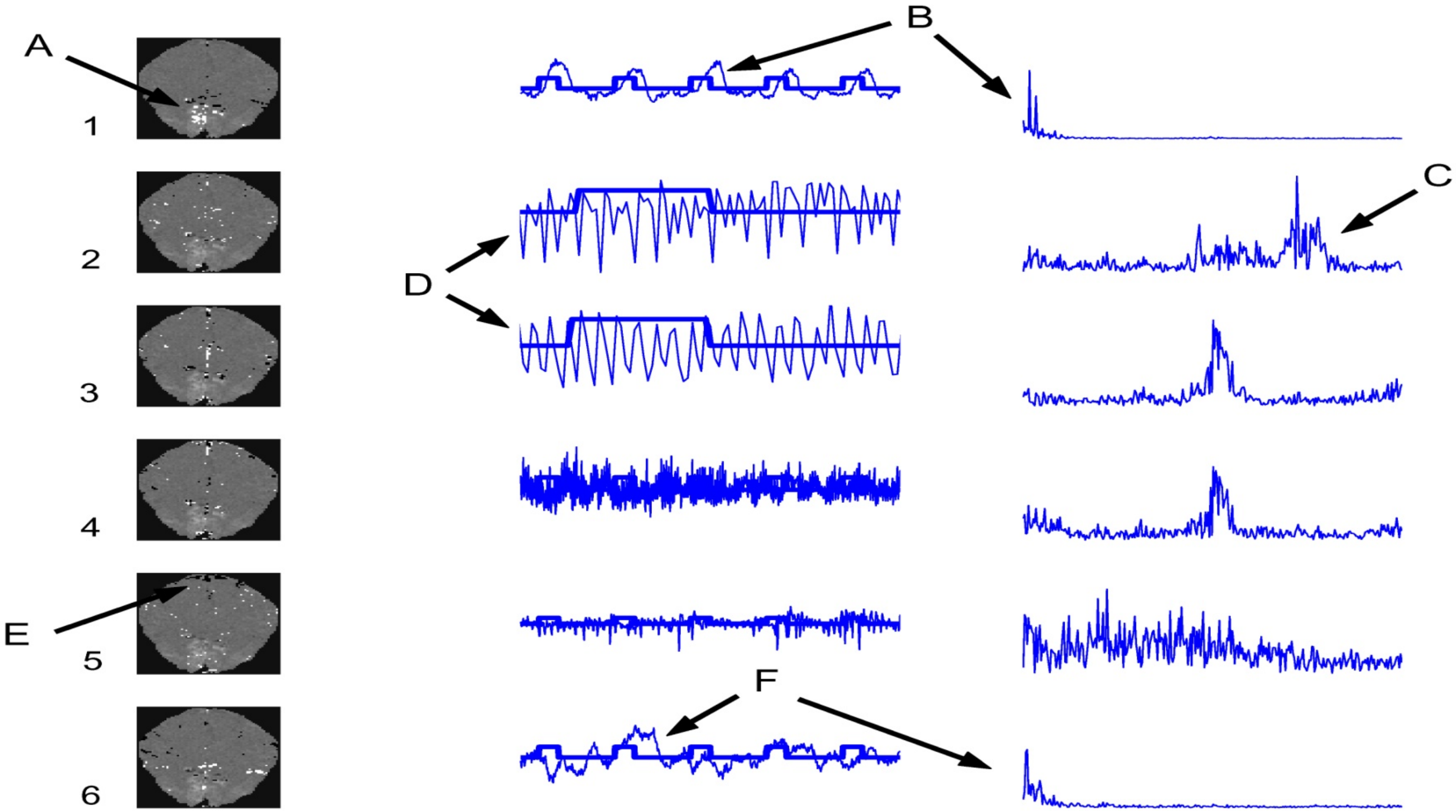


**FIG. 5.** Data set II. The principal components corresponding to the nine largest covariance eigenvalues for a sequence of 300 fMRI scans of a visually stimulated subject. Stimulation takes places at scan times $\tau = 8–25$ s, $\tau = 42–59$ s, and $\tau = 75–92$ s relative to the start of this three-run sequence. Scan time sampling interval was TR = 0.33 s. The sequences have been smoothed for presentation, reducing noise and high-frequency physiological components. Note that most of the response is captured by the first principal component, showing a strong response to all three periods of stimulation. Using the generalization error estimates in Fig. 4, we find that only the time sequences corresponding to the first three components generalize.

# ICA: Assume S(k,t), S(k’,t) statistically independent

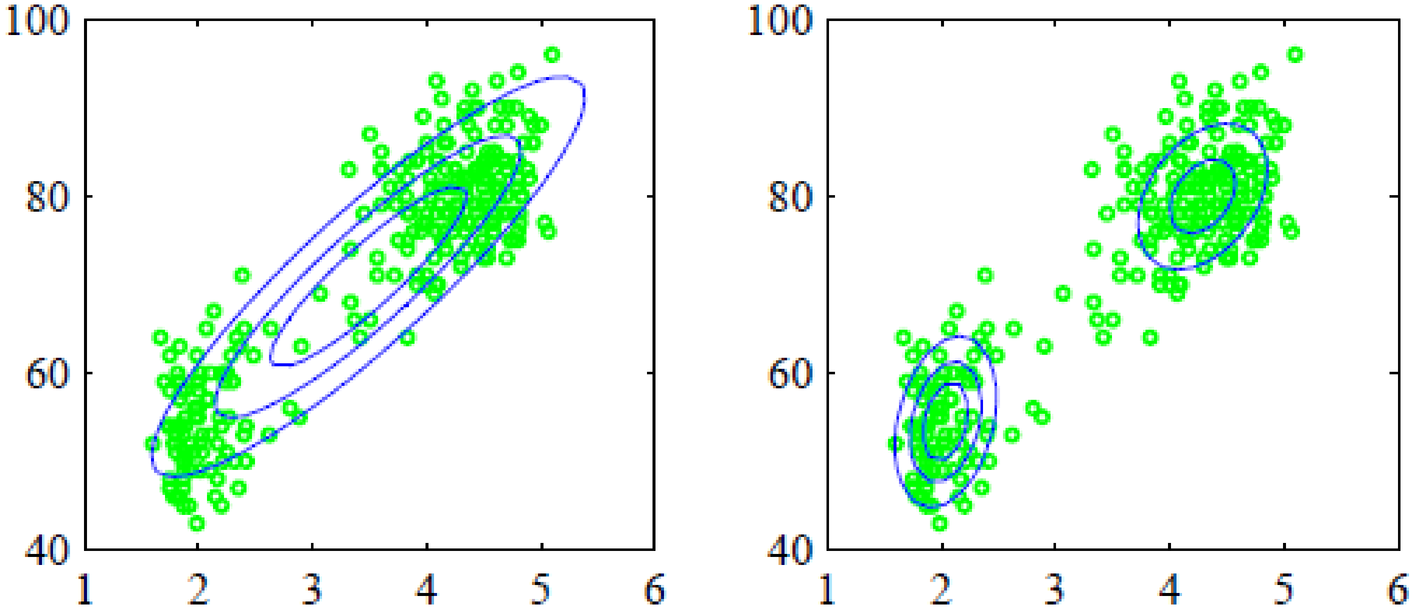


McKeown, Hansen, Sejnowski, Curr. Op. in Neurobiology (2003)

# Statistical machine learning

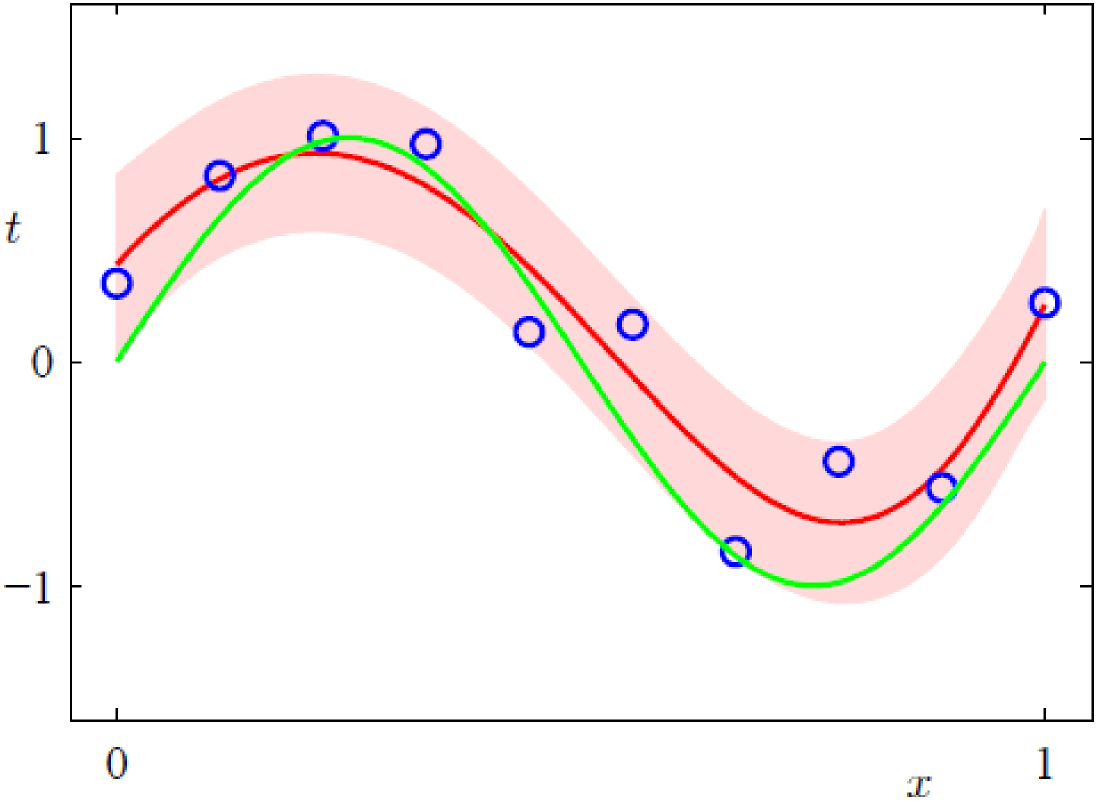
# Unsupervised learning

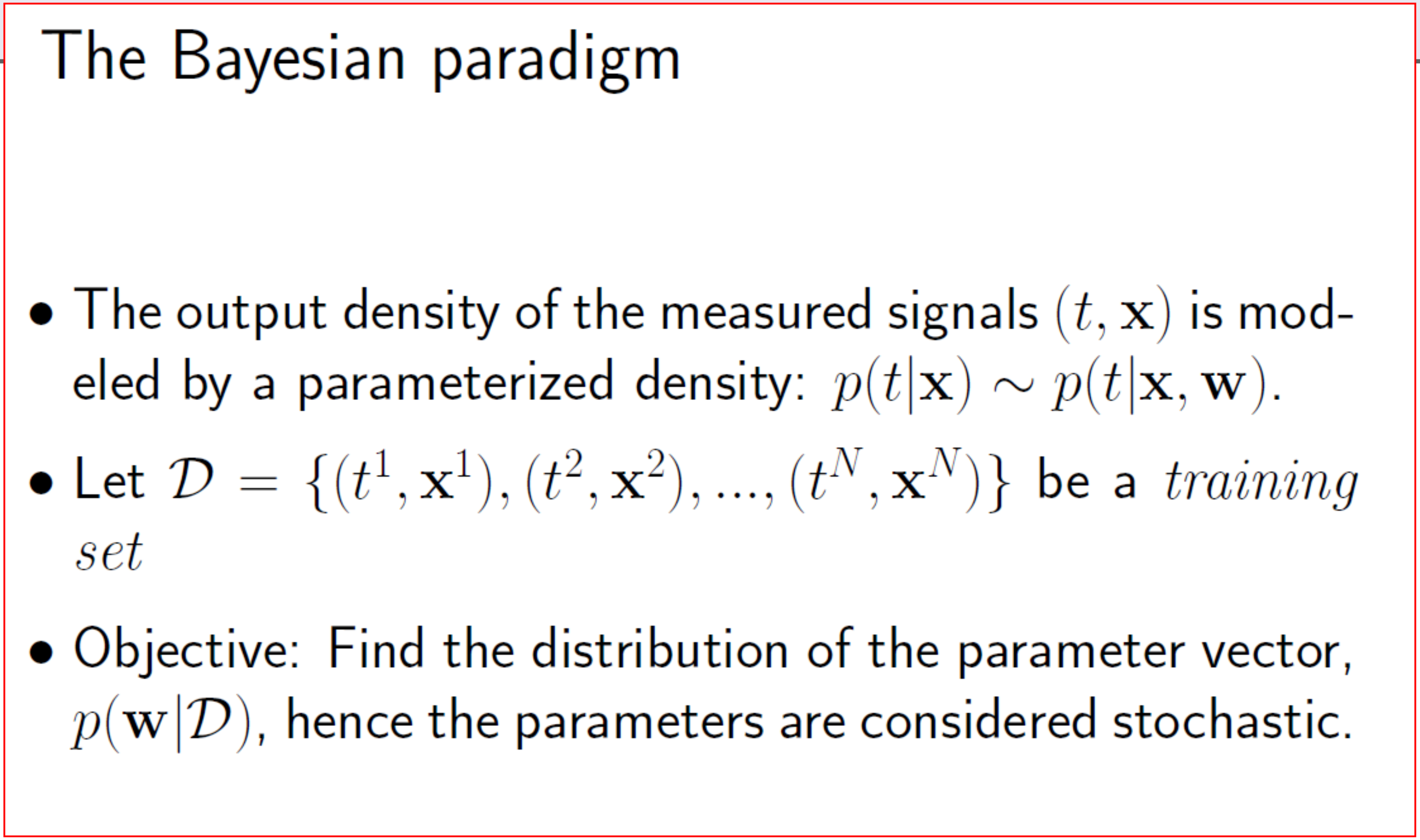
Unsupervised learning: Learning the distribution of a set of variables $p(\text{input})$.



# Supervised learning

- Supervised learning: Learning relations between sets of variables e.g. between input and output variables, conditional distributions $p(\mathrm{output}|\mathrm{input})$.





## The Bayesian paradigm

- The output density of the measured signals $(t, \mathbf{x})$ is modeled by a parameterized density: $p(t|\mathbf{x}) \sim p(t|\mathbf{x}, \mathbf{w})$.
- Let $\mathcal{D} = \{(t^1, \mathbf{x}^1), (t^2, \mathbf{x}^2), ..., (t^N, \mathbf{x}^N)\}$ be a *training set*
- Objective: Find the distribution of the parameter vector, $p(\mathbf{w}|\mathcal{D})$, hence the parameters are considered stochastic.

# The likelihood function

- Let $\mathcal{D} = \{(t^1, \mathbf{x}^1), (t^2, \mathbf{x}^2), ..., (t^N, \mathbf{x}^N)\}$ be the *training set*
- We use Bayes theorem

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

- The function $p(\mathcal{D}|\mathbf{w})$ is called the likelihood function (more correct the likelihood of the parameter vector $\mathbf{w}$). The density $p(\mathbf{w})$ is called the *a priori* or *prior* parameter distribution.

# The likelihood function

$$\boxed{p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}}$$

- If the prior is “flat” in the neighborhood of the peak of $p(\mathcal{D}|\mathbf{w})$, we have

$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w})$$

- ...and finding the most probable parameters is equivalent to finding the maximum likelihood parameters.

# Maximum likelihood and optimization

- For independent examples, $\mathcal{D} = \{(t^1, \mathbf{x}^1), (t^2, \mathbf{x}^2), ..., (t^N, \mathbf{x}^N)\}$, the likelihood function factorizes

$$p(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^{N} p(t^n|\mathbf{x}^n, \mathbf{w})p(\mathbf{x}^n) = p(\mathcal{D}_t|\mathcal{D}_{\mathrm{x}}, \mathbf{w}) * p(\mathcal{D}_{\mathrm{x}})$$

- Many algorithms are based on minimizing an index or cost function

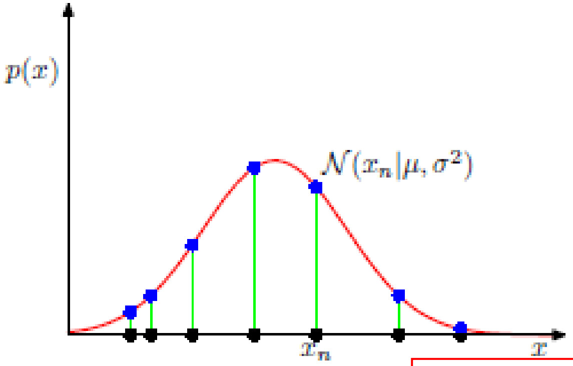
$$E(\mathbf{w}) = -\log p(\mathcal{D}_t|\mathcal{D}_{\mathrm{x}}, \mathbf{w}) = \sum_{n=1}^{N} -\log p(t^n|\mathbf{x}^n\mathbf{w})$$

# Maximum likelihood learning of mean and variance

Let the parameterized density be a 1D normal distribution

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

For independent examples, $\chi = \{x_1, x_2, x_3, ..., x_N\}$, the likelihood function becomes



$$p(\mathcal{D}|\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right)$$

$$E(\mu,\sigma^2) = \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

# Maximum likelihood mean and variance estimators

$$\frac{\partial E(\mu,\sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{n=1}^{N} -(x_n-\mu)$$

$$\frac{\partial E(\mu,\sigma^2)}{\partial \sigma^2} = \frac{N}{2}\frac{1}{\sigma^2} - \frac{1}{2(\sigma^2)^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

$$0 = \frac{1}{\widehat{\sigma^2}}\sum_{n=1}^{N} -(x_n-\widehat{\mu})$$

$$0 = \frac{N}{2}\frac{1}{\widehat{\sigma^2}} - \frac{1}{2(\widehat{\sigma^2})^2}\sum_{n=1}^{N}(x_n-\widehat{\mu})^2$$

$$\widehat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\widehat{\mu})^2$$

# Maximum likelihood for the multivariate Normal distribution

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).$$

$$\frac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

# Maximum likelihood and least squares

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu}).$$

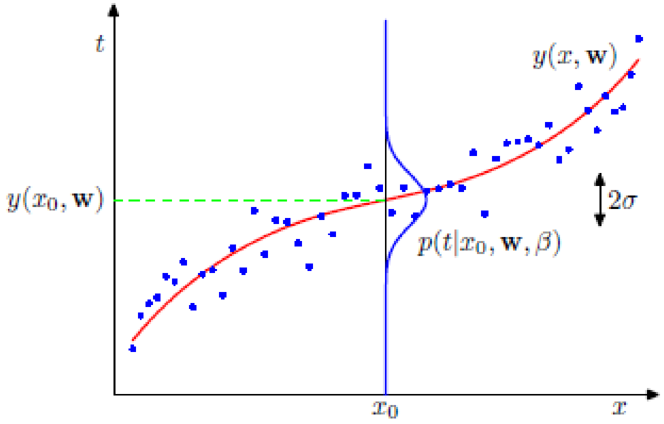
$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

$$\begin{aligned}\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu}\\ \mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}.\end{aligned}$$

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

# Maximum likelihood and least squares

$$
\begin{aligned}
t &= y(\mathbf{x};\mathbf{w}) + \nu \\
p(t|\mathbf{x},\sigma^2,\mathbf{w}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t - y(\mathbf{x};\mathbf{w}))^2\right) \\
p(\mathcal{D}_t|\mathcal{D}_\mathbf{x},\sigma^2,\mathbf{w}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(t^n - y(\mathbf{x}^n;\mathbf{w}))^2\right) \\
E(\mathbf{w},\sigma^2) &= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t^n - y(\mathbf{x}^n;\mathbf{w}))^2
\end{aligned}
$$

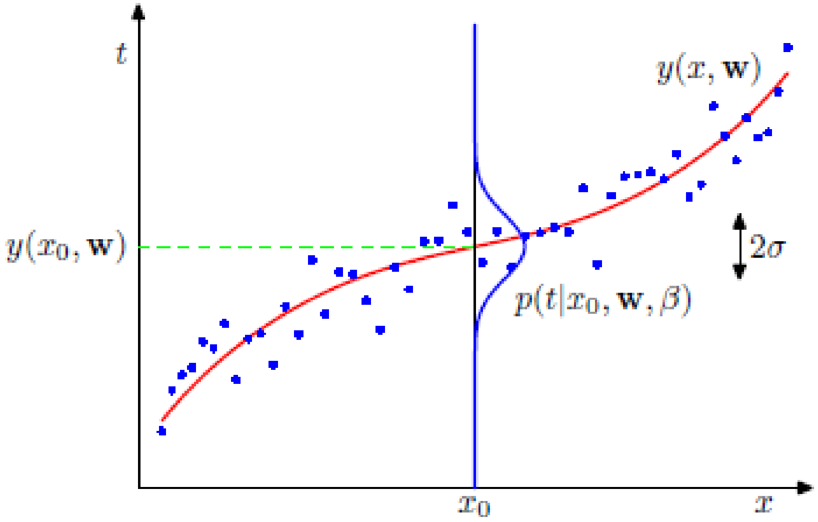


- Hence, maximizing the likelihood for Gaussian noise leads to a least squares problem (for $\mathbf{w}$).
- Note, the noise variance is given by ?

# Estimate of noise variance in least squares

$$E(\mathbf{w},\sigma^2) \;=\; \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t^n - y(\mathbf{x}^n;\mathbf{w}))^2$$

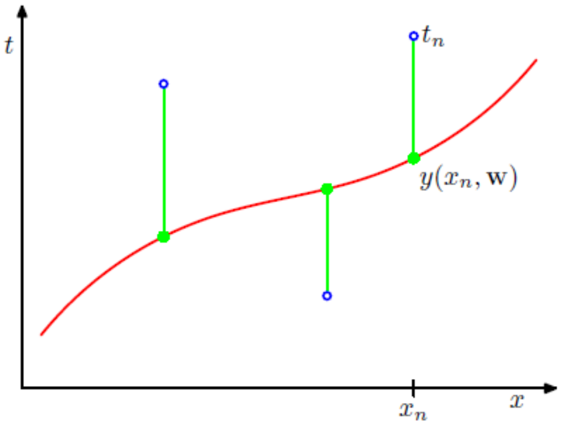
$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n=1}^{N}(t^n - y(\mathbf{x}^n;\mathbf{w}))^2$$



**Too optimistic (variance too small)!**

# Linear regression

Figure 1.3 The error function (1.2) corresponds to (one half of) the sum of the squares of the displacements (shown by the vertical green bars) of each data point from the function $y(x, \mathbf{w})$.



- Let the function be linear

$$y(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

- Let a training set be given by $\mathcal{D} = \{(t^1, \mathbf{x}^1), ..., (t^N, \mathbf{x}^N)\}$, the sum-of-squares approximation error is given by

$$E = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^n + w_0 - t^n)^2$$

- The optimal parameters are found by gradient based minimization,

$$\frac{\partial E}{\partial \mathbf{w}} = \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^n + w_0 - t^n) \mathbf{x}^n$$

$$\frac{\partial E}{\partial w_0} = \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^n + w_0 - t^n)$$

# Linear regresion

- equations to solve

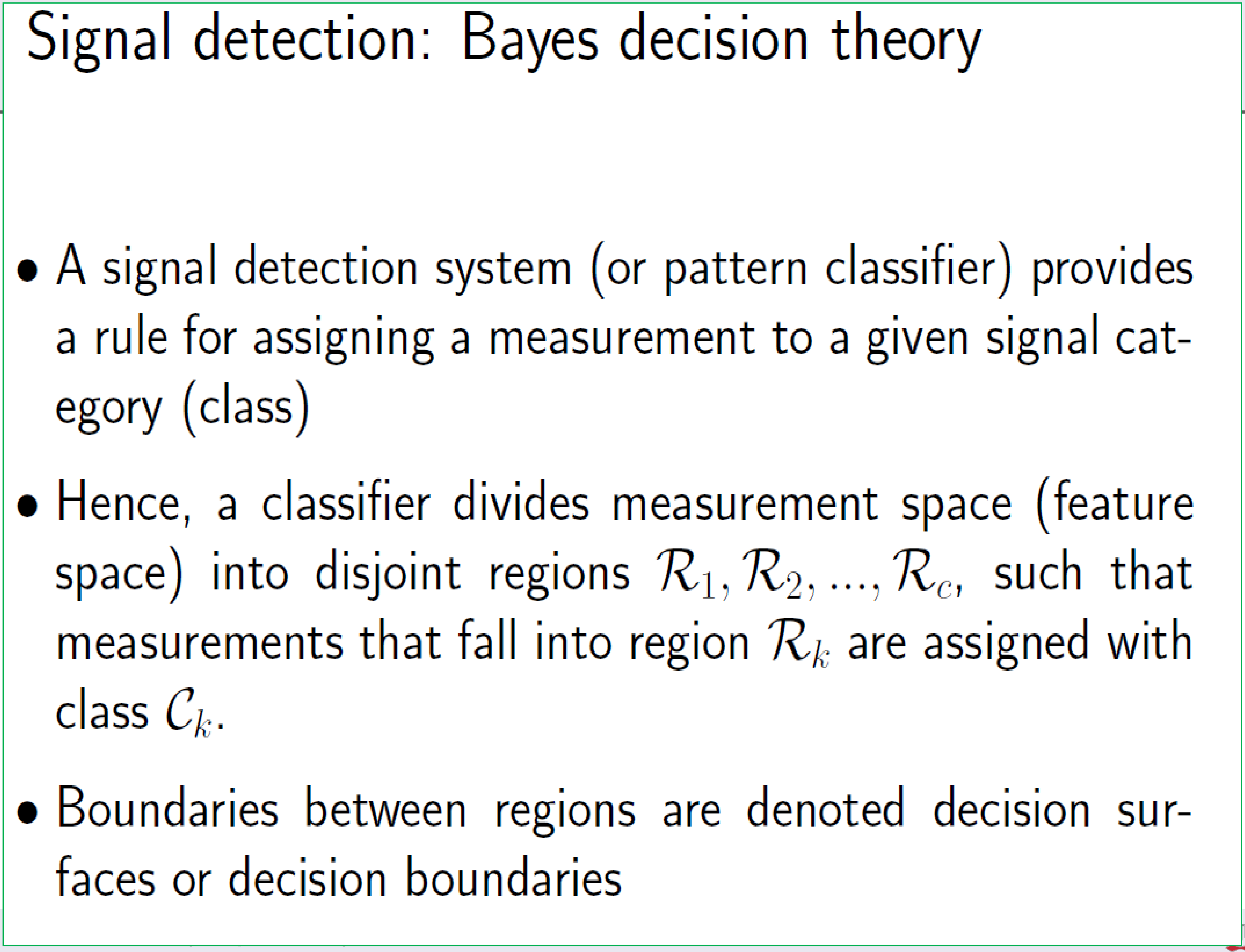
$$\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}^n + w_0 - t^n)(\mathbf{x}^n)^T \;=\; 0$$

$$\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}^n + w_0 - t^n) \;=\; 0$$

- the solution is given by in terms of $\boldsymbol{\mu} = (1/N)\sum \mathbf{x}^n$, and $\tau = (1/N)\sum t^n$

$$\mathbf{w} \;=\; \left(\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^n - \boldsymbol{\mu})(\mathbf{x}^n - \boldsymbol{\mu})^T\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}(t^n - \tau)\mathbf{x}^n\right)$$

$$w_0 \;=\; -\mathbf{w}^T\boldsymbol{\mu} + \tau$$



# Signal detection: Bayes decision theory

- A signal detection system (or pattern classifier) provides a rule for assigning a measurement to a given signal category (class)
- Hence, a classifier divides measurement space (feature space) into disjoint regions $\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_c$, such that measurements that fall into region $\mathcal{R}_k$ are assigned with class $\mathcal{C}_k$.
- Boundaries between regions are denoted decision surfaces or decision boundaries

# Signal Detection: Bayes decision theory

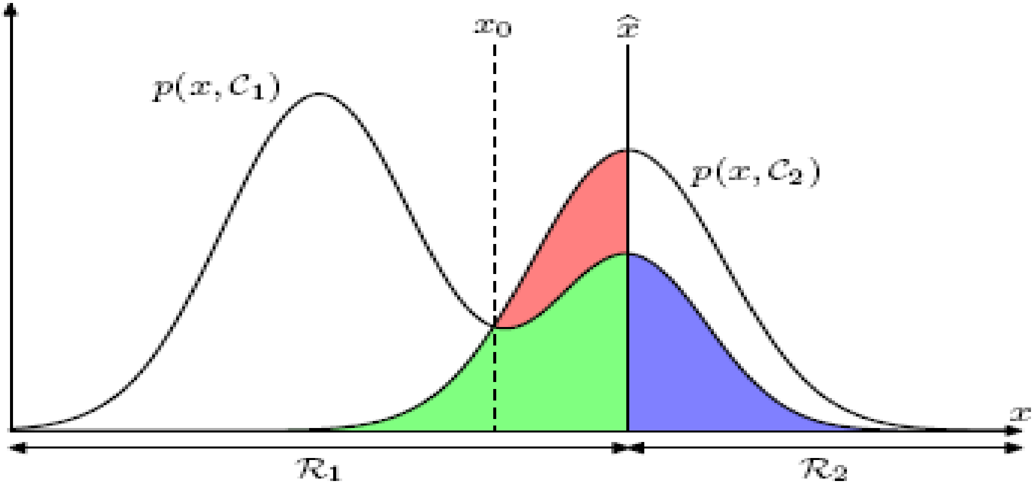
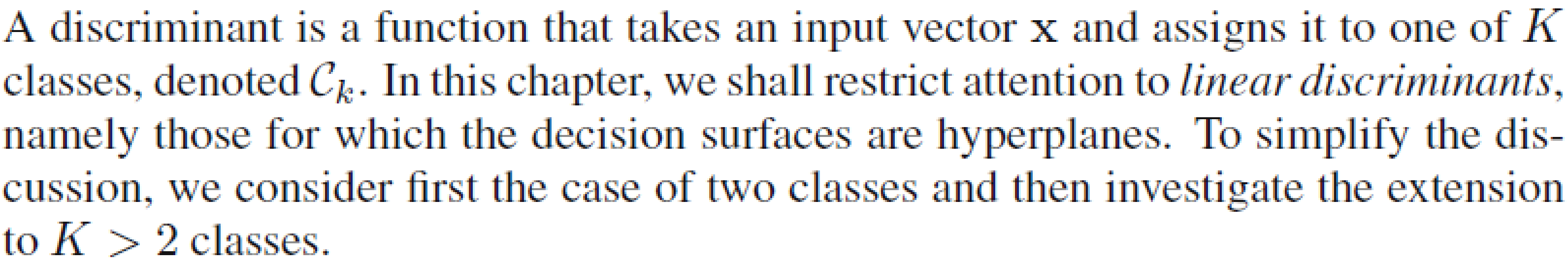


Figure 2: Schematic plot of the densities for a measured signal drawn from either of two populations $\mathcal{C}_1, \mathcal{C}_2$

$$\begin{aligned} P(\text{error}) &= P(x \in \mathcal{R}_2, \mathcal{C}_1) + P(x \in \mathcal{R}_1, \mathcal{C}_2) \\ &= P(x \in \mathcal{R}_2 | \mathcal{C}_1) P(\mathcal{C}_1) + P(x \in \mathcal{R}_1 | \mathcal{C}_2) P(\mathcal{C}_2) \\ &= \left( \int_{\mathcal{R}_2} p(x|\mathcal{C}_1) dx \right) P(\mathcal{C}_1) + \left( \int_{\mathcal{R}_1} p(x|\mathcal{C}_2) dx \right) P(\mathcal{C}_2) \end{aligned}$$

- The probability of error is minimized if we assign points to $\mathcal{R}_1$, whenever $p(x|\mathcal{C}_1)P(\mathcal{C}_1) > p(x|\mathcal{C}_2)P(\mathcal{C}_2)$

## Discriminant Functions



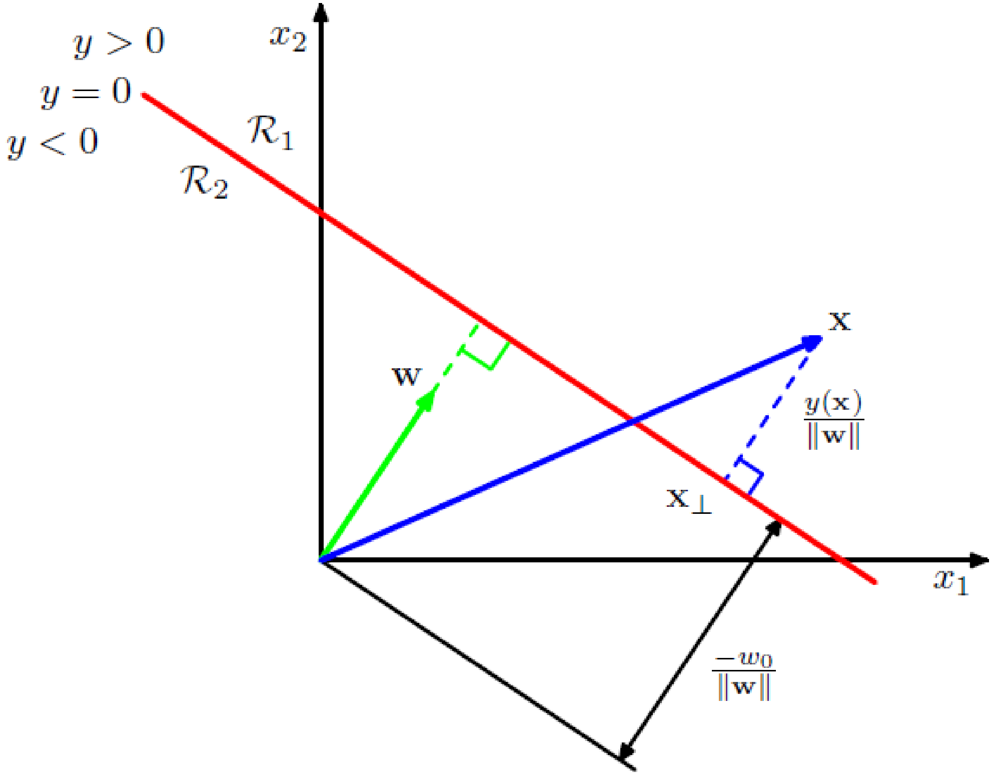
A discriminant is a function that takes an input vector $\mathbf{x}$ and assigns it to one of $K$ classes, denoted $\mathcal{C}_k$. In this chapter, we shall restrict attention to *linear discriminants*, namely those for which the decision surfaces are hyperplanes. To simplify the discussion, we consider first the case of two classes and then investigate the extension to $K > 2$ classes.

The simplest representation of a linear discriminant function

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

# Linear discriminant

Figure 4.1 Illustration of the geometry of a linear discriminant function in two dimensions. The decision surface, shown in red, is perpendicular to $\mathbf{w}$, and its displacement from the origin is controlled by the bias parameter $w_0$. Also, the signed orthogonal distance of a general point $\mathbf{x}$ from the decision surface is given by $y(\mathbf{x})/\|\mathbf{w}\|$.



# Fisher linear discriminant

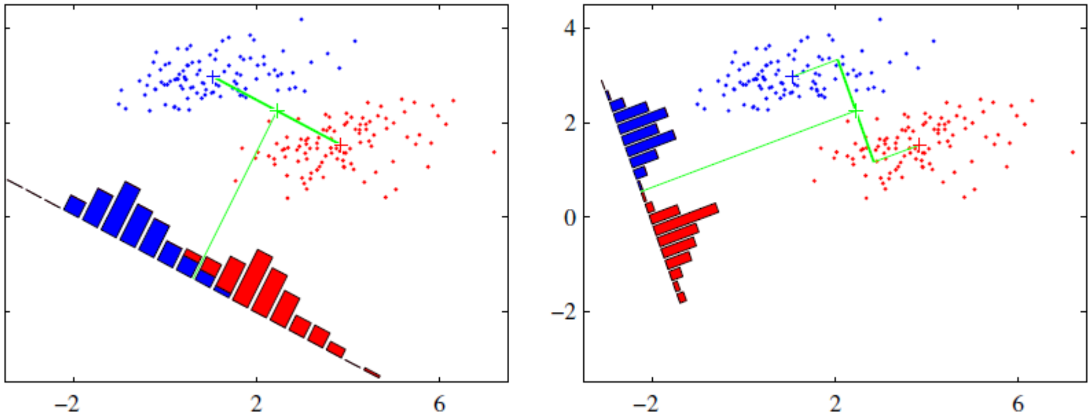
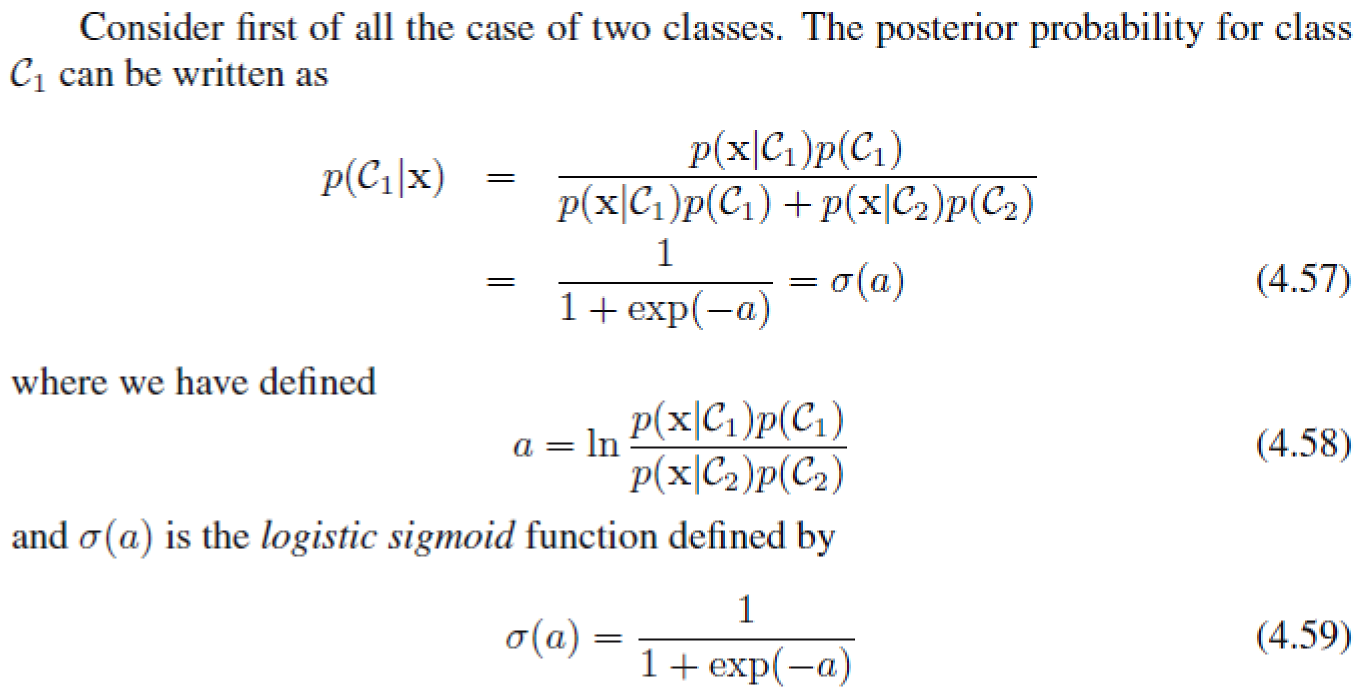


Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.



Consider first of all the case of two classes. The posterior probability for class $\mathcal{C}_1$ can be written as

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \end{aligned} \quad (4.57)$$

where we have defined

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \quad (4.58)$$

and $\sigma(a)$ is the *logistic sigmoid* function defined by

$$\sigma(a) = \frac{1}{1+\exp(-a)} \quad (4.59)$$

# Fisher linear discriminant

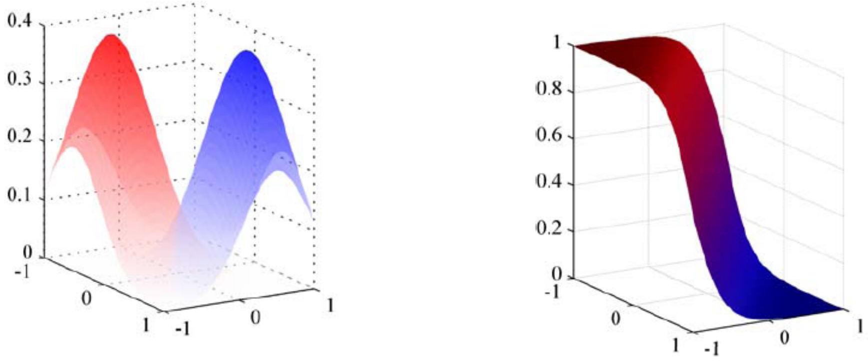


Figure 4.10 The left-hand plot shows the class-conditional densities for two classes, denoted red and blue. On the right is the corresponding posterior probability $p(\mathcal{C}_1|\mathbf{x})$, which is given by a logistic sigmoid of a linear function of $\mathbf{x}$. The surface in the right-hand plot is coloured using a proportion of red ink given by $p(\mathcal{C}_1|\mathbf{x})$ and a proportion of blue ink given by $p(\mathcal{C}_2|\mathbf{x}) = 1 - p(\mathcal{C}_1|\mathbf{x})$.

Let us assume that the class-conditional densities are Gaussian and then explore the resulting form for the posterior probabilities. To start with, we shall assume that all classes share the same covariance matrix. Thus the density for class $\mathcal{C}_k$ is given by

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}. \qquad (4.64)$$

Consider first the case of two classes. From (4.57) and (4.58), we have

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0) \qquad (4.65)$$

where we have defined

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) && (4.66) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}. && (4.67) \end{aligned}$$