Mixture Models

Václav Šmídl

April 24, 2018

Mixture of Gaussians

Probability distribution:

$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mu_k, \Sigma_k),$$

where μ_k, Σ_k are mean and covariance matrix of Gaussians weighted by α_k.

Mixture of Gaussians

Probability distribution:

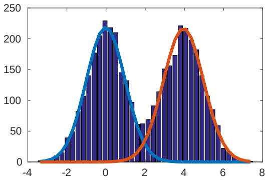
$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mu_k, \Sigma_k),$$

where μ_k, Σ_k are mean and covariance matrix of Gaussians weighted by α_k.

- universal approximation property
- non-uniqueness
  - combinatorial,
  - additive
- sampling,

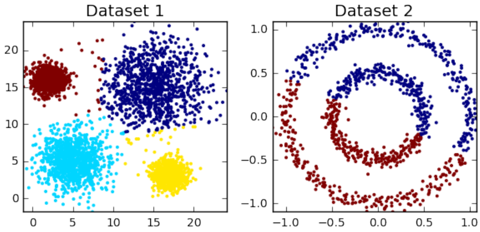
$$p(x) = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(4, 1),$$

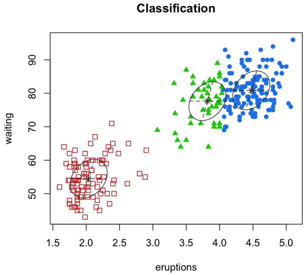
- maximum likelihood,



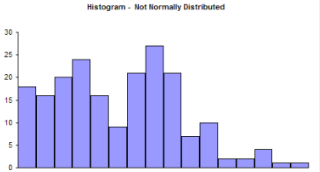
Uses of mixtures

Clustering (supervised, unsupervised, semi-supervised):





Density representation:



Classification:

Maximum likelihood

Probability distribution with parameters $\theta = \{\mu_k, \Sigma_k, \alpha_k\}_{k=1}^K$:

$$p(x_i|\theta) = \sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k),$$

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n p(x_i), \quad \log p(x_1, \ldots, x_n) = \sum_{i=1}^n \log p(x_i)$$

$$\log p(x_1, \ldots, x_n|\theta) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k) \right)$$

$$\mathcal{N}(\mu_k, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left(-\frac{1}{2}(x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right),$$

Finding $\hat{\theta} = \arg\max_\theta \log p(x_1, \ldots x_n)$:

$$\frac{d}{d\mu_k} \log p(x_{1:n}) = \sum_{i=1}^n \frac{\alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^K \mathcal{N}(x_i|\mu_k, \Sigma_k)\alpha_k} \left(\Sigma_k^{-1}(\mu_k - x_i)\right) \equiv 0,$$

and others.

Maximum likelihood [Baum, Welch]

Finding $\hat{\theta} = \arg\max_\theta \log p(x_1, \dots x_n)$ s.t. $\sum_k \alpha_k = 1$:

$$\frac{d}{d\mu_k} \log p(x_{1:n}) = \sum_{i=1}^{n} \frac{\alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)\alpha_k} \left(\Sigma_k^{-1}(\mu_k - x_i)\right) \equiv 0,$$

$$\frac{d}{d\alpha_k} \log p(x_{1:n}) = \sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)\alpha_k} + \lambda = 0,$$

where λ is a Lagrange multiplier for $\sum_k \alpha_k = 1$.

Maximum likelihood [Baum, Welch]

Finding $\hat{\theta} = \arg\max_\theta \log p(x_1, \ldots x_n)$ s.t. $\sum_k \alpha_k = 1$:

$$\frac{d}{d\mu_k} \log p(x_{1:n}) = \sum_{i=1}^{n} \frac{\alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)\alpha_k} \left(\Sigma_k^{-1}(\mu_k - x_i)\right) \equiv 0,$$

$$\frac{d}{d\alpha_k} \log p(x_{1:n}) = \sum_{i=1}^{n} \frac{\mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)\alpha_k} + \lambda = 0,$$

where λ is a Lagrange multiplier for $\sum_k \alpha_k = 1$.
Solution via alternating evaluation of

$$w_{i,k} = \frac{\alpha_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)\alpha_k},$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_i w_{i,k} x_i, \quad n_k = \sum_i w_{i,k},$$

$$\Sigma_k = \frac{1}{n_k} \sum_i w_{k,i} \left(x_i - \hat{\mu}_k\right) \left(x_i - \hat{\mu}_k\right)^T, \ \ \alpha_k = \frac{n_k}{n},$$

Peculiarity: $w_{i,k}$ is a conditional probability of x_i belonging to a class.
Any reason?

Mixture estimation

Consider latent variable $l \in \{\epsilon_1, \ldots, \epsilon_K\}$, $\epsilon_k = [0, 0, \ldots 1 \ldots 0]$. (1-of-n).

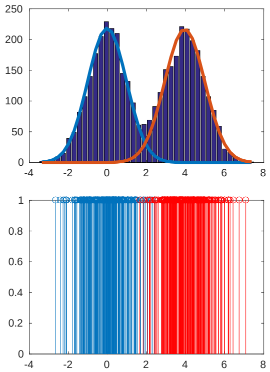
$$p(x, l) = p(x|l)p(l),$$

$$p(x|l) = \prod_k \mathcal{N}(\mu_k, \Sigma_k)^{l_k},$$

$$p(l_k = 1) = \alpha_k, \sum_k \alpha_k = 1$$

$$p(l) = \prod_{k=1}^{K} \alpha_k^{l_k}.$$

$$p(x) = \sum_k p(x|l = \epsilon_k)p(l = \epsilon_k).$$



Multinomial (Bernouli) distribution.

- Each data point has a label from which component is generated.
- Estimation of the joint distribution

Expectation maximization (EM) algorithm

Joint distribution:

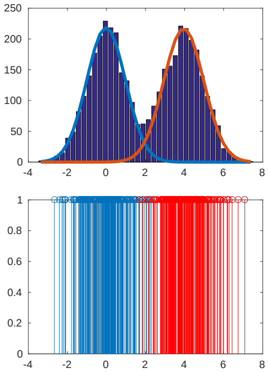
$$p(x, l) = p(x|l)p(l),$$

$$p(x|l) = \prod_k \mathcal{N}(x|\mu_k, \Sigma_k)^{l_k},$$

$$p(l_k = 1) = \alpha_k, \sum_k \alpha_k = 1.$$

Conditional distribution

$$p(l = \epsilon_k|x) = \frac{p(x, l)}{p(x)} = \frac{\mathcal{N}(x|\mu_k, \Sigma_k)\alpha_k}{\sum_k \mathcal{N}(x|\mu_k, \Sigma_k)\alpha_k}$$



Maximum likelihood with latent variable

General EM algorithm [Dempster, Laird, Rubin, 1977]:

$$\hat{\theta} = \arg\max_{\theta} \int p(x|\theta, l)p(l)dl,$$

can be (approximately) found by alternating:

E-step: $q(\theta|\theta^{(i)}) = \int \log p(x|\theta, l)p(l|\theta^{(i)})dl$

M-step: $\theta^{(i+1)} = \arg\max_{\theta} q(\theta|\theta^{(i)})$

Maximizing log-likelihood

$$p(x_1, l_1, \dots x_n, l_n|\theta) \propto \prod_{i=1}^{n}\prod_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{l_{k,i}} \alpha_k^{l_{k,i}}$$

$$\log p(x_1, l_1, \dots x_n, l_n|\theta) \propto \sum_{i=1}^{n}\sum_{k=1}^{K} l_{k,i} \left(\log(\mathcal{N}(x_i|\mu_k, \Sigma_k)) + \log \alpha_k\right)$$

$$\propto \sum_{i=1}^{n}\sum_{k=1}^{K} l_{k,i} \left(-\frac{1}{2}|\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^T\Sigma_k^{-1}(x_i - \mu_k) + \log$$

where

$$q(\theta|\theta^{(i)}) = \sum^{n}\sum^{K} \hat{l}_{k,i} \left(-\frac{1}{2}|\Sigma_k| - \frac{1}{2}(x_i - \mu_k)^T\Sigma_k^{-1}(x_i - \mu_k) + \log \alpha_k\right)$$

Expectation Maximization (EM) algorithm

[Dempster, Laird, Rubin, 1977]

Initialize: choose $\alpha_k^{(0)}, \mu_k^{(0)}, \Sigma_k^{(0)}, \forall k$

Iterate:

1. Compute expected labels:

$$p(l = \epsilon_k | x_i) = \hat{l}_{k,i} = \frac{\mathcal{N}(\mu_k, \Sigma_k)\alpha_k}{\sum_k \mathcal{N}(\mu_k, \Sigma_k)\alpha_k}$$

2. Recompute the component parameters

$$\hat{\mu}_k = \frac{1}{N_k} \sum_i \hat{l}_{k,i} x_i,$$

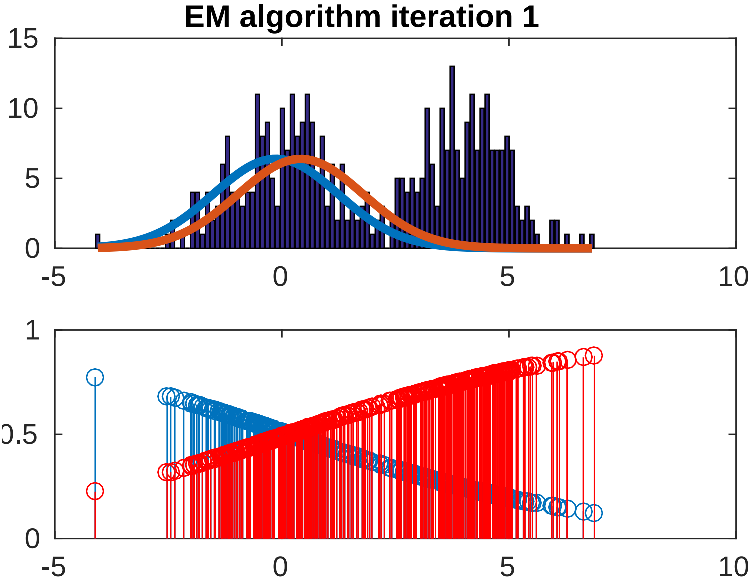
$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_i \hat{l}_{k,i} \left(x_i - \hat{\mu}_k\right) \left(x_i - \hat{\mu}_k\right)^T,$$

$$\hat{\alpha}_k = \frac{N_k}{N},$$

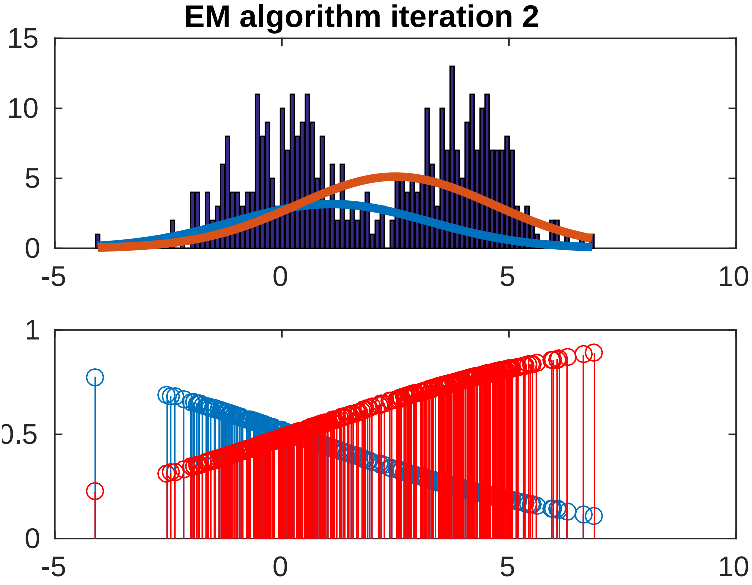
3. (Evaluate log-likelihood)

$$\log p(x) = \sum_i \log \left(\sum_{k=1}^{K} \mathcal{N}(\hat{\mu}_k, \hat{\Sigma}_k)\hat{\alpha}_k \right)$$

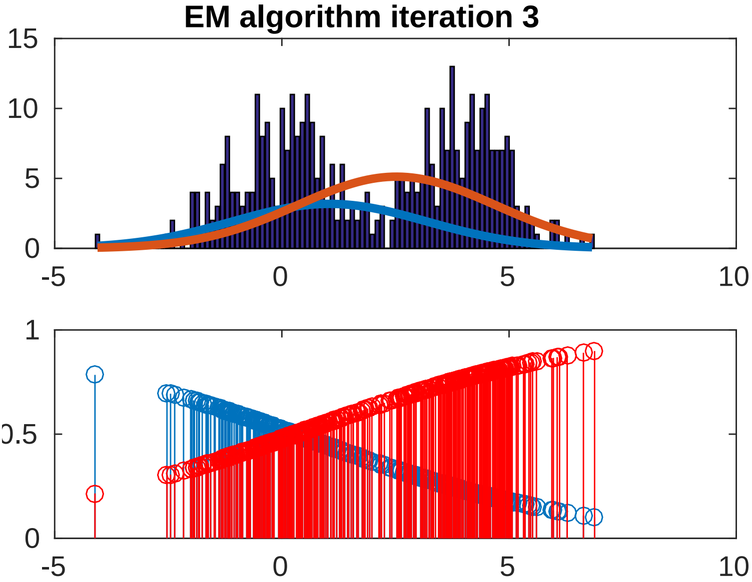
Expectation Maximization (EM) algorithm



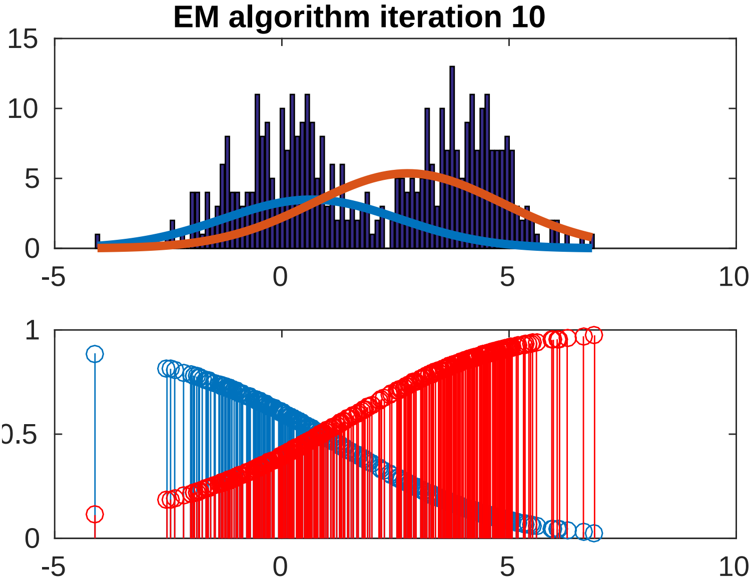
Expectation Maximization (EM) algorithm



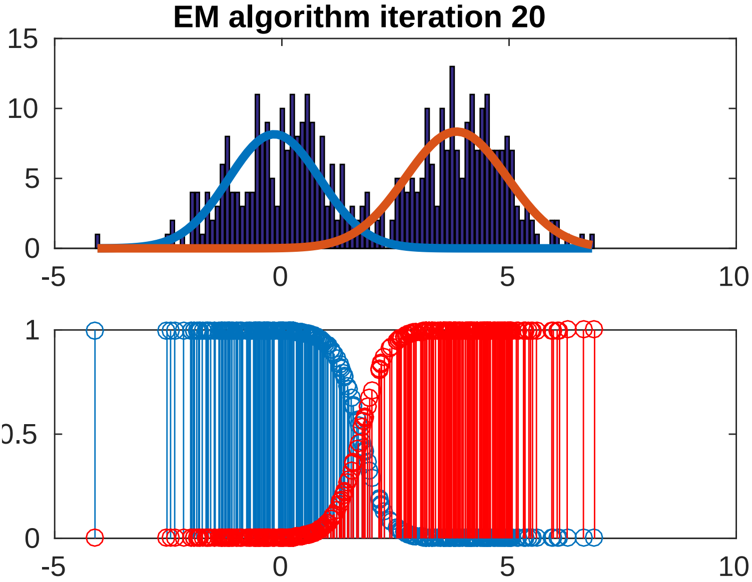
Expectation Maximization (EM) algorithm



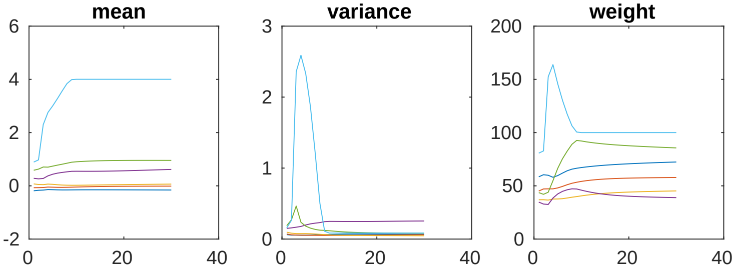
Expectation Maximization (EM) algorithm



Expectation Maximization (EM) algorithm



Expectation Maximization (EM) algorithm



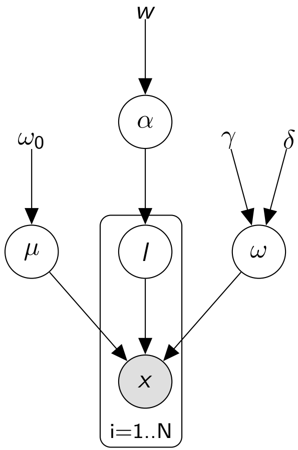
Bayesian treatment

Joint distribution:

$$p(x, l|\alpha) = p(x|l)p(l),$$
$$p(x|l) = \prod_k \mathcal{N}(\mu_k, \omega_k)^{l_k},$$
$$p(l_k = 1|\alpha) = \alpha_k, \sum_k \alpha_k = 1.$$

Priors

$$p(\mu_k|\omega_k) = \mathcal{N}(0, \infty),$$
$$p(\omega_k) = G(0, 0),$$
$$p(\alpha_k) = Di(w_k) = \frac{\Gamma(\sum_k w_k)}{\prod_k \Gamma(w_k)} \prod_k \alpha_k^{w_{0,k}-1},$$



Variational Bayes for Mixtures

Joint likelihood:

$$\log p(l, \omega, \mu, w) \propto \sum_{i,k} l_{i,k} \left[\frac{1}{2} (\log \omega_k - (x_i - \mu_k)\omega_k(x_i - \mu_k)) + \log \alpha_k \right] - \log \omega_k + (w_{0,k} - 1) \log \alpha_k,$$

Factors, with $n_k = \sum_i \hat{l}_{i,k}$

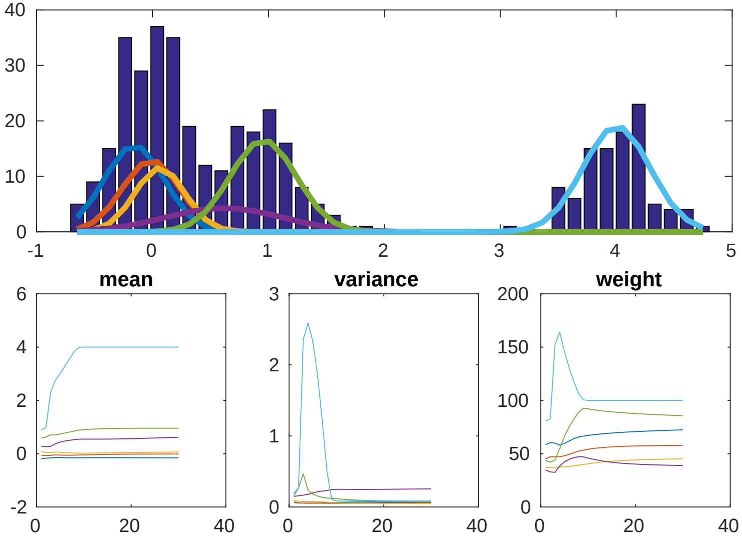
$$q(\omega_k|x) = G(\gamma, \delta), \qquad \gamma = n_k \qquad \delta = \sum_i \hat{l}_{i,k}(x_i - \hat{\mu}_k)^2 + n_k\hat{\sigma}_k,$$

$$q(\mu_k|x) = \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k), \qquad \hat{\sigma}_k = (n_k\hat{\omega}_k)^{-1} \qquad \hat{\mu}_k = \frac{1}{n_k}\sum_i l_{i,k}x_i.$$

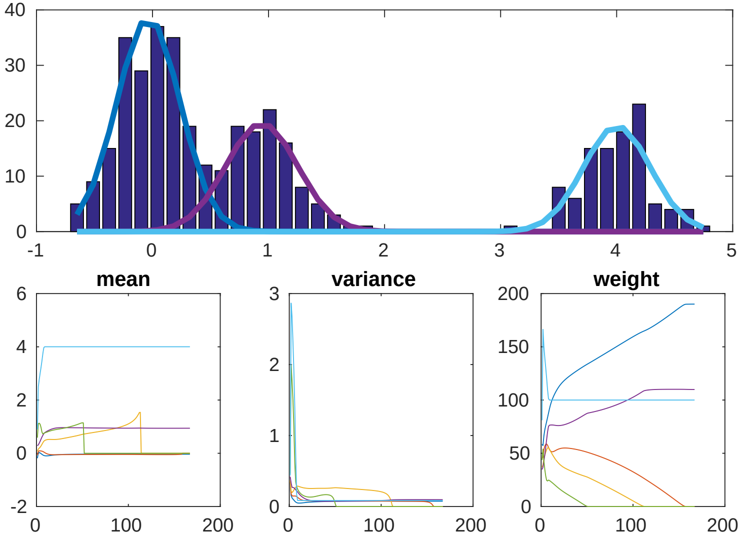
$$q(w|x) = Di(w) \qquad w_k = n_k$$

$$q(l_i|x) = Mu(\lambda) \qquad \hat{l}_{i,k} = \frac{\lambda_{i,k}}{\sum_i \lambda_{i,k}} \qquad \lambda_{i,k} = \exp\frac{1}{2}\left[-(x_i - \hat{\mu}_k)\hat{\omega}_k(x_i - \hat{\mu}_k) \langle \log \omega_k \rangle - \sigma_\mu \hat{\omega}_k\right]$$

EM: $\mu_{true} = \{0, 1, 4\}$, fit $K = 6$ components



VB: $\mu_{true} = \{0, 1, 4\}$ fit $K = 6$ components



Mixture of Gaussians in higher dimensions

Multivariate Gaussians in dimension d:

$$
\begin{aligned}
x &\sim \mathcal{N}(\mu, \Omega^{-1}), \\
\mu &\sim \mathcal{N}(\mu_0, (\tau\Omega)^{-1}) \\
\Omega &\sim \mathcal{W}(V, \nu),
\end{aligned}
$$

where $\mathcal{W}$ is the Wishart distribution with ν degrees of freedom.

Covariance matrix:

full covariance: effective number of data $n_k > d$, $O(d^2)$,

scaled identity: homogenous noise σI, (k-means),

diagonal: ignoring rotation of ellipses,

low rank: only selected principal components,

...

Mixture of Gaussians in higher dimensions

Initialization:

random: over what space? cubic...

LHS: latin hypercube sampling

Number of component:

very many: slow convergence

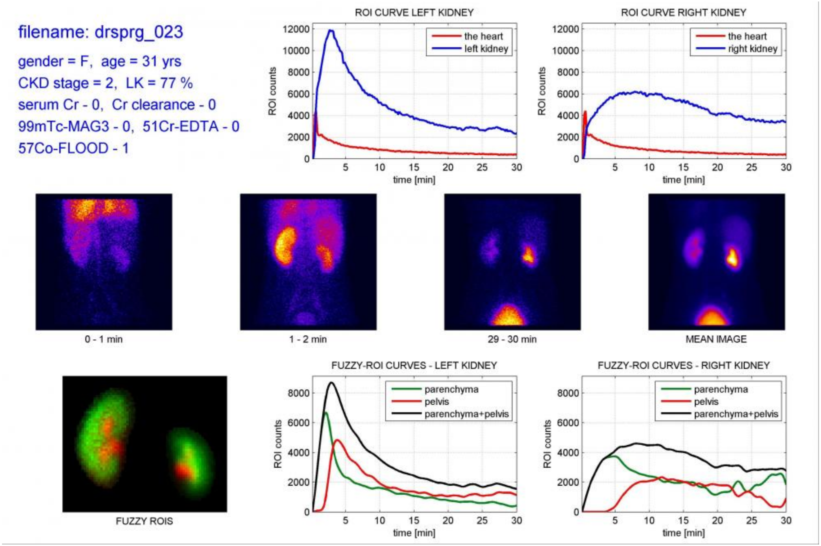
birth and death: random generation

split and merge: evaluate which component to split and/or which two components join into one.

problematic.

Challenge: Patlak Rutland plot

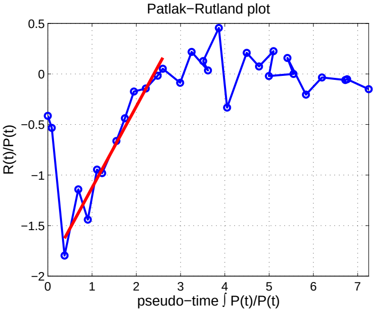
Sequence of scintigraphic images of kidneys.

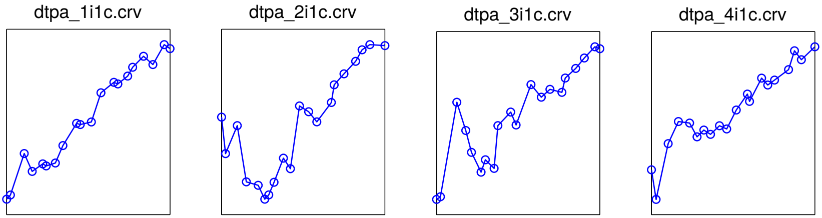


Challenge: Patlak Ruland plot

Patlak Rutland plot is a ratio of parenchyma curve over integral of heart curve.

- typically starts around 1min
- typically ends around 3min
- with outliers
- the slope is a diagnostically important





Mixtures

- Mixture of linear regressions

$$p(y|x) = \sum_k \alpha_k \mathcal{N}(X\theta_k, \sigma_k)$$

- Mixture of Gamma, Beta distributions – for positive support,
- Mixture of factor analyzers,
- Mixture of dynamic models,

Same basic principle:

- define latent variable with indicator of x being generated from each component.

Assignment I

Load data Patlak.mat
35 studies with:

- xpr x axis
- ypr y axis
- name name of he study
- int_start index where the linear part can start
- int_end index where the linear part should end

Assignment	points
find slope of linear part for all 35 studies	
a) built-in function	10
b) own code (WLS + WLS with 10$\times$variance)	25

Assignment II

Simulate 2d mixtures with components:

$$
\begin{aligned}
\mu_1 &= [1; 1], & \Sigma_1 &= \text{eye}(2), & \alpha_1 &= 0.3,\\
\mu_2 &= [-1; 1], & \Sigma_2 &= \text{eye}(2), & \alpha_2 &= 0.3,\\
\mu_3 &= [0; -1], & \Sigma_1 &= \text{diag}([2, 0.1]), & \alpha_3 &= 0.4.
\end{aligned}
$$

Estimation via	points
EM algorithm	20
VB algorithm	30