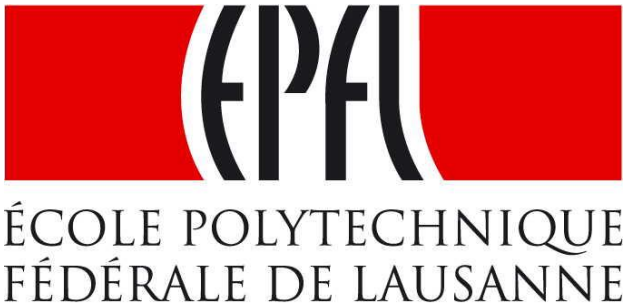
APPLIED MACHINE LEARNING

Probability Density Functions and Gaussian Mixture Models



Probability Density Functions

Lecture topics:

- Discrete and Continuous Random variables
- Joint, Conditional and Marginal probabilities
- Dependence and Correlation
- Gaussian Mixture Models
- Expectation- Maximization algorithm for Clustering

Discrete Probabilities

Random Variables (RV) are variables whose values are outcomes of a random phenomenon

Consider two variables *x* and *y* taking discrete values over the intervals $[1.....\mathrm{N}_x]$ and $[1.....\mathrm{N}_y]$ respectively.

$P(x=i)$: the probability that the variable *x* takes value *i*.

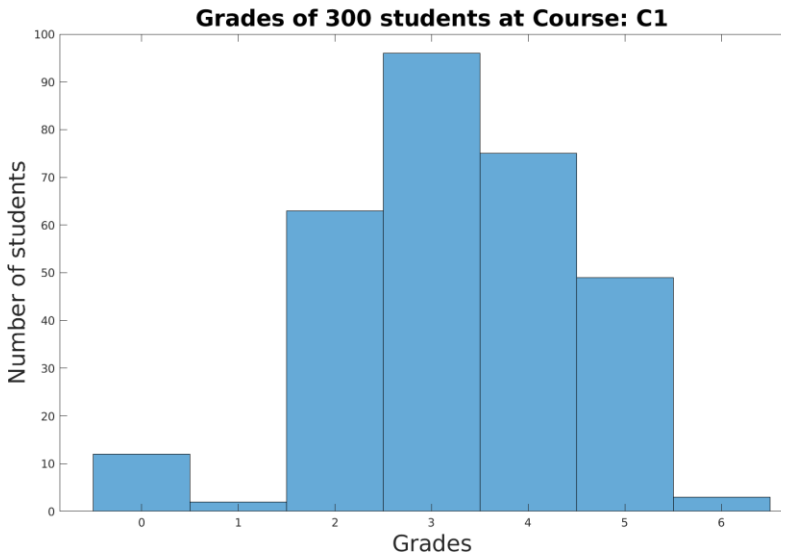
$0 \le P(x=i) \le 1, \ \forall i = 1,...,N_x,$

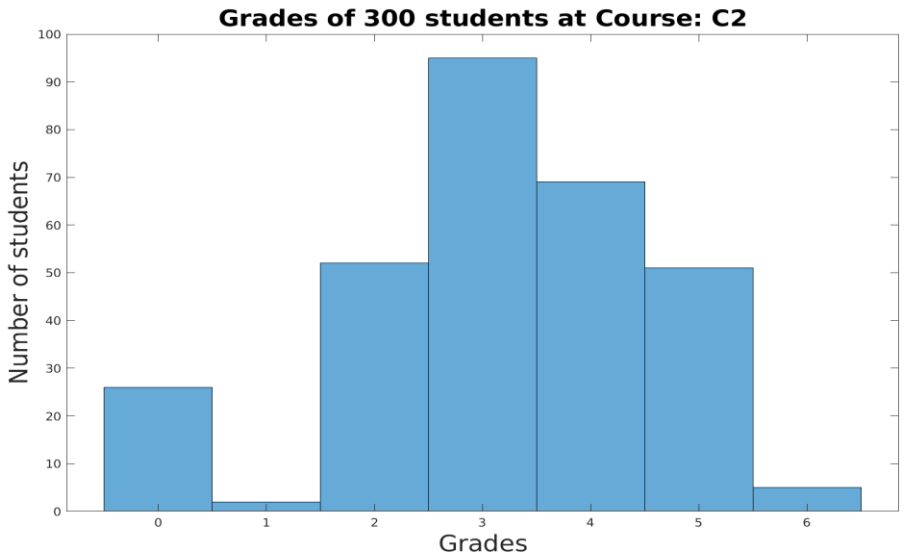
and $\sum_{i=1}^{N_x} P(x=i) = 1.$

Idem for $P(y=j), j = 1,...N_y$

Discrete Probability distributions

- Grades of 300 students in two courses C_1 and C_2

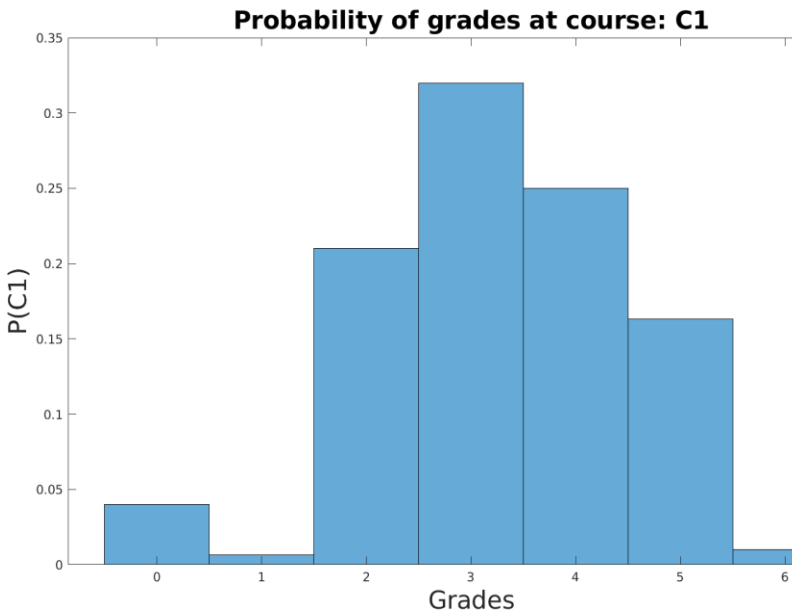


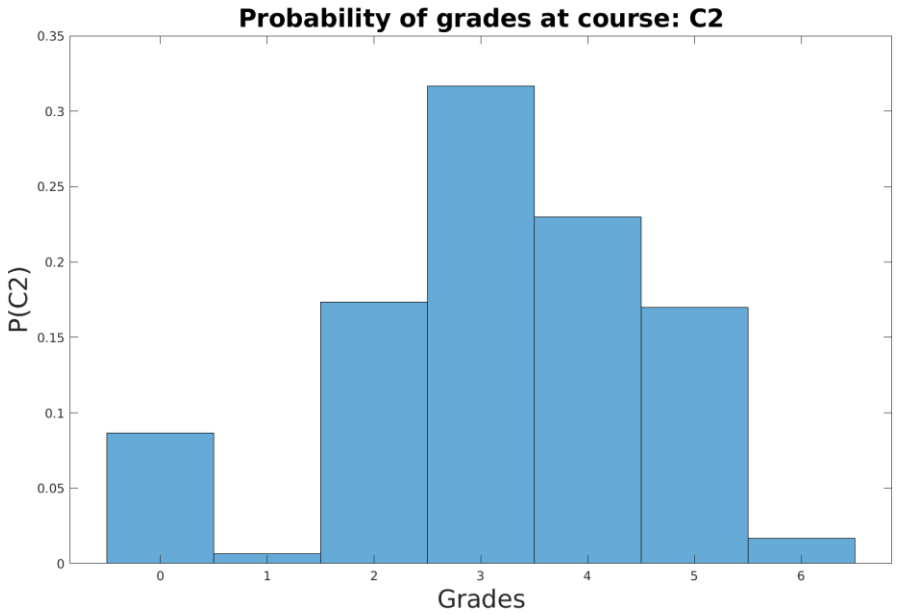


What is the $P(C_1 = 3)$?

Discrete Probability Distributions

- Probabilities of grades at courses C_1 and C_2





What is the probability of a student to receive a grade of three at both C_1 and C_2 ?

Discrete Probabilities

The *joint probability p(A,B)* that the two events A (variable *x* takes value x_i) *and* B (variable *y* takes value y_j) occur is expressed as:

$$P(A,B) = P(A \cap B) = P\left((x = x_i) \cap (y = y_j)\right)$$

$$P(A,B) = P(A \mid B)P(B)$$

Chain rule of probabilities:

$$P(A,B,C,D,...,Z) = P(A \mid B,C,D,...,Z)P(B \mid C,D,...,Z)P(C \mid D,...,Z)...P(Z)$$

P(A | B) is the *conditional probability* that *event A will take place given that event B already took place*

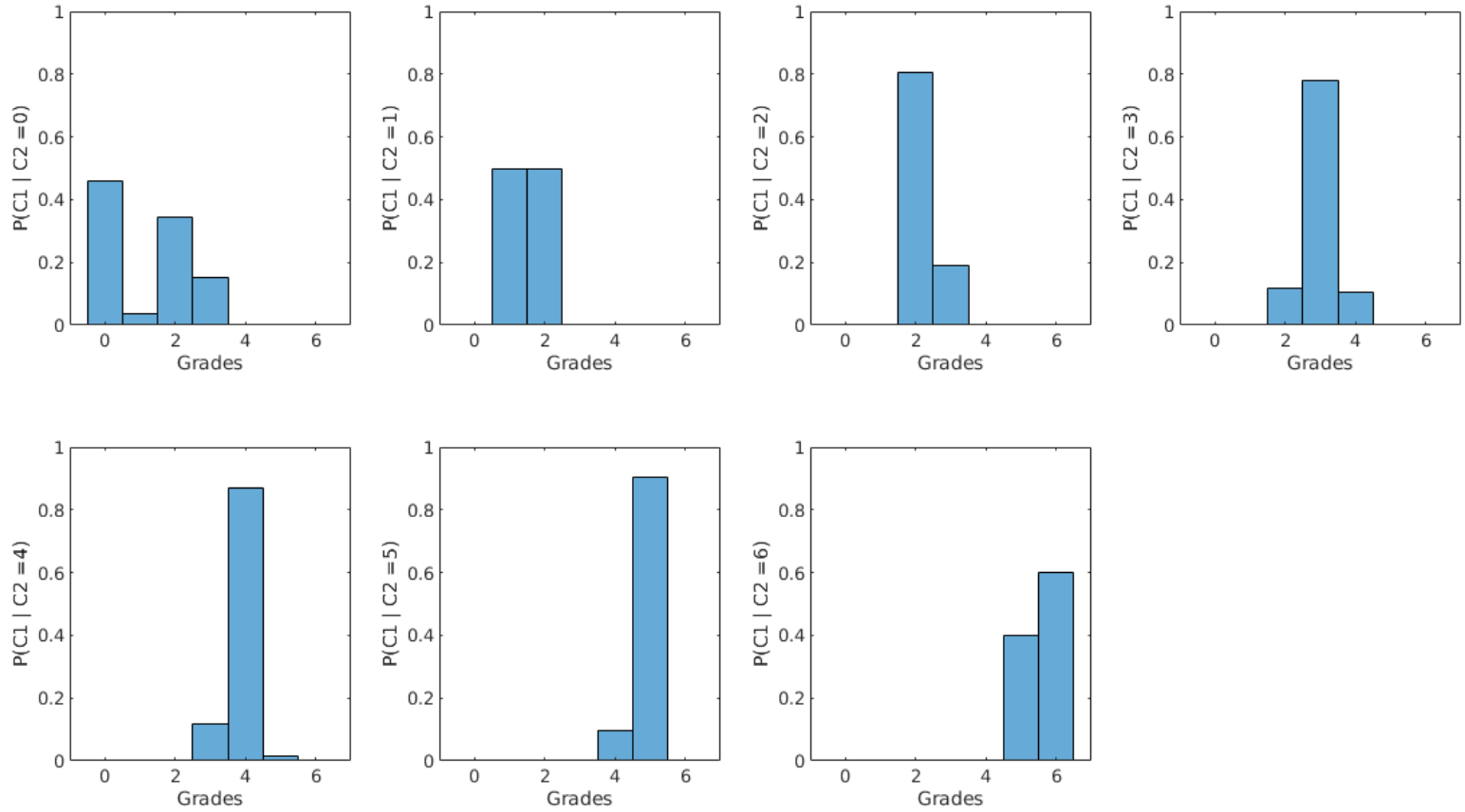
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

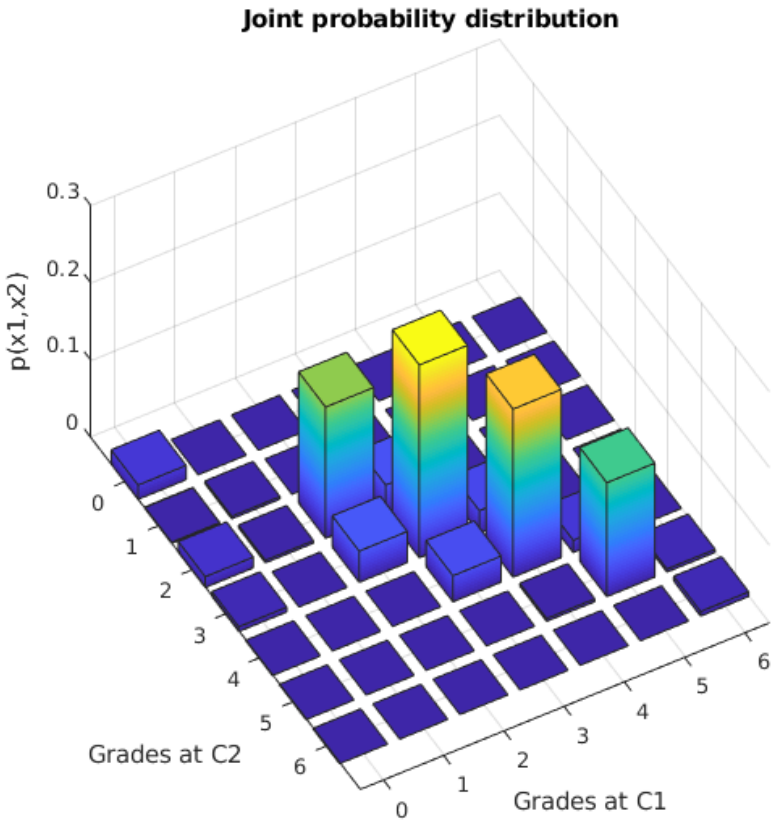
Conditional Probabilities of Discrete RV

How $P(C_1 = x \mid C_2 = y)$ is calculated?



Joint Probabilities of Discrete RV

$$P(C_1 = x , C_2 = y)$$



Discrete Probabilities

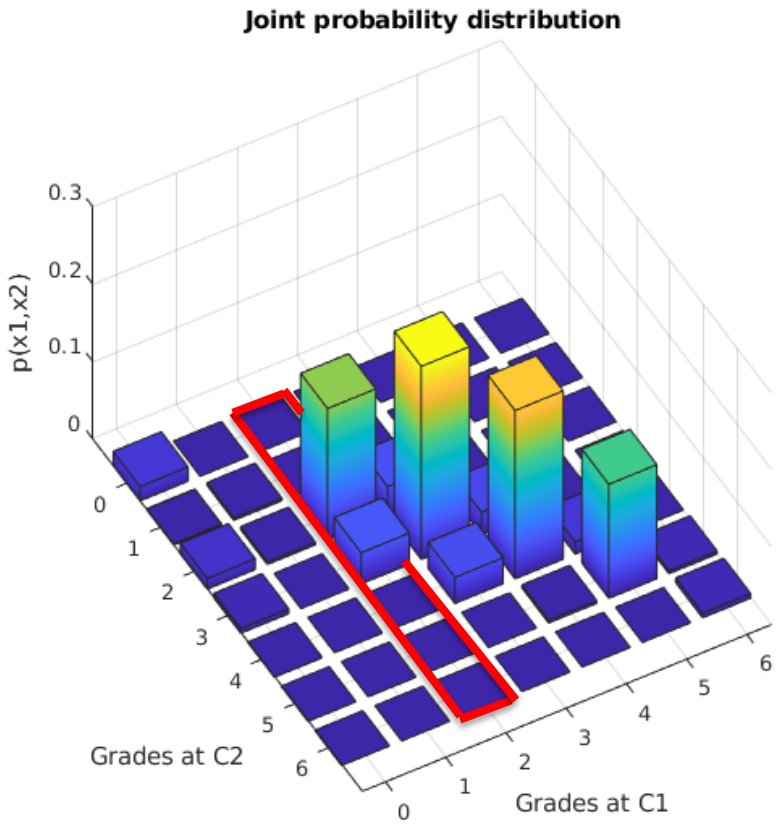
The *marginal probability* that variable x takes value x_i is given by:

$$P(X = x_i) := \sum_{y=1}^{N_y} P_{xy}(X = x, Y = y)$$

- To compute the marginal, one needs the joint distribution p(x,y).
- Often, one does not know it and can only estimate it.
- If x is a multidimensional variable → the marginal is a joint distribution!

Marginal Probabilities of Discrete RV

$$P(C_1 = 2) = \sum_{i=1}^{N_{C2}} P(C_1 = 2, C_2 = i)$$



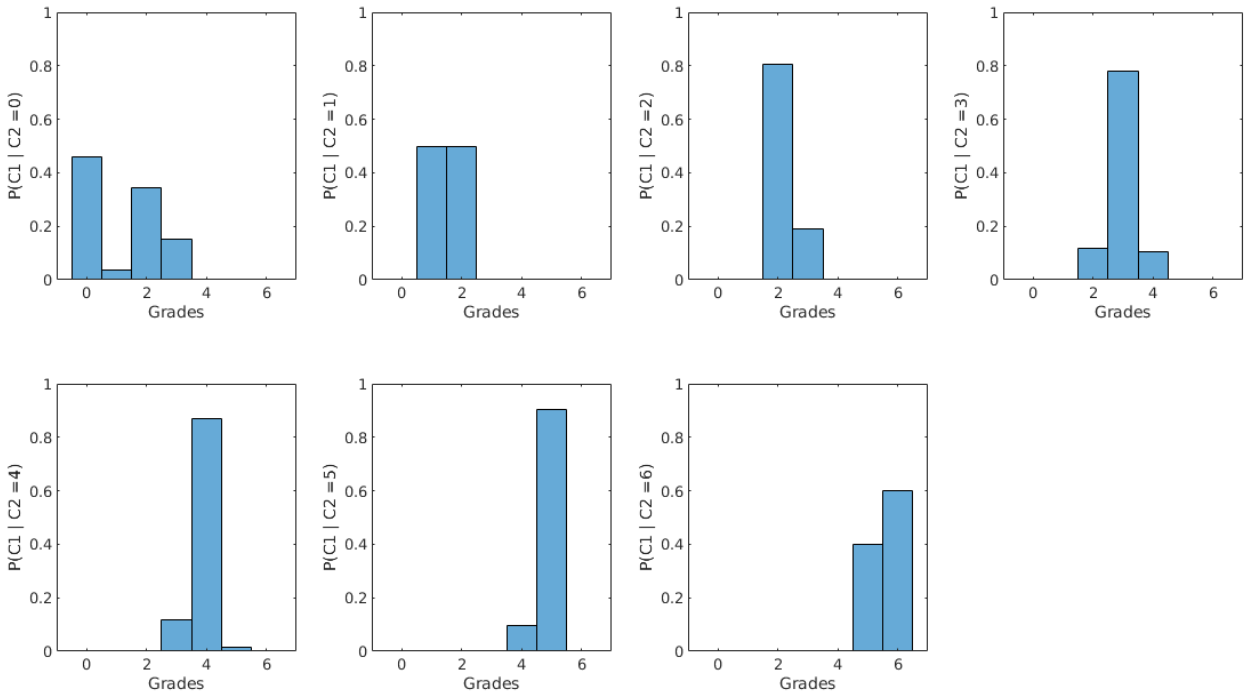
Conditional Probabilities and statistical independence

$$P(A,B) = P(A \mid B)P(B)$$

But if two RVs are statistically independent then:

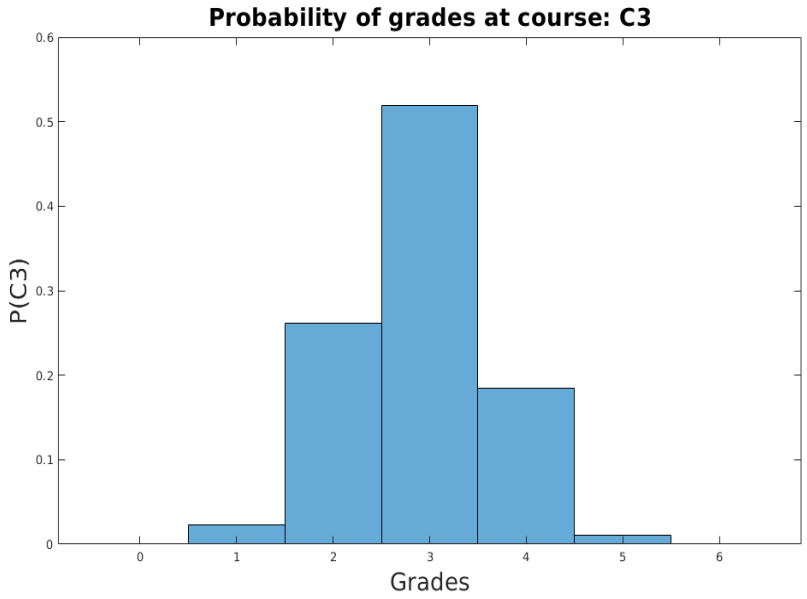
p(A,B)=p(A)p(B) p(A)=p(A|B)

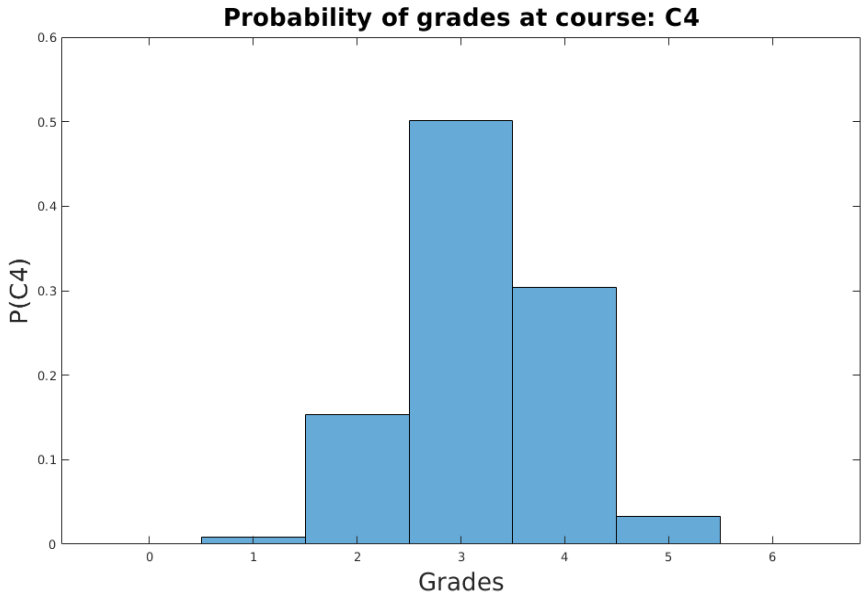
Are C1 and C2 intendent?



Dependent and Independent RVs

We are given data from 2 other Courses : C3, C4

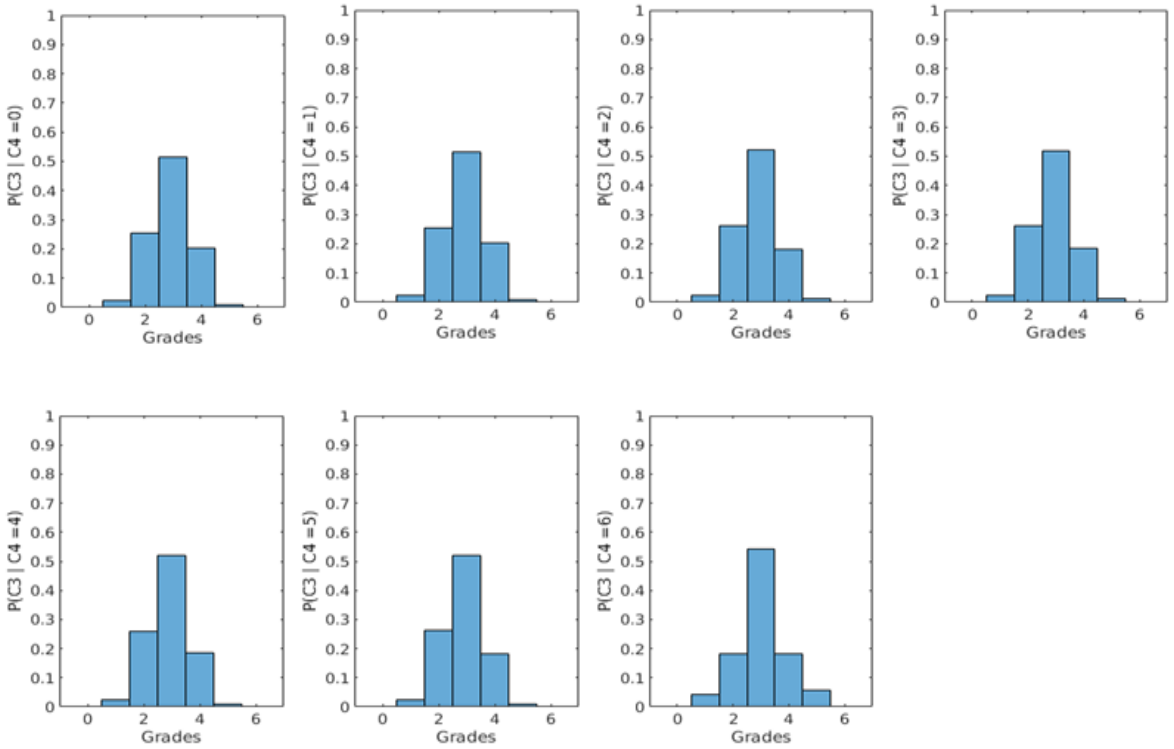




Dependent and Independent RVs

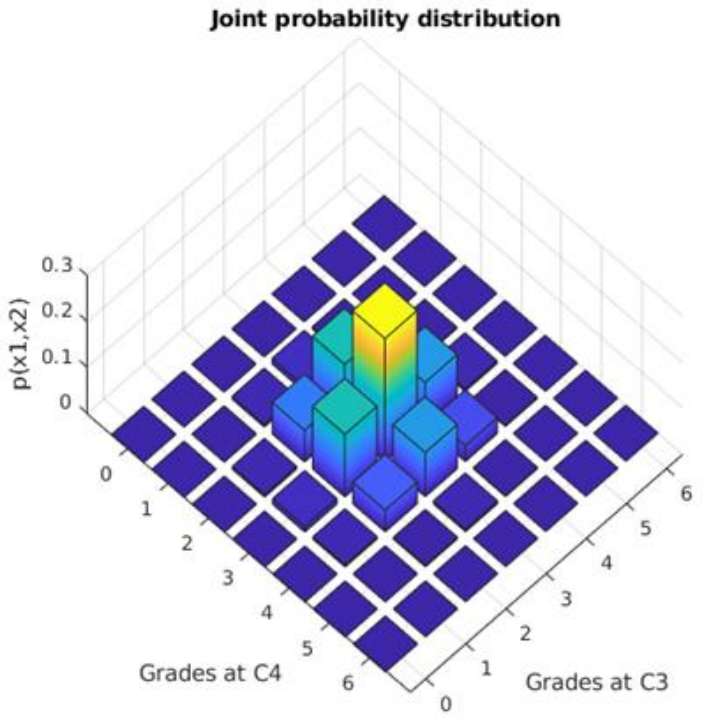
$$P(C_3 = x \mid C_4 = y)$$

Are C3 and C4 statistically dependent?

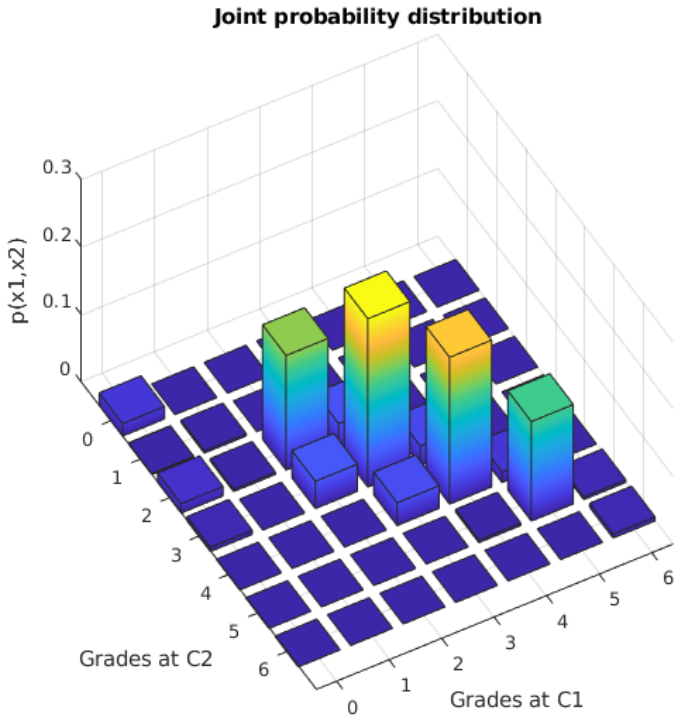


Dependent and Independent RVs

$P(C_3 , C_4)$



$P(C_1 , C_2)$



Joint Distribution and Curse of Dimensionality

The joint distribution is far richer than the marginals.

Pros of computing the joint distribution:
Provides statistical dependencies across all variables and the marginal distributions

Cons:
Computational costs grow exponentially with number of dimensions

Compute only the conditionals if we only care about dependecies across variables

Continuous Random Variables

Probability Distributions, Density Functions

p(x) a ***continuous function*** is the *probability density function or probability distribution function (PDF) (sometimes also called probability distribution* or simply *density)* of variable *x*.

It provides a relative measurement of how much dense the data are at x

$$p(x) \geq 0, \quad \forall x \in \mathbb{R}$$

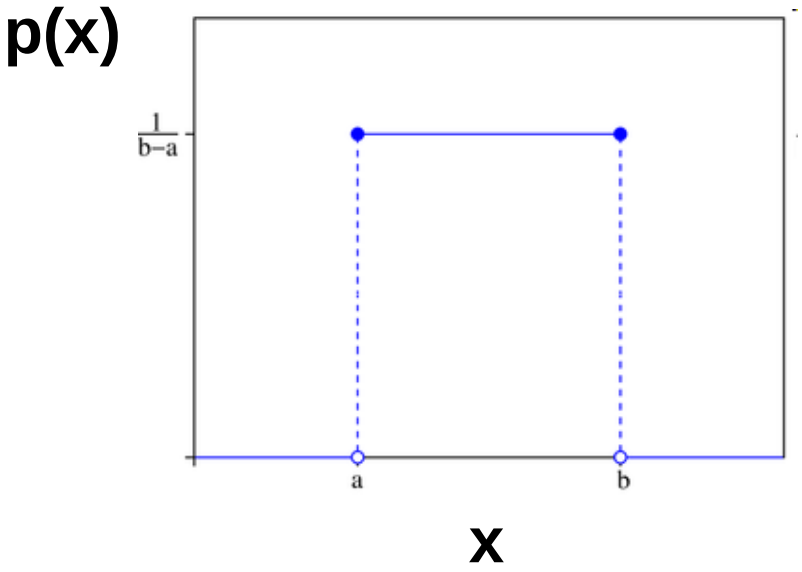
$$\int_{-\infty}^{\infty} p(x)dx = 1$$

Probability Distributions, Density Functions

The pdf is not bounded by 1.

It can grow unbounded, depending on the value taken by x.

The p(X=x) does not give the probabilities of a random continue variable. Those derive from the cumulative density function



Probability functions of continuous variables

The cumulative distribution function (or simply *distribution function*) of X is:

$$D_x\left(x^*\right) = P\left(x \leq x^*\right)$$

$$D_x\left(x^*\right) = \int_{-\infty}^{x^*} p(x)dx, \qquad x \in \mathbb{R}$$

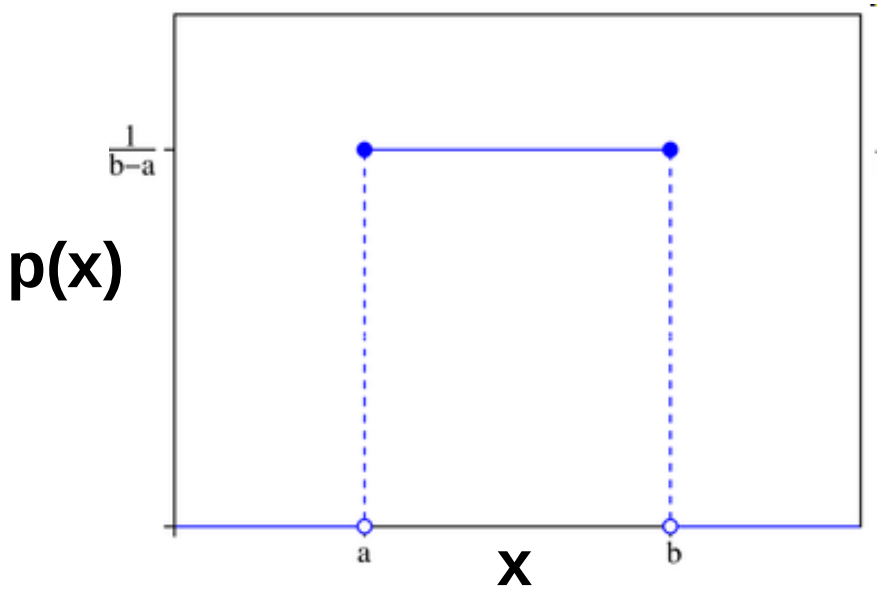
$p(x)\,dx$ ~ probability of x to fall within an infinitesimal interval $[x, x + dx]$

For obtaining the probability of a random continuous event to occur, we have to define bounds

$$P(x_l < x < x_u) = \int_{x_l}^{x_u} p(x)dx$$

Probability functions of continuous variables

Uniform distribution on x

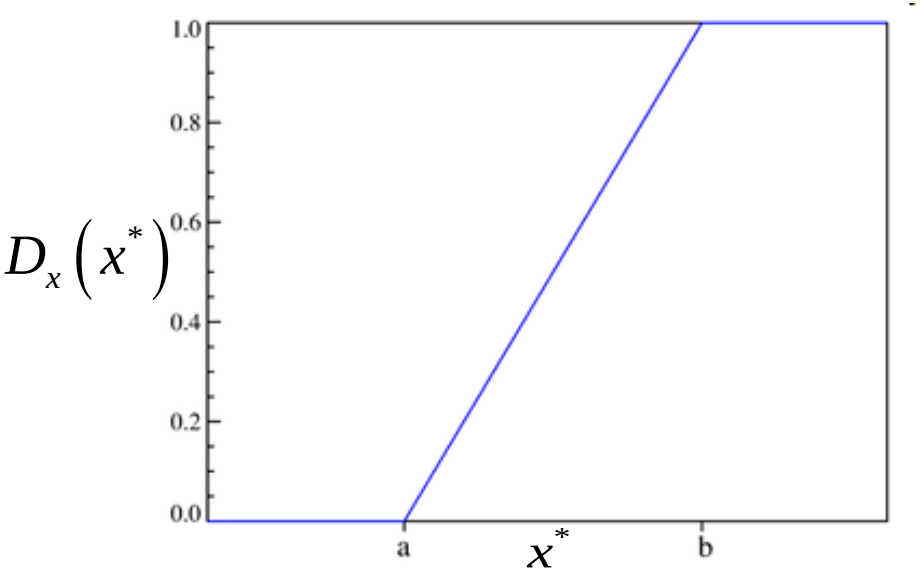


Probability that *x* takes a value in the subinterval [a,b] is given by:

$$P(x \le b) := D_x(x \le b) = \int_{-\infty}^{b} p(x)dx$$

$$P(a \le x \le b) = D_x(x \le b) - D_x(x \le a)$$

$$P(a \le x \le b) = \int_{a}^{b} p(x)dx = 1$$



Expectation

The *expectation* of the random variable *x* with probability P(x) (in the discrete case) and pdf p(x) (in the continuous case), also called the *expected value* or *mean,* is the mean of the observed value of *x* weighted by *p(x).* If *X* is the set of observations of *x,* then:

When x takes discrete values: $$\mu = E\{x\} = \sum_{x \in X} xP(x)$$

For continuous distributions: $$\mu = E\{x\} = \int_X x \cdot p(x) \cdot dx$$

Conditional Pdf and Statistical Independence

x_1 and x_2 are statistically independent if:

$$p(x_1 \mid x_2) = p(x_1) \quad \text{and} \quad p(x_2 \mid x_1) = p(x_2)$$

$$\Rightarrow \boxed{p(x_1, x_2) = p(x_1)\, p(x_2)}$$

---

x_1 and x_2 are uncorrelated if $\text{cov}(x_1, x_2) = 0.$

$$\text{cov}(x_1, x_2) = E\{x_1, x_2\} - E\{x_1\} E\{x_2\}$$

$$\Rightarrow E\{x_1 x_2\} = E\{x_1\} E\{x_2\}$$

The expectation of the product of RVs is equal to the product of their expectations.

Statistical independence and uncorrelation

Independent ⇏ **Uncorrelated**

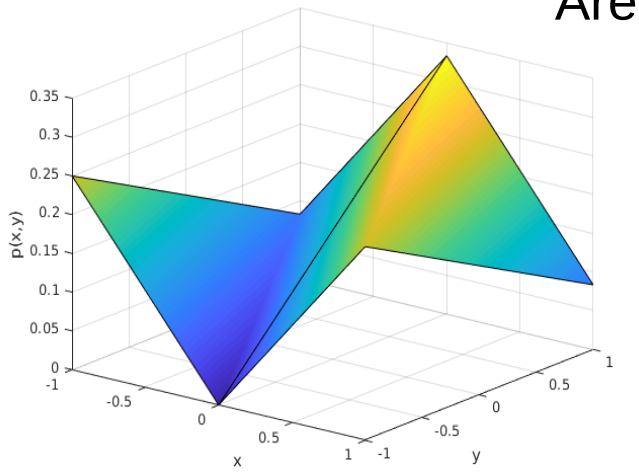
$$p(x_1, x_2) = p(x_1)\,p(x_2) \quad \Rightarrow \quad E\{x_1 x_2\} = E\{x_1\}\,E\{x_2\}$$

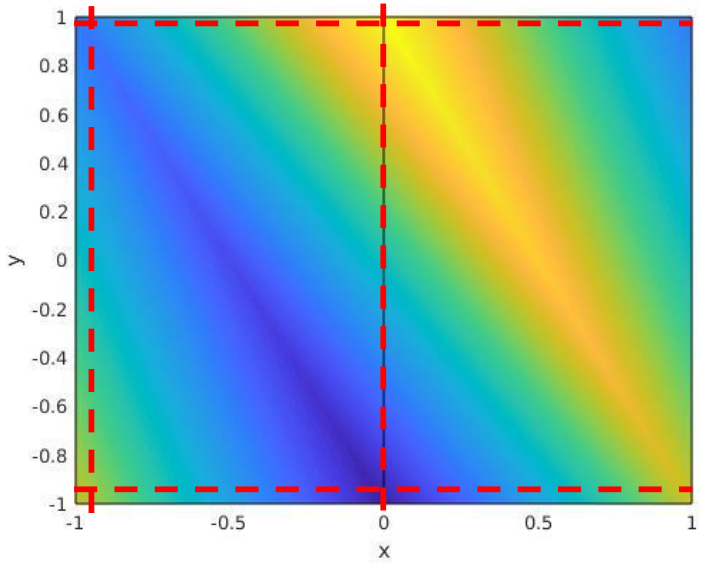
$$p(x_1, x_2) = p(x_1)\,p(x_2) \quad \not\Leftarrow \quad E\{x_1 x_2\} = E\{x_1\}\,E\{x_2\}$$

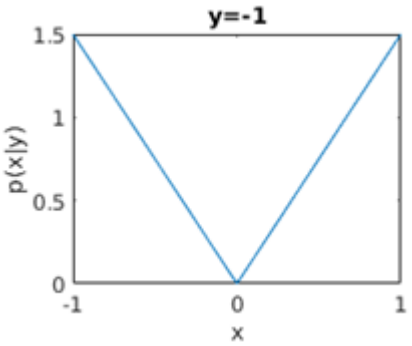
Statistical independence ensures uncorrelation.
The converse is not true

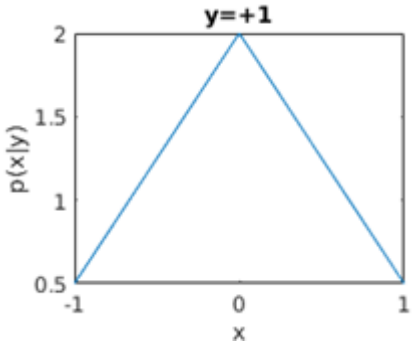
Statistical independence and uncorrelation

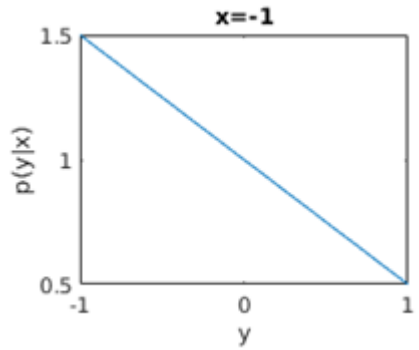
Are x,y correlated?
Are they dependent?

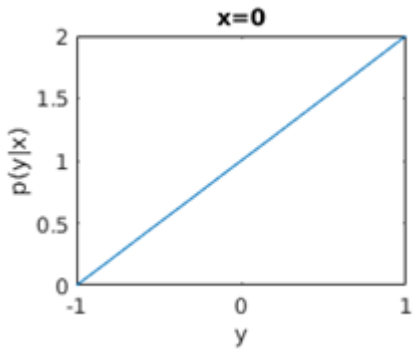












Statistical independence and uncorrelation

Are x_1 and x_2 uncorrelated?
Are x_1 and x_2 statistically independent?

	x_2=-1	x_2=0	x_2=1	Total
x_1=-1	3/12	0	3/12	**1/2**
x_1=1	1/12	4/12	1/12	**1/2**
Total	**1/3**	**1/3**	**1/3**	

$$p(x_1, x_2) = p(x_1)\,p(x_2) \quad \Rightarrow \quad E\{x_1 x_2\} = E\{x_1\}\,E\{x_2\}$$

$$p(x_1, x_2) = p(x_1)\,p(x_2) \quad \not\Leftarrow \quad E\{x_1 x_2\} = E\{x_1\}\,E\{x_2\}$$

$$E\{x\} = \sum_{x \in X} xP(x) \text{ and } E\{xy\} = \sum_{x,y} xyP(x,y)$$

Statistical independence and uncorrelation

	x_2=-1	x_2=0	x_2=1	Total
x_1=-1	3/12	0	3/12	**1/2**
x_1=1	1/12	4/12	1/12	**1/2**
Total	**1/3**	**1/3**	**1/3**	

x_1 and x_2 are uncorrelated but they are statistically dependent.

$$E\{x_1, x_2\} = E\{x_1\}E\{x_2\} = 0$$

$$\text{but } p(x_1 = -1, x_2 = 1) = 3/12 = 0.25$$

$$\neq$$

$$p(x_1 = -1)\, p(x_2 = 1) = 1/2 * 1/3 = 0.1667$$

Variance

σ^2, the variance of a distribution and measures the amount of spread of the data around the expected value:

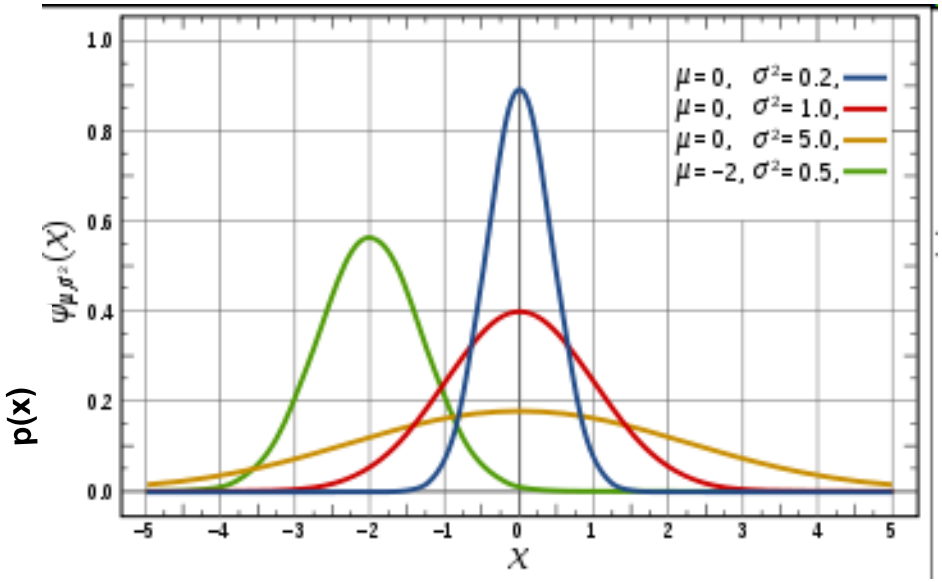
$$\sigma^2 = Var(x) = E\left\{\left(x-\mu\right)^2\right\} = E\left\{x^2\right\} - \left[E\left\{x\right\}\right]^2$$

σ is the *standard deviation* of x.

Parametric PDF

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

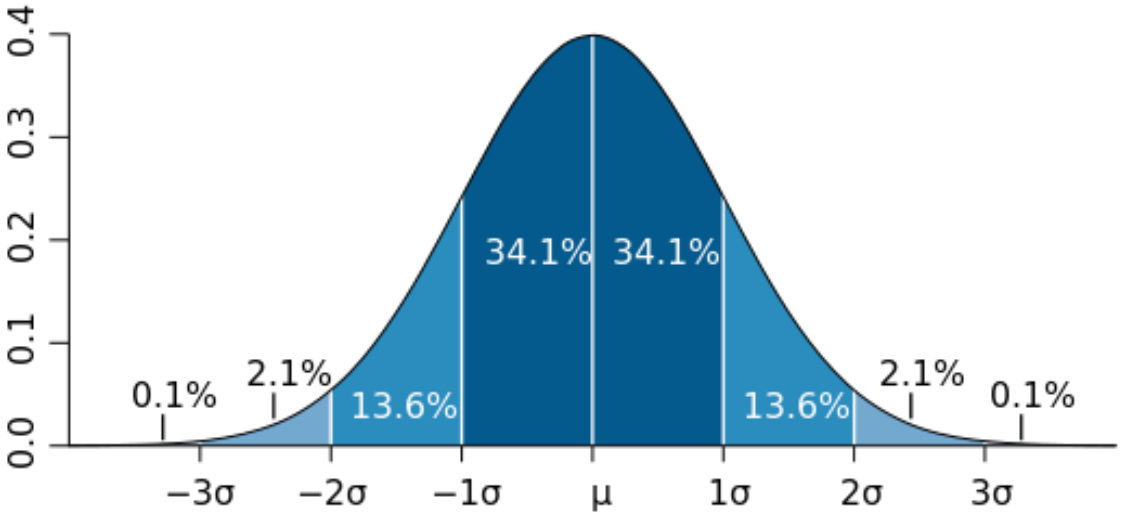
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, \quad \mu\text{:mean}, \quad \sigma^2\text{:variance}$$



The Gaussian function is entirely determined by its mean and variance. For this reason, it is referred to as a parametric distribution.

Illustrations from Wikipedia

Mean and Variance in PDF



~68% of the data are comprised between +/- 1 sigma
~96% of the data are comprised between +/- 2 sigma-s
~99% of the data are comprised between +/- 3 sigma-s

This is no longer true for arbitrary pdf-s!

Illustrations from Wikipedia

Multi-dimensional Gaussian Function

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

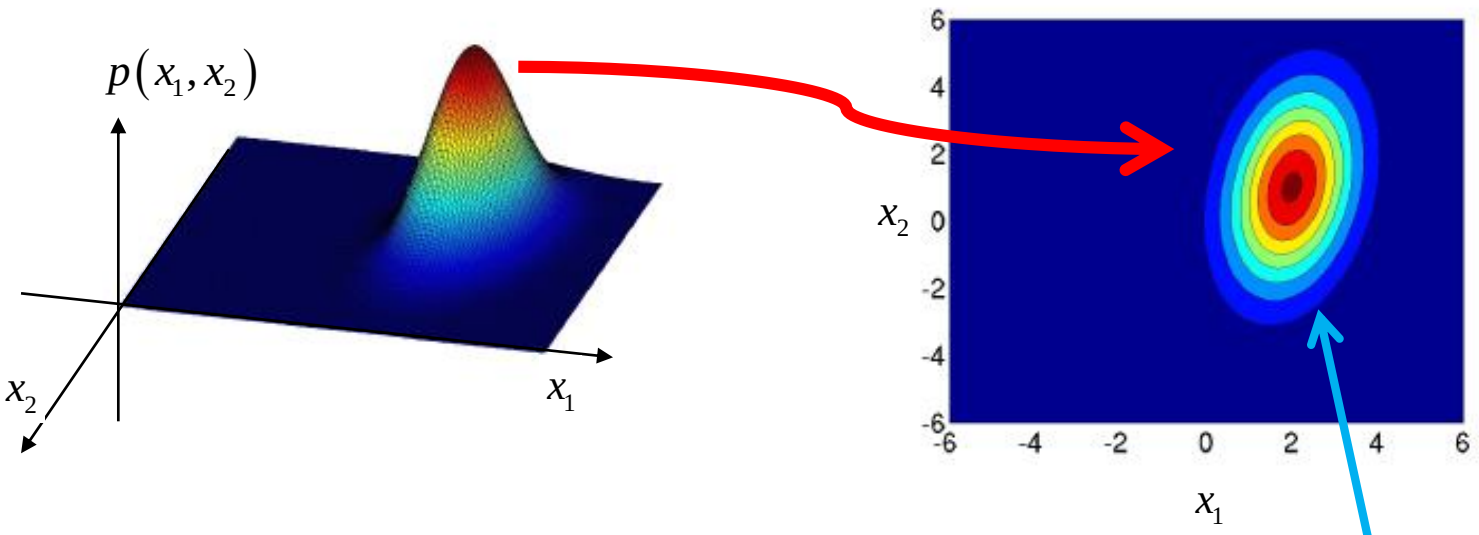
$$p(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}, \quad \mu\text{:mean}, \quad \sigma^2\text{:variance}$$

The multi-dimensional Gaussian or Normal distribution has a pdf 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×N covariance 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×N covariance matrix

Modeling Data with a Gaussian Function

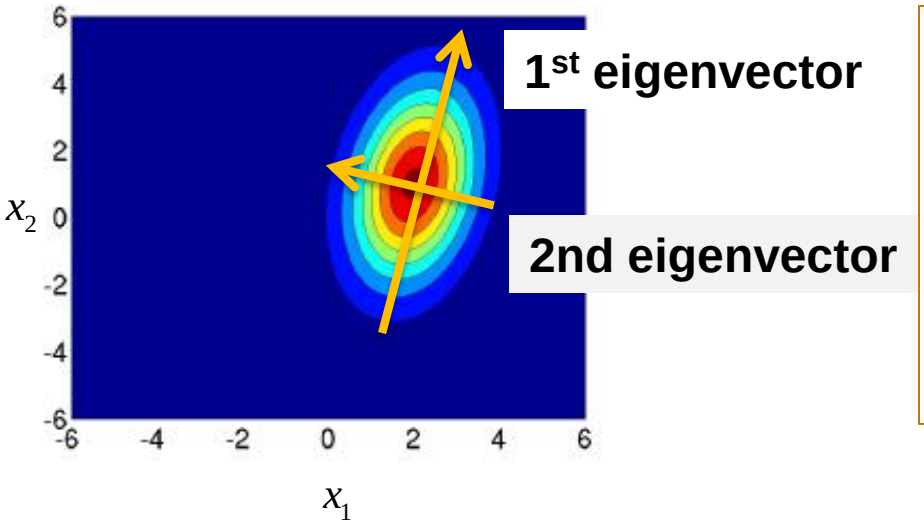
Construct covariance matrix from (centered) set of datapoints $X = \left\{x^i\right\}^{i=1...M}$:

$$\sum = \frac{1}{M} XX^T$$

$\sum$ is square and symmetric. It can be decomposed using the eigenvalue decomposition.

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

V : matrix of eigenvectors, $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ ..... & \\ 0 & \lambda_N \end{pmatrix}$: diagonal matrix composed of eigenvalues



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

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

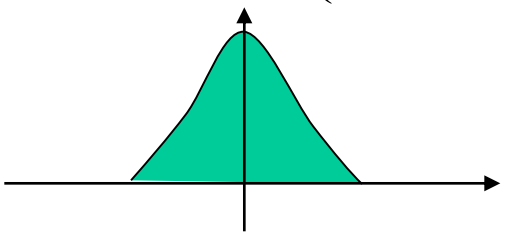
Each isoline corresponds to a scaling of the 1std ellipse.

Fitting a single Gauss function and PCA

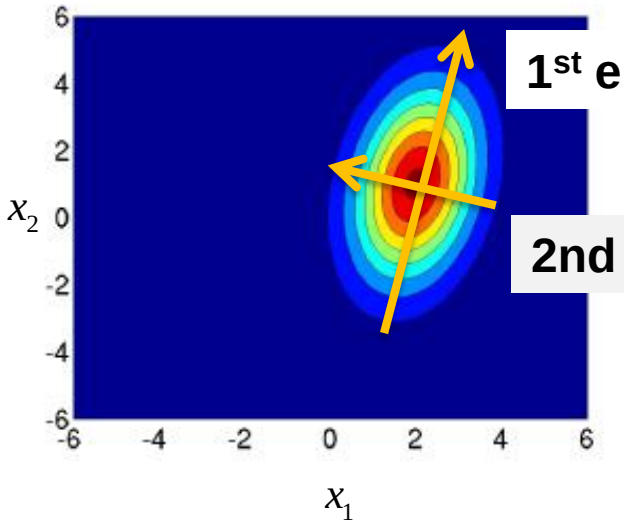
PCA Identifies a suitable representation of a multivariate data set by decorrelating the dataset.

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

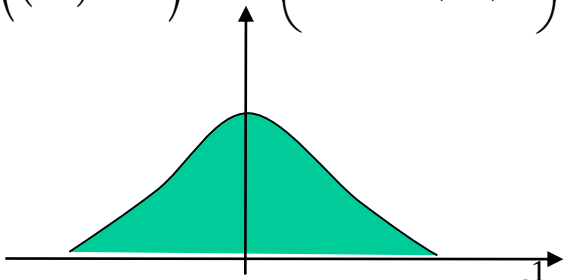
$$p\left(\left(e^2\right)^T X\right) \sim N\left(X; \mu_2, \left(\lambda_2\right)^{\frac{1}{2}}\right)$$



e^2



$$p\left(\left(e^1\right)^T X\right) \sim N\left(X; \mu_1, \left(\lambda_1\right)^{\frac{1}{2}}\right)$$



e^1

Marginal, Conditional in Pdf

Consider two random variables x_1 and x_2 with joint distribution $p(x_1, x_2)$, then the *marginal probability* of x_1 given x_1 is:

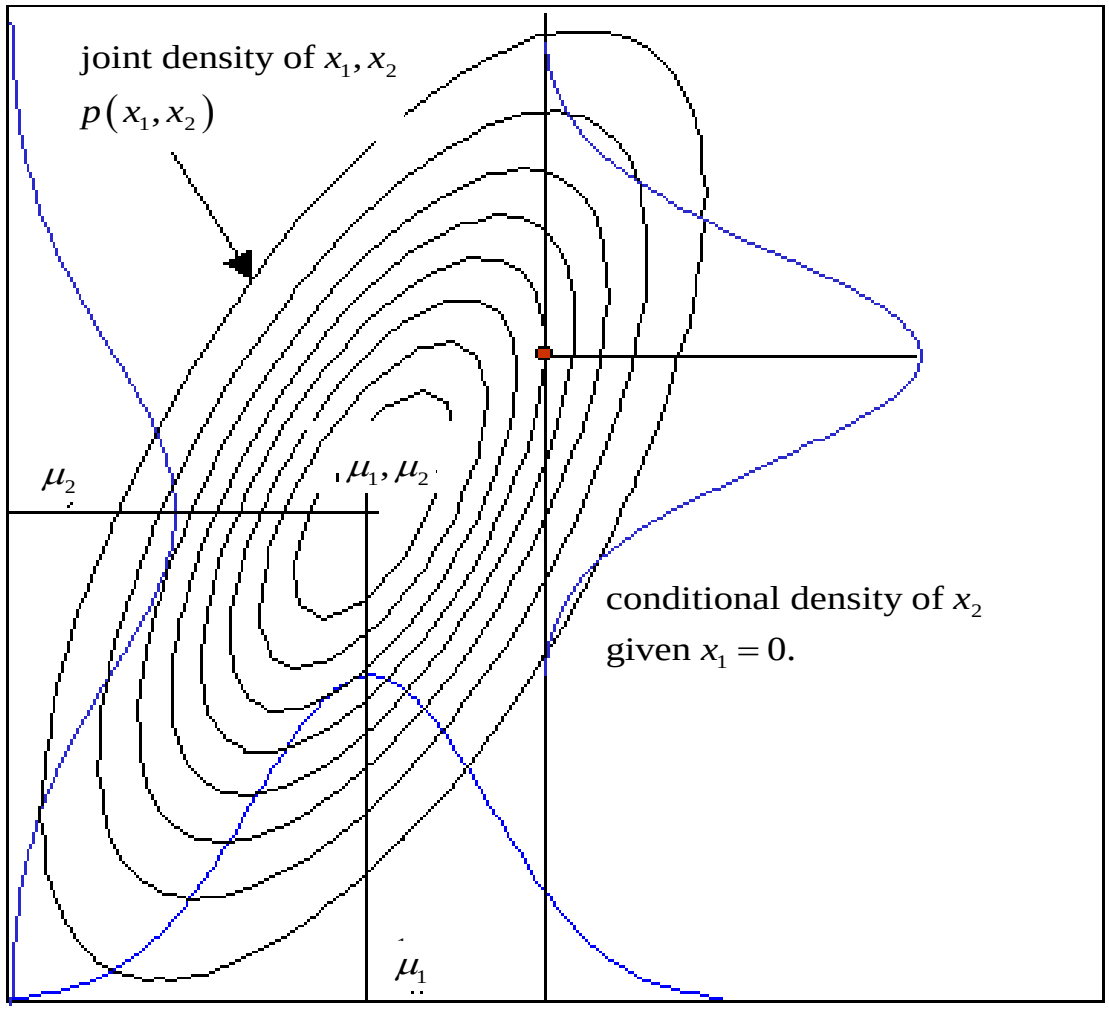
$$p(x_1) = \int p(x_1, x_2) dx_2$$

The conditional probability is given by:

$$p(x_2 \mid x_1) = \frac{p(x_1, x_2)}{p(x_1)} \Leftrightarrow p(x_2 \mid x_1) = \frac{p(x_1 \mid x_2) p(x_2)}{p(x_1)}$$

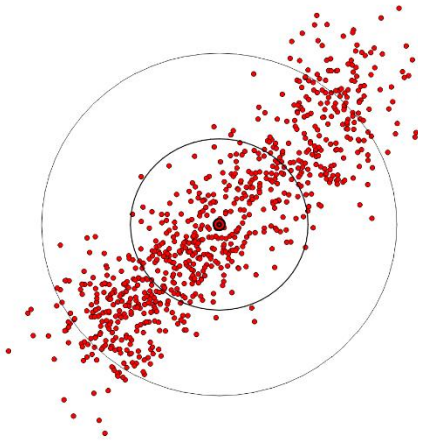
Marginal, Conditional Pdf of Gauss Functions

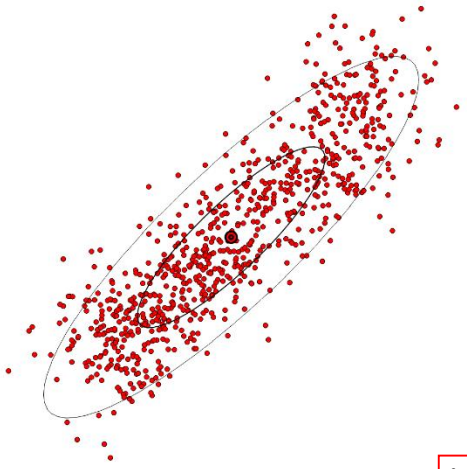
The conditional and marginal pdf of a multi-dimensional Gauss function are all Gauss functions!

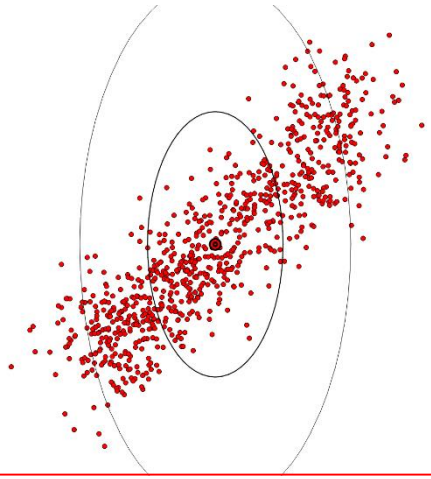


Illustrations from Wikipedia

Modeling Joint Densities from data using a Gaussian Distribution







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

if x is N-dimensional, then

μ is a $N-$dimensional mean vector

$\sum$ is a $N \times N$ covariance matrix

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 .

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

If the data are *independent and identically* distribution by *p,* then:

$$L(\Theta \mid X) = \prod_i^M p\left(x^1; \Theta\right)$$

The likelihood can be used to determine how well a particular model of *p(x)* models the data at hand.

Likelihood of Gaussian Pdf Paramatrization

Consider that the pdf of the dataset X is parametrized with parameters μ, Σ.

One writes: $p(X;\mu,\Sigma)$ or $p(X \mid \mu,\Sigma)$

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 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, search for parameters that maximize the likelihood.

Maximum Likelihood Optimization

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

For a multi-variate Gauss pdf-s, one can determine the mean and covariance matrix by solving:

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

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

Maximum Likelihood solutions:

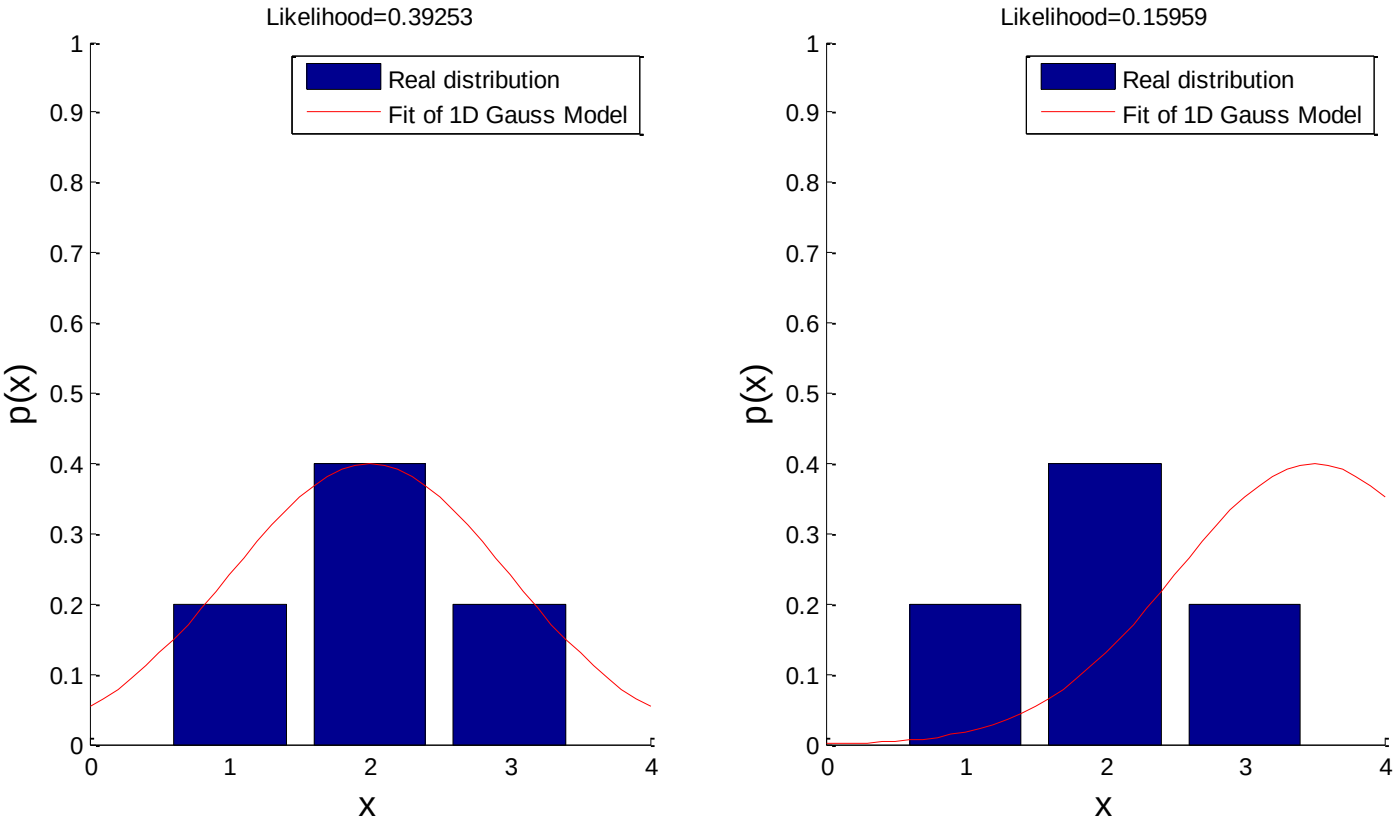
$$\mu_{ML} = \frac{1}{M}\sum_{i}^{M} x_i$$

$$\Sigma_{ML} = \frac{1}{M}\sum_{i}^{M}\left(x_i - \mu_{ML}\right)\left(x_i - \mu_{ML}\right)^T$$

If *p* is the Gauss pdf, then the above has an analytical solution (assuming that one has enough observations of x to draw the parameters from).

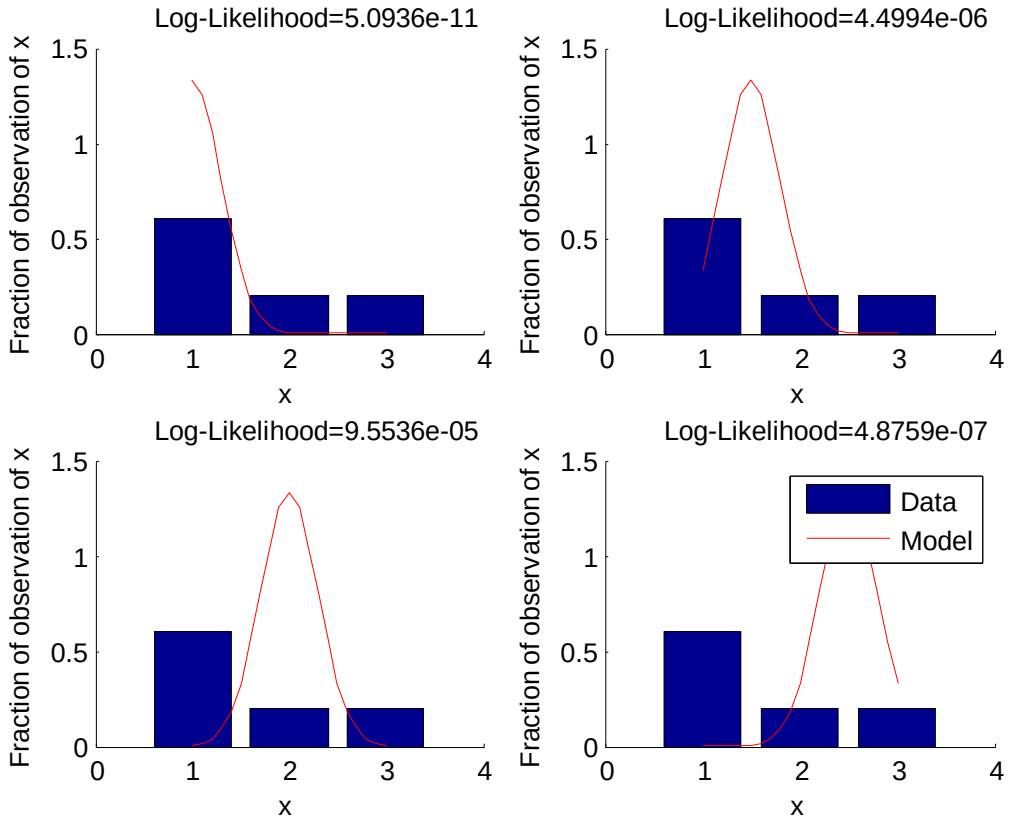
Usually instead of maximizing the likelihood, we minimize the negative log-likelihood

Likelihood Function



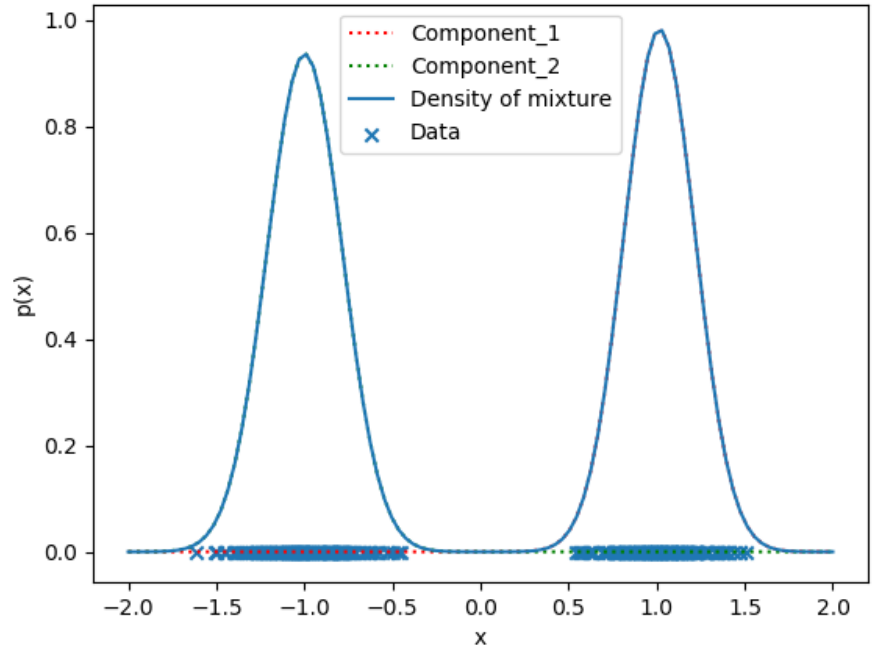
Values taken by the likelihood for two different fits using 1-D Gauss functions with different means.

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.

More complex density functions



Combination of K Gauss functions

$$p(x;\Theta) = \sum_{k=1}^{K} \alpha_k p_k\left(x;\mu^k,\Sigma^k\right) \text{ with } \Theta = \{\mu^1,\Sigma^1,\ldots\mu^K,\Sigma^K\},\ \alpha_k \in [0,1].$$

Linear weighted combination

K Gaussian Functions

Mixture of Gauss Functions

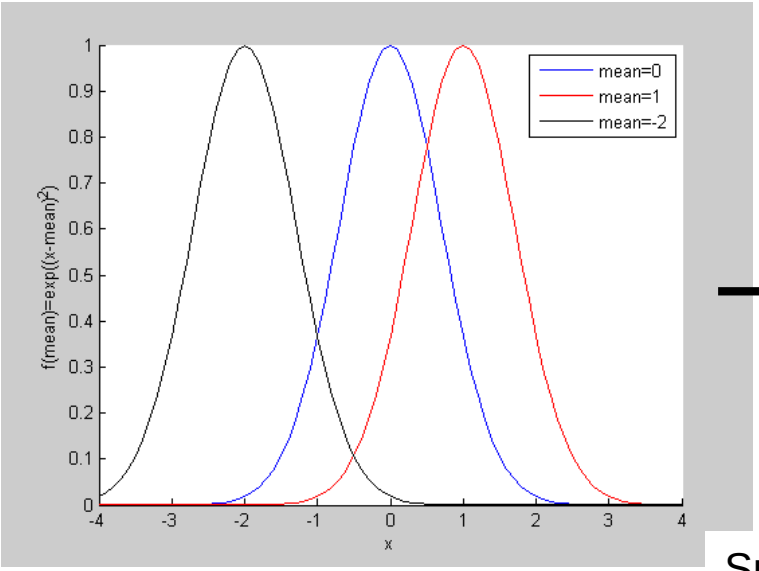
Combination of K Gauss functions

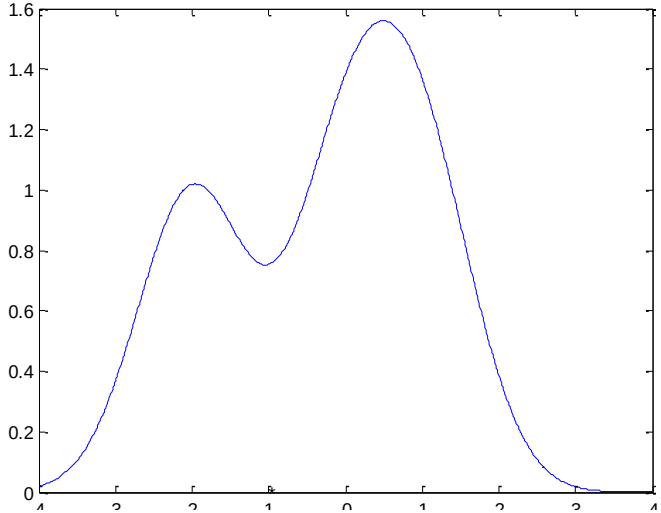
$$p(x;\Theta) = \sum_{k=1}^{K} \alpha_k p_k\left(x;\mu^k,\Sigma^k\right) \text{ with } \Theta = \{\mu^1,\Sigma^1,\ldots\mu^K,\Sigma^K\},\ \alpha_k \in [0,1].$$

Linear weighted combination

K Gaussian Functions

Here we use $K = 3,\ \alpha_1 = \alpha_2 = \alpha_3$





Superposition of the 3 Gauss functions with equal weight.

Mixture of Gauss Functions

Combination of K Gauss functions

$$p(x;\Theta)=\sum_{k=1}^{K}\alpha_k p_k\left(x;\mu^k,\Sigma^k\right) \text{ with } \Theta=\{\mu^1,\Sigma^1,\ldots\mu^K,\Sigma^K\},\ \alpha_k\in[0,1].$$

Linear weighted combination

K Gaussian Functions

To find the optimal parameters:

$$\max_{\Theta} L(\Theta\mid X)=\max_{\Theta} p(X\mid\Theta)$$

$$\max_{\Theta} L(\Theta\mid X)=\max_{\Theta}\prod_{i=1}^{M}\sum_{k=1}^{K}\alpha_k p_k\left(x^i;\mu^k,\Sigma^k\right)$$

No closed-form solution → Solve through Expectation-Maximization (E-M)
E-M is an *iterative* procedure to estimate the best set of parameters
It converges to **a local optimum** → Sensitive to initialization!

Expectation-Maximization (E-M)

EM is an iterative procedure:

0) Make a guess, pick a set of $\hat{\Theta}$ (initialization)

1) Compute likelihood $L\left(\hat{\Theta} \mid X, Z\right)$ (E-Step)

2) Update Θ by gradient ascent on $L\left(\Theta \mid X, Z\right)$

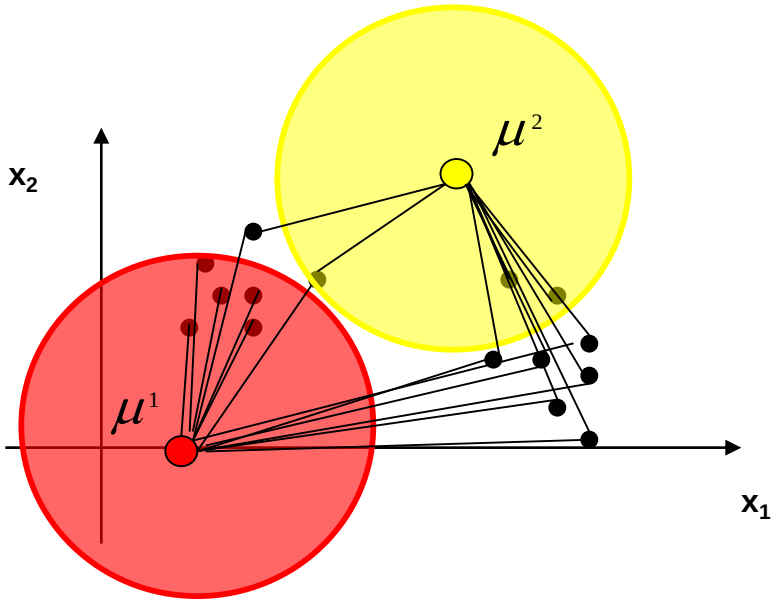
3) Iterate between steps 1 and 2 until reach plateau (no improvement on likelihood)

Ensured to converge to a local optimum only!
(see more details next slides)

From K-means Clustering to Density Modeling with Mixture of Gaussians

The algorithm of K-means is a simple version of Expectation-Maximization applied to a model composed of isotropic Gauss functions

K-means Clustering (probabilistic interpretation)



Computing the distance to the k-th centroid is equivalent to computing the probability that the data point has been generated by the k-th model.

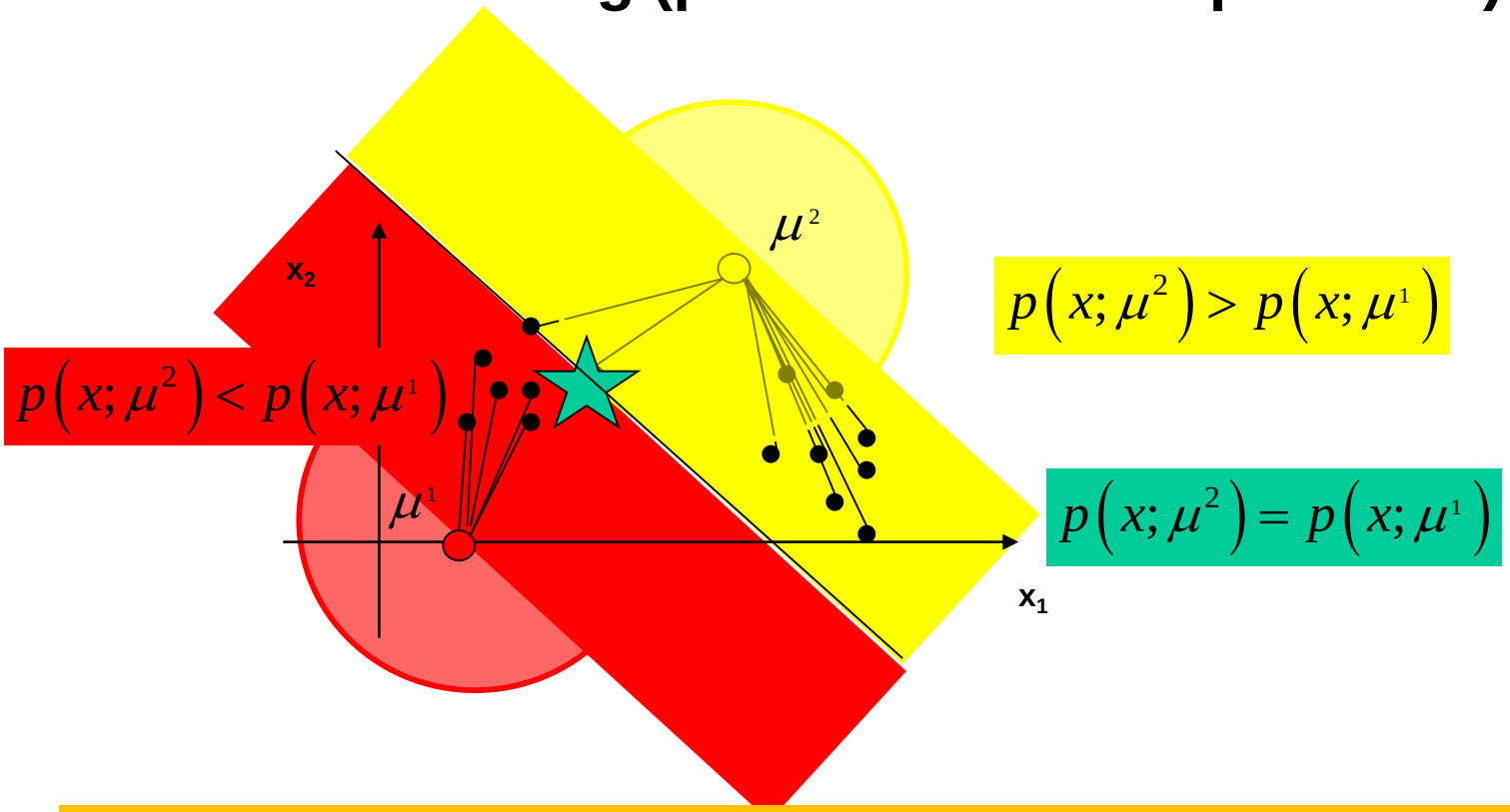
$$d\left(x^i, \mu^k\right) \sim p\left(x^i; \mu^k\right) = e^{-(x^i - \mu^k)^2}$$

The likelihood of the k-th model is:

$$L\left(X; \mu^k\right) = \prod_i e^{-(x^i - \mu^k)^2}$$

Assignment Step (E-step)**:**
~ Compute expectation of the equivalent Gaussian model with unity variance and centered on the centroid

K-means Clustering (probabilistic interpretation)



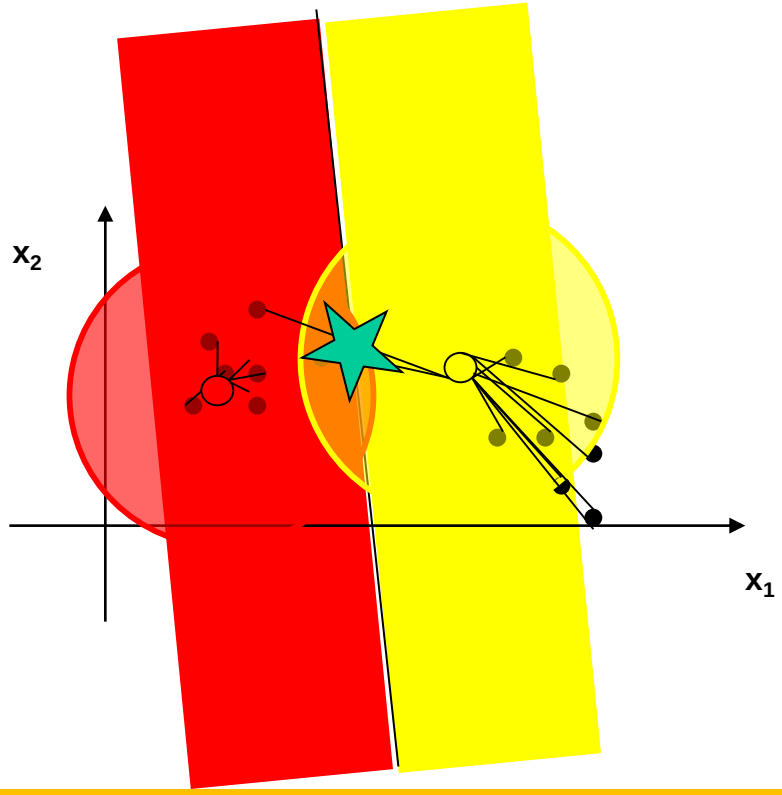
Assignment Step (E-step)**:**

~ Compute expectation of the equivalent Gaussian model with unity variance and centered on the centroid

- Assign the responsibility of each data point to its "closest" centroid.

~ Bayes' rule: If $p(x^i;\mu^2) > p(x^i;\mu^1)$, then x^i belongs to cluster 2.

K-means Clustering (probabilistic interpretation)



The new centroid is closer to the datapoints after the update step,

$$\mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k} \quad \Rightarrow d(x^i - \mu^k) \searrow$$

→ the likelihood of the k-th model increases.

$$L\left(X; \mu^k\right) = \prod_i e^{-(x^i - \mu^k)^2} \nearrow$$

Update Step (M-step)**:**
~ Maximize expectation of the equivalent Gaussian model with unity variance and centered on the centroid

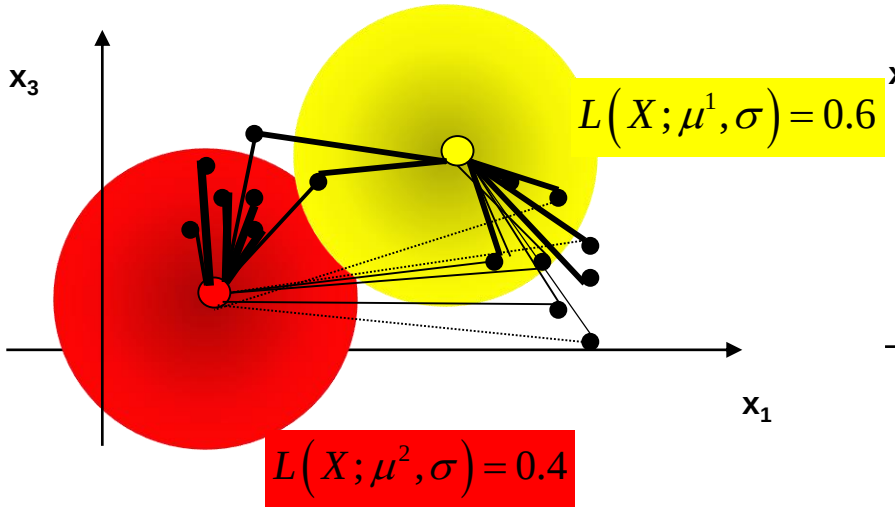
Soft-K-means (probabilistic interpretation)

Soft K-means can be seen as fitting the data distribution with a *mixture of isotropic (spherical) Gaussian pdf-s and same variance (the stiffness).*

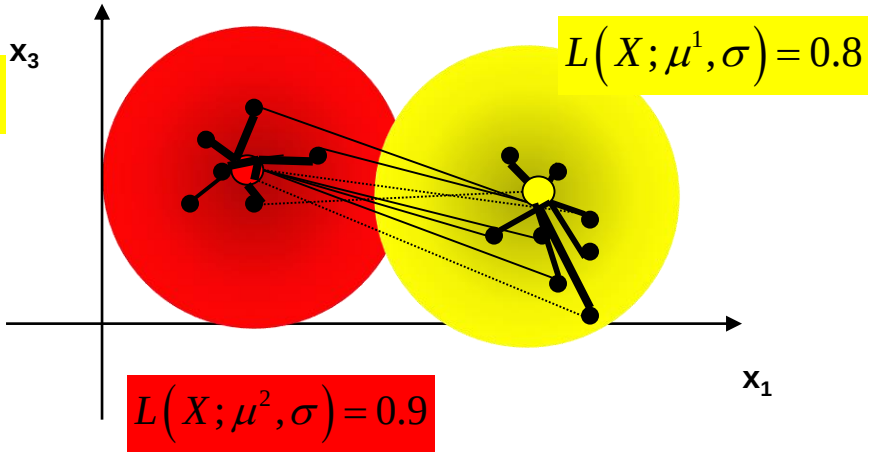
E-M updates the parameters of each Gaussian to optimize the likelihood that the Gaussians represent the distribution of the datapoints.

Likelihood of overall distribution (with uniform prior for each Gaussian):

$$L\left(X;\mu^1,\sigma,\mu^2,\sigma\right)=L\left(X;\mu^1,\sigma\right)+L\left(X;\mu^2,\sigma\right)$$



Poor fit



Better fit

Soft-K-means (probabilistic interpretation)

Assignment Step (E-step):

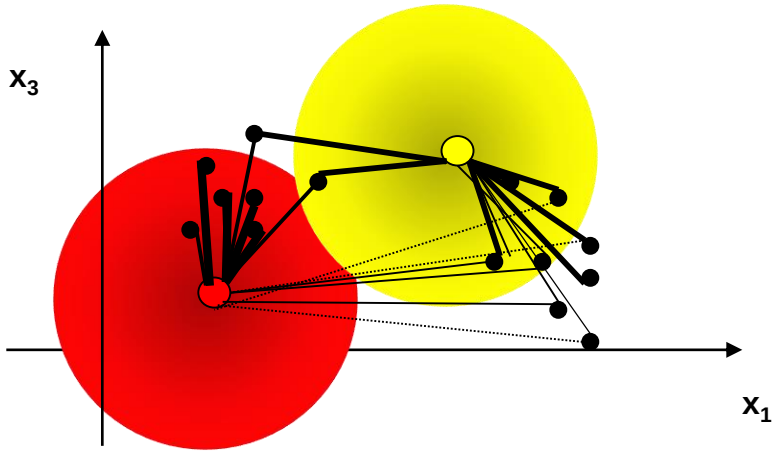
The responsibility factor gives a measure of the likelihood that cluster k generated the dataset.

r_i^k: responsibility of cluster k for point x^i

$$r_i^k = \frac{\alpha_k \, p\left(x^i; \mu^k, \sigma\right)}{\sum_{k'} \alpha_{k'} \, p\left(x^i; \mu^{k'}, \sigma\right)}, \qquad \alpha_k \in [0,1]$$

$p\left(x^i; \mu^k, \sigma\right) \in [0,1]$: Gauss pdf evaluated at x^i

Normalized over clusters: $\sum_k r_i^k = 1$



Relative importance of each of the K clusters (measure of number of datapoints in each cluster)

→ In GMM, we will see that this is a measure of the likelihood that the Gaussian *k* (or cluster *k*) generated the whole dataset.

One step towards Gaussian Mixture Model with Spherical Gaussians

Update Step (M-Step):

r_i^k: responsibility of cluster k for point x^i

$$r_i^k = \frac{\alpha_k \, p\left(x^i ; \mu^k, \sigma^k\right)}{\sum_{k'} \alpha_{k'} \, p\left(x^i ; \mu^{k'}, \sigma^{k'}\right)}, \qquad \alpha_k \in [0,1]$$

$p\left(x^i ; \mu^k, \sigma^k\right) \in [0,1]$: Gauss pdf evaluated at x^i

Normalized over clusters: $\sum_k r_i^k = 1$

$$\mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k}$$

$$\left(\sigma^k\right)^2 = \frac{\sum_i r_i^k \left\| x^i - \mu^k \right\|^2}{N \cdot \sum_i r_i^k}$$

$$\alpha_k = \frac{\sum_i r_i^k}{\sum_k \sum_i r_i^k}$$

Relative importance of each of the K clusters (measure of number of datapoints in each cluster)

→ In GMM, we will see that this is a measure of the likelihood that the Gaussian *k* (or cluster *k*) generated the whole dataset.

This fits a mixture of *spherical* Gaussians. The variance of each Gauss pdf fits the spread of the data around its mean.

From spherical to diagonal Gaussian pdf-s.

Update Step (M-Step):

r_i^k: responsibility of cluster k for point x^i

$$r_i^k = \frac{\alpha_k \, p\left(x^i; \mu^k, \sigma_j^k\right)}{\sum_{k'} \alpha_{k'} \, p\left(x^i; \mu^{k'}, \sigma_j^{k'}\right)}$$

$j = 1, \ldots N$: dimension of dataset

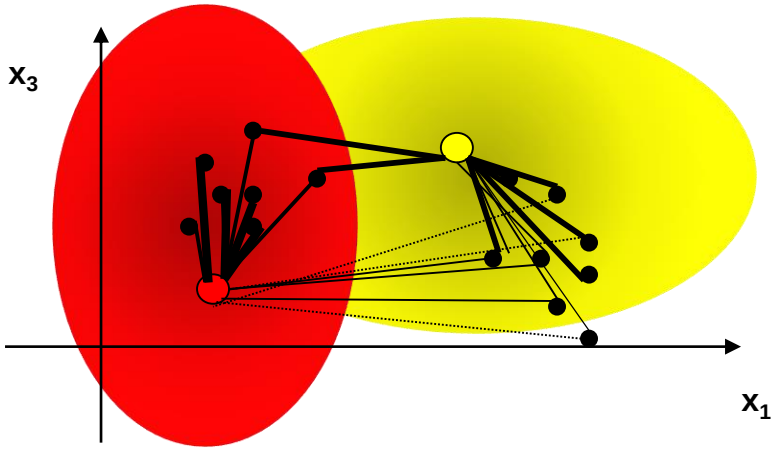
$p\left(x^i; \mu^k, \sigma^k\right) \in [0,1]$: Gauss pdf evaluated at x^i

Normalized over clusters: $\sum_k r_i^k = 1$

$$\mu^k = \frac{\sum_i r_i^k x^i}{\sum_i r_i^k}$$

$$\left(\sigma_j^k\right)^2 = \frac{\sum_i r_i^k \left(x_j^i - \mu_j^k\right)^2}{\sum_i r_i^k}$$

$$\alpha_k = \frac{\sum_i r_i^k}{\sum_k \sum_i r_i^k}$$



One covariance element per dimension, but still aligned with the axes of the original frame of reference.

Clustering with Mixture of Gaussians

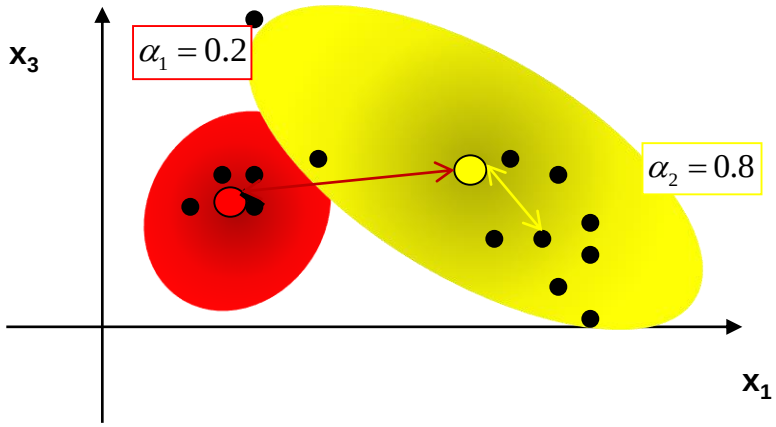
Likelihood of the mixture of Gaussians: $L\left(X;\left\{\mu^k,\Sigma^k\right\}_{k=1}^{K}\right)=\sum_{k=1}^{K}\alpha_k\cdot p\left(X;\mu^k,\Sigma^k\right)$

with $p\left(X;\mu^k,\Sigma^k\right)\sim\prod_{i=1}^{M}e^{-\left(x^i-\mu_k\right)^T\left(\Sigma^k\right)^{-1}\left(x^i-\mu_k\right)}$ (unormalized likelihood of Gaussian k)

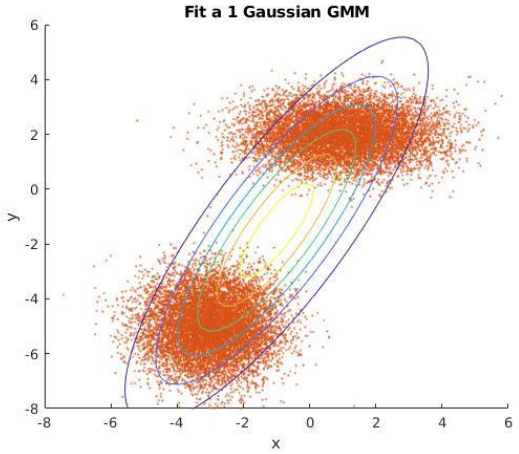
μ^k,Σ^k: mean and covariance matrix of Gaussian k

The mixing Coefficients are normalized. $\sum_{k=1}^{K}\alpha_k=1$

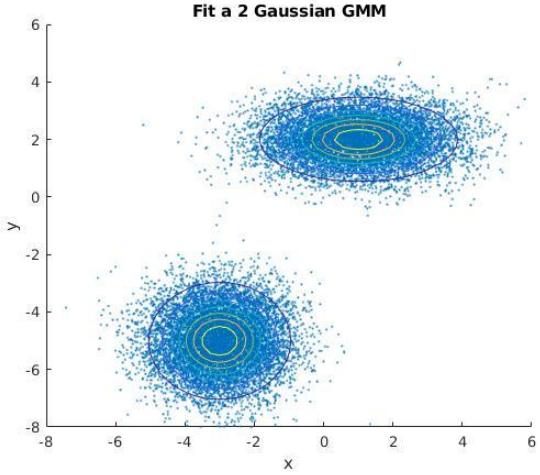
$$\alpha_k\sim\frac{1}{M}\sum_{i=1}^{M}\frac{p\left(x^i;\mu^k,\Sigma^k\right)}{\sum_k p\left(x^i;\mu^k,\Sigma^k\right)}$$



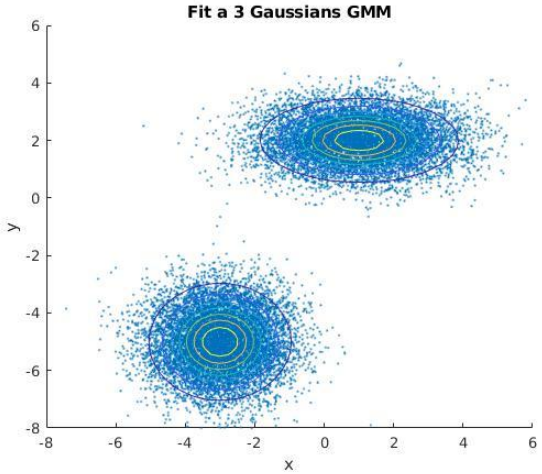
Clustering with Mixture of Gaussians



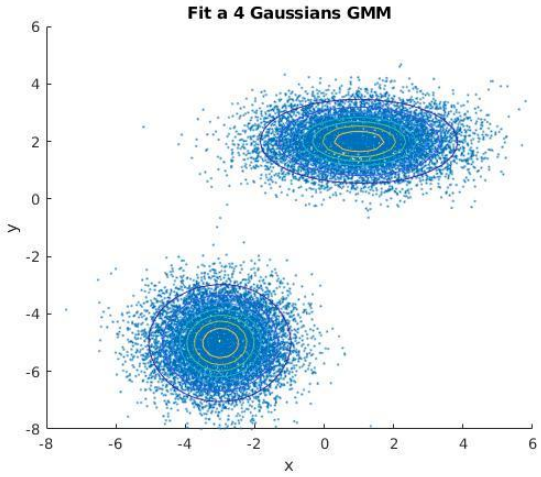
Log-Likelihood = -87825



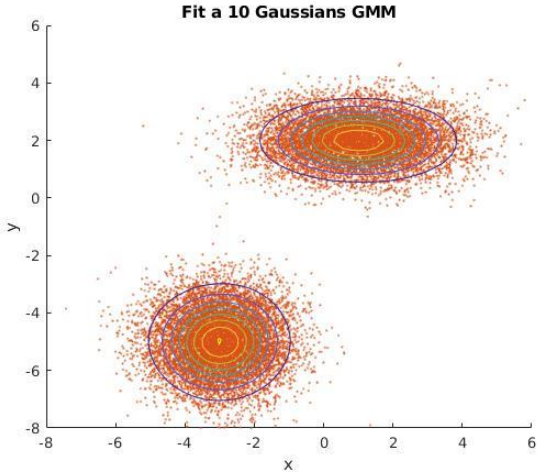
Log-Likelihood = - 70610



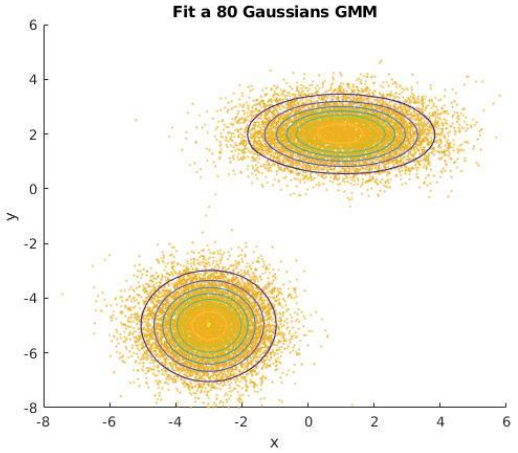
Log-Likelihood = -70610



Log-Likelihood = -70604



Log-Likelihood = -70601



Log-Likelihood = -70581

Gaussian Mixture Modeling with Expectation-Maximization

The parameters of a GMM are the means, covariance matrices and priors:

$$\Theta = \left\{ \mu^1, ...\mu^K, \sum^1, ...\sum^K, \alpha_1, ...\alpha_K \right\}$$

Estimation of all the parameters can be done through *Expectation-Maximization* (E-M). E-M tries to find the optimum of the likelihood of the model given the data, i.e.:

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

Expectation-Maximization

One usually can safely assume that the datapoints are i.i.d. (identically and independently distributed).

$$\max_{\Theta} \; p\left(X \mid \Theta\right) = \max_{\Theta} \; \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p\left(x^i; \mu^k, \Sigma^k\right)$$

Computing the log of the likelihood yields the same optimum:

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

$$\max_{\Theta} \; \log \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \cdot p\left(x^i; \mu^k, \Sigma^k\right) = \max_{\Theta} \; \sum_{i=1}^{M} \log \left(\sum_{k=1}^{K} \alpha_k \cdot p\left(x^i; \mu^k, \Sigma^k\right) \right)$$

No close-form solution unlike the case for one Gaussian.
See derivation of E-M for GMM in the annexes posted on the website

E-M Steps for GMM

Initialization:

The priors $\alpha_1,..,\alpha_k$ can be uniform for starters.

The means $\mu^1,..,\mu^K$ can be initialized with K-means.

Calculate the initial value of the likelihood

$$p\left(X \mid \Theta^{(t)}\right) = \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_{k(t)} \cdot \mathcal{N}\left(x^i; \mu^{k(t)}, \Sigma^{k(t)}\right)$$

Expectation Step (E-step):

Evaluate responsibilities of each cluster *k* over each sample x^i *using current parameters*

$$r_i^k = \frac{\alpha_k \mathcal{N}\left(x^i; \mu^k, \Sigma^k\right)}{\sum_{k'} \alpha_{k'} \mathcal{N}\left(x^i; \mu^{k'}, \Sigma^{k'}\right)}$$

E-M Estimate for Gaussian Mixture Models

Maximization (Update step) Step (M-step):

Recompute the means, covariances matrices and prior probabilities so as to maximize the log – likelihood of the current estimate : $\log\left(L\left(\Theta^{(t)} \mid X\right)\right)$

$$\mu_k^{(t+1)} = \frac{1}{M_k}\sum_{i=1}^{M} r_k^i x_i$$

$$\alpha_k^{(t+1)} = \frac{M_k}{M}$$

$$\sum_k^{(t+1)} = \frac{1}{M_k}\sum_{i=1}^{M} r_k^i (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T$$

where: $M_k = \sum_{i=1}^{M} r_k^i$

The E and M steps alternate until the log-likelihood reaches a plateau

Clustering Steps of E-M for Gaussian Mixture Models

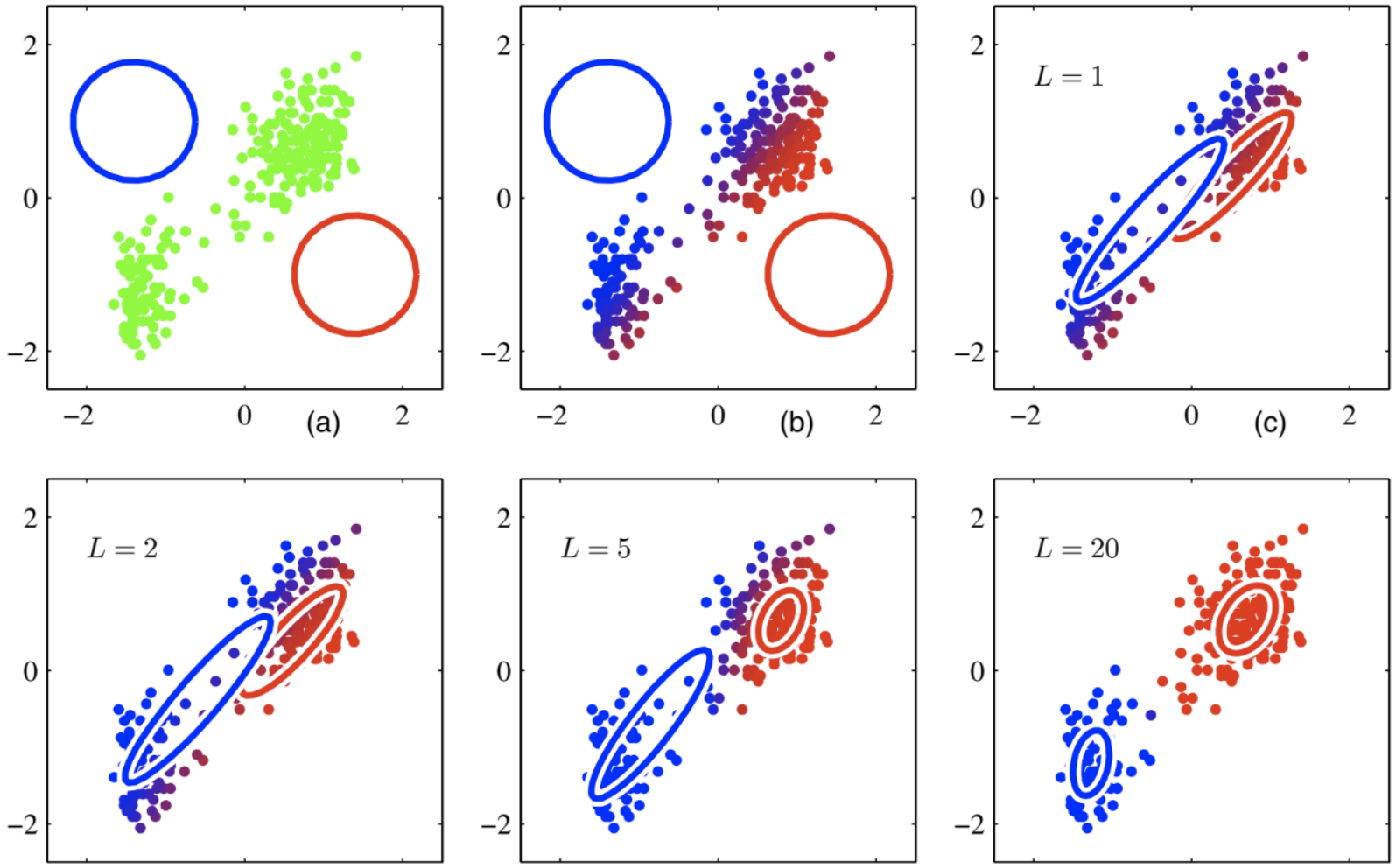
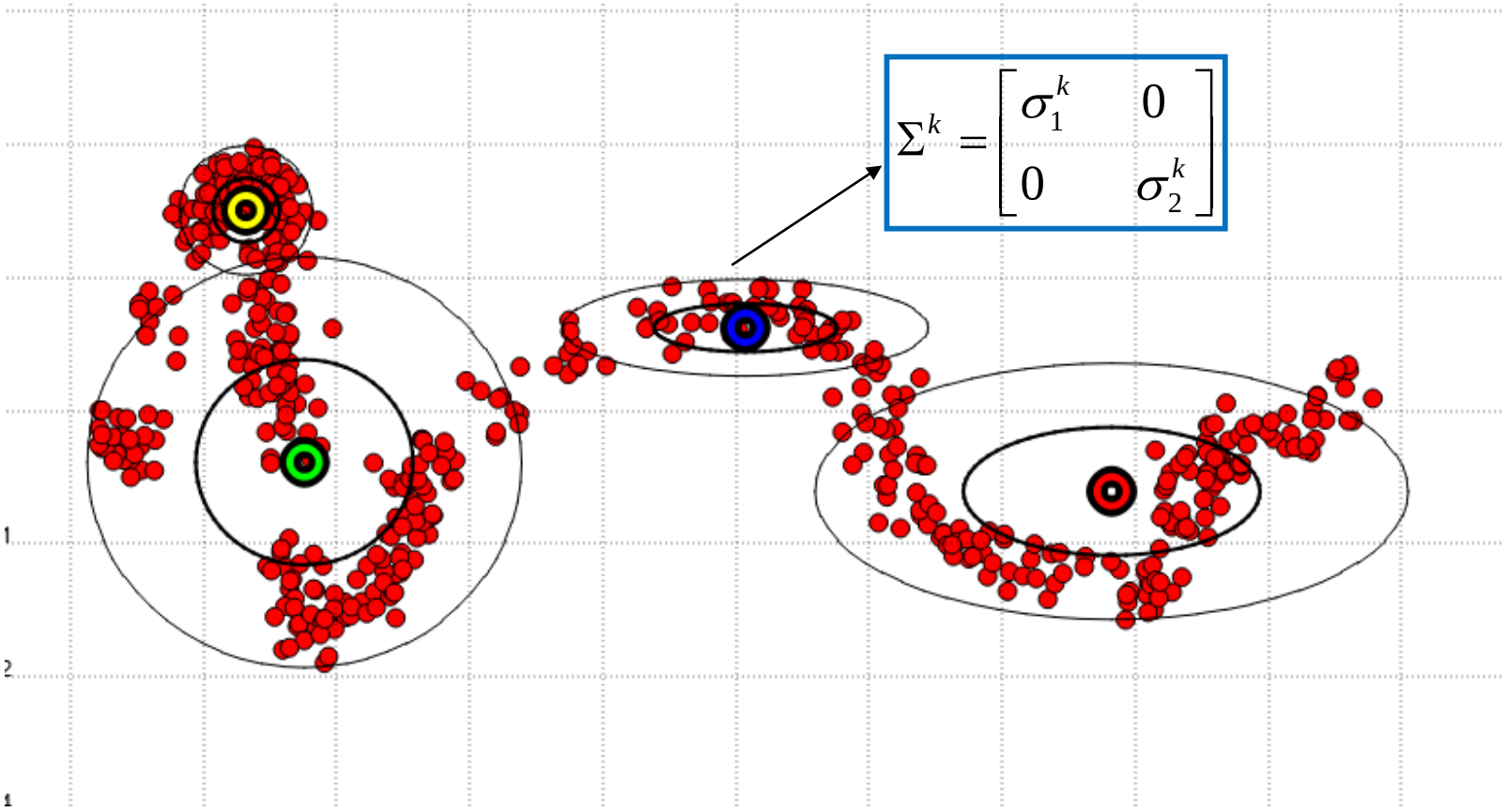


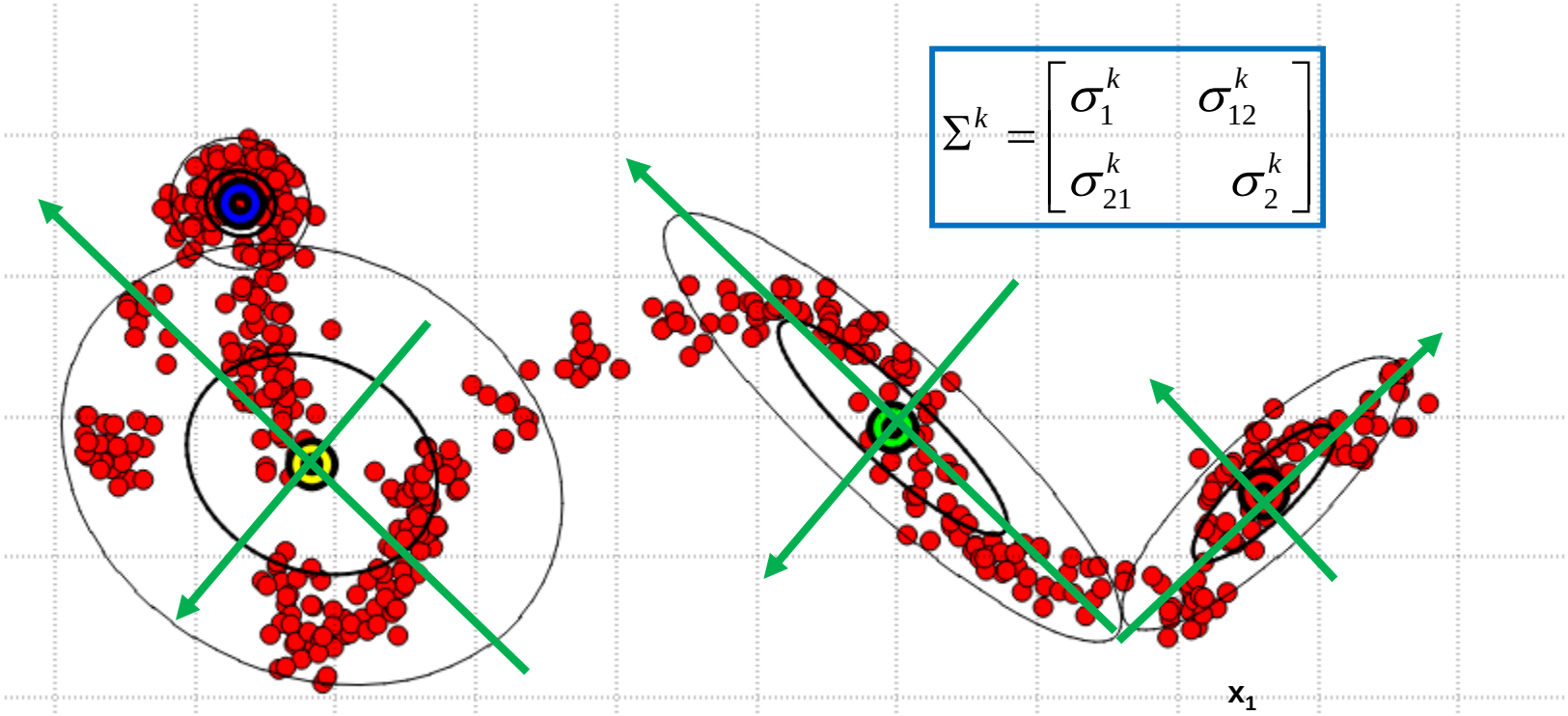
Image from : “Pattern Recognition and Machine Learning”, C.Bishop p.[437]

Fitting data with a Diagonal Mixture of Gaussians



Mixture of diagonal Gaussians (i.e. the covariance matrices of the Gaussians are diagonal) can only fit Gaussians **whose axes are aligned with the data axes**.

Fitting data with Mixtures of Full Gaussians



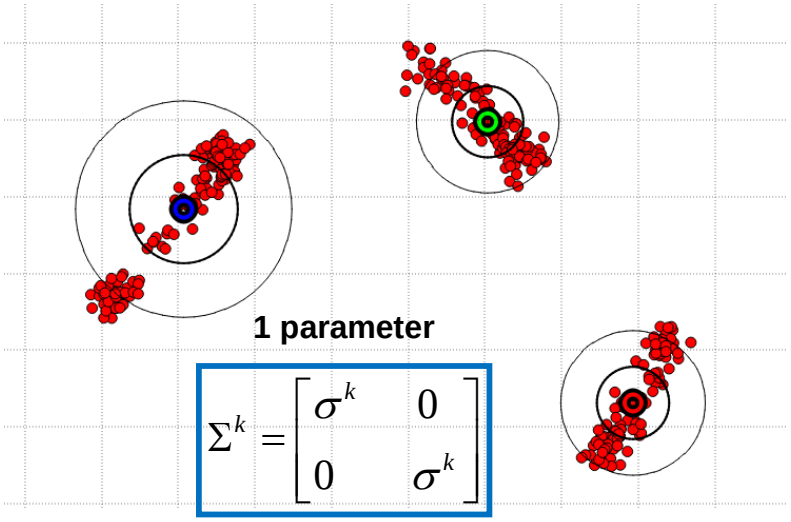
Gaussian Mixture Models (GMM) can learn mixtures of Gaussians with arbitrary (full) covariance matrices.

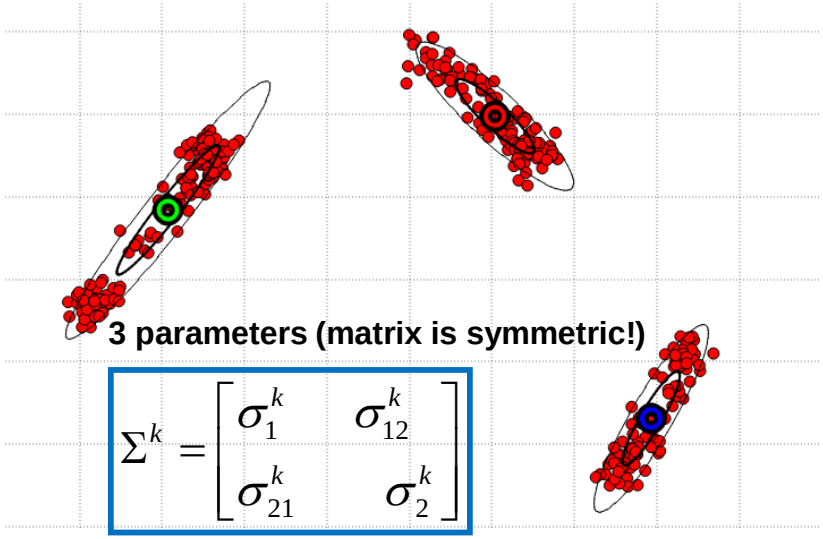
→ Gaussian Mixture Model can exploit local correlations and adapt the covariance matrix of each Gaussian so that it aligns with the local direction of correlation.

→ Each Gaussian performs a local linear PCA.

Tradeoff between computation costs and better fit

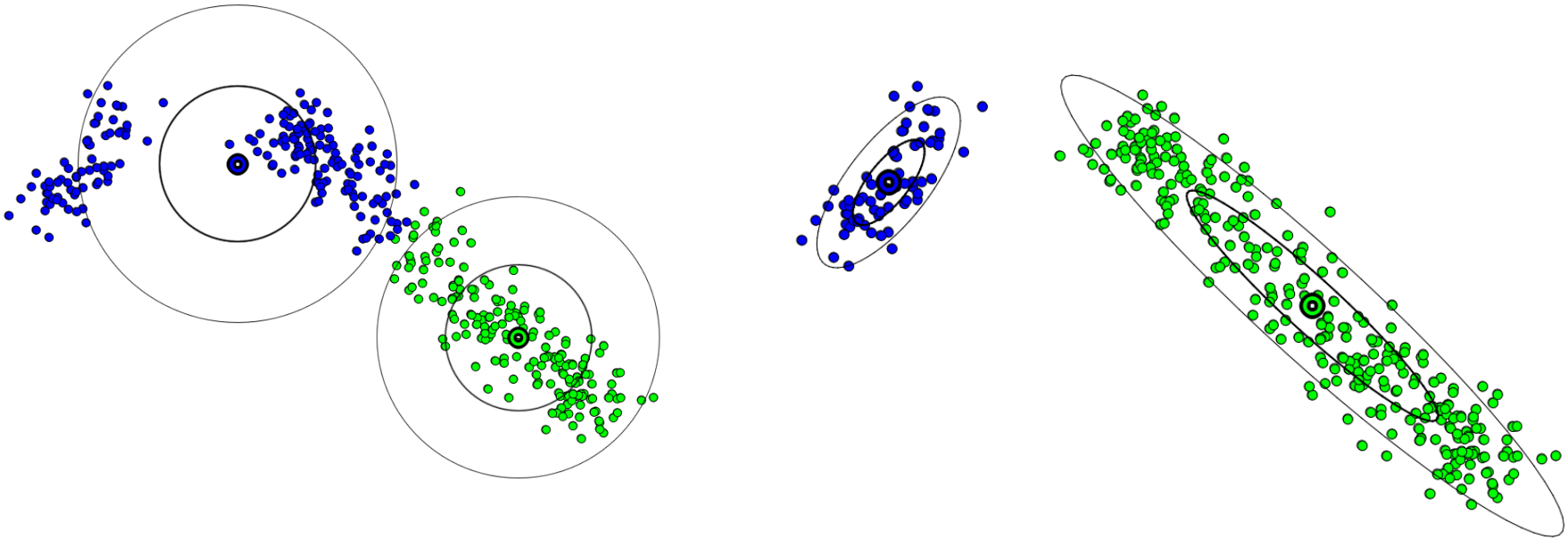
- In addition to better fit the local non-linearities, GMM may also reduce the number of Gaussians required to fit the data.
- But this comes at the cost of an increase in the number of parameters: *Full covariance matrices require N*(N+1)/2 parameters against N for diagonal matrices and 1 for spherical matrices.*





How to derive an algorithm for fitting data with complex mixtures of Gaussians?

Clustering with Mixtures of Gaussians



Clustering with Mixtures of Gaussians using spherical Gaussians (left) and non spherical Gaussians (i.e. with full covariance matrix) (right).

Notice how the clusters become elongated along the direction of the clusters (the grey circles represent the first and second variances of the distributions).

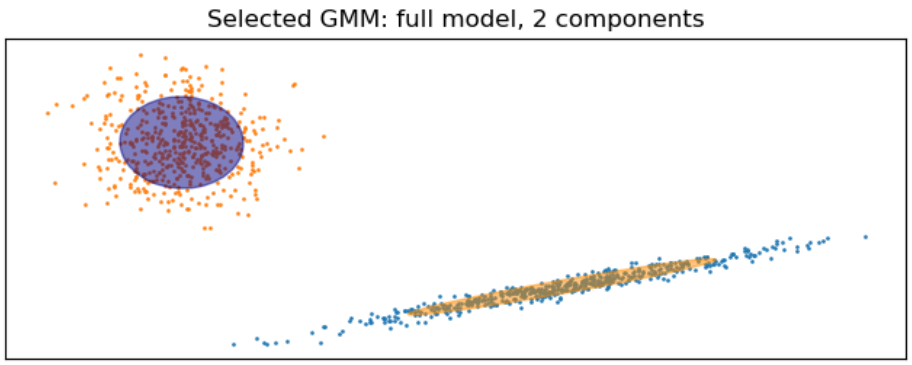
Hyper-parameter optimization in GMMs

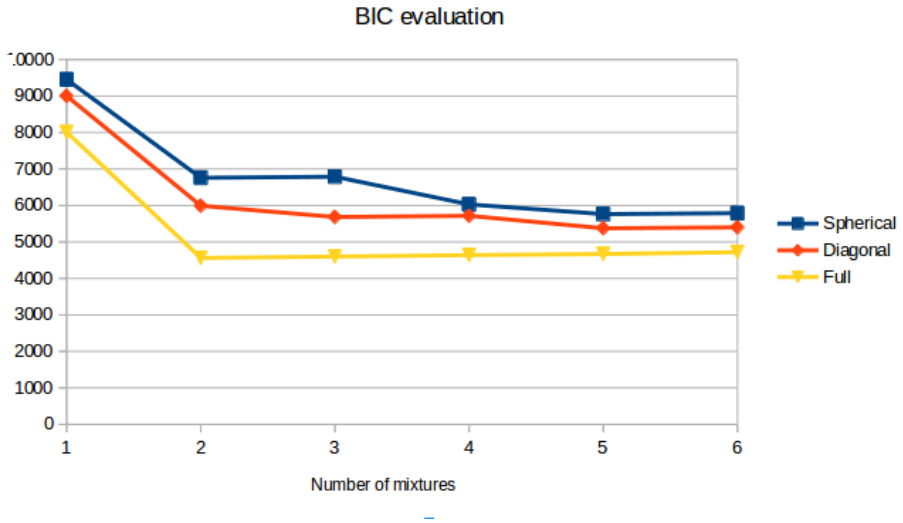
The selection is performed using the same criteria and way as k-means

$$AIC = -2\ln(L) + 2B$$

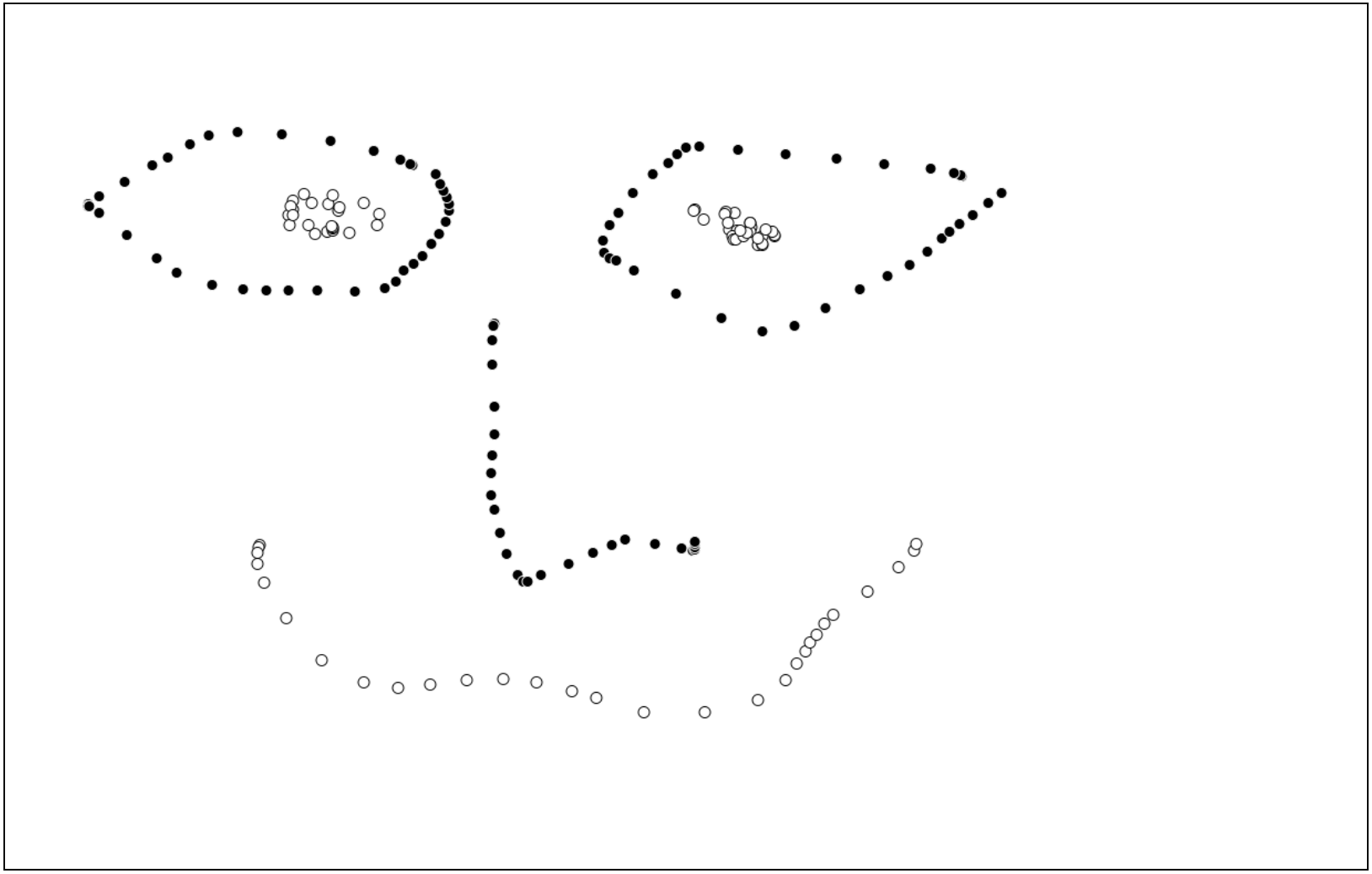
$$BIC = -2\ln(L) + \ln(M)B$$

But in the case of GMM, you should evaluate different choices of covariance matrix

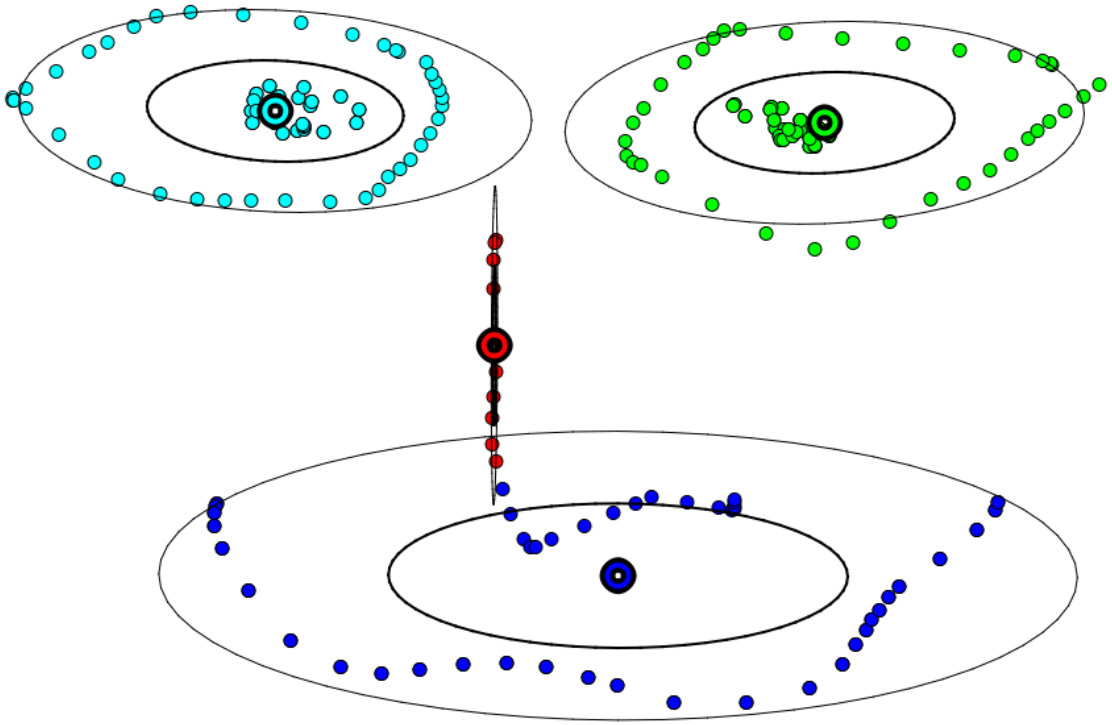




Gaussian Mixture Model



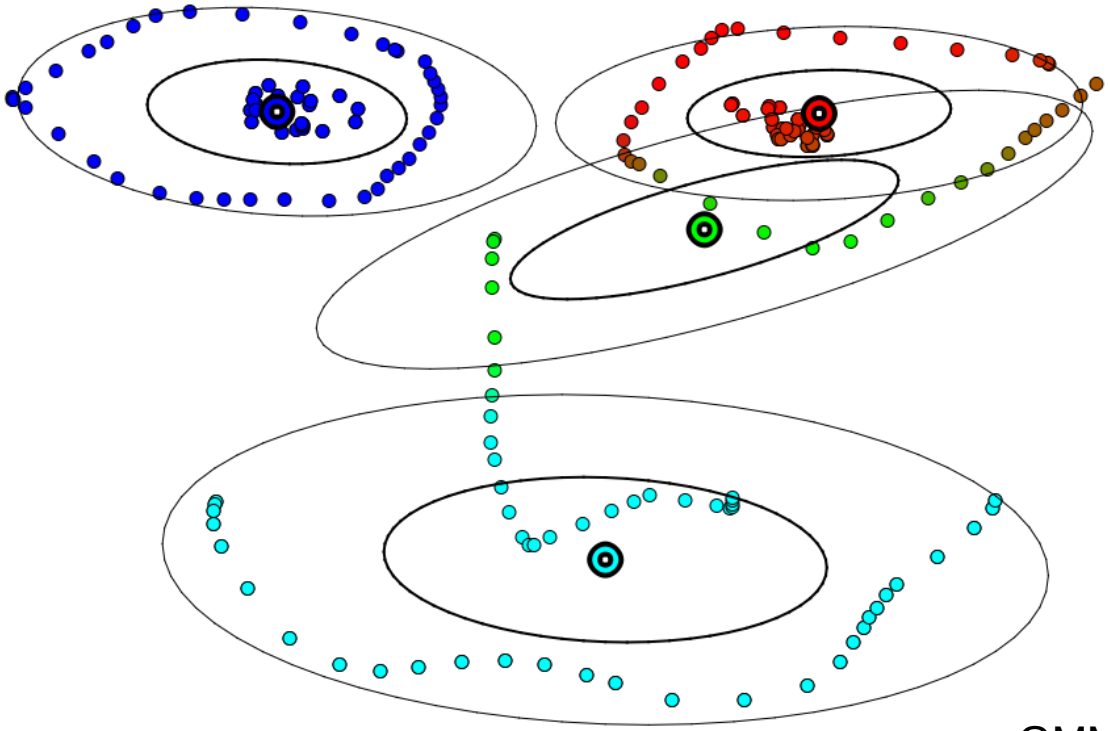
Gaussian Mixture Model



GMM using 4 Gaussians
with random initialization

Gaussian Mixture Model

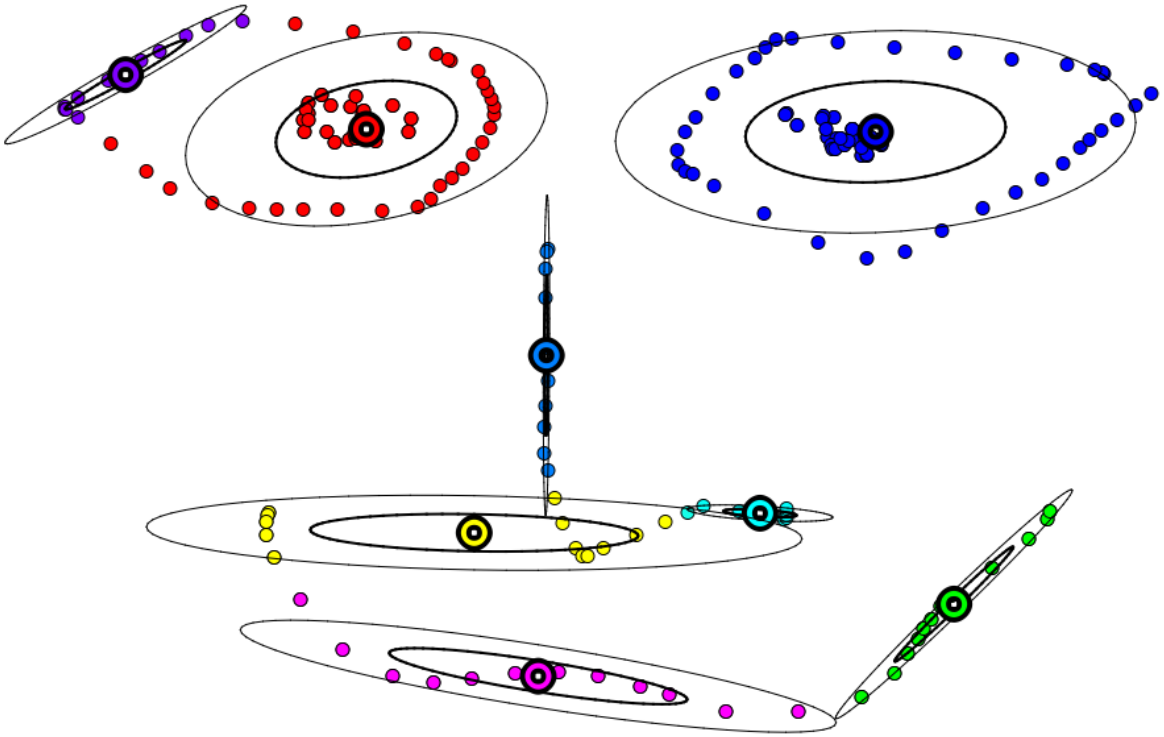
Expectation Maximization is very sensitive to initial conditions:



GMM using 4 Gaussians with new random initialization

Gaussian Mixture Model

Very sensitive to choice of number of Gaussians. Number of Gaussians can be optimized iteratively using AIC or BIC (see later slides):



Here, GMM using 8 Gaussians

Summary of this lecture

This class revisited some basic notions of statistics, with standard definitions of pdf, cdf, marginal and conditional distributions.

It emphasized the notion of statistical independence and how one can recognize it numerically and geometrically from looking at the distributions.

It exemplified these concepts with multi-dimensional Gaussian functions.

Finally, it introduced the notion of maximum likelihood fit first with one Gaussian and then with a mixture of Gaussians.