3. Medical Image Segmentation

- 3.1 Taxonomy of segmentation algorithms
- 3.2 Validation of segmentation algorithms
- 3.3 Deterministic Filtering & Thresholding Approaches
- 3.4 Probabilistic Imaging Model
- 3.5 **Expectation Maximisation for GMM**
- 3.6 Image classification with bias field
- 3.7 Variational Bayes EM
- 3.8 STAPLE Algorithm

Expectation Maximisation Algorithm

- Iterative approach for estimating parameters of (Gaussian) Mixture parameters
- General Idea :
  - New criterion : Add unknown variable u (posterior) and add constraint (KL divergence)
  - Alternate maximization performed in closed form : equivalent to lower bound maximization

Alternate maximisation

- Replace Log-Likelihood with a criterion easier to optimize but with additional unknowns
- Log-(marginal) likelihood :

$$L(\theta) = \log \Lambda(\theta) = \sum_n \log p(x_n|\theta) = \sum_n \log(\sum_k \pi_k \mathcal{N}(x_n; \mu_k, \sigma_k))$$

- New criterion F(θ, u):
  - Add $u = \{u_{nk}\}$ as unknown. u is a vector of u_{nk} which is the posterior probability

$$\mathrm{F}(\theta, u) = \mathrm{L}(\theta) - D_{KL}\big(u||p(z|x)\big)$$

- By maximizing F with respect to u,

$$u_{nk} = p(z_{nk} = 1|x_n)$$

Why is it easier to optimize $\mathrm{F}(\theta, u)$?

- General result :
  - X = observed random variable
  - Z = hidden random variable
  - Joint probability $p(x_n, z_n) = p(x_n|z_n)p(z_n) = p(z_n|x_n)p(x_n)$
  - Constraint on u_{nk}: $\sum_k u_{nk} = 1$
  - Log likelihood : $L(\theta) = \sum_n \log p(x_n) = \sum_n \sum_k u_{nk} \log p(x_n)$
  - New criterion :

$$\mathrm{F}(\theta, u) = \sum_n \sum_k u_{nk} \log p(x_n) - \sum_n \sum_k u_{nk} \log u_{nk}/p(z_{nk}|x_n)$$

$$\boxed{\mathrm{F}(\theta, u) = \sum_n \sum_k u_{nk} \log p(x_n, z_{nk}) - \sum_n \sum_k u_{nk} \log u_{nk}}$$

We have « relaxed » the optimization problem by introducing
a new unknown variable

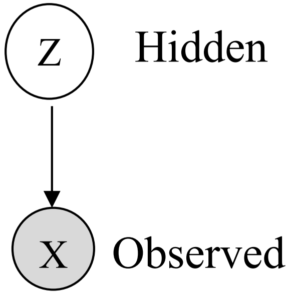
Interpretation

- New criterion involves 2 terms :

$$\mathrm{F}(\theta,u) = \underbrace{\sum_n \sum_k u_{nk} \log p(x_{nk}, z_{nk})}_{Q(\theta,u)} - \underbrace{\sum_n \sum_k u_{nk} \log u_{nk}}_{\mathbb{H}(u)}$$

  - F(θ, u) is the *variational lower bound*
  - -F(θ, u) is the *variational free energy*= average energy - entropy
  - $Q(\theta, u) = \sum_n \sum_k u_{nk} \log p(x_{nk}, z_{nk}) = \mathbb{E}_U(\log p(X, Z))$ is the expectation of the complete likelihood
  - $\mathbb{H}(u) = -\sum_n \sum_k u_{nk} \log u_{nk}$ is the **entropy** of the approximate posterior probability
- $Q(\theta, u)$ is easier to optimize wrt θ because it involves complete likelihood = likelihood of observed and hidden variables

Evidence Lower Bound



- General result :
  - For any inverse problem where Z is the hidden variable and X observed variable :

$$\boxed{\begin{aligned}&\log p(X) - D_{KL}\big(u \| p(Z|X)\big)\\ &= \mathbb{E}_u(\log p(X,Z)) + \mathbb{H}(u)\end{aligned}}$$

- Variational lower bound :

$$\log \mathrm{p}(X) \geq \mathbb{E}_u(\log p(X,Z)) + \mathbb{H}(u)$$

Case of Gaussian Mixtures

- Log likelihood

$L(\theta) = \log \Lambda(\theta) = \sum_n \log(\sum_k \pi_k \mathcal{N}(x_n; \mu_k, \sigma_k))$

- Function of parameters :

$$Q(\theta, u) = \sum_n \sum_k u_{nk} \log \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)$$

- Note that we have sum of log instead of log of sums !
- Criterion F$(\theta, u) = \mathrm{Q}(\theta, u) + \mathbb{H}(u)$is known as **Hathaway criterion**

EM Algorithm

- The algorithm optimizes alternatively between u and θ = coordinate ascent

$$\mathrm{F}(\theta,u) = \mathrm{L}(\theta) - D_{KL}\big(u||p(z|x)\big) = \mathrm{Q}(\theta,u) + \mathbb{H}(u)$$

- Constraints : $\sum_k \pi_k = 1 \quad \sum_k u_{nk} = 1$
- E-step
  - maximize $F(\theta,u)$ wrt u

$$\text{Compute } \boxed{\mathrm{u_{nk}} = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}}$$

  - Equivalent to minimizing KL divergence between u and posterior probability

M-Step

- M-step : maximize $F(\theta, u)$ or equivalently $Q(\theta, u)$ wrt $\theta = \{\theta_S, \theta_I\}$
  - Optimize with respect to mean μ_k

$$\frac{\partial Q}{\partial \mu_k} = 0 \Rightarrow \boxed{\mu_k = \frac{\sum_{n=1}^{N} u_{nk}\, x_n}{\sum_{n=1}^{N} u_{nk}}}$$

  - Optimize with respect to covariance Σ_k

$$\frac{\partial Q}{\partial \Sigma_k} = 0 \Rightarrow \boxed{\Sigma_k = \frac{\sum_{n=1}^{N} u_{nk}\,(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{i=1}^{N} u_{nk}}}$$

  - Optimize with respect to prior probabilities

$$\frac{\partial Q}{\partial \pi_k} = 0 \Rightarrow \boxed{\pi_k = \frac{1}{N}\sum_{n=1}^{N} u_{nk}}$$

EM Algorithm for GMM

- Iterative scheme
  - Make initial guesses for the parameters
  - Alternate between the following two stages:

1. E-step: evaluate posterior u_{nk}

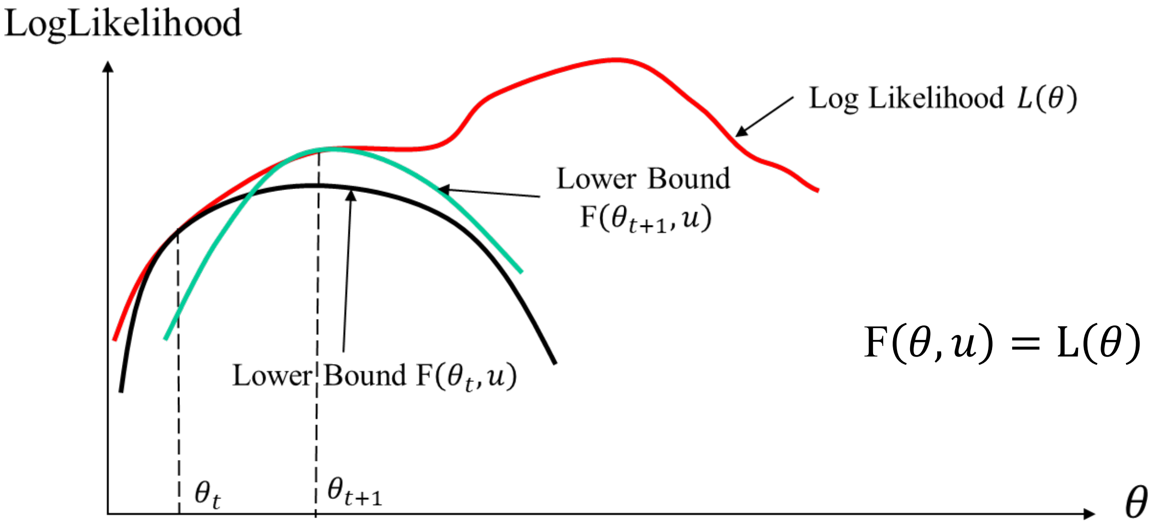
$$\mathrm{u}_{\mathrm{nk}} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_n|\mu_k, \Sigma_k)}$$

2. M-step: update parameters (μ_k, Σ_k, π_k) using ML results

$$\mu_k = \frac{\sum_{n=1}^{N} u_{nk} \; x_n}{\sum_{n=1}^{N} u_{nk}} \qquad \Sigma_k = \frac{\sum_{n=1}^{N} u_{nk} \, (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{i=1}^{N} u_{nk}} \qquad \pi_k = \frac{1}{N} \sum_{n=1}^{N} u_{nk}$$

EM as Iterated Lower Bound Maximisation

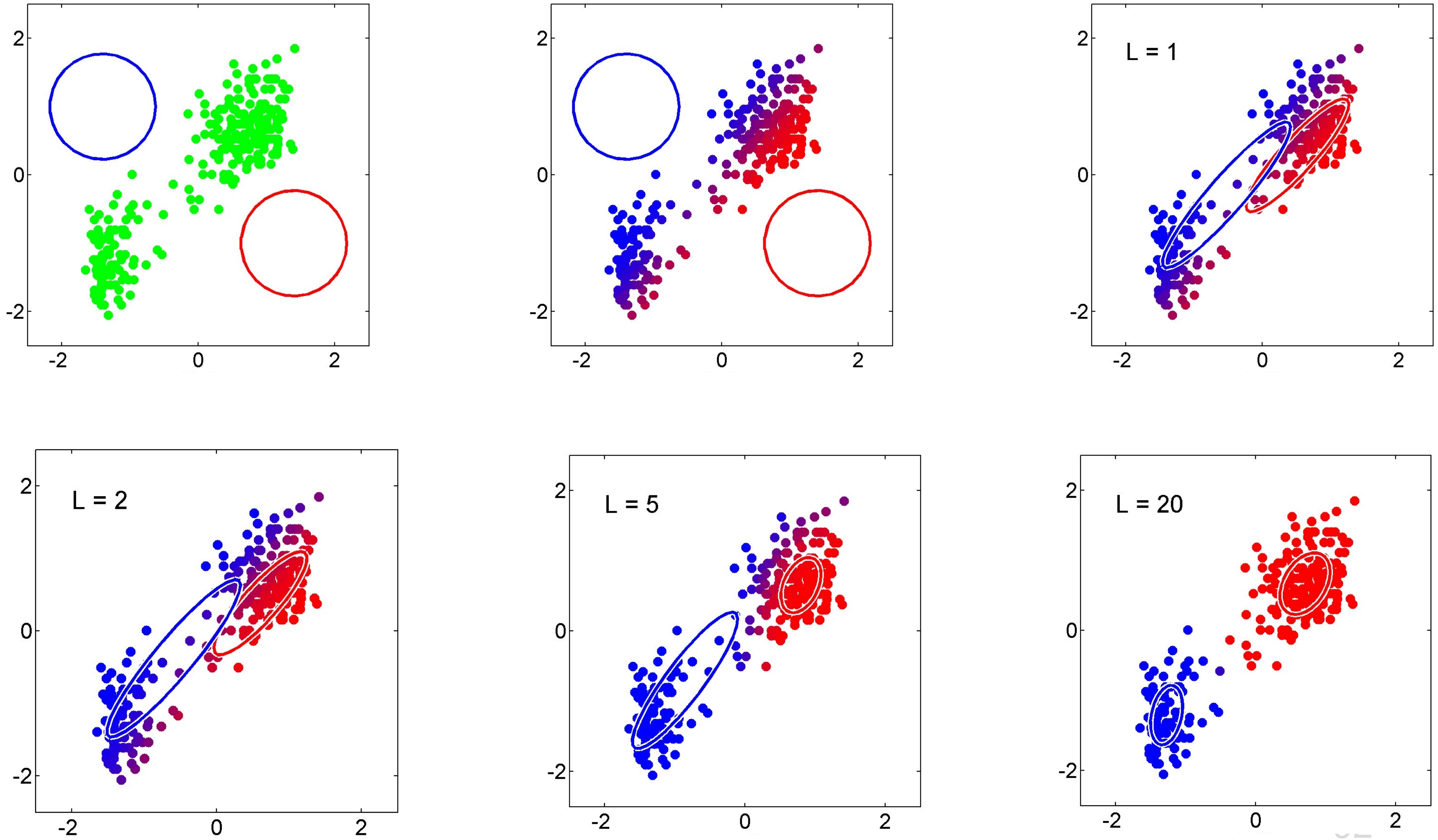
- Equivalent view of EM algorithm :
  - E-step leads to $u = p(z|x)$ and therefore makes $L(\theta_t) = F(\theta_t, u)$.
  - $F(\theta, u)$ is a lower bound of Log-likelihood $L(\theta)$ since Kullback Leibler divergence is positive
  - M-step optimizes $F(\theta, u)$ with respect to θ which is easier to maximize than log likelihood



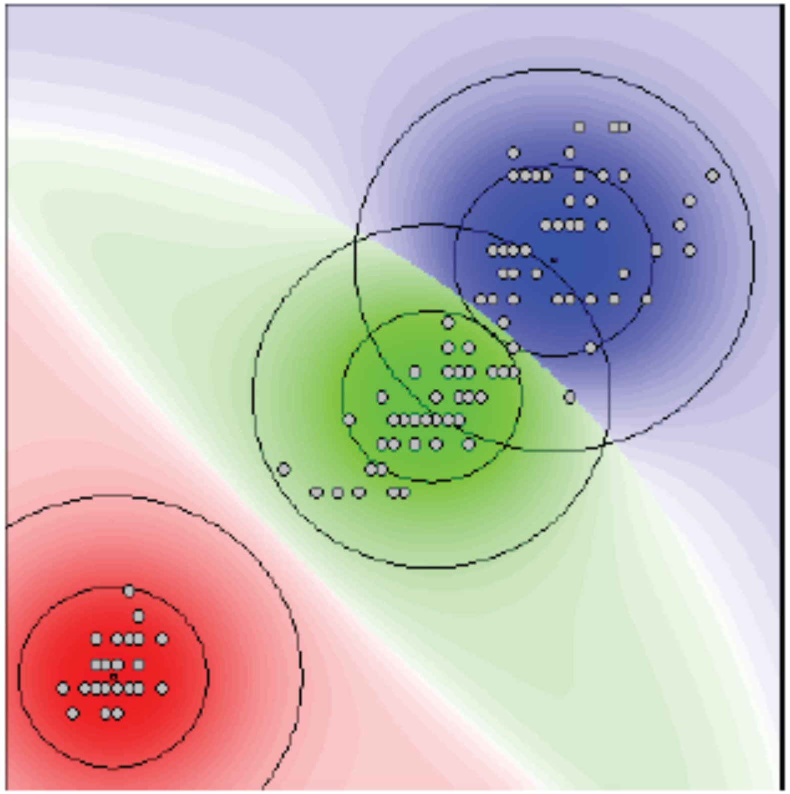
$$L(\theta) \geq F(\theta, u)$$

$$F(\theta, u) = L(\theta) - D_{KL}\big(u || p(z|x)\big) = Q(\theta, u) + \mathbb{H}(u)$$

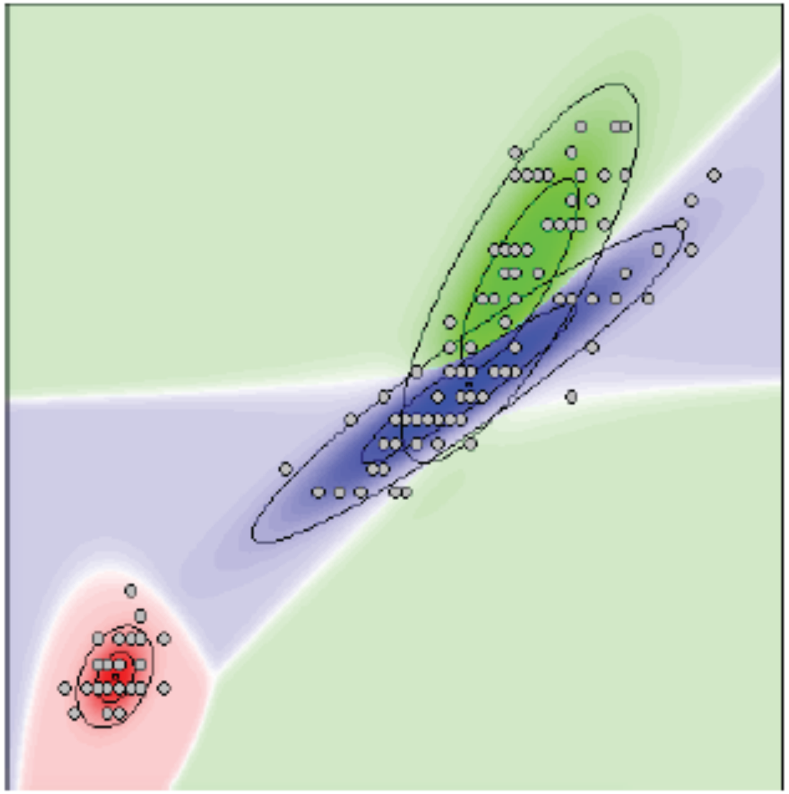
Example of EM with 2 Gaussian distributions



EM on Iris data



equal prior, spherical

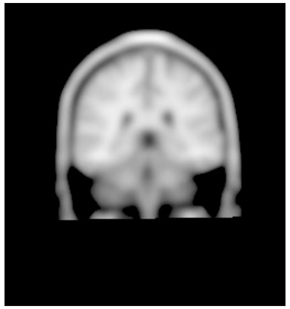


equal prior, ellipsoidal

Class Priors

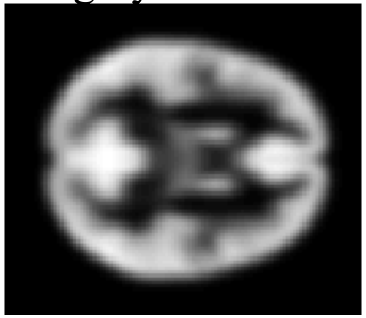
- Initial hypothesis : homogeneous priors $p(z_{nk} = 1) = \pi_k$ is estimated
- Priors may be given by atlas registered on images. In this case θ_S are the registration parameters

Atlas



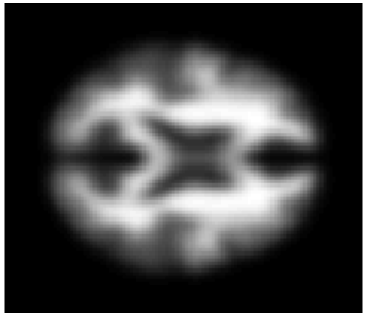
T1 template

Prior $p(z_{n1})$ on grey matter



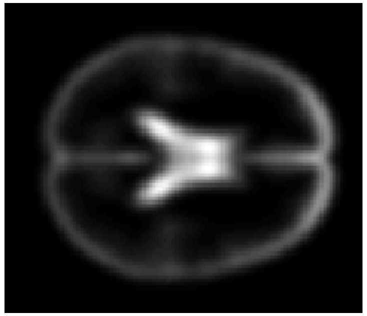
gray matter

Prior $p(z_{n2})$ on White matter



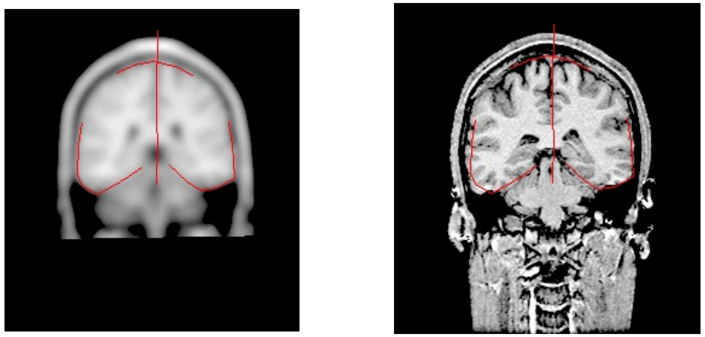
white matter

Prior $p(z_{n3})$ on cerebro spinal fluid



csf

Affinely Registered Atlas



Courtesy of D. Vandermeulen

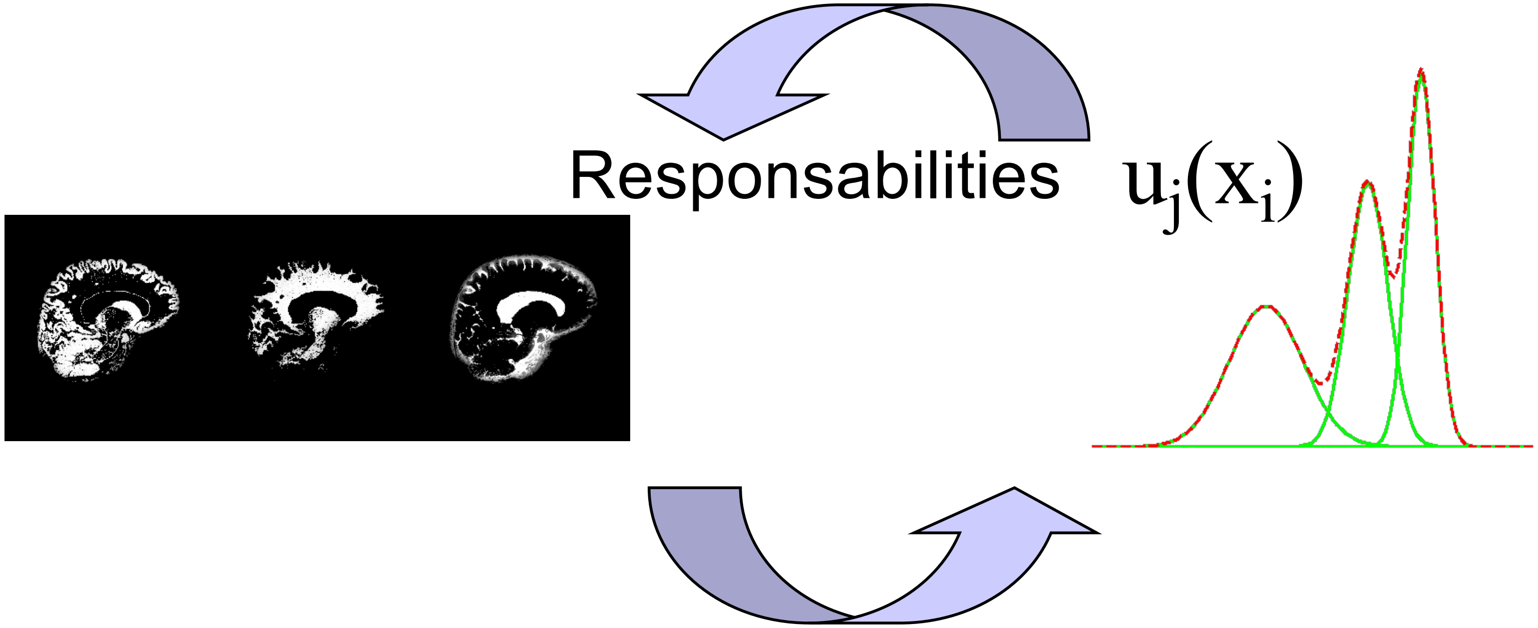
Example : BrainWeb at MNi

http://www.bic.mni.mcgill.ca/brainweb/

EM for Image Intensity Classification

- Use the EM algorithm [Dempster77,Wells94] :

Expectation-Maximisation

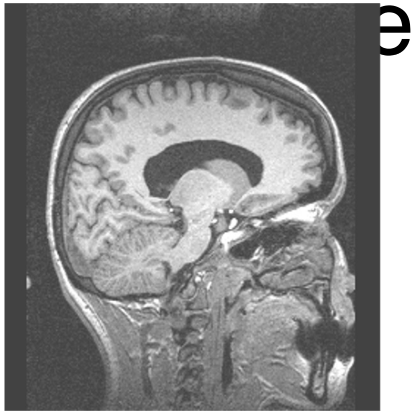


Responsabilities $u_j(x_i)$

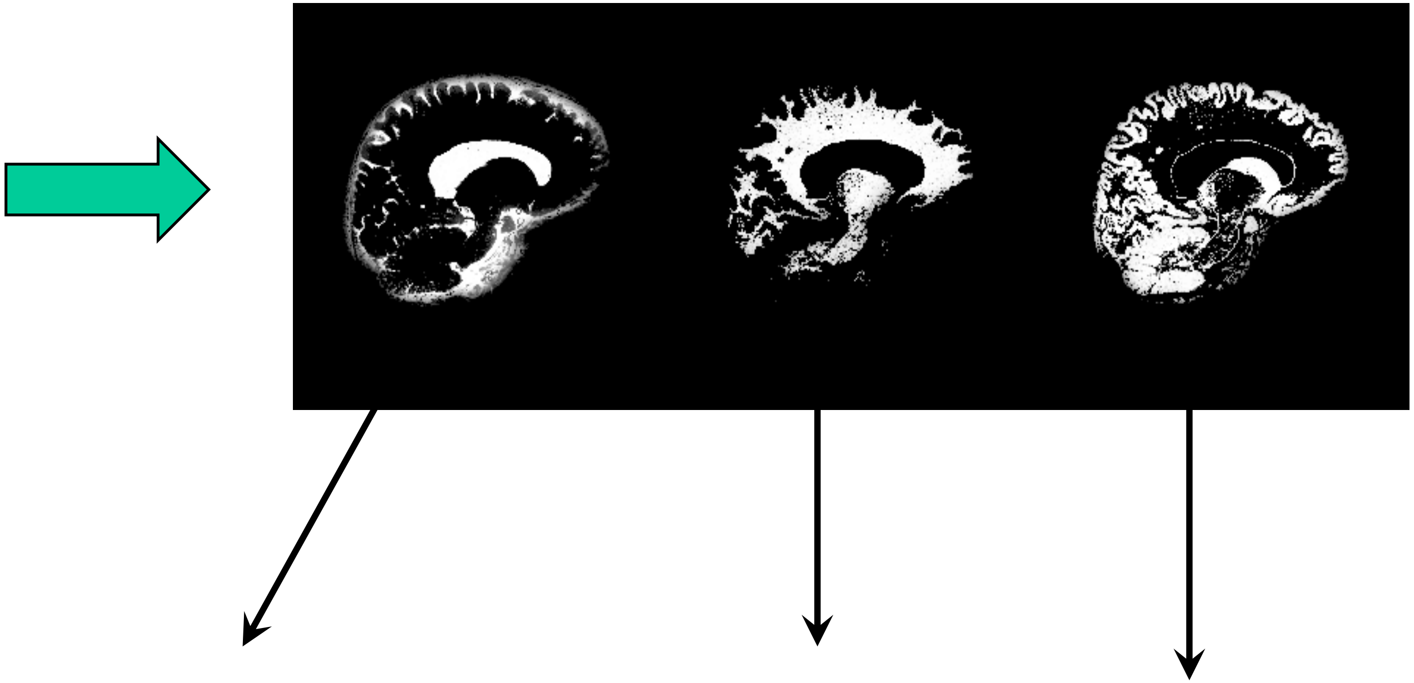
Mixture Param. μ_κ Σ_κ π_κ

Brain Tissue Classification

- Typical application : use MR cerebral i e



Courtesy of D. Vandermeulen

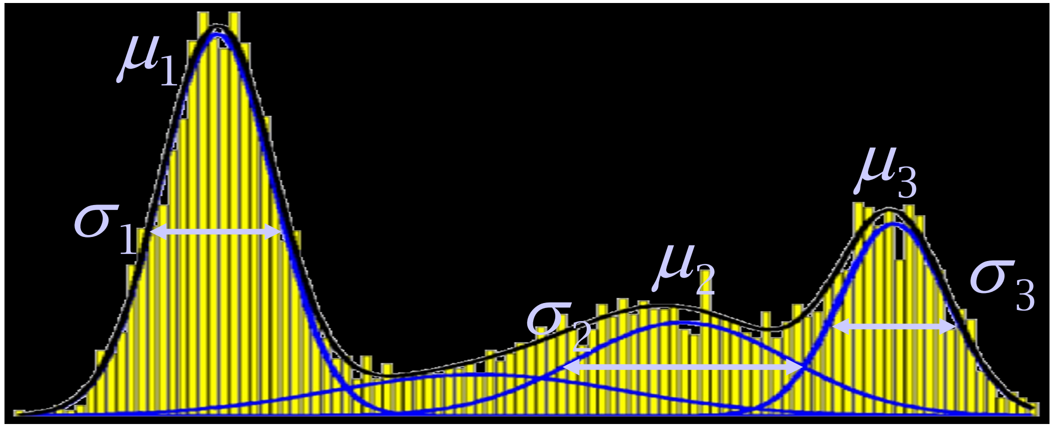


Cerebro-spinal fluid Grey matter White matter

3 Classes

Scalar feature
= Intensity

EM Classification - Algorithm



Starting estimates for responsibilities

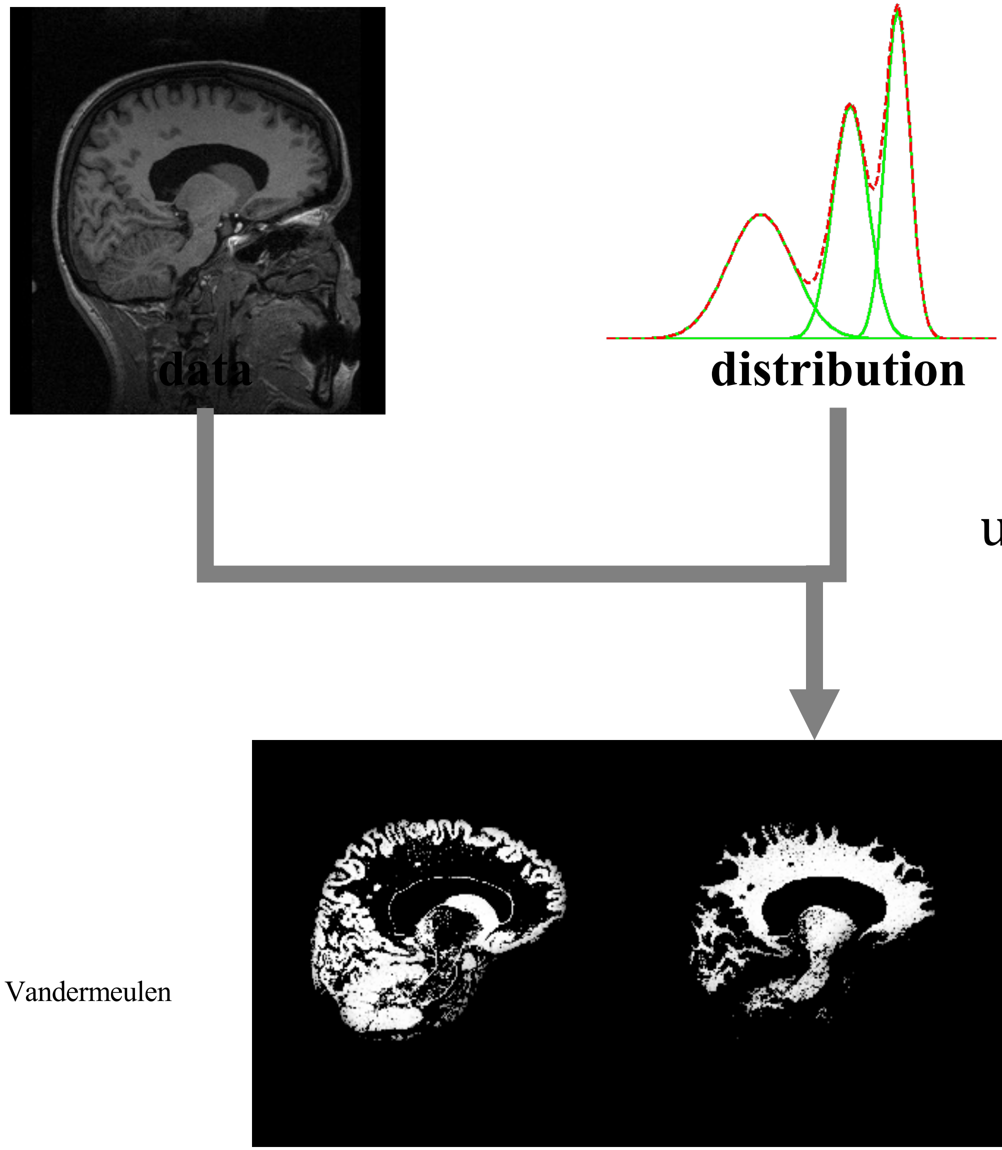
Compute Gaussian parameters from Responsibilities

Compute Responsibilities from Gaussian parameters

Converged ?
No Yes

STOP

Stage 1: Expectation

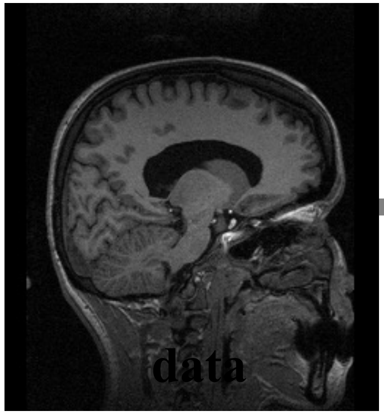


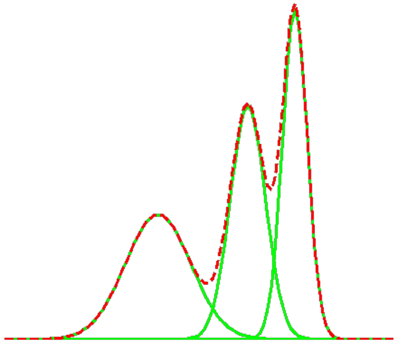
Compute
Responsibilities

$$u_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{k=1}^{K} \pi_k \, \mathcal{N}(x_n|\mu_k,\Sigma_k)}$$

Courtesy of D. Vandermeulen

Stage 2: Maximization





distribution

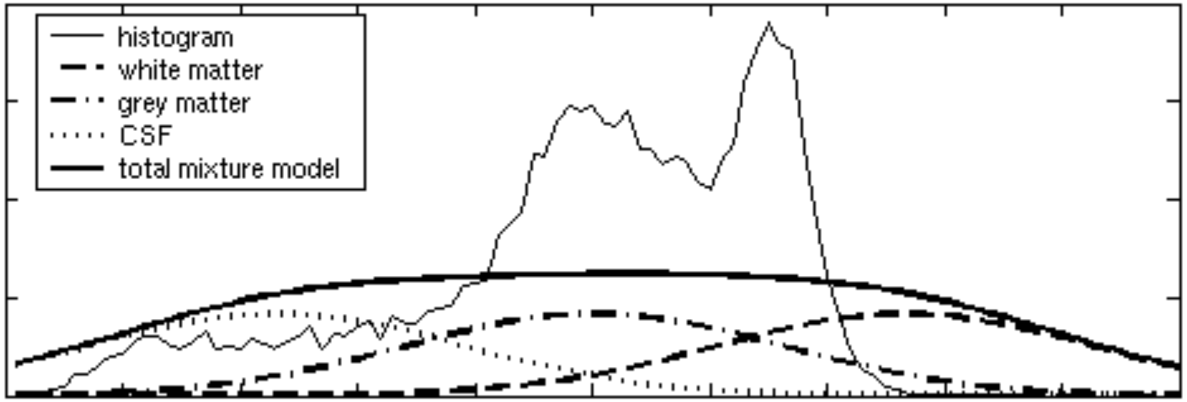
$$\mu_k = \frac{\sum_{n=1}^{N} u_{nk} \; x_n}{\sum_{n=1}^{N} u_{nk}}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} u_{nk} \, (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{i=1}^{N} u_{nk}}$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} u_{nk}$$

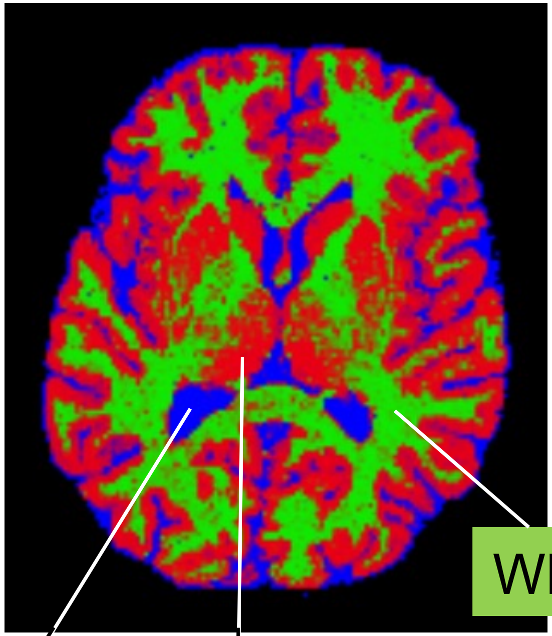
Courtesy of D. Vandermeulen

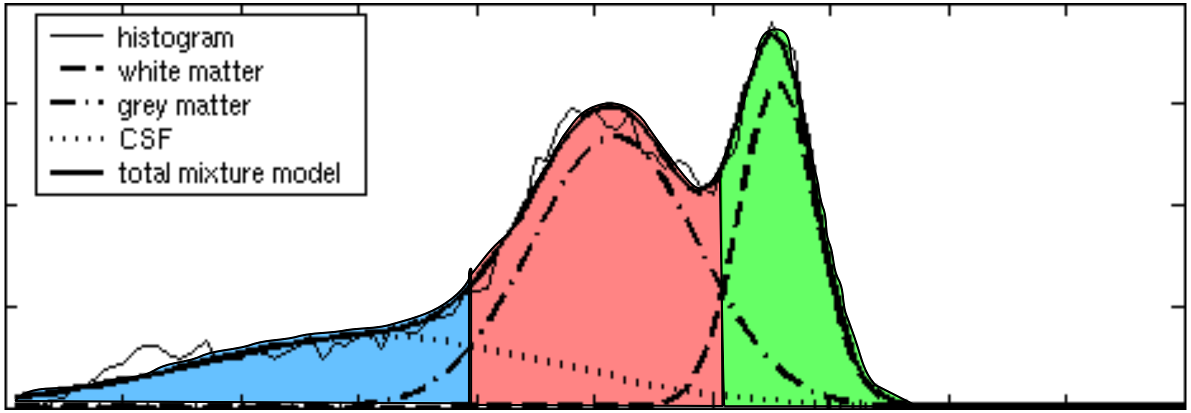
Iterations EM



Courtesy of K. Van Leemput

Results





White matter

CSF

Grey Matter

$$\hat{\mathbf{l}} = \arg\max_{\mathbf{l}} p(\mathbf{l}|\mathbf{d}, \hat{\boldsymbol{\theta}})$$

Courtesy of K. Van Leemput

GMM and K-Means

- GMM with :
  - Isotropic variance $\Sigma_k = \epsilon\, Id$
  - Uniform prior : $\pi_k = \frac{1}{K}$
- Expectation of complete Lik. : $Q(\theta) = -\sum_n \sum_k \frac{u_{nk}|x_n - \mu_k|^2}{2\epsilon}$
- Same as Fuzzy-Cmeans with m=1
- Same as K-means when :
  - $\epsilon \to 0$
  - $u_{nk} \in \{0,1\}$

$$\mathrm{u_{nk}} = \frac{\exp\left(-|x_n - \mu_k||^2/2\epsilon\right)}{\sum_{j=1}^{K} \exp\left(-||x_n - \mu_j||^2/2\epsilon\right)} \to r_{nk} \in \{0,1\}$$

K Means functional

- K Means algorithm consists in optimizing the functional :
  - $\mathrm{J}(r,\mu) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \, ||x_n - \mu_k||^2$
  - With the constraint that $r_{nk} \in \{0,1\}$ and $\sum_{k=1}^{K} r_{nk} = 1 \;\; \forall n$
- J can be seen as
  - minimizing the correlation between the assignment and the distance to cluster center
  - Minimizing the compactness of the clusters

K Means optimization

- Perform alternate optimization :
  - Consider μ_k fixed and optimize on r_{nk}
    - For each data x_n choose which r_{nk} is 1

E-Step

$$r_{nk} = \begin{cases} 1 \text{ if } \mathrm{k} = \arg min_j \|x_n - \mu_k\| \\ 0 \text{ otherwise} \end{cases}$$

  - Consider r_{nk} fixed and optimize on μ_k

M-Step

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{n=1}^{N} r_{nk}(\mu_k - x_n) = 0$$

$$\mu_k = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_{nk}}$$

Good Initial Seeds (kmeans++)

- Choose the centers as far away as possible from each other but in a random manner.
- Algorithm :
  - Choose one center at random μ_1
  - While $k \leq K$
    - Compute $\mathrm{d_n} = \arg min_{j<k} \|x_n - \mu_j\|^2$ the minimum distance of data x_n to the already chosen centers
    - Pick μ_k among data with probability proportional to d_n
    - k=k++

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. “Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms“, 2007 , pp. 1027–1035

Issues with EM for GMM

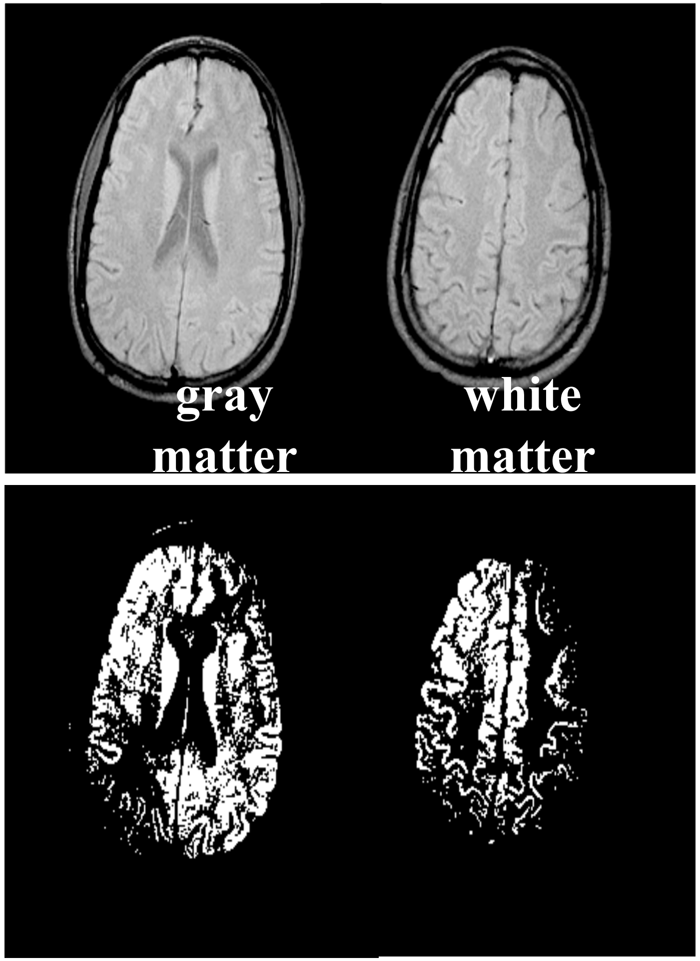
- Presence of bias field in MR images
- EM leads to only local maxima of Log-likelihood
- Functional admits trivial solutions (zero covariance centered at data points) that can lead to bad estimate
- The covariance matrix Σ_k should be invertible which is not guaranteed (may use pseudo-inverse)
- How to choose the number of classes
- How to make the estimation robust to outliers ?

3. Medical Image Segmentation

- 3.1 Taxonomy of segmentation algorithms
- 3.2 Validation of segmentation algorithms
- 3.3 Deterministic Filtering & Thresholding Approaches
- 3.4 Probabilistic Imaging Model
- 3.5 Expectation Maximisation for GMM
- 3.6 **Image classification with bias field**
- 3.7 Variational Bayes EM
- 3.8 STAPLE Algorithm

Issue : presence of MR bias field

MR images are corrupted by a smooth intensity non-uniformity (bias).



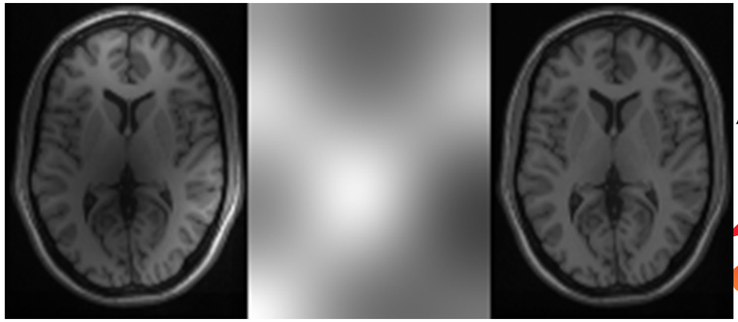


Image with bias artefact

Corrected image

MR bias field estimation methods

- homogeneous (water) phantom measurements
  - non-retrospective
  - bias is patient-dependent
  - same MRI parameter settings for patient and phantom
- analytic correction of antenna receptor profile (idem)
- bias field estimation helps segmentation => segmentation helps bias field estimation?
  - Dawant et al: manual selection (or 1 iter.) of WM points + LSQ spline fit
  - Meyer et al.: region based, too many degrees of freedom
  - Wells et al.: EM-based estimation of bias+classification, requires pre-set MRI intensity model-parameters

Biased Gaussian Mixture model

- Bias field is modeled as a additive or multiplicative Noise
- Bias field parameters with smooth linear combination of smooth basis functions

C_k are the parameters controlling the bias field

$$b(x) = \sum_{l=1}^{M} C_l \phi_l\big(r(x)\big)$$

r(x) is the position of voxel of intensity x

- Convenient choice : additive noise (but not realistic)

$$p(x) = \sum_{k} \pi_k \mathcal{N}\left(x - \sum_{l=1}^{M} C_l \phi_l\big(r(x)\big) \,|\mu_k, \sigma_k\right)$$

- Use Log of image to cope with additive noise

Bias Description

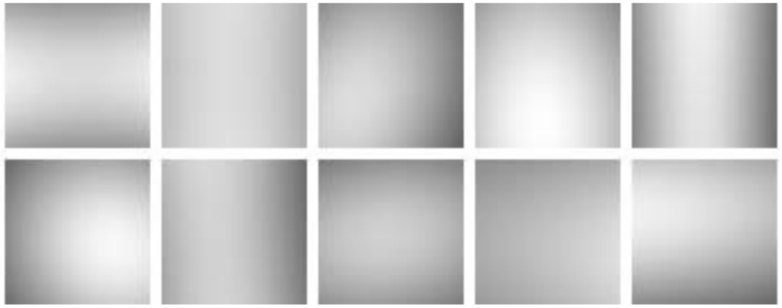
- Bias is described as a combination of slowly varying polynomials :

$$b(x) = \sum_{l=1}^{M} C_l \phi_l\big(r(x)\big)$$

r(x) is the position of voxel x

C_l are the parameters controlling the bias field

- ϕ_l(r(x)) is a slowly varying polynomial :
  - For instance :
    - ϕ_0(x,y,z)=1
    - ϕ_1(x,y,z)= x-tx/2
    - ϕ_2(x,y,z)=(y-ty/2)
    - ϕ_3(x,y,z)=(z-tz/2)
    - ϕ_4(x,y,z)=(x-tx/2)*(y-ty/2)
    - ……



Bias field with brain mask

Bias Description

- Bias field is modeled as an additive or multiplicative noise

$$b(x) = \sum_{l=1}^{M} C_l \phi_l\big(r(x)\big)$$

  - Multiplicative Noise :

☹ No closed form solution

$$p(x|\theta) = \sum_{k} \pi_k \mathcal{N}\left(x * \left(\sum_{l=1}^{M} C_l \phi_l\big(r(x)\big) \right) |\mu_k, \sigma_k \right)$$

  - Additive Noise using the Log of the image intensity:

☺ closed form solution

$$p(x|\theta) = \sum_{k} \pi_k \mathcal{N}\left(x - \sum_{l=1}^{M} C_l \phi_l\big(r(x)\big) \,|\mu_k, \sigma_k \right)$$

Extended EM Algorithm

- Include new parameters
  - Bias description C={C$_l$}
- Define Extended $Q(u,\theta) = \mathbb{E}_U(\log p(X,Z))$

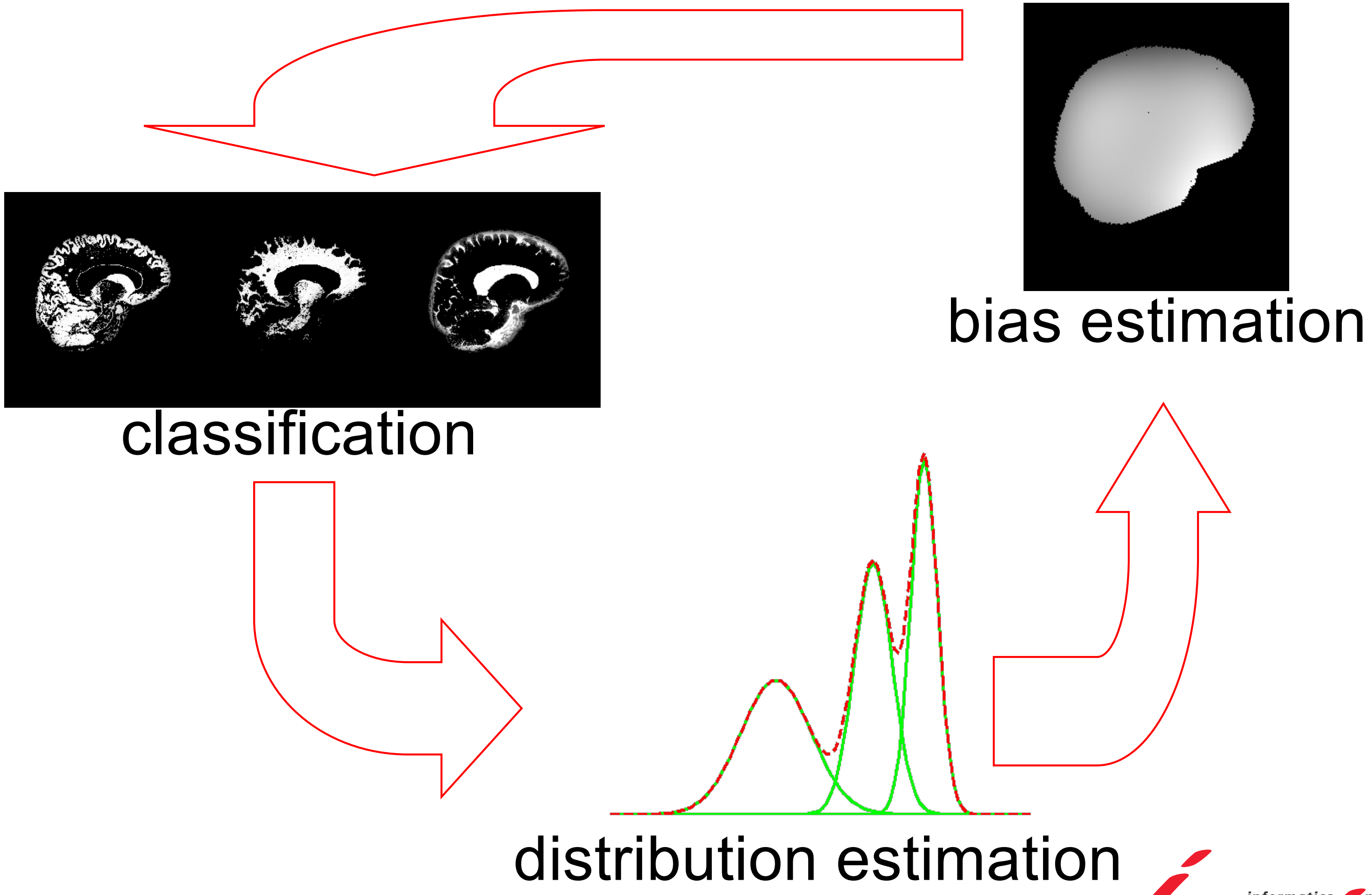
$$Q(u,\theta) = \sum_n \sum_k \log\left(u_{nk}\mathcal{N}\left(x_n - \sum_{l=1}^{M} C_l \phi_l(r_n) \,|\mu_k, \Sigma_k\right)\right)$$

r_n= position of voxel n

- Define new posterior probabilities :

$$u_{nk} = p(z_{nk} = 1|\theta, C) = \frac{\pi_k \mathcal{N}\left(x_n - \sum_{l=1}^{M} C_l \phi_l(r_n) \,|\mu_k, \Sigma_k\right)}{\sum_j^K \pi_j \mathcal{N}\left(x_n - \sum_{l=1}^{M} C_l \phi_l(r_n) \,|\mu_j, \Sigma_j\right)}$$

Extended EM algorithm

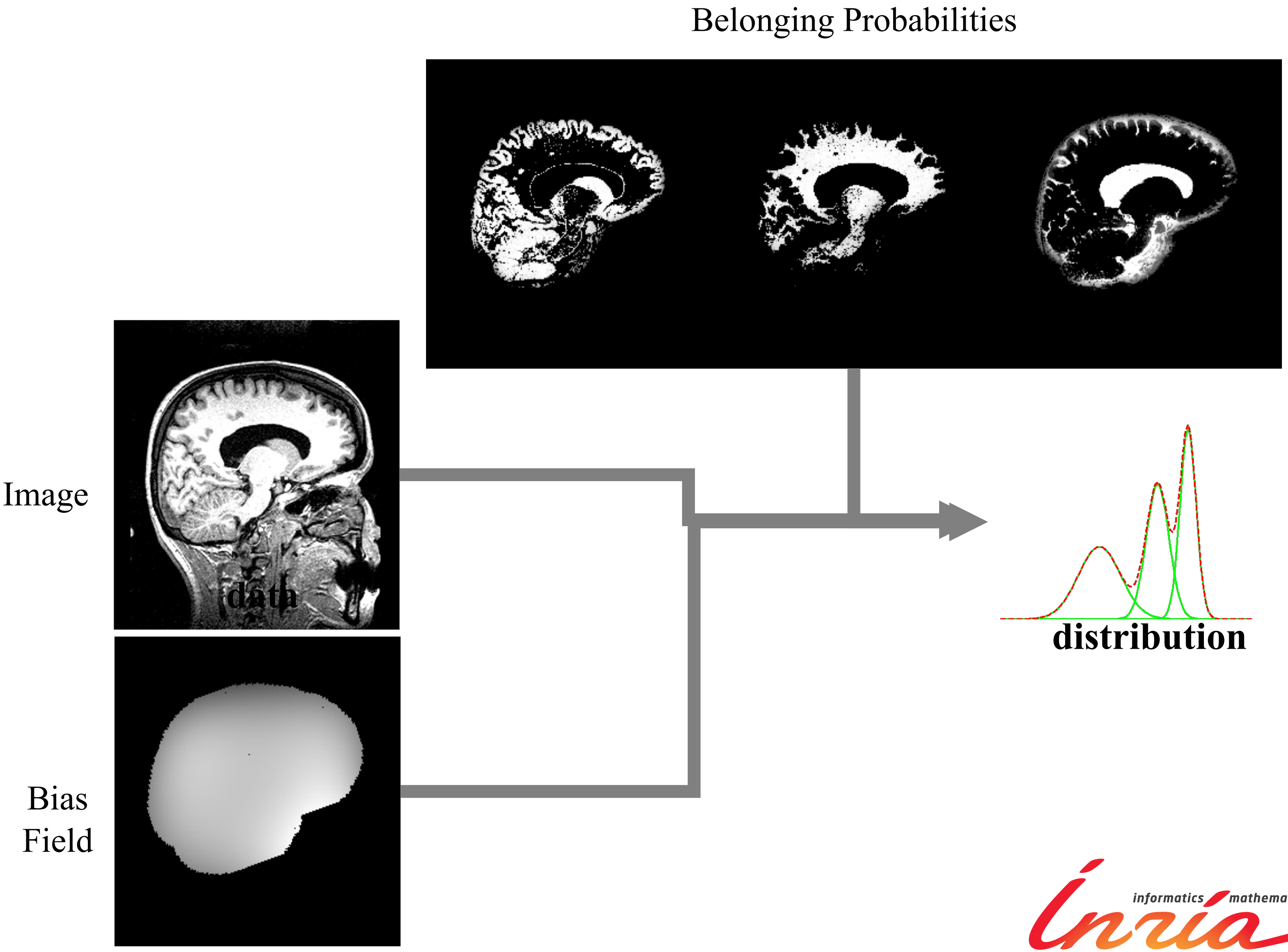


bias estimation

classification

distribution estimation

Step 1: distribution estimation



Belonging Probabilities

Image

data

Bias Field

distribution

Step1: distribution estimation

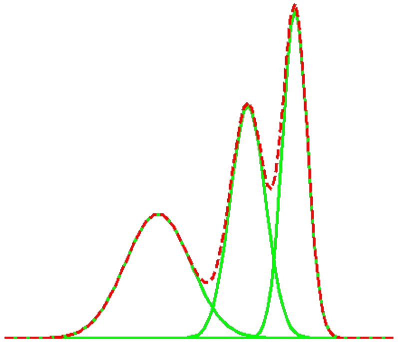
- Compute mean and variance based on the bias corrected intensity

$$\mu_k = \frac{\sum_{n=1}^{N} u_{nk}\left(x_n - \sum_{l=1}^{M} c_l \phi_l(r_n)\right)}{\sum_{n=1}^{N} u_{nk}}$$

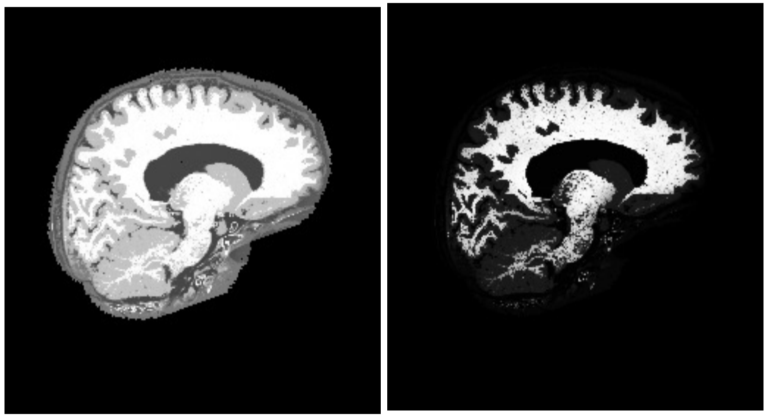
$$\Sigma_k = \frac{\sum_{n=1}^{N} u_{nk}\left(x_n - \sum_{l=1}^{M} c_l \phi_l(r_n) - \mu_k\right)^2}{\sum_{n=1}^{N} u_{nk}}$$

Step 2: bias field estimation

Belonging Probabilities



distribution



Reminder : Linear Least Square Problems

- Linear Least Square :
  - For any rectangular matrix X, Find X such that $\hat{X} = \arg\min_X ||AX - B||^2$
  - If rank(A) is full then $\hat{X} = (A^T A)^{-1} A^T \ B$
    $A^+ = (A^T A)^{-1} A^T$ is the pseudo-inverse
- Weighted Least Square
  - For any diagonal matrix W Find X such that $\hat{X} = \arg\min_X (AX - B)^T W (AX - B)$
  - If rank(A) is full then $\hat{X} = (A^T W A)^{-1} A^T W \ B$

Bias Field Estimation

- Estimate Bias Field to minimize :

$$Q(\theta, C) = -\sum_n \sum_k u_{nk} \log\big(\pi_k \mathcal{N}(x_n - \sum_l C_l \phi_l(r_n)\,;\mu_k, \Sigma_k)\big) \;\; + \ldots$$

$$Q(\theta, C) = -\sum_n \sum_k u_{nk} \left(\frac{x_n - \sum_l C_l \phi_l(r_n) - \mu_k}{2\Sigma_k}\right)^2 + \ldots$$

- C solution of a weighted least square problem

$$Q(\theta, C) = (AC - R)^T W (AC - R) + \ldots$$

Bias Field Estimation

- Estimate Bias Field in the least square sense

$$\begin{bmatrix} C_1 \\ \vdots \\ C_M \end{bmatrix} = (A^T W A)^{-1} A^T W R$$

where

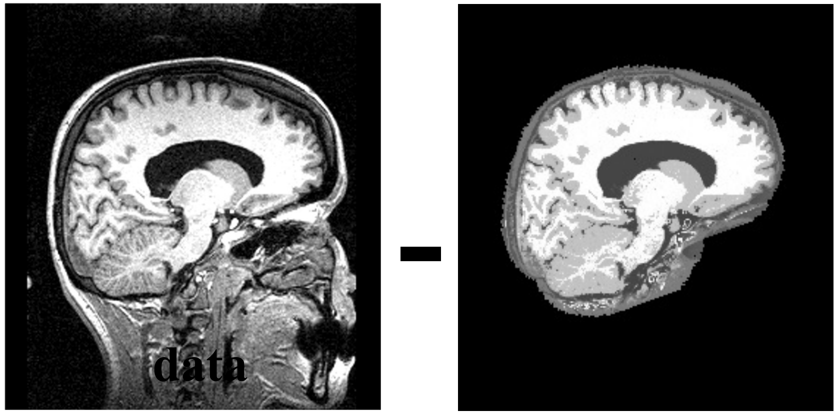
$$A = \begin{bmatrix} 1 & \dots & \phi_M(r_1) \\ 1 & \dots & \vdots \\ 1 & \dots & \phi_M(r_n) \end{bmatrix}$$

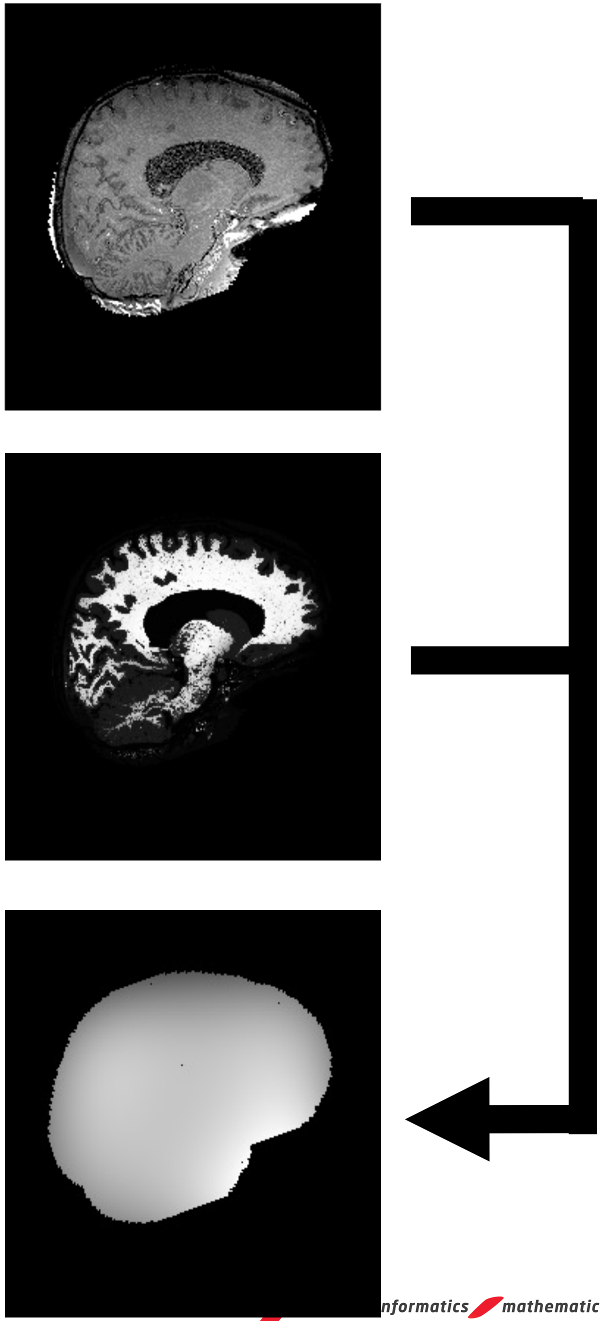
$$W = \begin{bmatrix} W_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & W_N \end{bmatrix} \qquad W_n = \sum_{j=1}^{K} \frac{u_{nj}}{\Sigma_j}$$

Difference between

Intensity and expected intensity
without any bias

$$R = \begin{bmatrix} x_1 - \dfrac{\sum_{k=1}^{K} u_{1k}\mu_k / \Sigma_k}{\sum_{k=1}^{K} u_{1k} / \Sigma_k} \\ x_2 - \dfrac{\sum_{k=1}^{K} u_{2k}\mu_k / \Sigma_k}{\sum_{k=1}^{K} u_{2k} / \Sigma_k} \\ \text{M} \\ \\ \end{bmatrix}$$

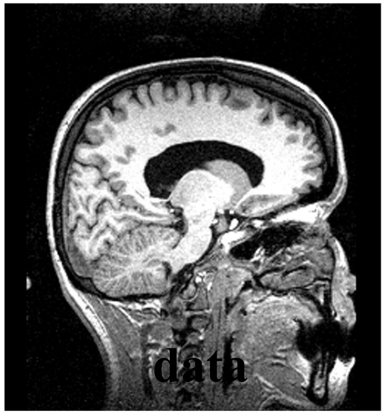


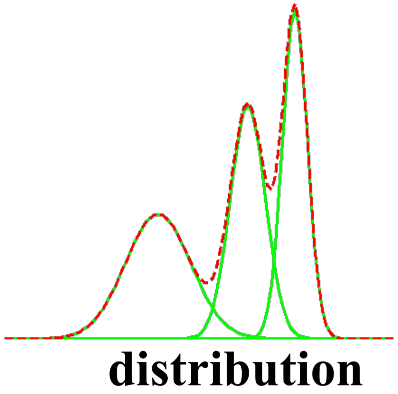


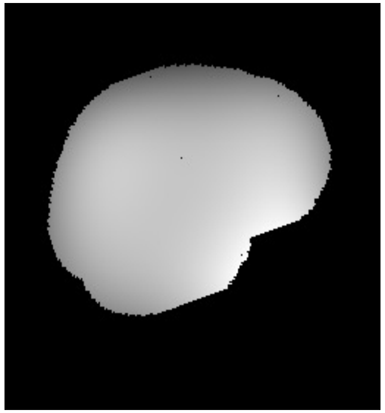
$$\begin{bmatrix} C_1 \\ \vdots \\ C_M \end{bmatrix} = (A^T W A)^{-1} A^T W R$$

$$R = \begin{bmatrix} x_1 - \dfrac{\sum_{k=1}^{K} u_{1k} \mu_k / \Sigma_k}{\sum_{k=1}^{K} u_{1k} / \Sigma_k} \\ x_2 - \dfrac{\sum_{k=1}^{K} u_{2k} \mu_k / \Sigma_k}{\sum_{k=1}^{K} u_{2k} / \Sigma_k} \\ \vdots \end{bmatrix}$$

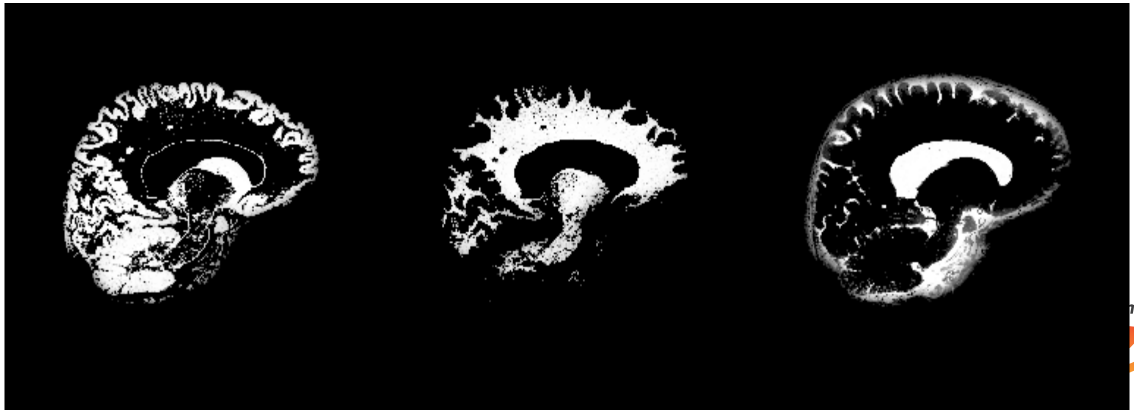
Step 3: classification



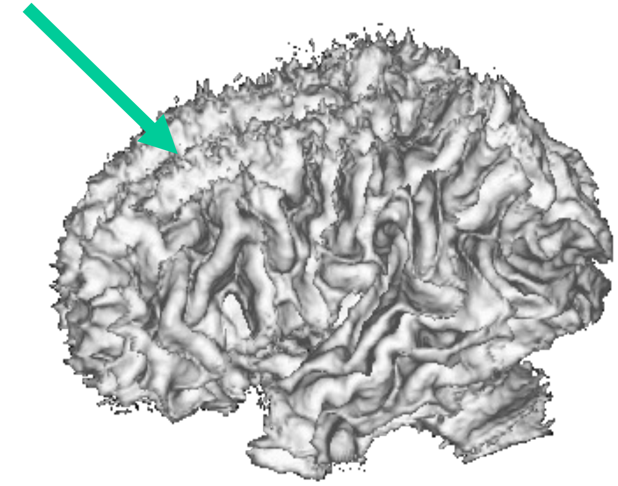




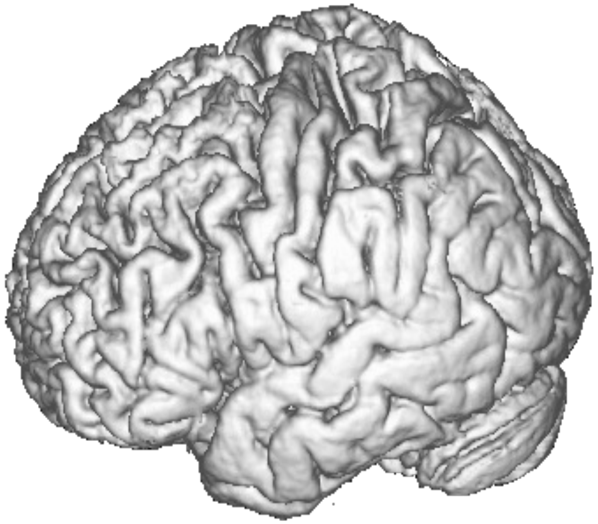
$$u_{nk} = \frac{\pi_k G\left(x - \sum_l C_l \phi_l (r_n) \,|\, \mu_k , \Sigma_k \right)}{\sum_{j=1}^{K} \pi_j G\left(x - \sum_l C_l \phi_l (r_n) \,|\, \mu_j , \Sigma_j \right)}$$

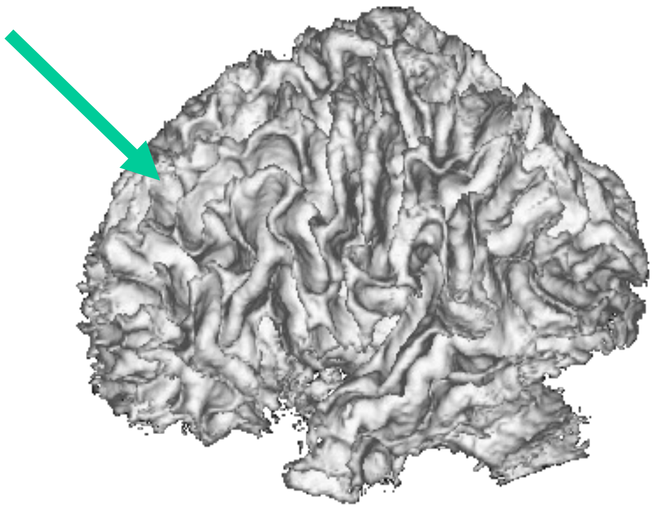


Results

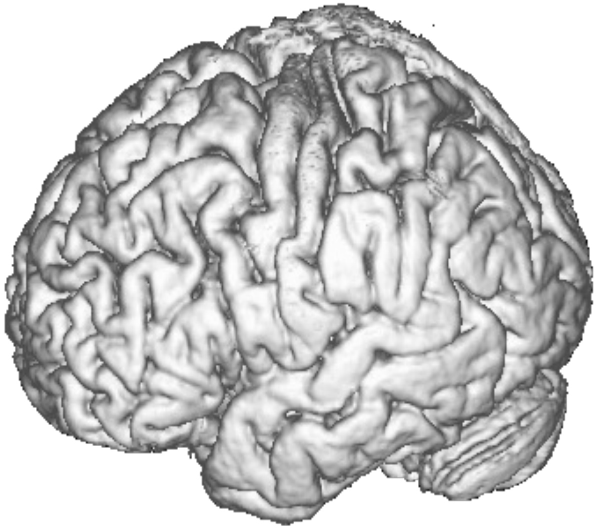


without
bias correction





with
bias correction



white matter surface

Courtesy of D. Vandermeulen

gray matter surface

3. Medical Image Segmentation

- 3.1 Taxonomy of segmentation algorithms
- 3.2 Validation of segmentation algorithms
- 3.3 Deterministic Filtering & Thresholding Approaches
- 3.4 Probabilistic Imaging Model
- 3.5 Expectation Maximisation for GMM
- 3.6 Image classification with bias field
- 3.7 **Variational Bayes EM**
- 3.8 STAPLE Algorithm

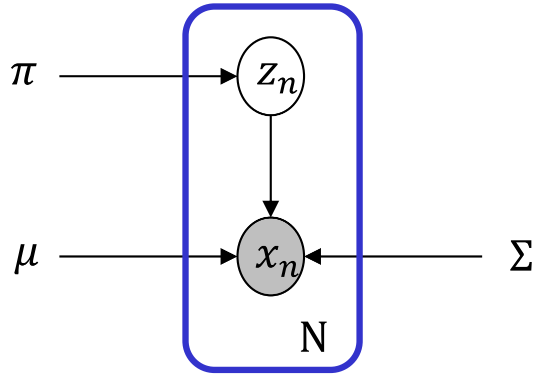
Addressing limitations of EM

- Limitations on the EM algorithm :
  - 1) Trivial solutions of the maximum Likelihood (Dirac on data points)
  - 2) Non invertibility of Covariance matrices Σ_k
  - 3) Most fix the number of classes prior to the algorithm
- To address those limitations :

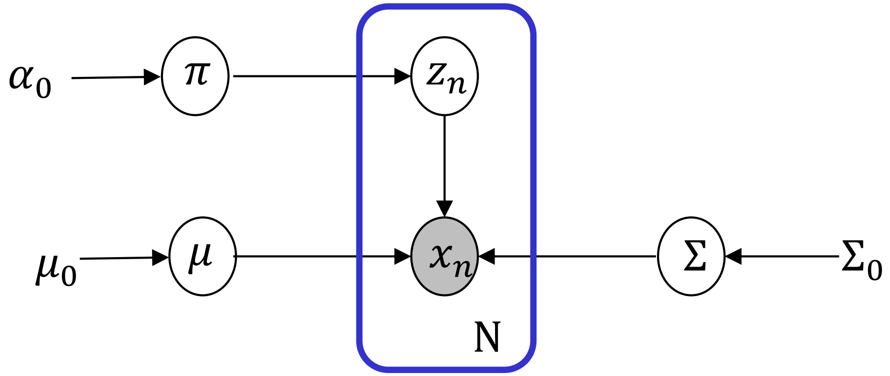
add priors on parameters -> Variational Bayes EM

Variational Bayes Gaussian Mixture

- Parameters θ are now random variables
- Define (hyper)prior probabilities $p(\theta)$ to “regularize” their estimated values



Regular Gaussian Mixture Model



Variational Gaussian Mixture Model

$\alpha_0, \mu_0, \Sigma_0$ are hyperparameters

How to choose the parameter prior distributiom ?

- Convenient choice : “conjugate prior”
  - Lead to closed form expression of posterior

$$p(\theta|X\,) = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta')p(\theta')d\theta'}$$

- Example :

Likelihood distribution $p(X\|\theta)$	**Parameters**	**Conjugated Prior $p(\theta)$**	**Posterior $p(\theta\|X)$**
Gaussian	μ, Σ	Normal x Wishart	Normal x Wishart
Multinomial	π	Dirichlet	Dirichlet

- Existence of conjugated priors for the “exponential family”

Variational Bayes
Gaussian Mixture

- Extended EM algorithm :
  - replace maximization of likelihood $p(X|\theta)$ with joint probability $p(X,\theta) = p(X|\theta)p(\theta)$

  - Keep lower bound : $\log p(X,\theta) \geq \mathbb{E}_u(\log p(X,Z|\theta)) + H(u) + \log p(\theta)$

  - Same E-step, but modified M-steps

VBGM : prior on covariance

- Add prior on K inverse Covariance (precision) Matrices Σ_k^{-1}:
  - Objectives :
    - Remove trivial solutions
    - Make Σ_k invertible
  - Wishart Distribution $p(\Sigma_k^{-1}) = \mathcal{W}(\Sigma_k^{-1}; \Sigma_k^0, \beta_k)$ centered on fixed precision (inverse covariance) matrix Σ_0.
  - New M-Step : $\Sigma_k = \frac{\sum_{n=1}^N u_{nk}(x_n - \mu_k)(x_n - \mu_k)^T + \Sigma_k^0}{\sum_{n=1}^N u_{nk} + 1 + 2\beta_k - (d+1)}$

VBGM : prior on mixture coefficients

- Add prior on K mixture coefficient $\{\pi_k\} = \pi$:
  - Objectives :
    - Remove small clusters
  - Dirichlet distribution $p(\pi_k) = \mathcal{D}(\pi; \alpha_k)$ where $\alpha_k$is a positive scalar
  - If α_k <<1 then this prior is sparsity inducing : $p(\pi_k)$ is either 0 or close to 1
  - New M-step : $\pi_k = \frac{\sum_{n=1}^{N} u_{nk} + \alpha_k - 1}{\sum_{k=1}^{K} \alpha_k + N - K}$

VBGM : prior on mean intensity

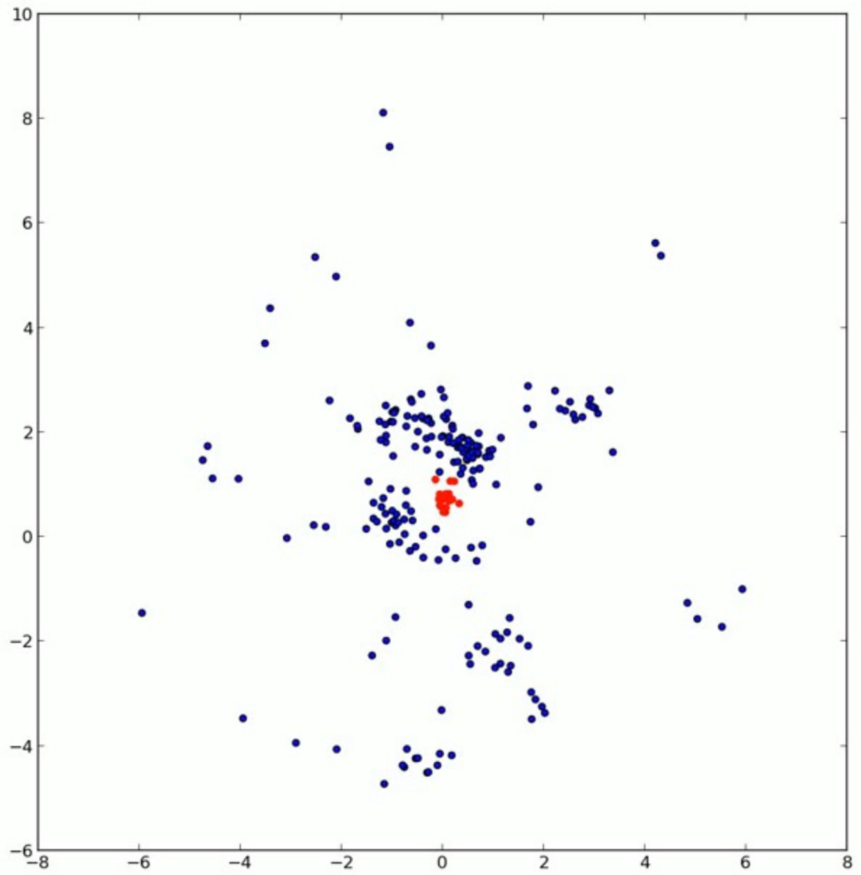
- Add prior on K mean values μ_k :
  - Objectives :
    - Constraint mean values to be within certain range given by (μ_k^0, Σ_k^0)
  - Prior is Gaussian distribution characterized by a mean and covariance matrix $p(\mu_k) = \mathcal{N}\left(\mu_k; \mu_k^0, \left(\Sigma_k^0\right)^{-1}\Sigma_k\right)$
  - New M-step :

$$\mu_k = \frac{\sum_n^N u_{nk} x_n + \Sigma_k^0 \mu_k^0}{\sum_{n=1}^N u_{nk} + tr(\Sigma_k^0)}$$

VBGM : estimating number of classes

- Initialize GMM with many classes
- Perform VB iterations
  - Dirichlet prior will put move some mixture coefficient to 0 ->
- Remove classes with $\pi_k \approx 0$
- Start again VB iterations until $\pi_k > T$ such that each class has enough samples

Variational Bayes Gaussian Mixture : Example

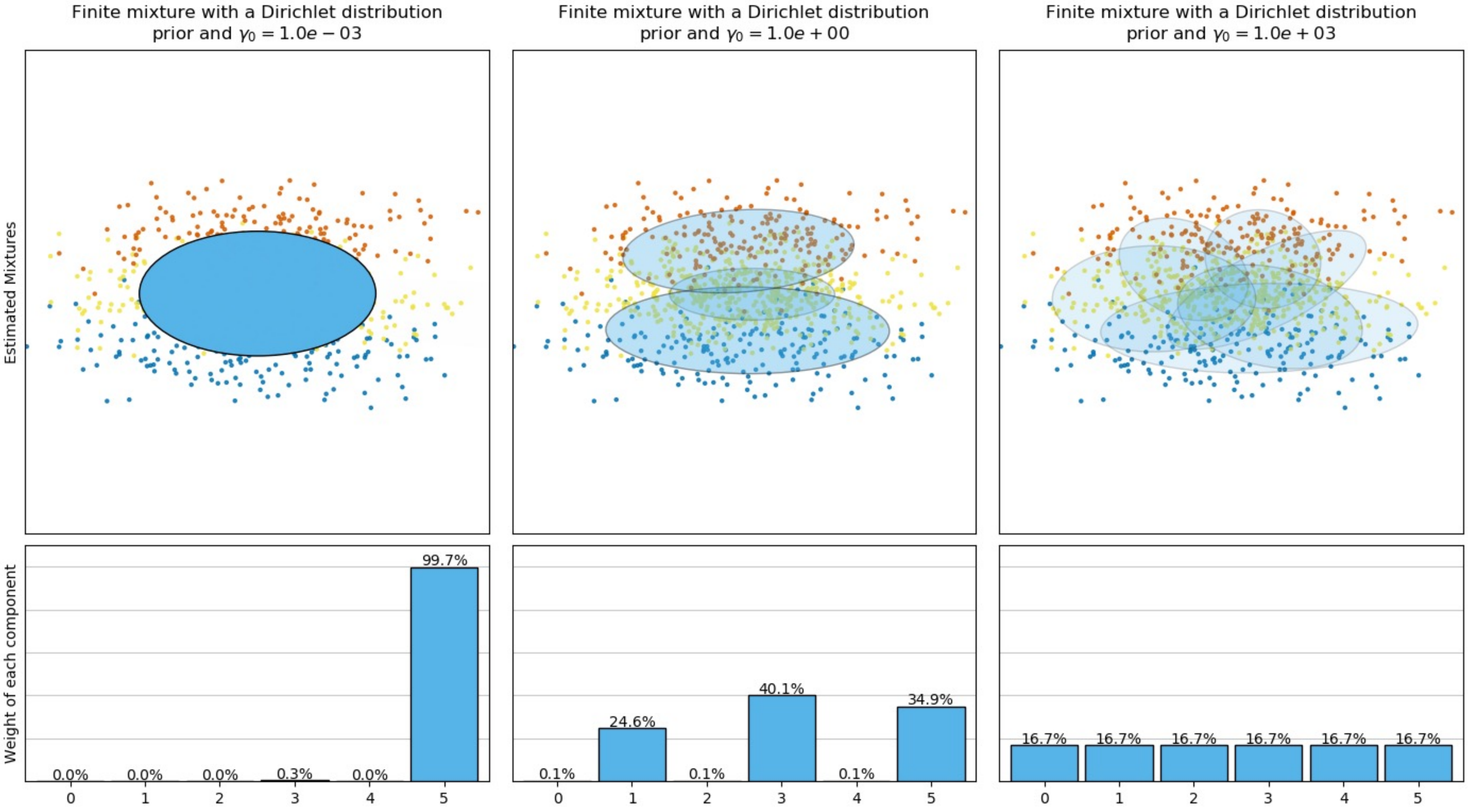


Source : J. Mc Inerney https://www.youtube.com/watch?v=jijtOcVl0Kw

Effect of Dirichlet Prior on the number of components

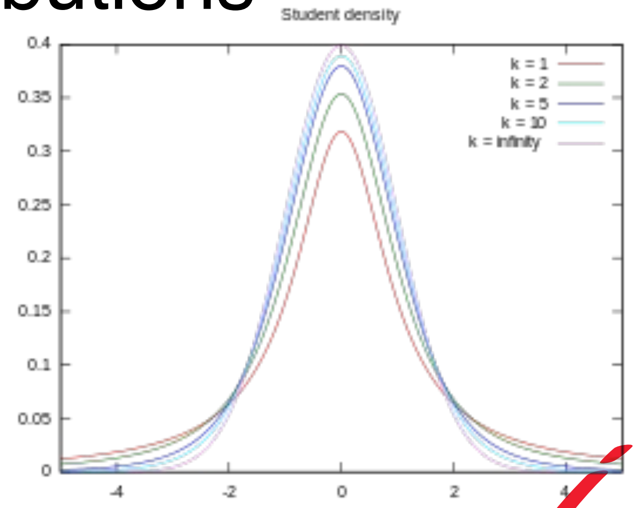
Source : scikit-learn documentation http://scikit-learn.org/stable/modules/mixture.html#bgmm

Initial number of clusters is 5



Other extensions of Gaussian Mixture models

- Dealing with number of classes
  - Use Dirichlet process equivalent to using infinite number of classes
- Robustness to outliers
  - Replace mixture of Gaussians with mixture of Student distributions



$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

ν is the degrees of freedom

$Student$ tends towards normal distribution
As $\nu \rightarrow \infty$

Medical Imaging :
Connexity and Shape Constrained Image segmentation

Hervé Delingette
Epione Team
Herve.Delingette@inria.fr

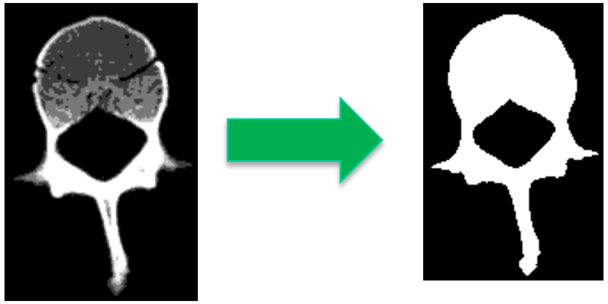
4. Connexity and Shape Constrained Image segmentation

- **4.1 Label Connexity Hypothesis : Markov Random Field**
  - **Definition of prior**
  - **Graph cut algorithm**
  - **Neighborhood EM**
  - **Grab Cut**
- 4.2 Introduction to shape and deformable Models
- 4.3 Snakes algorithm
- 4.4 Level Set Algorithm
- 4.5 Point Distribution Model
- 4.6 Multi-atlas Algorithm

Image Segmentation Approaches

Intensity Only

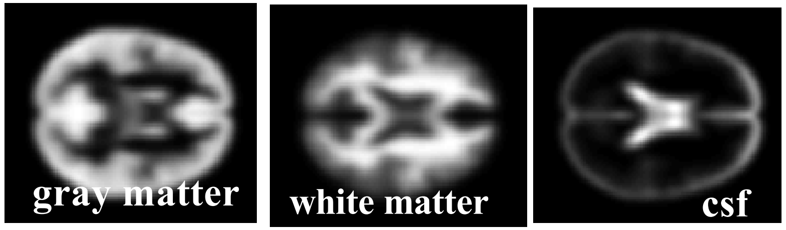
Thresholding



Contrast Agent in CT

Lesions in CT / MR

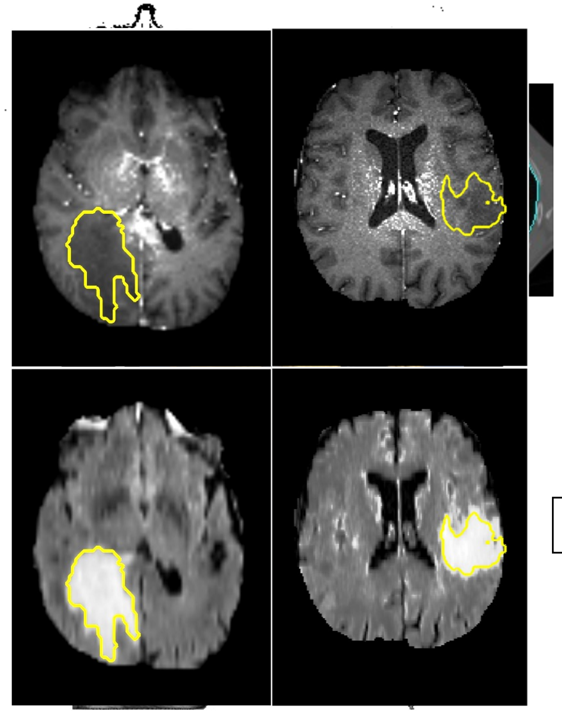
Intensity-only Classification



No Typical Shape

Typical Shape

Image Segmentation Approaches



Intensity and connexity between regions

Mathematical Morphology

MRF (graph cuts, RW, watershed)

Machine Learning (RF, SVM, Boost,ML)

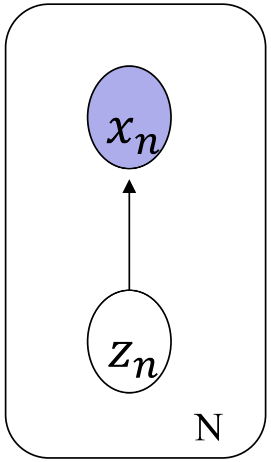
Vessels / tumors / bones /lesions

Grey / White matter in MR

No Typical Shape

Typical Shape

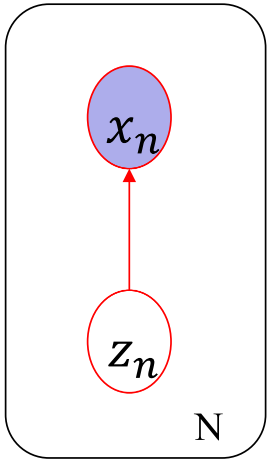
MoG Segmentation Hypothesis



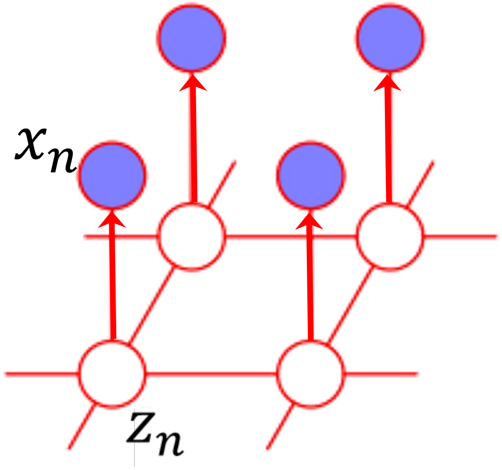
- So far considered independent voxels
  - Z_n variable specifying the class of voxel n
  - X_n variable representing the intensity
- Class membership only dependent on voxel intensity (thresholding)
- But may not be realistic in the presence of noise & partial volume effect

MRF Segmentation Hypothesis

- In Markov Random Fields :
  - Label variables z_n are no longer independent but depend on their neighbors
  - Intensity variables x_n only depends on the class label (variable z_n)



Mixture of Gaussian



Markov Random Field

Markov Random Field

- Intensity prior depends on neighboring values :

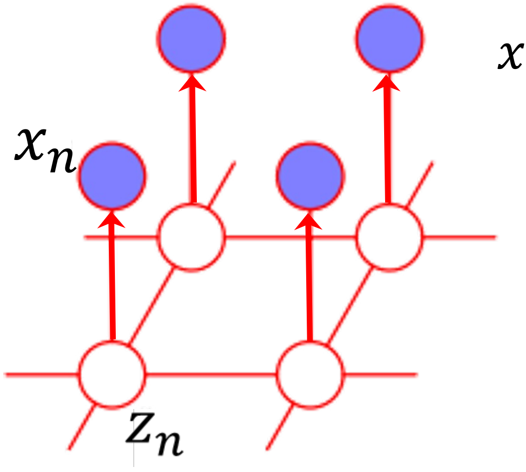
$$p(Z_n|Z_{-n}) = p(Z_n|Z_{N(n)})$$

Label at voxel n

Set of Labels of
all image voxels except
Voxel n

Labels of
Neighboring voxels
Of voxel n

- Graphical Model



x_n are independent only if z_n are known
(conditional independence)

$$p(X) \neq \prod_n p(x_n)$$

$$p(X|Z) = \prod_n p(x_n|z_n)$$

Challenges in MRF

- Posterior probability is no longer tractable

$$p(Z|X) = \frac{p(X|Z)p(Z)}{\sum_{Z'} p(X|Z')p(Z')}$$

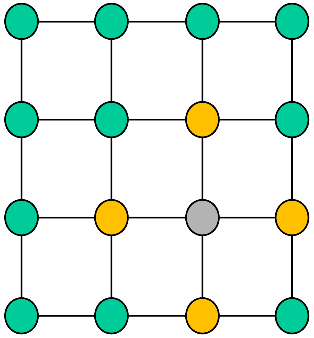
Intractable sum over 2^N terms

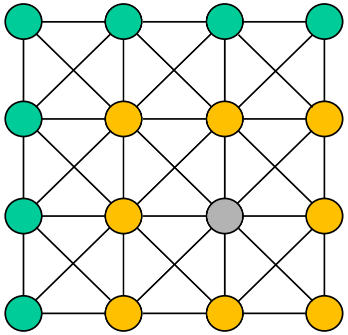
$$p(z_n|X) = \underbrace{\sum_{z_1} \sum_{z_2} \cdots \sum_{z_{n-1}} \sum_{z_{n+1}} \sum_{z_N}}_{\text{Intractable marginalization over N-1 term}} p(Z|X)$$

Definition of Label Prior in MRF

- Images seen as Graph

4 neighborhood





8- neighborhood

- Label Prior $p(Z)$ depends on neighborhood :
  - 2D images : 4 or 8 neighborhood
  - 3D images : 6, 18 or 26 neighborhood

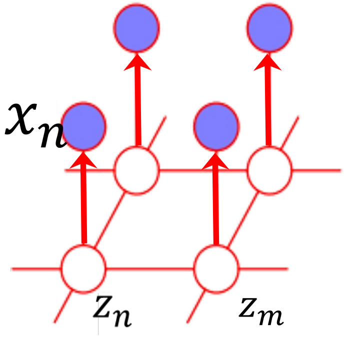
Definition of Label Prior in MRF

- Label prior $p(Z)$ is defined on a graph

  4 neighborhood : $p(Z_n|Z_{-n}) = f(Z_{n-1}, Z_{n+1}, Z_{n-R}, Z_{n+R})$

- Hammersley-Clifford theorem gives the expression of $p(Z)$:
  - There exists functions ψ and ϕ such that

$$\log p(Z|\theta) = \frac{-1}{T} \underbrace{\sum_{edges(n,m)} \psi(z_n, z_m, \theta)}_{\text{Binary term}} - \frac{1}{T^*} \underbrace{\sum_n \phi(z_n, \theta)}_{\text{Unary term}}$$



$\psi(z_n, z_m, \theta)$ is any function of 2 Binary vectors : it enforces how likely are two labels are different

$\phi(z_n, \theta) = \phi_n$

Gives how likely voxel n belongs to class k

Potts Model for Label Prior

- Idea : neighboring voxels should have similar labels.
- Definition Ising when K=2 :
  - One hot encoding : $Z_n = (Z_{n1}, Z_{n2} \dots Z_{nK})^T$
  - $\psi(z_n, z_m, \theta) = -\sum_{k=1}^{K} f_{nm} z_{nk}\, z_{mk}$,
  - In another words :
    - $\psi(z_n, z_m, \theta) = -f_{nm}$ if $Z_n = Z_m$ and $\psi(z_n, z_m, \theta) = 0$ if $Z_n \neq Z_m$,
- Alternative 1 : $\psi(z_n, z_m, \theta) = f_{nm} \| Z_n - Z_m \|^2$
- Coefficient definition : neighboring voxels having similar intensity should have the same labels.

$$f_{nm} = \exp -\beta (x_n - x_m)^2$$

Joint Probability in MRFs

- Definition of joint probability :
  - $p(X,Z|\theta) = p(Z)p(X|Z)$
- Log joint probability

$$\Lambda(Z,\theta) = \log p(X,Z|\theta) = \log p(Z|\theta) + \log p(X|Z,\theta)$$

Conditional independence ➡

$$\Lambda(Z,\theta) = \log p(Z|\theta) + \sum_n \log p(x_n|z_n,\theta)$$

Categorical variable ➡

$$\Lambda(Z,\theta) = \log p(Z|\theta) + \sum_n \sum_k z_{nk} \log p(x_n|z_{nk} = 1,\theta)$$

Energy

$$-\Lambda(Z,\theta) = \frac{1}{T} \underbrace{\sum_{edges(n,m)} \psi(z_n,z_m,\theta)}_{\text{Binary term}} + \frac{1}{T^*} \underbrace{\sum_n \phi(z_n,\theta) - \sum_n \sum_k z_{nk} \log p(x_n|z_{nk} = 1,\theta)}_{\text{Unary terms}}$$

Algorithms for solving MRF

- Many existing algorithms :
  - 1) **Graph cut Algorithm** :
    - Fast
    - solve for hard memberships z_{nk}
    - Unique solution for K=2 if some constraints on f_{nm} are met
    - Several extensions for K>2
  - 2) **Neighborhood EM**
    - solve for soft memberships $\mathrm{p}(z_n|x_n)$
    - Simple Extension of GMM
    - Fixed point Iterative method
  - 3) **Grab Cut**

Graph cuts

- Binary case & Ising model :
  - 2 labels case $y_i \in \{0,1\}$
  - Minimize energy :
    $E(Y) = \sum_{i,j} c_{ij} y_i (1 - y_j) + \sum_i d_i y_i$, with $d_i > 0$
  - Submodular constraint for unique solution

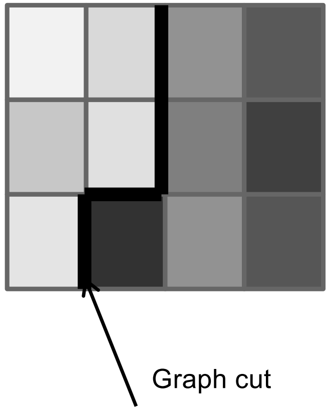
$$c_{ij} + c_{ji} \geq 0$$

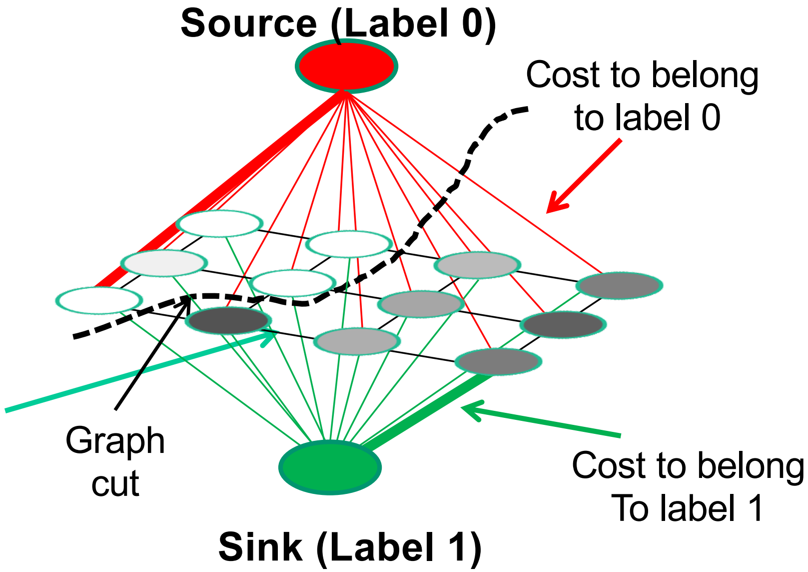
  - Minimize $E(Y)$ ⟺ Minimize a graph cut

Combinatorial problem

D.M. Greig, B.T. Porteous and A.H. Seheult (1989), *Exact maximum a posteriori estimation for binary images*, Journal of the Royal Statistical Society Series B, **51**, 271–279.

Graph Cut





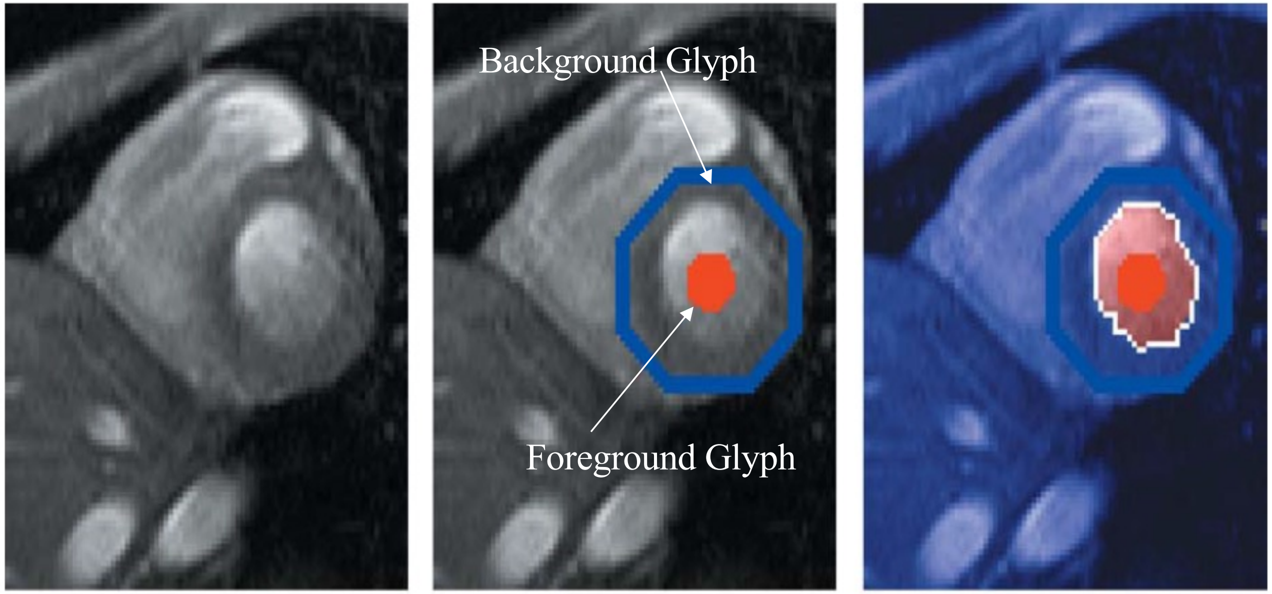
- Minimal graph cut :
  - Set of edges whose removal create several connected components:
  - Cost of a cut :

$$cut(A, B) = \sum_{p \in A, q \in B} c_{p,q}$$

⟺

Maximize the flux between the source and the sink nodes

Interactive Segmentation Algorithm



Source : Boykov & Gareth Funka-Lea Graph Cuts and Efficient N-D Image Segmentation

Manual glyph from user to guide segmentation

Graph cut Segmentation

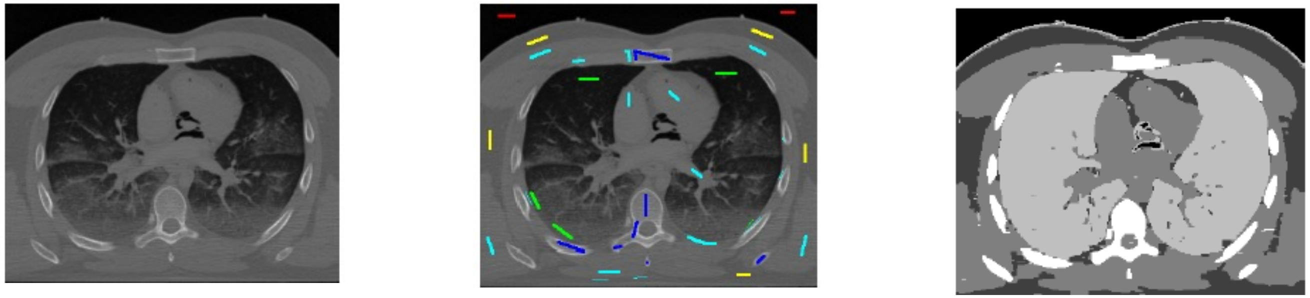
- Combinatorial algorithm for graph cut :

  Ford & Fulkerson Algorithm (1951)

  BoyKov & Kolmogorov Algorithm (2004)

  Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9):1124–1137, September 2004.

- Multi Label Segmentation with α-expansion algorithm [Veksler 99] [Boykov 99]



R. Kéchichian, S. Valette, M. Desvignes, R. Prost: **Efficient multi-object segmentation of 3D medical images using clustering and graph cuts.** ICIP 2011

Neighborhood EM

- Hypothesis :
  - Posterior probability $p(z_n|X\ \ )$ is intractable therefore estimate an approximation
  - Each tissue class is represented by a Gaussian distribution

$$p(x_n|z_{nk} = 1) = \mathcal{N}(x_n|\theta_k)$$

  - The label prior is a Potts model and global prior per class

$$\log \mathrm{p(Z)} = -\frac{\beta}{2}\sum_k \sum_{edges(m,n)} c_{nm}\ z_{nk}z_{mk} + \sum_n \sum_k \pi_k z_{nk}$$

C. Ambroise , M. Dang , G. Govaert: Clustering of Spatial Data by the EM Algorithm. In geoENV I-Geostatistics for Environmental Applications (1997), pp. 493-504.

Mean Field approximation

- A.ka Variational Bayes approach
  - Look for an approximation of posterior parameters as product $q(Z) = \{q_n\}$ of factorized terms $p(Z = \{z_n\}|X) \approx \prod_n q_n(z_n)$
  - Therefore NK unknown q_{nk} s.t

$$\mathrm{q_n}(z_n) = \sum_k q_{nk} z_{nk} \ \ \& \sum_k q_{nk} = 1 = \sum_{z_n} q_n(z_n)$$

  - Find the set q which minimizes the Kullback Leibler divergence between q and true posterior $p(Z|X)$

Mean Field Criterion

- Reminder EM criterion for GMM :
  - Maximize =F(π,θ,u)

$$\mathrm{F}(\pi,\theta,u)=\mathrm{L}(\pi,\theta)-D_{KL}\big(u||p(z|x)\big)=\mathrm{Q}(\theta,u)+\mathrm{H}(u)$$

- Evidence Lower bound :

$$D_{KL}\big(q||p(Z|X)\big)=-\log p(X)-\mathbb{E}_q\,(\log p(X,Z))-H(q)$$

- Neighborhood EM criterion same as GMM but with additional term $R(q)$

minimize $D_{KL}\big(q|p(Z|X)\big)=-\mathrm{H}(q)+R(q)-\mathrm{Q}(\mathrm{q})+\log p(X)$

  - Where $R(q)=\frac{\beta}{2}\sum_k\sum_{edges(n,m)}c_{nm}\,q_{nk}q_{mk}$

Neighborhood EM

- Only E-step changed compared to regular EM for GMM
- New E-step :
  - Fixed point iteration

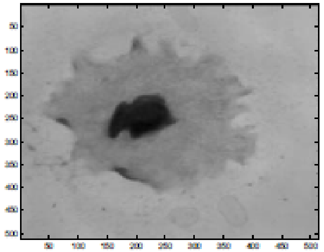
$$q_{nk} = \frac{\pi_k \mathcal{N}(x_n|\theta_k) \exp \beta \sum_m c_{mn} q_{nm}}{\sum_l \pi_l \mathcal{N}(x_n|\theta_l) \exp \beta \sum_m c_{mn} q_{nm}}$$

- Same M-step

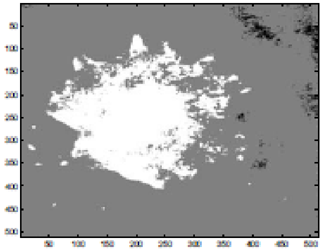
$$\mu_k = \frac{\sum_{n=1}^{N} q_{nk} \; x_n}{\sum_{n=1}^{N} q_{nk}} \qquad \Sigma_k = \frac{\sum_{n=1}^{N} q_{nk} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{i=1}^{N} q_{nk}} \quad \pi_k = \frac{1}{N} \sum_{n=1}^{N} q_{nk}$$

Neighborhood EM

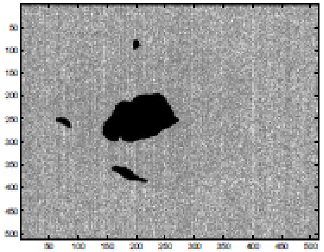
(a) Original image



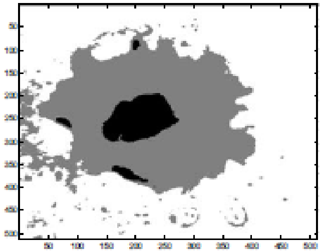
(b) EM



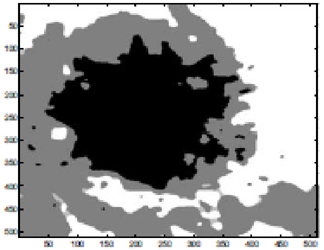
(c) Gibbsian EM



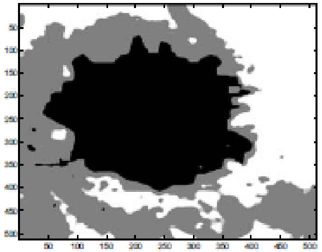
(d) Neighborhood EM with $\beta = 0.5$



(e) Neighborhood EM with $\beta = 1.0$

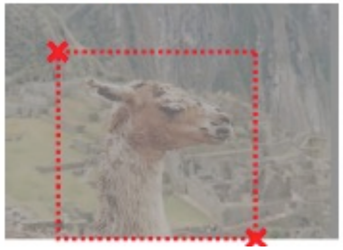


(f) Neighborhood EM with $\beta = 2.0$



Grab Cut

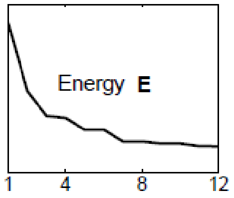
- Algorithm combines :
  - Model intensity of foreground and background as mixture of Gaussians (vs one Gaussian for each class)
  - Iterate between :
    - hard segmentation using graph cuts
    - Estimation of Gaussian components



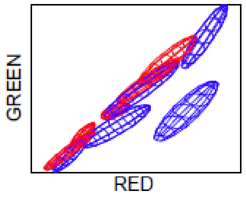
Input + bounding box

(f)

Segmentation

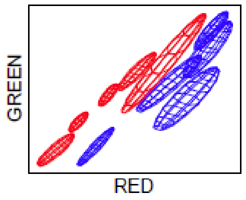


(a)



(b)

Initial GMM



(c)

Final GMM

Grab Cut Examples

Available in MS Office !!

Difficult Examples

Camouflage & Low Contrast

Fine structure

Harder Case

Initial Rectangle

Initial Result

Grabcut: Interactive foreground extraction using iterated graph cuts, Carsten Rother, V. Kolmogorov, Andrew Blake, Siggraph 2004

4. Connexity and Shape Constrained Image segmentation

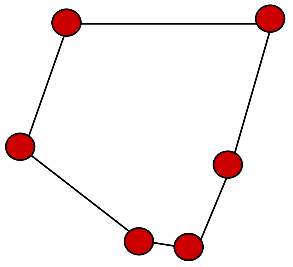
- 4.1 Label Connexity Hypothesis : Markov Random Field
- **4.2 Introduction to shape and deformable Models**
- 4.3 Snakes algorithm
- 4.4 Level Set Algorithm
- 4.5 Point Distribution Model
- 4.6 Multi-atlas Algorithm

Shape Constraints in Image Segmentation

- MRFs enforce connectivity between neighboring voxels : region approach
- Deformable shapes / models :
  - Work on boundaries between regions -> dual approach
  - Define constraints on the boundaries :
    - Minimize length
    - Minimize curvature
    - Shape constraints

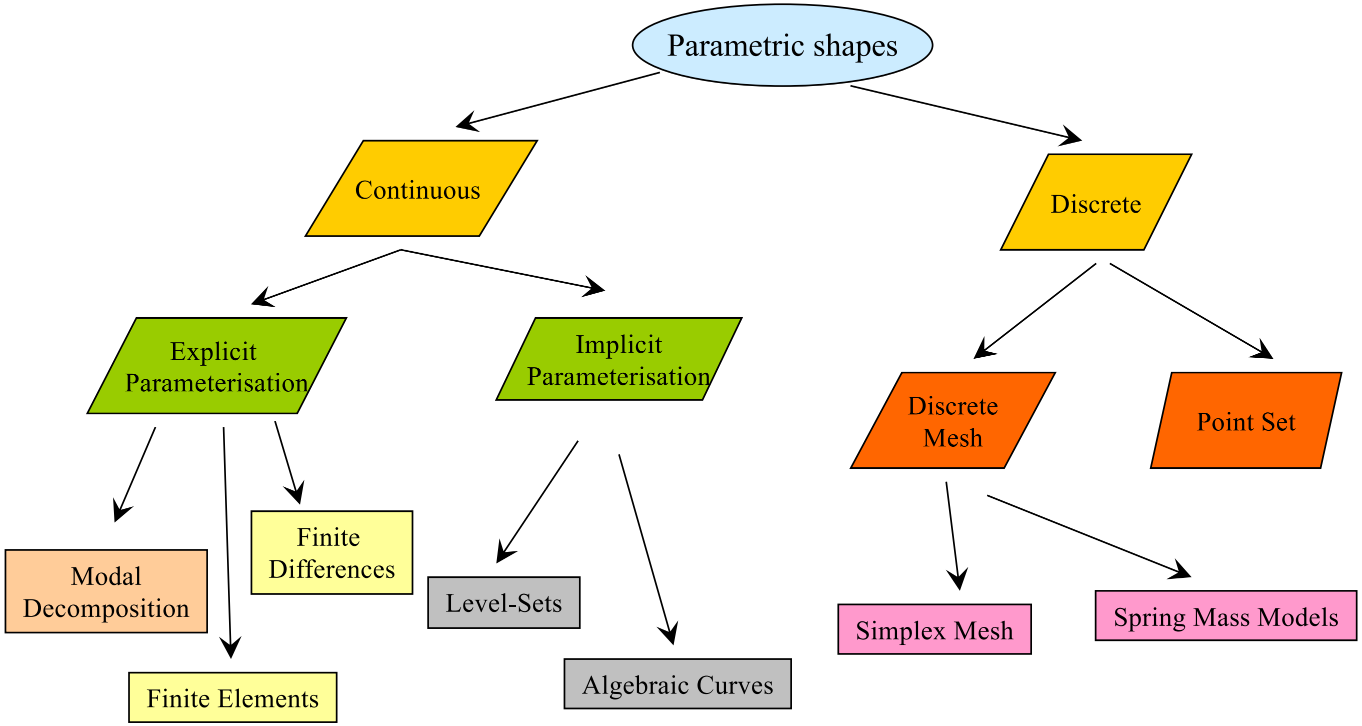
Parametric Shape representation

- Parametric representation of a shape :
  - Shape controlled by (intrinsic) parameters
- Examples :
  - Vertex position of a mesh
  - Scalar field for level sets
  - Fourier coefficients,…



Deformation in the object space

Shape representation


Parametric shapes

Continuous

Discrete

Explicit Parameterisation

Implicit Parameterisation

Discrete Mesh

Point Set

Modal Decomposition

Finite Differences

Finite Elements

Level-Sets

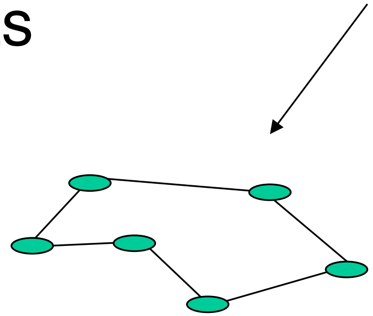
Algebraic Curves

Simplex Mesh

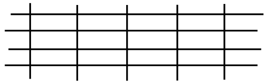
Spring Mass Models

Shape representation As Template Transformation

- Template Transformation :
  - Define a single shape instance in $\mathbb{R}^n$ as template
  - Parameterise the deformation of the embedding space $\phi(x)$: $\mathbb{R}^n \rightarrow \mathbb{R}^n$
- Examples :
  - Rigid Transformation (translation + rotation)
  - Affine Transformation (translation + linear transform)



Shape instance= template



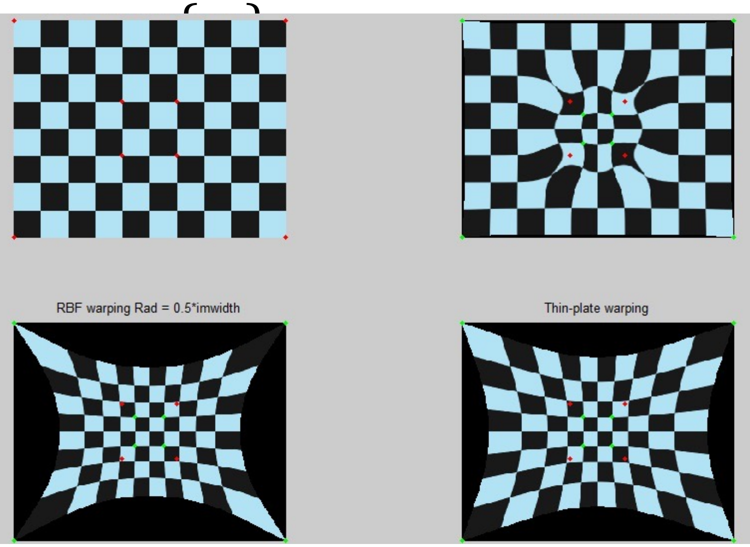
Define $\phi(x)$ as an affine transform

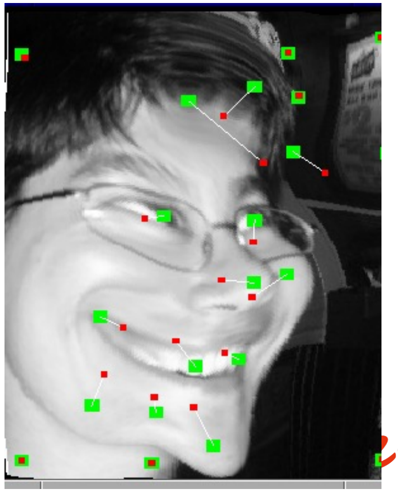
Simple Transformations

T_{reg}	**Description**	**Degrees of Freedom**
2D Rigid	Translation + Rotation	2+1=**3**
2D Similarity	Translation + Rotation + Scale	3+1=**4**
2D Affine	Translation + Linear	2+4=**6**
3D Rigid	Translation + Rotation	3+3=**6**
3D Similarity	Translation + Rotation + Scale	6+1=**7**
3D Affine	Translation + Linear	3+9=**12**

Complex Transformations

- Radial Basis functions :
  - Basis $\psi(x) = \psi(\|x\|)$ which only depend on distance : example : Gaussian, thin plate spline, B-spline
  - Define N control points x_i
  - Define $\phi(x)$ as $\phi(x) = \sum_i^N \psi(x - x_i) y_i$ parameterized by





Shape Optimization

- If {θ} are parameters in the shape space (parametric representation)

➡ Framework of deformable templates

- If {θ} are parameters in the space of geometric transformations

➡ Framework of Image Registration

- Often includes both frameworks