Tutorial on Expectation-Maximization: Application to Segmentation of Brain MRI

Maria Murgasova

May 6, 2007

Contents

1 Introduction

Expectation-maximization is a very popular framework for different classification problems. It became extremely popular in segmentation of brain MRI during last decade. However, it is not easy to understand the underlying theory. In this tutorial I offer complete definitions and descriptions of the widely-used EM-based classification methods with focus on segmentation of brain MRI. I will not leave any holes - so hopefully if you read carefully, you will have a deep understanding of the method!

2 Model for brain MRI

Magnetic resonance imaging (MRI) is a medical imaging technique very well suited for analyzing human soft tissue anatomy. It provides high resolution 3D volumetric data with high intensity contrast between soft tissues. Normal adult brain MR images are theoretically piecewise constant with small number of classes - of which only 3 are usually of interest - white matter (WM), gray matter (GM) and cerebro-spinal fluid (CSF). However, the contrast between tissues depends on how the image is acquired and it is often difficult to find ideal radio-frequency and gradient pulses in practice. Image is further corrupted by electronic noise and intensity inhomogeneity of od the magnetic field which results in slow continuous change of intensity across the image. Situation is further complicated by partial volume effect, when multiple tissues are present in one voxel. This creates problems especially along the boundary of cortical GM and CSF as the complicated shape of cortex and a lot of small spaces filled with CSF result in large volume of voxels containing mixed intensities of GM and CSF .Originally, researchers tried to segment the brain images using tresholding, edge extraction and region growing. These algorithms were vulnerable to noise and artifacts and therefore more robust statistical methods were needed. In this tutorial we will describe current state-of-the-art brain segmentation class of statistical methods based on expectation-maximization framework.

3 Statistical model of the brain segmentation problem

Let the voxel intensities of the brain MR image be denoted $Y = \{y_1, ..., y_n\}$ where the image consists of n voxels. These intensities are called *observed data* and can be viewed as realization of a random variable Y. The real

labeling of the image is not known and is therefore called *hidden data* $Z = \{z_1, ..., z_n\}$. It is assumed that the observed data (the image Y) are generated by the hidden labeling and parameters Φ and image Y is described by conditional probability density function (PDF) $p(Y|Z, \Phi)$. The parameters Φ can describe either the PDF itself, noise or bias field, depending on model.

Y and Z can be viewed as n-dimensional random variables $Y = (Y_1, ..., Y_n)$ and $Z = (Z_1, ..., Z_n)$. Then each voxel intensity y_i is a realization of random variable Y_i and labeling of this voxel z_i is a realization of random variable Z_i. The conditional probability distribution function describing Y_i is $p(Y_i|Z, \Phi)$. As intensity y_i depends only on the label z_i, we have

$$p(Y_i|Z, \Phi) = p(Y_i|Z_i, \Phi)$$

The simplest model assumes that the conditional distribution function for each class is constant (the same intensity value for all voxels of the given tissue class) but corrupted by noise or other factors with Gaussian distribution. We can describe this relationship by formula:

$$y_i = \mu_k + n_i$$

where μ_k is the mean intensity of k^{th} tissue class and n_i is a random sample generated by Gaussian probability distribution function $G(., 0, \sigma)$ with zero mean and variance σ. This means that y_i is a random sample generated by Gaussian probability density function $G(., \mu_k, \sigma)$.

If we assume that other factors may make the intensity variation of the different tissue types different we can use a different variance for each tissue class. Then the conditional probability density function of the random variable Y_i is

$$p(Y_i = y|Z_i = k, \Phi) = G(y, \mu_k, \sigma_k) \tag{1}$$

As the labeling is not known, it is useful to express PDF of Y_i depending only on parameters Φ using total probability theorem:

$$p(Y_i|\Phi) = \sum_{k=1}^{K} p(Y_i|Z_i = k, \Phi)p(Z_i = k|\Phi) \tag{2}$$

The term $p(Z_i = k|\Phi)$ is the *prior probability* that the voxel i belongs to the tissue class k. The term $p(Y_i|Z_i = k, \Phi)$ is called the *likelihood*. If we assume that prior probability is constant for all voxels (does not depend on spatial position of the voxel) we obtain widely used *Gaussian mixture model*.

4 Gaussian Mixture Model

Let us assume that the conditional probability density function for each tissue class is Gaussian (Eq. (1)), the prior probability that the voxel i belongs to tissue class k is spatially constant

$$p(Z_i = k|\Phi) = c_k \tag{3}$$

and parameters Φ are unknown means and variances of Gaussian probability density functions for intensity of each tissue class and weights of the mixture

$$\Phi = (\mu_1, \sigma_1, c_1, ..., \mu_K, \sigma_K, c_K)$$

Using equations 1, 2 and 3, the probability density function for intensity of voxel i can be expressed as a weighted mixture of K Gaussian PDFs given by parameters Φ

$$p(Y_i = y|\Phi) = \sum_{k=1}^{K} G(y, \mu_k, \sigma_k)c_k \tag{4}$$

where

$$G(y, \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{\frac{(y-\mu_k)^2}{2\sigma_k^2}}$$

and c_k are the weights.

In case of Gaussian mixture model all voxels are considered to be independent samples with PDF defined by Eq. 4, which does not depend on spatial position of the voxel and therefore all the voxels have the same PDF. Consequently, the normalised histogram of the image can be considered as an estimate of this PDF. The task can be intuitively described as fitting the Gaussians to the image histogram - see Fig. 4

Fitting the Gaussians means finding the set of parameters Φ (means, variances and weights) which describe the PDF approximated by the image histogram the best. This can be formalized by using *maximum likelihood* principle.

5 Maximum likelihood

When parameters Φ are known, the function $p(Y_i = y|\Phi)$ (Eq. 4) is probability density function for voxel i having intensity value y. However, in our case, the intensity of voxel i is known to be y_i and parameters Φ are to be found. The function $p(Y = y_i|\Phi)$ is called *likelihood function* and for each value of parameters Φ the likelihood function returns the value of likelihood that the observed intensity y_i was generated given the parameters Φ.

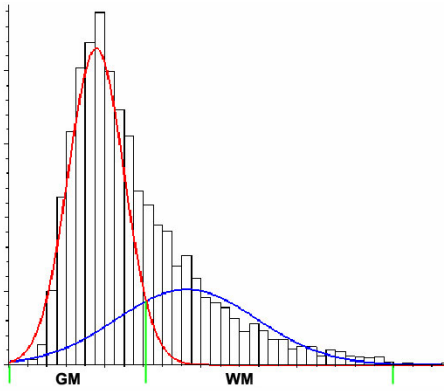

Figure 1: Gaussian mixture model - fitting the mixture of Gaussians to the normalized image histogram

Similarly, we can calculate the likelihood of the whole image $p(Y|\Phi)$. The *maximum likelihood* method is a procedure which finds the such a value of parameters Φ which maximizes the likelihood of the observed image:

$$\hat{\Phi} = \underset{\Phi}{\operatorname{argmax}}\, p(Y|\Phi) \tag{5}$$

It is usually assumed that random variables $Y_1, ..., Y_n$ are independent and therefore the likelihood function of the image Y can be expressed as:

$$p(Y|\Phi) = \prod_{i=1}^{n} p(Y_i|\Phi)$$

Calculating the likelihood in practice means multiplying of a large number of numbers smaller than one end the resulting value is therefore very small. Therefore it is more convenient to work with *log likelihood* as the product is transformed to a sum. Logaritmic function $\log_e(x)$ is increasing and continuous and therefore the solution of Eq. 5 can be found as maximum log-likelihood $L(\Phi)$:

$$\hat{\Phi} = \underset{\Phi}{\operatorname{argmax}} \log_e p(Y|\Phi) = \underset{\Phi}{\operatorname{argmax}}\, L(\Phi) \tag{6}$$

The log-likelihood of the observed image can be further expressed as

$$L(\Phi) = \log_e p(Y|\Phi) = \log_e \prod_{i=1}^{n} p(Y_i|\Phi)$$

and together with Eq. 2 we have

$$L(\Phi) = \sum_{i=1}^{n} \log_e \sum_{k=1}^{K} p(Y_i = y_i | Z_i = z_i, \Phi) p(Z_i = k | \Phi)$$

In case of Gaussian mixture model the log-likelihood becomes

$$L(\Phi) = \sum_{i=1}^{n} \log_e \sum_{k=1}^{K} G(y_i, \mu_k, \sigma_k) c_k$$

When Gaussian PDF is used, the log-likelihood can be maximized by finding partial derivatives for each parameter and setting it equal to zero. Such an expression for parameter μ_j is given by:

$$\frac{\partial}{\partial \mu_j}(L(\Phi)) = 0$$

Differentiating, substituting $p(Y_i = y_i | Z_i = k, \Phi)$ by Gaussian distribution and using Bayes formula

$$p(Z_i = j | Y_i = y_i, \Phi) = \frac{p(Y_i = y_i | Z_i = j, \Phi) p(Z_i = j | \Phi)}{\sum_{k=1} p(Y_i = y_i | Z_i = k, \Phi) p(Z_i = k | \Phi)} \tag{7}$$

yields

$$\sum_{i=1}^{n} p(Z_i = j | Y_i = y_i, \Phi)(y_i - \mu_j) = 0$$

The term $p(Z_i = j | y_i, \Phi)$ is called *posterior probability* and expresses the probability that the voxel i belongs to tissue type j. It is also called *soft assignment* or *soft segmentation*, meaning that voxels are not strictly assigned to only one tissue class. The set

$$P_k = \{p(Z_i = j | Y_i = y_i, \Phi), i = 1, ..., n\}$$

is called a *probability map* for tissue k (see Fig. **??**b,c,d). At the same time, posterior probabilities are the estimate of the underlying segmentation or the hidden data.

Let us denote

$$p_{ik} = p(Z_i = k | Y_i = y_i, \Phi)$$

After rearranging we have an expression for μ_j:

$$\mu_j = \frac{\sum_{i=1}^{n} y_i p_{ij}}{\sum_{i=1}^{n} p_{ij}} \tag{8}$$

Similarly, we can obtain the expressions for variances σ_j and weights c_j

$$\sigma_j^2 = \frac{\sum_{i=1}^{n}(y_i - \mu_j)^2 p_{ij}}{\sum_{i=1}^{n} p_{ij}} \tag{9}$$

$$c_j = \frac{1}{n}\sum_{i=1}^{n} p_{ij} \tag{10}$$

The equation for soft segmentation can be obtaind from Eq. 7, 1, 3:

$$p_{ij} = \frac{G(y_i, \mu_j, \sigma_j)c_j}{\sum_{k=1}^{K} G(y_i, \mu_k, \sigma_k)c_k} \tag{11}$$

The segmentation can now be calculated using iterative process of interleaving estimation of parameters (Eq. 8, 9, 10) and estimation of soft segmentation (Eq. 11). This technique is called *expectation maximization algorithm* and is guaranteed to converge to maximum likelihood estimate $\hat{\Phi}$ under certain conditions.

6 Expectation-Maximization Algorithm

The expectation-maximization algorithm (EM) [**?**] is a general technique for finding maximum likelihood parameter estimates in problems with hidden data. Observed data are known and hidden data can be observed only indirectly through the observed data. The EM tries to find maximum likelihood parameter estimates by first estimating the hidden data based on current parameter estimates. The estimated complete data (observed and hidden data) are then used to estimate the parameters through maximizing the likelihood of the complete data. In other words, EM finds maximum likelihood estimate for observed data through maximizing the likelihood of the complete data.

As we have shown in previous section, in case of GMM the maximum likelihood of the observed data can be calculated directly. However, the hidden data is involved in the resulting equations. This example leads us to the following intuitive description of EM:

Expectation Maximization Algorithm

- **E-step** or expectation step:

  Calculate the estimate $\mathbf{p}^{(m+1)}$ of the hidden data Z from the observed data Y and and current parameter estimate $\Phi^{(m)}$.

- **M-step** or maximization step:

  Calculate the maximum likelihood parameters $\Phi^{(m+1)}$ for the current estimate of the complete data $(\mathbf{y}, \mathbf{p}^{(m+1)})$

EM algorithm iterates between E-step and M-step and converges to maximum likelihood parameter estimate $\hat{\Phi}$ for observed data Y (Eq. 5).

To clarify the notation: The capital letter denotes the random variable and small letters its realization (the value). The PDF associated with random variable X is $p(X = x)$. When the value is unknown we can write the PDF in the short form $p(X)$. In contrast when the value is known (e.g. in case of likelihood function) we can use short notation $p(x)$. In case of multivariate random variable $X = (X_1, ..., X_n)$ we denote the realization of X by a bold letter $X = \mathbf{x} = (x_1, ..., x_n)$.

By using the previous informal definition of the EM and equations for GMM from previous section we can derive the following algorithm:

Gaussian mixture model via EM

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{G(y_i, \mu_j^{(m)}, \sigma_j^{(m)}) c_j^{(m)}}{\sum_{k=1}^{K} G(y_i, \mu_k^{(m)}, \sigma_k^{(m)}) c_k^{(m)}}$$

- **M-step**:

$$\begin{aligned}
\mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} y_i p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
(\sigma_j^{(m+1)})^2 &= \frac{\sum_{i=1}^{n} (y_i - \mu_j^{(m+1)})^2 p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
c_j^{(m+1)} &= \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(m+1)}
\end{aligned}$$

7 Definition of the EM algorithm

In previous section we described the simplified version of EM so that we could clearly explain how is it used. However, if the log-likelihood cannot be maximized by direct differentiation, the more general definition needs to be used.

General definition of Expectation Maximization Algorithm

- **E-step**:

  Calculate a function $Q(.|\Phi^{(m)})$ based on current parameter estimate $\Phi^{(m)}$:

$$Q(\Phi|\Phi^{(m)}) = E(\log_e p(Y, Z|\Phi)|\mathbf{y}, \Phi^{(m)})$$

- **M-step**: Choose such a value for $\Phi^{(m+1)}$ so that it maximizes the function $Q(.|\Phi^{(m)})$

$$\Phi^{(m+1)} = \underset{\Phi}{\operatorname{argmax}}\, Q(\Phi|\Phi^{(m)})$$

EM algorithm iterates between E-step and M-step and converges to maximum likelihood parameter estimate $\hat{\Phi}$ for observed data Y (Eq. 5) under certain conditions not stated in this report.

The function $Q(.|\Phi^{(m)})$ represents the expected value of the log-likelihood from the complete data also called *complete log-likelihood*

$$L_c(\Phi) = log_e p(Y, Z|\Phi)$$

and defines a lower bound to log-likelihood from the observed data $L(.)$ (Eq. 6) (also called *incomplete log-likelihood*) [1]. Maximizing the lower bound leads to increasing of the incomplete log-likelihood. The lower bound is tightened to the likelihood at every iteration until its maximum converges to a local maximum of the log-likelihood.

The expected value of the complete log-likelihood $E(\log_e p(Y, Z|\Phi)|\mathbf{y}, \Phi^{(m)})$ is summed over all values of Z given the observed data and the current parameter estimate. Let us define the *conditional expected value*:

$$E(f(A)|B) = \sum_{\forall a} p(A = a|B) f(a)$$

assuming that a is a realisation of a discrete random variable A. We will use the following well-known properties of expected values:

$$E(A + B|C) = E(A|C) + E(B|C)$$

$$E(dA|C) = dE(A|C)$$

where A,B,C are random variables and d is a constant.

Let us now demonstrate using of the formal definition of EM on GMM and segmentation (or multiple clustering) problem in general. The complete log-likelihood can be further expressed as:

$$L_c(\Phi) = \log_e p(\mathbf{y}, \mathbf{z}|\Phi) = \log_e p(\mathbf{y}|\mathbf{z}, \Phi) + \log_e p(\mathbf{z}|\Phi) =$$

$$= \sum_{i=1}^{n} \log_e p(y_i|z_i, \Phi) + \sum_{i=1}^{n} \log_e p(z_i|\Phi) =$$

$$= \sum_{i=1}^{n} \log_e p(y_i|z_i, \Phi)p(z_i|\Phi) = \sum_{i=1}^{n} \log_e p(y_i, z_i|\Phi)$$

To be able to calculate expectation value over Z we need to amend the definition of the hidden data. Instead of simply assigning the number k we will now consider Z_i to take value from the set of k-dimensional unit vectors $\{e_1, ..., e_K\}$ where $z_i = e_k = (0, ..., 0, 1, 0, ..., 0)$ means that i^{th} voxel belongs to tissue k. Let us denote $z_i = (z_{i1}, ..., z_{iK})$. The complete log likelihood then becomes:

$$L_c(\Phi) = \sum_{i=1}^{n} \sum_{j=1}^{K} z_{ij} \log_e p(y_i, Z_i = e_j|\Phi) = \sum_{i=1}^{n} z_i^T \mathbf{V}(y_i|\Phi)$$

where vector $\mathbf{V}(y_i|\Phi) = (\log_e p(y_i, Z_i = e_1|\Phi), ..., \log_e p(y_i, Z_i = e_K|\Phi))$ is now constant in variable z_i and symbol ^{T} means transposed vector. We can now calculate the function $Q(.|\Phi^{(m)})$:

$$Q(\Phi|\Phi^{(m)}) = E(L_c(\Phi)|\mathbf{y}, \Phi^{(m)}) = E(\sum_{i=1}^{n} z_i^T \mathbf{V}(y_i|\Phi)|\mathbf{y}, \Phi^{(m)}) =$$

$$= \sum_{i=1}^{n} E(z_i|\mathbf{y}, \Phi^{(m)})^T \mathbf{V}(y_i|\Phi)$$

According to definition of conditional expected value $E(z_i|\mathbf{y}, \Phi^{(m)})$ can be further expressed as

$$E(z_i|\mathbf{y}, \Phi^{(m)}) = \sum_{j=1}^{K} p(Z_i = e_j|y_i, \Phi^{(m)})e_j = \sum_{j=1}^{K} p_{ij}^{(m+1)} e_j$$

where

$$p_{ij}^{(m+1)} = p(Z_i = e_j|y_i, \Phi^{(m)})$$

is a soft assignment of voxel i to tissue j at $(m+1)^{st}$ iteration. Finally, the function $Q(.|\Phi^{(m)})$ can be expressed as

$$Q(\Phi|\Phi^{(m)}) = \sum_{i=1}^{n} \sum_{j=1}^{K} p_{ij}^{(m+1)} \log_e p(y_i, Z_i = e_j|\Phi) =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{K} p_{ij}^{(m+1)} \log_e p(y_i|Z_i = e_j, \Phi) + \sum_{i=1}^{n} \sum_{j=1}^{K} p_{ij}^{(m+1)} \log_e p(Z_i = e_j|\Phi)$$

It is now obvious that for the segmentation problem, calculating of $Q(.|\Phi^{(m)})$ in E-step is equivalent to calculating soft assignment according to Bayes rule (eq. 7) and M-step maximizes the resulting expression for $Q(.|\Phi^{(m)})$. In many cases this can be done by direct partial differentiation resulting in system of linear equation. The EM for segmentation problem can be summarized as follows:

Segmentation via EM

- **E-step**:

  Calculate the probability maps (soft segmentation) $\mathbf{p}^{(m+1)}$ given the observed voxel intensities $\mathbf{y}$ and parameter estimate $\Phi^{(m)}$:

$$p_{ij}^{(m+1)} = \frac{p(y_i|Z_i = e_j, \Phi^{(m)})p(Z_i = e_j|\Phi^{(m)})}{\sum_{k=1}^{K} p(y_i|Z_i = e_k, \Phi^{(m)})p(Z_i = e_k|\Phi^{(m)})}$$

- **M-step**:

  Estimate the parameters $\Phi^{(m+1)}$ based on probability maps $\mathbf{p}^{(m+1)}$ and the observed voxel intensities $\mathbf{y}$

$$\Phi^{(m+1)} = \underset{\Phi}{\operatorname{argmax}}\, Q(\Phi, \Phi^{(m)})$$

$$Q(\Phi, \Phi^{(m)}) = \sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij}^{(m+1)}(\log_e p(y_i|Z_i = e_j, \Phi) + \log_e p(Z_i = e_j|\Phi))$$

It is now easy to prove that the formal definition yealds the same equation for GMM as the direct differentiation of incomplete log-likelihood.

8 Incorporating the probabilistic brain atlas into EM segmentation

It is well known that EM algorithm is sensitive to initialisation of parameters $\Phi^{(0)}$ as it can only find the local extremum. Different initial parameter estimates will yeald different results. In case of brain MR images it is possible to estimate the initial parameters from the histogram. The other possibility is to use a probabilistic atlas.

When initializing the EM with probabilistic atlas, the atlas is usually aligned with the image data using the affine transformation. The atlas consists of four probability maps $P_0, ..., P_3$ for white matter, grey matter, csf and

brainmask. The atlas represents the prior spatial information about human brain strutcure common to all human brains. Let us denote p_{ij}^{atlas} the probability that i^{th} voxel belongs to j^{th} tissue class. We can initialize the EM in the E-step of the iteration zero with prior information from the atlas as follows:

$$p_{ij}^{0} = p_{ij}^{atlas}$$

After initialisation, the Gaussian mixture model can be used to calculate the segmentation.

However, the classic GMM does not produce satisfactory results for the brain segmentation. Even if non-brain tissues are extracted in pre-processing step and only 3 brain tissues are left in the image, the noise will severly affect the resulting segmentation. Therefore GMM works only on well-defined images with low level of noise. More robust algorithm can be obtained when the probabilistic atlas is used not only to initialise but also to spatially constrain the segmentation process. Consequently, the voxels are classified based not only on intensity but also on spatial position.

Van Leemput *et al.*([2], [3]) amends GMM by using the atlas as prior information at each iteration. The prior information is fixed and does not change with iterations:

$$p(Z_i = e_j|\Phi) = p_{ij}^{atlas}$$

With this assuption the resulting equation of EM algorithm become:

EM segmentation by Van Leemput

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{G(y_i, \mu_j^{(m)}, \sigma_j^{(m)}) p_{ij}^{atlas}}{\sum_{k=1}^{K} G(y_i, \mu_k^{(m)}, \sigma_k^{(m)}) p_{ik}^{atlas}}$$

- **M-step**:

$$\begin{aligned} \mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} y_i p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\ (\sigma_j^{(m+1)})^2 &= \frac{\sum_{i=1}^{n} (y_i - \mu_j^{(m+1)})^2 p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \end{aligned}$$

In GMM the mixture weights c_k are changed at each iteration to reflect the proportion of the image volume classified as the k^{th} tissue type. In Van Leemput's model these weights are fixed to values from atlas, but they vary with the position of the voxel. Ashburner *et al.* [4] combines both approaches:

EM segmentation by Ashburner (SPM)

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{G(y_i, \mu_j^{(m)}, \sigma_j^{(m)}) c_{ij}^{(m)}}{\sum_{k=1}^{K} G(y_i, \mu_k^{(m)}, \sigma_k^{(m)}) c_{ik}^{(m)}}$$

- **M-step**:

$$\begin{aligned}
\mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} y_i p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
(\sigma_j^{(m+1)})^2 &= \frac{\sum_{i=1}^{n} (y_i - \mu_j^{(m+1)})^2 p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
c_{ij}^{(m+1)} &= \frac{p_{ij}^{atlas} \sum_{l=1}^{n} p_{lj}^{(m+1)}}{\sum_{l=1}^{n} p_{lj}^{atlas}}
\end{aligned}$$

This method is implemented in widely used SPM software package [5].

9 Multichannel segmentation

Very often more than one modality of MRI brain image is available, typically T1-weighted, T2-weigted and PD images. When correctly aligned, multiple intensity values can enhance the segmentation process and reduce the impact of artefacts such as noise or bias field.

Let $\mathbf{y}_i = (y_{i1}, ..., y_{iR})$ be image intensities of R different modalities. Let us denote $\mu_k = (\mu_{k1}, ..., \mu_{kR})$ with μ_{kr} representing the mean intensity of k^{th} tissue in modality r and $\mathbf{\Sigma_k}$ the covariance matrix for all the modalities of k^{th} tissue class. We will assume multivariate Gaussian mixture PDF for the image, parallel with the single channel version:

$$G(\mathbf{y}_i, \mu_k, \mathbf{\Sigma}_k) = \frac{1}{\sqrt{2\pi}^R |\mathbf{\Sigma}_k|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(\mathbf{y}_i - \mu_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{y}_i - \mu_k)}$$

Van Leemput *et al.*([2], [3]) describes the multichannel version of their algorithm:

Multichannel EM segmentation by Van Leemput

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{G(\mathbf{y}_i, \mu_j^{(m)}, \mathbf{\Sigma}_j^{(m)}) p_{ij}^{atlas}}{\sum_{k=1}^{K} G(\mathbf{y}_i, \mu_k^{(m)}, \mathbf{\Sigma}_k^{(m)}) p_{ik}^{atlas}}$$

- **M-step**:

$$\begin{aligned}
\mu_{jr}^{(m+1)} &= \frac{\sum_{i=1}^{n} y_{ir} p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\
(\Sigma_j^{(m+1)})_{rs}^2 &= \frac{\sum_{i=1}^{n} (y_{ir} - \mu_{jr}^{(m+1)})(y_{is} - \mu_{js}^{(m+1)}) p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}}
\end{aligned}$$

10 Incorporating the bias field correction into EM

MR images are corrupted by a low-frequency spatially varying artifact known as the *bias field*. This is caused by equipment limitation and patient-induced electrodynamic interactions.

Let $I = (I_1, ..., I_n)$ be the observed intensities of the image and $I^* = (I_1^*, ..., I_n^*)$ the ideal intensities and $B = (B_1, ..., B_n)$ is the bias field. Then the degradation effect at each voxel can be expressed as

$$I_i = I_i^* B_i$$

Let $Y = (Y_1, ..., Y_n)$ and $Y^* = (Y_1^*, ..., Y_n^*)$ denote respectively the log transformed observed and ideal intensities. The logarithmic transformation changes the multiplicative bias field to additive:

$$Y_i = Y_i^* + B_i$$

We can therefore model PDF of the voxel intensity

$$p(y_i | Z_i = e_k, \Phi, B) = G(y_i - b_i, \mu_k, \sigma_k)$$

The low frequency charasteristics of the bias field B can be modeled by a linear combination of smooth basis functions $\Psi_l(x)$

$$b_i = \sum_{l=1}^{L} a_l \Psi_l(pos(i))$$

where $pos(i)$ denotes the 3D position of voxel i and $A = (a_1, ..., a_k)$ the bias field parameters. The basis functions can be polynomial, the model can represent the splines or any other smooth functions.

Parallel to deriving equations for GMM, the bias field can be estimated in EM framework by maximizing the incomplete log likelihood. Setting the partial derivations of incomplete log likelihood $p(Y|\Phi, A)$ for parameters μ_j, σ_j and a_l and using Van Leemput's model results in the following algorithm:

EM segmentation with bias field correction by Van Leemput

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{G(y_i - \sum_{l=1}^{L} a_l^{(m)} \Psi_l(pos(i)), \mu_j^{(m)}, \sigma_j^{(m)}) p_{ij}^{atlas}}{\sum_{k=1}^{K} G(y_i - \sum_{l=1}^{L} a_l^{(m)} \Psi_l(pos(i)), \mu_k^{(m)}, \sigma_k^{(m)}) p_{ik}^{atlas}}$$

- **M-step**:

1. **Gaussian distribution parameters estimation**

$$\begin{aligned} \mu_j^{(m+1)} &= \frac{\sum_{i=1}^{n} (y_i - \sum_{l=1}^{L} a_l^{(m)} \Psi_l(pos(i))) p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \\ (\sigma_j^{(m+1)})^2 &= \frac{\sum_{i=1}^{n} (y_i - \sum_{l=1}^{L} a_l^{(m)} \Psi_l(pos(i)) - \mu_j^{(m+1)})^2 p_{ij}^{(m+1)}}{\sum_{i=1}^{n} p_{ij}^{(m+1)}} \end{aligned}$$

2. **Bias correction**

$$(A^{(m+1)})^T = (F^T W^{(m+1)} F)^{-1} F^T W^{(m+1)} R^{(m+1)} \tag{12}$$

$$F = \begin{pmatrix} \Psi_1(pos(1)) & \Psi_2(pos(1)) & . & . & . & \Psi_L(pos(1)) \\ \Psi_1(pos(2)) & \Psi_2(pos(2)) & & & & \Psi_L(pos(2)) \\ . & & & . & & . \\ . & & & & . & . \\ . & & & & & . \\ \Psi_1(pos(N)) & \Psi_2(pos(N)) & . & . & . & \Psi_L(pos(N)) \end{pmatrix}$$

$$W^{(m+1)} = \begin{pmatrix} \sum_{k=1}^{K} w_{1k}^{(m+1)} & 0 & . & . & . & 0 \\ 0 & \sum_{k=1}^{K} w_{2k}^{(m+1)} & & & & 0 \\ . & & & . & & . \\ . & & & & . & . \\ . & & & & & . \\ 0 & 0 & . & . & . & \sum_{k=1}^{K} w_{Nk}^{(m+1)} \end{pmatrix}$$

$$w_{ik}^{(m+1)} = \frac{p_{ik}^{(m+1)}}{(\sigma_k^{(m+1)})^2}$$

$$\tilde{y}_i^{(m+1)} = \frac{\sum_{k=1}^{K} w_{ik}^{(m+1)} \mu_{ik}^{(m+1)}}{\sum_{k=1}^{K} w_{ik}^{(m+1)}}$$

$$R = \begin{pmatrix} y_1 & - & \tilde{y}_1^{(m+1)} \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ y_N & - & \tilde{y}_N^{(m+1)} \end{pmatrix}$$

The bias correction step can be interpreted as follows: The estimated soft segmentation and gaussian distribution parameters can be used to reconstruct the image estimate $\tilde{Y} = (\tilde{y_1}, ..., \tilde{y_N})$ which is not corrupted by the bias field. When subtracted from the observed image, the *residual image* R is calculated. From the residual image the bias field is estimated. The matrix F represents the discretized geometry of the bias field and W is an inverse covariance matrix.

11 Maximum a posteriori principle

The idea of estimating the bias field in the EM framework was originally proposed by Wells *et al* [6]. In his method the Gaussian distribution parameters are assumed to be known (they are estimated from the histogram in preprocessing step) and EM is only used to estimate the the bias field using the *maximum a posteriori* (MAP) principle instead of ML principle (eq. 5):

$$\hat{\Phi} = \underset{\Phi}{\operatorname{argmax}}\, p(\Phi|Y)$$

Bayes rule can be applied:

$$p(\Phi|Y) = \frac{p(Y|\Phi)p(\Phi)}{p(Y)}$$

As $p(Y)$ does not depend on Φ the MAP principle can be expressed as:

$$\hat{\Phi} = \underset{\Phi}{\operatorname{argmax}}\, p(\Phi|Y)p(\Phi)$$

The lower bound function for MAP estimation $Q_{MAP}(\Phi, \Phi^{(m)})$ can be expressed as:

$$Q_{MAP}(\Phi, \Phi^{(m)}) = Q(\Phi, \Phi^{(m)}) + \log p(\Phi)$$

In Wells' method [6] the parameters Φ directly represent the bias field $B = (b_1, ..., b_n)$. As prior term for for the parameters $\log p(\Phi)$ is included in the optimisation function, the n-dimensional zero mean Gaussian distribution

$$p(\Phi) = p(B) = G(B, \mathbf{0}, \Sigma_B)$$

can be assumed for the bias field and no parametric model is therefore needed. The equation for the bias field estimation step (eq. 12) will be then replaced by

$$(B^{(m+1)})^T = (W^{(m+1)} + \Sigma_B^{-1}))^{-1} W^{(m+1)} R^{(m+1)}$$

This equation differs from (eq. 12) by adding the smootness constraint Σ_B^{-1} and setting F to unit matrix as no parametric model for the bias field is assumed.

Let us further define the *mean residual image* $\bar{R}^{(m+1)}$

$$\bar{R}^{(m+1)} = W^{(m+1)} R^{(m+1)}$$

which means that

$$\bar{R}_i^{(m+1)} = \sum_{k=1}^{K} \frac{p_{ik}^{(m+1)} (y_i - \mu_k^{(m+1)})}{\sigma_k^{(m+1)}}$$

The bias field estimation step can be then simplified to

$$(B^{(m+1)})^T = H \bar{R}^{(m+1)}$$

where H is a linear operator. In practise the linear operator H can be approximated by a linear low-pass filter. Wells uses the following efficient implementation

$$b_i^{(m+1)} = \frac{[F \bar{R}^{(m+1)}]_i}{[F W^{(m+1)} \mathbf{1}]_i}$$

where F is a low-pass filter and $\mathbf{1} = (1, ..., 1)^T$.

12 Including neighbourhood information in EM

The impact of noise on the resulting image can be reduced by including the neigbourhood voxels information instead of single voxel intensity. Zhang *et*

al. [7] model the set of labels and voxel intensities as Markov random fields (MRF). This model prefers homogeneous neighbourhoods (neighbouring voxels are likely to be of the same tissue class) and therefore encourages smooth homogeneous areas and suppresses noise which is consistent with the connected structure of the brain tissues. However, incorporating Markov Random Fields into the EM framework is more complicated and computationally expensive. On the other hand their method does not require a probabilistic atlas. The segmentation process is initialised using a discriminant measure based tresholding method.

Van Leemput also includes the MRF regularisation step. In addition to noise-reducing effect, the knowledge of the neigbouring labels can help reduce errors in classification - such as voxel surrounded by non-brain tissues should not be classified as grey matter. The implementation of this step is quite complicated and according to our experience the improvement in Van Lemput's model is minimal and therefore we will omit the equations in this report.

13 Partial volume effect and other extensions

Most EM-based methods assume the Gaussian distribution for tissue probability distributions as it is easy to incorporate into EM framework and reasonably describes the real tissue intensity distributions. However, this assumption results in some misclassification as the real distribution is not strictly Gaussian. This is most pronounced in CSF distribution as very large boundary with GM results in distribution distortion thanks to partial volume effect, see Fig. **??**. Wells *et al.* [6] suggest to estimate the real distributions instead of the means and variances of the Gaussian distributions. This extended method is not an instance of EM algorithm but it is still reported to be robust in practice. Some recent studies propose to improve the segmentation results by including a model for the partial volume effect in the EM framework [8].

Rajapakse *et al.*[9] estimates intensity distributions locally to compensate natural tissue intensity variation.

Inevitable misclassifications resulting from overlaps in tissue intensity distributions can be partially avoided by spatially constraining the segmentation process with prior information given by an aligned probabilistic atlas at each iteration as suggested by Van Leemput *et al.* ([2], [3]); however this makes the methods very sensitive to correct alignment of the atlas with the image. This is almost impossible for subjects with distorted anatomy or child brains which significantly differ from adult brains [10]. To solve this problem, [11]

and [12] suggest to warp the atlas to subject by non-rigid registration. In [13] Pohl reports difficulties with non-rigidly registering the template to image which is consistent with our own experience. Recently, researchers tried to overcome the problem of correct alignment of the atlas with image by iteratively warping a deformable atlas (a subject with attached labeling) and refining the segmentation at the same time [14], [**?**].

14 Simultaneous segmentation and registration

The most recent trend in brain MRI segmentation is towards simultaneous registration and segmentation of MR images. Segmentation and registration are often complementary in succeeding or failing in certain areas of the brain. Recent methods incorporate both processes into one probabilistic framework [14], [13], [15] which are extensions of the segmentation algorithms by Ashburner *et al* [16], Wells *et al* [6] and Zhang *et al* [7].

Including the registration process in the EM framework results in additional registration estimation step as part of M-step. As an example we include the equations for the extension of Well's MAP segmentation developed by Pohl *et al* [13]. In this method the Gaussian distribution parameters are assumed to be known and parameters $\Phi = (\beta, \alpha)$ consist of bias field parameters β and registration parameters α.

Simultaneous segmentation and registration by Pohl

- **E-step**:

$$p_{ij}^{(m+1)} = \frac{p(y_i|Z_i = e_j, \beta^{(m)})p(Z_i = e_j|\alpha^{(m)})}{\sum_{k=1}^{K} p(y_i|Z_i = e_k, \beta^{(m)})p(Z_i = e_k|\alpha^{(m)})}$$

- **M-step**:

$$\alpha^{(m+1)} = \underset{\alpha}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij}^{(m+1)} \log_e p(Z_i = e_j|\alpha) + log_e p(\alpha)$$

$$\beta^{(m+1)} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij}^{(m+1)} (\log_e p(y_i|Z_i = e_j, \beta) + \log_e p(\beta))$$

The model results in two separate estimation steps in M-step. For the bias correction step the methods mentioned above can be used. The registration parameters can represent any kind of registration - rigid, affine or non-rigid. To find them, very often optimisation methods have to be employed as it is usually not possible to find closed expression for the registration estimation step.

15 Summary of the EM approach

We have shown that EM algorithm is a very flexible optimisation framework suitable for the problem of brain segmentation. The underlying model can be specified according the specific requirements of the given task. The approach is feasable as the M-step can be divided into several separate estimation steps which can be combined as needed. We can summarize the EM approach as follows:

EM approach for brain segmentation

- **E-step**: Estimate the soft segmentation given the current estimate of parameters. This may include using neighbourhood statistics such as MRF as well as partial volume estimation.
- **M-step**: Estimate the parameters which can consist of a comination of the following steps:
  1. Estimate the intensity disribution parameters for each tissue class
  2. Estimate the bias correction parameters
  3. Estimate the registration parameters

References

[1] T. Minka. "Expectation-maximization as lower bound maximization.", 1998.

[2] K. V. Leemput, F. Maes, D. Vandermeulen et al. "Automated model-based tissue classification of MR images of the brain." *IEEE Transactions on Medical Imaging* **18(10)**, pp. 897–908, 1999.

[3] K. V. Leemput, F. Maes, D. Vandermeulen et al. "Automated model-based bias field correction of MR images of the brain." *IEEE Transactions on Medical Imaging* **18(10)**, pp. 885–896, 1999.

[4] R. Frackowiak, K. Friston, C. Frith et al. *Human Brain Function*. Academic Press, second edition, 2003.

[5] Statistical Parametric Mapping. www.fil.ion.ucl.ac.uk/spm.

[6] W. M. Wells III, W. E. L. Grimson, R. Kikinis et al. "Adaptive segmentation of MRI data." *IEEE Transactions on Medical Imaging* **15(4)**, pp. 429–442, 1996.

[7] Y. Zhang, M. Brady & S. Smith. "Segmentation of brain MR images through a hidden markov random field model and the expectation maximization algorithm." *IEEE Transactions on Medical Imaging* **20(1)**, pp. 45–57, 2001.

[8] N. Joshi & M. Brady. "A non-parametric mixture model for partial volume segmentation of MR images." In *British Machine Vision Conference*. 2005.

[9] J. L. R. J. C. Rajapakse, J. N. Giedd. "Statistical Approach to Segmentation of Single Channel Cerebral MR Images." *Medical Imaging, IEEE Transactions on* **16(2)**, pp. 176–186, 1997.

[10] M. Wilke, V. Schmithorst & S. Holland. "Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data." *Magnetic Resonance in Medicine* **50(4)**, pp. 749–757, 2003.

[11] K. Pohl, W. Wells, A. Guimond et al. "Incorporating non-rigid registration into expectation maximization algorithm to segment MR images." In *Proc. of the 5th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*, volume 2488 of *Lecture Notes in Computer Science*, pp. 564–572. 2002.

[12] E. D'Agostino, F. Maes, D. Vandermeulen et al. "Non-rigid atlas-to-image registration by minimization of class-conditional image entropy." In *Proc. of the 7th Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, Part I*, pp. 745–753. 2004.

[13] K. Pohl. *Prior Information for Brain Parcellation*. Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[14] J. Ashburner & K. Friston. "Unified segmentation." *NeuroImage* **26**, pp. 839–851, 2005.

[15] C. Xiaohua. *Simultaneous Segmentation and Registration of Medical Images*. Ph.D. thesis, University of Oxford, 2005.

[16] J. Ashburner. *Computational Neuroanatomy.* Ph.D. thesis, University College London, 2000.