APPLIED MACHINE LEARNING

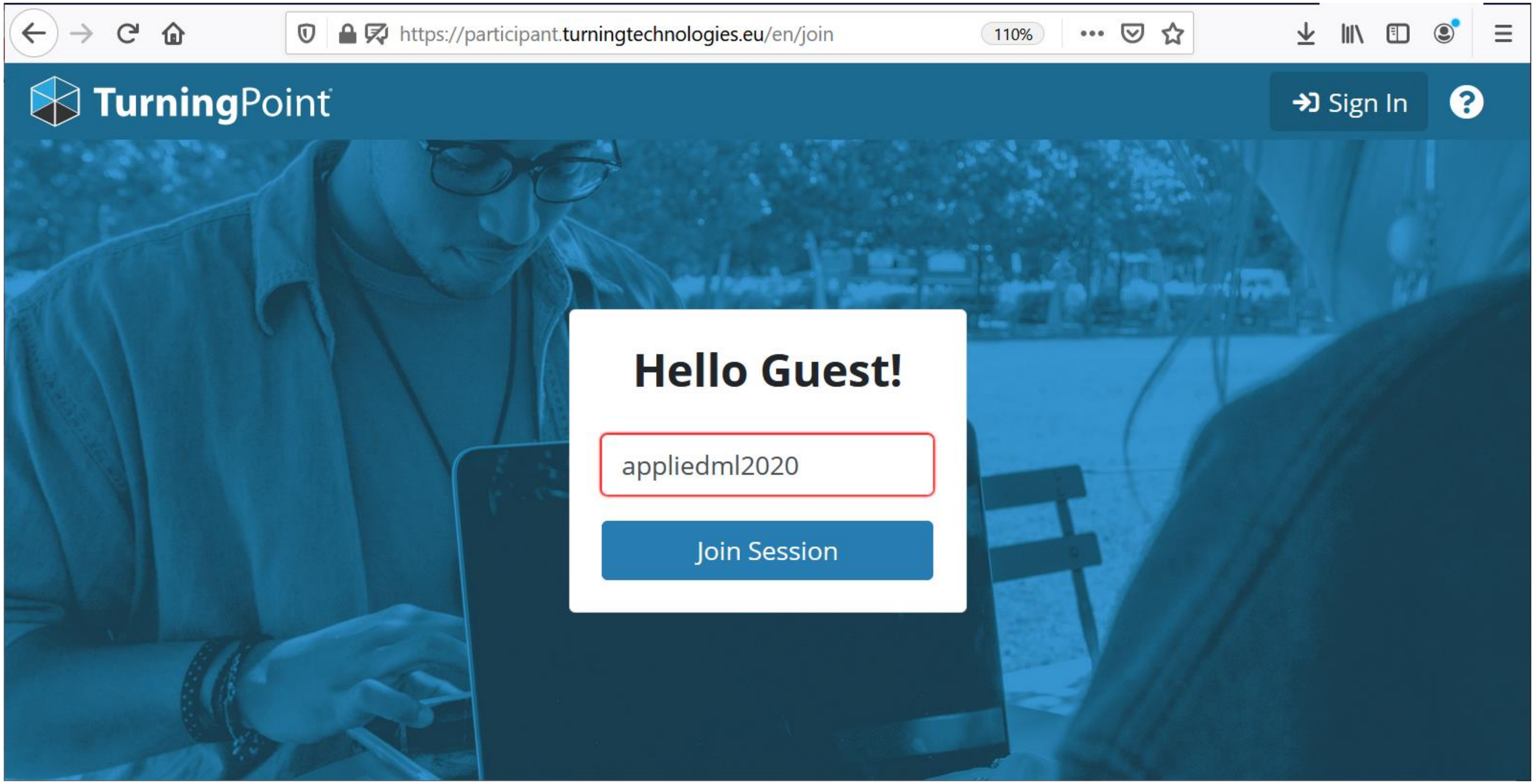
Pdf, GMM and E-M

Interactive Lecture

Launch polling system

https://participant.turningtechnologies.eu/en/join

Acces as GUEST and enter the session id: *appliedml2020*

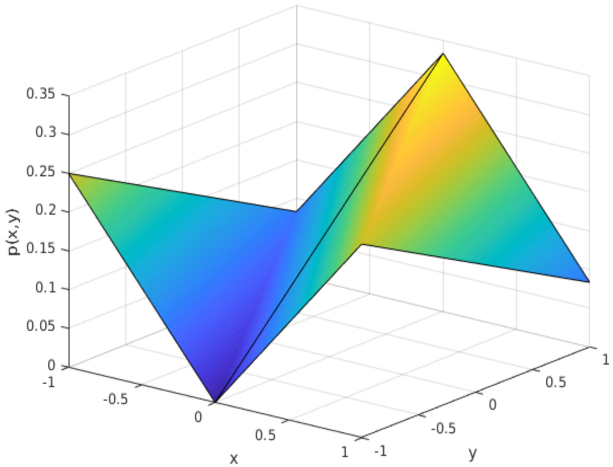


Statistical Independence and uncorrelatedness

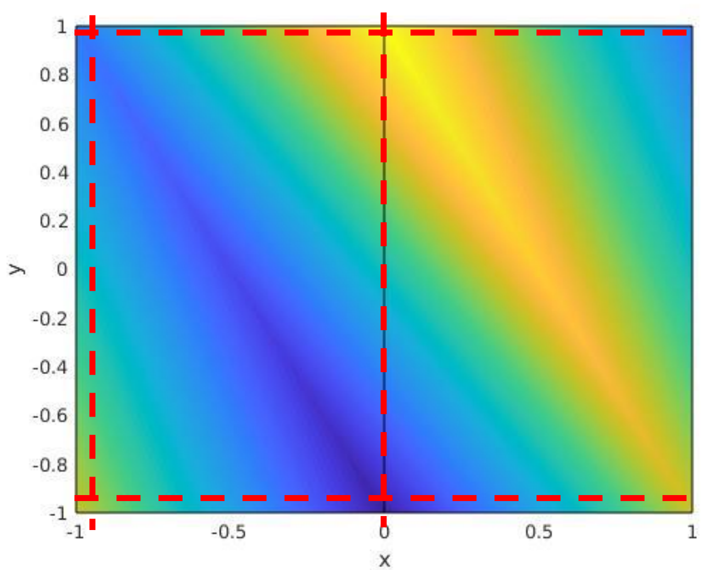
Are x,y correlated?
Are x,y statist. dependent?

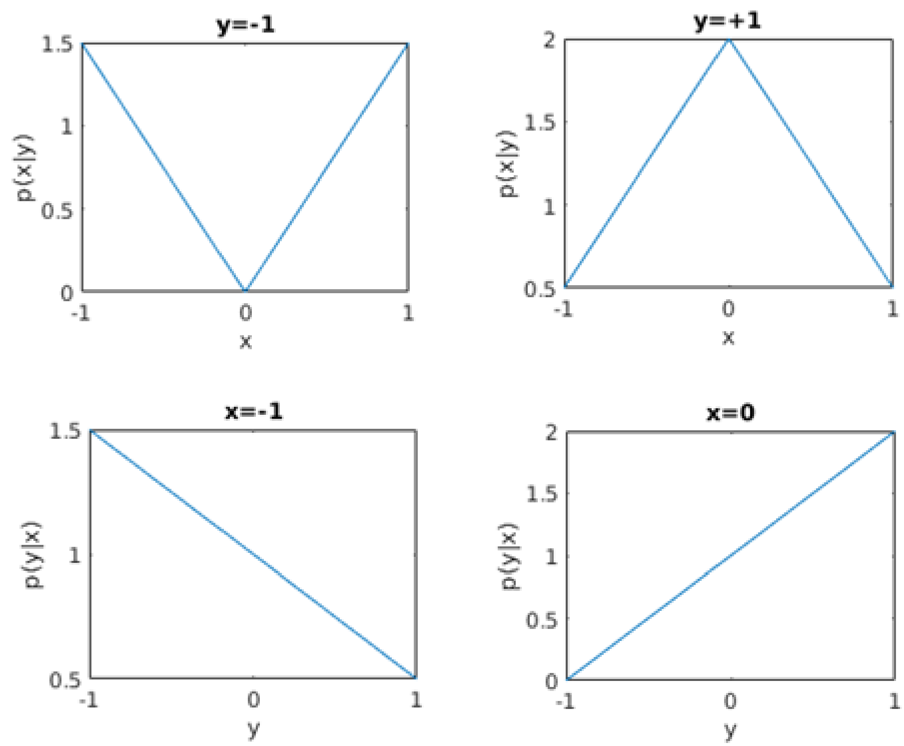
Uncorrelated: $E\{x,y\} = E\{x\}E\{y\}$

Statistical Ind.: $p(x,y) = p(x)p(y)$



$$E\{x,y\} = E\{x\}E\{y\} = 0$$





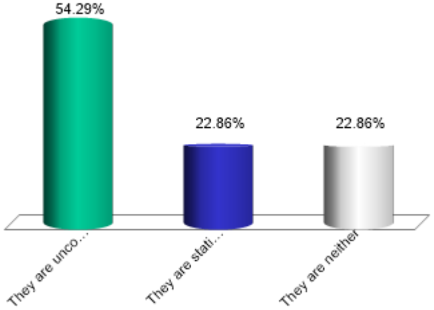
Statistical Independence and uncorrelatedness

Joint probabilities over two variables x_1, x_2

	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**	

Are x_1 and x_2 uncorrelated and statistically independent?

A. They are uncorrelated
B. They are statistically independent
C. They are neither



Statistical Independence and uncorrelatedness

Joint probabilities over two variables x_1, x_2

	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**	

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

$$\sum_{i,j=1}^{3} (x_1 = i, y = j)\, p\big((x_1 = i), (y = j)\big) = \underbrace{\sum_{i=1}^{3} (x_1 = i)\, p\big((x_1 = i)\big)}_{0} \underbrace{\sum_{i=1}^{3} (x_2 = i)\, p\big((x_2 = i)\big)}_{0}$$

Both sums are zero $\Rightarrow$ x_1, x_2 : uncorrelated

$$p(x_1 = -1, x_2 = 1) = 3/12 = 0.25$$

$$\neq$$

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

x_1 and x_2 are uncorrelated but not statistically independent.

Statistical Independence and uncorrelatedness

Independent $\rightleftarrows$ **Uncorrelated**

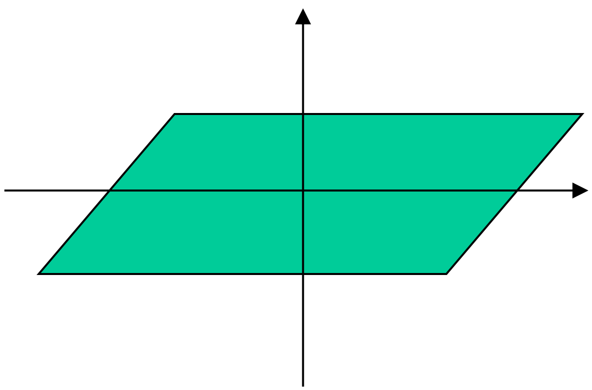
$$p(x_1, x_2) = p(x_1)\, p(x_2) \quad \Rightarrow \quad E\{x_1, x_2\} = E\{x_1\} E\{x_2\}$$

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

Statistical independence ensures uncorrelatedness.
The converse is not true.

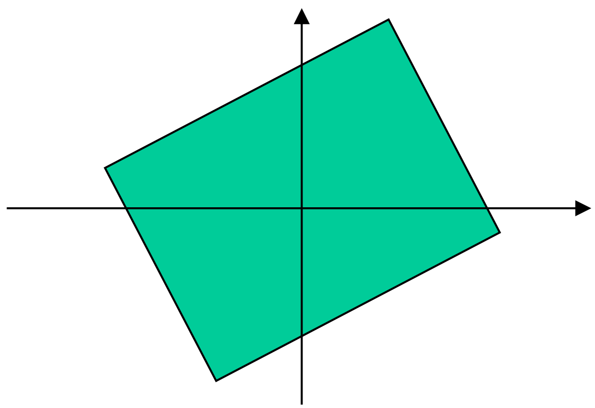
ICA: Preprocessing – Whitening & Independent Component Identification

Original Distribution



Uncorrelated distribution:

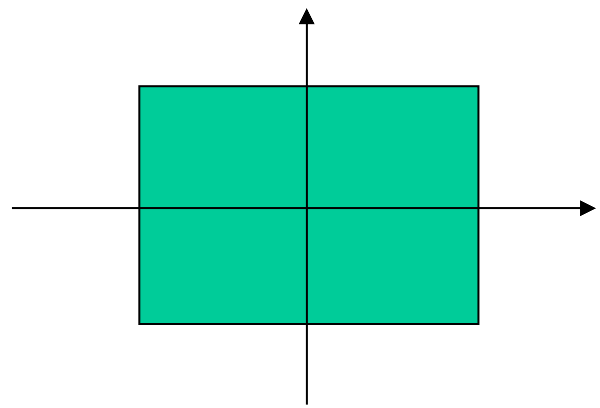
$E\{x_1, x_2\} = E\{x_1\}\, E\{x_2\}$



Whitening preprocessing:

$E\{XX^T\} = I$

Statistically Indep. Distr.



After projection on independent components

Linear Correlation

Two variables x_1, x_2 are correlated if:

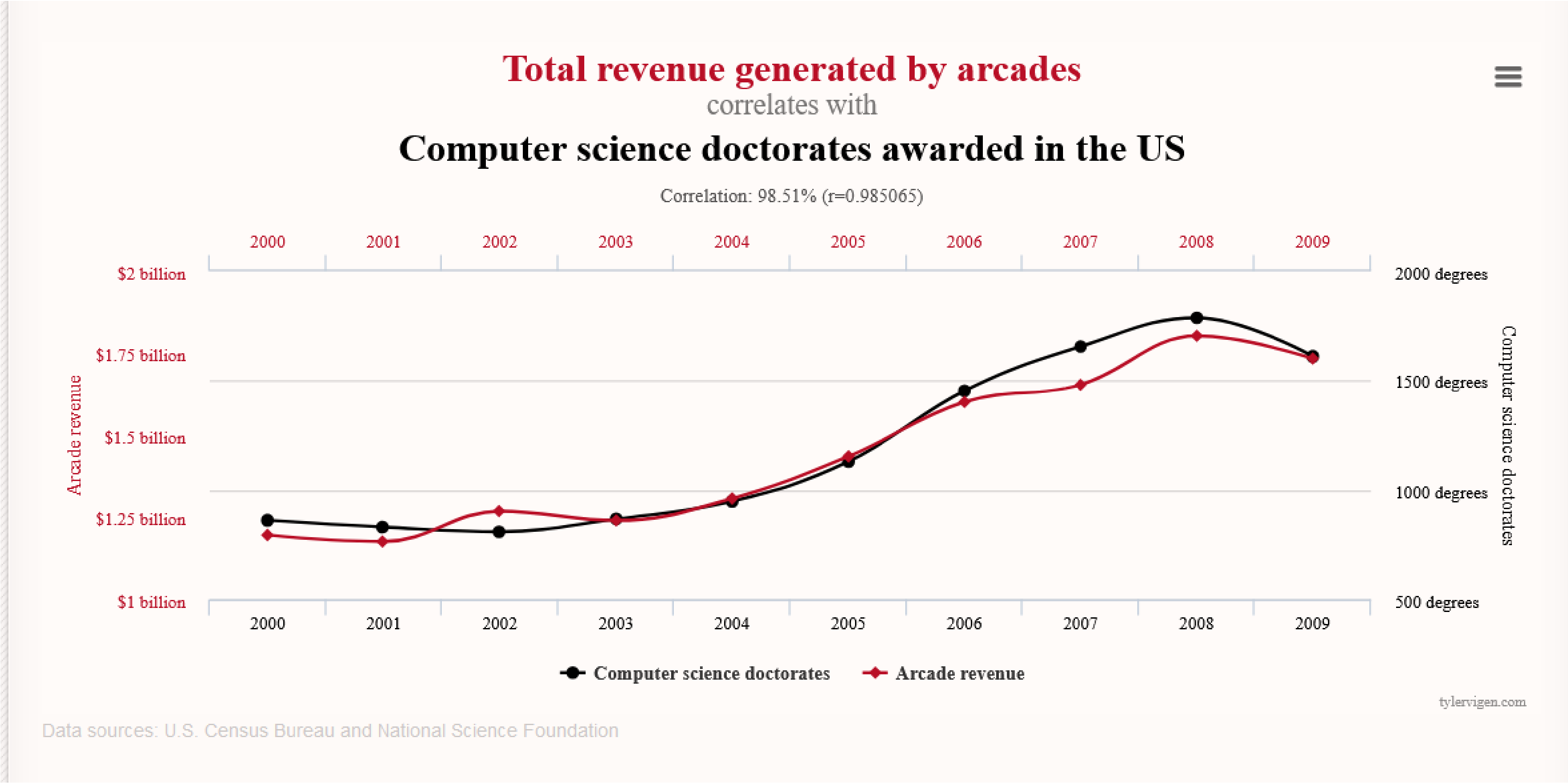
$$corr\left(x_1, x_2\right) = \frac{\mathrm{cov}\left(x_1, x_2\right)}{\mathrm{var}\left(x_1\right)\mathrm{var}\left(x_2\right)} \neq 0$$

$corr\left(x_1, x_2\right) >< 0:$ positive / negative correlation

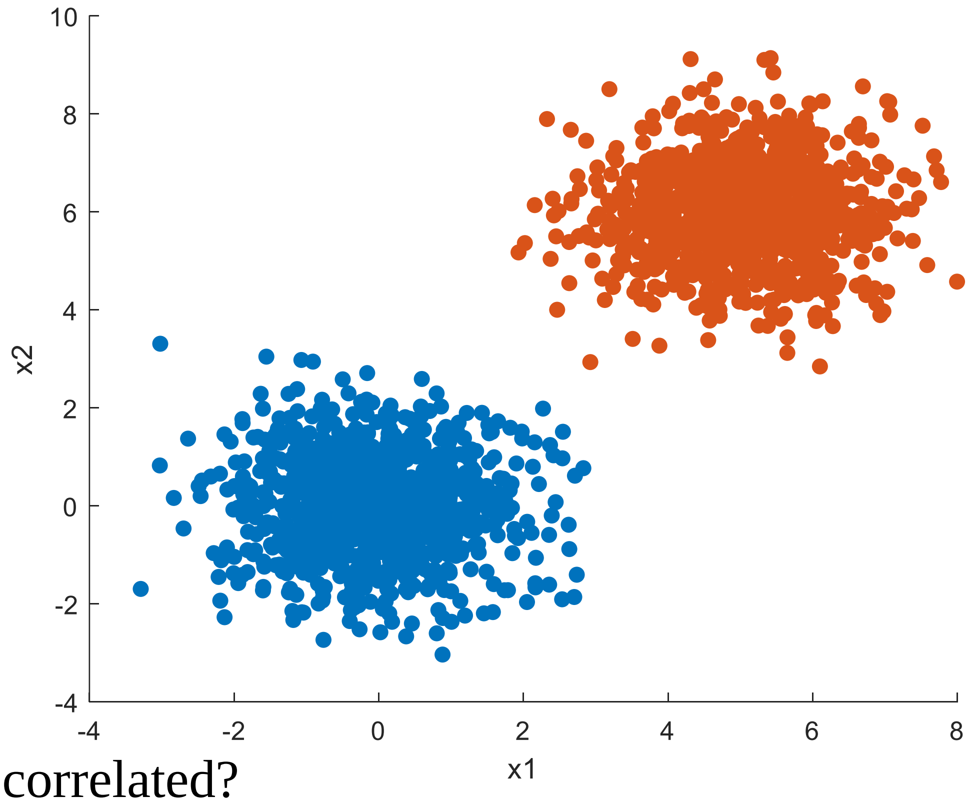
$\left|corr\left(x_1, x_2\right)\right| = 1:$ perfectly correlated

$\left|corr\left(x_1, x_2\right)\right| < 0.5$ weakly correlated

Real and Spurious Correlations



Spurious Correlations, T. Vygen

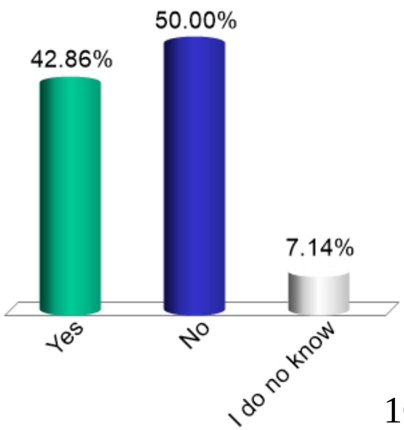


Are x_1 and x_2 correlated?

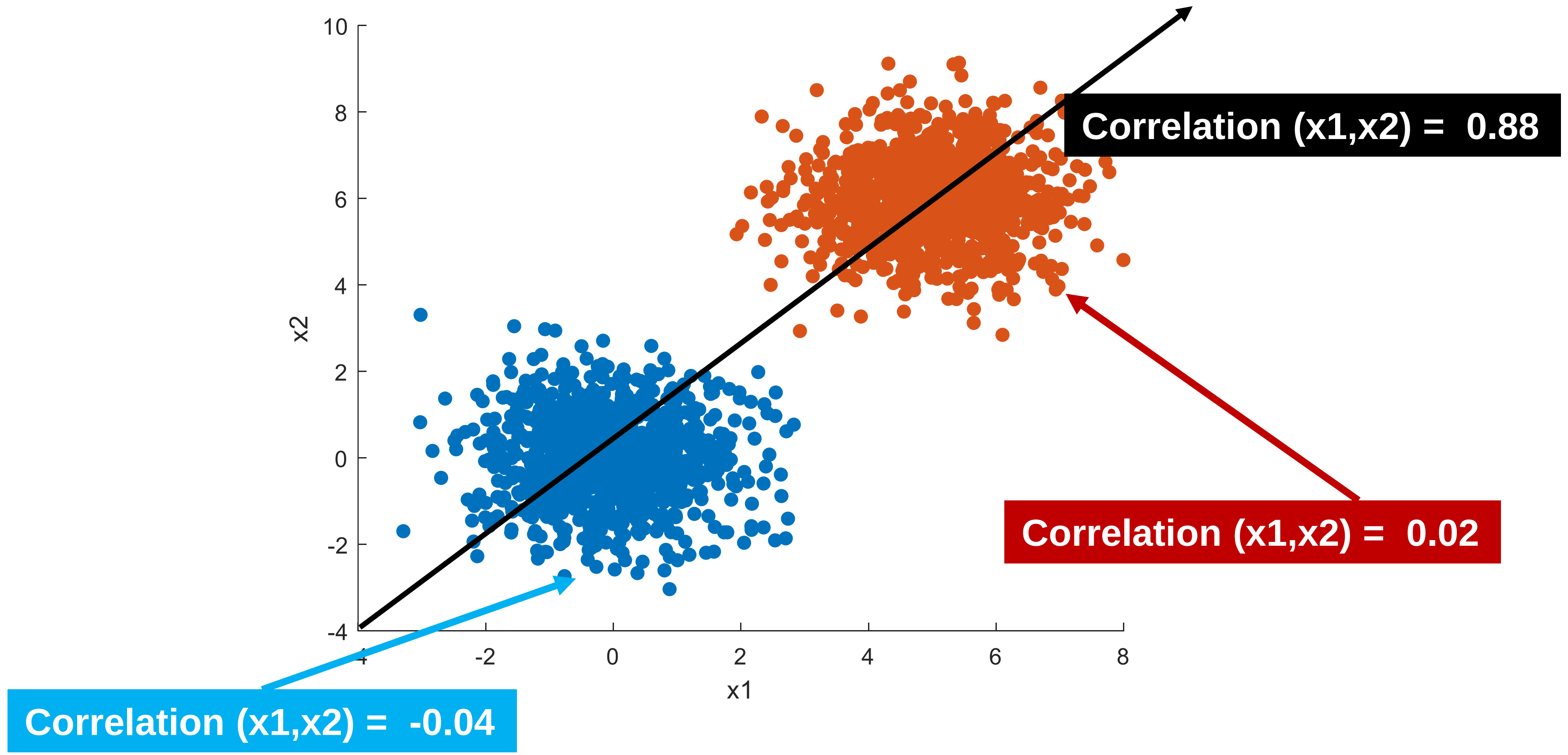
A. Yes

B. No

C. I do no know



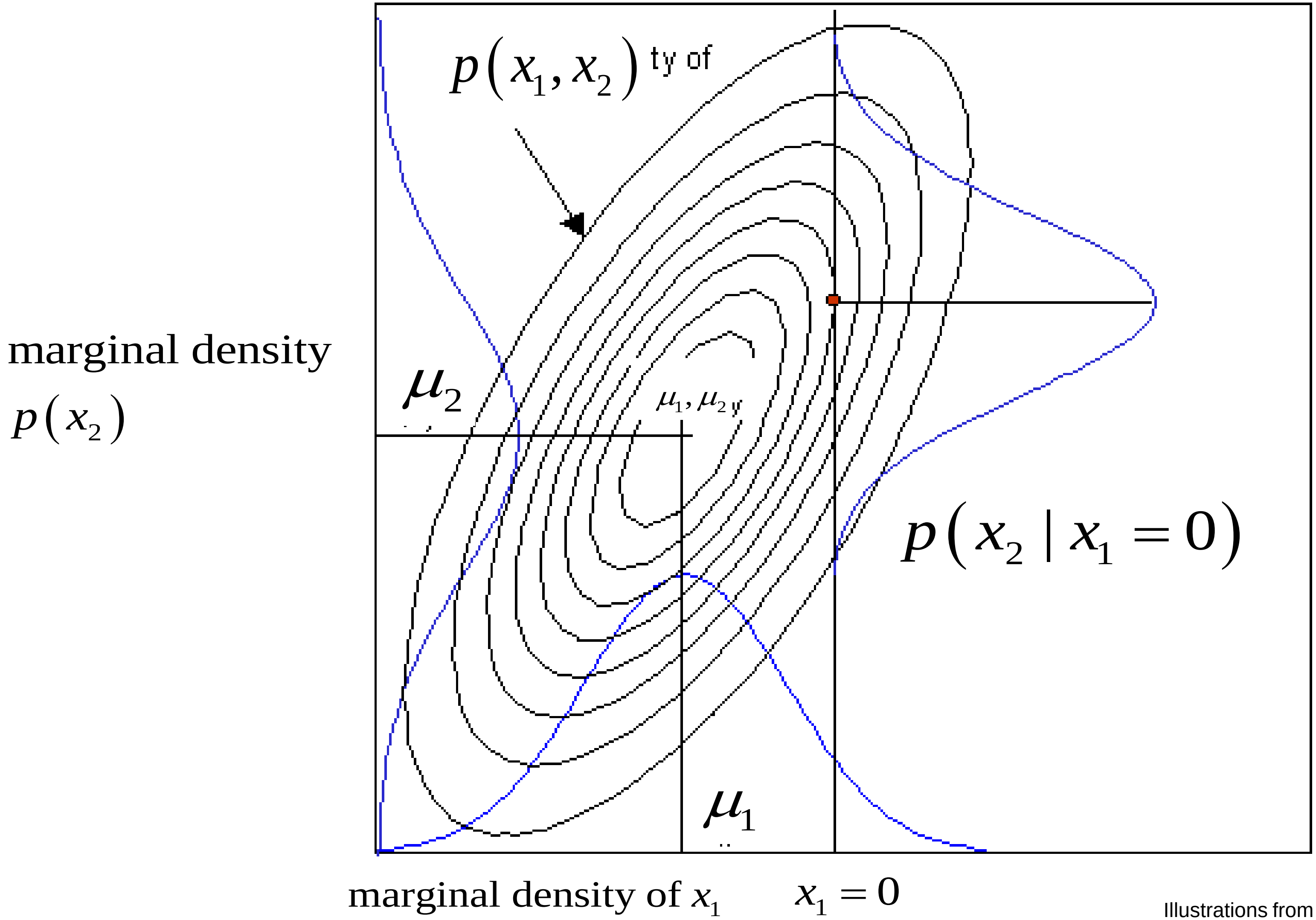
Correlations



Spurious correlations as we compare two groups of data that come from two different distributions

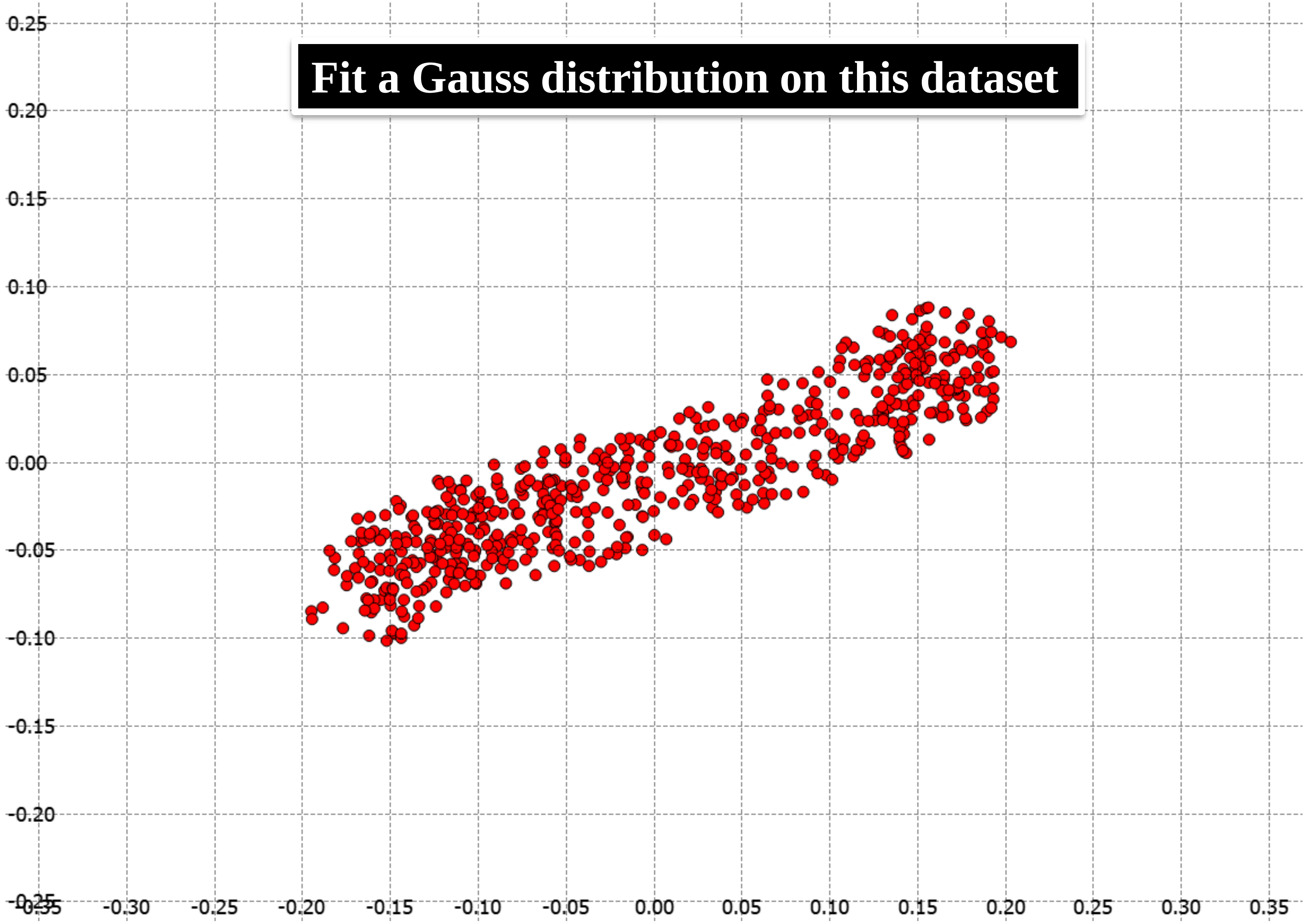
Marginal, Conditional Pdf of Gauss Functions

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

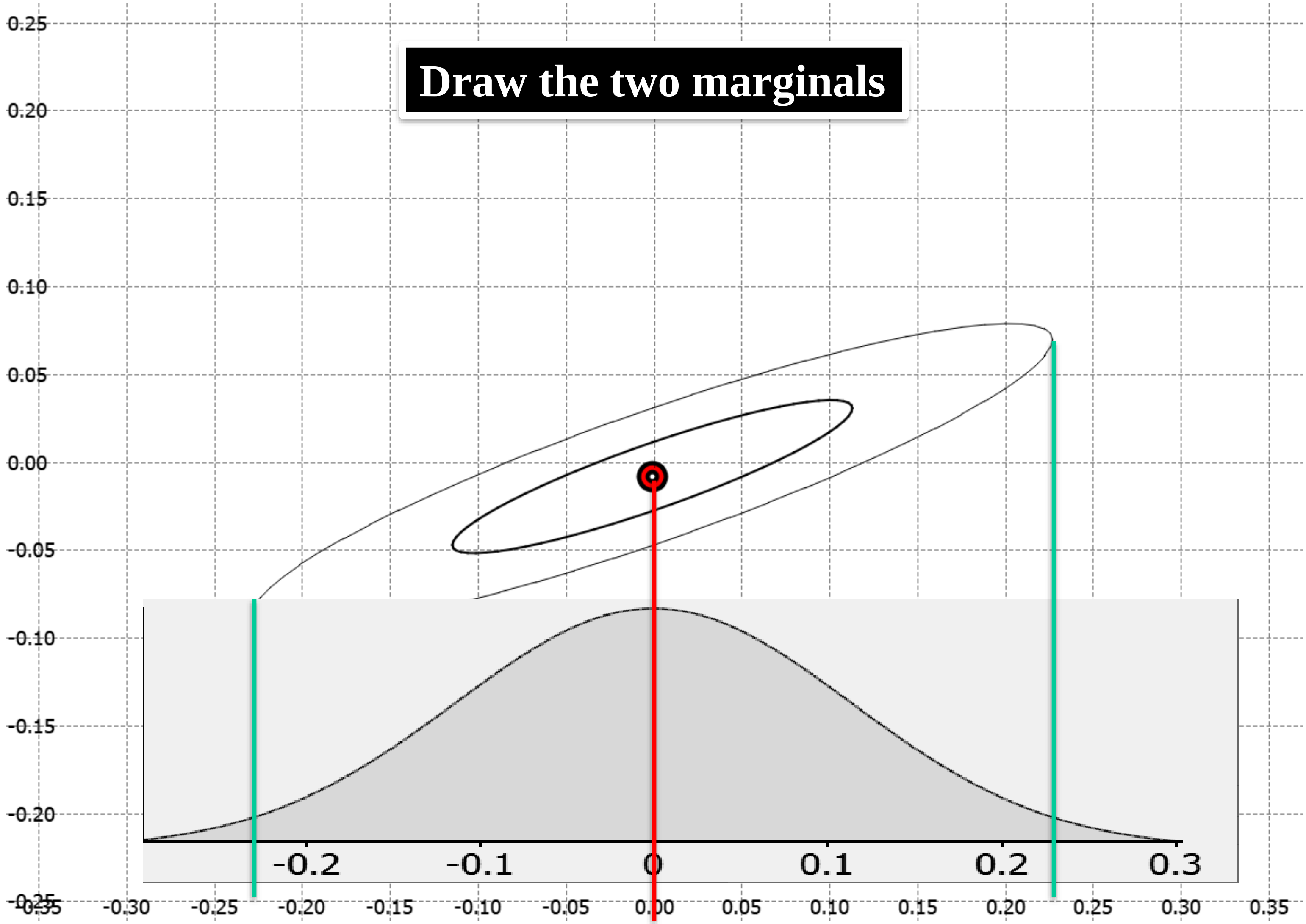


Illustrations from Wikipedia

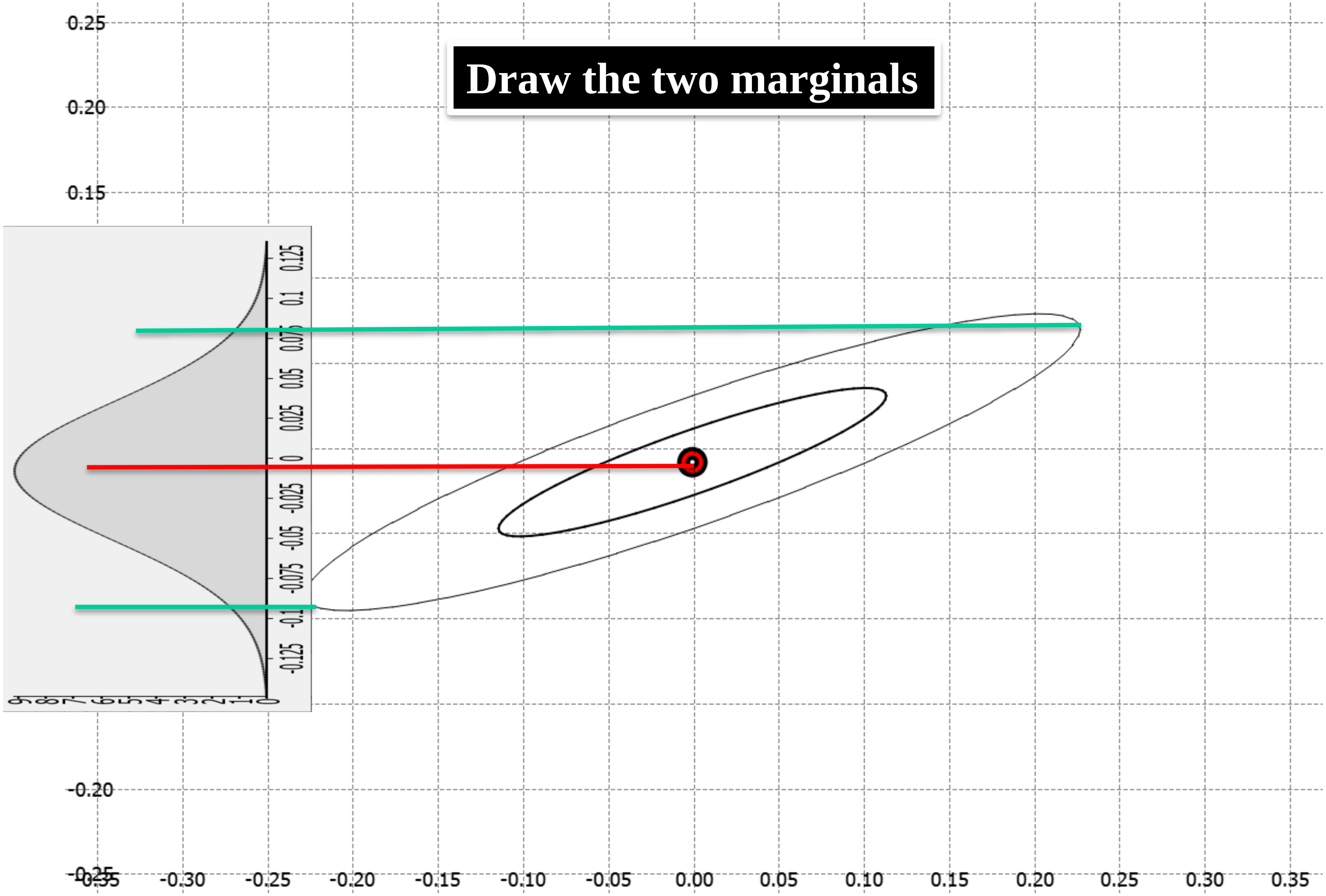
Marginal, Conditional Pdf of Gauss Functions



Marginal, Conditional Pdf of Gauss Functions



Marginal, Conditional Pdf of Gauss Functions



Likelihood

Likelihood for a single unnormalized Gauss function, given a set of M datapoints $\{x^i\}_{i=1}^{M}$

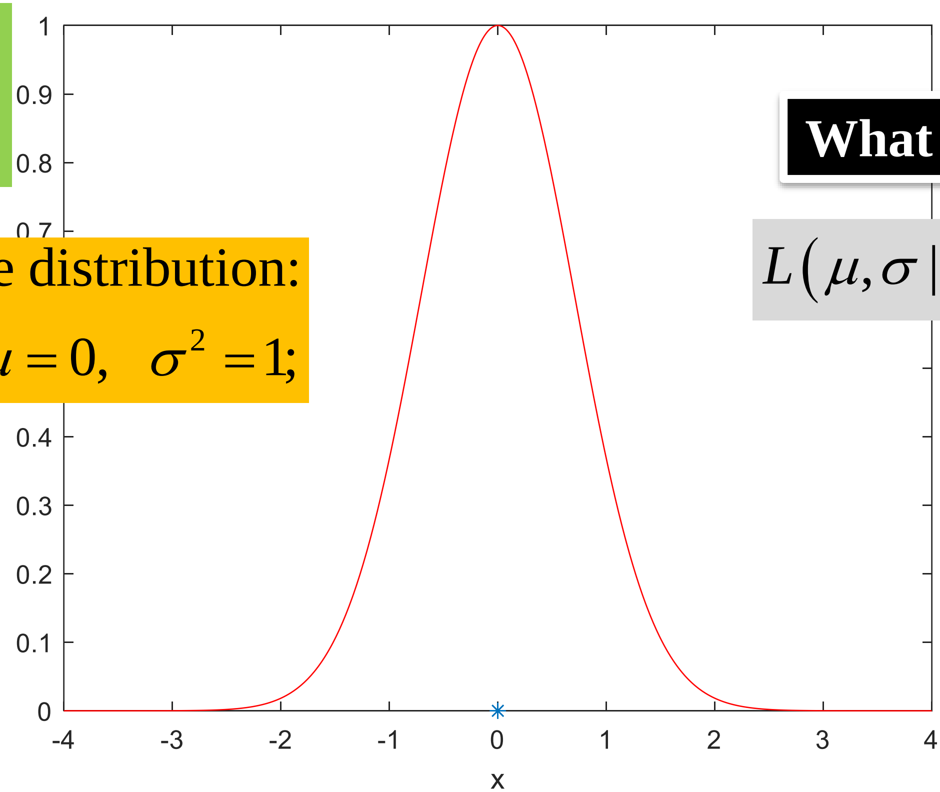
$$L(\mu,\sigma|X) = \prod_{i=1}^{M} p\left(x^i;\mu,\sigma\right) = e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

Dataset:

$X = \left[x^1\right],\ x^1 = 0;$

Sampled from the distribution:

$x \sim p\left(x;\mu,\sigma\right);\ \mu = 0,\ \ \sigma^2 = 1;$



What is the likelihood?

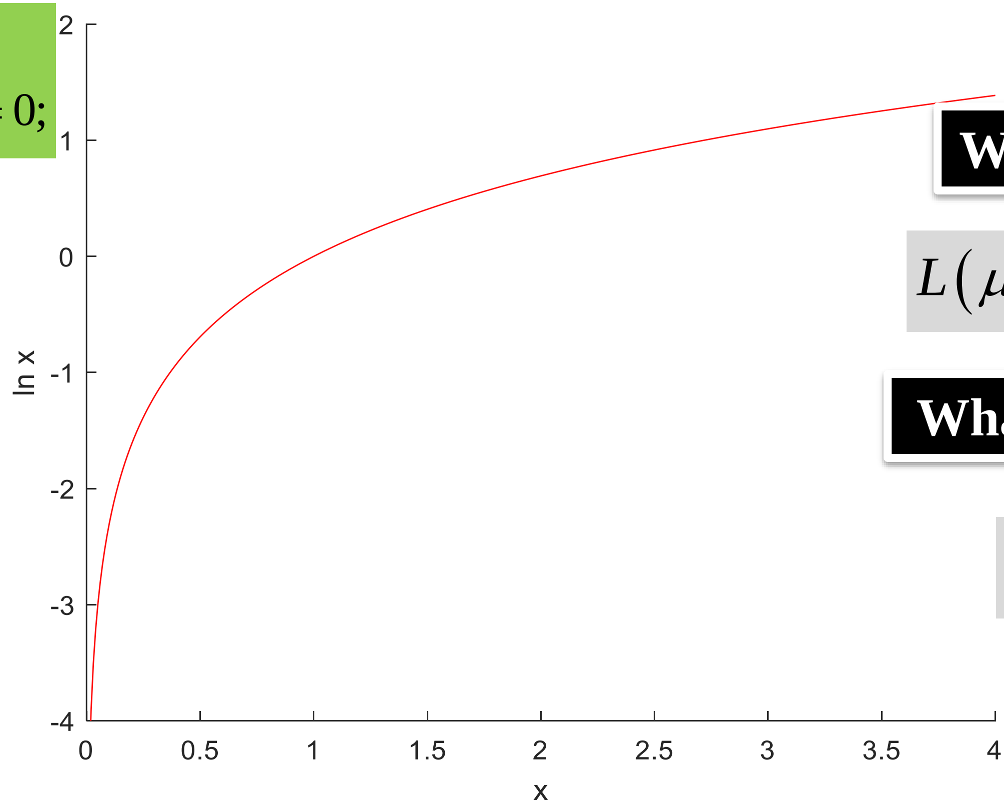
$$L\left(\mu,\sigma \mid X\right) = p\left(x^1;\mu,\sigma\right) = 1$$

Here: we consider an un-normalized Gauss function, also called Radial Basis Functions (RBF).

Likelihood

Dataset:

$X = \left[x^1 \right],\ x^1 = 0;$



What is the likelihood?

$$L\left(\mu,\sigma \mid X\right) = p\left(x^1;\mu,\sigma\right) = 1$$

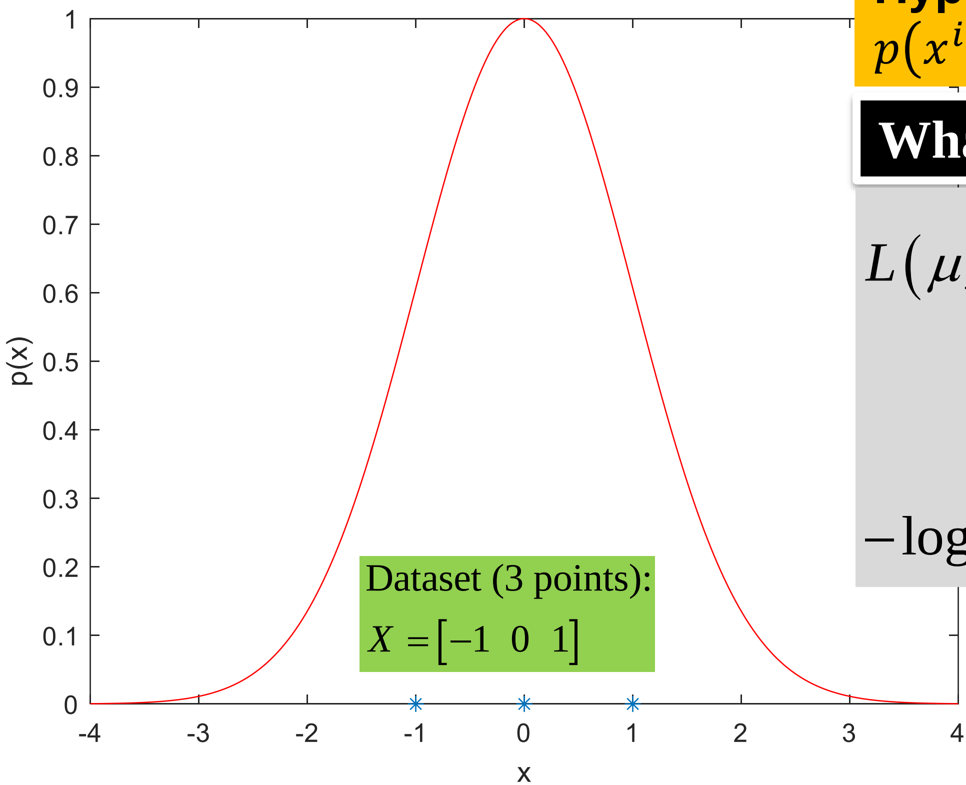
What is the log-likelihood?

$$\log p\left(x^1;\mu,\sigma\right) = 0$$

Likelihood

Likelihood for a single unnormalized Gauss function, given a set of M datapoints $\{x^i\}_{i=1}^{M}$

$$L(\mu,\sigma|X) = \prod_{i=1}^{M} p(x^i;\mu,\sigma) = e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$



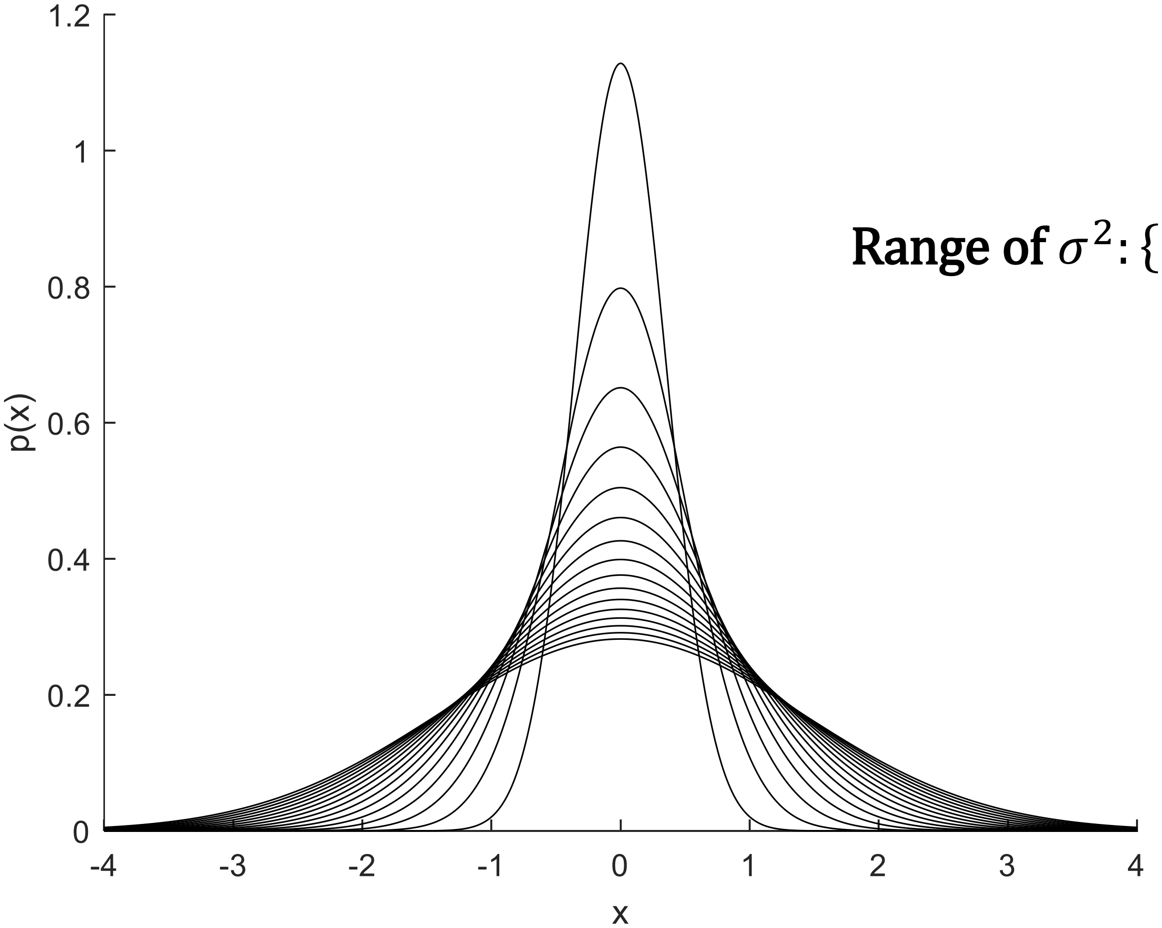
Dataset (3 points):
$X = [-1 \;\; 0 \;\; 1]$

Hypothetical Model:
$p(x^i;\mu,\sigma);\ \mu = 0, \quad \sigma^2 = 2;$

What is the likelihood?

$$L(\mu,\sigma \mid X) = \prod_{i=1}^{3} p(x^i;\mu,\sigma)$$
$$= 0.8*1*0.8$$
$$= 0.6$$
$$-\log L(\mu,\sigma \mid X) = 0.5$$

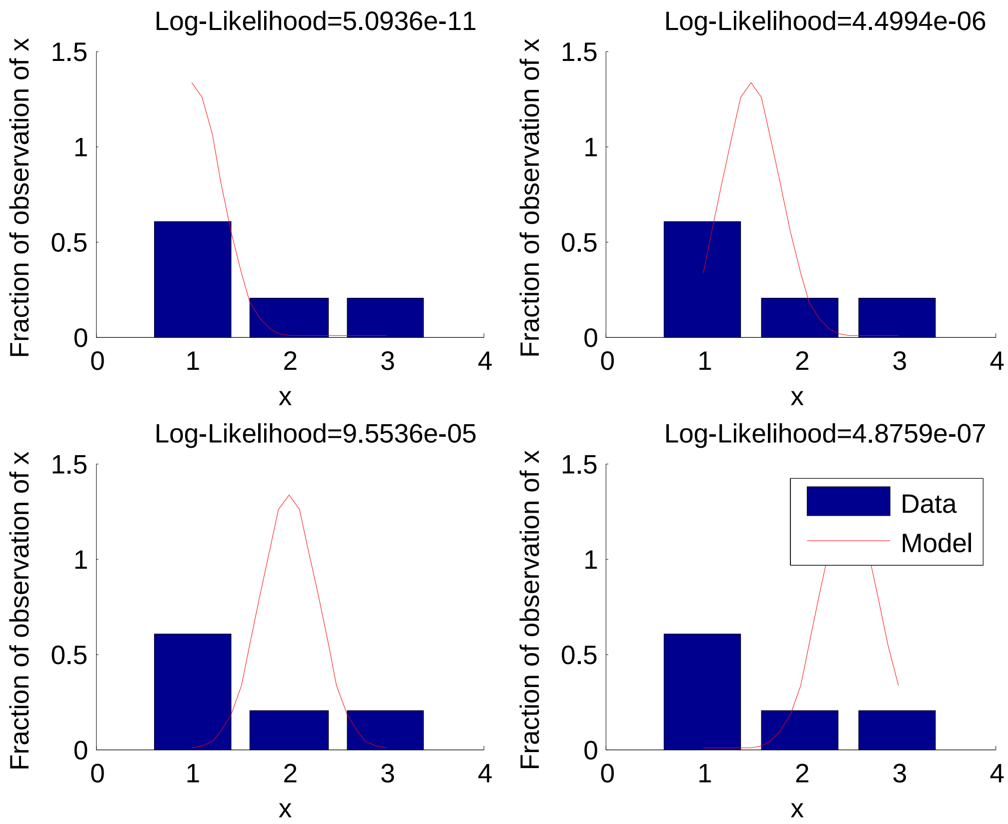
Gauss pdfs



Range of σ^2: $\{0.25, \ldots\ldots, 2\}$

Pdf for the normalized distributions

Likelihood Function for data not gauss distributed

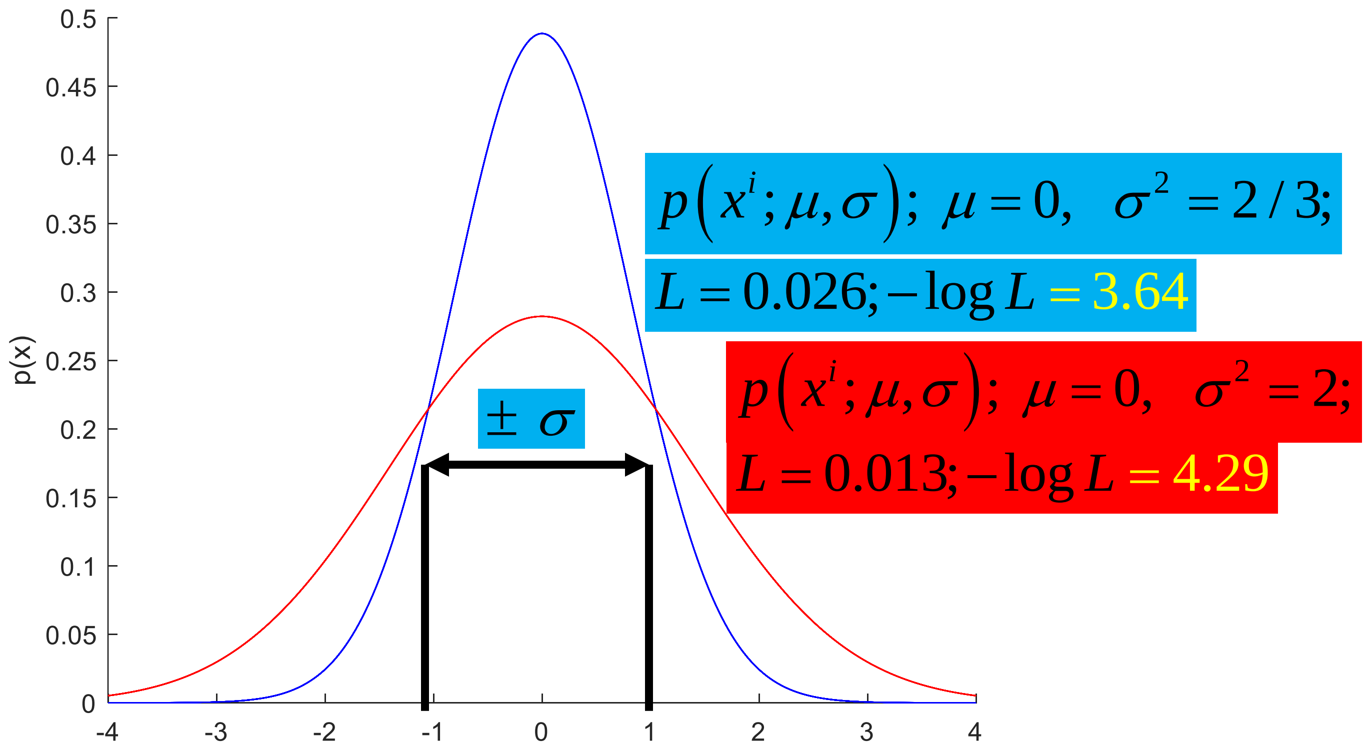


Log-Likelihood for a series of Gauss functions applied to datasets with pdfs that do not follow a Gauss distribution. The Likelihood increases as the fit is closer to the real mean of the data, even if this may appear as a poorer fit.

Maximum Likelihood

Likelihood for a single Gauss function, given a set of M datapoints $\left\{x^i\right\}_{i=1}^{M}$

$$L(\mu,\sigma \mid X)=\prod_{i=1}^{M} p\left(x^i;\mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$



Pdf for the normalized distributions

Maximum Likelihood

Likelihood for a single Gauss function, given a set of M datapoints $\left\{x^i\right\}_{i=1}^{M}$

$$L(\mu,\sigma \mid X) = \prod_{i=1}^{M} p\left(x^i;\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$

The blue solution is the optimal value on the likelihood we can ever get for such a distribution of points.

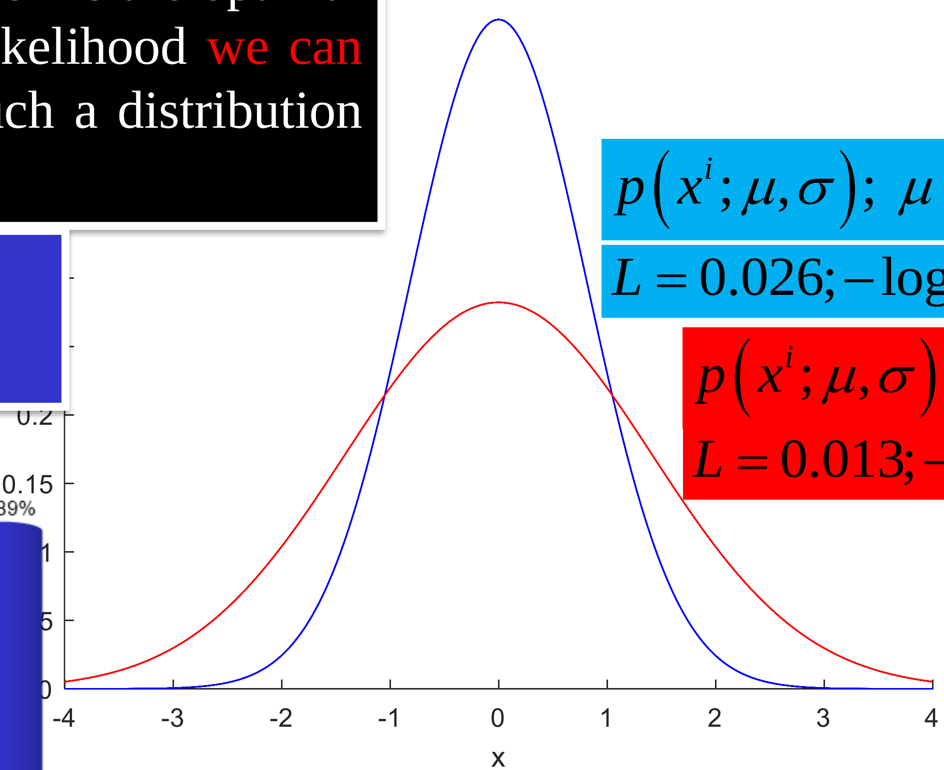
A. True
B. False

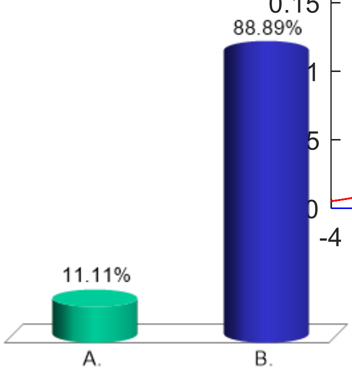
$p\left(x^i;\mu,\sigma\right);\ \mu=0,\ \ \sigma^2=2/3;$

$L=0.026; -\log L = 3.64$

$p\left(x^i;\mu,\sigma\right);\ \mu=0,\ \ \sigma^2=2;$

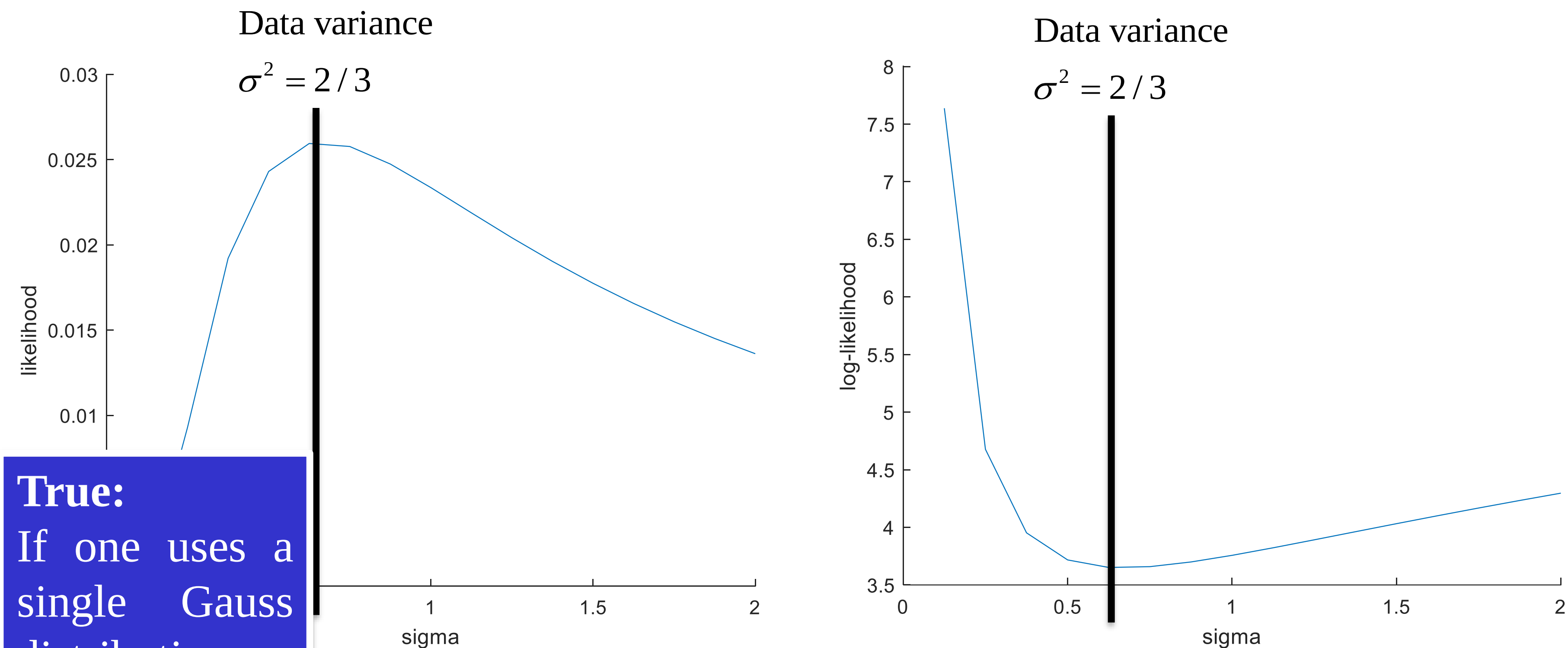
$L=0.013; -\log L = 4.29$





Maximum Likelihood

The maximum of the likelihood, minimum of -log-likelihood, is obtained for a distribution with same variance as that of the data.



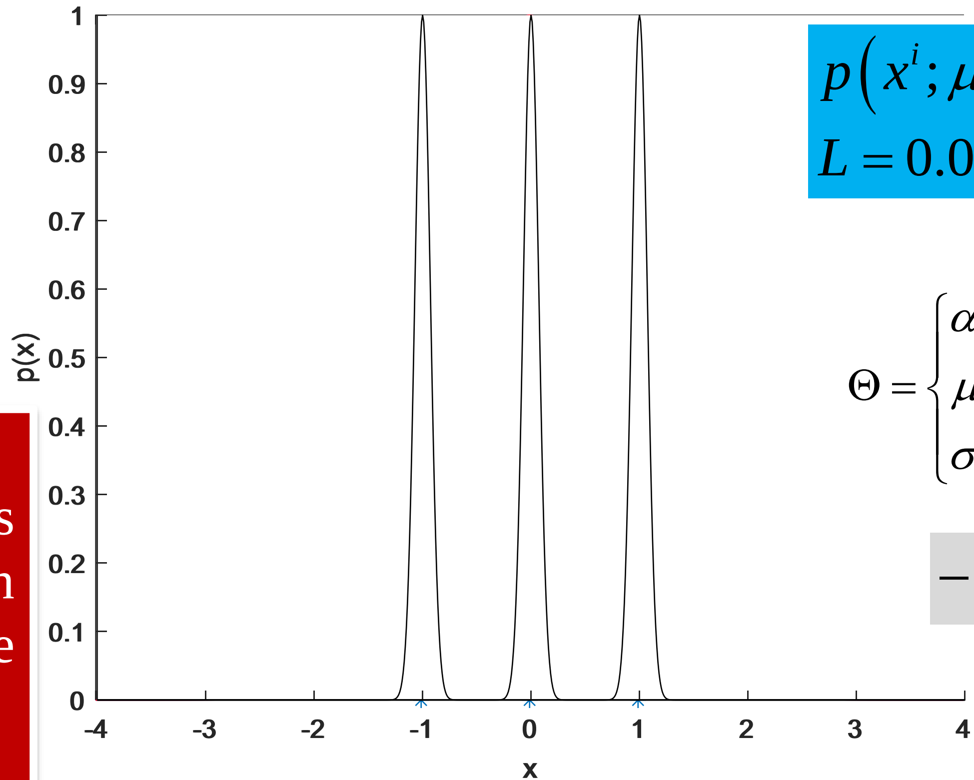
True:
If one uses a single Gauss distribution as model fit.

Real values for the normalized distributions

Maximum Likelihood

Likelihood for a single Gauss function, given a set of M datapoints $\left\{x^i\right\}_{i=1}^{M}$

$$L\left(\Theta \mid X\right) = \prod_{i=1}^{M} \sum_{k=1}^{3} \alpha_k \, p\left(x^i; \mu_k, \sigma_k\right)$$



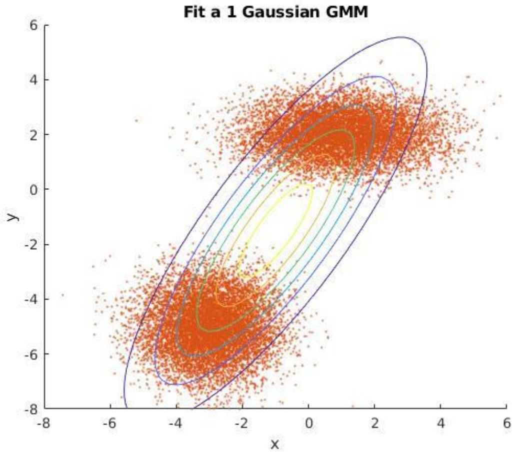
$$p\left(x^i; \mu, \sigma\right); \;\; \mu = 0, \quad \sigma^2 = 2/3;$$
$$L = 0.026; -\log L = 3.64$$

$$\Theta = \begin{Bmatrix} \alpha_1 = 1/3, \alpha_2 = 1/3, \alpha_3 = 1/3, \\ \mu_1 = -1, \mu_3 = 0, \mu_3 = 1, \\ \sigma_1 = 0.01, \sigma_2 = 0.01, \sigma_3 = 0.01 \end{Bmatrix}$$

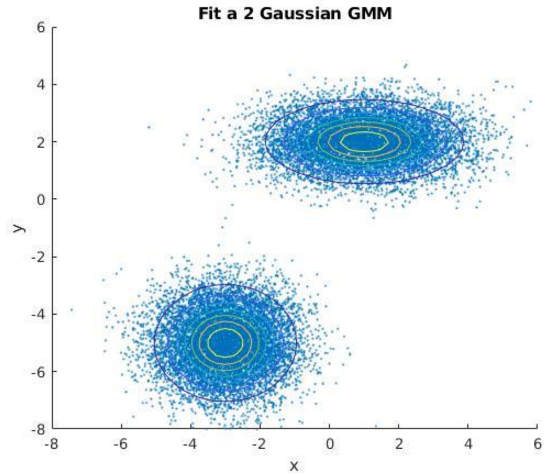
$$-\log L\left(\Theta \mid X\right) = 3.29$$

False:
A better fit is obtained when using more complex distributions.

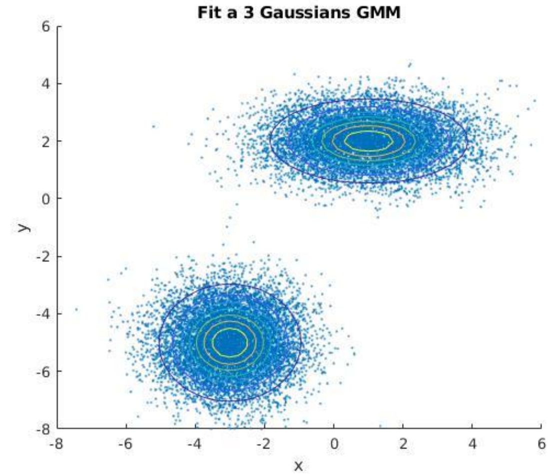
Maximum likelihood with Mixture of Gaussians



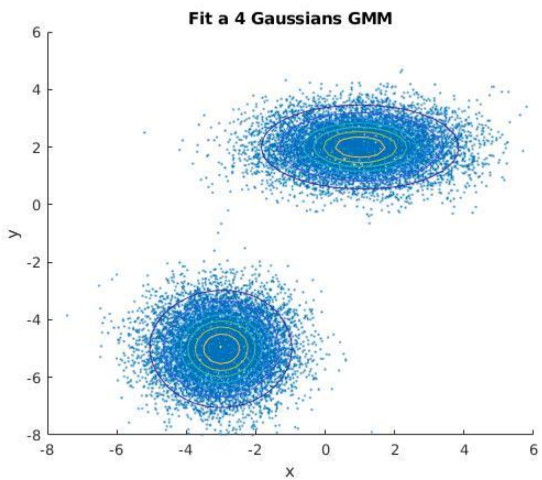
Log-Likelihood = -87825



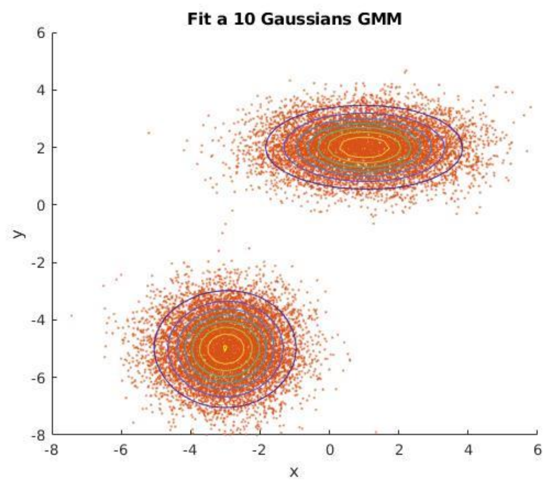
Log-Likelihood = - 70610



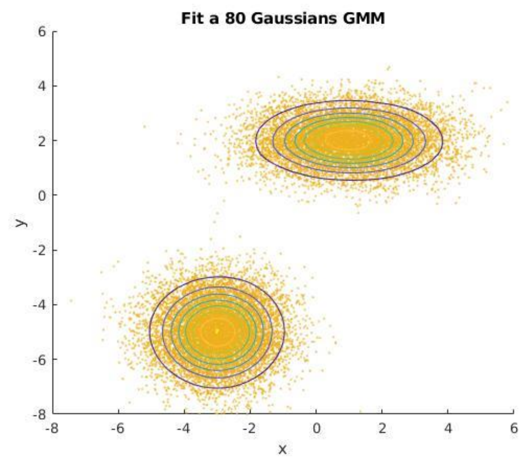
Log-Likelihood = -70610



Log-Likelihood = -70604



Log-Likelihood = -70601



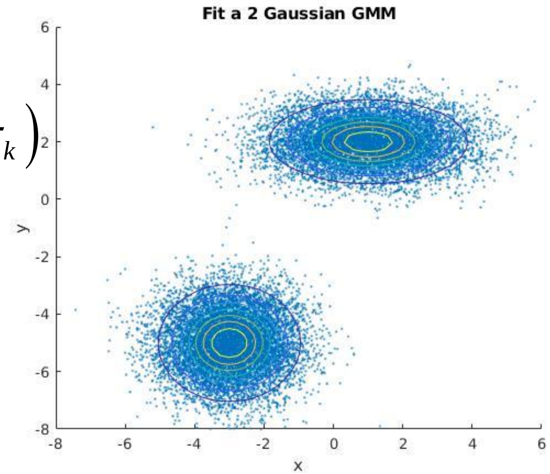
Log-Likelihood = -70581

Maximum likelihood with Mixture of Gaussians

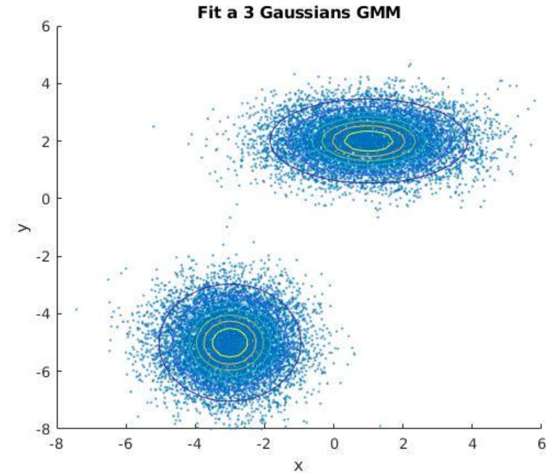
Likelihood identical:

$$L\left(\Theta \mid X\right) = \prod_{i=1}^{M} \sum_{k=1}^{K} \alpha_k \, p\left(x^i ; \mu_k, \sigma_k\right)$$

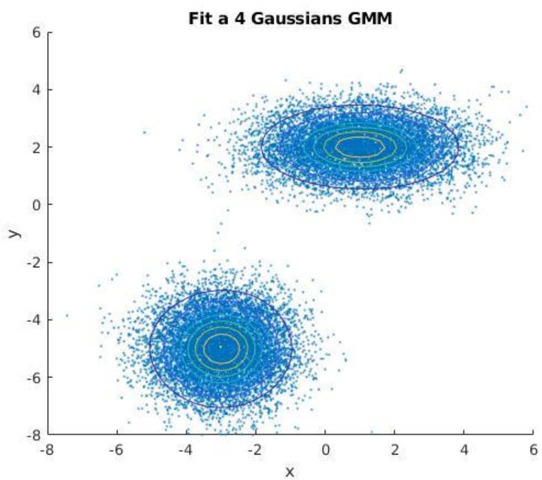
$$\sum_{k=1}^{K} \alpha_k = 1$$



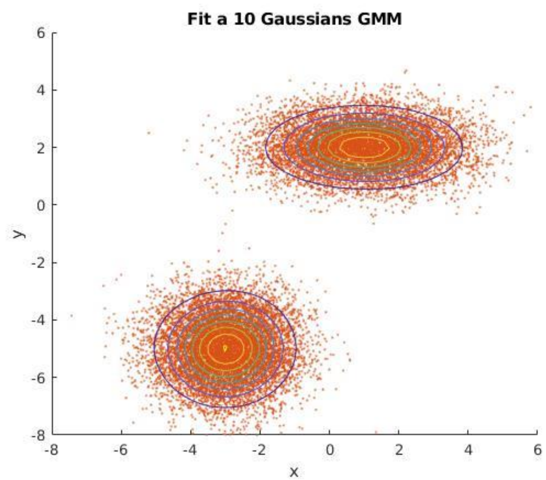
Log-Likelihood = - 70610



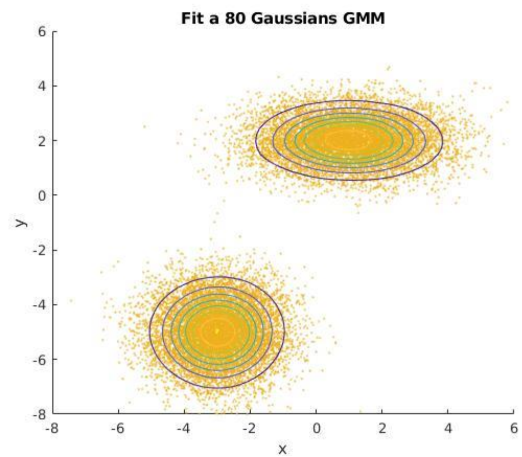
Log-Likelihood = -70610



Log-Likelihood = -70604

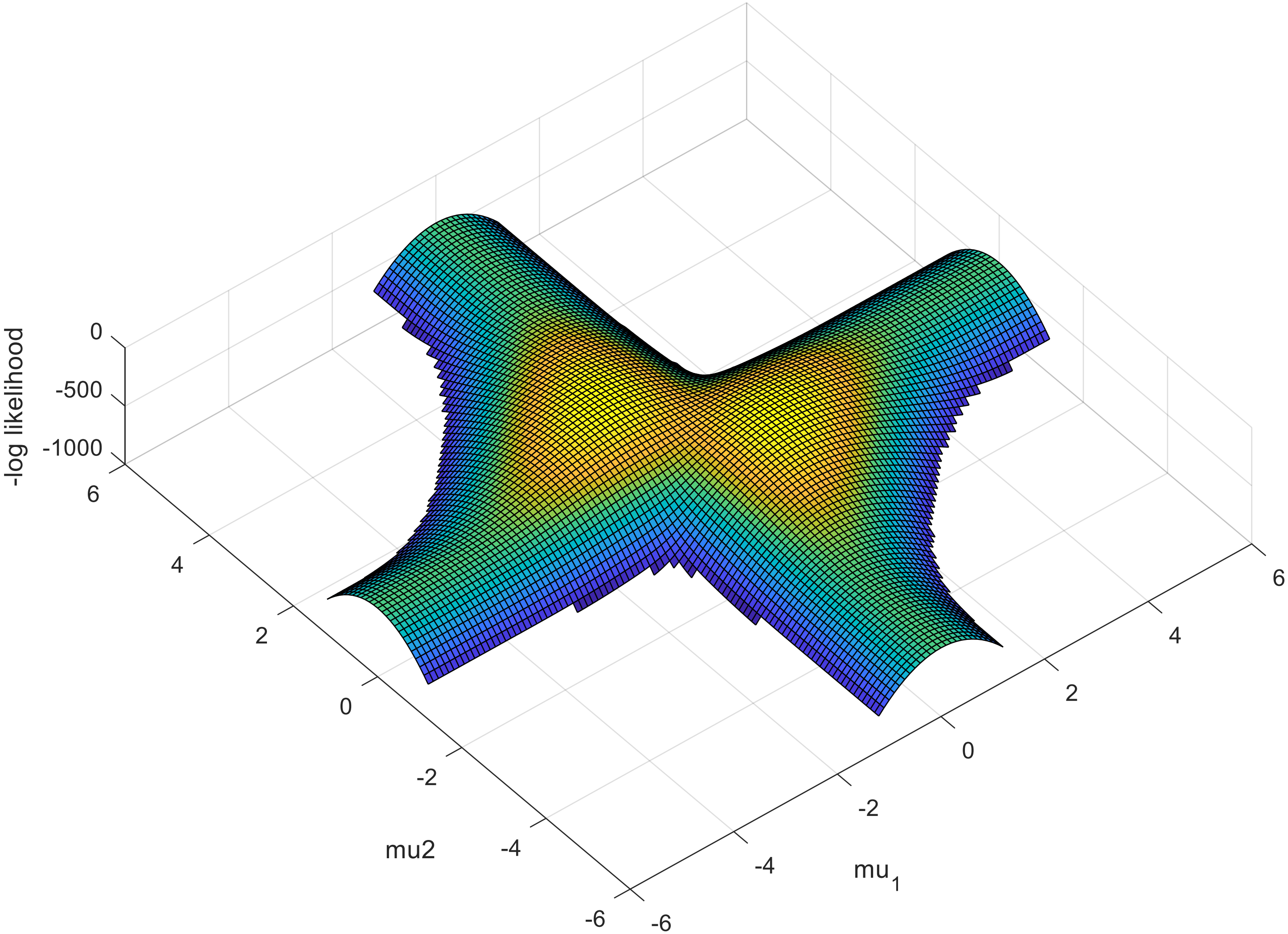


Log-Likelihood = -70601



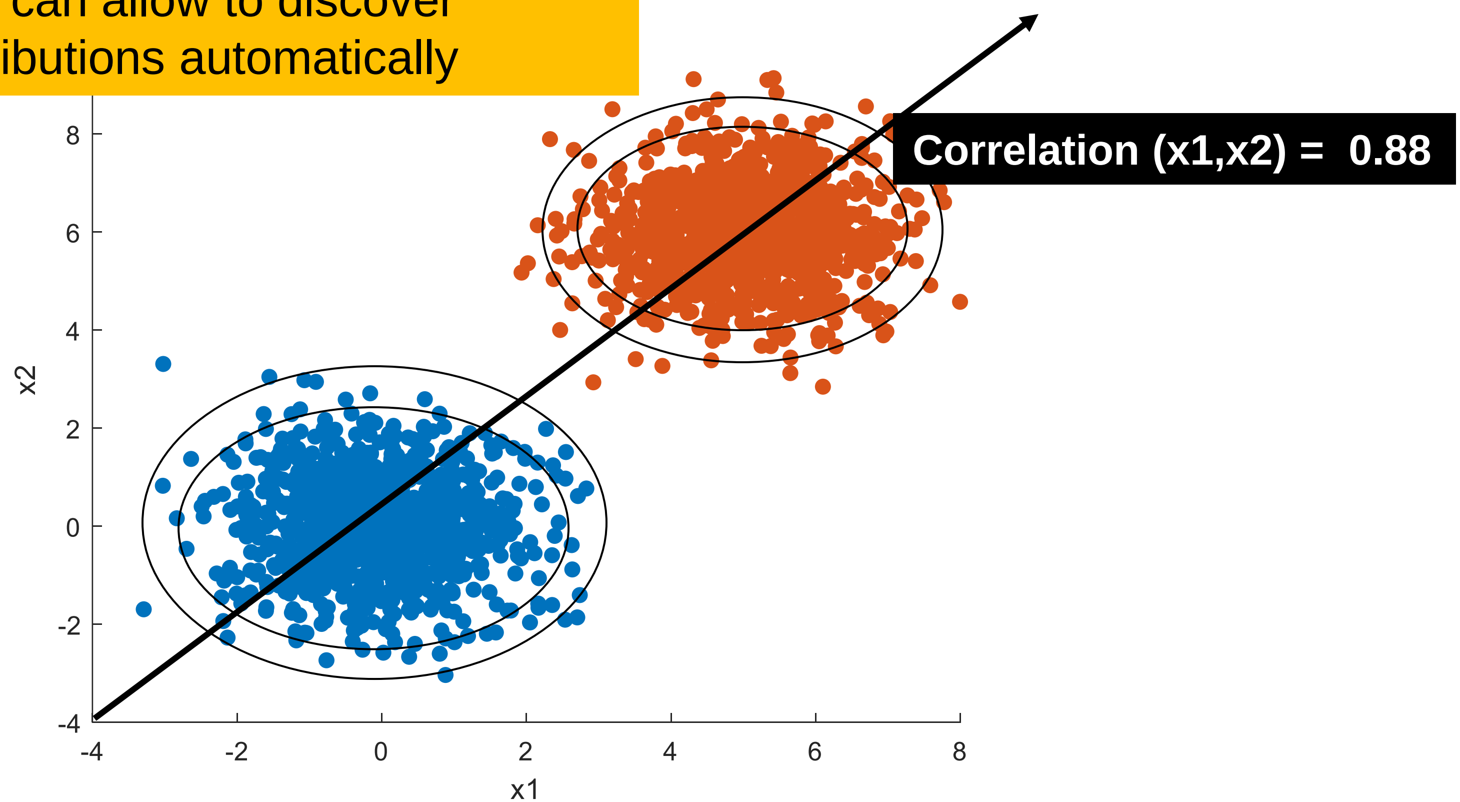
Log-Likelihood = -70581

Non-convexity of the likelihood



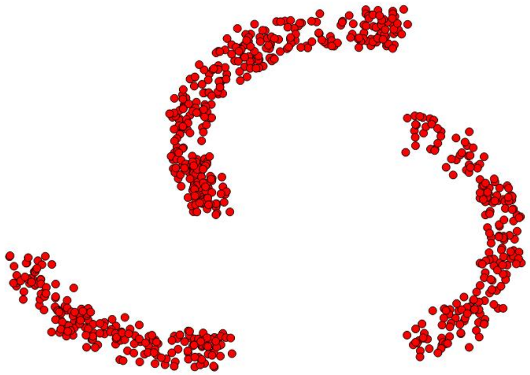
Correlations

Clustering can allow to discover these distributions automatically

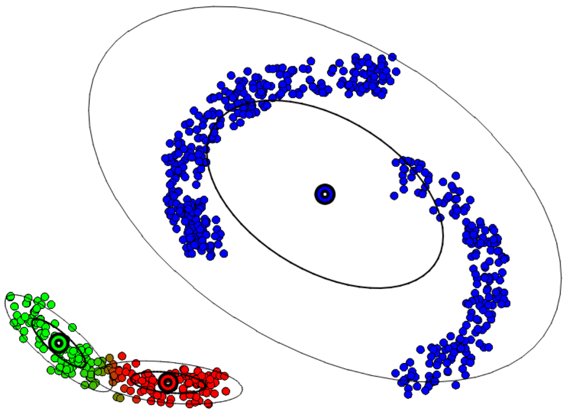


Spurious correlations as we compare two groups of data that come from two different distributions

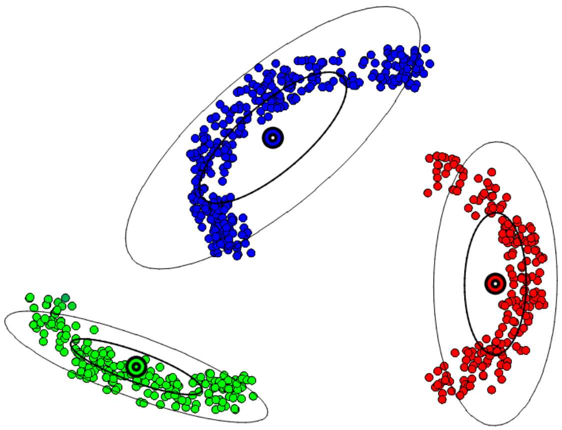
Clustering with Gaussian Mixture Models



Original data

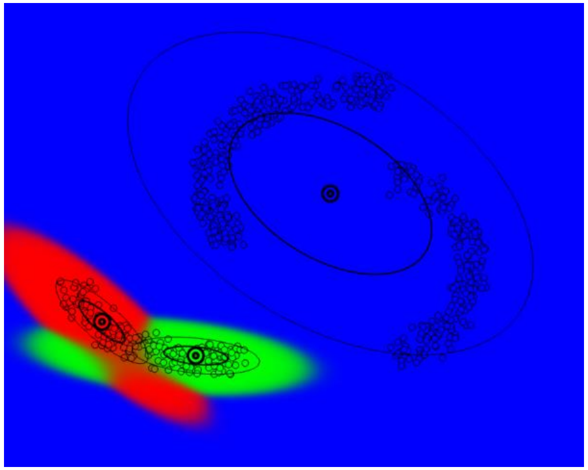


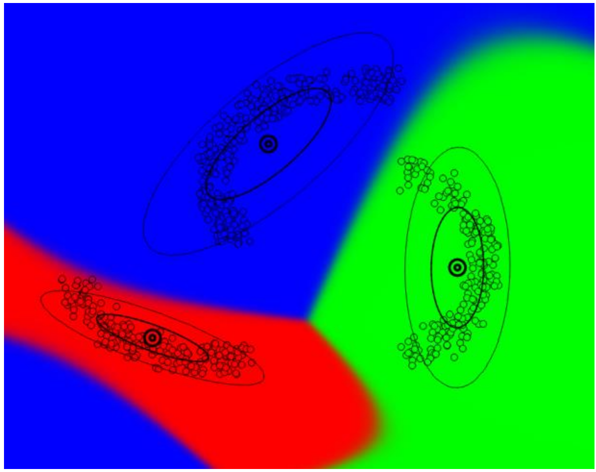
Solution 1



Solution 2

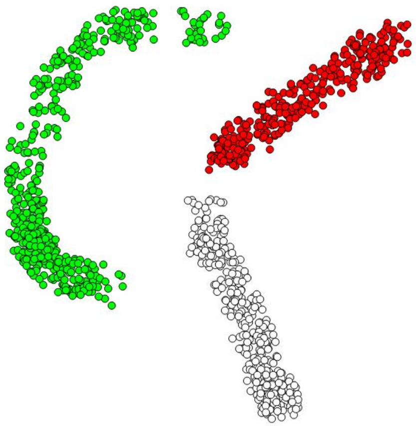
Compute boundary across clusters by comparing likelihood of each cluster, i.e. of each Gauss function.



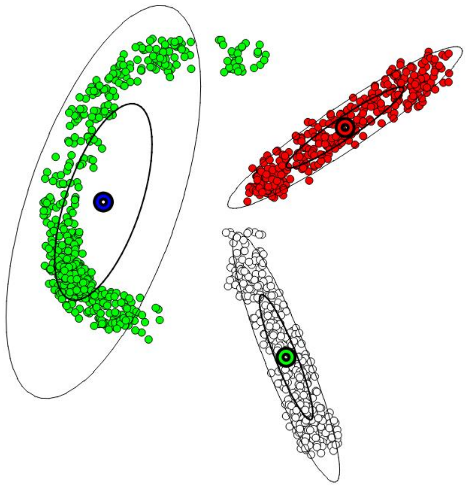


Boundaries

Which Model?



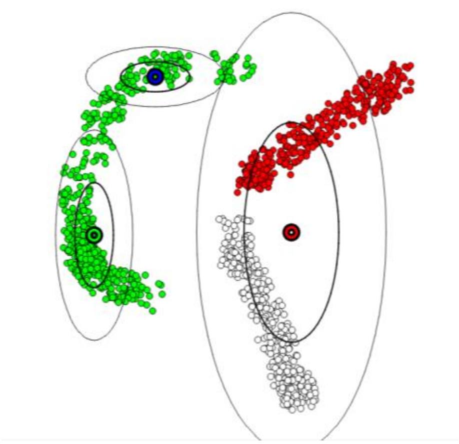
Original data



Solution with full matrices

Full matrices:

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



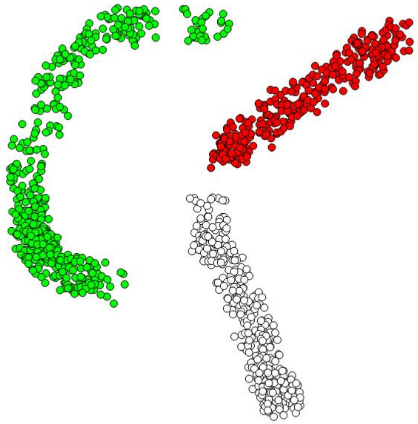
Diagonal matrices:

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

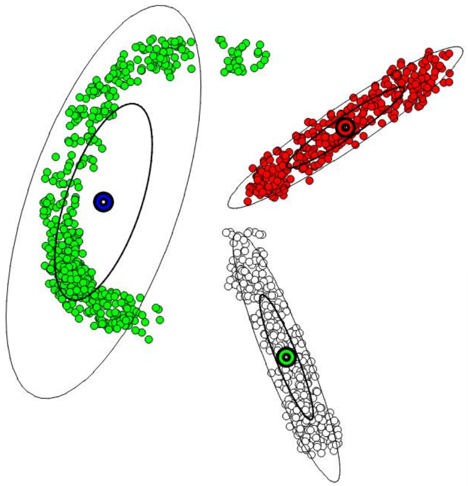
Solution with diagonal matrices

Metrics to choose model

1180 sample points



Original data



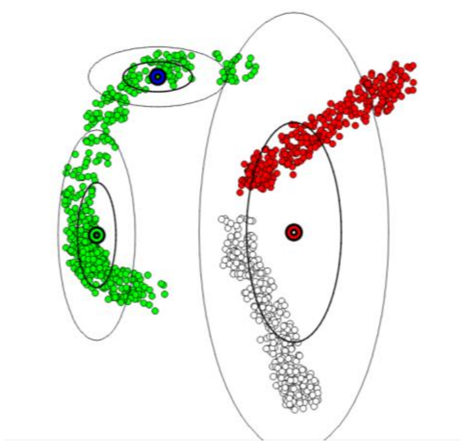
Solution with full matrices

Parameters: 3*(2+3)=15

Likelihood: 2719,
AIC: - 5419
BIC: -5373

Even if it requires more parameters, the gain on likelihood is important
→ Optimal solution

Which of the two solutions would get the best values on AIC or BIC?



Solution with diagonal matrices

Parameters: 3*(2+2)=12

Likelihood: 1999,
AIC: - 2912
BIC: -2972

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

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

B=# Parameters