APPLIED MACHINE LEARNING

Fitting Data with One Multi-dimensional Gaussian Function

Multi-dimensional Gaussian Function

The uni-dimensional Gaussian or Normal distribution is a pdf given by:

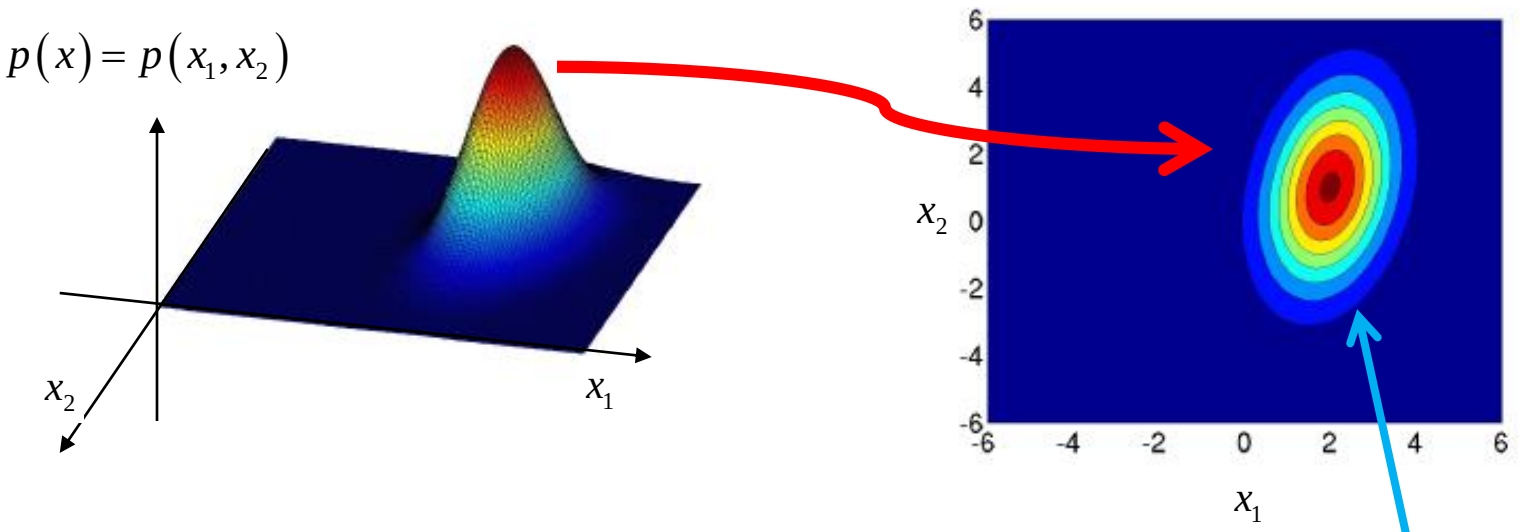
$$p(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, \quad \mu\text{:mean}, \quad \sigma^2\text{:variance}$$

The multi-dimensional Gaussian distribution is given by:

$$p(x;\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)}$$

if x is N-dimensional, then
μ is a $N-$dimensional mean vector
Σ is a $N\times N$ matrix

2-dimensional Gaussian Pdf



$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}$$

Isolines: $p(x) = cst$

if x is N-dimensional, then

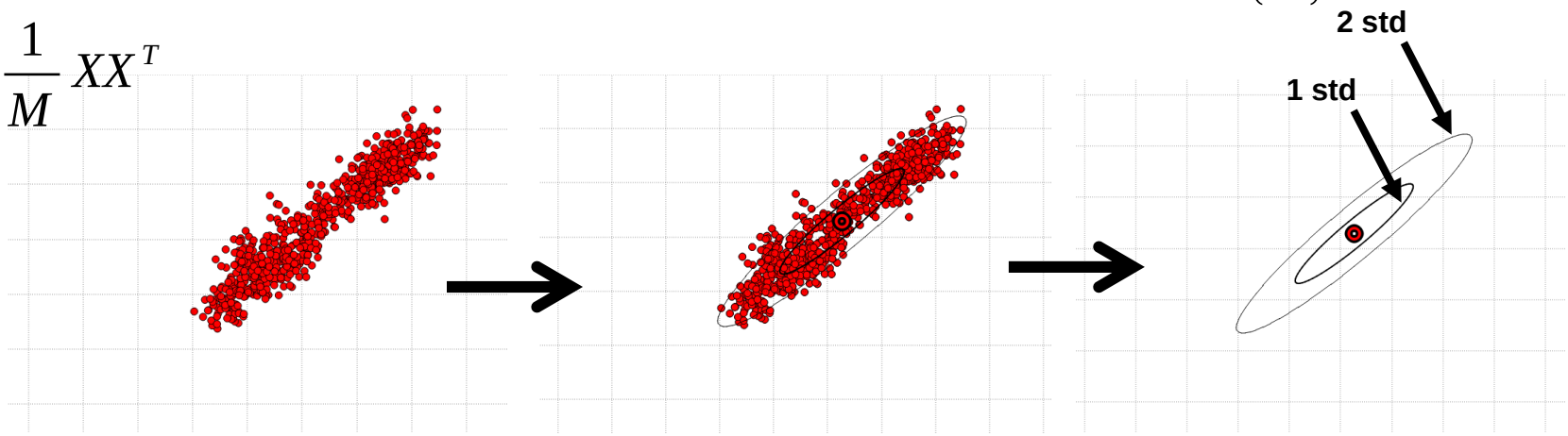
μ is a $N-$dimensional mean vector

Σ is a $N \times N$ matrix

Modeling Data with a Gaussian Function

Construct covariance matrix from (centered) set of datapoints $X = \left\{ x^i \right\}^{i=1...M}$:

$$\Sigma = \frac{1}{M} X X^T$$



$$p(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)}$$

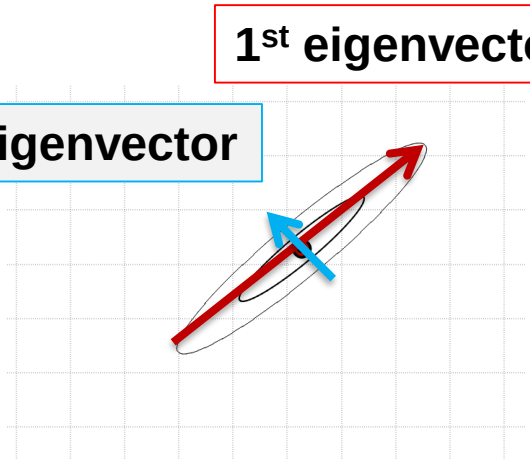
if x is N-dimensional, then

μ is a $N-$dimensional mean vector

Σ is a $N \times N$ covariance matrix

Modeling Data with a Gaussian Function

$$\Sigma = \frac{1}{M} X X^T$$



Σ is square and symmetric. It can be decomposed using the eigenvalue decomposition.

$\Sigma = V \Lambda V^T$,

V : matrix of eigenvectors,

$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ \ldots\ldots & \\ 0 & \lambda_N \end{pmatrix}$: diagonal matrix composed of eigenvalues

$\bar{\Sigma} = \Lambda = \begin{pmatrix} \lambda_1 & 0 \\ \ldots\ldots & \\ 0 & \lambda_N \end{pmatrix}$

diagonal matrix once data projected onto eigenvectors

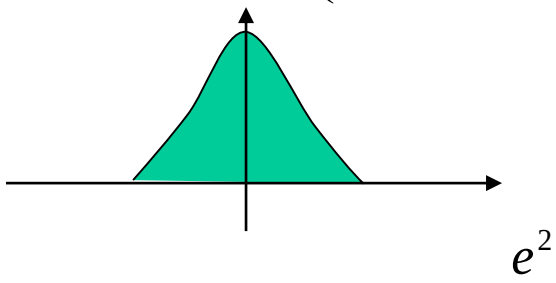
For the 1-std ellipse, the axes' lengths are equal to:

$\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, with $\Sigma = V \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} V^T$.

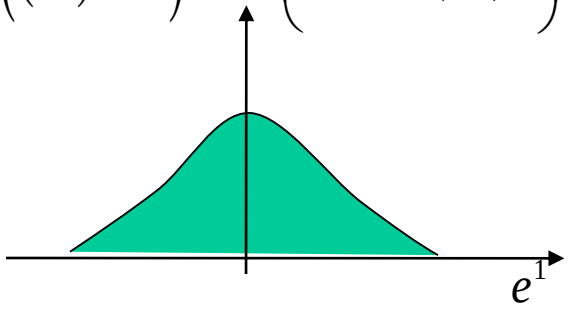
Each isoline corresponds to a scaling of the 1std ellipse.

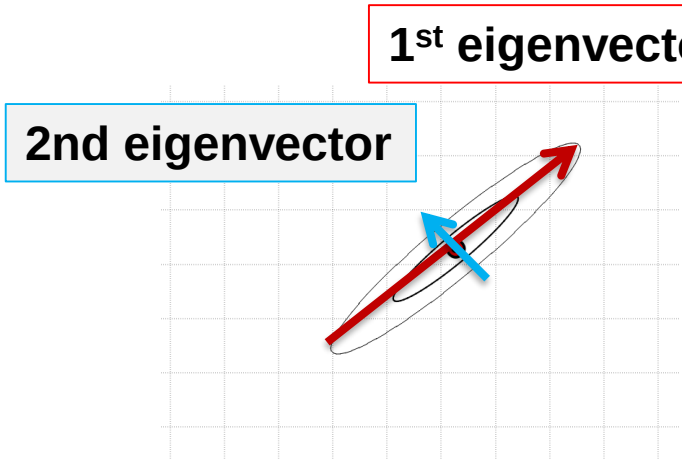
Modeling Data with a Gaussian Function

$$p\left(\left(e^{2}\right)^{T} X\right) \sim N\left(X; \mu_{2}, \left(\lambda_{2}\right)^{\frac{1}{2}}\right)$$



$$p\left(\left(e^{1}\right)^{T} X\right) \sim N\left(X; \mu_{1}, \left(\lambda_{1}\right)^{\frac{1}{2}}\right)$$

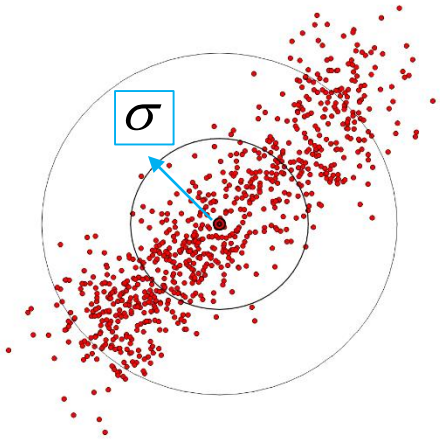




When projected onto e^1 and e^2, the set of datapoints follow two Normal distributions.

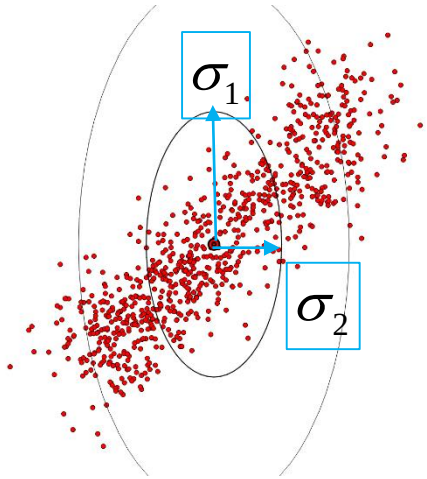
6

Which model?



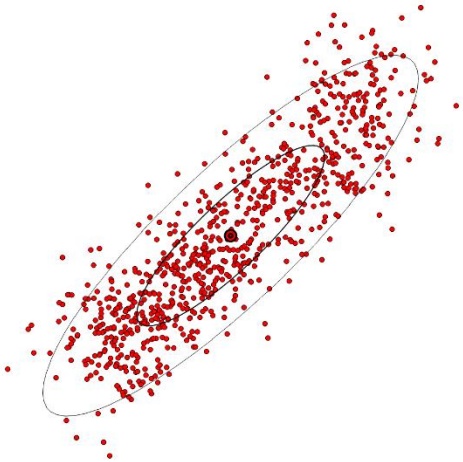
Spherical

$$\Sigma = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}$$



Diagonal

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}$$



Full

$$\Sigma = \begin{bmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{21} & \sigma_2 \end{bmatrix}$$

Need a method to derive optimal parameters (mean and covariance matrix)

Likelihood Function

The Likelihood function or *Likelihood (for short) determines the joint probability density of observing the set X* of *M* datapoints, if each datapoint has been generated by the pdf *p(x) with parameters* Θ.

$$L(\Theta \mid X) = p\left(x^1, x^2, ..., x^M; \Theta\right) \quad X = \left\{x^i\right\}_{i=1}^{M},$$

The likelihood determines how well a particular choice of pdf *p(x)* models the data.

Likelihood of Gaussian Pdf Paramatrization

A Gaussian pdf is parametrized with parameters μ, Σ.

The *likelihood function* (short – *likelihood*) of the model parameters is given by:

$$L(\mu,\Sigma \mid X) := p(X;\mu,\Sigma)$$

Measures probability of observing *X* if the distribution of *p(X)* is parametrized with μ, Σ.

If all datapoints are identically and independently distributed (i.i.d.)

$$L(\mu,\Sigma \mid X) = \prod_{i=1}^{M} p(x^i;\mu,\Sigma)$$

To determine the best fit, one must search for parameters that maximize the likelihood.

Maximum Likelihood Optimization

The principle of *maximum likelihood* consists of finding the optimal parameters of a pdf that maximize the likelihood function / maximizing the probability of the data given the model and its parameters.

For a Gauss pdf, one determines the mean and covariance matrix by solving:

$$\max_{\mu,\Sigma} L\left(\mu,\Sigma \mid X\right) = \max_{\mu,\Sigma} p\left(X \mid \mu,\Sigma\right)$$

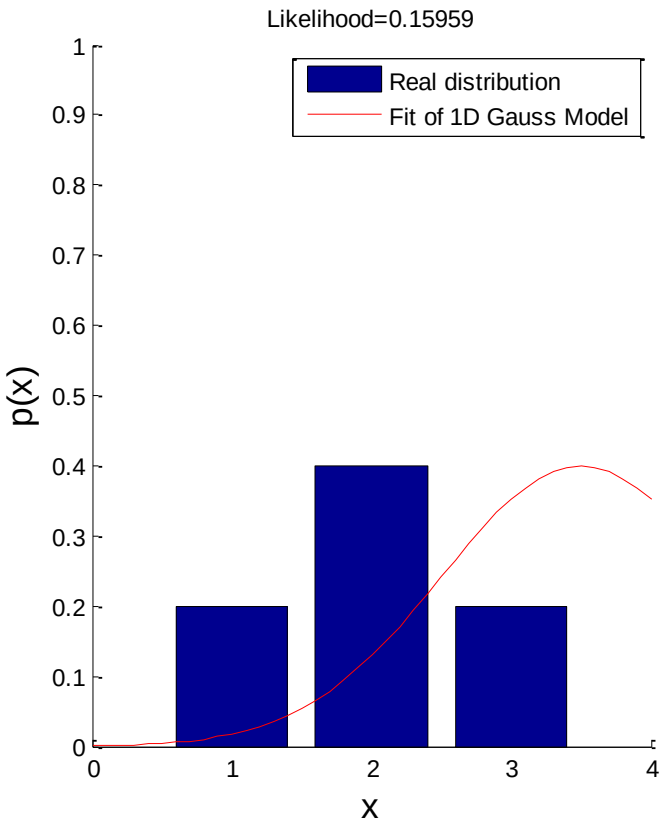
Computing the log of the likelihood yields the same optimum:

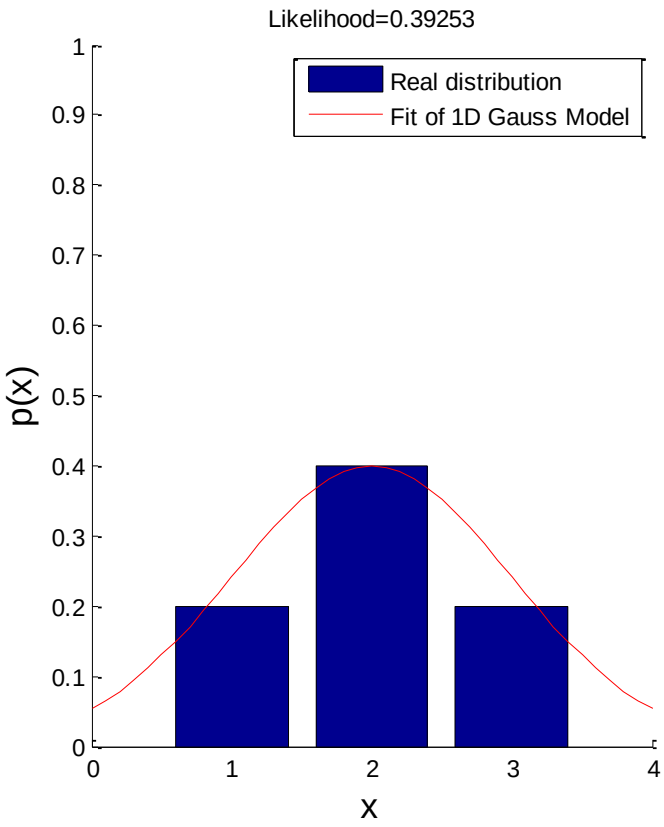
$$\max_{\Theta} \; p\left(X \mid \Theta\right) = \max_{\Theta} \; \log p\left(X \mid \Theta\right)$$

$$\frac{\partial}{\partial \mu} \log p\left(X \mid \mu,\Sigma\right) = 0 \quad \text{and} \quad \frac{\partial}{\partial \Sigma} \log p\left(X \mid \mu,\Sigma\right) = 0$$

The optimum is the mean and covariance of the data.

Likelihood Function

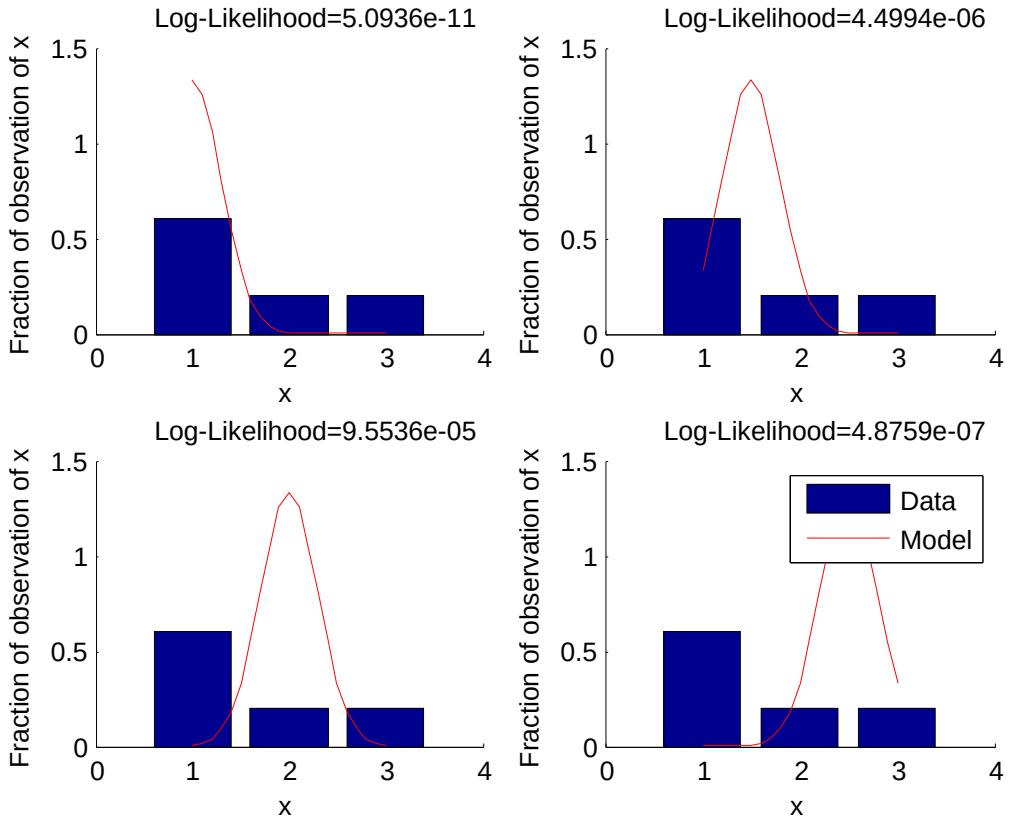




The optimum is the mean and covariance of the data.

(Instead of maximizing the likelihood, minimize the negative log-likelihood)

Likelihood Function



Log-Likelihood for a series of Gauss functions applied to datasets with pdfs that do not follow a Gauss distribution. The Likelihood increases as the fit is closer to the real mean of the data, even if this may appear as a poorer fit.