UiO : Matematisk institutt

Det matematisk-naturvitenskapelige fakultet

STK-4051/9051 Computational Statistics Spring 2022
Chaper 4

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

Last time

- Examples IRLS, combinatorial optimization
- EM algorithm $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E[\log f_Y(\boldsymbol{y}|\boldsymbol{\theta})|\,\boldsymbol{x},\boldsymbol{\theta}^{(t)}\,]$
  - Missing data (Moths)
  - Proof of increasing log likelihood when using EM algorithm

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

The generality of the proof is related to the connection between optimization of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to achieve an increased value of $f_X(\boldsymbol{x}|\boldsymbol{\theta})$

If $f_X(\boldsymbol{x}|\boldsymbol{\theta})$ is not “well behaved” we might get a local optimum

The problem of missing data

- You design an experiment such that your likelihood is:

$$f_Y(\boldsymbol{y}|\boldsymbol{\theta})$$

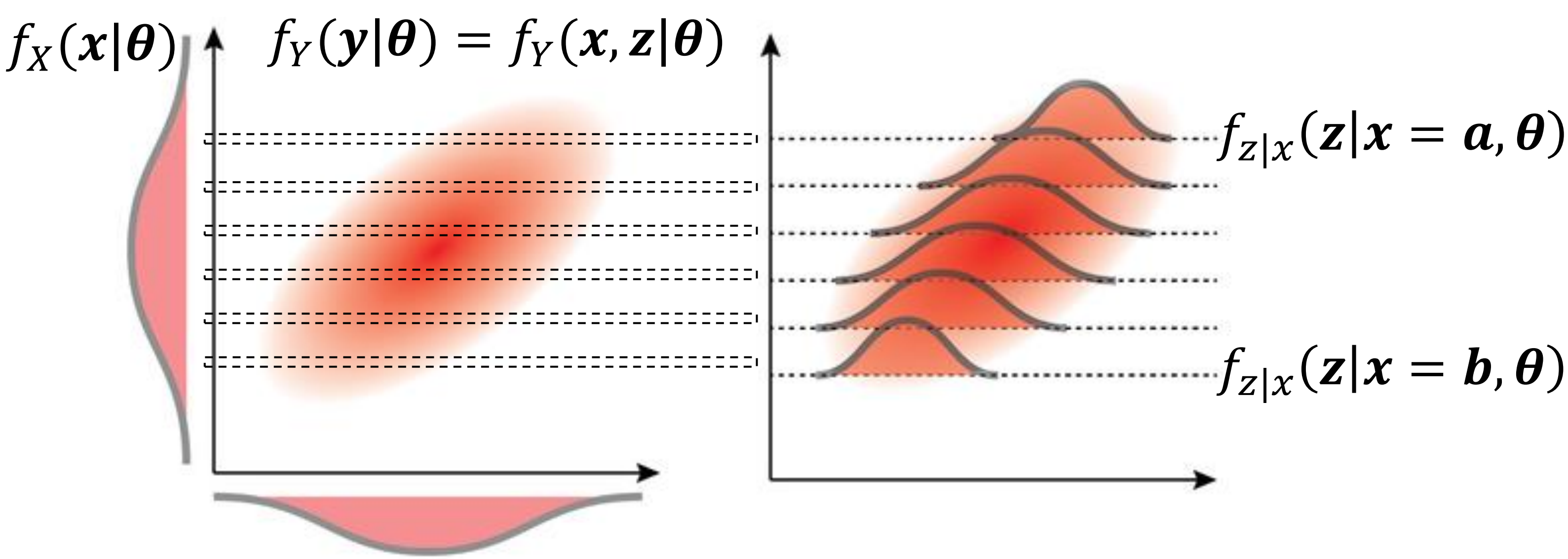
You know how to link your parameters to the data

- Then the data gets back and parts of your dat are missing, you get $\boldsymbol{x}$, you do not get $\boldsymbol{z}$ now you need to link the parameters to the «actual observations»:

$$f_X(\boldsymbol{x}|\boldsymbol{\theta}) = \int f(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})d\boldsymbol{z}$$

- Luckily you have learned about EM algorithm

Joint, marginal and conditional distribution



$$f_X(\boldsymbol{x}|\boldsymbol{\theta}) = \int_z f_Y(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{\theta})dz$$

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

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

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

EM recap

- Notation:
  - $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$ are complete data
  - $\boldsymbol{X}$ observed,
  - $\boldsymbol{Z}$ missing (or hidden)
  - Have $f_Y(\boldsymbol{y}|\boldsymbol{\theta})$
  - Want $\max_{\theta} f_X(\boldsymbol{x}|\boldsymbol{\theta})$

$$\boxed{f_X(\boldsymbol{x}|\boldsymbol{\theta}) = \int_z f_Y(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})dz}$$

Marginal likelihood

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

Complete likelihood

We maximize:

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

Expected value of the complete log likelihood
given the observed data
using the current estimate of the parameter

The Q function vs marginal log likelihood

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

The expected value of the complete log likelihood
given the observed data and the curent estimate of parameter

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

What we want to optimize $\ell(\boldsymbol{\theta}|\boldsymbol{x})$

What EM tells us to optimize

Thm Last time: by optimizing $Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right)$ we optimize $\log(f_X(x|\boldsymbol{\theta}))$

Today

- EM for Mixture Gaussian clustering
- EM in Exponential family
- Variance estimate in EM
- Bootstrap
- EM for hidden Markov model

- Stochastic gradient decent
  - What it is
  - Proof of convergence

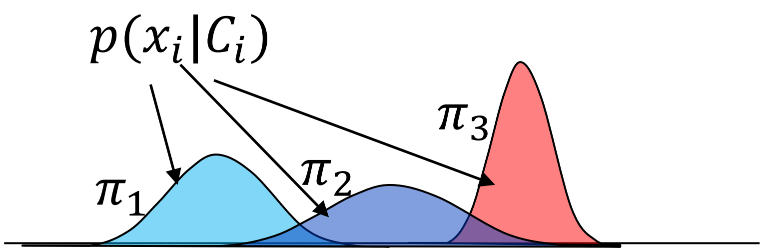
Example: Mixture Gaussian clustering

Class probability

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

Conditional probability given class

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

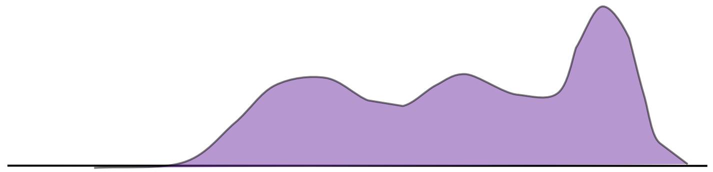


Joint distribution

$$p(x_i, C_i = k) = \pi_k \phi(x_i; \mu_k, \sigma_k) = \prod_{l=1}^{K} \left(\pi_k \phi(x_i; \mu_k, \sigma_k)\right)^{I(l=k)}$$

Marginal distribution of x_i

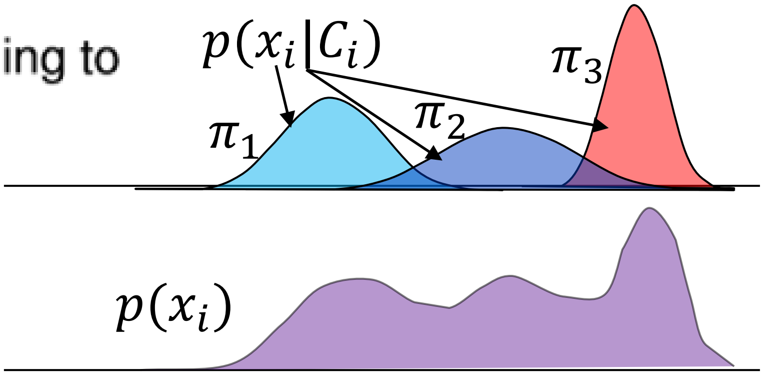
$$p(x_i) = \sum_l \pi_l \phi(x_i; \mu_l, \sigma_l)$$



Complete log likelihood for one individual

$$\log(p(x_i, C_i = k)) = \log(\pi_k) + \log(\phi(x_i; \mu_k, \sigma_k))$$

Example: Mixture Gaussian clustering



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

$$\Pr(C_i = k) = \pi_k, \quad k = 1, \dots, K$$
$$X_i | C_i = k \sim N(\mu_k, \sigma_k)$$

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

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

- 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- 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)

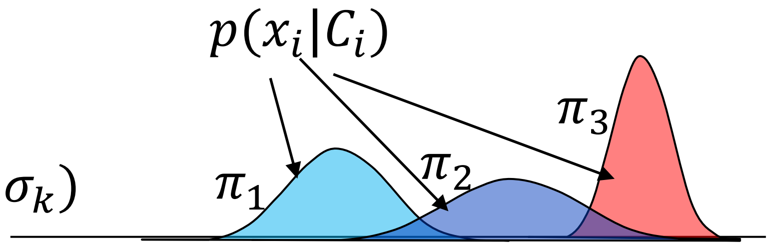
$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E\left[\sum_{i=1}^{n}\sum_{k=1}^{K} I(C_i = k)[\log(\pi_k) + \log[\phi(x_i;\mu_k,\sigma_k^2)]|\mathbf{x},\boldsymbol{\theta}^{(t)}\right] \\
=& \sum_{i=1}^{n}\sum_{k=1}^{K} E[I(C_i = k|\mathbf{x},\boldsymbol{\theta}^{(t)})][\log(\pi_k)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

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

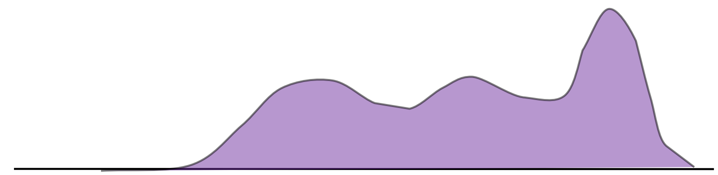
$$\boxed{\Pr(C_i = k|\boldsymbol{x}, \theta^{(t)}) = \frac{\pi_k^{(t)}\phi(x_i, \mu_k^{(t)}, \sigma_k^{(t)})}{\sum_l \pi_l^{(t)}\phi(x_i, \mu_l^{(t)}, \sigma_l^{(t)})}}$$

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

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



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



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

M-step- Mixture Gaussian

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

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

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

$$\Downarrow$$

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

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

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

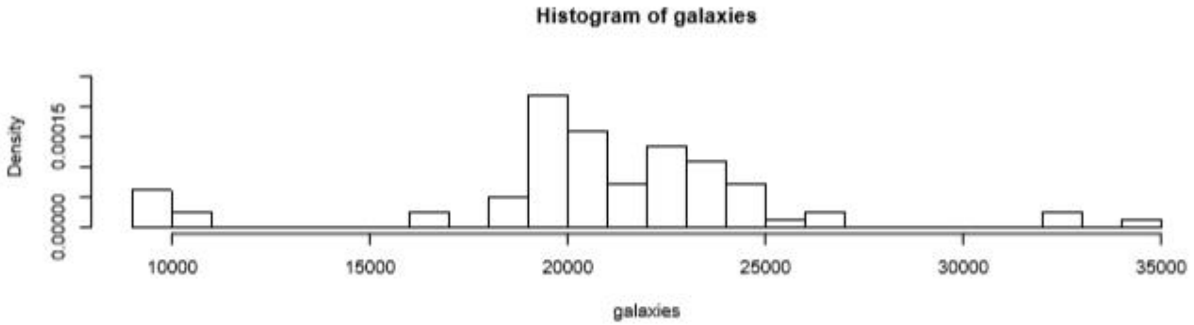
Similarly

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

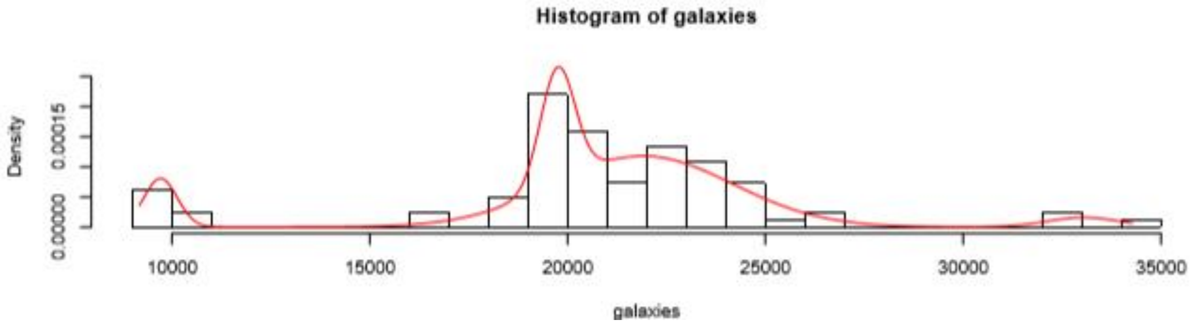
$$(\sigma_k^2)^{(t+1)} = \frac{1}{n\pi_k^{(t+1)}} \sum_{i=1}^{n} \Pr(C_i = k|\mathbf{x}, \theta^{(t)})(x_i - \mu_k^{(t+1)})^2$$

Examples galaxy

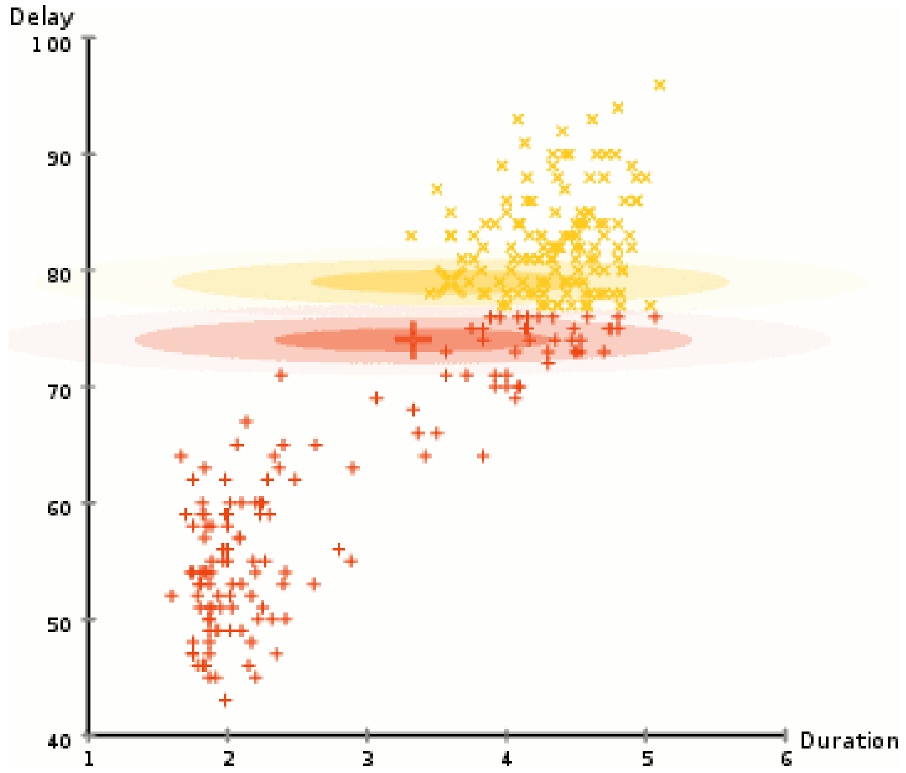
- A numeric vector of velocities in km/sec of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. Multimodality in such surveys is evidence for voids and superclusters in the far universe.



- `galaxies_EM.R`



EM clustering of Old Faithful eruption data.



By Chire - Own work, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=20494862

EM in exponential family

- The Exponential family:

$$f_y(\mathbf{y}|\boldsymbol{\theta}) = c_1(\mathbf{y})c_2(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^T\mathbf{s}(\mathbf{y})\}$$

- Includes
  - binomial, multinomial, Poisson, Gaussian, Gamma,...
- $\boldsymbol{s}(\boldsymbol{y})$ is a sufficient statistic:

$$\begin{aligned} f_s(\boldsymbol{s}|\boldsymbol{\theta}) &= \int_{\boldsymbol{y}:\boldsymbol{s}(\boldsymbol{y})=\boldsymbol{s}} f_y(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{y} \\ &= \int_{\boldsymbol{y}:\boldsymbol{s}(\boldsymbol{y})=\boldsymbol{s}} c_1(\boldsymbol{y})c_2(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^T\boldsymbol{s}(\mathbf{y})\}d\boldsymbol{y} \\ &= c_2(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^T\boldsymbol{s}\}\int_{\boldsymbol{y}:\boldsymbol{s}(\boldsymbol{y})=\boldsymbol{s}} c_1(\boldsymbol{y})d\boldsymbol{y} \\ &= c_2(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^T\boldsymbol{s}\}g(\boldsymbol{s}) \end{aligned}$$

$$f(\boldsymbol{y}|\boldsymbol{s};\boldsymbol{\theta}) = \frac{f_y(\mathbf{y}|\boldsymbol{\theta})}{f_s(\boldsymbol{s}|\boldsymbol{\theta})} = \frac{c_1(\mathbf{y})c_2(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^T\mathbf{s}\}}{c_2(\boldsymbol{\theta})\exp\{\boldsymbol{\theta}^T\boldsymbol{s}\}g(\boldsymbol{s})} = \frac{c_1(\mathbf{y})}{g(\boldsymbol{s})}$$

which do not depend on $\boldsymbol{\theta}$!

Why? We can do computations in advance and just identify terms afterwards

Simplifies a lot of standard problems

The EM algorithms in exponential families E & M

- Log-likelihood

$$l(\boldsymbol{\theta}) = \log c_1(\boldsymbol{y}) + \log c_2(\boldsymbol{\theta}) + \boldsymbol{\theta}^T \boldsymbol{s}(\boldsymbol{y})$$

- E-step:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = k + \log c_2(\boldsymbol{\theta}) + \int \boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}$$

- M-step:

What is this?

$$\frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \boxed{\frac{\mathbf{c}_2'(\boldsymbol{\theta})}{c_2(\boldsymbol{\theta})}} + \int \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}$$

$$\boxed{\frac{\partial}{\partial \theta} Q(\theta|\theta^{(t)}) = 0}$$

$$\int_{\boldsymbol{y}} c_1(\boldsymbol{y}) c_2(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\} d\boldsymbol{y} = 1 \Leftrightarrow$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int_{\boldsymbol{y}} c_1(\boldsymbol{y}) c_2(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\} d\boldsymbol{y} = 0 \Leftrightarrow$$

$$\int_{\boldsymbol{y}} c_1(\boldsymbol{y}) [c_2'(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\} + c_2(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\} \boldsymbol{s}(\boldsymbol{y})] d\boldsymbol{y} = 0 \Leftrightarrow$$

$$\frac{\mathbf{c}_2'(\boldsymbol{\theta})}{c_2(\boldsymbol{\theta})} + E[\mathbf{s}(Y); \boldsymbol{\theta}] = 0$$

$$\boxed{\int s(y) f_{z|x}(z|x, \theta^{(t)}) dz = E[s(Y); \theta]}$$

$$\boxed{\frac{\boldsymbol{c}_2'(\theta)}{c_2(\theta)} = -E[s(Y); \theta]}$$

1st term: Multiply with $\frac{c_2(\theta)}{c_2(\theta)}$ and put $\frac{c_2'(\theta)}{c_2(\theta)}$ outside integral, What remains inside integrates to 1

Results

Compute this integral with your old theta
A bit sloppy to say that this is the E-step
the real E-step is to compute the
maximizing function:

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

In the exponential family it turns out that
what you need for computations is
the expectation of the sufficient statistics

- Algorithm

$$\begin{aligned} &\text{E-step} \quad \boldsymbol{s}^{(t)} = E[\boldsymbol{s}(\boldsymbol{Y})|\boldsymbol{x};\boldsymbol{\theta}^{(t)}] \\ &\text{M-step} \quad \boldsymbol{\theta}^{(t+1)} \text{ solves } E[\boldsymbol{s}(\boldsymbol{Y})|\boldsymbol{\theta}] = \boldsymbol{s}^{(t)} \end{aligned}$$

$E[s(\boldsymbol{Y})|\boldsymbol{x},\theta]$ is the conditional expectation of the complete data given the observed data.

$E[s(\boldsymbol{Y})|\theta]$ is the unconditional expectation of the complete data

- Peppered Moths
  - Multinomial distribution part of exponential family with $\boldsymbol{\theta} = (\log p_C, \log p_I, \log p_T)$
  - Sufficient statistics:

$$\begin{aligned} S_1 &= 2n_{CC} + n_{CI} + n_{CT} & E[S_1] &= 2nP_C \\ S_2 &= 2n_{II} + n_{CI} + n_{IT} & E[S_2] &= 2nP_I \\ S_3 &= 2n_{TT} + n_{CT} + n_{IT} & E[S_3] &= 2nP_T \end{aligned}$$

  - Gives directly

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

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

Variance estimate in EM (4.2.3.4)

- Many approaches
- Approximation using, information matrix

$$\text{var}\,(\hat{\theta}) \approx \boldsymbol{J}_X(\boldsymbol{\theta})^{-1}$$

  - $\boldsymbol{J}_X(\boldsymbol{\theta}) = -\ell''(\boldsymbol{\theta}|\boldsymbol{x})$ (observed information matrix)
  - Louis method (4.2.3.1), Just the part about complete and missing information
  - The SEM algorithm (4.2.3.2)
  - Empirical information (4.2.3.4)
  - Numerical Differentiation (4.2.3.5)
- Bootstrapping (4.2.3.3 for EM)
  - General approach (9.1 & 9.2)

$$\log f_X(\mathbf{x}|\boldsymbol{\theta}) = \log f_Y(\mathbf{y}|\boldsymbol{\theta}) - \log f_{Z|X}(\mathbf{z}|\mathbf{x}, \theta)$$

Missing information (4.2.3.1)

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) = Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) - H\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right)$$

So you do not differentiate with respect to second argument

$$-\ell''(\boldsymbol{\theta}|\boldsymbol{x}) = -\boldsymbol{Q}''(\boldsymbol{\theta}|\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=\boldsymbol{\theta}} + \boldsymbol{H}''(\boldsymbol{\theta}|\boldsymbol{\omega})\Big|_{\boldsymbol{\omega}=\boldsymbol{\theta}}$$

$$\boldsymbol{J_X}(\boldsymbol{\theta}) = \boldsymbol{J_Y}(\boldsymbol{\theta}) - \boldsymbol{J_{Z|X}}(\boldsymbol{\theta})$$

Observed information Complete information Missing information

- Nice way of understanding the information loss in missing data
- Sometimes easier to compute $\boldsymbol{J_Y}(\boldsymbol{\theta})$ and $\boldsymbol{J_{Z|X}}(\boldsymbol{\theta})$

Empirical information

- The (expected) Fisher information is defined by

$$\mathbf{I}(\boldsymbol{\theta}) = E\{\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{X})\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{X})\}$$

- Further, since $E[\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{X})] = \mathbf{0}$, we have

$$\mathbf{I}(\boldsymbol{\theta}) = \text{var}[\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{X})]$$

- If we have IID data,

$$\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x}) = \frac{\partial \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{\partial \log f_{\mathbf{X}_i}(\mathbf{x}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \equiv \sum_{i=1}^{n} \boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x}_i)$$

- We can estimate the information for one observation, $\mathbf{I}_1(\boldsymbol{\theta})$ by

$$\widehat{\mathbf{I}}_1(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x}_i)\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x}_i)^T - \frac{1}{n^2}\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x})\boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x})^T$$

while information for all data can be estimated by

$$\widehat{\mathbf{I}}(\boldsymbol{\theta}) = n \cdot \widehat{\mathbf{I}}_1(\boldsymbol{\theta})$$

But how do we compute $\ell'(\boldsymbol{\theta}|\boldsymbol{x})$?

Computing the score function in EM

$$\ell(\boldsymbol{\theta}|\boldsymbol{x}) = Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) - H\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right)$$

- giving

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}|\mathbf{x}) &= H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ &\leq H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \\ &= Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}^{(t)}|\mathbf{x}) \end{aligned}$$

  so $\boldsymbol{\theta}^{(t)}$ is a max point of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}|\mathbf{x})$.

- Assuming smooth functions,

$$\boxed{\frac{\partial}{\partial \boldsymbol{\theta}}\left(Q\left(\boldsymbol{\theta}\middle|\boldsymbol{\theta}^{(t)}\right) - \ell(\boldsymbol{\theta}|\boldsymbol{x})\right) = \mathbf{0}}$$

$$Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \boldsymbol{\ell}'(\boldsymbol{\theta}|\mathbf{x})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$$

- $Q'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$ typically calculated in the M-step of the EM-algorithm!

Bootstrapping (9.1-9.2.2)

General for exchangeable observations $(\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$,
e.g. iid from $f(\boldsymbol{x} \mid \theta)$

$$\hat{F}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{x}_i < \boldsymbol{x})$$

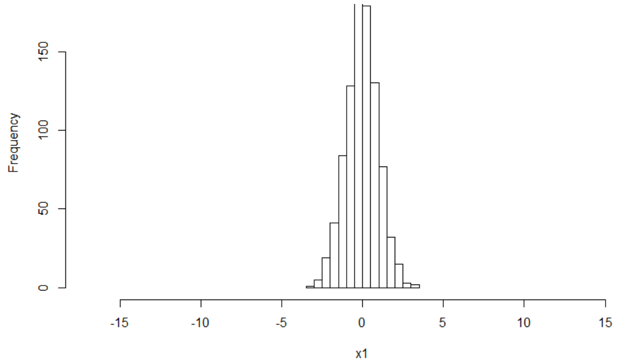
The target parameter is $\theta = T(F)$
We make the **estimate** $\hat{\theta} = T(\hat{F})$ (plug in)
or just $\hat{\theta} = R(\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$ (some function of data)

- In frequentist inference, the randomness in the estimator comes form the uncertainty in the sampled values. This uncertainty is modelled by the probability density $f(\boldsymbol{x} \mid \theta)$.
- We could compute the uncertainty by generating many samples from $f(\boldsymbol{x} \mid \theta)$, and recompute the estimator,
  - but we need many samples from true distribution, we only have one ☹
  - And we do not know the value of θ ☹.
- Two solutions
  - We can get approximate sample from $\hat{F}(\boldsymbol{x})$ [nonparametric bootstrap]
  - We can sample from the distribution $f(\boldsymbol{x} \mid \hat{\theta})$ [parametric bootstrap]

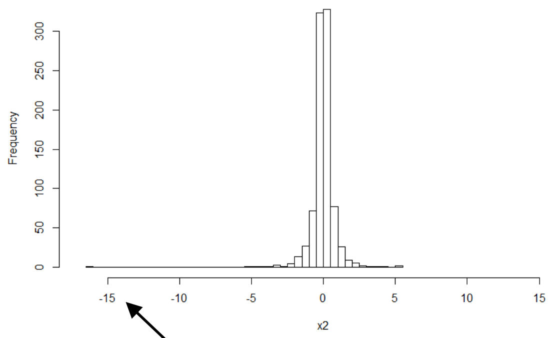
Example: In a symmetric distribution should you estimate the center using the mean or the median?

- If data have a normal distribution the theory says mean.
- But what if the distribution is not known to be normal?

Case 1: normal



Case 2 Student-t



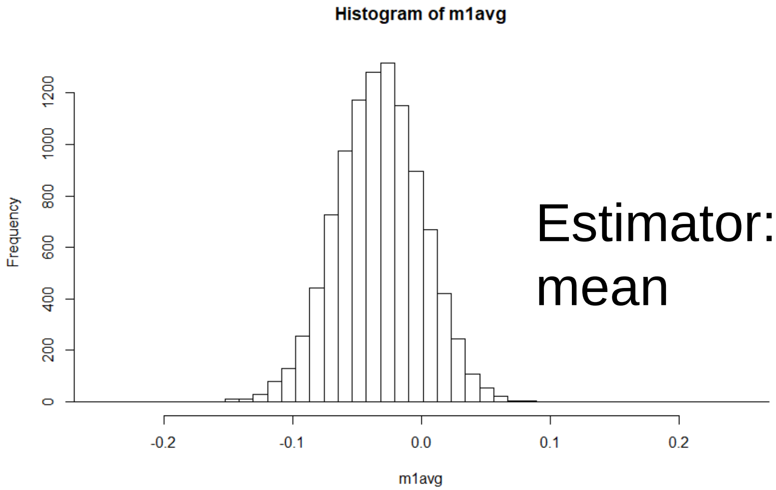
One very low value

Bootstrap code:

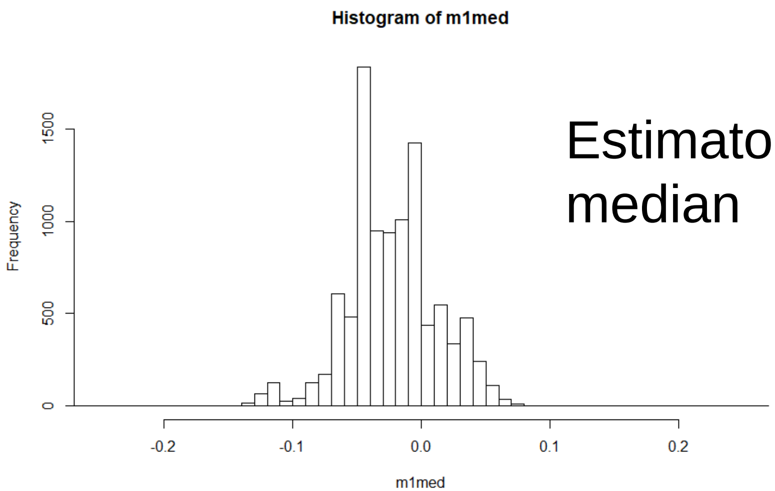
```
m1avg=rep(0,B)
m1med=rep(0,B)

for(b in 1:B)
{
  ind=sample(1:n,n,replace=TRUE)
  m1avg[b]=mean(x1[ind])
  m1med[b]=median(x1[ind])
}
```

Case 1: normal



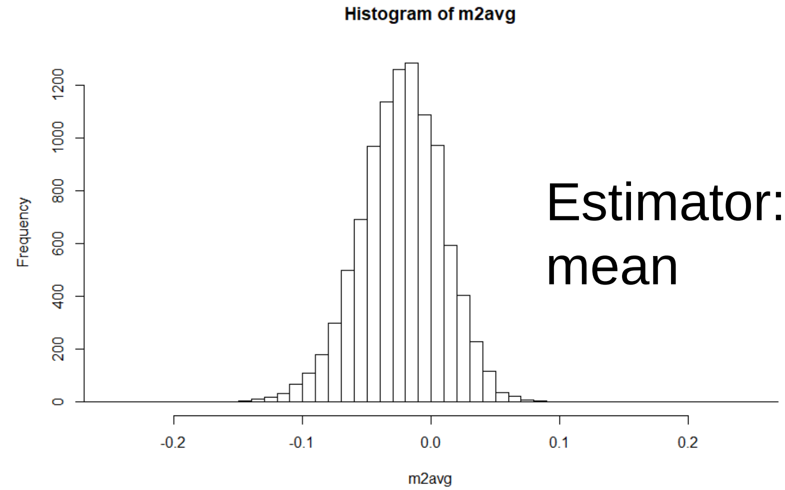
Estimator: mean



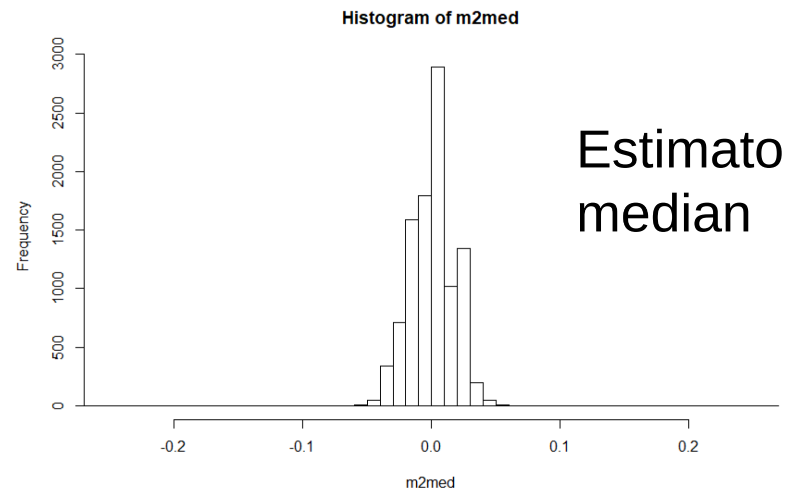
Estimator: median

```
mean(m1avg),mean(m1med),sd(m1avg), sd(m1med))
-0.03323639 -0.02249270  0.03307427  0.03477931
```

Case 2 Student-t



Estimator: mean



Estimator: median

```
:(mean(m2avg),mean(m2med),sd(m2avg), sd(m2med))
 -0.023154073  0.000701744  0.031569047  0.017299653
```

Bootstrapping EM algorithm

- General Bootstrap
- We have $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$, i.e. a way to compute an estimate

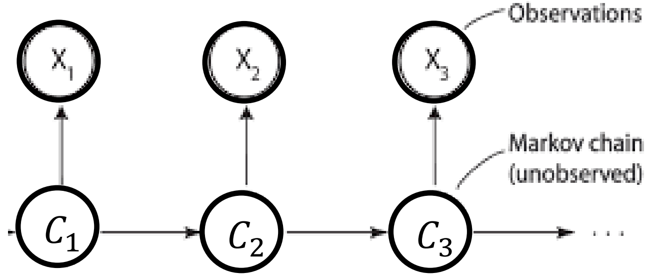
- Algorithm:
  - For j=1,…,B
    - Generate sample $\{\boldsymbol{x_1^*}, \dots, \boldsymbol{x_n^*}\}$, from an approximation of $f_X(\boldsymbol{x}|\theta)$ (parametric / nonparametric)
    - Calculate $\widehat{\boldsymbol{\theta_j}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{x_1^*}, \dots, \boldsymbol{x_n^*})$
  - $\{\widehat{\boldsymbol{\theta_j}}\}_{j=1}^{B}$ can be seen as a samples from the sampling distribution of $\widehat{\boldsymbol{\theta}}$
- Compute variance, quantiles, etc. empirically from $\{\widehat{\boldsymbol{\theta_j}}\}_{j=1}^{B}$

- For the EM algorithm
  - $\widehat{\boldsymbol{\theta}}(\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$ is computed by EM algorithm, $\widehat{\boldsymbol{\theta}}_{EM}(\boldsymbol{x_1}, \dots, \boldsymbol{x_n})$
  - Parametric: Sample $\{\boldsymbol{y_1^*}, \dots, \boldsymbol{y_n^*}\}$ iid $\sim f_Y(\boldsymbol{y}|\theta)$, keep only $\boldsymbol{x}$, i.e. $\{\boldsymbol{x_1^*}, \dots, \boldsymbol{x_n^*}\}$
  - Nonparametric: Sample $\{\boldsymbol{x_1^*}, \dots, \boldsymbol{x_n^*}\}$ with replacement from $\{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\}$

EM-Hidden Markov model

- Assume now $Y_i = (X_i, C_i)$ where i refer to timepoint.
- Model:

$$\begin{aligned}\Pr(C_1 = k) =& \pi_1(k)\\ \Pr(C_i = k|C_{i-1} = j) =& p(k|j),\\ X_i|C_i = k \sim& N(\mu_k, \sigma_k) = f_k(x_i)\end{aligned}$$



- $\{C_i\}$ is a Markov chain and hidden/missing
- Complete likelihood:

$$\begin{aligned}L(\theta) =& \pi_1(c_1)\phi(x_1;\mu_{c_1},\sigma_{c_1})\prod_{i=2}^n p(c_i|c_{i-1})\phi(x_i;\mu_{c_i},\sigma_{c_i})\\ \ell(\theta) =& \log(\pi_1(c_1)) + \log[\phi(x_1;\mu_{c_1},\sigma_{c_1})]+\\ &\sum_{i=2}^n [\log[p(c_i|c_{i-1})] + \log[\phi(x_i;\mu_{c_i},\sigma_{c_i})]]\\ =&\sum_{k=1}^K I(c_1 = k)[\log(\pi_1(k)) + \log[\phi(x_1;\mu_k,\sigma_k)]+\\ &\sum_{i=2}^n\sum_{k=1}^K\sum_{j=1}^K I(c_i = k, c_{i-1} = j)[\log[p(k|j)] + \log[\phi(x_i;\mu_k,\sigma_k)]]\end{aligned}$$

EM - Hidden Markov model

- We get

$$Q(\theta|\theta^{(t)}) = \sum_{k=1}^{K} \Pr(C_1 = k|\mathbf{x}, \theta^{(t)})[\log(\pi_1(k)) + \log[\phi(x_1; \mu_k, \sigma_k)]+$$

$$\sum_{i=2}^{n}\sum_{k=1}^{K}\sum_{j=1}^{K} \Pr(C_i = k, C_{i-1} = j|\mathbf{x}, \theta^{(t)}) \left[\log[p(k|j)] + \log[\phi(x_i; \mu_k, \sigma_k)]\right]$$

- Main problem now: Calculation of

$$\Pr(C_1 = k|\mathbf{x}, \theta^{(t)})$$

$$\Pr(C_i = k, C_{i-1} = l|\mathbf{x}, \theta^{(t)}), \quad i = 2, \ldots, n$$

General idea in hidden Markov Model

Any time All data

$P(C_i|\boldsymbol{x}_{1:n})$

- We compute «forward» (filtering)

$P(C_{i-1}|\boldsymbol{x}_{1:(i-1)})$ Update $P(C_i|\boldsymbol{x}_{1:(i-1)})$ Predict $P(C_i|\boldsymbol{x}_{1:i})$ Update $P(C_{i+1}|\boldsymbol{x}_{1:i})$ Predict $\ldots$ $\boxed{P(C_n|\boldsymbol{x}_{1:n})}$ Update

- Then we compute backward (smoothing)

$P(C_{i+1}|\boldsymbol{x}_{1:n})$ $P(C_i|\boldsymbol{x}_{1:n})$ $P(C_{i-1}|\boldsymbol{x}_{1:n})$ $\cdots$ $P(C_1|\boldsymbol{x}_{1:n})$

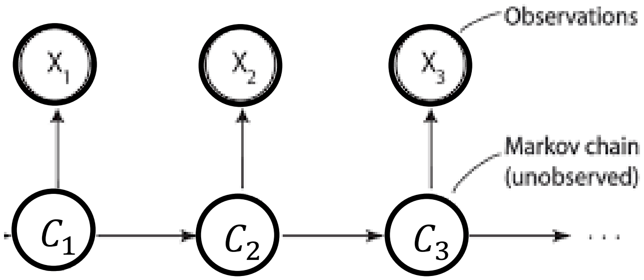
- At the end we combine to get

$$P(C_i, C_{i-1}|\boldsymbol{x}_{1:n})$$

Hidden Markov model

- We have

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= f(\boldsymbol{x}|\boldsymbol{\theta}) \\
&= f(x_1|\boldsymbol{\theta}) \prod_{i=2}^{n} f(x_i|\boldsymbol{x}_{1:i-1};\boldsymbol{\theta})
\end{aligned}
$$



where

$$
f(x_1|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(x_1;\boldsymbol{\theta})
$$

$$
f(x_i|\boldsymbol{x}_{1:i-1};\boldsymbol{\theta}) = \sum_{k=1}^{K} q_{i|i-1}(k) f_k(x_i|\boldsymbol{\theta})
$$

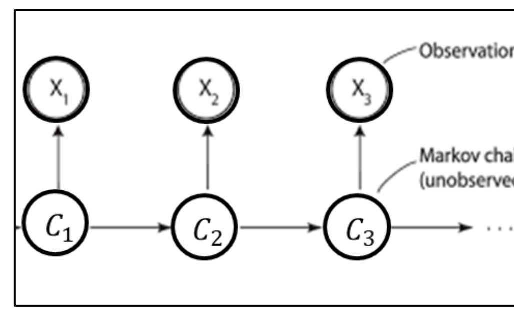
HMM forward equations

- Define $q_{i|j}(k) = \Pr(C_i = k|x_{1:j}), x_{1:j} = (x_1, ..., x_j)$.
- Initialization

$$q_{1|1}(k) = \Pr(C_1 = k|x_1) = \frac{\pi_1(k)f_k(x_1)}{\sum_{j=1}^{K} \pi_1(j)f_j(x_1)}$$

- Prediction:

$$\begin{aligned} q_{i|i-1}(k) &= \Pr(C_i = k|x_{1:i-1}) \\ &= \sum_{j=1}^{K} \Pr(C_i = k|C_{i-1} = j, x_{1_{i-1}}) \Pr(C_{i-1} = j|x_{1:i-1}) \\ &= \sum_{j=1}^{K} p(k|j)q_{i-1|i-1}(j) \end{aligned}$$



- Updating:

$$\begin{aligned} q_{i|i}(k) = \Pr(C_i = k|x_{1:i}) &= \frac{\Pr(C_i = k|x_{1:i-1})p(x_i|C_i = k)}{p(x_i|x_{1:i-1})} \\ &\propto q_{i|i-1}(k)f_k(x_i) \end{aligned}$$

Backward equations

- $q_{n|n}(k) = \Pr(C_n = k|x_{1:n})$ obtained from forward equations
- Going backwards:

$$q_{i|n}(k) = \Pr(C_i = k|x_{1:n})$$

Expand

$$= \sum_{\ell=1}^{K} \Pr(C_i = k, C_{i+1} = \ell|x_{1:n})$$

$$= \sum_{\ell=1}^{K} \Pr(C_i = k|C_{i+1} = \ell, x_{1:n}) \Pr(C_{i+1} = \ell|x_{1:n})$$

Recognize

$$= \sum_{\ell=1}^{K} \Pr(C_i = k|C_{i+1} = \ell, x_{1:i}) q_{i+1|n}(\ell)$$

Bayes formula

$$= \sum_{\ell=1}^{K} \frac{\Pr(C_i = k|x_{1:i}) \Pr(C_{i+1} = \ell|C_i = k, x_{1:i})}{\Pr(C_{i+1} = \ell|x_{1:i})} q_{i+1|n}(\ell)$$

Because of the Markov structure

$$= P(C_{i+1} = l|C_i = k)$$

Recognize

$$= \sum_{\ell=1}^{K} \frac{q_{i|i}(k) p(\ell|k)}{q_{i+1|i}(\ell)} q_{i+1|n}(\ell)$$

Sequence probability

- Needed $\Pr(C_i = k, C_{i-1} = \ell | x_{1:n})$ within EM

$$\Pr(C_i = k, C_{i-1} = \ell | x_{1:n})$$
$$= \Pr(C_i = k | x_{1:n}) \Pr(C_{i-1} = \ell | C_i = k, x_{1:n})$$

Recognize
$$= q_{i|n}(k) \Pr(C_{i-1} = \ell | C_i = k, x_{1:i-1})$$

Bayes formula
$$= q_{i|n}(k) \frac{\Pr(C_{i-1} = \ell | x_{1:i-1}) \Pr(C_i = k | C_{i-1} = \ell, x_{1:i-1})}{\Pr(C_i = k | x_{1:i-1})}$$

Recognize
$$= q_{i|n}(k) \frac{q_{i-1|i-1}(\ell) p(k|\ell)}{q_{i|i-1}(k)}$$

HMM - M-step

- Estimation of probabilities

$$\pi_1^{(t+1)}(k) = \Pr(C_1 = k|x_{1:n}, \theta^{(t)})$$

$$p^{(t+1)}(k|j) = \frac{\sum_{i=2}^{n} \Pr(C_i = k, C_{i-1} = j|x_{1:n}, \theta^{(t)})}{\sum_{i=2}^{n} \Pr(C_{i-1} = j|x_{1:n}, \theta^{(t)})}$$

- Estimation of parameters in Gaussian distribution

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

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