Mixture Models

Václav Šmídl

March 20, 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:

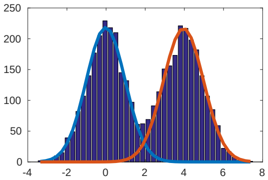
$$p(x) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mu_k, \Sigma_k),$$

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

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

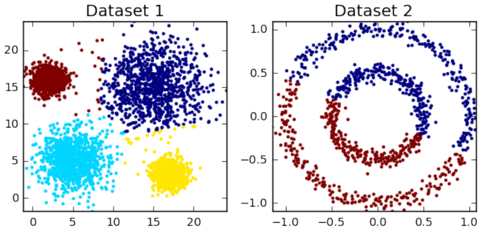
$$p(x) = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(4, 1),$$

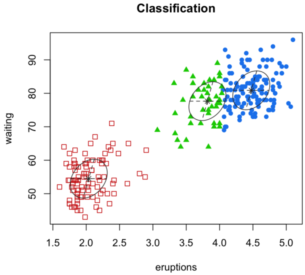
- maximum likelihood,

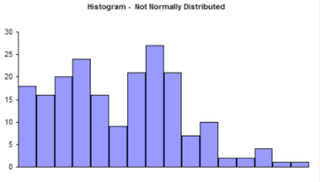


Uses of mixtures

Clustering (supervised, unsupervised, semi-supervised):







Density representation:

Classification:

Mixture estimation

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

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

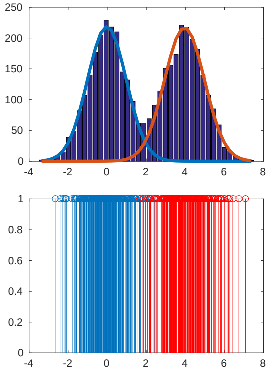
$$p(x|l) = \prod_k \mathcal{N}(\mu_k, \Sigma_k)^{l_k},$$

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

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

Multinomial (Bernouli) distribution.

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



Expectation maximization (EM) algorithm

Joint distribution:

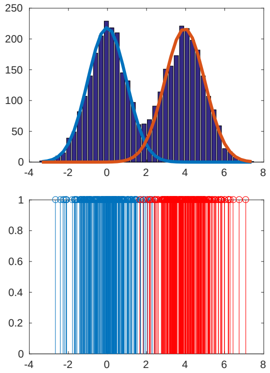
$$p(x, l) = p(x|l)p(l),$$

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

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

Conditional distribution

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



Maximum likelihood

Maximizing log-likelihood

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

$$\frac{d}{d\mu_k}\log p(x) = \frac{1}{\sum_{k=1}^{K}\mathcal{N}(\mu_k, \Sigma_k)\alpha_k}\alpha_k\mathcal{N}(\mu_k, \Sigma_k)\left(-\Sigma_k^{-1}(\mu_k - x)\right),$$

$$= p(l = \epsilon_k|x)\left(-\Sigma_k^{-1}(\mu_k - x)\right)$$

After observing N points

$$\frac{d}{d\mu_k}\log p(x) = \sum_{i=1}^{n} p(l = \epsilon_k|x_i)\left(-\Sigma_k^{-1}(\mu_k - x_i)\right)$$

$$\hat{\mu}_k = \frac{1}{N_k}\sum_i p(l = \epsilon_k|x_i)x_i$$

$$N_k = \sum_i p(l = \epsilon_k|x_i)$$

$p(l = \epsilon_k|x_i)$ is a function of $\mu_k, \Sigma_k, \forall k$.

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

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

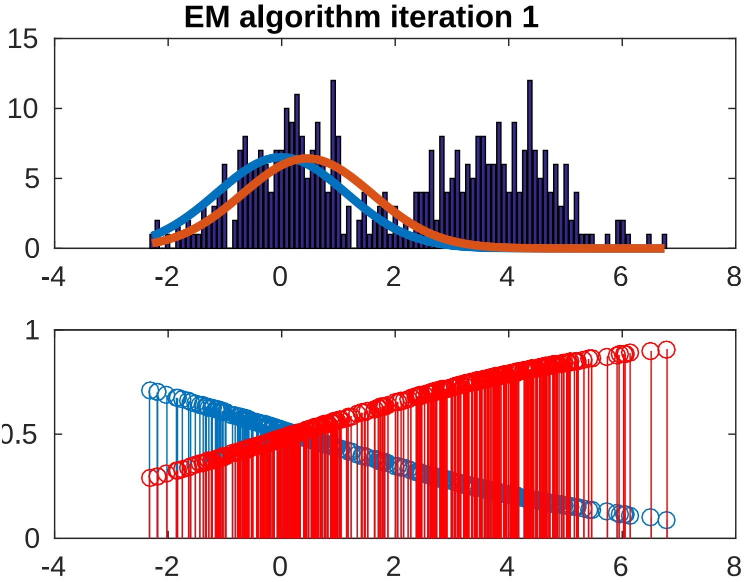
$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_i \hat{l}_{k,i} \left(x_i - \hat{\mu}_k\right) \left(x_i - \hat{\mu}_k\right)^T,$$

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

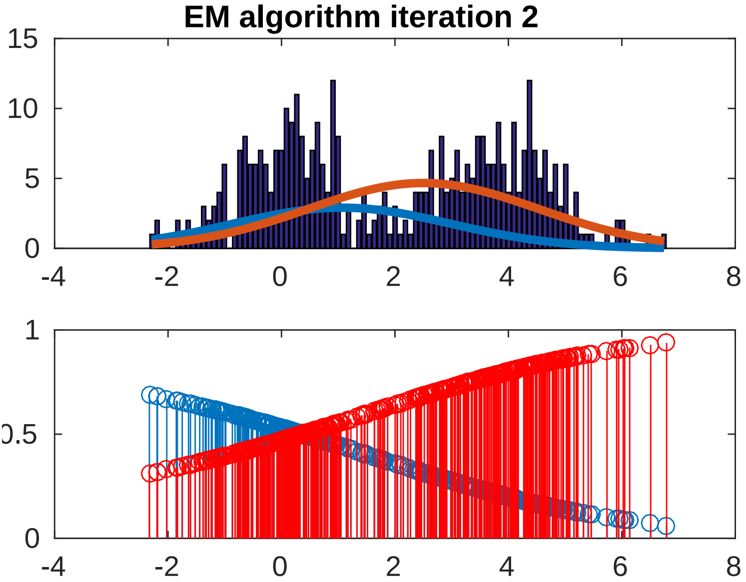
3. (Evaluate log-likelihood)

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

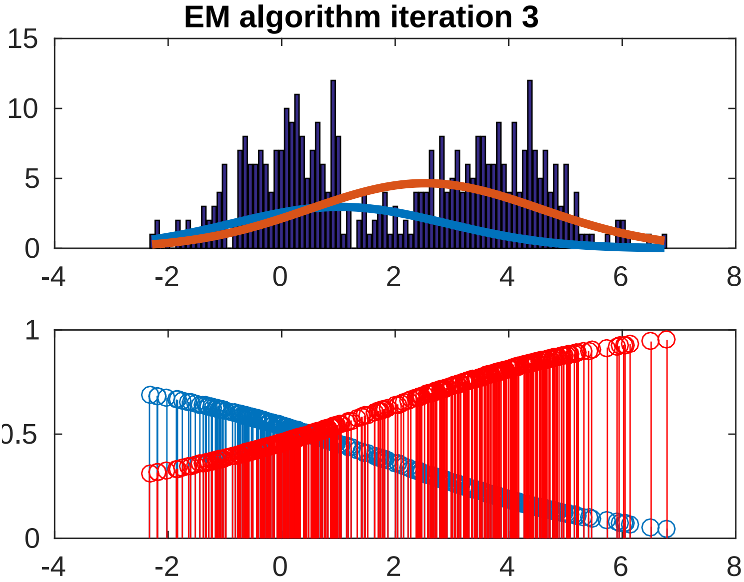
Expectation Maximization (EM) algorithm



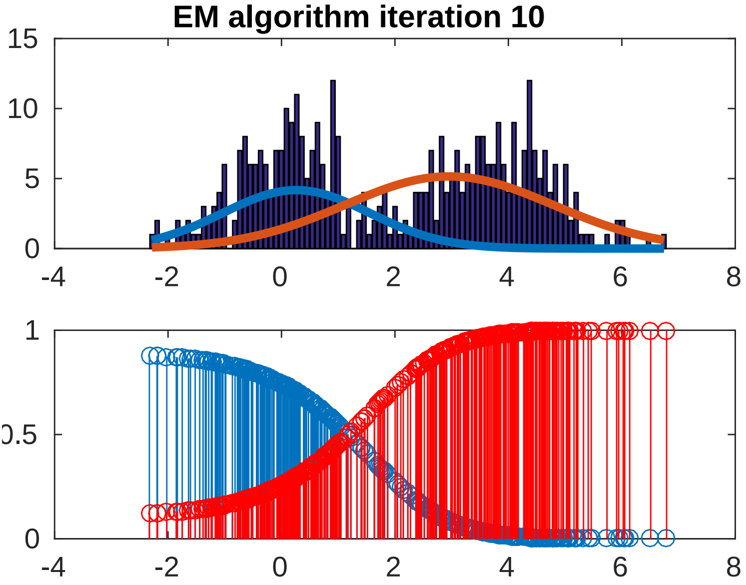
Expectation Maximization (EM) algorithm



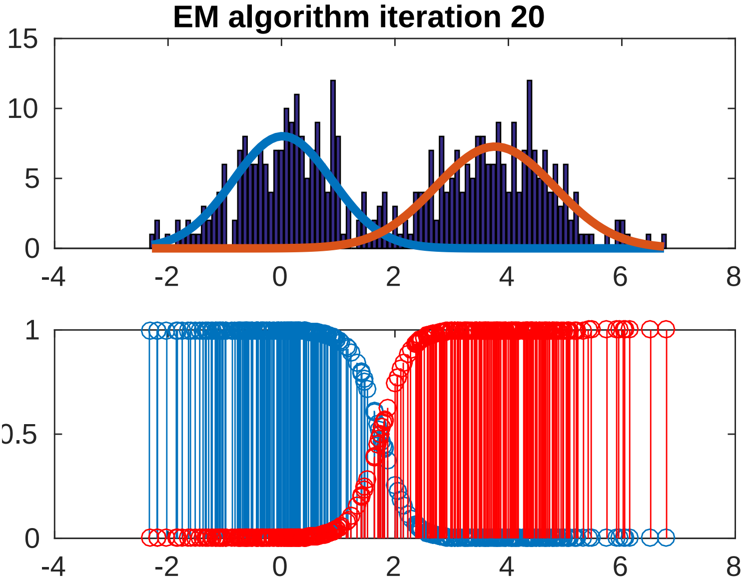
Expectation Maximization (EM) algorithm



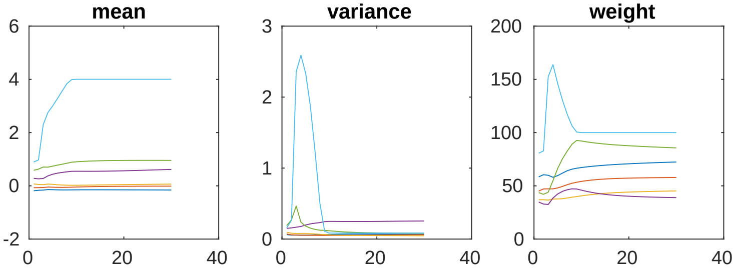
Expectation Maximization (EM) algorithm



Expectation Maximization (EM) algorithm



Expectation Maximization (EM) algorithm



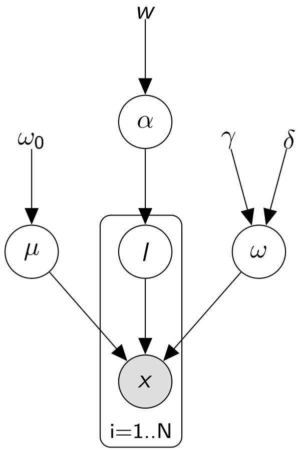
Bayesian treatment

Joint distribution:

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

Priors

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



Variational Bayes for Mixtures

Joint likelihood:

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

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

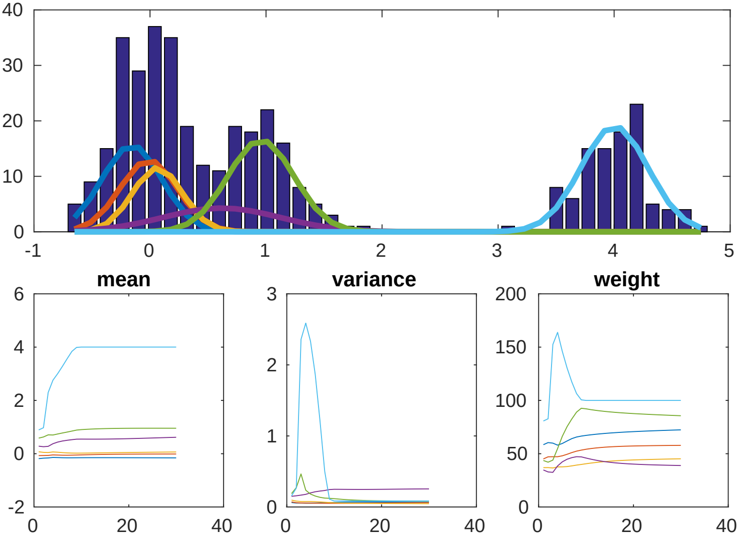
$$q(\omega_k|x) = G(\gamma, \delta), \qquad \gamma = n_k \qquad \delta = \sum_i \hat{I}_{i,k}(x_i - \hat{\mu}_k)^2 + n_k \hat{\sigma}_k,$$

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

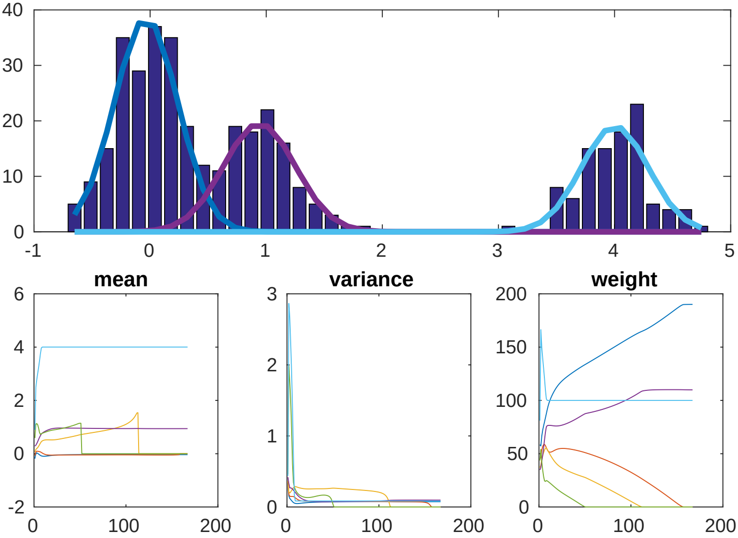
$$q(w|x) = Di(w) \qquad w_k = n_k$$

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

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

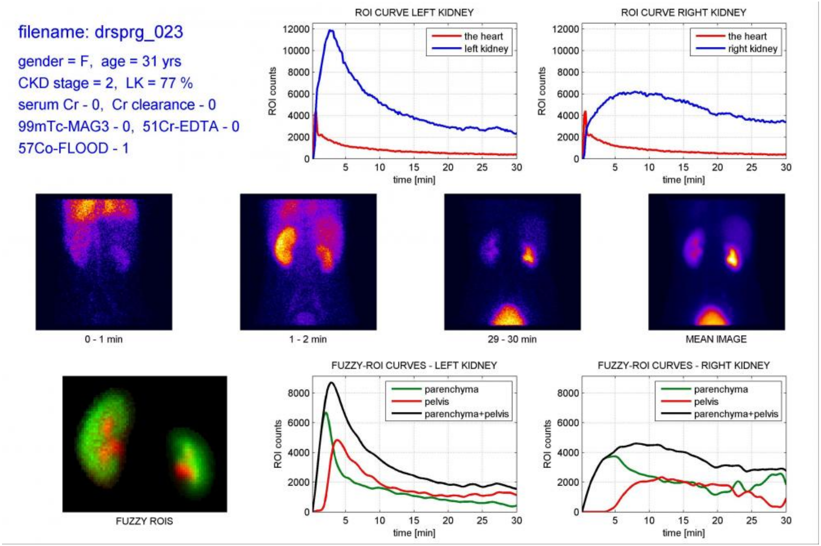
birth and death: random generation

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

problematic.

Challenge: Patlak Rutland plot

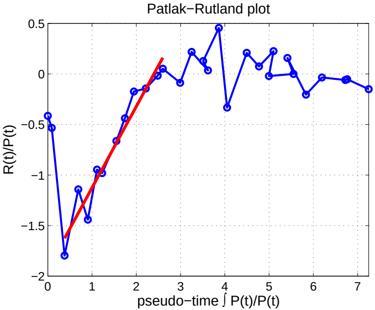
Sequence of scintigraphic images of kidneys.

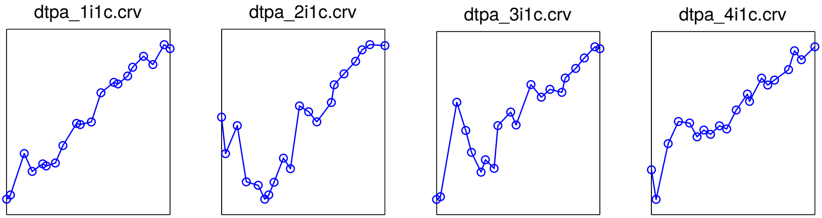


Challenge: Patlak Ruland plot

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

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	25