

# Segmenting Magnetic Resonance Images via Hierarchical Mixture Modelling

Carey E. Priebe  
<*cep@jhu.edu*>

Center for Imaging Science  
&  
Department of Mathematical Sciences  
Johns Hopkins University

University of Florida  
January, 2004

# Collaborators

Michael I. Miller & J. Tilak Ratnanather, JHU CIS

Washington University School of Medicine, St. Louis

Office of  
Naval Research  
800 N. Quincy Street, Arlington, VA 22217-5660



<http://www.mts.jhu.edu/~priebe/akm4mri.html>

# Outline

## Segmenting Magnetic Resonance Images via Hierarchical Mixture Modelling

# Outline

## Segmenting Magnetic Resonance Images via Hierarchical Mixture Modelling

- MR Imaging  
& Cingulate Gyrus Segmentation

# Outline

## Segmenting Magnetic Resonance Images via Hierarchical Mixture Modelling

- MR Imaging  
& Cingulate Gyrus Segmentation
- Wash. U. Brain Data

# Outline

## Segmenting Magnetic Resonance Images via Hierarchical Mixture Modelling

- MR Imaging  
& Cingulate Gyrus Segmentation
- Wash. U. Brain Data
- Mixture Estimation

# Outline

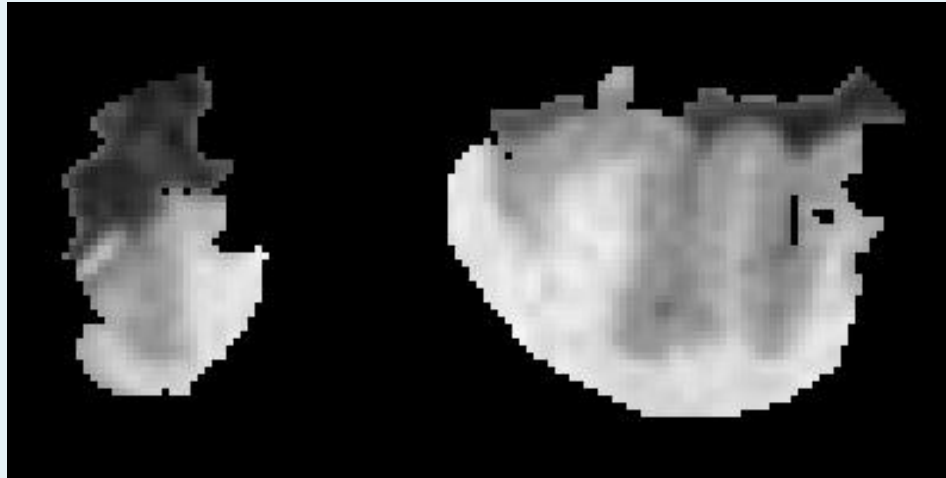
## Segmenting Magnetic Resonance Images via Hierarchical Mixture Modelling

- MR Imaging  
& Cingulate Gyrus Segmentation
- Wash. U. Brain Data
- Mixture Estimation
- Segmentation Results

# Cingulate Gyrus

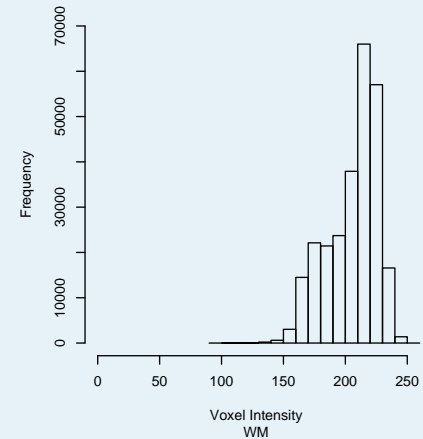
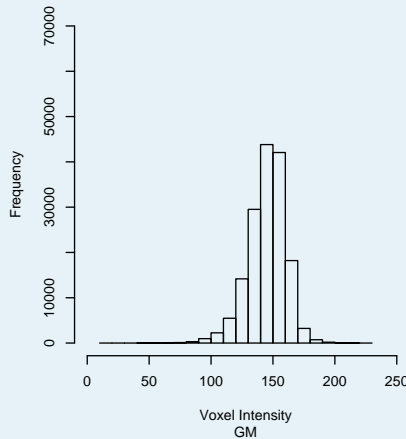
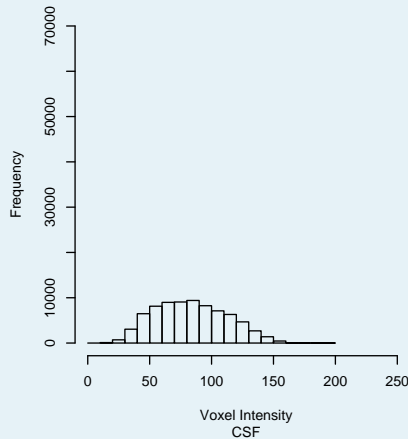
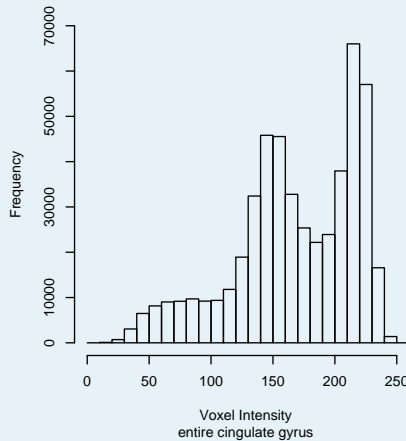


# Hand-segmentation



Original MRI (top) and hand-segmentation (bottom).  
CSF (dark gray), GM (light gray), WM (white); black is unsegmented.

# Histograms



Frequency histograms for s2006:  
entire cingulate gyrus, right hemisphere (top), and  
CSF (bottom left), GM (bottom center), and WM (bottom right).

# Table 1

Sample sizes for 10 cingulate gyri:  
Number of voxels

<b>SZ</b>	s1002	s1003	s1009	s1010	s1013
<i>CSF</i>	150128	90035	91452	85063	121138
<i>GM</i>	210636	186529	136856	124227	130717
<i>WM</i>	333341	359746	263401	243651	274943
<i>total</i>	694105	636310	491709	452941	526798

<b>NV</b>	s2002	s2003	s2004	s2006	s2007
<i>CSF</i>	45709	108617	84402	76886	125409
<i>GM</i>	154993	181538	169583	161138	131230
<i>WM</i>	193633	320293	276141	264607	339410
<i>total</i>	394335	610448	530126	502631	596049

# Methodology Overview

# Methodology Overview

- (1) For each subject/class pair in the available training data set, estimate the marginal subject-specific class-conditional probability densities. Notice that for this supervised step (since these are training images, individual voxel class labels are available) semiparametric mixture complexity estimation is appropriate.

# Methodology Overview

- (1) For each subject/class pair in the available training data set, estimate the marginal subject-specific class-conditional probability densities. Notice that for this supervised step (since these are training images, individual voxel class labels are available) semiparametric mixture complexity estimation is appropriate.
- (2) Find the “closest” training model to the (unlabeled) test data.

# Methodology Overview

- (1) For each subject/class pair in the available training data set, estimate the marginal subject-specific class-conditional probability densities. Notice that for this supervised step (since these are training images, individual voxel class labels are available) semiparametric mixture complexity estimation is appropriate.
- (2) Find the “closest” training model to the (unlabeled) test data.
- (3) Fit a mixture to the test data, using the training model obtained in step (2) to determine the class-conditional mixture complexities and starting locations.

# Methodology Overview

- (1) For each subject/class pair in the available training data set, estimate the marginal subject-specific class-conditional probability densities. Notice that for this supervised step (since these are training images, individual voxel class labels are available) semiparametric mixture complexity estimation is appropriate.
- (2) Find the “closest” training model to the (unlabeled) test data.
- (3) Fit a mixture to the test data, using the training model obtained in step (2) to determine the class-conditional mixture complexities and starting locations.
- (4) Classify voxels from the test data using the plug-in Bayes rule, where the mixture component class labels are inherited from the selected training model but the mixture itself is estimated from the test data.

Priebe, C.E., and Marchette, D.J. (2000),  
“Alternating Kernel and Mixture Density Estimates,”  
*Computational Statistics and Data Analysis*, 35, 43–65.

James, L.F., Priebe, C.E., and Marchette, D.J. (2001),  
“Consistent Estimation of Mixture Complexity,”  
*The Annals of Statistics*, 29, 1281–1296.

# Modelling(a)

For each subject  $j \in \mathcal{J} := \{1, \dots, J\}$ , consider magnetic resonance voxel observations  $\mathcal{X}_j := \{X_{j1}, \dots, X_{jn_j}\}$ ; the subject-specific sample sizes are denoted by  $n_j$ . Let the marginal probability density function for subject  $j$  voxel observations be given by

$$f_j = \sum_{c \in \mathcal{C}} \pi_{jc} f_{jc}.$$

That is,  $f_j = \pi_{jC} f_{jC} + \pi_{jG} f_{jG} + \pi_{jW} f_{jW}$ .

The subject-specific class-conditional marginals are denoted  $f_{jc}$ , and the subject-specific class-conditional mixing coefficients  $\pi_{jc}$  are nonnegative and sum to unity.

We assume that the  $X_{ji}$  are identically distributed according to  $f_j$ , but not independent.

## Modelling(b)

Each subject-specific class-conditional marginal  $f_{jc}$  is itself modelled as a mixture of normals;

$$f_{jc} = \sum_{t=1}^{k_{jc}} \pi_{jct} \varphi_{jct}.$$

Thus the subject-specific marginals  $f_j$  are modelled as hierarchical mixtures – mixtures of Gaussian mixtures;

$$f_j = \sum_{c \in \mathcal{C}} \pi_{jc} \sum_{t=1}^{k_{jc}} \pi_{jct} \varphi_{jct}.$$

The subject-specific class-conditional mixture complexities  $k_{jc}$  are to be estimated from the data.

## Estimation(a)

Given class-labelled training data, we have available subject-specific class-conditional sample sizes  $n_{jc}$  such that  $n_j = \sum_c n_{jc}$ .

For the subject-specific mixing coefficients  $\pi_{jc}$  we use the empirical estimate

$$\hat{\pi}_{jc} = n_{jc}/n_j,$$

the ratio of the subject-specific class-conditional sample size to the total subject-specific sample size.

Let  $\mathcal{X}_{jc}$  denote that subset of  $\mathcal{X}_j$  for which the class label is  $c$ .

## Estimation(b)

We estimate the subject-specific class-conditional mixture complexities  $k_{jc}$ , mixing coefficients  $\pi_{jct}$ , and mixture components  $\varphi_{jct}$ , via AKM. This estimation is semiparametric; the mixture complexities  $\hat{k}_{jc}$  are estimated from the data.

AKM employs an iterative estimation scheme, comparing the  $k$ -component mixture estimate against the  $k + 1$ -component mixture estimate. When the improvement obtained by adding a  $k + 1$ st component is negligible (less than some penalty term) the iteration is terminated and the resulting  $k$ -component mixture is used as the estimate. This general version of model selection – looking for the “elbow” or “knee” in a complexity vs. penalized estimation accuracy curve – is quite common. A distinguishing feature of AKM is the process of using successive kernel estimates to guide the successive mixture estimates.

## Estimation(c)

The filtered kernel estimator extends the basic kernel estimator by allowing multiple bandwidths driven by a pilot mixture model.

Given a Gaussian mixture

$$f = \sum_{t=1}^k \pi_t \varphi_t$$

and a bandwidth  $h$ , define the filtered kernel estimator  $\tilde{f}(\cdot; \mathcal{X}, f, h)$  based on the mixture  $f$  and using the data  $\mathcal{X} = \{X_1, \dots, X_n\}$  to be

$$\tilde{f}(x; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^k \frac{\pi_t \varphi_t(X_i)}{f(X_i) h \sigma_t} \varphi_0\left(\frac{x - X_i}{h \sigma_t}\right)$$

where  $\sigma_t^2$  is the variance of the  $t$ th component of the mixture  $f$  and  $\varphi_0$  is the standard zero mean, unit variance normal.

This allows for different bandwidths, guided by the mixture model  $f$ .

## Estimation(d)

Let  $\hat{f}^1$  be the single normal component with mean and variance given by the sample moments of  $\mathcal{X}_{j_c}$  – that is,  $\hat{f}^1$  is the trivial one-component mixture. Let  $\tilde{f}^1$  be the filtered kernel estimate based on the mixture  $\hat{f}^1$  – that is,  $\tilde{f}^1$  is the standard kernel estimate using the normal reference rule to determine the bandwidth. For  $k = 2, 3, \dots$ , define in turn first  $\hat{f}^k$  to be the  $k$ -component mixture best matched to the nonparametric estimate  $\tilde{f}^{k-1}$ ;

$$\hat{f}^k := \arg \min_{f \in \mathcal{F}^k} \|f - \tilde{f}^{k-1}\|_2^2,$$

where  $\|f - g\|_2^2 := \int_{-\infty}^{\infty} (f(x) - g(x))^2 dx$  is the integrated squared error and  $\mathcal{F}^k$  is the class of  $k$ -component Gaussian mixtures.

Subsequently define  $\tilde{f}^k$  to be the filtered kernel estimate based on the mixture  $\hat{f}^k$ .

# Estimation(e)

Let

$$\ell(f, \mathcal{X}) := \log \prod_{x \in \mathcal{X}} f(x) = \sum_{x \in \mathcal{X}} \log f(x).$$

The estimate of mixture complexity is given by

$$\hat{k}_{j_c} = \arg \min \{k \in \{1, 2, \dots\} : \ell(\hat{f}^{k+1}, \mathcal{X}_{j_c}) - \ell(\hat{f}^k, \mathcal{X}_{j_c}) < a(n_{j_c}, k + 1)\}.$$

The penalty term  $a(n_{j_c}, k + 1)$  in the above equation – a function of sample size and model complexity – is *the* key practical issue in this version of model selection.

## Spatial Dependence(a)

Because there is correlation amongst the voxel observations and thus  $\ell(\cdot, \cdot)$  defined above is not the log-likelihood, simple interpretation of the penalty term  $a(n_{jc}, k + 1)$  as a function of sample size calls for conditional covariance modelling of the three-dimensional spatial process.

Covariograms for the subject-specific class-conditional random fields indicate that significant spatial correlation is evident at a distance of up to ten voxels.

(Isotropy is indicated as well.)

This suggests that the effective sample sizes are significantly smaller than the number of voxels  $n_{jc}$ .

## Spatial Dependence(b)

One way to look at the effect of dependent observations on our estimate of mixture complexity is to note that, if the observations in the subject-specific class-conditional sample  $\mathcal{X}_{jc}$  are positively correlated, then the terms  $\ell(\hat{f}^{k+1}, \mathcal{X}_{jc})$  and  $\ell(\hat{f}^k, \mathcal{X}_{jc})$  in the mixture complexity estimation equation should be scaled accordingly. That is, if the “effective number of independent observations” is  $n'_{jc} := n_{jc}/\delta$  for  $1 \leq \delta \leq n_{jc}$ , then substituting  $\delta^{-1}\ell(\cdot, \cdot)$  into the mixture complexity estimation equation will account for the correlation. The case  $\delta = 1$  represents independence and, as (positive) spatial correlation increases,  $\delta$  increases as well. This can be presented as

$$\hat{k}_{jc} = \arg \min \{k \in \{1, 2, \dots\} : \ell(\hat{f}^{k+1}, \mathcal{X}_{jc}) - \ell(\hat{f}^k, \mathcal{X}_{jc}) < \delta a(n_{jc}/\delta, k+1)\}.$$

# Spatial Dependence(c)

Estimating  $\delta$  is a sticky wicket.

Furthermore, Occam's Razor may suggest a preference to err on the side of parsimony.

Nevertheless, as will be seen in the experimental results presented herein, a significant performance improvement can be obtained by employing even a crude covariogram-based estimate  $\hat{\delta}$  in the complexity selection methodology.

Our rule of thumb is to choose  $\hat{\delta}$  to be the number of voxels in a ball whose radius is given by the distance at which the omnidirectional empirical covariogram drops below some threshold, rounded to the nearest integer.

## Spatial Dependence(d)

Using  $a(n_{jc}, k + 1) = 3 \log(n)$ , we arrive at the estimate of mixture complexity

$$\hat{k}_{jc} = \arg \min \{k \in \{1, 2, \dots\} : \ell(\hat{f}^{k+1}, \mathcal{X}_{jc}) - \ell(\hat{f}^k, \mathcal{X}_{jc}) < \hat{\delta} 3 \log(n_{jc}/\hat{\delta})\}.$$

The mixture model estimate of the marginal probability density for subject  $j$ , tissue class  $c$ , is given by

$$\hat{f}_{jc} = \hat{f}^{\hat{k}_{jc}} = \arg \min_{f \in \mathcal{F}^{\hat{k}_{jc}}} \|f - \hat{f}^{\hat{k}_{jc}-1}\|_2^2,$$

the  $\hat{k}_{jc}$ -component mixture identified in the AKM procedure.

Results will be reported below for three choices of  $\hat{\delta}$ : 1, 25, and 99.

## Experiment: Training(a)

We consider segmenting the cingulate gyrus for subject  $\gamma \in \mathcal{J}$ .

In this case, we have  $J - 1$  training subjects corresponding to indices  $j \in \mathcal{J}, j \neq \gamma$ .

For  $j \in \mathcal{J} \setminus \{\gamma\}$ , we obtain

$$\hat{f}_j = \sum_{c \in \mathcal{C}} \frac{n_{jc}}{n_j} \hat{f}_{jc} = \sum_{c \in \mathcal{C}} \frac{n_{jc}}{n_j} \sum_{t=1}^{\hat{k}_{jc}} \hat{\pi}_{jct} \hat{\varphi}_{jct}$$

via AKM.

For  $\gamma$  we have the test data set  $\mathcal{X}_\gamma$ .

## Experiment: Training(b)

Write

$$\mathcal{L}(\gamma, j) := \prod_{x \in \mathcal{X}_\gamma} \hat{f}_j(x).$$

Choose the index  $\gamma^*$  specifying the training subject “closest” to the test data set;

$$\gamma^* := \arg \max_{j \in \mathcal{J} \setminus \{\gamma\}} \mathcal{L}(\gamma, j).$$

This provides us with class-conditional complexity estimates  $\hat{k}_{\gamma^*c}$ , as well as initial conditions obtained from the  $\hat{f}_{\gamma^*c}$  and  $n_{\gamma^*c}/n_{\gamma^*}$ , for use in modelling  $\mathcal{X}_\gamma$ .

## Experiment: Training(c)

We can view this procedure as utilizing a type nearest neighbor classifier to determine class-conditional complexity estimates and initial conditions for the unsupervised modelling of test subject  $\gamma$ .

We find the (training) subject “closest” to the (test) observation, and use this to guide the segmentation.

This provides a hedge against the oversmoothing of the model that might result from combining all the observations into a single model.

(One can, of course, use something other than the pseudo-likelihood for determining the “closest” training subject  $\gamma^*$ .)

## Experiment: Testing I(a)

Given  $\gamma^*$  selected as above, we estimate  $f_\gamma$  by estimating the parameters

$$\theta = \left[ \begin{array}{l} \pi_{\gamma C}, \pi_{\gamma G}, \\ \pi_{\gamma C1}, \dots, \pi_{\gamma C(\hat{k}_{\gamma C}-1)}, \mu_{\gamma C1}, \sigma_{\gamma C1}^2, \dots, \mu_{\gamma C\hat{k}_{\gamma C}}, \sigma_{\gamma C\hat{k}_{\gamma C}}^2, \\ \pi_{\gamma G1}, \dots, \pi_{\gamma G(\hat{k}_{\gamma G}-1)}, \mu_{\gamma G1}, \sigma_{\gamma G1}^2, \dots, \mu_{\gamma G\hat{k}_{\gamma G}}, \sigma_{\gamma G\hat{k}_{\gamma G}}^2, \\ \pi_{\gamma W1}, \dots, \pi_{\gamma W(\hat{k}_{\gamma W}-1)}, \mu_{\gamma W1}, \sigma_{\gamma W1}^2, \dots, \mu_{\gamma W\hat{k}_{\gamma W}}, \sigma_{\gamma W\hat{k}_{\gamma W}}^2 \end{array} \right]'$$

via

$$\hat{f}_\gamma := \arg \max_{x \in \mathcal{X}_\gamma} \prod_{c \in \mathcal{C}} \sum_{t=1}^{\hat{k}_{\gamma c}} \pi_{\gamma ct} \varphi_{\gamma ct}(x).$$

Notice that this involves estimation of  $3 \sum_{c \in \mathcal{C}} \hat{k}_{\gamma^* c} - 1$  parameters.

## Experiment: Testing I(b)

Algorithmically, this  $M$ -estimate is obtained using the EM algorithm with the training model  $\hat{f}_{\gamma^*}$  as the starting point;

$$\hat{f}_{\gamma} := EM(\mathcal{X}_{\gamma}; \hat{f}_{\gamma^*}).$$

The estimate obtained thusly can be written as

$$\hat{f}_{\gamma} = \sum_{c \in \mathcal{C}} \hat{\pi}_{\gamma c} \hat{f}_{\gamma c} = \sum_{c \in \mathcal{C}} \hat{\pi}_{\gamma c} \sum_{t=1}^{\hat{k}_{\gamma c}} \hat{\pi}_{\gamma ct} \hat{\varphi}_{\gamma ct}.$$

Class complexities and component labels are inherited from  $\hat{f}_{\gamma^*}$ , so that (for  $c = G$ , for instance) the mixing coefficient  $\hat{\pi}_{\gamma G}$  indicates the amount (proportion) of gray matter in the test image while  $\hat{f}_{\gamma G} = \sum_{t=1}^{\hat{k}_{\gamma G}} \hat{\pi}_{\gamma Gt} \hat{\varphi}_{\gamma Gt}$  provides a model for that gray matter.

## Experiment: Testing II(a)

We next consider the Bayes plug-in classifier

$$g(x) = \arg \max_{c \in \mathcal{C}} \hat{\pi}_{\gamma c} \hat{f}_{\gamma c}(x).$$

The voxel  $x$  is to be labelled as belonging to the class which maximizes posterior probability of class membership.

## Experiment: Testing II(b)

The Bayes plug-in classifier is operationally equivalent to the likelihood ratio test procedure given by considering

$$LRT_{\mathbf{C}/\mathbf{G}}(x) = \frac{\hat{\pi}_{\gamma\mathbf{C}} \hat{f}_{\gamma\mathbf{C}}(x)}{\hat{\pi}_{\gamma\mathbf{G}} \hat{f}_{\gamma\mathbf{G}}(x)} =: r_1(x)$$

and

$$LRT_{\mathbf{G}/\mathbf{W}}(x) = \frac{\hat{\pi}_{\gamma\mathbf{G}} \hat{f}_{\gamma\mathbf{G}}(x)}{\hat{\pi}_{\gamma\mathbf{W}} \hat{f}_{\gamma\mathbf{W}}(x)} =: r_2(x).$$

## Experiment: Testing II(c)

Our automatic segmentation is then given by the following rules:

- $r_1(x) > 1$  implies voxel  $x$  is to be labelled as CSF.
- $r_2(x) < 1$  implies voxel  $x$  is to be labelled as WM.
- $r_1(x) < 1$  &  $r_2(x) > 1$  implies voxel  $x$  is to be labelled as GM.
- $r_1(x) > 1$  &  $r_2(x) < 1$  should not occur (CSF  $<^{st}$  GM  $<^{st}$  WM).

## Table 2(a)

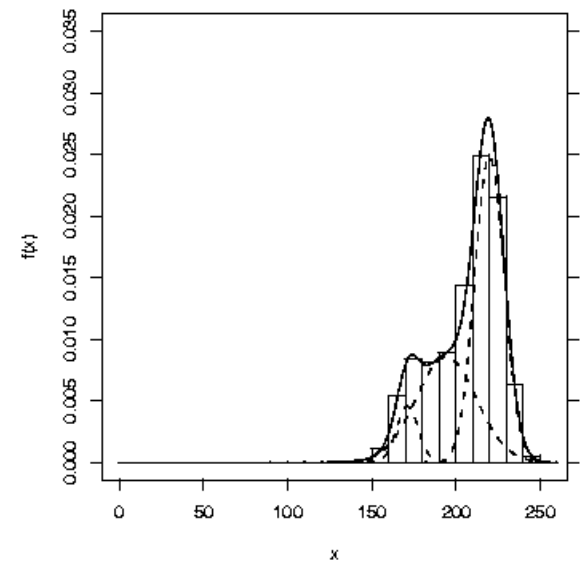
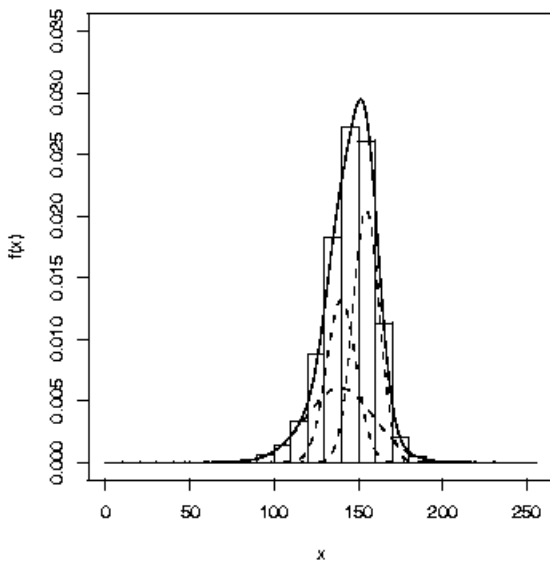
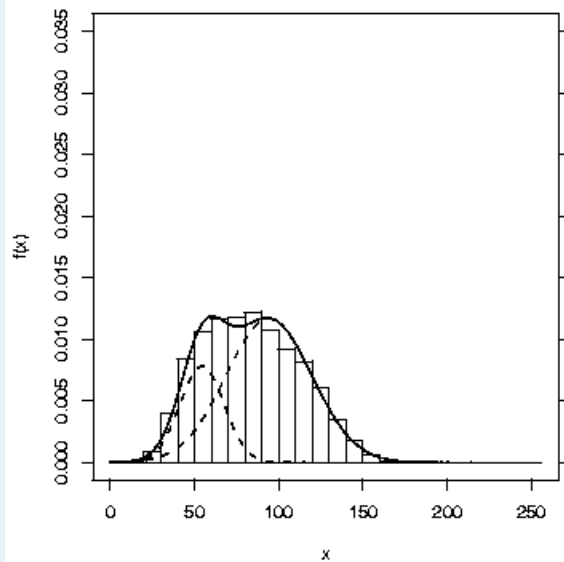
Mixture complexity estimation results for 10 cingulate gyri:  
Estimated number of components

SZ	s1002	s1003	s1009	s1010	s1013
$\hat{\delta} = 1$					
<i>CSF</i>	3	1	4	3	3
<i>GM</i>	2	16	3	1	3
<i>WM</i>	3	13	5	16	17
$\hat{\delta} = 25$					
<i>CSF</i>	3	1	2	3	3
<i>GM</i>	1	2	3	1	2
<i>WM</i>	3	2	3	4	3
$\hat{\delta} = 99$					
<i>CSF</i>	1	1	1	1	1
<i>GM</i>	1	1	2	1	2
<i>WM</i>	1	2	3	3	3

## Table 2(b)

Mixture complexity estimation results for 10 cingulate gyri:  
Estimated number of components

NV	s2002	s2003	s2004	s2006	s2007
$\hat{\delta} = 1$					
<i>CSF</i>	2	26	3	2	2
<i>GM</i>	1	1	1	3	3
<i>WM</i>	13	6	16	12	11
$\hat{\delta} = 25$					
<i>CSF</i>	1	2	2	2	2
<i>GM</i>	1	1	1	3	3
<i>WM</i>	2	3	4	3	5
$\hat{\delta} = 99$					
<i>CSF</i>	1	2	1	1	1
<i>GM</i>	1	1	1	1	2
<i>WM</i>	2	2	2	3	3



Mixtures ( $\hat{\delta} = 25$ ) and histograms for *s2006*.

From left to right: CSF, GM, WM.

Estimated model complexities are:  $\hat{k}_{CSF} = 2$ ,  $\hat{k}_{GM} = 3$ ,  $\hat{k}_{WM} = 3$ .

## Table 3

Segmentation results for 10 cingulate gyri:  
Probability of misclassification

<b>SZ</b>	s1002	s1003	s1009	s1010	s1013
<i>PV1</i>	0.15	0.09	0.23	0.20	0.28
<i>PV2</i>	0.09	0.06	0.13	0.11	0.18
<i>AKM</i> ( $\hat{\delta} = 25$ )	0.13	0.07	0.12	0.09	0.11
<i>AKM</i> ( $\hat{\delta} = 99$ )	0.10	0.07	0.13	0.10	0.15

<b>NV</b>	s2002	s2003	s2004	<b>s2006</b>	s2007
<i>PV1</i>	0.15	0.15	0.16	<b>0.20</b>	0.24
<i>PV2</i>	0.09	0.09	0.11	<b>0.12</b>	0.15
<i>AKM</i> ( $\hat{\delta} = 25$ )	0.10	0.07	0.10	<b>0.09</b>	0.11
<i>AKM</i> ( $\hat{\delta} = 99$ )	0.09	0.07	0.10	<b>0.08</b>	0.10

## Kronecker Quote

*“The wealth of your practical experience  
with sane and interesting problems  
will give to mathematics  
a new direction and a new impetus.”*

*— Leopold Kronecker to Hermann von Helmholtz*