UiO : Matematisk institutt
Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2021
Chaper 4 (part 1)

Instructor: Odd Kolbjørnsen, oddkol@math.uio.no

Missing data

- Data often (partly) missing
- Censored data (ex 2.3): Time to event not completely known
- Classification of images: Classes to some pixels known, unknown for most of the pixels
- Clustering: Data to be allocated to groups, group membership unknown
- If complete data, Likelihood "often easy"
- Likelihood becomes complicated when data are missing
- Notation:
  - $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$ are complete data
  - $\boldsymbol{X}$ observed,
  - $\boldsymbol{Z}$ missing
  - $\boldsymbol{X} = M(\boldsymbol{Y})$ is observed part
  - Have $f_Y(\boldsymbol{y}|\boldsymbol{\theta})$
  - Want $\max_{\theta} f_X(\boldsymbol{x}|\boldsymbol{\theta})$

$$\boxed{f_X(\boldsymbol{x}|\boldsymbol{\theta}) = \int_{y:M(y)=x} f_Y(\boldsymbol{y}|\boldsymbol{\theta})dy = \int_z f_Y(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})dz}$$

$$\boxed{f_X(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{f_Y(\boldsymbol{y}|\boldsymbol{\theta})}{f_{Z|X}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}}$$

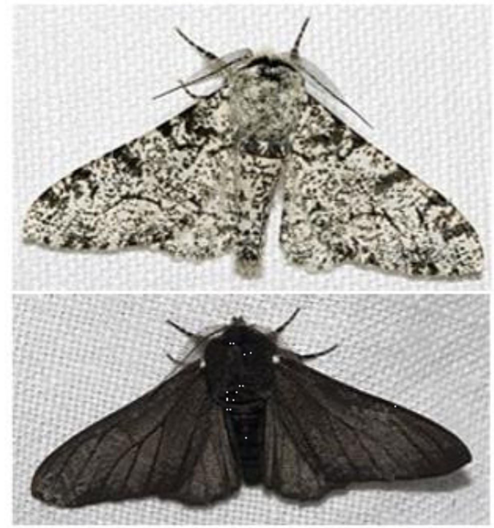
EM algorithm

- Main idea: Iterate between
  - Estimate Z given X, θ (E-step)
  - Estimate θ given (X, Z) (M-step)
- Formally a bit more complicated
  - If complete data, we want to maximize $\log L(\theta|Y)$
  - $\log L(\theta|Y)$ unknown, but given a current value $\theta^{(t)}$ we can estimate it by

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left[\log L(\boldsymbol{\theta}|\boldsymbol{Y}) \;\middle|\; \boldsymbol{x}, \boldsymbol{\theta}^{(t)}\right] \\
&= E[\log f_Y(y|\theta)|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}] \\
&= \int_z \log f_Y(y|\theta)]\, f_{z|x}(z|x, \theta^t) dz
\end{aligned}
$$

- Algorithm:
  1. E-step: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$
  2. M-step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ wrt $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(t+1)}$.
  3. Return to E-step unless a stopping criterion has been met

Peppered Moths - Example



- Color based on one single gene
- Three different allels (C,I,T)
- C is dominant to I, and I is dominant to T

TT	Light-colored
II,IT	Intermediate
CC,CI,CT	Black coloring

- Observing color, interest in frequency of C, I, T
-
- Assume frequencies p_C, p_I, p_T, $p_C + p_I + p_T = 1$

Color	Probability
White	p_T^2
Intermediate	$p_I^2 + 2p_Ip_T$
Black	$p_C^2 + 2p_Cp_I + 2p_Cp_T$

- Observed $(n_T, n_I, n_C) = (341, 196, 85)$
- Complete $(n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$

Options	Count
CC	1
CI (IC)	2
CT (TC)	2
II	1
IT (TI)	$\boxed{2}$
TT	1

Peppered Moths - Likelihood

- Complete data $(n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$
- Complete likelihood (multinomial distribution)

$$
\begin{aligned}
f_{\mathbf{Y}}(\mathbf{y}|\mathbf{p}) =& \frac{n!}{n_{CC}!n_{CI}!n_{CT}!n_{II}!n_{IT}!n_{TT}!} p_C^{2n_{CC}}(2p_Cp_I)^{n_{CI}}(2p_Cp_T)^{n_{CT}}p_I^{2n_{II}}(2p_Ip_T)^{n_{IT}}p_T^{2n_{TT}} \\
=& \frac{n!}{n_{CC}!n_{CI}!n_{CT}!n_{II}!n_{IT}!n_{TT}!} 2^{n_{CI}+n_{CT}+n_{IT}} \times \\
& p_C^{2n_{CC}+n_{CI}+n_{CT}} p_I^{2n_{II}+n_{CI}+n_{IT}} p_T^{2n_{TT}+n_{CT}+n_{IT}}
\end{aligned}
$$

- Complete log-likelihood

$$
\begin{aligned}
\log\{f_{\mathbf{Y}}(\mathbf{y}|\mathbf{p})\} = \log\left(\frac{n!}{n_{CC}!n_{CI}!n_{CT}!n_{II}!n_{IT}!n_{TT}!}\right) + \\
[n_{CI}+n_{CT}+n_{IT}]\log(2) + [2n_{CC}+n_{CI}+n_{CT}]\log(p_C) + \\
[2n_{II}+n_{CI}+n_{IT}]\log(p_I) + [2n_{TT}+n_{CT}+n_{IT}]\log(p_T)
\end{aligned}
$$

- $Q(\mathbf{p}|\mathbf{p}^{(t)}) = E[\log\{f_{\mathbf{Y}}(\mathbf{y}|\mathbf{p})\}|n_C, n_I, n_T, \mathbf{p}^{(t)}]$
- Note: First term do not depend on $\mathbf{p} = (p_C, p_I, p_T)$, not needed in the optimization step!

Peppered Moths – updating E & M

- Complete log-likelihood

$$
\begin{aligned}
Q(\mathbf{p}|\mathbf{p}^{(t)}) = & \text{Const} + E[n_{CI} + n_{CT} + n_{IT}|\boldsymbol{p}^{(t)}]\log(2) + \\
& E[2n_{CC} + n_{CI} + n_{CT}|\boldsymbol{p}^{(t)}]\log(p_C) + \\
& E[2n_{II} + n_{CI} + n_{IT}|\boldsymbol{p}^{(t)}]\log(p_I) + \\
& E[2n_{TT} + n_{CT} + n_{IT}|\boldsymbol{p}^{(t)}]\log(p_T)
\end{aligned}
$$

$$
\mathrm{E}[N_{CC}|n_C, n_I, n_T, \boldsymbol{p}^{(t)}] = n_{CC}^{(t)} = \frac{n_{\mathrm{C}}(p_{\mathrm{C}}^{(t)})^2}{(p_{\mathrm{C}}^{(t)})^2 + 2p_{\mathrm{C}}^{(t)}p_{\mathrm{I}}^{(t)} + 2p_{\mathrm{C}}^{(t)}p_{\mathrm{T}}^{(t)}}
$$

Expectation

- Updating:

$$
p_{\mathrm{C}}^{(t+1)} = \frac{2n_{\mathrm{CC}}^{(t)} + n_{\mathrm{CI}}^{(t)} + n_{\mathrm{CT}}^{(t)}}{2n} \qquad p_{\mathrm{I}}^{(t+1)} = \frac{2n_{\mathrm{II}}^{(t)} + n_{\mathrm{IT}}^{(t)} + n_{\mathrm{CI}}^{(t)}}{2n}
$$

$$
p_{\mathrm{T}}^{(t+1)} = \frac{2n_{\mathrm{TT}}^{(t)} + n_{\mathrm{CT}}^{(t)} + n_{\mathrm{IT}}^{(t)}}{2n},
$$

Maximization

- `Moth_EM.R`

$$\boxed{\mathrm{E}[N_{CC}|n_C, n_I, n_T, \boldsymbol{p}^{(t)}] = n_{CC}^{(t)} = \frac{n_{\mathrm{C}}(p_{\mathrm{C}}^{(t)})^2}{(p_{\mathrm{C}}^{(t)})^2 + 2p_{\mathrm{C}}^{(t)}p_{\mathrm{I}}^{(t)} + 2p_{\mathrm{C}}^{(t)}p_{\mathrm{T}}^{(t)}}}$$

X could be either C,T, or I

$$E\left(N_{CC}\middle|n_c, n_I, n_T, \boldsymbol{p}^{(t)}\right) = n_c \cdot P(CC|\ CX, \boldsymbol{p}^{(t)})$$

$$= n_c \cdot \frac{P(CC\ \&\ CX|\boldsymbol{p}^{(t)})}{P(CX|\boldsymbol{p}^{(t)})}$$

$$= n_c \frac{P(CC|\boldsymbol{p}^{(t)})}{P\left(CC\middle|\boldsymbol{p}^{(t)}\right) + P\left(CI\middle|\boldsymbol{p}^{(t)}\right) + P\left(CT\middle|\boldsymbol{p}^{(t)}\right)}$$

$$P(CC\ ) = p_c^2$$

$$P(CI\ ) = 2p_c p_I$$

$$P(CI\ ) = 2p_c p_T$$

Insert to get result

Moths in R

Data = (85, 196, 341)

```
> show(c(p.old,l.old,NA))
[1] 0.3333333 0.3333333 0.3333333 0.0000000        NA
> more = TRUE
> while(more){
+     n = allele.e(x,p)
+     p = allele.m(x,n)
+     l = loglik(p,n)
+     more = abs(l-l.old)>eps
+     R = sum((p-p.old)^2)/sum(p.old^2)
+     more = R > eps
+     show(c(p,l,R))
+     l.old = l
+     p.old = p
+ }
[1]   0.08199357   0.23740622   0.68060021 -90.55303903   0.57890393
[1]   0.071248952   0.197869614   0.730881433 -68.467059735   0.007993122
[1]  7.085204e-02  1.903604e-01  7.387876e-01 -6.526257e+01  2.058264e-04
[1]  7.083746e-02  1.890227e-01  7.401398e-01 -6.474409e+01  6.163093e-06
[1]  7.083693e-02  1.887869e-01  7.403762e-01 -6.465487e+01  1.894317e-07
[1]  7.083691e-02  1.887454e-01  7.404177e-01 -6.463926e+01  5.851928e-09
>
> ## OUTPUT
> p    # FINAL ESTIMATE FOR ALLELE PROBABILITIES (p.c, p.i, p.t)
[1] 0.07083691 0.18874537 0.74041772
```

Convergence EM

- Iterations increases log likelihood
  - Jensen's inequality $\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \log f_X(\boldsymbol{x}|\boldsymbol{\theta})$
  - For convex $f(x)$, we have:

$$f\big(E(X)\big) \leq E(f(X))$$

- Convergence order $\beta > 0$:

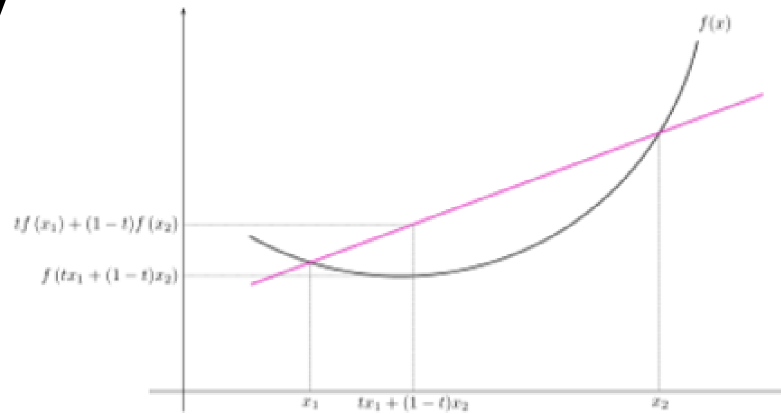
How fast iteration $x^{(t)}$ approaches the true solution x^*

$$\epsilon^{(t)} = x^{(t)} - x^*$$

$$\lim_{t\to\infty} |\epsilon^{(t)}| \to 0$$

$$\boxed{\lim_{t\to\infty} \frac{|\epsilon^{(t+1)}|}{\left|\epsilon^{(t)}\right|^{\beta}} = c}$$

Jensen's inequality



- Convex functions

$$f(tx_1+(1-t)x_2) \leq tf(x_1)+(1-t)f(x_2)$$

for $t \in [0, 1]$

- Finite form

$$f(\sum t_i x_i) \leq \sum t_i f(x_i), \quad t_i \text{ positive}, \sum t_i = 1$$

- Infinite form ($g(\cdot)$ non-negative, integrable):

$$f\left(\frac{1}{b-a}\int_a^b g(x)dx\right) \leq \frac{1}{b-a}\int_a^b f(g(x))dx$$

- Probabilistic form ($g(\cdot)$ density):

$$f(E^g[X]) \leq E^g[f(X)]$$

Ill prove this next week in exercise

Iterations increase the value of, $\ell(\boldsymbol{\theta}|\boldsymbol{x}) = \log(f_X(x|\theta))$

$$f_{z|x}(\mathbf{z}|\mathbf{x},\theta) = \frac{f_y(\mathbf{y}|\theta)}{f_x(\mathbf{x}|\theta)}$$

$$\Downarrow$$

$$\log f_x(\mathbf{x}|\boldsymbol{\theta}) = \log f_y(\mathbf{y}|\boldsymbol{\theta}) - \log f_{z|x}(\mathbf{z}|\mathbf{x},\theta)$$

$$\Downarrow$$

$$E[\log f_x(\mathbf{x}|\boldsymbol{\theta})] = E[\log f_y(\mathbf{y}|\boldsymbol{\theta})] - E[\log f_{z|x}(\mathbf{z}|\mathbf{x},\theta)]$$

Any expectation

- If expectation with respect to $\mathbf{Z}|(\mathbf{x},\boldsymbol{\theta}^{(t)})$,

$$\begin{aligned}\log f_x(\mathbf{x}|\boldsymbol{\theta}) &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - E[\log f_{z|x}(\mathbf{z}|\mathbf{x},\theta)|\mathbf{x},\boldsymbol{\theta}^{(t)}] \\ &= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\end{aligned}$$

$$Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = E\left[\log f_Y(\boldsymbol{y}|\boldsymbol{\theta})|\,\boldsymbol{x},\boldsymbol{\theta}^{(t)}\right]$$

$$H\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = E\left[\log f_{Z|X}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})|\,\boldsymbol{x},\boldsymbol{\theta}^{(t)}\right]$$

Select :
Expectation with respect to the distribution the missing data have under the
under the current estimate of the parameter

Proof: $H\left(\theta^{(t)}\middle|\theta^{(t)}\right) \geq H\left(\theta\middle|\theta^{(t)}\right)$ for any θ

$$\boxed{H\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) = E\left[\log f_{z|x}(\boldsymbol{z}|x,\theta)\,|\, x,\theta^{(t)}\right]}$$

$$H\left(\theta^{(t)}\middle|\theta^{(t)}\right) - H\left(\theta\middle|\theta^{(t)}\right) = E\left\{\log f_{z|x}\left(z\middle|x,\theta^{(t)}\right) - \log f_{z|x}(z|x,\theta)\right\}$$

$$= E\left\{-\log \frac{f_{z|x}(z|x,\theta)}{f_{z|x}\left(z\middle|x,\theta^{(t)}\right)}\right\} \geq -\log E\left\{\frac{f_{z|x}(z|x,\theta)}{f_{z|x}\left(z\middle|x,\theta^{(t)}\right)}\right\}$$

Jensen's

$$= -\log \int \frac{f_{z|x}(z|x,\theta)}{f_{z|x}\left(z\middle|x,\theta^{(t)}\right)} f_{z|x}\left(z\middle|x,\theta^{(t)}\right) dz$$

$$= -\log \int f_{z|x}(z|x,\theta)\, dz = -\log E\{1\} = 0$$

Proof of increasing likelihood

$$\log f_X(\mathbf{x}|\theta^{(t+1)}) - \log f_X(\mathbf{x}|\theta^{(t)})$$
$$= \underbrace{Q(\theta^{(t+1)}, \theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)})} \underbrace{- [H(\theta^{(t+1)}, \theta^{(t)}) - H(\theta^{(t)}, \theta^{(t)})]}$$

In maximization step choose the $\theta^{(t+1)}$ such that it improves the old $\theta^{(t)}$
$> 0.$
If you are not able to improve Q, you have converged

Select $\theta = \theta^{(t+1)}$ and apply result from previous page. Result holds for any θ in particular for $\theta = \theta^{(t+1)}$
≥ 0

$$H(\theta^{(t)}|\theta^{(t)}) \geq H(\theta|\theta^{(t)})$$
$$\Updownarrow$$
$$-H(\theta|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)}) \geq 0$$

$$\boxed{\log f_x(x|\theta^{(t+1)}) > \log f_x(x|\theta^{(t)})}$$

Convergence order (good to know, but need not derive)

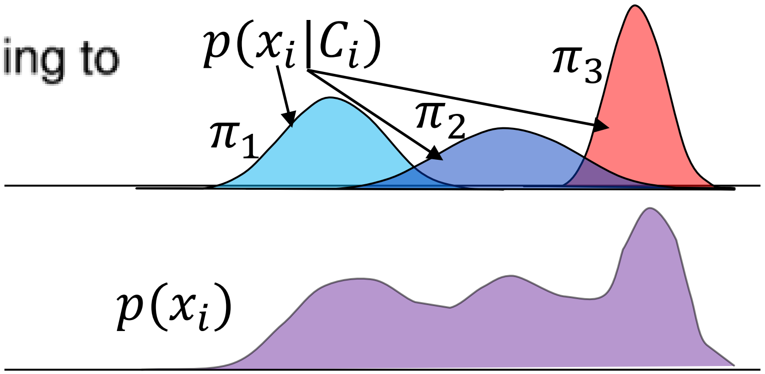
- The EM algorithm defines a mapping $\boldsymbol{\theta}^{(t+1)} = \Psi(\boldsymbol{\theta}^{(t)})$
- When the EM algorithm converges, $\widehat{\boldsymbol{\theta}} = \Psi(\widehat{\boldsymbol{\theta}})$
- Tayler expansion:

$$\begin{aligned}
\boldsymbol{\varepsilon}^{(t+1)} \equiv& \boldsymbol{\theta}^{(t+1)} - \widehat{\boldsymbol{\theta}} \\
=& \Psi(\boldsymbol{\theta}^{(t)}) - \Psi(\widehat{\boldsymbol{\theta}}) \\
\approx& \Psi(\boldsymbol{\theta}^{(t)}) - [\Psi(\boldsymbol{\theta}^{(t)}) + \Psi'(\boldsymbol{\theta}^{(t)})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(t)})] \\
=& \Psi'(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta}^{(t)} - \widehat{\boldsymbol{\theta}}) \\
=& \Psi'(\boldsymbol{\theta}^{(t)})\boldsymbol{\varepsilon}^{(t)}
\end{aligned}$$

- Convergence order β if $\lim_{t\to\infty} \frac{||\boldsymbol{\varepsilon}^{(t+1)}|}{|\boldsymbol{\varepsilon}^{(t)}|^{\beta}} = \rho$

- $p = 1$: $\lim_{t\to\infty} \frac{|\varepsilon^{(t+1)}|}{|\varepsilon^{(t)}|} = \Psi'(\hat{\theta})$, linear convergence
- $p > 1$: Still linear if $-\ell''(\hat{\boldsymbol{\theta}}|\mathbf{x})$ is positive definite
- (Newton's method has convergence order $\beta = 2$)

Example: Mixture Gaussian clustering



- Assume $\mathbf{Y}_i = (X_i, C_i)$ are distributed according to

$$\Pr(C_i = k) = \pi_k, \quad k = 1, \ldots, K$$

$$X_i | C_i = k \sim N(\mu_k, \sigma_k)$$

- The C_i's are missing
- Complete log-density:

$$\begin{aligned}\log f(\mathbf{y}_i) &= \log(\pi_{c_i}) + \log[\phi(x_i; \mu_{c_i}, \sigma_{c_i})] \\ &= \sum_{k=1}^{K} I(c_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k)]]\end{aligned}$$

- Complete log-likelihood:

$$\log f_Y(\mathbf{y}|\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(c_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)]$$

E-step- Mixture Gaussian

- Complete log-likelihood:

$$\log f_Y(\mathbf{y}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} I(c_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)]$$

- E-step (the C_i's the only stochastic part)

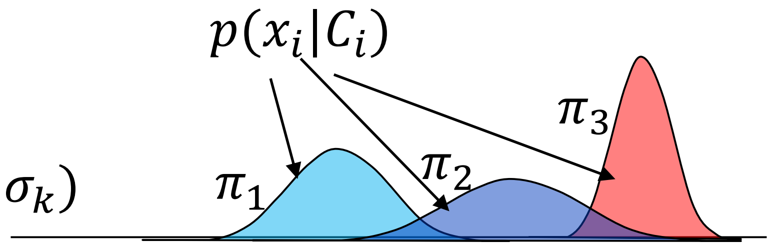
$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E\left[\sum_{i=1}^{n}\sum_{k=1}^{K} I(C_i = k)[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k^2)]|\mathbf{x}, \boldsymbol{\theta}^{(t)}\right] \\
=& \sum_{i=1}^{n}\sum_{k=1}^{K} E[I(C_i = k|\mathbf{x}, \boldsymbol{\theta}^{(t)})][\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k)] \\
=& \sum_{i=1}^{n}\sum_{k=1}^{K} \Pr(C_i = k|\mathbf{x}, \boldsymbol{\theta}^{(t)})[\log(\pi_k) + \log[\phi(x_i; \mu_k, \sigma_k)]
\end{aligned}$$

$$\Pr(C_i = k|\mathbf{x}, \boldsymbol{\theta}^{(t)}) = \frac{\pi_k^{(t)}\phi(x_i, \mu_k^{(t)}, \sigma_k^{(t)})}{\sum_l \pi_l^{(t)}\phi(x_i, \mu_l^{(t)}, \sigma_l^{(t)})}$$

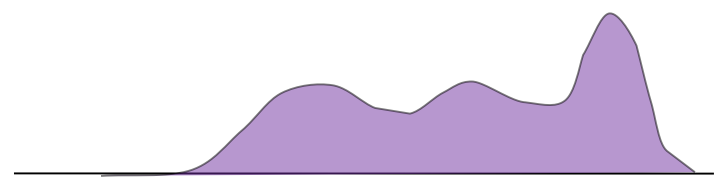
$$\boxed{\Pr(C_i = k|\mathbf{x}, \boldsymbol{\theta}^{(t)}) = \frac{\pi_k^{(t)}\phi(x_i, \mu_k^{(t)}, \sigma_k^{(t)})}{\sum_l \pi_l^{(t)}\phi(x_i, \mu_l^{(t)}, \sigma_l^{(t)})}}$$

$$P(C_i = k) = \pi_k$$

$$p(x_i|C_i = k) = \phi(x_i; \mu_k, \sigma_k)$$



$$p(x_i) = \sum_l \pi_l \phi(x_i; \mu_l, \sigma_l)$$



$$P(C_i = k\,|X_i = x_i) = \frac{p(C_i = k\ \&\ X_i = x_i)}{p(X_i = x_i)} = \frac{P(C_i = k)\, p(x_i|C_i = k)}{p(x_i)}$$

M-step- Mixture Gaussian

- M-step: Taking into account $\sum_{k=1}^{K}\pi_k = 1$:

$$Q_{lagr}(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\Pr(C_i = k|\mathbf{x},\theta^{(t)})[\log(\pi_k)+\log[\phi(x_i;\mu_k,\sigma_k^2)]+\lambda(1-\sum_{k=1}^{K}\pi_k)$$

$$\frac{\partial}{\partial\pi_k}Q_{lagr}(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\Pr(C_i = k|\mathbf{x},\theta^{(t)})\pi_k^{-1} - \lambda$$

$$\Downarrow$$

$$\pi_k^{(t+1)} = \frac{\sum_{i=1}^{n}\Pr(C_i = k|\mathbf{x},\theta^{(t)})}{\lambda}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Pr(C_i = k|\mathbf{x},\theta^{(t)})$$

$$\sum_{k=1}^{K}\sum_{i=1}^{n}\frac{\Pr(C_i = k|x,\theta^{(t)})}{\lambda} = 1$$

$$\frac{1}{\lambda}\sum_{i=1}^{n}\underbrace{\sum_{k=1}^{K}\Pr(C_i = k|x,\theta^{(t)})}_{=1} = 1$$

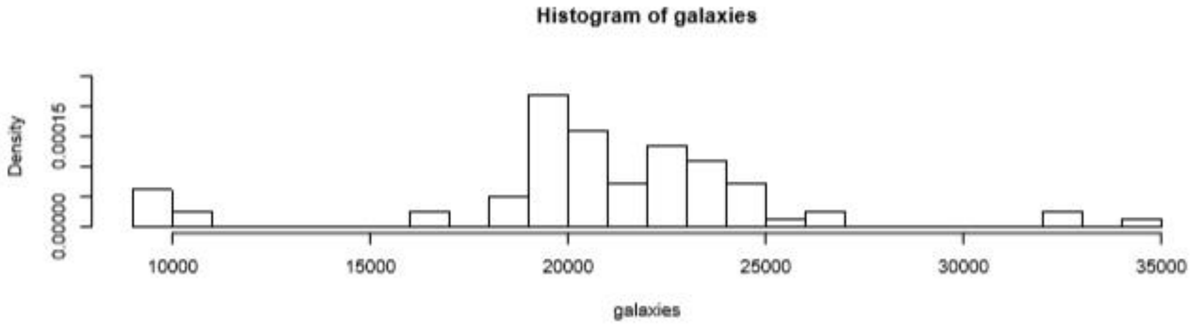
Similarly

$$\mu_k^{(t+1)} = \frac{1}{n\pi_k^{(t+1)}}\sum_{i=1}^{n}\Pr(C_i = k|\mathbf{x},\theta^{(t)})x_i$$

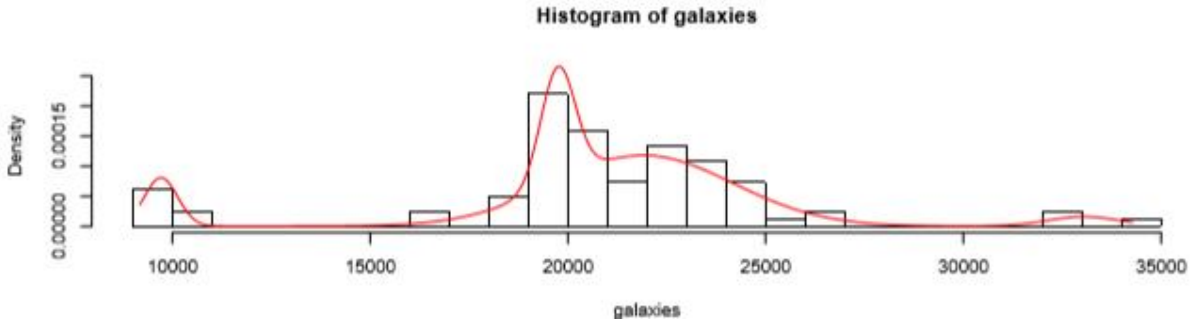
$$(\sigma_k^2)^{(t+1)} = \frac{1}{n\pi_k^{(t+1)}}\sum_{i=1}^{n}\Pr(C_i = k|\mathbf{x},\theta^{(t)})(x_i - \mu_k^{(t+1)})^2$$

Examples galaxy

- A numeric vector of velocities in km/sec of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. Multimodality in such surveys is evidence for voids and superclusters in the far universe.



- `galaxies_EM.R`



EM clustering of Old Faithful eruption data.