Clustering: Models and Algorithms

Shikui Tu

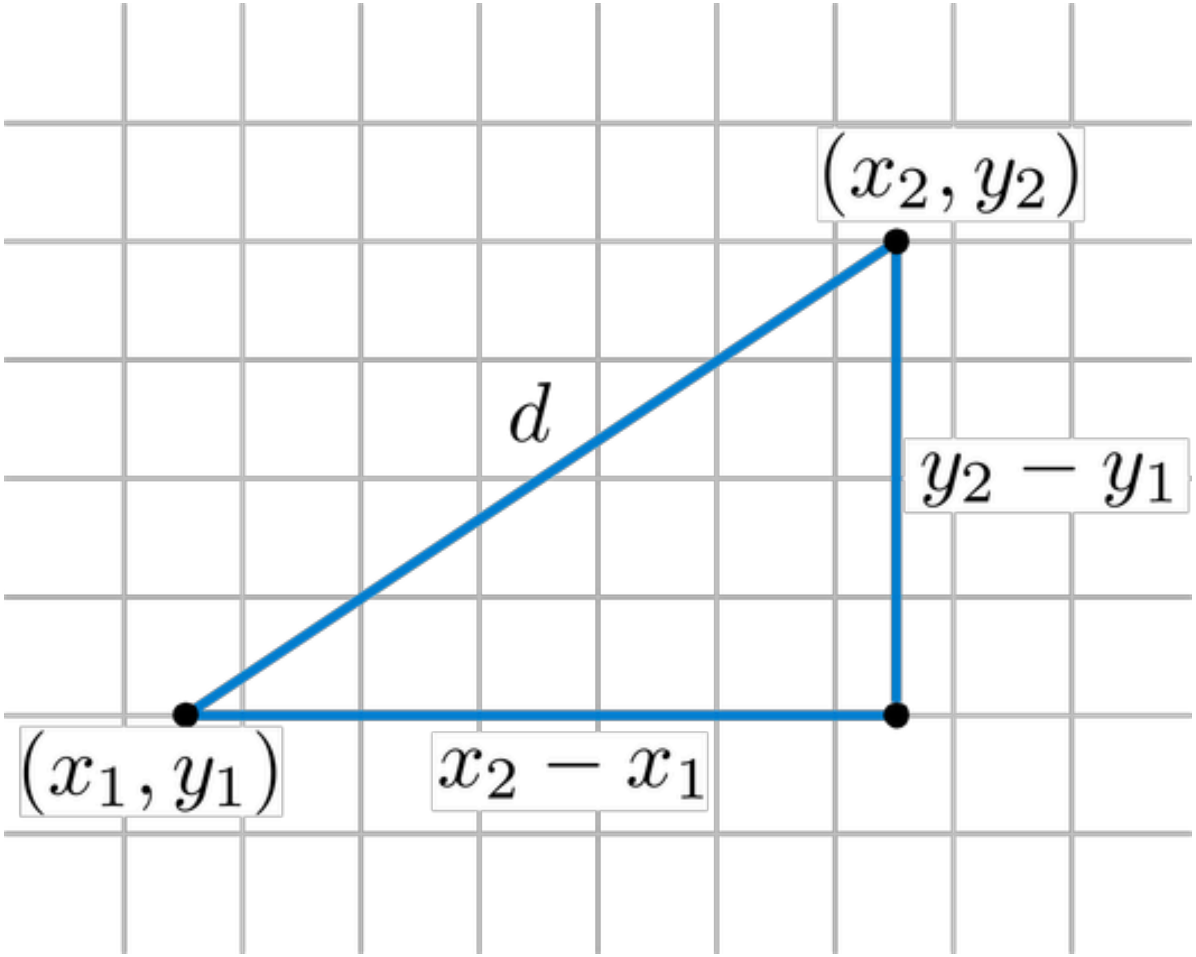
Shanghai Jiao Tong University

2021-03-16

Outline

- **Gaussian Mixture Models (GMM)**
- Expectation-Maximization (EM) for maximum likelihood
- Gaussian Mixture Models (GMM)
  - From generation process perspective

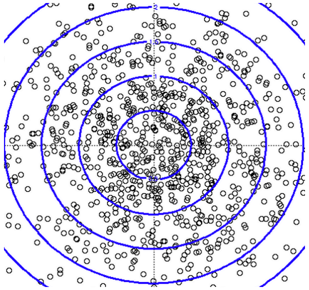
Euclidean Distance



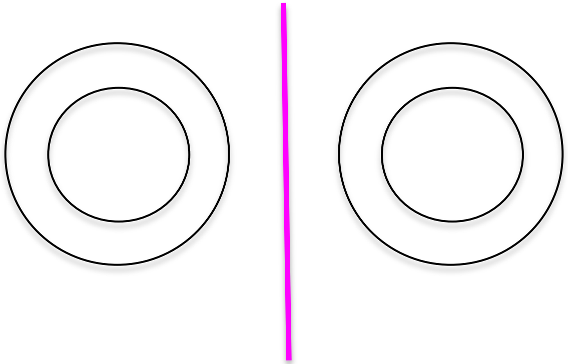
Glossary. (n.d.). Retrieved February 13, 2015, from http://rosalind.info/glossary/euclidean-distance/

Euclidian distance may not be a good measure for some data

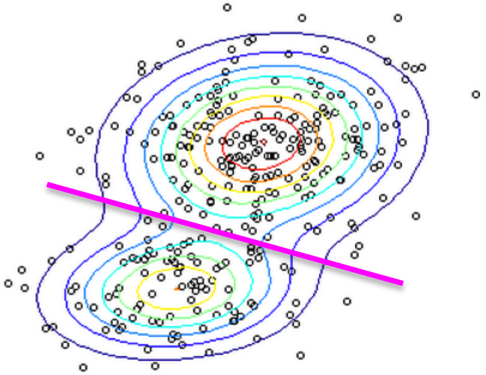
Euclidean distance



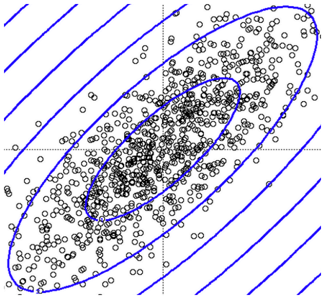
Equal distance line

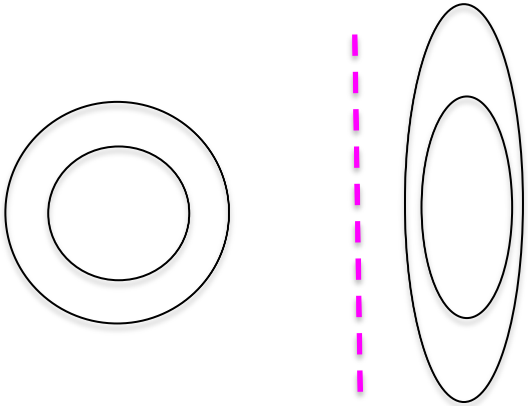


In general



Mahalanobis distance





Distances at different directions could be different!

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T\Sigma^{-1}(\vec{x}-\vec{y})}.$$

Σ is the covariance matrix

More Distance Measures

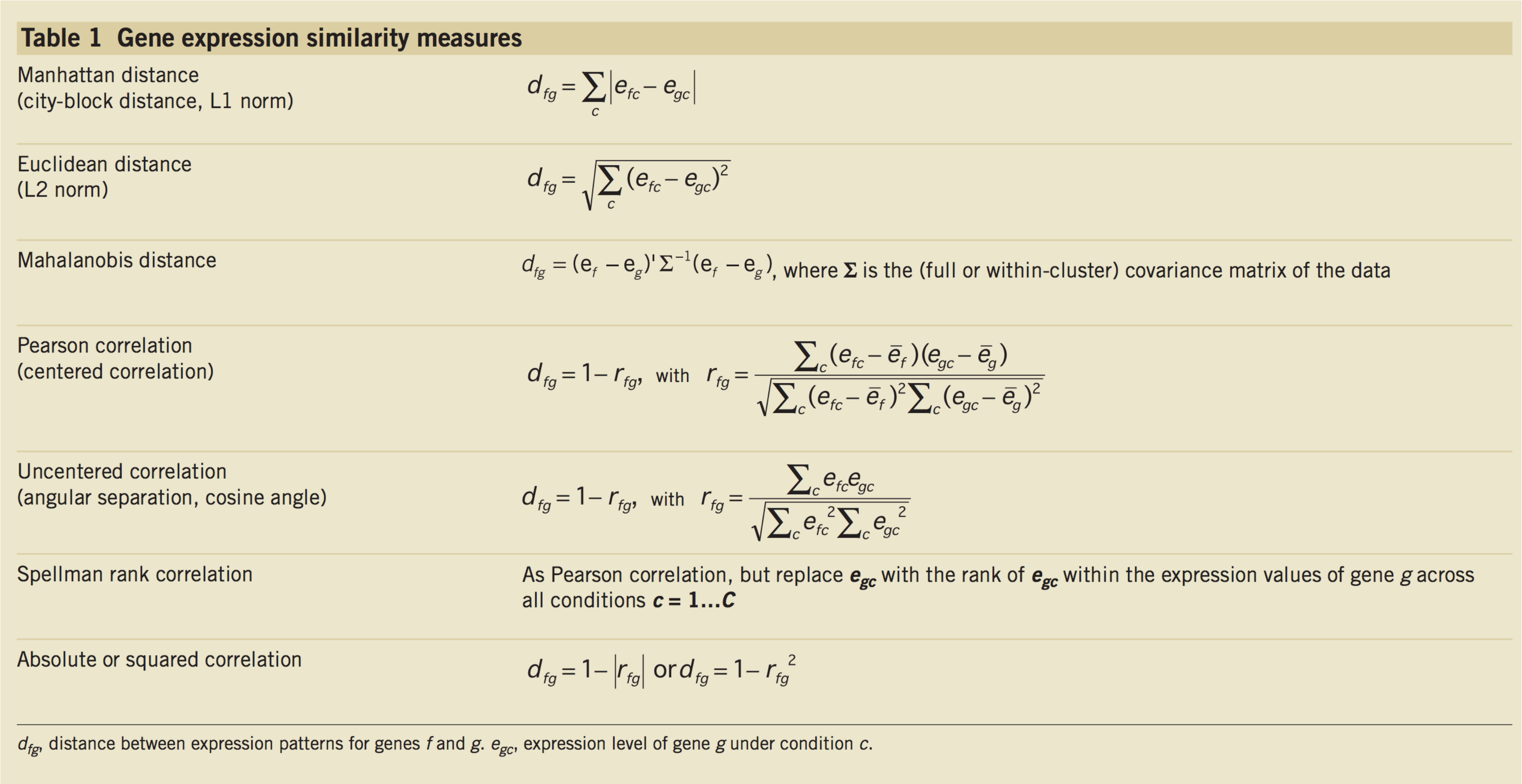


Table 1 Gene expression similarity measures

Manhattan distance (city-block distance, L1 norm)	$d_{fg} = \sum_c \lvert e_{fc} - e_{gc} \rvert$
Euclidean distance (L2 norm)	$d_{fg} = \sqrt{\sum_c (e_{fc} - e_{gc})^2}$
Mahalanobis distance	$d_{fg} = (e_f - e_g)' \Sigma^{-1} (e_f - e_g)$, where $\mathbf{\Sigma}$ is the (full or within-cluster) covariance matrix of the data
Pearson correlation (centered correlation)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$
Uncentered correlation (angular separation, cosine angle)	$d_{fg} = 1 - r_{fg}$, with $r_{fg} = \frac{\sum_c e_{fc} e_{gc}}{\sqrt{\sum_c e_{fc}^2 \sum_c e_{gc}^2}}$
Spellman rank correlation	As Pearson correlation, but replace $\boldsymbol{e_{gc}}$ with the rank of $\boldsymbol{e_{gc}}$ within the expression values of gene *g* across all conditions $\boldsymbol{c = 1 \ldots C}$
Absolute or squared correlation	$d_{fg} = 1 - \lvert r_{fg} \rvert$ or $d_{fg} = 1 - r_{fg}^2$

d_{fg}, distance between expression patterns for genes *f* and *g*. e_{gc}, expression level of gene *g* under condition *c*.

D'haeseleer, P. (2005). How does gene expression clustering work? Nat Biotech *23*, 1499–1501.

From distance to probability

distance

likely

$$\| x - \mu \|^2 \longrightarrow \exp\{-\lambda \| x - \mu \|^2\}$$

"The closer, the more likely."

$$\Big\downarrow$$

Sum or integral to be one

Probability

It is more powerful to consider everything in probability framework!

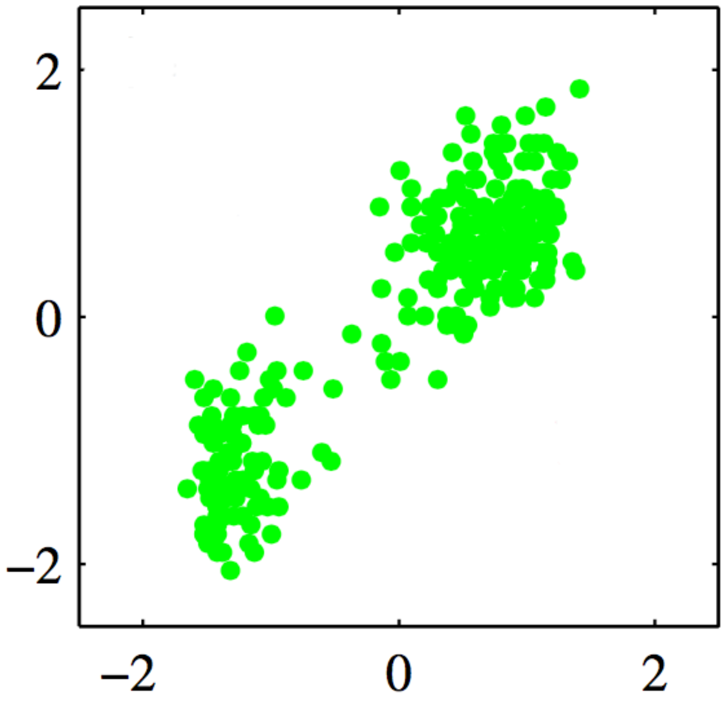
$$\frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Gaussian distribution with the Mahalanobis distance

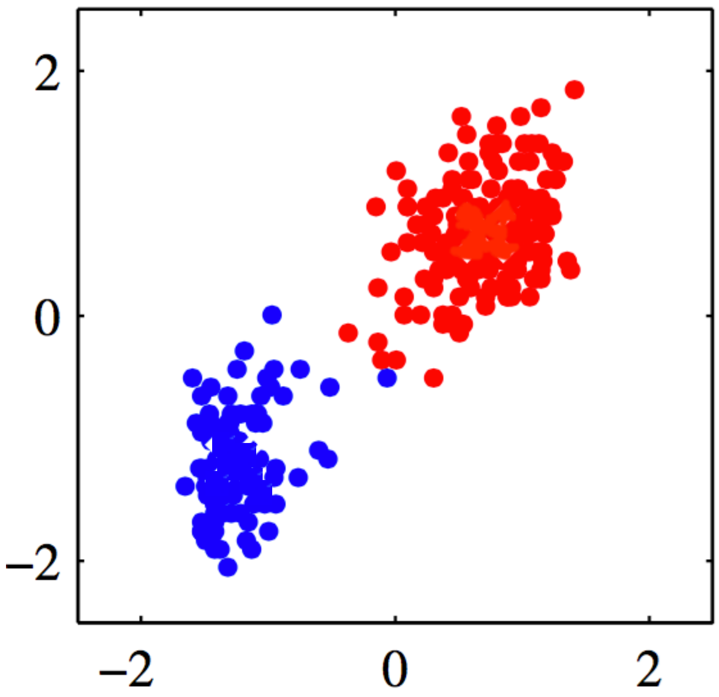
$$D_M(x) = \sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

Review the clustering problem again

We have the following data:

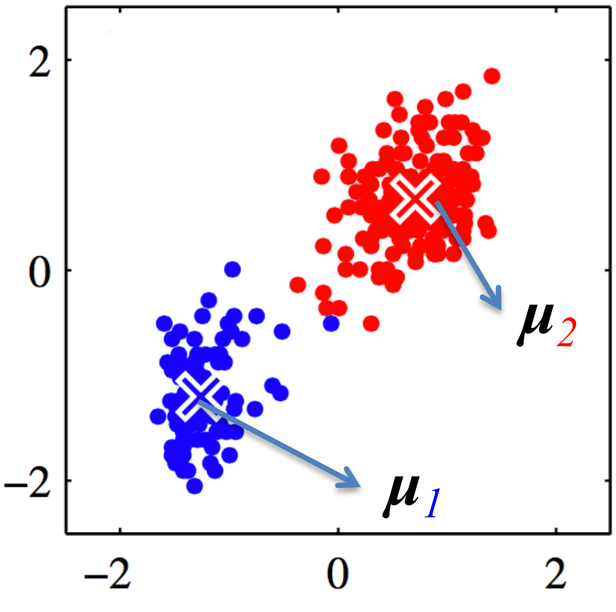


We want to cluster the data into two clusters (red and blue)

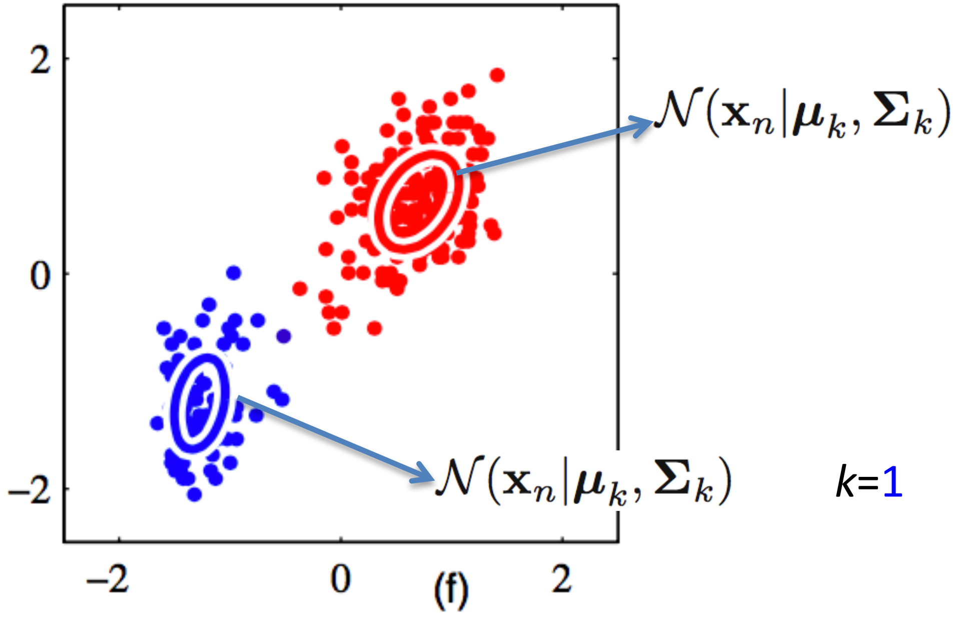


Instead if using {$\boldsymbol{\mu}_{1,}$ $\boldsymbol{\mu}_{2}$}, each cluster is represented as a Gaussian distribution

K-means



Gaussian Mixture Model (GMM)



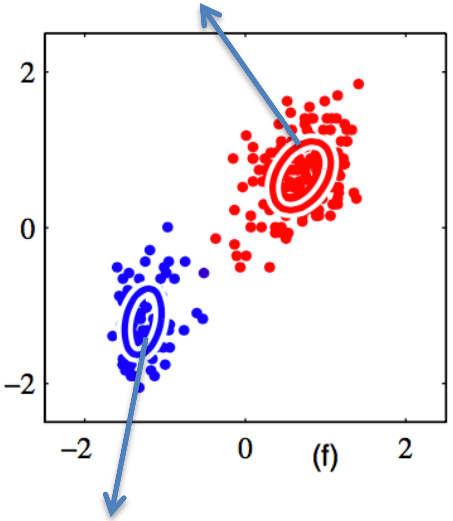
k=2

k=1

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

Gaussian Mixture Model (GMM)

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *k*=2



$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=1

We use $z_k = 1$ to indicate a point $\mathbf{x}$ belongs to cluster k

$$\mathbf{z} = (z_1, \ldots, z_K) \qquad z_k \in \{0, 1\} \qquad \sum_k z_k = 1$$

Assume the points in the same cluster follow a **Gaussian distribution**

$$\boxed{p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

A mixing weight for each cluster:

$$\boxed{p(z_k = 1) = \pi_k} \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

prior probability of point belonging to a cluster

So, we get a distribution for the data point $\mathbf{x}$:

$$\boxed{p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

From minimizing sum of square distances to finding maximum likelihood

minimize

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

maximize likelihood

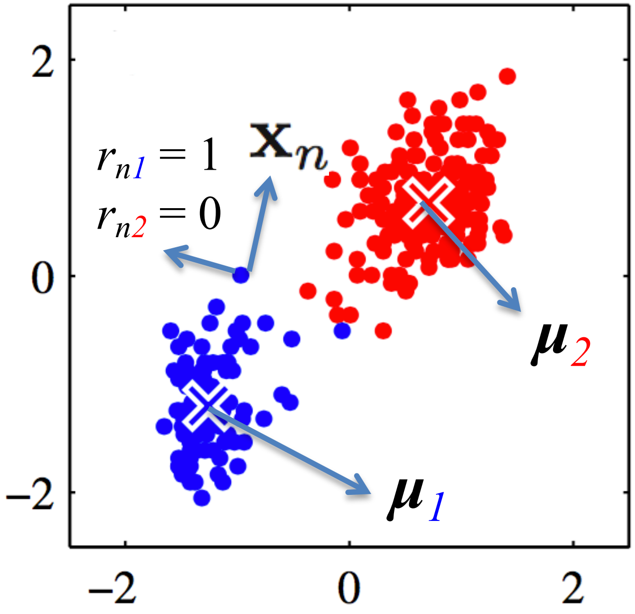
$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$

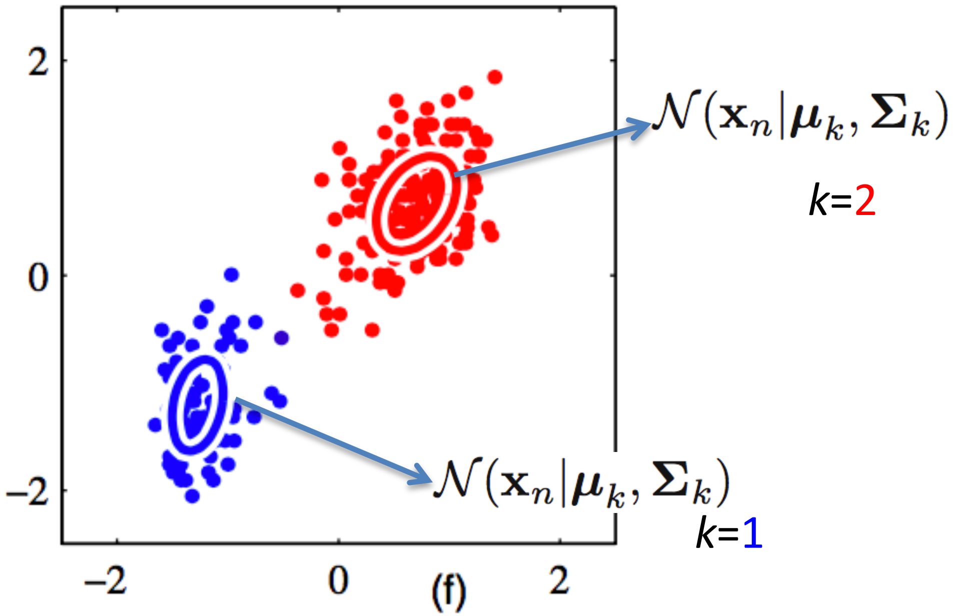
$X = \{x_1, \dots, x_N\}$

$\pi = \{\pi_1, \dots, \pi_K\}$

$\mu = \{\mu_1, \dots, \mu_K\}$

$\Sigma = \{\Sigma_1, \dots, \Sigma_K\}$





Remember: **The closer the distance, the more likely the probability.**

Maximum likelihood

Given a data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^{\mathrm{T}}$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, we can estimate the parameters of the distribution by maximum likelihood. The log likelihood function is given by

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

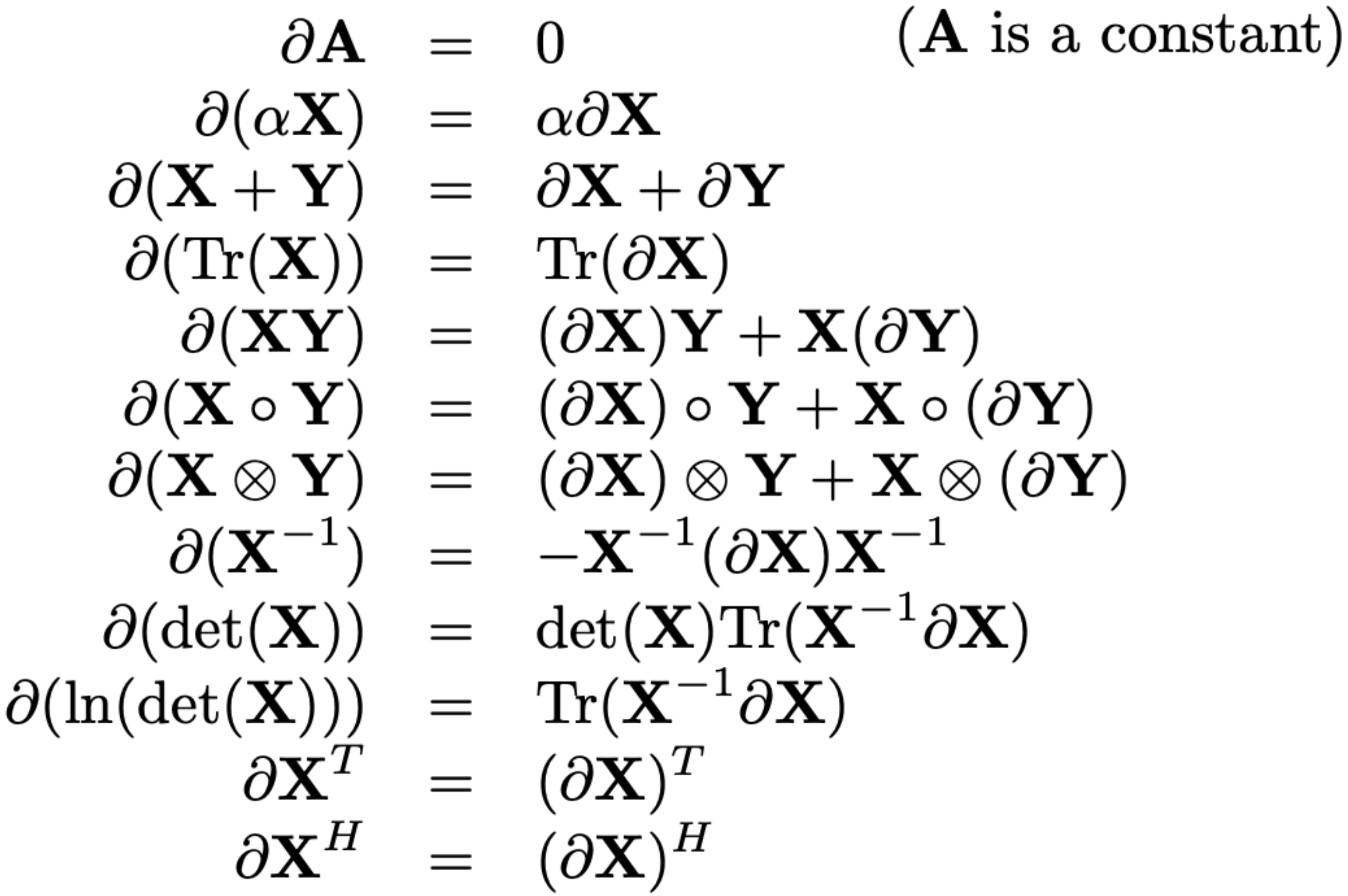
Maximizing the log-likelihood function:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0 \quad \longrightarrow \quad \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$$

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

$\boldsymbol{\mu}_{\mathrm{ML}}$ and $\boldsymbol{\Sigma}_{\mathrm{ML}}$ are the maximum likelihood estimates of the mean and the co-variance matrix.

Matrix-cook-book



$$
\begin{aligned}
\partial\mathbf{A} &= 0 \qquad\qquad\qquad\qquad (\mathbf{A}\text{ is a constant})\\
\partial(\alpha\mathbf{X}) &= \alpha\partial\mathbf{X}\\
\partial(\mathbf{X}+\mathbf{Y}) &= \partial\mathbf{X}+\partial\mathbf{Y}\\
\partial(\mathrm{Tr}(\mathbf{X})) &= \mathrm{Tr}(\partial\mathbf{X})\\
\partial(\mathbf{X}\mathbf{Y}) &= (\partial\mathbf{X})\mathbf{Y}+\mathbf{X}(\partial\mathbf{Y})\\
\partial(\mathbf{X}\circ\mathbf{Y}) &= (\partial\mathbf{X})\circ\mathbf{Y}+\mathbf{X}\circ(\partial\mathbf{Y})\\
\partial(\mathbf{X}\otimes\mathbf{Y}) &= (\partial\mathbf{X})\otimes\mathbf{Y}+\mathbf{X}\otimes(\partial\mathbf{Y})\\
\partial(\mathbf{X}^{-1}) &= -\mathbf{X}^{-1}(\partial\mathbf{X})\mathbf{X}^{-1}\\
\partial(\det(\mathbf{X})) &= \det(\mathbf{X})\mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X})\\
\partial(\ln(\det(\mathbf{X}))) &= \mathrm{Tr}(\mathbf{X}^{-1}\partial\mathbf{X})\\
\partial\mathbf{X}^{T} &= (\partial\mathbf{X})^{T}\\
\partial\mathbf{X}^{H} &= (\partial\mathbf{X})^{H}
\end{aligned}
$$

Outline

- Gaussian Mixture Models (GMM)

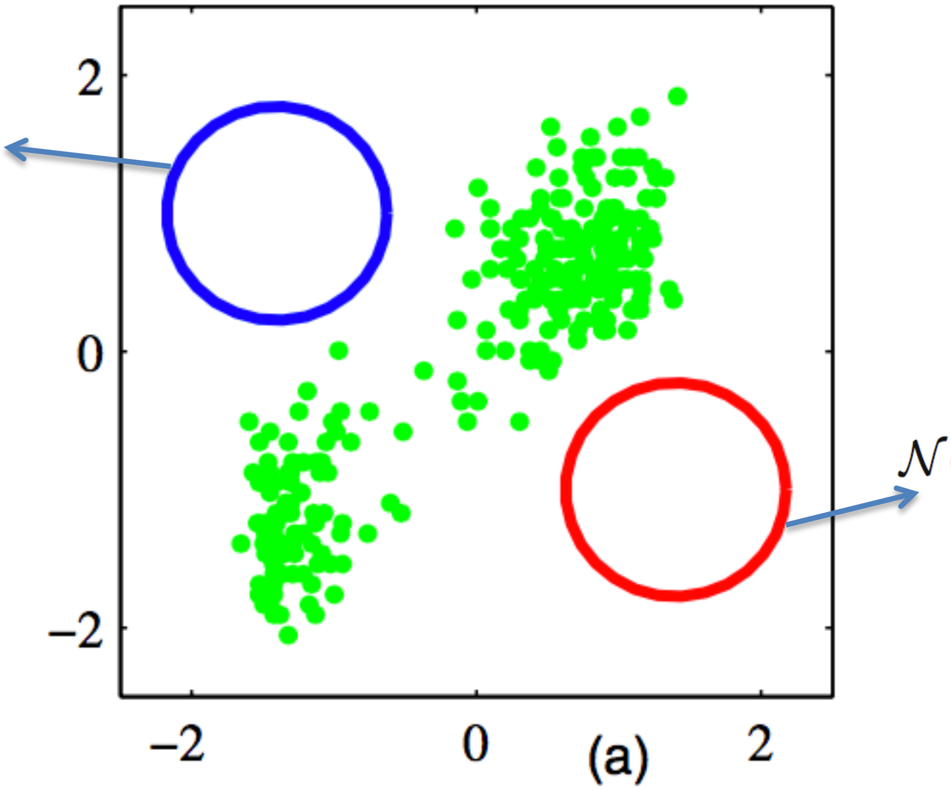
- **Expectation-Maximization (EM) for maximum likelihood**

- Gaussian Mixture Models (GMM)
  - From generation process perspective

Expectation-Maximization (EM) algorithm for maximum likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Initialization



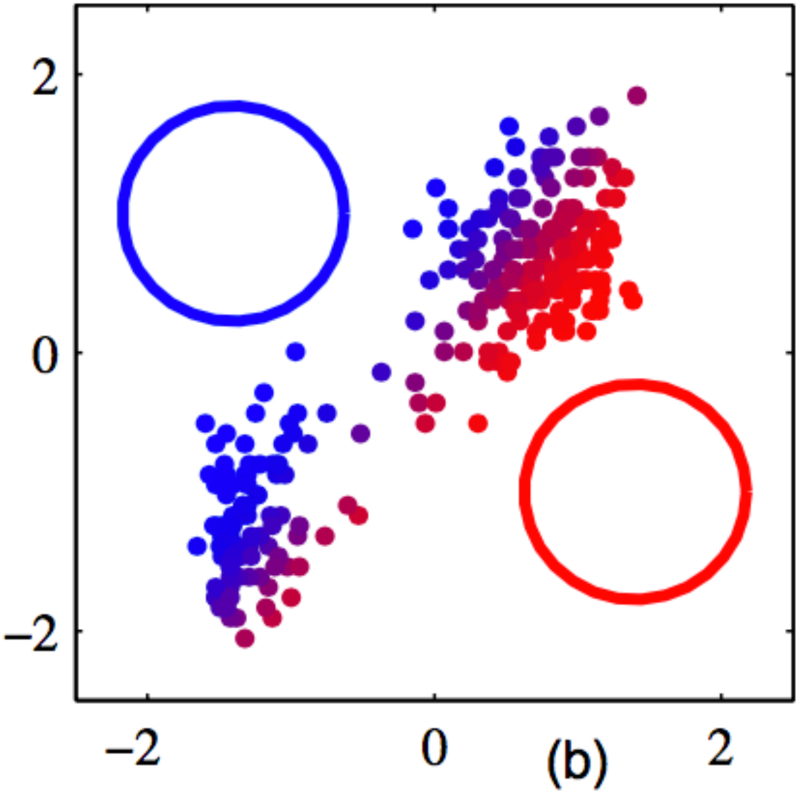
$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=1

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=2

E Step



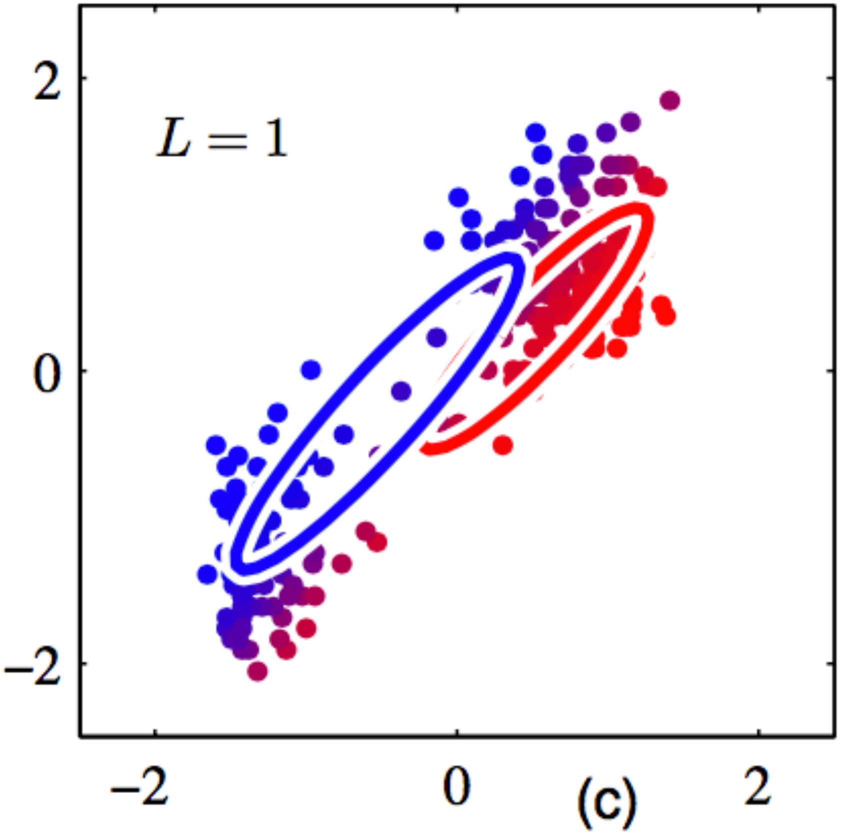
When the parameters are given, the assignments of the points can be calculated by the posterior probability, i.e., the probability of a data point belonging to a cluster once we have observed the data point.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Soft assignment:
A point fractionally belongs to two clusters.

For example,
0.2 belong to cluster 1
0.8 belong to cluster 2

M Step



When the assignments $\gamma(z_{nk})$ of the points to the clusters are known, parameters could be calculated for each cluster (Gaussian) separately.

Mixing weight π_k: the proportion of number of points in cluster k within all data points

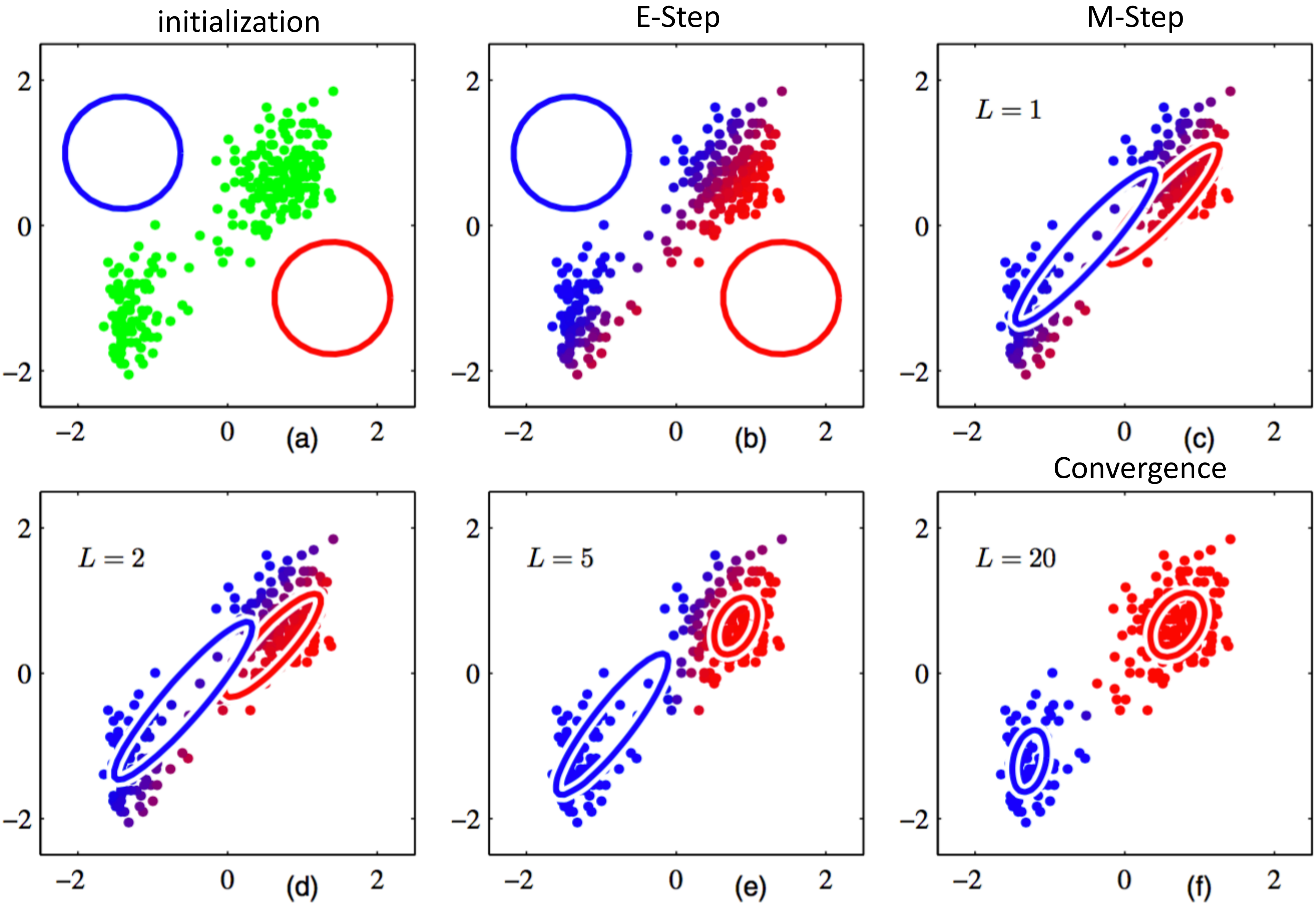
$$\pi_k = \frac{N_k}{N} \quad ; \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

$\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$: the mean and the covariance matrix are calculated for each cluster

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

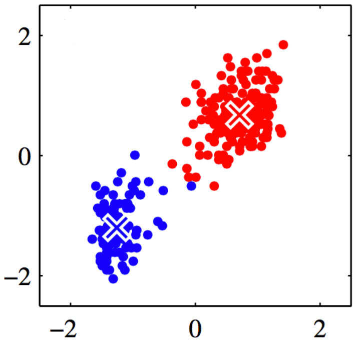
L denotes the number of cycles of the EM algorithm.



L denotes the number of cycles of E-Step and M-Step.

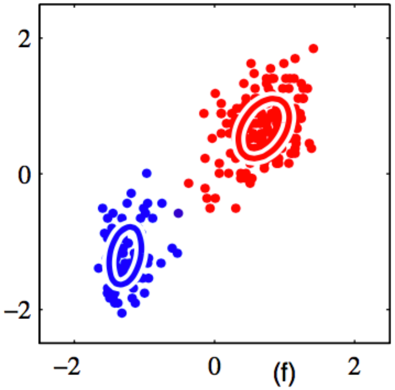
Relation to K-means

$$\|\mathbf{x}-\boldsymbol{\mu}_k\|^2$$



$$\{\mu_k\}$$

$$\Sigma_k = \varepsilon\mathbf{I}$$



GMM considers covariance and mixing weights.

$$p(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}-\boldsymbol{\mu}_k\|^2\right\}$$

One-in-K assignment

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

Soft assignment

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n-\boldsymbol{\mu}_j\|^2/2\epsilon\}}$$

Summary for the EM algorithm for GMM

- Does it find the global optimum?
  - No, like K-means, EM only finds the nearest local optimum and the optimum depends on the initialization

- GMM is more general then K-means by considering mixing weights, covariance matrices, and soft assignments.

- Like K-means, it does not tell you the best K.

Outline

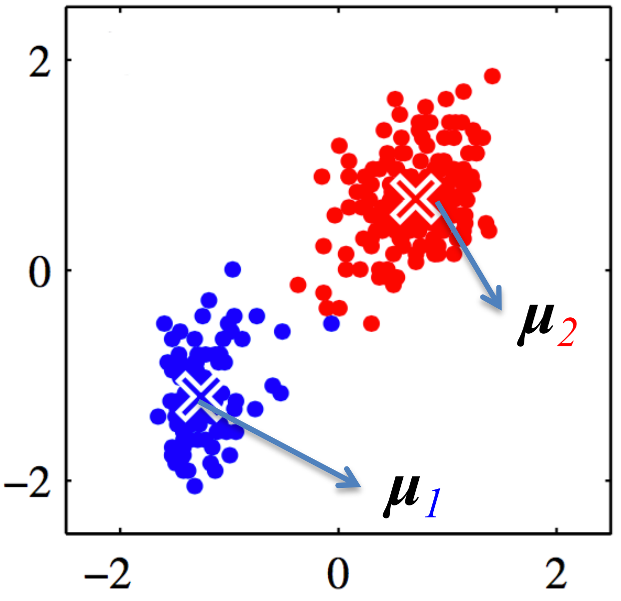
- Gaussian Mixture Models (GMM)

- Expectation-Maximization (EM) for maximum likelihood

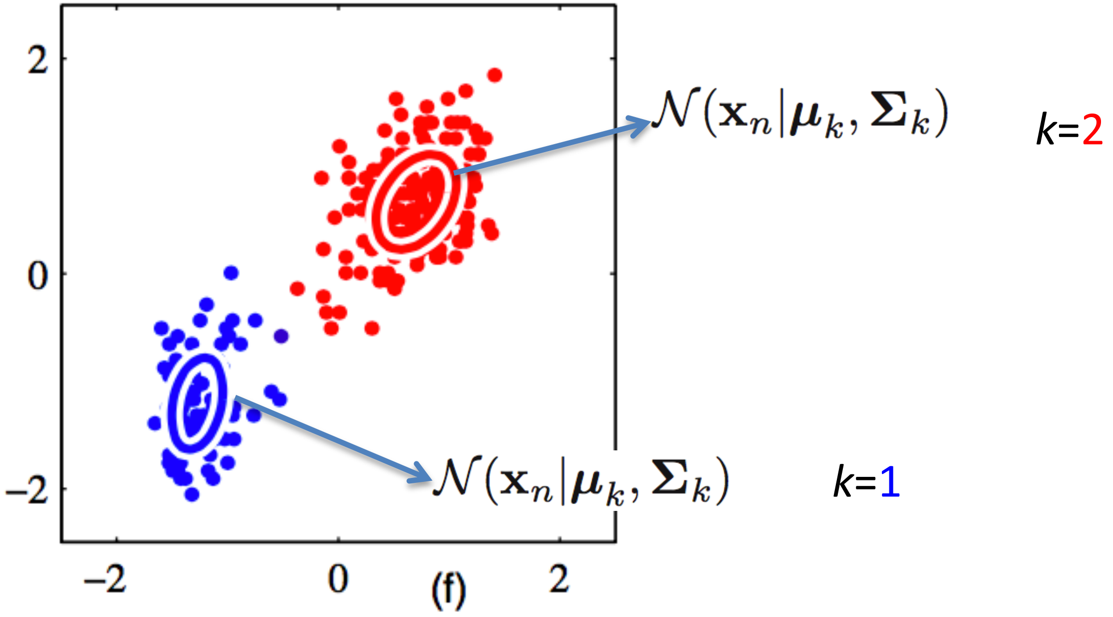
- Gaussian Mixture Models (GMM)
  - **From generation process perspective**

Instead if using $\{\boldsymbol{\mu}_{1,}\ \boldsymbol{\mu}_2\}$, each cluster is represented as a Gaussian distribution

K-means



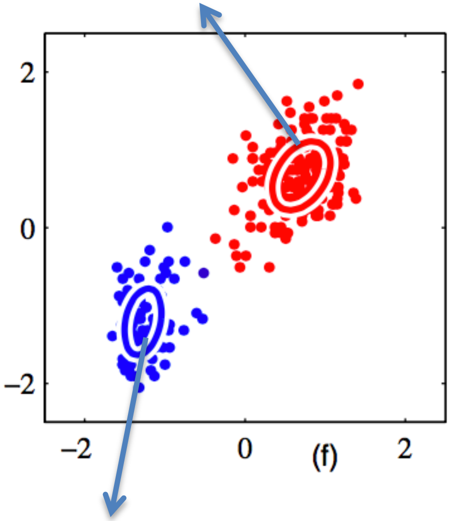
Gaussian Mixture Model (GMM)



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

Gaussian Mixture Model (GMM)

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *k*=2



$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=1

We use $z_k = 1$ to indicate a point $\mathbf{x}$ belongs to cluster k

$$\mathbf{z} = (z_1, \ldots, z_K) \qquad z_k \in \{0, 1\} \qquad \textstyle\sum_k z_k = 1$$

Assume the points in the same cluster follow a **Gaussian distribution**

$$\boxed{p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

A mixing weight for each cluster:

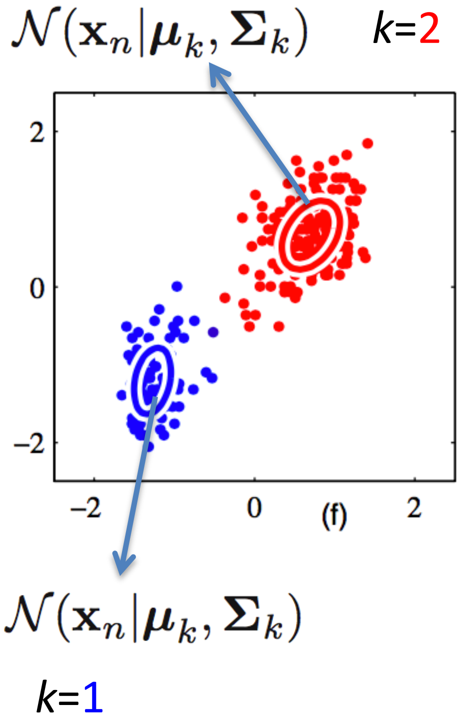
$$\boxed{p(z_k = 1) = \pi_k} \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$$

prior probability of point belonging to a cluster

So, we get a distribution for the data point $\mathbf{x}$:

$$\boxed{p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

Introduce a latent variable



$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ *k*=2

$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

k=1

We use $z_k = 1$ to indicate a point $\mathbf{x}$ belongs to cluster k

$\mathbf{z} = (z_1, \ldots, z_K)$ $\quad z_k \in \{0, 1\}$ $\quad \sum_k z_k = 1$

A mixing weight for each cluster:

$\boxed{p(z_k = 1) = \pi_k}$ $\quad 0 \leqslant \pi_k \leqslant 1 \quad \sum_{k=1}^{K} \pi_k = 1$

prior probability of point belonging to a cluster

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

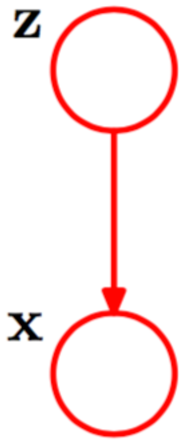
Assume the points in the same cluster follow a **Gaussian distribution**

$$\boxed{p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

Gaussian Mixture Model (GMM)

Generative process

- Randomly sample a **z** from a categorical distribution [π_1, …, π_K];
- Generate $\boldsymbol{x}$ according to Gaussian distribution $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$



Graphical representation of $\mathbf{p}(\mathbf{x}, \mathbf{z}) = \mathbf{p}(z)\mathbf{p}(\mathbf{x}|\mathbf{z})$

So, we get a distribution for the data point **x**:

$$\boxed{p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

From minimizing sum of square distances to finding maximum likelihood

minimize

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

maximize likelihood

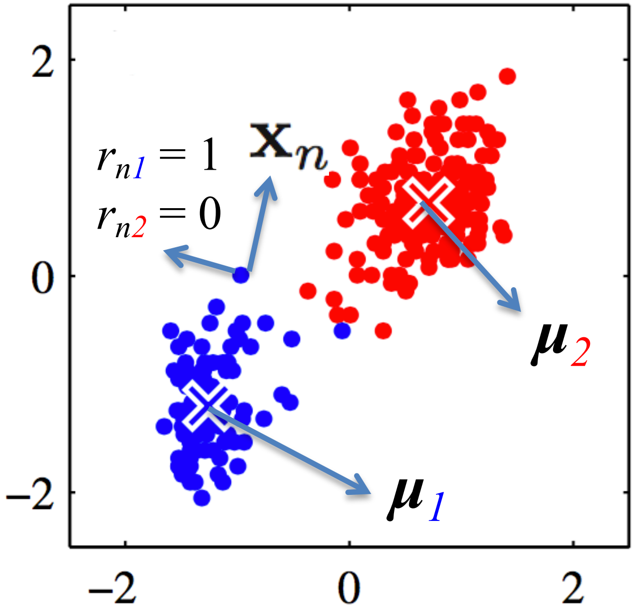
$$p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$$

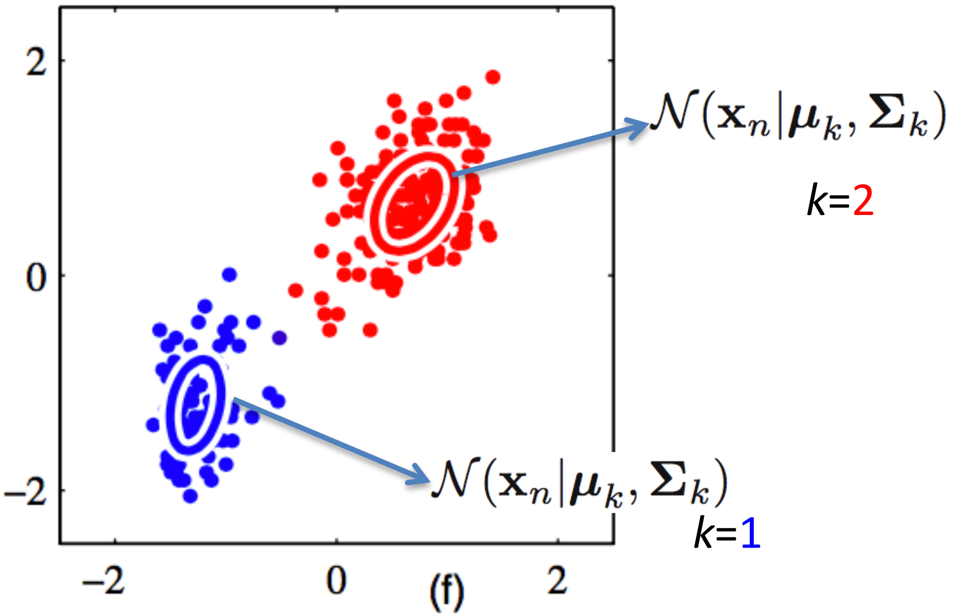
$X = \{x_1,\dots,x_N\}$

$\pi = \{\pi_1,\dots,\pi_K\}$

$\mu = \{\mu_1,\dots,\mu_K\}$

$\Sigma = \{\Sigma_1,\dots,\Sigma_K\}$





Remember: **The closer the distance, the more likely the probability.**

Thank you!