# Multivariate Robust clustering via mixture models

Florence Forbes, Darren Wraith, Senan Doyle, Eric Frichot

INRIA-LJK équipe MISTIS (Grenoble)

# Model-based multivariate clustering

## Model-based clustering: Mixture of gaussians

$$f(y_i) = \sum_{k=1}^{K} \pi_k f_k(y_i;\theta_k)$$
$$\text{With } f_k(y_i;\theta_k) \sim \mathcal{N}(\mu_k, \Sigma_k)$$

- Univariate and multivariate
- Decomposition of the covariance matrix for flexibility in shape, volume and orientation (Banfield and Raftery 93, Celeux and Govaert 95)

$$\Sigma_k = \lambda_k \; D_k \; A_k D_k^T$$
$$\text{with } \lambda_k : \text{volume}$$
$$D_k : \text{orientation}$$
$$A_k : \text{shape}$$

Convenient computational tractability:

EM algorithm

+ additional minimization algorithm for some decompositions (see Celeux, Govaert 95)

# Robust clustering

In some applications:

- The tails of the normal distributions are shorter than appropriate or
- Parameter estimations are affected by atypical observations (outliers)

➡ Fit a mixture of t-distribution (Student distribution) $f_k(y_i;\theta_k) \sim t(y_i;\mu_k,\Sigma_k,\nu_k)$

- Univariate and multivariate
- Additional degree of freedom (dof) parameter $\nu \to \infty \Rightarrow f_k \to \mathcal{N}$ distribution

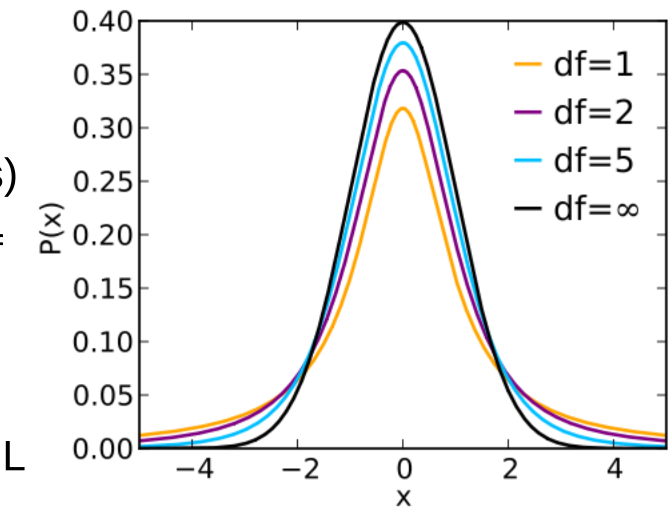
The dof can be seen as a robustness tuning parameter

Convenient computational tractability:

EM algorithm with additional missing variables (class+ weights)

+ additional numerical procedure for the ML estimate of the dof (see MacLachlan and Peel 2000)

In contrast to the Gaussian case, no closed-form solution for ML




But a useful representation of the t-distribution as an infinite mixture of scaled Gaussians

# Scaled mixture of Gaussians

The EM algorithm for t-mixtures is based on the following construction of the t-distribution

$$t(y;\mu,\Sigma,\nu) = \frac{\Gamma((M+\nu)/2)}{\Gamma(\nu/2)\ (\nu\pi)^{M/2}}\ |\Sigma|^{-1/2}\ \left[1+\delta^2/\nu\right]^{-(\nu+M)/2}$$

with $\delta^2 = (y-\mu)^T\Sigma^{-1}(y-\mu)$ the squared Mahalanobis distance

$M$: dimensionality of $y$
$\Gamma$: Gamma function

➡ $$t(y\ \mu,\Sigma,\nu) = \int_0^\infty \mathcal{N}(y;\mu,\Sigma/w)\ \mathcal{G}(w;\nu/2,\nu/2)\ dw$$

Another construction (equivalent):

$$X \sim \mathcal{N}(0,\Sigma) \text{ and } V \sim \mathcal{X}^2(\nu) = \mathcal{G}(\nu/2, 1/2)$$

$$Y = X \times \sqrt{(\tfrac{\nu}{V})} + \mu \quad \sim \quad t(\mu,\Sigma,\nu)$$

$$\frac{V}{\nu} \sim \mathcal{G}(\nu/2,\nu/2)$$

# EM algorithm for t-mixtures (MoT)

- Observations: $\mathbf{y} = \{\mathbf{y}_1 \dots \mathbf{y}_N\}$ where $\mathbf{y}_i = \{y_{i1} \dots y_{iM}\}$
- Missing data: $\mathbf{z} = \{\mathbf{z}_1 \dots \mathbf{z}_N\}$ with $\mathbf{z}_i \in \{e_1 \dots e_K\}$ (K classes)
- Additional missing data: $\mathbf{w} = \{\mathbf{w}_1 \dots \mathbf{w}_N\}$

$$\mathbf{y}_i | w_i, \mathbf{z}_i = e_k \sim \mathcal{G}(\mathbf{y}_i; \mu_k, \frac{\Sigma_k}{w_i})$$
$$w_i | \mathbf{z}_i = e_k \sim \Gamma(\frac{\nu_k}{2}, \frac{\nu_k}{2})$$

$$Z_i \sim \mathcal{M}(\pi_1 \dots \pi_K) \text{ independent}$$

- Unknown parameters: $\psi = \{\mu_k, \Sigma_k, \nu_k, \pi_k\}$

Expectation Maximization (EM) for maximum likelihood ( $\psi$ )

Iteration $r$

**E-step:** compute $p(\mathbf{z}, \mathbf{w} | \mathbf{y}; \psi^{(r)})$

**M-step:** $\psi^{(r+1)} = \arg\max_{\psi \in \underline{\Psi}} E[\log p(\mathbf{Z}, \mathbf{W}, \mathbf{y}; \psi) | \mathbf{y}; \psi^{(r)}]$

# EM algorithm for t-mixtures (MoT)

Iterate:

(E) $\left\{ \begin{array}{l} \text{Compute } q_{Z_i}^{(r)}(e_k) \text{ posterior class membership probabilities, for all } i, k \\ \text{Compute } \bar{w}_{ik}^{(r)} \text{ as} \end{array} \right.$

$$\bar{w}_{ik}^{(r)} = \frac{\nu_k^{(r)} + M}{\nu_k^{(r)} + \delta(y_i, \mu_k^{(r)}, \Sigma_k^{(r)})}$$

(M) $\left\{ \begin{array}{l} \text{Compute the dof } \nu_k^{(r+1)} \text{ as a solution of an equation} \\ \text{Compute the gaussian means and variances using} \end{array} \right.$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\ \bar{w}_{ik}^{(r)}\ y_i}{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\ \bar{w}_{ik}^{(r)}}$$

and $$\Sigma_k^{(r+1)} = \frac{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\ \bar{w}_{ik}^{(r)}\ (y_i - \mu_k^{(r+1)})(y_i - \mu_k^{(r+1)})^T}{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\bar{w}_{ik}^{(r)}}$$

# Illustrations

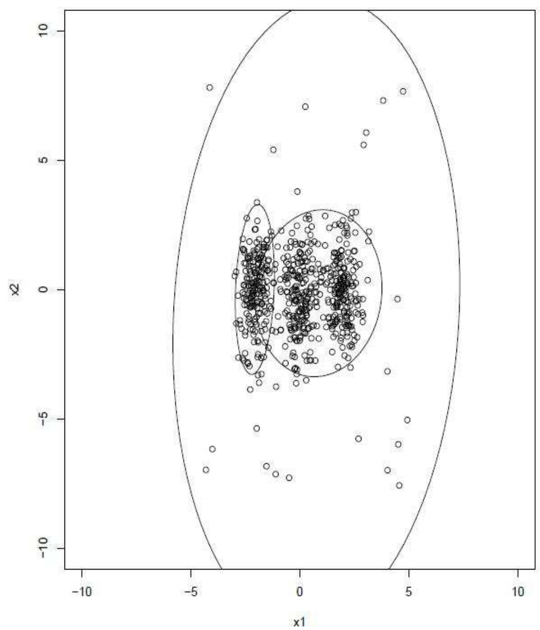



Figure 1: (Asymptotic) ellipsoids for the three clusters obtained by fitting a mixture of $g = 3$ normal components to three normal groups plus uniformly distributed noise.

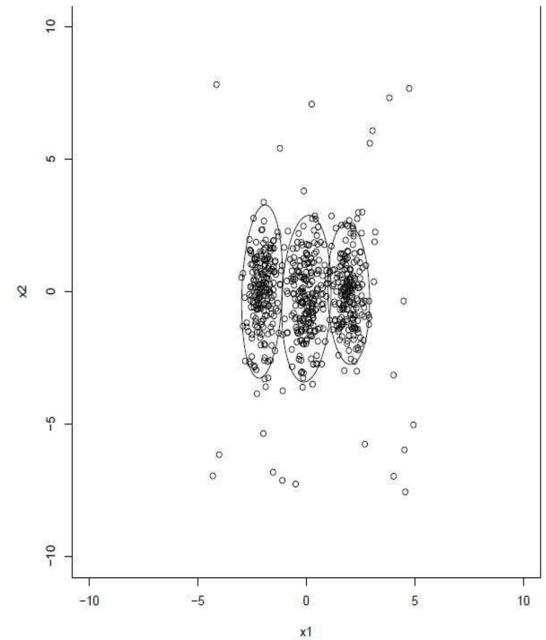



Figure 2: (Asymptotic) ellipsoids for the three clusters obtained by fitting a mixture of $g = 3$ $t$ components to three normal groups plus uniformly distributed noise.

MacLachlan, Ng, Bean 2006

# Robust Bayesian clustering

Student mixtures + priors on parameters : Bayesian Student mixtures

Inference by Variational Bayes EM

Advantage: automatic and robust model selection

References:

Svensen and Bishop: no prior on the dof and variational approximation qz qw qtheta

Archambeau and Verleysen: uniform prior on dof and “better” approximation qzw qtheta

Takekawa et Fukai: improved on Archambeau and Verleysen with exponential prior on dof

A lot of robust approaches to clustering via mixture models has been based on mixture of Student distribution

Goal: explore the scale mixture framework further

# Multivariate heavy tail distributions

Many ways to generalize the univariate Student distribution (in the Student spirit):

• The standard way has one particular disadvantage as a model for data: all its marginals are Student but have the same dof and hence the same amount of tailweight.

$$t(y;\mu,\Sigma,\nu) = \frac{\Gamma((M+\nu)/2)}{\Gamma(\nu/2)\ (\nu\pi)^{M/2}}\ |\Sigma|^{-1/2}\ \left[1+\delta^2/\nu\right]^{-(\nu+M)/2}$$

$$\text{with } \delta^2 = (y-\mu)^T\Sigma^{-1}(y-\mu)$$

$$X \sim \mathcal{N}(0,\Sigma) \text{ and } V \sim \mathcal{X}^2(\nu)$$

$$Y = X \times \sqrt{(\tfrac{\nu}{V})} + \mu \quad \sim \quad t(\mu,\Sigma,\nu)$$

• Product of independent t-distributions: varying dof but no correlation

• Jones 2002: a dependent bivariate t distribution with marginals of different dof. Extension to multivariate? Joint density not tractable?

• Eltoft et al. 2006: new multivariate scale mixture of Gaussians, more general than Student

$$X \sim \mathcal{N}(0, Id_M) \text{ and } \Lambda \text{ pos.def } M\times M \text{ with } |\Lambda| = 1,$$
$$Z \text{ a scalar positive variable with pdf to be chosen (eg } \Gamma \text{ or } \mathcal{X})$$
$$Y = \mu + \Lambda^{1/2}\frac{X}{\sqrt{Z}}$$

# New multivariate heavy tail distributions

## Several equivalent constructions:

### 1. Gaussian scale mixtures

$$f(y;\mu,\Sigma,\theta_1...\theta_M) = \int_0^\infty \int_0^\infty \mathcal{N}(y;\mu,DW^{-1}AD^T)\, g_1(w_1;\theta_1)...g_M(w_M;\theta_M)\, dw_1...dw_M$$

$$W = diag(w_1,...w_M) \qquad \Sigma = DAD^T$$

Student like: $g_1(w_1;\theta_1) = \mathcal{G}(w_1;\nu_1/2,\nu_1/2)\ldots g_M(w_M;\theta_M) = \mathcal{G}(w_M;\nu_M/2,\nu_M/2)$

Pearson Type VII like: $g_1(w_1;\theta_1) = \mathcal{G}(w_1;\alpha_1,\gamma_1)\ldots g_M(w_M;\theta_M) = \mathcal{G}(w_M;\alpha_M,\gamma_M)$

### 2. Generative construction (useful for simulation)

$X \sim \mathcal{N}(0, Id_M)$
and for $m = 1...M$, $Z_m \sim g_m(z;\theta_m)$
all independent (positive variables)

$$\tilde{X} = (\frac{X_1}{\sqrt{Z_1}}, ... \frac{X_M}{\sqrt{Z_M}})^T$$

Then $Y = \mu + \Sigma^{-1/2}\tilde{X}$

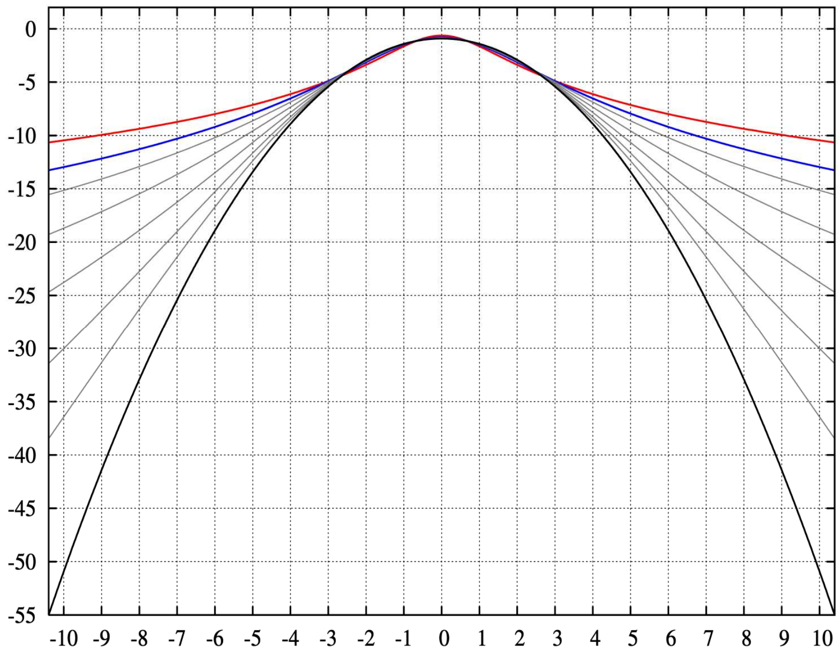
$X \sim \mathcal{N}(0, A)$
and for $m = 1...M$, $Z_m \sim g_m(z;\theta_m)$
all independent (positive variables)

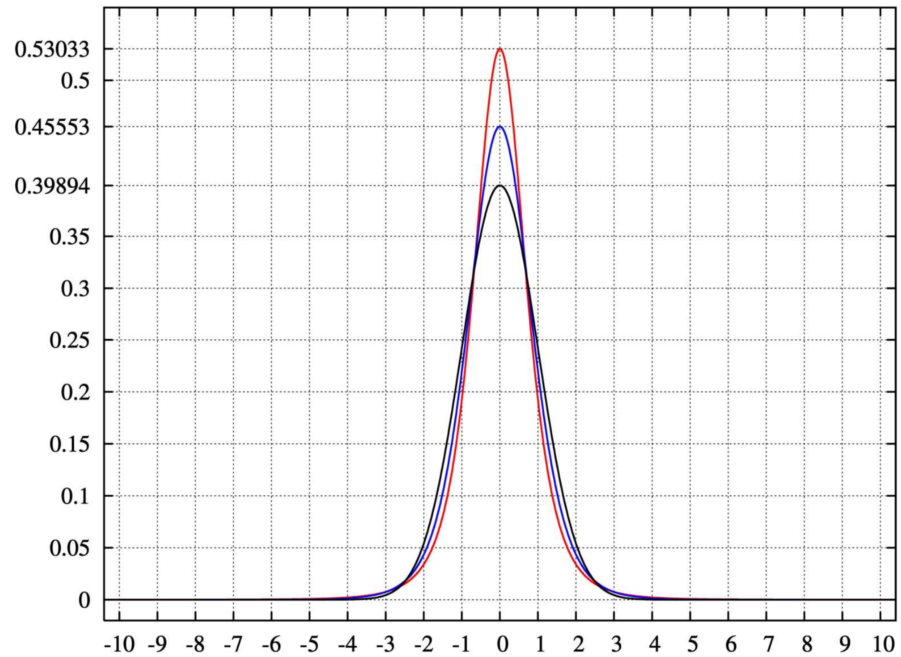
$$\tilde{X} = (\frac{X_1}{\sqrt{Z_1}}, ... \frac{X_M}{\sqrt{Z_M}})^T$$

Then $Y = \mu + D\tilde{X}$

# Univariate Pearson type VII distribution

Log density and density for different parameters (varying kurtosis ie. sharpness of peak)



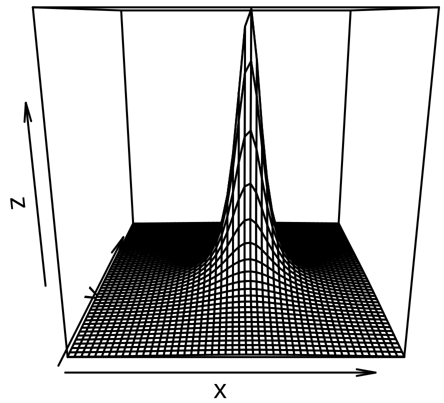



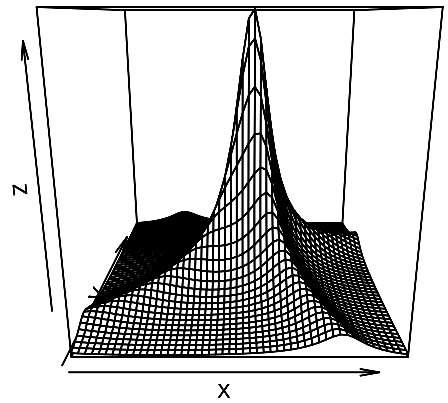

Gaussian distribution in black

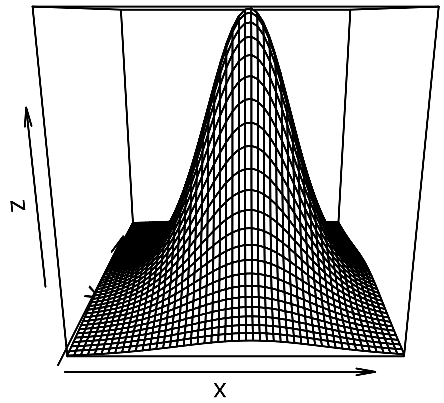
# Multiple DoF Student distributions

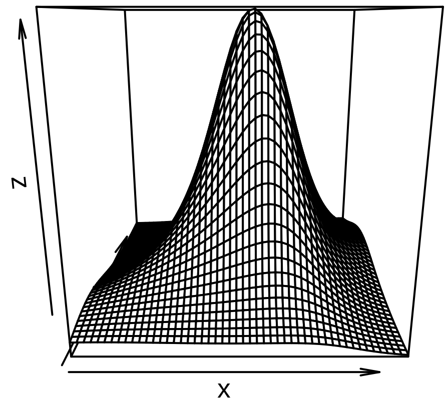
Student
$\nu = 0.1$ (top left) and
$\nu = 5$ (bottom left)

Student like:
$\nu_1 = \nu_2 = 0.1$ and $\theta = \pi/3$ (top right)
$\nu_1 = 1, \nu_2 = 10, \theta = \pi/3$ (bottom right)



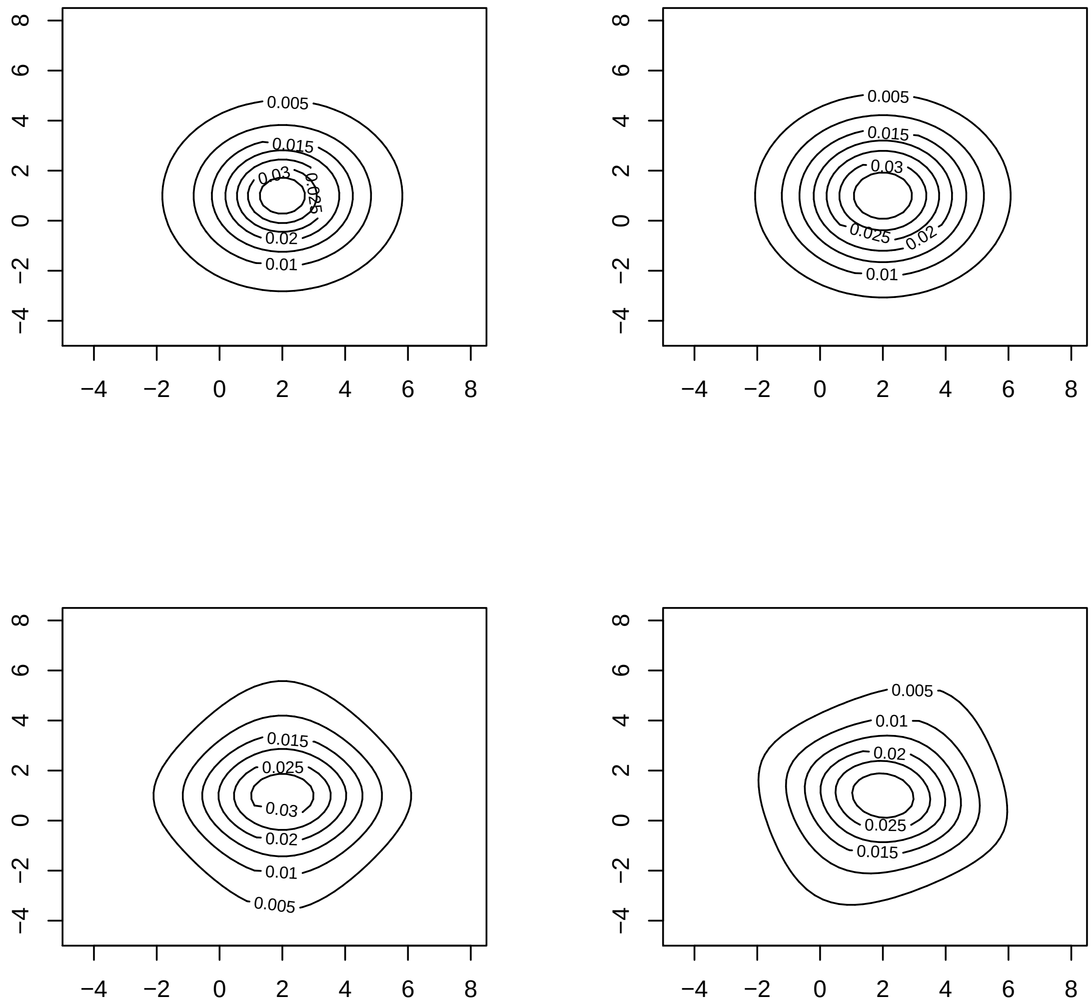

Multivariate Student

Multiple dof Multivariate Student

# Multiple DoF Student distributions

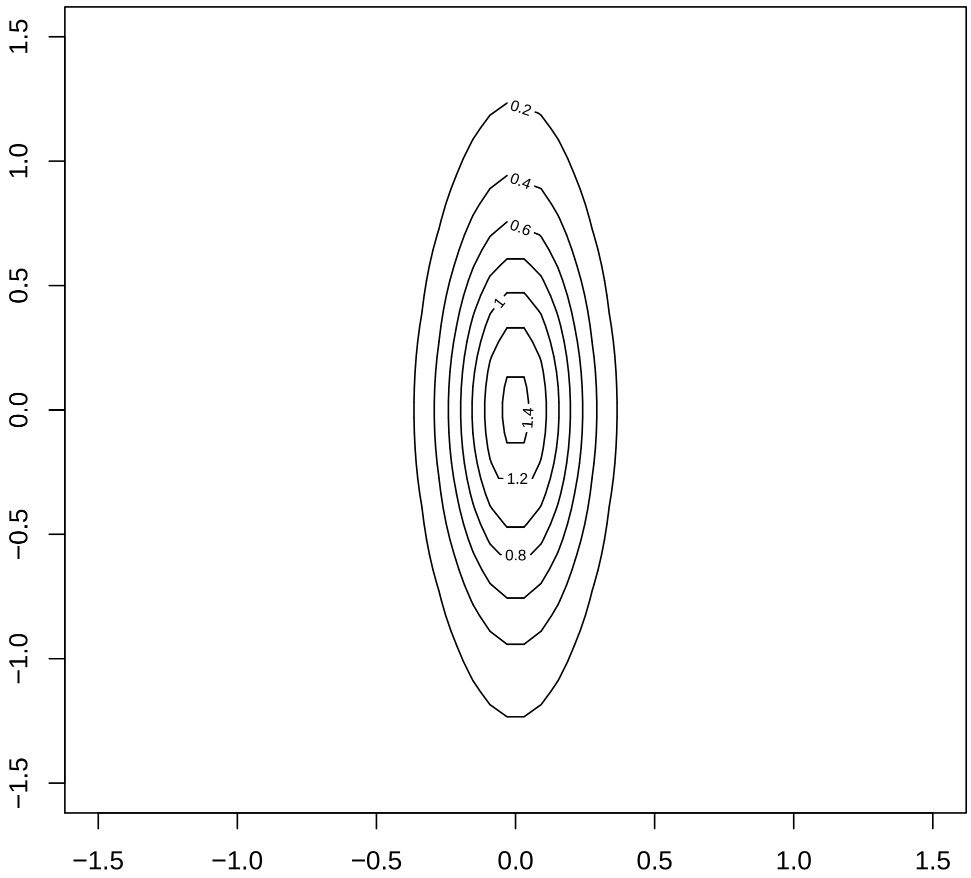
Bivariate Student:
$\nu = 2$ (left) and $\nu = 10$ (right)

New Bivariate Student like
for $\nu_1 = 2$, $\nu_2 = 10$:
$\theta = 0$ (left) and $\theta = \pi/8$ (right)



# Multivariate Pearson like distributions

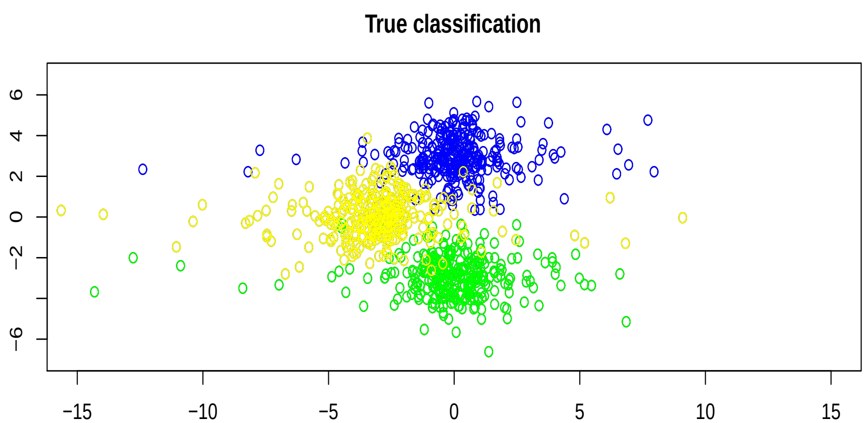
$$A_1 = 0.15,\ A_2 = 1,\ \theta = 0,\ \alpha_1 = 0.2,\ \alpha_1 = 1,\ \beta_1 = 5,\ \beta_2 = 10$$

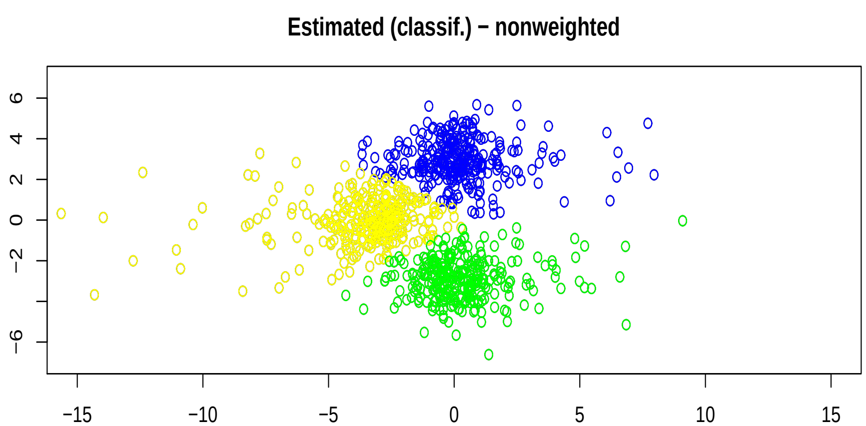


# Application to clustering: mixtures

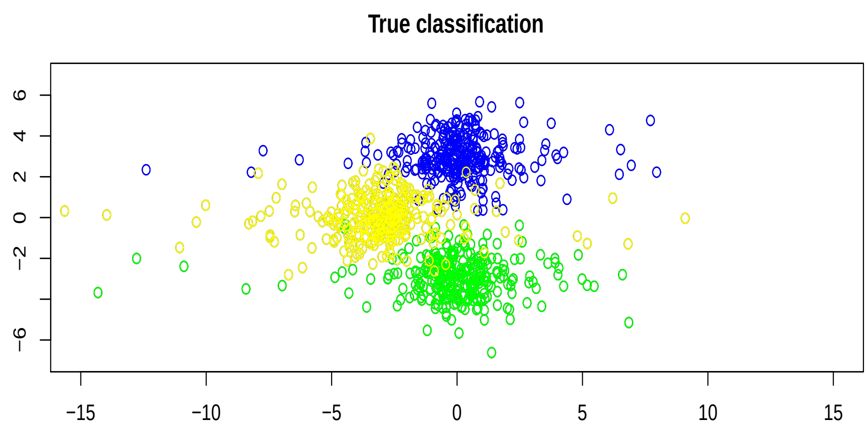
3 Clusters generated from the product of two univariate Student distribution with $\nu = 2$ and $\nu = 30$

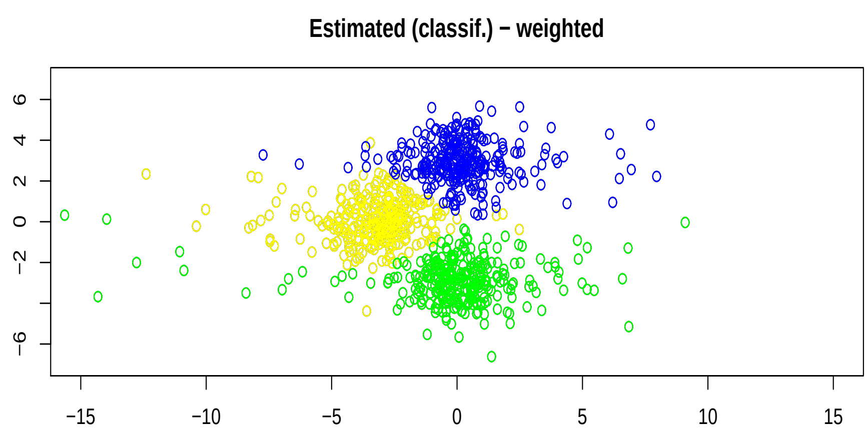
Standard MoT




MoMultipleDoFT: the different tailweights are better modeled



# Further generalization

**Data item dependent weight** priors and spatial class prior

**K-component mixture of M-dim. t-distributions** | **weighted model**

Data augmentation:

$\mathbf{w} = \{w_1 \dots w_N\}$
with $w_i > 0$ (independent of m)

$\{\mathbf{z}_1 \dots \mathbf{z}_N\}$ independent

$\mathbf{y}_i | w_i, \mathbf{z}_i = e_k \sim \mathcal{G}(\mathbf{y}_i; \mu_k, \frac{\Sigma_k}{w_i})$

$w_i | \mathbf{z}_i = e_k \sim \Gamma(\frac{\nu_k}{2}, \frac{\nu_k}{2})$

Ex. $\nu_k = \nu \quad \forall k$
$\Longrightarrow w_i$ independent of $\mathbf{z}_i$

$\mathbf{w} = \{\mathbf{w}_1 \dots \mathbf{w}_N\}$ with $\mathbf{w}_i = \{w_{i1} \dots w_{iM}\}$

$W_i = Diag(w_{i1} ... w_{iM})$

$\mathbf{z}$ Markovian

$\mathbf{y}_i | \mathbf{w}_i, \mathbf{z}_i = e_k \sim \mathcal{G}(\mathbf{y}_i; \mu_k, D_k W_i^{-1} A_k D_k^T))$

$w_{im} \sim \Gamma(\alpha_{im}, \gamma_{im})$ independent of k

For a standard mixture, we would need
$\alpha_{im}, \gamma_{im}$ independent of $i$
$\Longrightarrow$ inappropriate for lesion detection

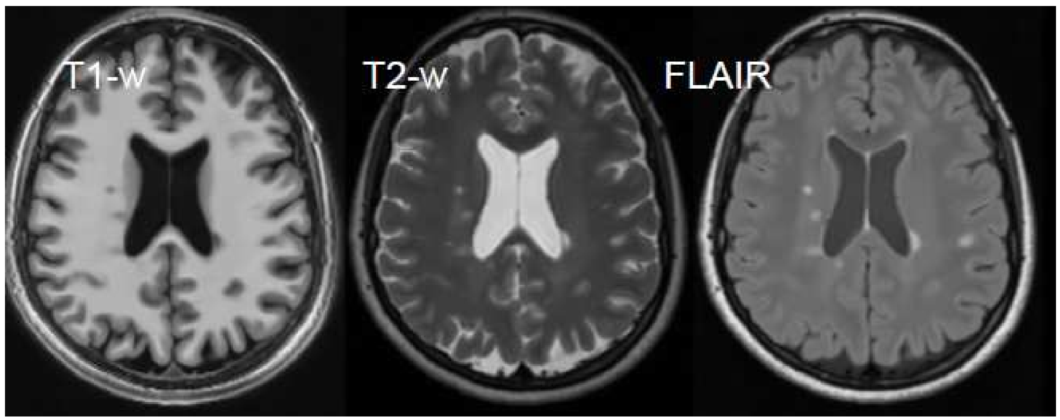
**MOTIVATION for such a generalization**

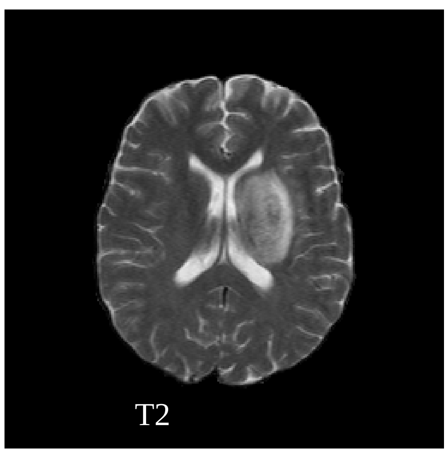
# Modelling lesions: inliers vs outliers

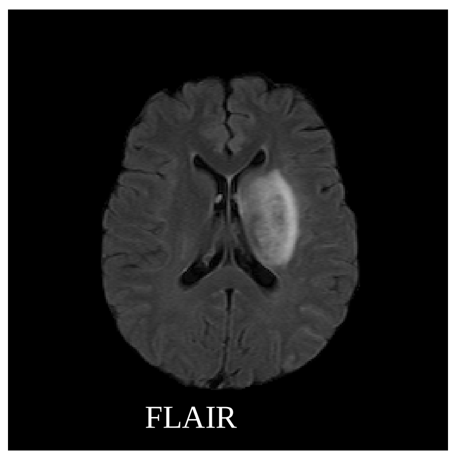
Explicit modelling usually avoided:

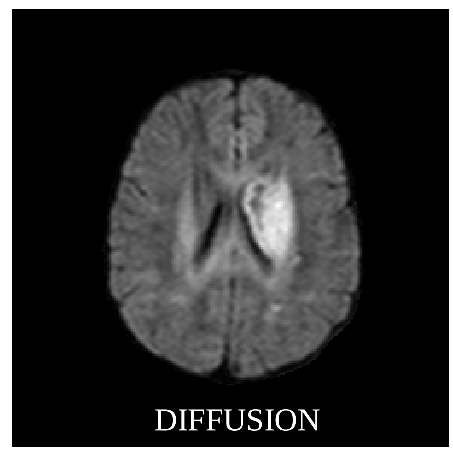
1) Widely varying and inhomogeneous appearance (tumors, stroke)

2) Lesion size can be small (MS lesions)



# Modelling lesions: inliers vs outliers

→ prevent accurate model parameter estimation

→ bad lesion delineation

**In most approaches:** lesion voxels identified as outliers wrt a normal brain model (a priori)

**Our approach (incorporation strategy):**

- Modify the segmentation model so that lesion voxels become inliers
- Make the estimation of the lesion class possible
- Use an additional weight field

# Reasons for using weights

1) To bias the model toward lesion identification: voxel specific weights

- eg. duplicate intensity values typical of the lesion

2) To weight the information content of each sequences: modality specific weights

- Multiple MR volumes are commonly modelled via multivariate Gaussian intensity distributions
- But all the sequences have equal importance

⇒ Optimally combine sequences to take into account

– a priori (expert) knowledge
– the targeted task
– the type of lesion

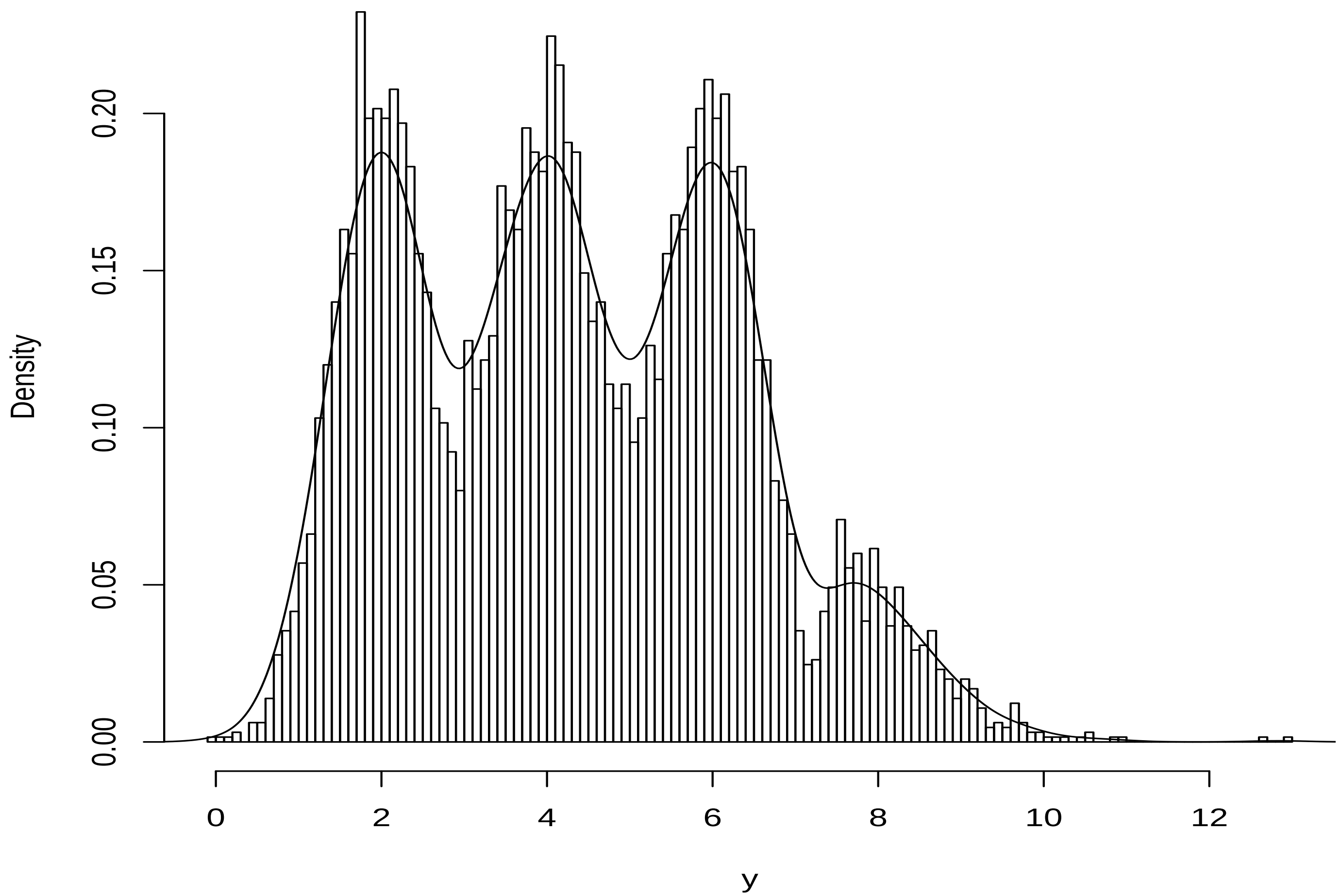
**Weight choice? Bayesian framework:**

- incorporate a priori on relevant information content of each sequence
- a weighting scheme modified adaptively

# Robustness to non Gaussian components

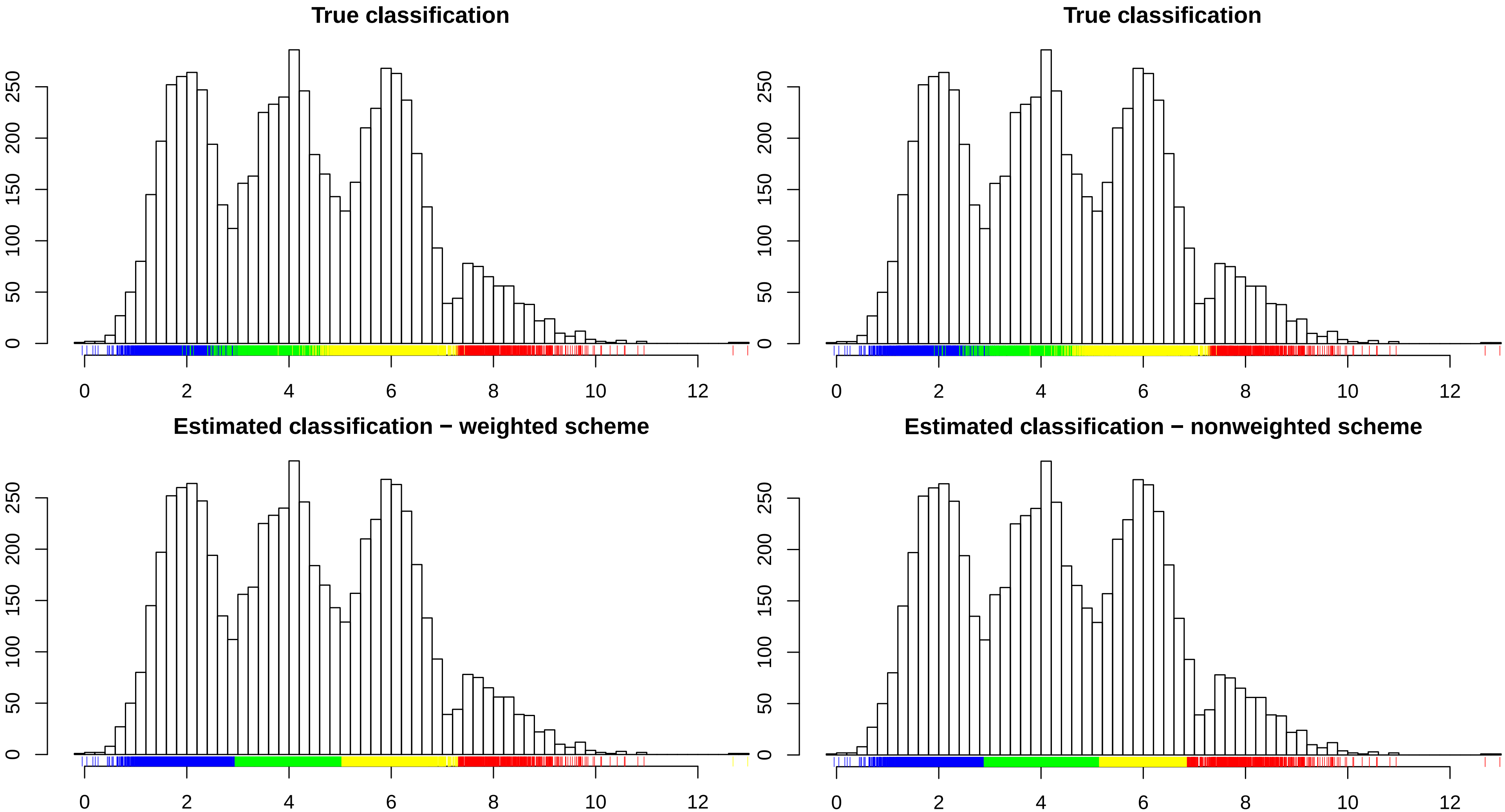
**Illustration: non spatial, data point dependent weight, no expert, priors G(1,1)**

1. To assess the ability to deal with varying cluster shapes




3 Gaussians and 500 data points from a $\mathcal{G}(2,2)$ to the right

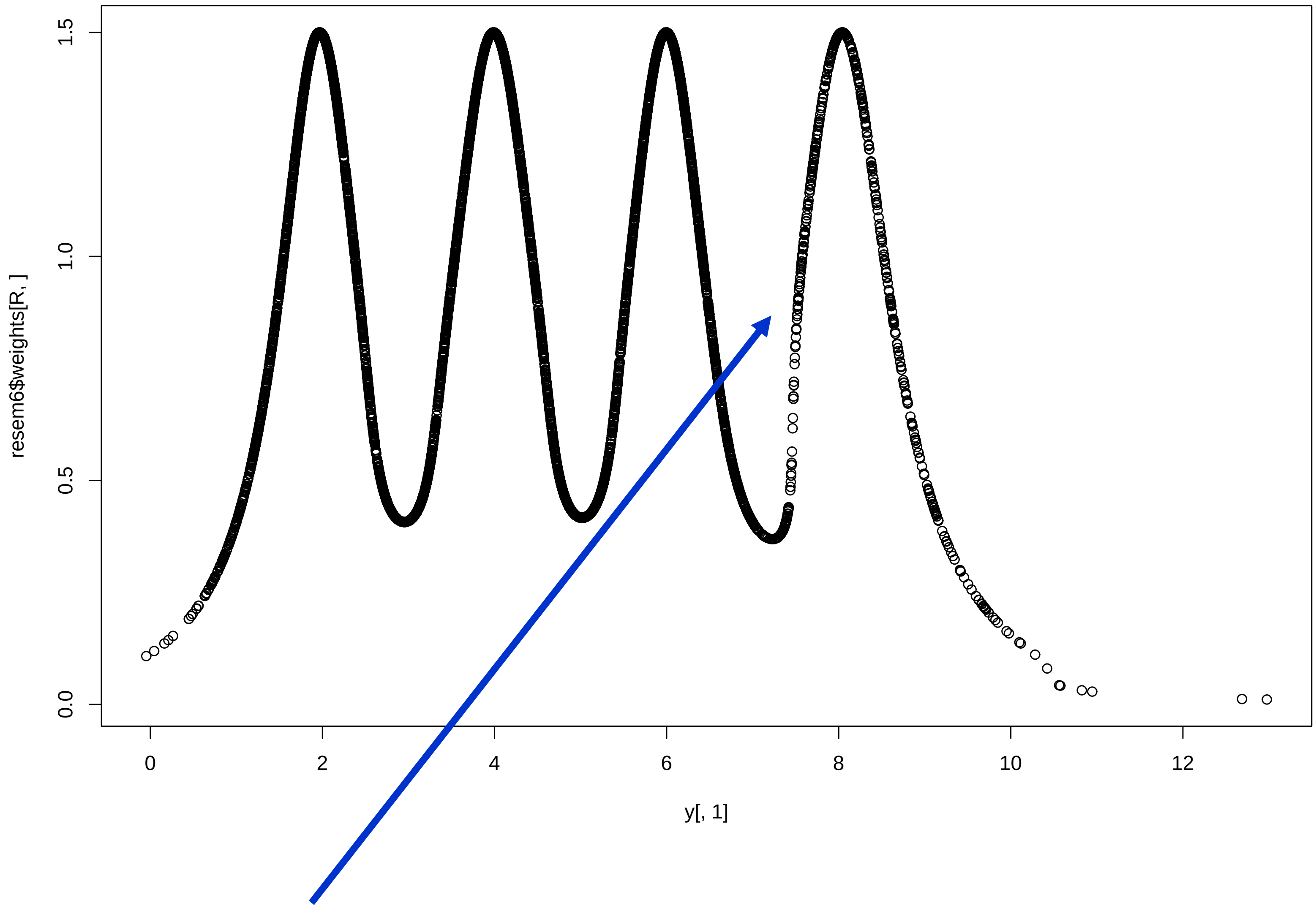
# Weighted G(1,1) vs non weighted approach




Better classification: 520 versus 500 (true)

MoG: the 4th component is over fitted, 650 points vs 500 (true)

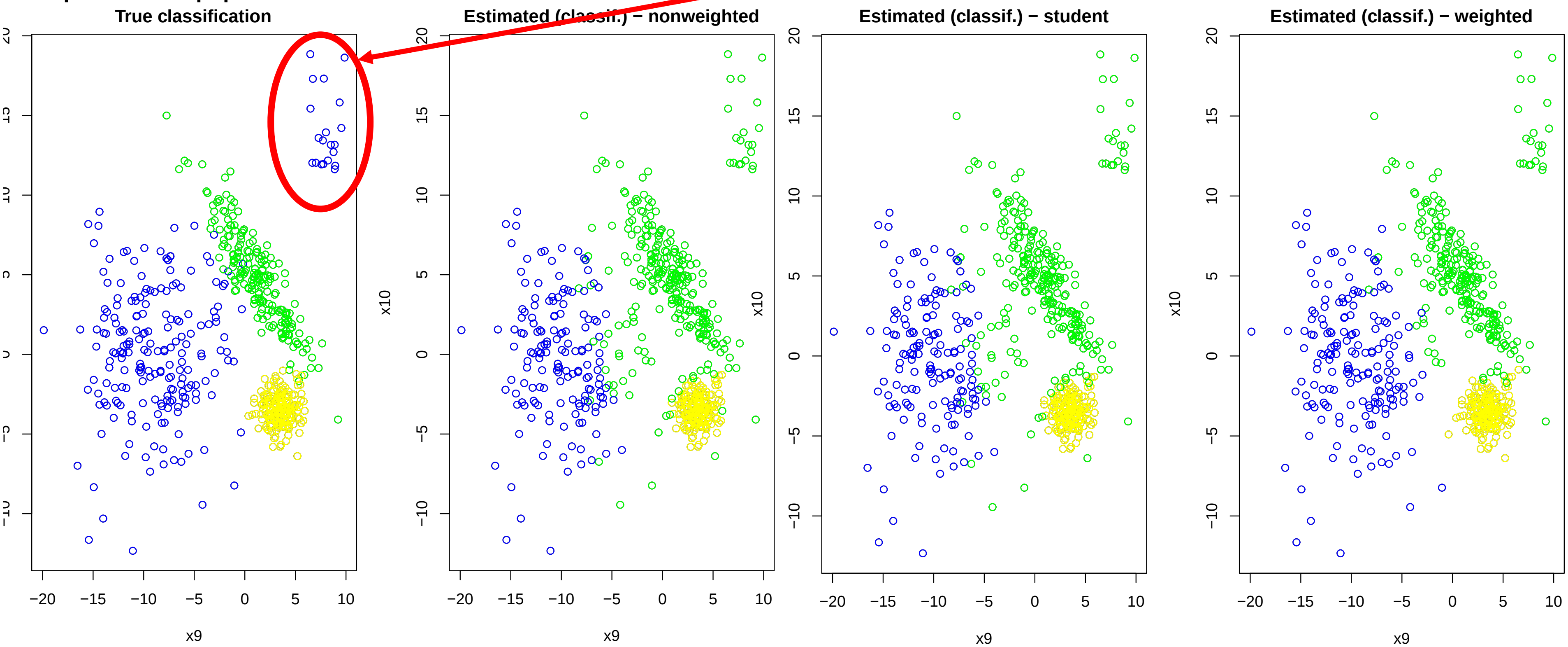
# Robustness to shape variability




The weights adjust to the data allowing slight deviations from a Gaussian distribution

# Robustness to outliers (grouped) Prior: $\mathcal{G}(1,1)$

3 bivariate Gaussians (600 points) contaminated with 20 points from a uniform distribution in a parallelepiped




$$\pi_i = 1/3, \quad i = 1,\ldots,3, \quad \mu_1 = (-9,0), \quad \mu_2 = (1,5), \quad \mu_3 = (3.5,-3.5)$$

$$\Sigma_1 = \begin{pmatrix} 16 & 0 \\ 0 & 16 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 8.5 & -7.5 \\ -7.5 & 8.5 \end{pmatrix}, \quad \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The data item dependent weight model is less sensitive to outliers

Effect of the weights: allow a long tail on one side and a truncated distribution on the other side (green component) => Flexible shape clusters…

# Lesion detection or semi-supervised context

## Principe de la méthode d'incorporation

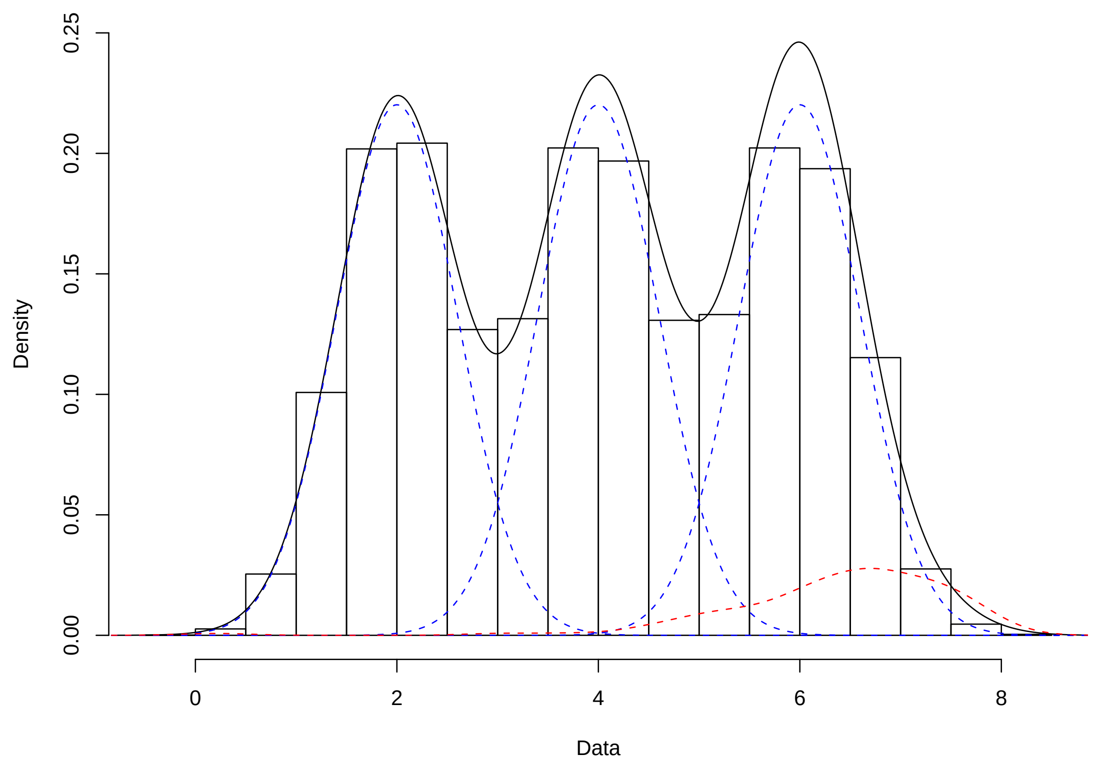
- Détermination d'une région d'intéret, (à pondérer):
  - Les voxels candidats à appartenir à la lesion
  - Les voxels sélectionnés sont les moins représentatifs du modèle "sain", ie. les outliers
  - Ou sélection par un expert (semi-supervisé) [Graph cut]
  - Ou utilisation de règles (faux positifs) [STREM]

- Segmentation initiale:
  - Classe lésion= région d'intéret
  - Les autres voxels sont segmentés en 3 classes

⇒ Variante de EM pour modèle de mélange à K=4 classes avec pondérations avec loi a priori sur les poids

# Supervised context: non spatial illustration

1. To assess the ability of the weighted approach to detect a small non Gaussian component




3 Gaussian components (5000 data points each)

and

a small Beta(10,2) shifted by 6 units representing 100 data points (proportion=0.066)

The smallest component is "of interest" (eg. lesions)

Procedure: choose $\mathcal{L}$ data points from the fourth component (supervised) and use a $\mathcal{G}(\alpha, \beta)$ prior for the corresponding weights variables

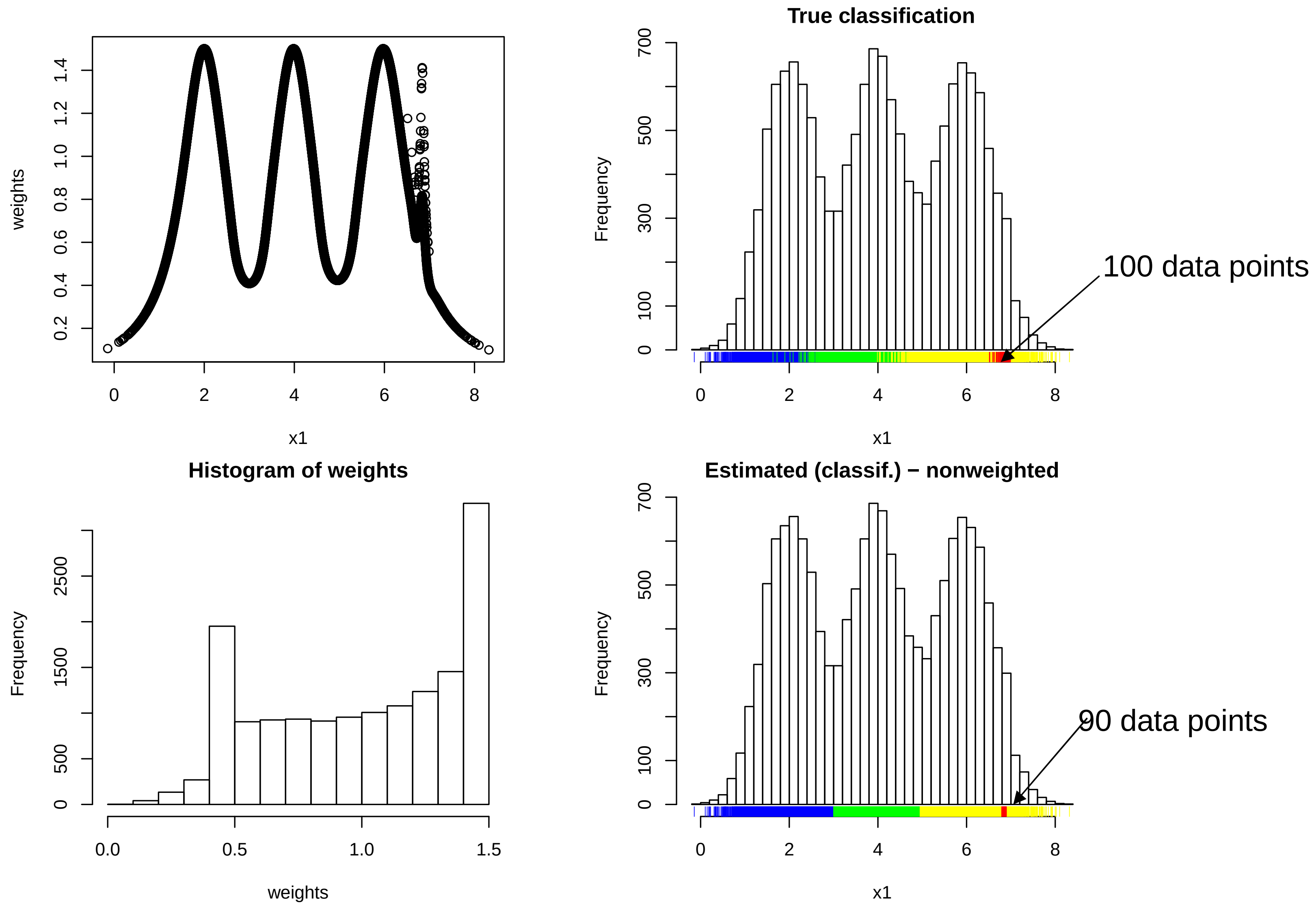
# Supervised context

| | Number of points classified to 4th component | | |
|---|---|---|---|
| Prior parameters for $w$ | $\mathcal{L} = 10$ | $\mathcal{L} = 50$ | $\mathcal{L} = 100$ |
| $\alpha = 1.0; \beta = 1.0$ | 0 | 0 | 0 |
| $\alpha = 1.5; \beta = 1.0$ | 25 | 90 | 147 |
| $\alpha = 3.0; \beta = 1.0$ | 50 | 223 | |

Note: with a Gaussian mixture model (K=4), no points in the $4^{th}$ component

# Supervised context

Clustering results for $\mathcal{L} = 50$ and $\mathcal{G}(1.5, 1)$



# Spatial case: A weighted Hidden Markov model

N voxels (3D) x M modalities (T1, T2, Flair images)

- Observations: $\mathbf{y} = \{\mathbf{y}_1 \dots \mathbf{y}_N\}$ where $\mathbf{y}_i = \{y_{i1} \dots y_{iM}\}$
- Labels: $\mathbf{z} = \{\mathbf{z}_1 \dots \mathbf{z}_N\}$ with $\mathbf{z}_i \in \{e_1 \dots e_K\}$ (K tissues)
- Weights: $\mathbf{w} = \{\mathbf{w}_1 \dots \mathbf{w}_N\}$ with $\mathbf{w}_i = \{w_{i1} \dots w_{iM}\}$
  sequence and voxel specific

**Spatial dependencies between voxels:**

the joint distribution is a Markov random field (MRF)

$$p(\mathbf{y}, \mathbf{z}, \mathbf{w}; \psi) \propto \exp(H(\mathbf{y}, \mathbf{z}, \mathbf{w}; \psi)) \qquad \psi = \{\beta, \phi\}$$

$$H(\mathbf{y}, \mathbf{z}, \mathbf{w}; \psi) = H_Z(\mathbf{z}; \beta) + H_W(\mathbf{w}) + \sum_{i \in V} \log g(\mathbf{y}_i | \mathbf{z}_i, \mathbf{w}_i; \phi)$$

Missing data term

Parameter prior term

Data driven term, based on intensities

**Data term:** $g(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}_i; \phi) = \mathcal{G}(y_i; \mu_{z_i}, D_{z_i} W_i^{-1} A_{z_i} D_{z_1}^T)$

If all weights are 1, a standard multivariate (diagonal) Gaussian case is recovered

**Missing data term:** $H_Z(\mathbf{z};\beta) = \sum_{i=1}^{N} (< \mathbf{z}_i, \xi > + \sum_{j \in \mathcal{N}(i)} \eta \ < \mathbf{z}_i, \mathbf{z}_j >)$

$\mathcal{N}(i)$: voxels neighboring $i$

$\beta = \{\xi, \eta\}$ with $\xi = {}^t(\xi_1, \ldots, \xi_K)$ and $\eta > 0$

Potts model with external field $\xi$, interaction parameter $\eta$

**Parameter prior term:** $p(\mathbf{w}) = \prod_{m=1}^{M} p(\mathbf{w}_m)$ $\quad \mathbf{w}_m = \{w_{1m} \ldots w_{Nm}\}$

1) $\sum_{i=1}^{N} w_{im} = N$ $\quad$ A dirichlet distribution for $p(\mathbf{w}_m)$

2) the $w_{im}$ are independent $\quad w_{im} \sim \Gamma(\alpha_{im}, \gamma_{im})$

$\alpha_{im} = \gamma_{im} w_{im}^{exp} + 1$ $\quad$ $w_{im}^{exp}$ is the mode of the prior for $w_{im}$

# Estimation by variational EM

An alternating maximization view of EM: $F(q,\psi) = E_q[\log p(\mathbf{y},\mathbf{Z},\mathbf{W}\ ;\ \psi)] + I[q]$

[Neal&Hinton98]

$I[q] = -E_q[\log q(\mathbf{z}, W)]$ (entropy of $q$)

$q \in \mathcal{D}$ a distribution on $\mathcal{Z} \times \mathcal{W}$

$$\textbf{E-step:}\ q^{(r)} = \arg\max_{q\in\mathcal{D}} F(q, \psi^{(r)})$$

$$\textbf{M-step:}\ \psi^{(r+1)} = \arg\max_{\psi\in\underline{\Psi}} F(q^{(r)}, \psi)$$

Variational approximation:

Exact E-step leads to $q^{(r)}(\mathbf{z},\mathbf{w}) = p(\mathbf{z},\mathbf{w}|\mathbf{y};\psi^{(r)})$ intractable

EM variant (Variational EM): $q(\mathbf{z},\mathbf{w}) = q_Z(\mathbf{z})\ q_W(\mathbf{w})$

The E-step is solved over a restricted class of pdfs (that factorize)

The E-step is further approximated by its decomposition in 2 sub-steps (Incremental EM [Neal&Hinton98])

Modified GAM procedures [Byrne&Gunawardana05]

# Variational E-step

$$\textbf{E-Z:}\quad q_Z^{(r)} = \arg\max_{q_Z \in \mathcal{D}_Z} F(q_W^{(r-1)}\ q_Z; \psi^{(r)})$$

$$\textbf{E-W:}\quad q_W^{(r)} = \arg\max_{q_W \in \mathcal{D}_W} F(q_W\ q_Z^{(r)}; \psi^{(r)})$$

$$\textbf{E-Z:}\quad q_Z^{(r)} \propto \exp\left(E_{q_W^{(r-1)}}[\log p(\mathbf{z}|\mathbf{y}, \mathbf{W}; \psi^{(r)}]\right)$$

$$\textbf{E-W:}\quad q_W^{(r)} \propto \exp\left(E_{q_Z^{(r)}}[\log p(\mathbf{w}|\mathbf{y}, \mathbf{Z}; \psi^{(r)})]\right)$$

## For the weighted Markov model

$$p(\mathbf{y}, \mathbf{z}, \mathbf{w}; \psi) \text{ is Markovian } \Longrightarrow \text{ all conditionals are Markovian}$$

$$p(\mathbf{z}|\mathbf{y}, \mathbf{w}; \psi) \text{ is Markovian: } H(\mathbf{z}|\mathbf{y}, \mathbf{w}; \psi) = H_Z(\mathbf{z}; \beta) + \sum_{i \in V} \log g(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}_i; \phi)$$

$$p(\mathbf{w}|\mathbf{y}, \mathbf{z}; \psi) \text{ is Markovian: } H(\mathbf{w}|\mathbf{y}, \mathbf{z}; \psi) = H_W(\mathbf{w}) + \sum_{i \in V} \log g(\mathbf{y}_i|\mathbf{z}_i, \mathbf{w}_i; \phi)$$

# In practice (for diagonal covariances)

Fix $\eta$ (interaction parameter) and the expert weights $w_{im}^{exp}$ (modes of the weight priors) and $\gamma_{im}$ (variances of the weight priors)

Iterate:

(E) Compute $q_Z^{(r)}(\mathbf{z})$ using mean-field approximation or variants

Compute $\bar{w}_{im}^{(r)}$ as

$$\bar{w}_{im}^{(r)} = \frac{\alpha_{im} + \frac{1}{2}}{\gamma_{im} + \frac{1}{2} \sum_{k=1}^{K} \delta(y_{im}, \mu_{km}^{(r)}, s_{km}^{(r)})\ q_{Z_i}^{(r)}(e_k)}$$

with $\alpha_{im} = \gamma_{im} w_{im}^{exp} + 1$

$$\delta(y_{im}, \mu_{km}^{(r)}, s_{km}^{(r)}) = \frac{(y_{im} - \mu_{km}^{(r)})^2}{s_{km}^{(r)}}$$

(Mahalanobis distance)

(M) Compute (Newton) $\xi$ (external field parameter) using mean field

Compute the gaussian means and variances using

$$\mu_{km}^{(r+1)} = \frac{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\ \bar{w}_{im}^{(r)}\ y_{im}}{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\ \bar{w}_{im}^{(r)}}$$

and $$s_{km}^{(r+1)} = \frac{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\ \bar{w}_{im}^{(r)}\ (y_{im} - \mu_{km}^{(r+1)})^2}{\sum_{i=1}^{N} q_{Z_i}^{(r)}(e_k)\bar{w}_{im}^{(r)}}$$

# Choosing the expert weights

Expert knowledge difficult to formalize into weight values

Proposed setting:

$$
\begin{aligned}
w_{im}^{exp} &= w_{\mathcal{L}} > 1 \quad \forall i \in \mathcal{L} \\
w_{im}^{exp} &= 1 \quad \forall i \notin \mathcal{L}
\end{aligned}
$$

$\mathcal{L}$ is obtained by applying the model
with $K = 3$, $w_{im}^{exp} = 1$, $\gamma_{im} = 1 \quad \forall i, m$
$\eta = 0$ is ok

→ Identify outliers by thresholding (Chi2 percentile)
the estimated weights (typicality)

# Experiments

Parameters to tune:

$$Chi^2 \text{ percentile fixed to } 99\%$$
$$w_{\mathcal{L}} = 10$$
$$\gamma_{im} = \gamma = 10 \quad \forall i, m \quad \text{(prior variances)}$$

**Simulated data (BrainWeb) with MS lesions: T1,T2,PD sequences, 1mm$^3$**

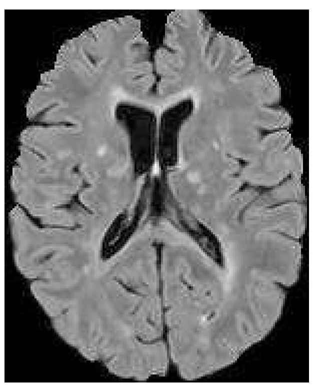
| Method | 3% | 5% | 7% | 9% |
|---|---|---|---|---|
| Mild lesions (0.02% of the voxels) | | | | |
| AWEM | **72 (+5)** | 55 (-15) | **39 (+5)** | **22 (+18)** |
| [G] | 67 | **70** | 34 | 0 |
| EMS | 56 | 33 | 13 | 4 |
| [R] | 52 | NA | NA | NA |
| Moderate lesions (0.18% of the voxels) | | | | |
| AWEM | **86 (+7)** | 80 (-1) | **77 (+18)** | **73 (+36)** |
| [G] | 72 | **81** | 59 | 29 |
| EMS | 79 | 69 | 52 | 37 |
| [R] | 63 | NA | NA | NA |
| Severe lesions (0.52% of the voxels) | | | | |
| AWEM | **93 (+8)** | 88 (0) | **78 (+6)** | **74 (+33)** |
| [G] | 79 | **88** | 72 | 41 |
| EMS | 85 | 72 | 56 | 41 |
| [R] | 82 | NA | NA | NA |

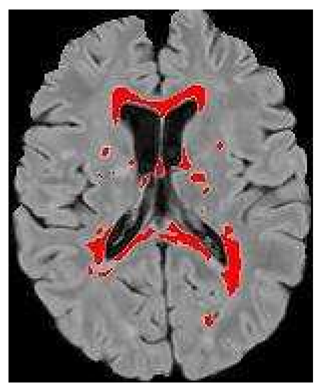
# Real data sets

2 Patients with MS: Flair,T1,T2 sequences, 1mm$^2$x3mm

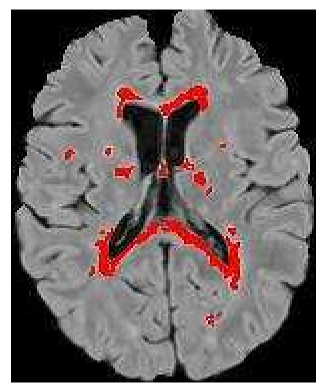
76%

Ground truth





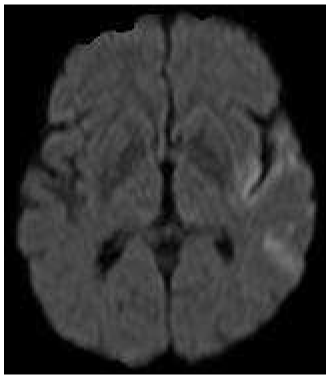
58%

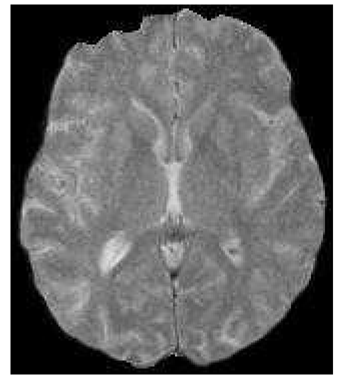


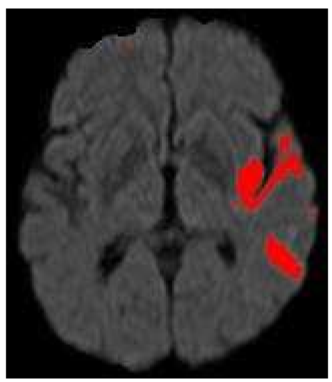
Ground truth

# Real data sets

1 Patient with stroke: DW , Flair, T2 sequences, 1mm$^2$x5mm







63%

Ground truth

# Future work

- Extension to full covariance matrices: temporal multi-sequence data, eg. patient follow-up
- Other prior for the weights: eg. MRF prior
- Other expert weighting schemes, possibly lesion specific
- Extension to handle intensity inhomogeneities
- Sensitivity analysis: initialization, parameter tuning etc.
- Evaluation in a semi-supervised context
- Add lesion specific information: atlas, rules etc.