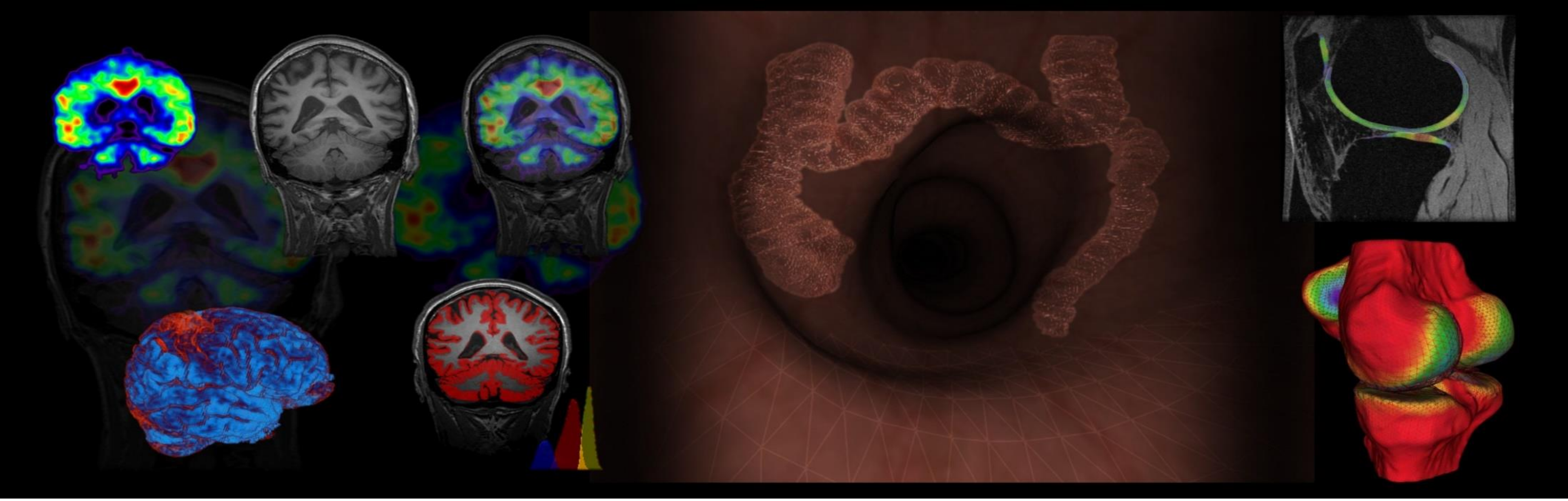


Maximum Likelihood
Expectation Maximization Algorithm

Part of the Slides are from P. Bourgeat (CSIRO)
and ICCV 2003

Math reminder: Rules of Probability

Let X and Y be random variables.

1. Sum rule

$$\sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}) = p(\mathbf{X})$$

2. Product rule

$$p(\mathbf{X},\mathbf{Y}) = p(\mathbf{X}\,|\,\mathbf{Y})\,p(\mathbf{Y})$$

3. Bayes theorem

$$p(\mathbf{Y}\,|\,\mathbf{X}) = \frac{p(\mathbf{X}\,|\,\mathbf{Y})\,p(\mathbf{Y})}{p(\mathbf{X})}$$

4. Using the sum and product rules, this can also be written

$$p(\mathbf{Y}\,|\,\mathbf{X}) = \frac{p(\mathbf{X}\,|\,\mathbf{Y})\,p(\mathbf{Y})}{\sum_{\mathbf{Y}} p(\mathbf{X}\,|\,\mathbf{Y})\,p(\mathbf{Y})}$$

Likelihood

- We have a density function $p(\mathbf{X} \mid \boldsymbol{\theta})$ that is governed by the set of parameters $\boldsymbol{\theta}$ (p could be a set of Gaussians and $\boldsymbol{\theta}$ their means and covariance)

- We also have a data set of size N supposedly drawn from this distribution, i.e., $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N)$

- If these data vectors are independent and identically distributed (i.i.d.) with distribution *p*, the resulting density for the samples is

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid \mathbf{X})$$

- $L(\boldsymbol{\theta} \mid \mathbf{X})$ is called the likelihood of the parameters given the data

Maximum Likelihood

- In the maximum likelihood problem, our goal is to find the $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta} \mid \mathbf{X})$

- We wish to find $\boldsymbol{\theta}^{opt}$ where $\boldsymbol{\theta}^{opt} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta} \mid \mathbf{X})$

- Often, we maximize the log of the likelihood since it is equivalent but easier to manipulate.

$$L(\boldsymbol{\theta} \mid \mathbf{X}) = \prod_{i=1}^{N} p(\mathbf{x}_i \mid \boldsymbol{\theta})$$

$$\log(L(\boldsymbol{\theta} \mid \mathbf{X})) = \sum_{i=1}^{N} \log(p(\mathbf{x}_i \mid \boldsymbol{\theta}))$$

Maximum Likelihood: Simple Example

- Simple case: if $p(\mathbf{X} \mid \boldsymbol{\theta})$ is a single Gaussian distribution with

$$\boldsymbol{\theta} = (\mu, \sigma^2)$$

- then we can set the derivative of $\log(L(\boldsymbol{\theta} \mid \mathbf{X}))$ to zero, and solve directly for μ and σ^2.

$$p(x_1, \ldots, x_n | \mu, \sigma) = \prod \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/(2\sigma^2)}$$

$$= \frac{(2\pi)^{-n/2}}{\sigma^n} \exp\left[-\frac{\sum (x_i - \mu)^2}{(2\sigma^2)} \right]$$

$$\ln p = -\frac{1}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum (x_i - \mu)^2}{(2\sigma^2)}$$

Maximum Likelihood: Simple Example

$$\ln p = -\frac{1}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum (x_i - \mu)^2}{(2\sigma^2)}$$

For the mean:

$$\frac{\partial(\ln p)}{\partial\mu} = \frac{\sum (x_i - \mu)}{\sigma^2} = 0$$

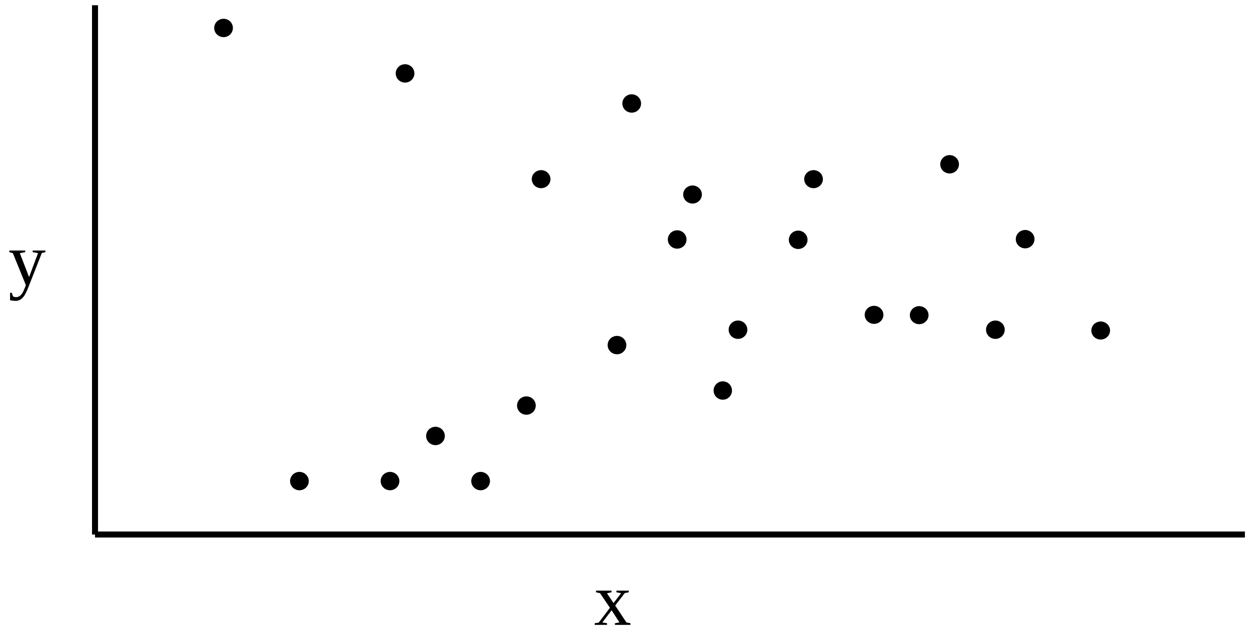
$$\boxed{\mu = \frac{\sum x_i}{n}}$$

For the standard deviation:

$$\frac{\partial(\ln p)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{\sum (x_i - \mu)^2}{\sigma^3} = 0$$

$$\boxed{\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}}$$

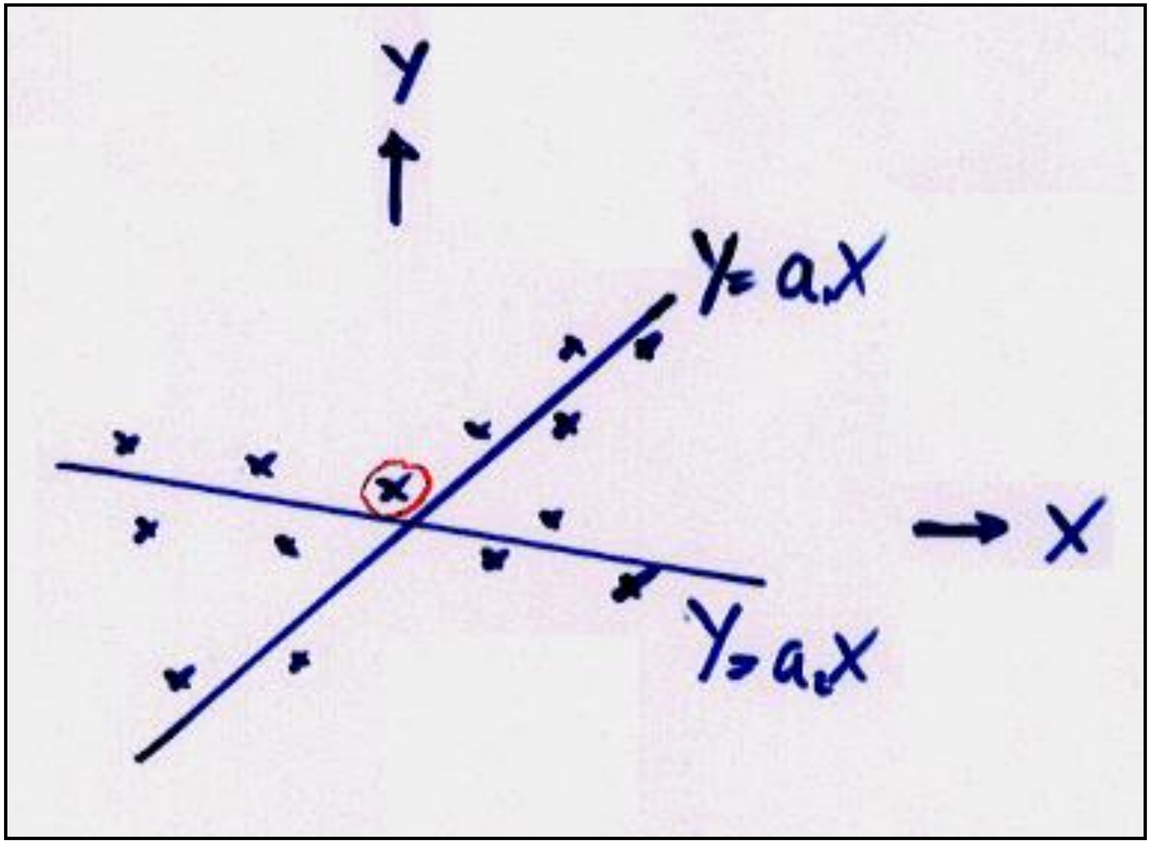
Model fitting
Fitting two lines to observed data



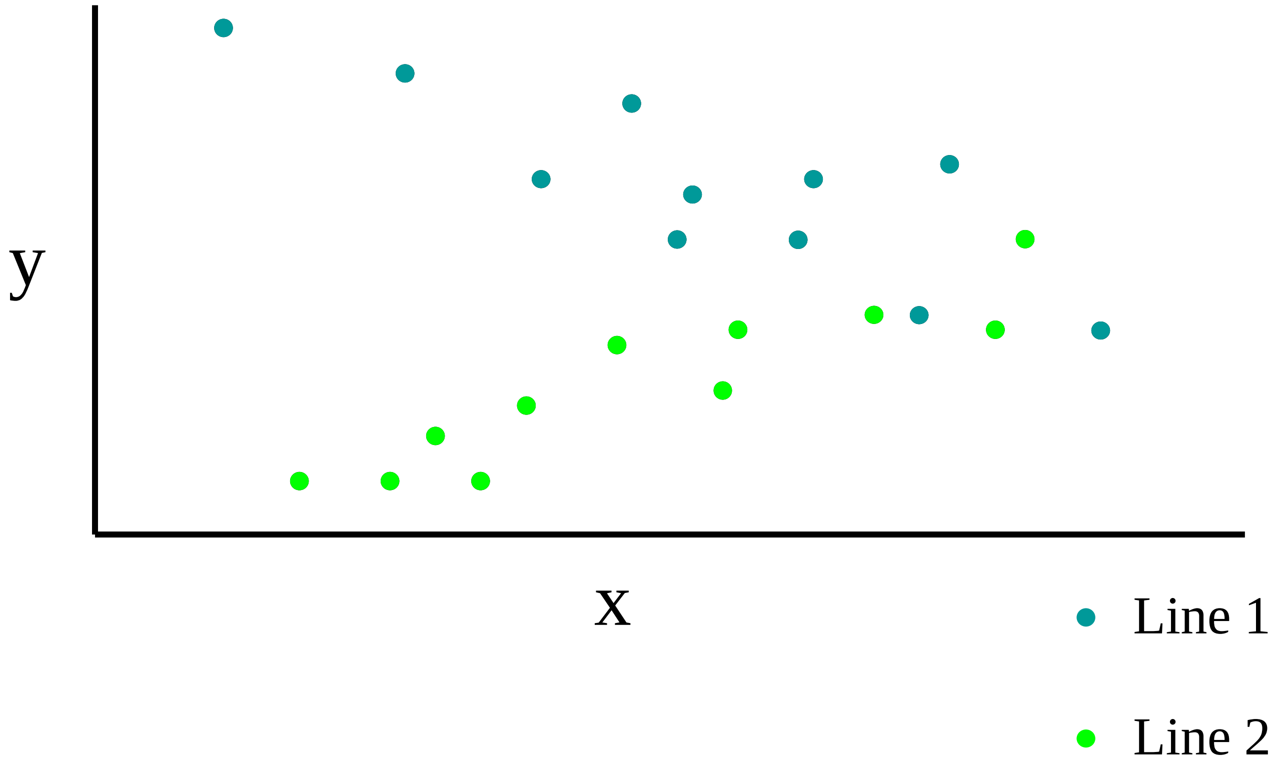
MLE for fitting a line pair

Lines $Y = a_1 X + w$ or $Y = a_2 X + w$, with $w \sim \mathcal{N}(0, 1)$.

(a form of mixture dist. for $\boldsymbol{Y}$)

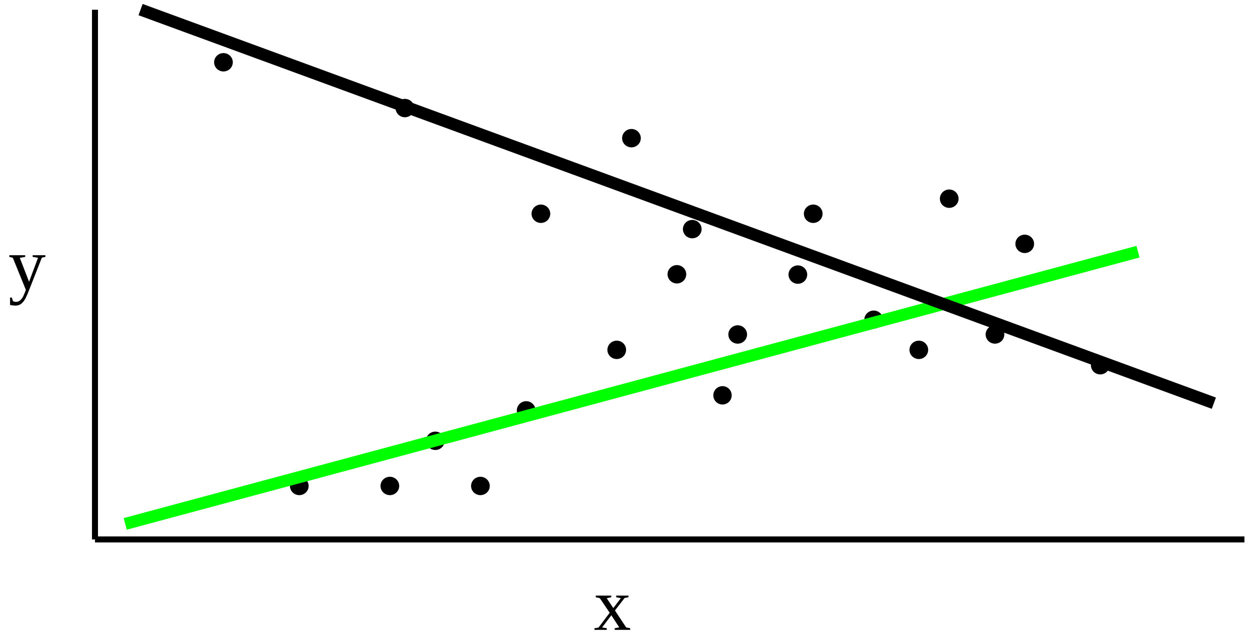


Fitting two lines: on the one hand…



If we knew which points went with which lines, we'd be back at the single line-fitting problem, twice.

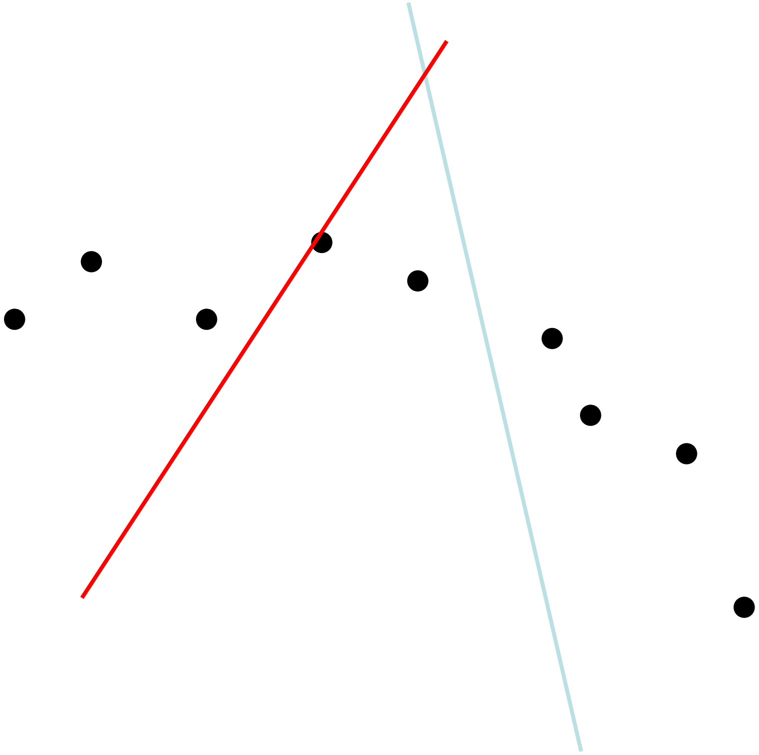
Fitting two lines, on the other hand…



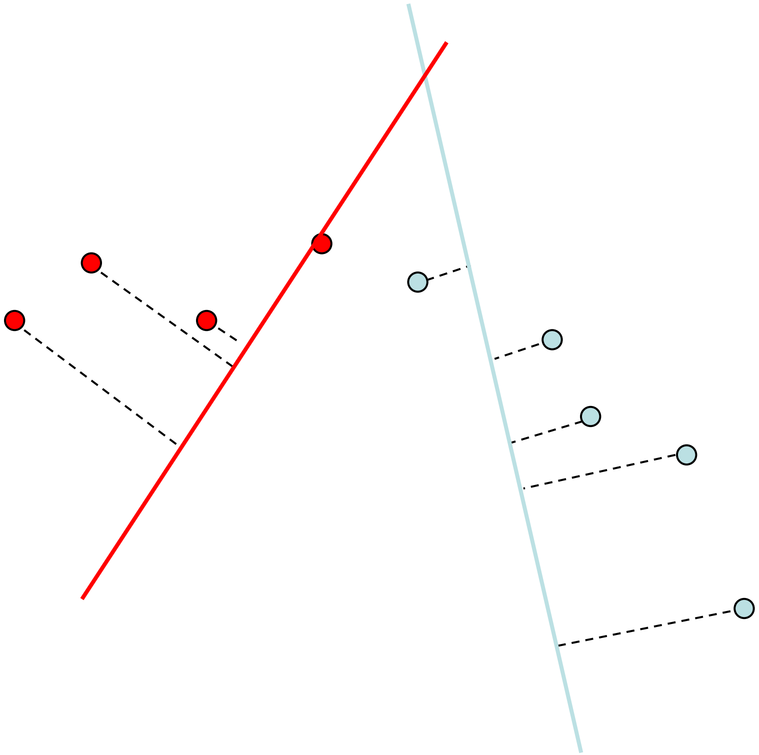
We could figure out the probability that any point came from either line if we just knew the two equations for the two lines.

Expectation Maximization (EM): a solution to chicken-and-egg problems

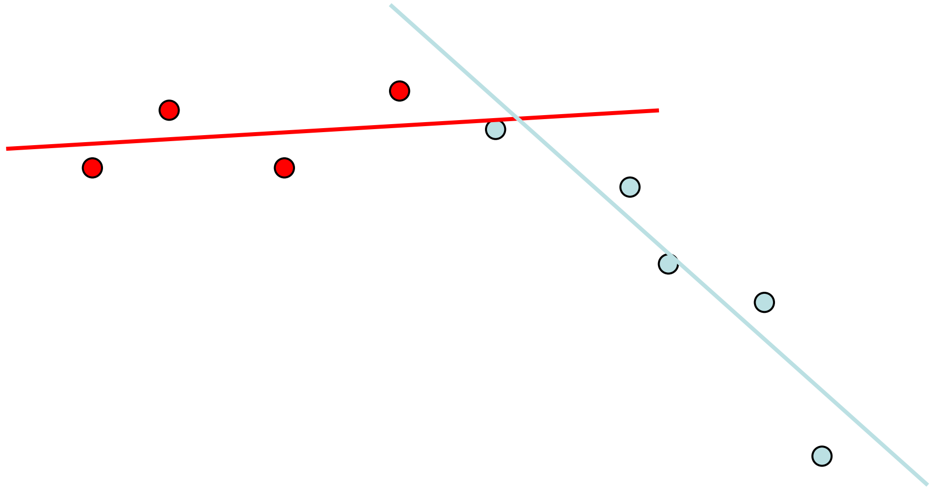
EM example:



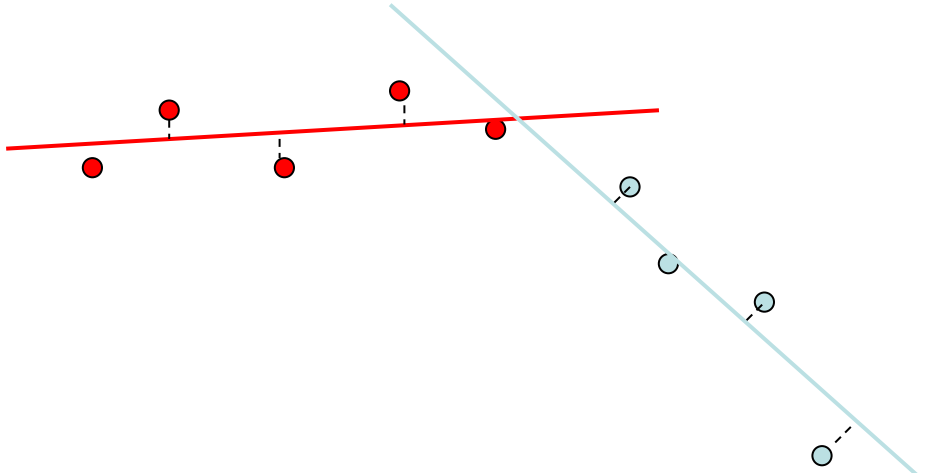
EM example:



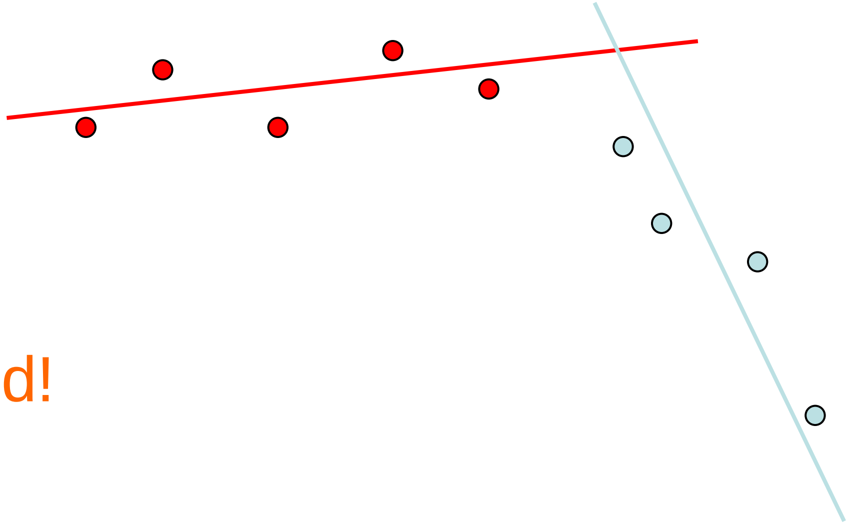
EM example:



EM example:



EM example:

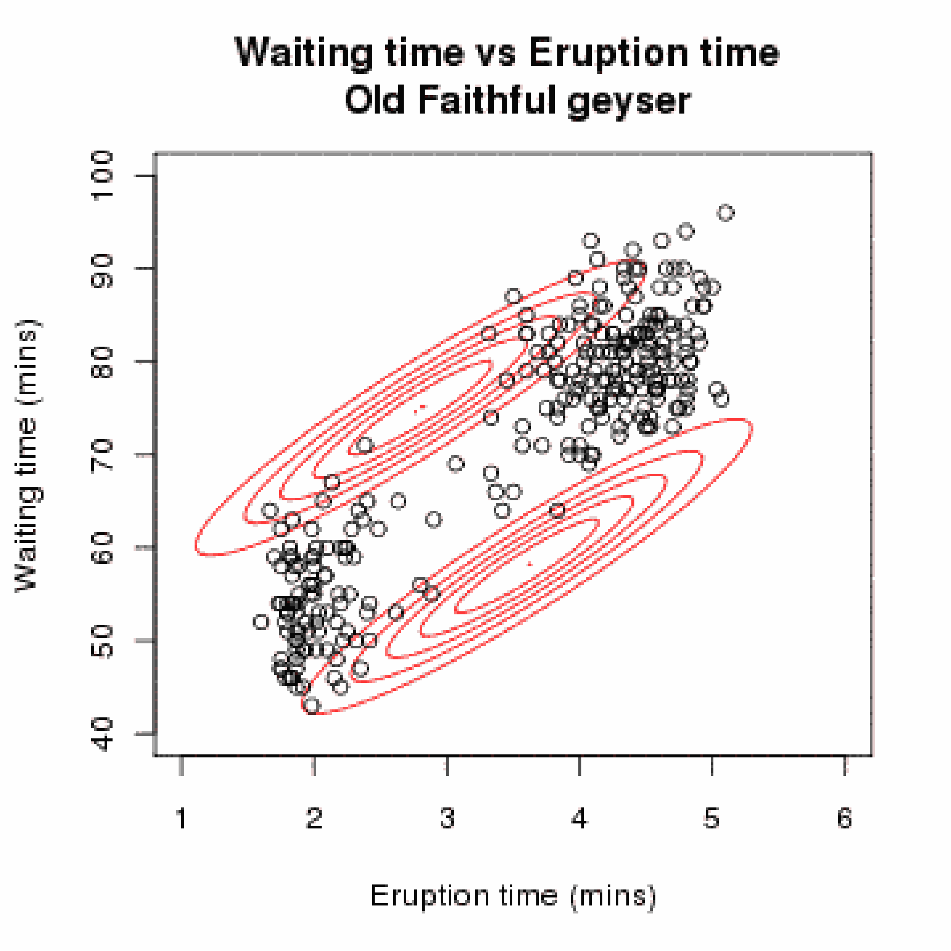


Converged!

Expectation Maximisation Algorithm

- The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data.
- In ML estimation, we wish to estimate the model parameter(s) for which the observed data are the most likely.
- Each iteration of the EM algorithm consists of two processes: The E-step,and the M-step.
  - In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters.
  - In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step are used in lieu of the actual missing data.
- Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

The Expectation Maximization Algorithm



The Expectation Maximization Algorithm

- Given: observed data $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N)$
- Goal: describe probability distribution by a parametric model *M* with parameters $\boldsymbol{\theta}$: $p(\mathbf{X} \mid M, \boldsymbol{\theta}) = p(\mathbf{X} \mid \boldsymbol{\theta})$
- The expectation maximization (EM) algorithm is a method to determine the optimal parameters $\boldsymbol{\theta}_{opt}$ by *Maximum Likelihood* (ML) for models with hidden (i.e. unobserved) variables $\mathbf{Y}$:

$$\boldsymbol{\theta}_{opt} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X} \mid \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{Y}} p(\mathbf{X}, \mathbf{Y} \mid \boldsymbol{\theta})$$

The EM algorithm is widely applicable

Many powerful probabilistic models can be formulated as models with hidden variables that are accessible with the EM algorithm.

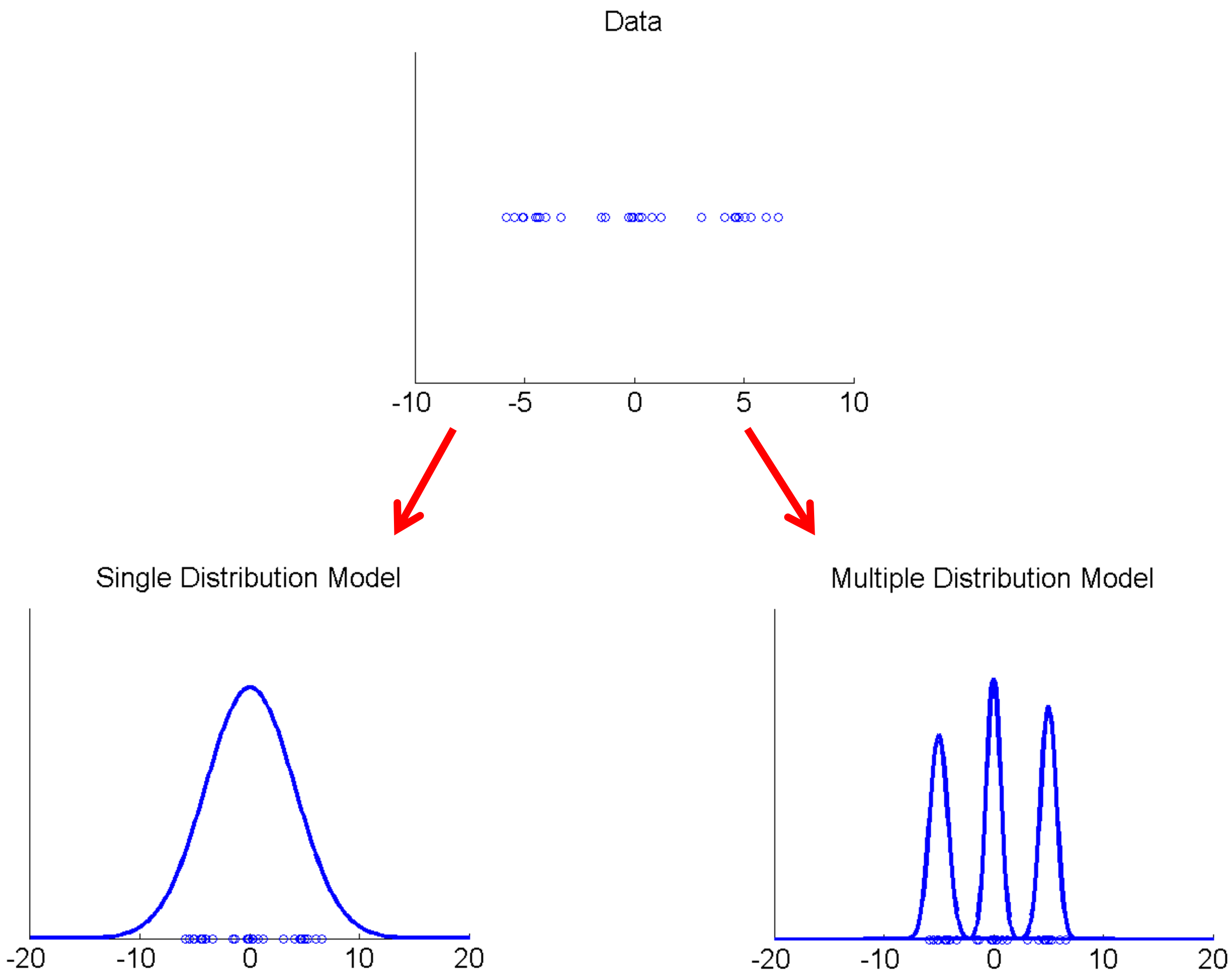
- Gaussian mixture model (and other mixture models)

$$p(\mathbf{x}\,|\,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_{\mathbf{k}} N(\mathbf{x}\,|\,\boldsymbol{\mu}_k,\Sigma_k)$$

$$= \sum_{k=1}^{K} p(\mathbf{x}\,|\,y=k,\mu_k,\Sigma_k)\;\; p(y=k)$$

Mixture weight

Mixture Model



Mixture Model

- A Mixture Model is a weighed sum of distributions

$$p(\mathbf{x}) = \pi_1 p_1(\mathbf{x}) + \pi_2 p_2(\mathbf{x}) + \cdots + \pi_K p_K(\mathbf{x})$$

- Weights are called “mixing coefficients”

- Mixing coefficients must sum to 1 so that the resulting distribution has the properties of a probability distribution

$$\sum_{k=1}^{K} \pi_{\mathbf{k}} = 1$$

- Each distribution must be a proper distribution

$$\int p_k(\mathbf{x}) d\mathbf{x} = 1$$

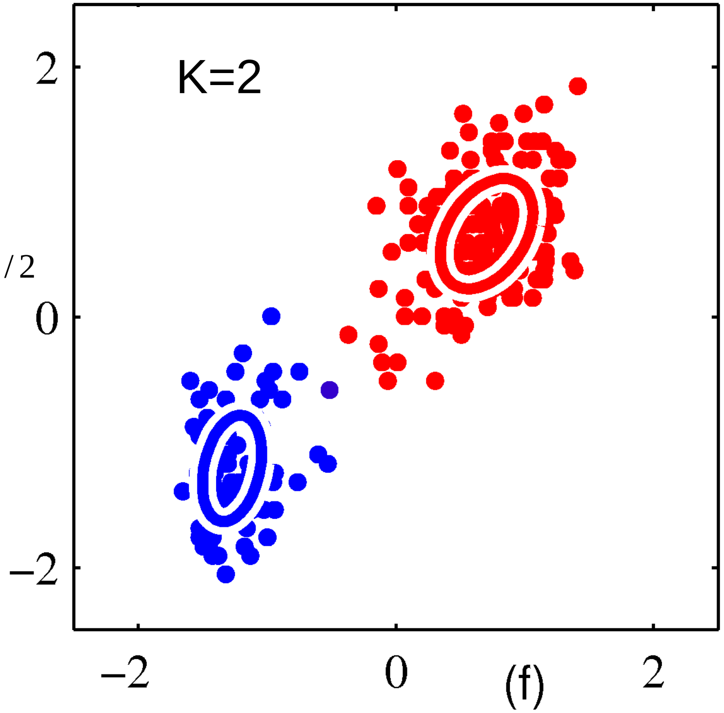
Example: Gaussian mixture model

N observed variables $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N)$, K mixture components

Likelihood for single observation

$$p(\mathbf{x}_n \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_{\mathbf{k}} N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)$$

$$N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} \left|\Sigma\right|^{1/2}} e^{-(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \Sigma^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)/2}$$



(f)

Likelihood for all observations

$$p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_{\mathbf{k}} N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)$$

Log Likelihood

$$\log\left(p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_{\mathbf{k}} N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)\right)$$

Gaussian Mixtures: Responsabilities

The likelihood for the Gaussian mixture is

$$p(\mathbf{x}_n \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_{\mathbf{k}} N(\mathbf{x}_n \mid \mu_k, \Sigma_k) = \sum_{\mathbf{y}_n} p(\mathbf{y}_n) p(\mathbf{x}_n \mid \mathbf{y}_n) = \sum_{\mathbf{z}_n} p(\mathbf{x}_n, \mathbf{y}_n)$$

The probability for observation x_n being generated by component k is (using Bayes‘ Theorem):

$$p(y_n = k \mid \mathbf{x}_n) = \frac{p(\mathbf{x}_n \mid y_n = k) p(y_n = k)}{\sum_{l=1}^{K} p(\mathbf{x}_n \mid y_n = l) p(y_n = l)} = \frac{\pi_k N(\mathbf{x}_n \mid \mu_k, \Sigma_k)}{\sum_l \pi_l N(\mathbf{x}_n \mid \mu_l, \Sigma_l)} = r_{nk}$$

These r_{nk} are called the responsibilities of component k for observation x_n.

It is the value of the k^{th} function evaluated at x_n divided by the sum of all of the weighted distributions evaluated at x_n.

Try to maximize the likelihood directly...

$$L(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log\Big(p(\mathbf{X}\,|\,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})\Big) = \sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\pi_{\mathbf{k}}\,N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)\right)$$

Maximizing wrt $\boldsymbol{\mu}$

$$\boxed{\frac{d\ln(u)}{dx} = \frac{1}{u}\frac{du}{dx}}$$

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N}\frac{\pi_k}{\sum_l \pi_l N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_l,\Sigma_l)}\,\frac{\partial N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)}{\partial \boldsymbol{\mu}_k} = 0$$

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N}\underbrace{\frac{\pi_k N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)}{\sum_l \pi_l N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_l,\Sigma_l)}}_{r_{nk}}\Sigma^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

$$\boxed{\frac{d\exp(u)}{dx} = \frac{du}{dx}\exp(u)}$$

Multiply with Σ^{1}

$$\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = \sum_{n=1}^{N} r_{nk}\mathbf{x}_n - \overbrace{\left(\sum_{n=1}^{N} r_{nk}\right)}^{N_k = \sum_{n=1}^{N} r_{nk}}\mu_k = 0$$

$$\boxed{\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}$$

This looks like the usual ML estimator for the mean, except that each point x_n contributes only with weight r_{nk}

Try to maximize the likelihood directly...

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

Maximizing wrt Σ_k, or rather wrt Σ_k^{-1} yields :

$$\frac{d\ln(u)}{dx} = \frac{1}{u}\frac{du}{dx}$$

$$\frac{\partial L}{\partial \Sigma_k^{-1}} = \sum_{n=1}^{N} \frac{\pi_k}{\sum_k \pi_{\mathbf{k}} N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)} \frac{\partial N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)}{\partial \Sigma_k^{-1}} = 0$$

$$= \sum_{n=1}^{N} r_{nk}\left[\frac{1}{|\Sigma_k^{-1}|^{1/2}} \frac{\partial |\Sigma_k^{-1}|^{1/2}}{\partial \Sigma_k} + \frac{\partial}{\partial \Sigma_k}\left(-\frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^t \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right)\right]$$

…

$$\boxed{\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^t}$$

This looks like the usual ML estimator for the covariance matrix, except that each point x_n contributes only with weight r_{nk}

Try to maximize the likelihood directly...

$$L(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \log\left(p(\mathbf{X}\mid\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})\right) = \sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\pi_{\mathbf{k}}N(\mathbf{x}_n\mid\boldsymbol{\mu}_k,\Sigma_k)\right)$$

Maximizing wrt π_k under the constraint $g(\boldsymbol{\pi}) = 1 - \sum_{k=1}^{K}\pi_k = 0$

Maximizing with constraints: Use method of *Lagrange multipliers*.

A necessary condition for $L(\boldsymbol{\pi})$ to have a maximum under the constaint $g(\boldsymbol{\pi})$ is

$$\exists\lambda\in R : \nabla L(\boldsymbol{\pi}) + \lambda\nabla g(\boldsymbol{\pi}) = 0$$

$$\Rightarrow \quad \frac{\partial L(\boldsymbol{\pi})}{\partial\pi_k} + \lambda\frac{\partial g(\boldsymbol{\pi})}{\partial\pi_k} = \sum_{n=1}^{N}\frac{N(\mathbf{x}_n\mid\boldsymbol{\mu}_k,\Sigma_k)}{\sum_l\pi_l N(\mathbf{x}_n\mid\boldsymbol{\mu}_l,\Sigma_l)} + \lambda = 0$$

Try to maximize the likelihood directly...

$$\Rightarrow \qquad \frac{\partial L(\boldsymbol{\pi})}{\partial \pi_k} + \lambda \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_k} = \sum_{n=1}^{N} \frac{N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_l \pi_l N(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l)} + \lambda = 0$$

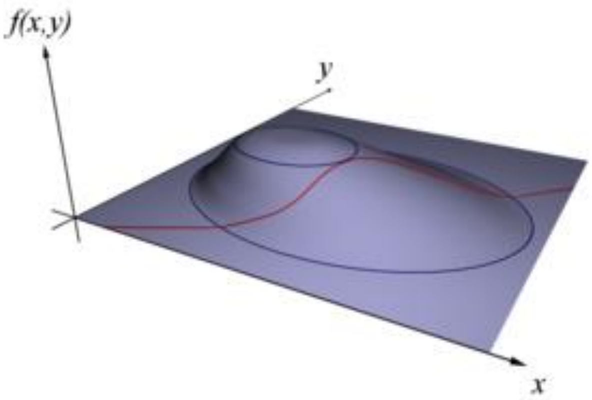
Factoring π_l :

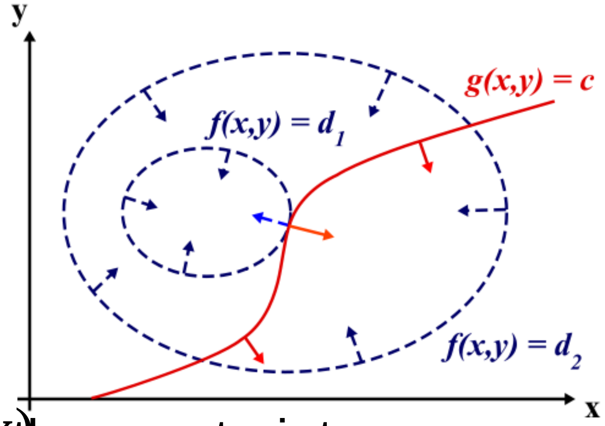
$$\pi_l = \frac{1}{\lambda} \sum_{n=1}^{N} \frac{N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_l N(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l)}$$

$$1 = \sum_{k=1}^{K} \pi_k = \frac{1}{\lambda} \sum_{n=1}^{N} \frac{\sum_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_l N(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l)} = \frac{1}{\lambda} \sum_{n=1}^{N} 1 \qquad \Rightarrow \qquad \lambda = N$$

Multipliying by π_k and rearranging yields $\boxed{\pi_k = \frac{N_k}{N}}$ with $N_k = \sum_{n=1}^{N} r_{nk}$

Lagrange Multipliers





Let x* be a maximum of $f(\mathbf{x})$ under the constraint $g(\mathbf{x})=0$

- $\nabla g(\mathbf{x})$ must be perpendicular to the g(x)=0 hypersurface (red). Otherwise, it would have a component parallel to it, and moving along that component would not leave g(x) equal to 0.
- $\nabla f(\mathbf{x})$ must be perpendicular to the g(x)=0 hypersurface. Otherwise, f(x) would be further maximized by moving along the gradient's component parallel to the hypersurface.
- Therefore, $\nabla g(\mathbf{x})$ and $\nabla f(\mathbf{x})$ must be parallel to each other, or, in other words

$$\boxed{\nabla f(\mathbf{x}) = -\lambda \nabla g(\mathbf{x}) \quad \text{for some} \quad \lambda \in R}$$

Try to maximize the likelihood directly...

Conditions for ML solution:

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n \qquad \Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^t \qquad \pi_k = \frac{N_k}{N}$$

with

$$r_{nk} = \frac{\pi_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_l \pi_l N(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l)} \qquad N_k = \sum_{n=1}^{N} r_{nk}$$

Problem: the require r_{nk} knowledge of $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$

Hence, the equations do not allow a closed form solution!

Solution: iterate between estimating and r_{nk} $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$

The EM Algorithm for Gaussian mixtures

1. Initialize $\pi_k, \boldsymbol{\mu}_k, \Sigma_k$ randomly

2. Expectation step: evaluate the responsibilities r_{nk}

$$r_{nk} = \frac{\pi_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{old}, \Sigma_k^{old})}{\sum_l \pi_l N(\mathbf{x}_n \mid \boldsymbol{\mu}_l^{old}, \Sigma_l^{old})} \quad \text{and} \quad N_k = \sum_{n=1}^{N} r_{nk}$$

3. Maximization step: update parameters

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \mathbf{x}_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k^{new})(\mathbf{x}_n - \boldsymbol{\mu}_k^{new})^t$$

$$\pi_k^{new} = \frac{N}{N_k}$$

4. Evaluate likelihood and check for convergence

The EM Algorithm – derivation

Given: a probabilistic model with hidden variables: $p(\mathbf{X}\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})$

Goal: estimate ML parameters

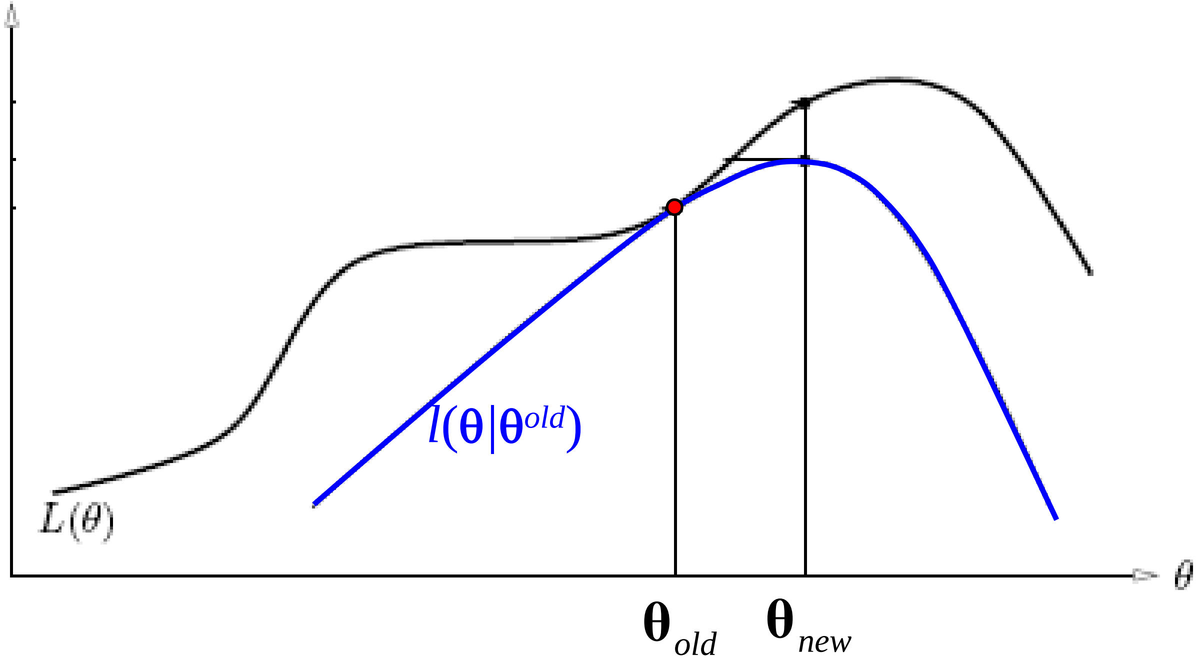
Log of a sum is hard to derive

$$\boldsymbol{\theta}_{opt} = \arg\max_{\boldsymbol{\theta}} \big(L(\boldsymbol{\theta}\,|\,\mathbf{X})\big) = \arg\max_{\boldsymbol{\theta}} \big(\log p(\mathbf{X}\,|\,\boldsymbol{\theta})\big) = \arg\max_{\boldsymbol{\theta}} \left(\log \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})\right)$$

Idea: Iteratively improve estimation of θ by maximizing a function $l(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{old})$ for which

$$l(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{old}) \le L(\boldsymbol{\theta})$$

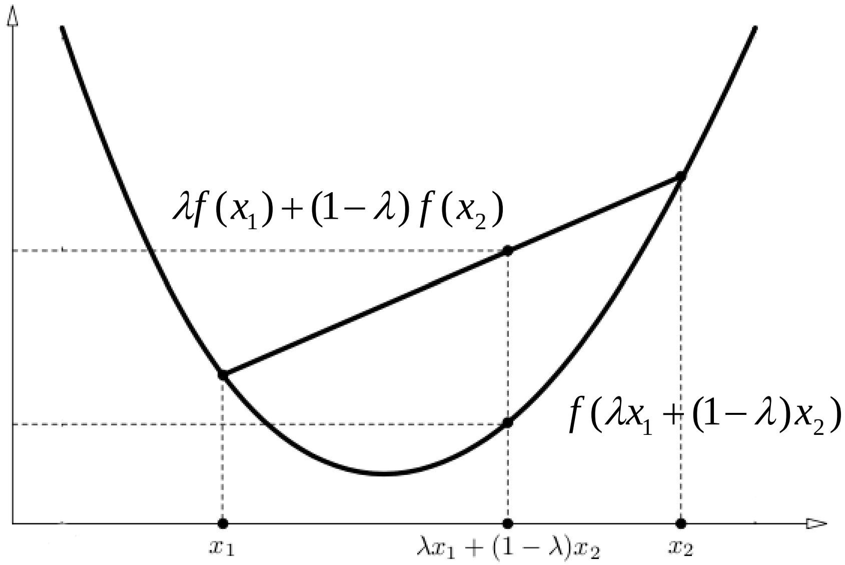
$$l(\boldsymbol{\theta}^{old}\,|\,\boldsymbol{\theta}^{old}) = L(\boldsymbol{\theta}^{old})$$



Jenssen‘s inequality for convex functions

Let *f(x)* be a function with *f“(x)*≥0 everywhere. Then *f(x)* can be shown to be *convex*, which means that for all $\lambda \in [0,1]$:

$$f(\lambda x_1 + (1-\lambda)x_2) \\ \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$



$\lambda f(x_1) + (1-\lambda) f(x_2)$
$f(\lambda x_1 + (1-\lambda)x_2)$
x_1
$\lambda x_1 + (1-\lambda)x_2$
x_2

- *Jenssen‘s inequality*: Let *f(x)* be convex. Then, for all $\lambda_k \in [0,1]$ with $\Sigma_k \lambda_k = 1$, the following can be proved by mathematical induction:

$$f(\sum_{k=1}^{K} \lambda_k x_k) \leq \sum_{k=1}^{K} \lambda_k f(x_k)$$

- If f‘‘(x)≤0 everywhere, it is said to be *concave*. In that case, Jenssens inequality is inverted. The log function is concave, therefore

$$\log(\sum_{k=1}^{K} \lambda_k x_k) \geq \sum_{k=1}^{K} \lambda_k \log(x_k)$$

The EM Algorithm – derivation

Goal: Maximize $L(\boldsymbol{\theta})$

$$L(\boldsymbol{\theta}) = \log \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y} \mid \boldsymbol{\theta}) = \log \sum_{\mathbf{Y}} \underline{p(\mathbf{Y} \mid \mathbf{X},\boldsymbol{\theta}^{old})} \frac{p(\mathbf{X},\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y} \mid \mathbf{X},\boldsymbol{\theta}^{old})}$$

arbitrary probability distribution of the hidden variables $\mathbf{Y}$

Apply Jenssen's inequality

$$\log\Big(\sum_{k=1}^{K} \lambda_k x_k\Big) \geq \sum_{k=1}^{K} \lambda_k \log(x_k) \quad \text{for} \quad \sum_{k=1}^{K} \lambda_k = 1$$

We have transformed the log of a sum into the sum of logs which is easier to derive

$$L(\boldsymbol{\theta}) \geq \sum_{\mathbf{Y}} p(\mathbf{Y} \mid \mathbf{X},\boldsymbol{\theta}^{old}) \log \frac{p(\mathbf{X},\mathbf{Y} \mid \boldsymbol{\theta})}{p(\mathbf{Y} \mid \mathbf{X},\boldsymbol{\theta}^{old})} = l(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{old})$$

Note that

$$l(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{old}) = \overbrace{\sum_{\mathbf{Y}} p(\mathbf{Y} \mid \mathbf{X},\boldsymbol{\theta}^{old}) \log p(\mathbf{X},\mathbf{Y} \mid \boldsymbol{\theta})}^{Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{old})} + const.$$

EM Algorithm

Given observations $\mathbf{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ and a model $p(\mathbf{X}\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})$,

find the parameters $\boldsymbol{\theta}_{opt}$ which maximize the likelihood (or the posterior):

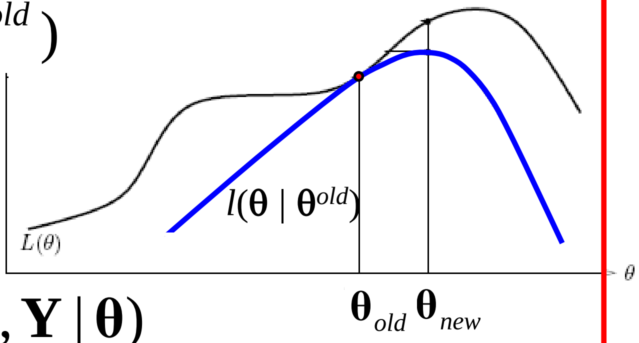
$$\boldsymbol{\theta}_{opt} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}\,|\,\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{Y}} p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})$$

1. Initialize parameters $\boldsymbol{\theta}_{opt}$ randomly

2. <u>E</u>xpectation step: evaluate

$$p(\mathbf{Y}\,|\,\mathbf{X},\boldsymbol{\theta}^{old}) = \frac{p(\mathbf{X}\,|\,\mathbf{Y},\boldsymbol{\theta}^{old})\,p(\mathbf{Y}\,|\,\boldsymbol{\theta}^{old})}{\sum_{\mathbf{Y}'} p(\mathbf{X}\,|\,\mathbf{Y}',\boldsymbol{\theta}^{old})\,p(\mathbf{Y}'\,|\,\boldsymbol{\theta}^{old})}$$

3. **<u>M</u>**aximization step: update parameters



$$\boldsymbol{\theta}^{new} = \arg\max_{\theta} \sum_{\mathbf{z}} p(\mathbf{Y}\,|\,\mathbf{X},\boldsymbol{\theta}^{old})\ \log p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\theta})$$

4. Evaluate likelihood; if not converged repeat from 2

EM Algorithm: Evaluating $l(\theta|\theta^{old})$

$$l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old}) = \sum_{\mathbf{Y}} p(\mathbf{Y}\,|\,\mathbf{X},\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})\ \log p(\mathbf{X},\mathbf{Y}\,|\,\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$= \sum_{\mathbf{Y}} p(\mathbf{Y}\,|\,\mathbf{X},\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})\ \log \prod_{n=1}^{N} \left(\pi_{y_n} N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_{y_n},\Sigma_{y_n})\right)$$

$$= \sum_{y_1,\ldots,y_N=1} p(\mathbf{Y}\,|\,\mathbf{X},\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})\ \sum_{n=1}^{N} \left(\log \pi_{y_n} + \log N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_{y_n},\Sigma_{y_n})\right)$$

$$= \sum_{n=1}^{N} \sum_{y_1,\ldots,y_N=1} p(\mathbf{Y}\,|\,\mathbf{X},\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})\ \left(\log \pi_{y_n} + \log N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_{y_n},\Sigma_{y_n})\right)$$

$$= \sum_{n=1}^{N} \sum_{k=1} p(y_n = k\,|\,\mathbf{X},\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})\ \left(\log \pi_k + \log N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)\right)$$

$$= \sum_{n=1}^{N} \sum_{k=1} r_{nk} \left(\log \pi_k + \log N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)\right)$$

EM Algorithm: Evaluating the mean

$$l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_{\mathbf{k}} + \log N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)\right) + const.$$

Maximizing wrt $\boldsymbol{\mu}$ yields

$$0 = \frac{\partial l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} r_{nk} \frac{\partial \log N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\partial \boldsymbol{\mu}_k}$$

$$0 = \sum_{n=1}^{N} r_{nk} \frac{\partial}{\partial \boldsymbol{\mu}_k}\left(-\frac{1}{2\sigma}(\mathbf{x}_n - \boldsymbol{\mu}_k)^t \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)\right) \quad = \sum_{n=1}^{N} r_{nk} \Sigma_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$0 = \sum_{n=1}^{N} r_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k) = \sum_{n=1}^{N} r_{nk}\mathbf{x}_n - \boldsymbol{\mu}_k \underbrace{\sum_{n=1}^{N} r_{nk}}_{N_k}$$

$$\boxed{\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}$$

(M-step)

EM Algorithm: Evaluating the covariance matrix

$$l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log \pi_{\mathbf{k}} + \log N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)\right) + const.$$

Maximizing wrt Σ, or rather Σ^{-1}, yields

$$0 = \frac{\partial l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old})}{\partial \Sigma_k^{-1}}$$

$$0 = \sum_{n=1}^{N} r_{nk}\left[\frac{\partial}{\partial \Sigma_k}\log|\Sigma_k^{-1}|^{1/2} + \frac{\partial}{\partial \Sigma_k^{-1}}\left(-\frac{1}{2\sigma}(\mathbf{x}_n-\boldsymbol{\mu}_k)^t\Sigma_k^{-1}(\mathbf{x}_n-\boldsymbol{\mu}_k)\right)\right]$$

$$0 = \underbrace{\left(\sum_{n=1}^{N} r_{nk}\right)}_{N_k}\Sigma_k - \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^t$$

$$\boxed{\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^t}$$

(M-step)

EM Algorithm: Evaluating the mixing coefficients

$$l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old}) = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\left(\log\pi_{\mathbf{k}} + \log N(\mathbf{x}_n\,|\,\boldsymbol{\mu}_k,\Sigma_k)\right) + const.$$

Maximizing wrt π yields

Lagrange multiplier

$$0 = \frac{\partial}{\partial\pi_k}\left(l(\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{\pi}^{old},\boldsymbol{\mu}^{old},\boldsymbol{\Sigma}^{old}) - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)\right) = \sum_{n=1}^{N} r_{nk}\frac{\partial}{\partial\pi_k}\log\pi_{\mathbf{k}} - \lambda$$

$$\Rightarrow \pi_{\mathbf{k}} = \frac{1}{\lambda}\sum_{n=1}^{N} r_{nk} = \frac{N_k}{\lambda}$$

From the normalization condition for π we obtain

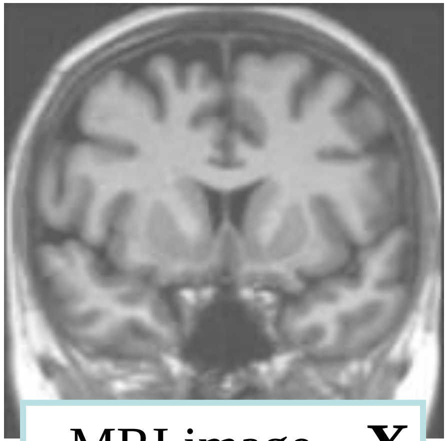
$$\boxed{\pi_{\mathbf{k}} = \frac{N_k}{N}}$$

(M-step)

Expectation Maximisation Algorithm

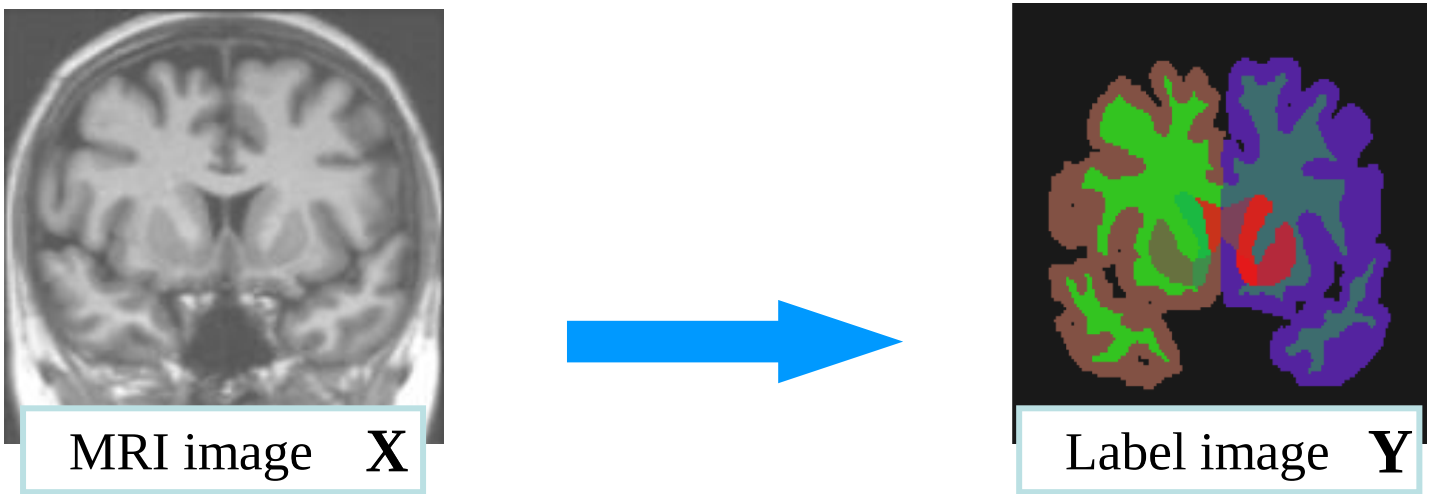
- Each iteration is guaranteed to increase the loglikelihood
- The algorithm is guaranteed to converge to a local maximum of the likelihood function.
- A modified form of the M-step is to find some $\boldsymbol{\theta}^{new}$ such that $Q(\boldsymbol{\theta}^{new} \mid \boldsymbol{\theta}^{old}) > Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{old})$

Application: Image Segmentation



MRI image $\mathbf{X}$

The problem to be solved

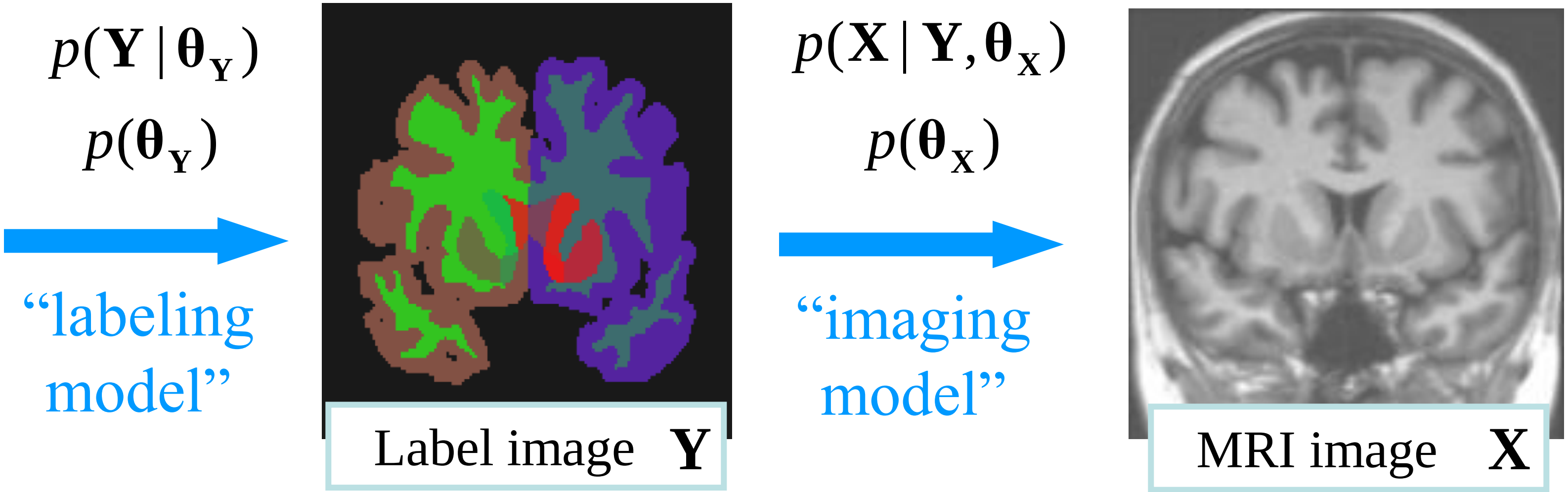


MRI image **X**

Label image **Y**

One solution: generative modeling

- Formulate a statistical model of how an MRI image is formed



$p(\mathbf{Y} | \boldsymbol{\theta}_\mathbf{Y})$

$p(\boldsymbol{\theta}_\mathbf{Y})$

“labeling model”

Label image $\mathbf{Y}$

$p(\mathbf{X} | \mathbf{Y}, \boldsymbol{\theta}_\mathbf{X})$

$p(\boldsymbol{\theta}_\mathbf{X})$

“imaging model”

MRI image $\mathbf{X}$

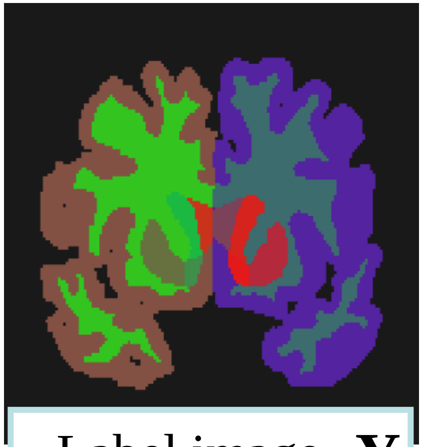
- The model depends on some parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_\mathbf{Y}, \boldsymbol{\theta}_\mathbf{X}\}$

Example: Gaussian mixture model

$p(\mathbf{Y} \mid \boldsymbol{\theta}_{\mathbf{Y}})$

$p(\boldsymbol{\theta}_{\mathbf{Y}})$

"labeling model"

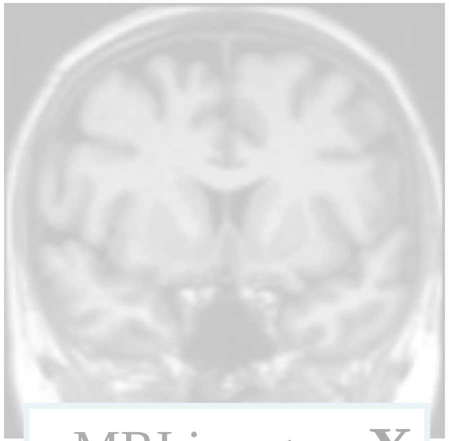


Label image $\mathbf{Y}$

$p(\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}_{\mathbf{X}})$

$p(\boldsymbol{\theta}_{\mathbf{X}})$

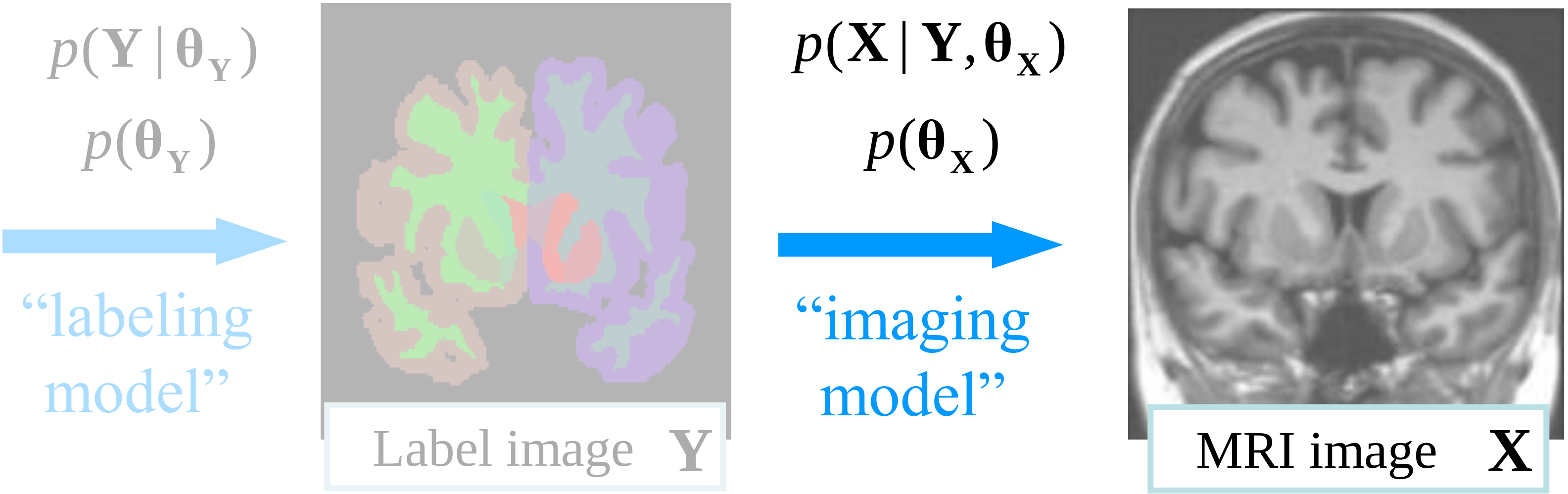
"imaging model"



MRI image $\mathbf{X}$

- The label in each voxel is drawn independently with a probability π_{k} for tissue type k
- Assume a uniform prior $p(\boldsymbol{\theta}_{\mathbf{Y}})$ for the labeling model parameters $\boldsymbol{\theta}_{\mathbf{Y}} = \{\pi_{\mathrm{k}}\}$

Example: Gaussian mixture model



$p(\mathbf{Y} \mid \boldsymbol{\theta}_{\mathrm{Y}})$

$p(\boldsymbol{\theta}_{\mathrm{Y}})$

“labeling model”

Label image $\mathbf{Y}$

$p(\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}_{\mathrm{X}})$

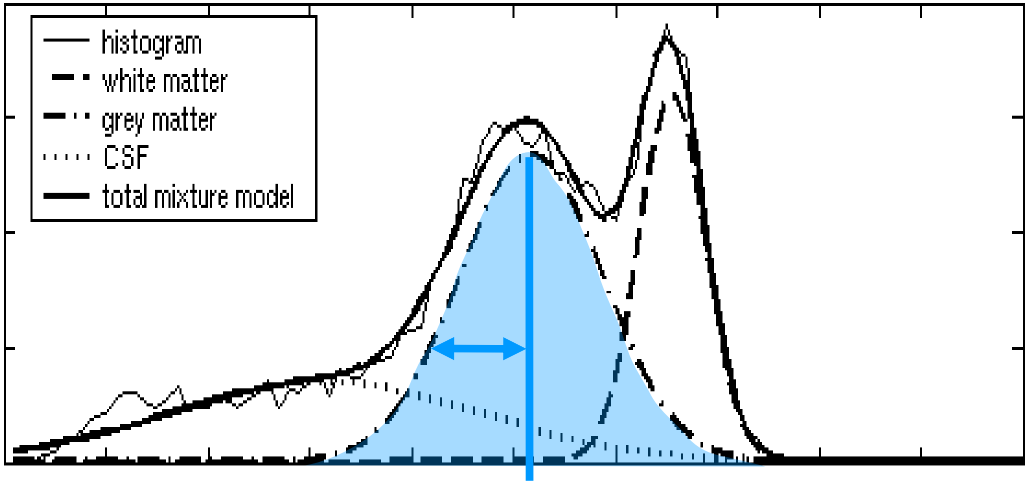
$p(\boldsymbol{\theta}_{\mathrm{X}})$

“imaging model”

MRI image $\mathbf{X}$

- The intensity in each voxel is drawn independently from a Gaussian distribution associated with its label
- The imaging model parameters are the mean μ_k and variance Σ_k of each Gaussian: $\boldsymbol{\theta}_{\mathrm{X}} = \{\mu_k, \Sigma_k\}$
- Assume a uniform prior $p(\boldsymbol{\theta}_{\mathrm{X}})$

Example: Gaussian mixture model



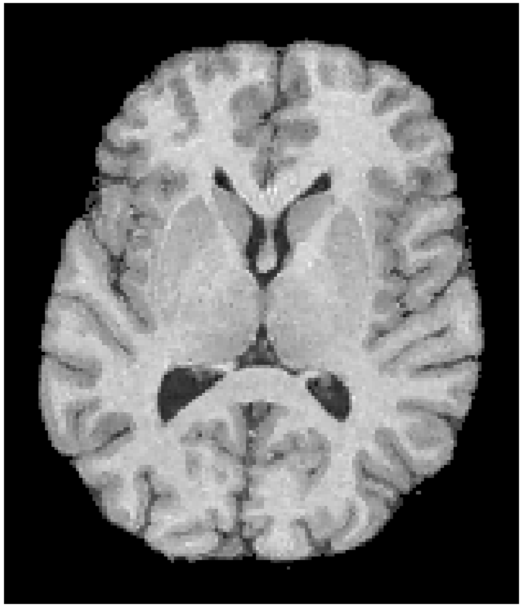
three labels

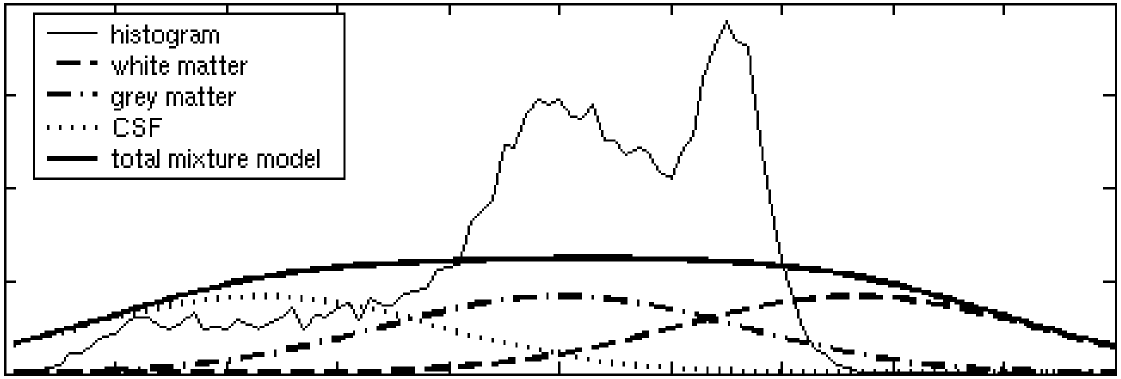
Model parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{\mathbf{Y}}, \boldsymbol{\theta}_{\mathbf{X}}\}$ are unknown

Relative weight of each Gaussian

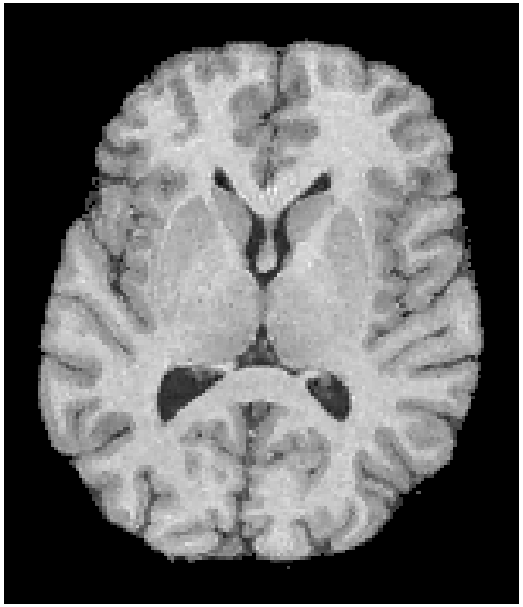
Mean and variance of each Gaussian

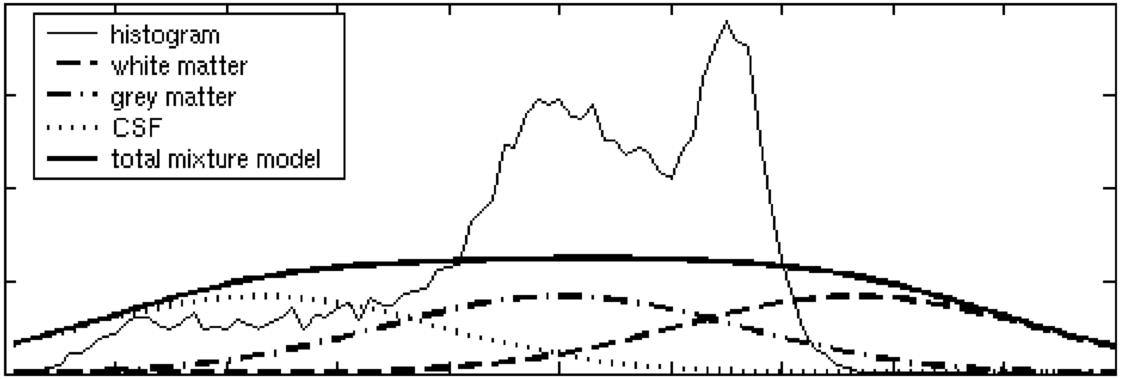
Optimization 1: parameter estimation



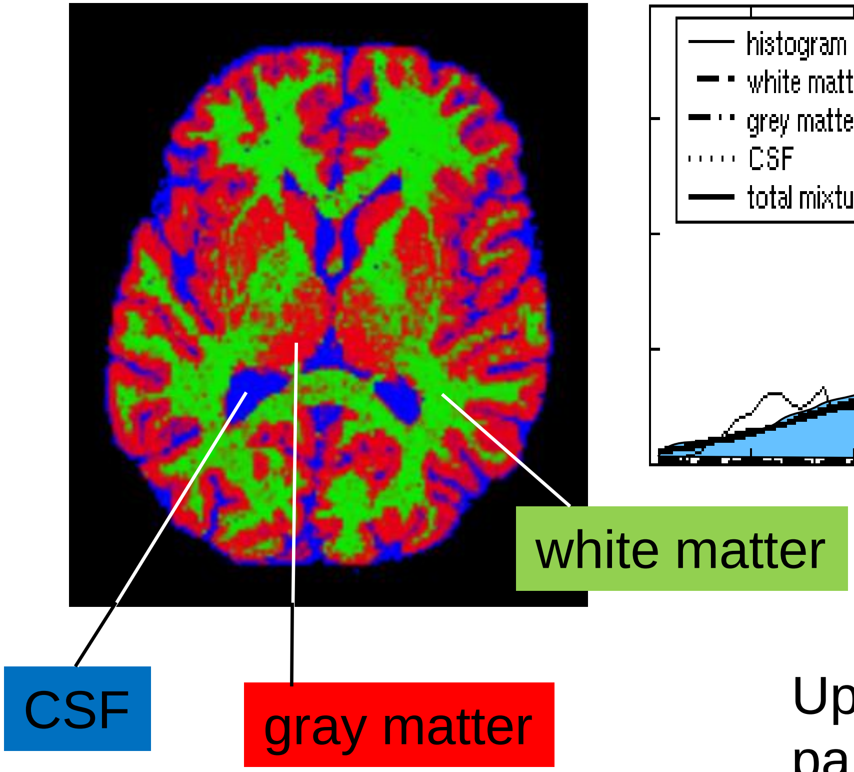


Optimization 1: parameter estimation





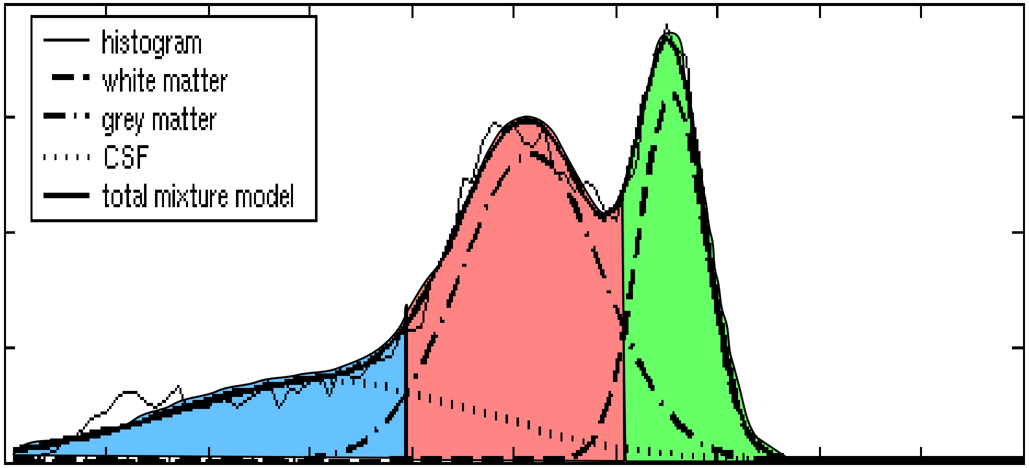
Optimization 2: segmentation



white matter

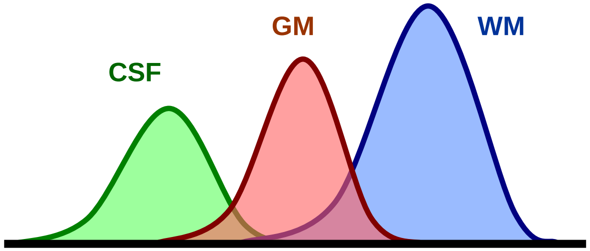
CSF

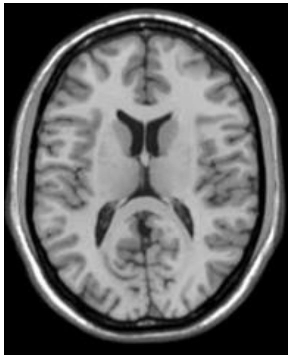
gray matter



Upon completion of the parameter estimation algorithm, assign each voxel to the Maximum a posteriori label

Expectation Step



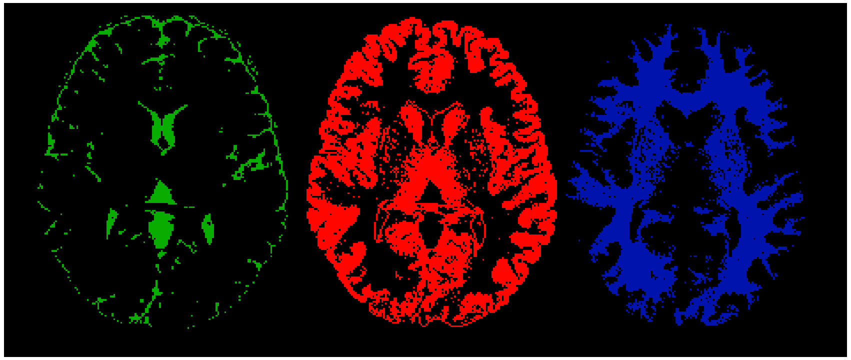


Computing the responsibilities based on the current parameter estimation

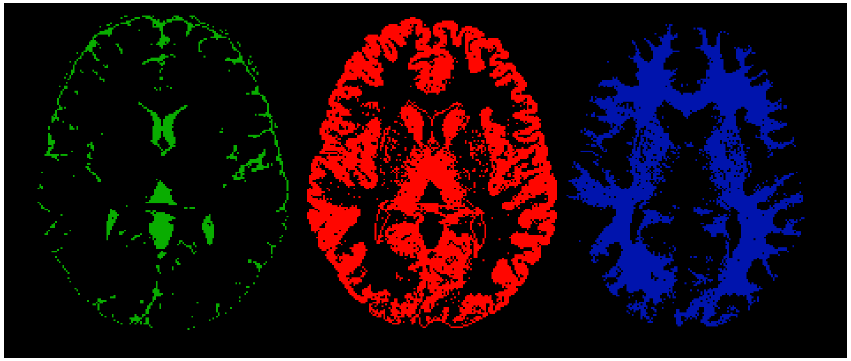
$$r_{nk} = \frac{\pi_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \Sigma_k)}{\sum_l \pi_l N(\mathbf{x}_n \mid \boldsymbol{\mu}_l, \Sigma_l)}$$

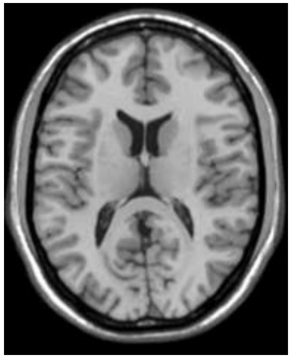
$$N_k = \sum_{n=1}^{N} r_{nk}$$

$$\pi_k^{new} = \frac{N}{N_k}$$

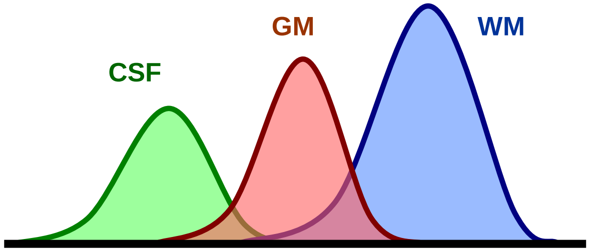


Maximisation Step





Parameter reestimation based on the current responsabilities



$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}\mathbf{x}_n$$

$$\Sigma_k^{new} = \frac{1}{N_k}\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n-\boldsymbol{\mu}_k^{new})(\mathbf{x}_n-\boldsymbol{\mu}_k^{new})^t$$