Computational statistics

Chapter 3: EM algorithm

Thierry Denœux
Université de technologie de Compiègne

August 2026

EM Algorithm

- An iterative optimization strategy useful when maximizing the likelihood is difficult, but:
  - There are missing (non-observed) data
  - If the missing data were observed, maximizing the likelihood would be easy.
- Many applications in statistics and econometrics.
- Can be very simple to implement. Can reliably find an optimum through stable, uphill steps.

Overview

Overview

Notation

$\mathbf{Y}$: Observed data

$\mathbf{Z}$: Missing data or latent variables

$\mathbf{X}$: Complete data $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$

$\boldsymbol{\theta}$: Unknown parameter

$L(\boldsymbol{\theta})$: observed-data likelihood, short for $L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta})$

$L_c(\boldsymbol{\theta})$: complete-data likelihood, short for $L(\boldsymbol{\theta}; \boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{\theta})$

$\ell(\boldsymbol{\theta}), \ell_c(\boldsymbol{\theta})$: observed and complete-data log-likelihoods

Q function

- Suppose we seek to maximize $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.
- Define $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ to be the expectation of the complete-data log-likelihood, conditional on the observed data $\mathbf{Y} = \mathbf{y}$. Namely

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) =& \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left\{ \ell_c(\boldsymbol{\theta}) \mid \mathbf{y} \right\} \\ =& \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left\{ \log f(\mathbf{X}; \boldsymbol{\theta}) \mid \mathbf{y} \right\} \\ =& \int \left[\log f(\mathbf{x}; \boldsymbol{\theta}) \right] f(\mathbf{z} \mid \mathbf{y}; \boldsymbol{\theta}^{(t)}) \, d\mathbf{z}. \end{aligned}$$

($f(\boldsymbol{x} \mid \boldsymbol{y}; \theta^{(t)}) = f(\boldsymbol{z} \mid \boldsymbol{y}; \theta^{(t)})$ because $\mathbf{Z}$ is the only random part of $\mathbf{X}$ once we are given $\mathbf{Y} = \mathbf{y}$.)

The EM Algorithm

Start with $\boldsymbol{\theta}^{(0)}$. Then

1. **E step**: Compute $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
2. **M step**: Maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. Set $\boldsymbol{\theta}^{(t+1)}$ equal to the maximizer of Q.
3. Increment t and return to the E step unless a stopping criterion has been met; e.g.,

$$\ell(\boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{\theta}^{(t)}) \leq \epsilon$$

or

$$\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\| \leq \epsilon.$$

Convergence of the EM Algorithm

- It can be proved that $L(\boldsymbol{\theta})$ increases after each EM iteration, i.e., $L(\boldsymbol{\theta}^{(t+1)}) \geq L(\boldsymbol{\theta}^{(t)})$ for $t = 0, 1, \ldots$.
- Consequently, the algorithm converges to a local maximum of $L(\boldsymbol{\theta})$ if the likelihood function is bounded above.
- Typically, we run the algorithm several times with random initial conditions, and we keep the results of the best run.

Example: mixture of normal and uniform distributions

- Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be an i.i.d. sample from a mixture of a normal distribution $\mathcal{N}(\mu, \sigma)$ and a uniform distribution $\mathcal{U}([-a, a])$, with pdf

$$f(y; \boldsymbol{\theta}) = \pi\phi(y; \mu, \sigma) + (1 - \pi)c \tag{1}$$

  where $\phi(\cdot; \mu, \sigma)$ is the normal pdf, $c = (2a)^{-1}$ is a known constant, π is the proportion of the normal distribution in the mixture and $\boldsymbol{\theta} = (\mu, \sigma, \pi)^T$ is the vector of parameters.
- Typically, the uniform distribution corresponds to outliers in the data. The proportion of outliers in the population is then $1 - \pi$.
- We want to estimate $\boldsymbol{\theta}$.

Observed and complete-data likelihoods

- Let $Z_i = 1$ if observation i is not an outlier, $Z_i = 0$ otherwise. We have $Z_i \sim \mathcal{B}(\pi)$.
- The vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ is the missing data.
- Observed-data likelihood:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} [\pi \phi(y_i; \mu, \sigma) + (1 - \pi) c]$$

- Complete-data likelihood:

$$\begin{aligned} L_c(\boldsymbol{\theta}) &= \prod_{i=1}^{n} f(y_i, z_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i \mid z_i; \mu, \sigma) f(z_i \,; \pi) \\ &= \prod_{i=1}^{n} \left[\phi(y_i; \mu, \sigma)^{z_i} c^{1 - z_i} \pi^{z_i} (1 - \pi)^{1 - z_i} \right] \end{aligned}$$

Derivation of function Q

- Complete-data log-likelihood:

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^{n} z_i \log \phi(y_i; \mu, \sigma) + \left(n - \sum_{i=1}^{n} z_i \right) \log c + \sum_{i=1}^{n} \left(z_i \log \pi + (1 - z_i) \log(1 - \pi) \right)$$

- It is linear in the z_i. Consequently, the Q function is simply

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} z_i^{(t)} \log \phi(y_i; \mu, \sigma) + \left(n - \sum_{i=1}^{n} z_i^{(t)} \right) \log c + \sum_{i=1}^{n} \left(z_i^{(t)} \log \pi + (1 - z_i^{(t)}) \log(1 - \pi) \right)$$

with $z_i^{(t)} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[Z_i | y_i]$.

EM algorithm

E-step: compute

$$z_i^{(t)} = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[Z_i \mid y_i] = \mathbb{P}_{\boldsymbol{\theta}^{(t)}}[Z_i = 1 \mid y_i]$$
$$= \frac{\phi(y_i; \mu^{(t)}, \sigma^{(t)})\pi^{(t)}}{\phi(y_i; \mu^{(t)}, \sigma^{(t)})\pi^{(t)} + c(1 - \pi^{(t)})}$$

M-step: Maximize $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$. We get

$$\pi^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} z_i^{(t)}, \quad \mu^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t)} y_i}{\sum_{i=1}^{n} z_i^{(t)}}, \quad \sigma^{(t+1)} = \sqrt{\frac{\sum_{i=1}^{n} z_i^{(t)} (y_i - \mu^{(t+1)})^2}{\sum_{i=1}^{n} z_i^{(t)}}}$$

Bayesian posterior mode

- Consider a Bayesian estimation problem with likelihood $L(\boldsymbol{\theta})$ and prior $f(\boldsymbol{\theta})$.
- The posterior density if proportional to $L(\boldsymbol{\theta})f(\boldsymbol{\theta})$. It can also be maximized by the EM algorithm.
- The E-step requires

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left\{ \ell_c(\boldsymbol{\theta}) \mid \mathbf{y} \right\} + \log f(\boldsymbol{\theta})$$

- The addition of the log-prior often makes it more difficult to maximize Q during the M-step.
- Some methods can be used to facilitate the M-step in difficult situations (see below).

Overview

Why does it work?

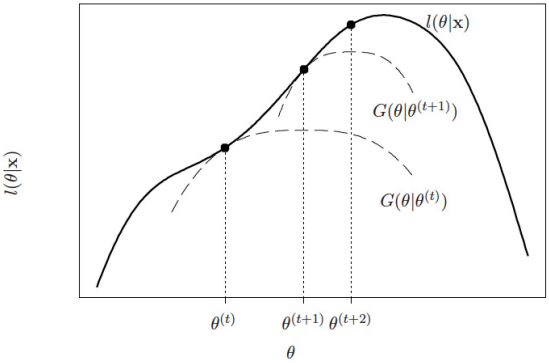
- Ascent: Each M-step increases the log-likelihood.
- Optimization transfer:

$$\ell(\boldsymbol{\theta}) \geq \underbrace{Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})}_{G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})}$$

- The last two terms in $G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ do not depend on $\boldsymbol{\theta}$, so Q and G are maximized at the same $\boldsymbol{\theta}$.
- Further, G is tangent to ℓ at $\boldsymbol{\theta}^{(t)}$, and lies everywhere below ℓ. We say that G is a minorizing function for ℓ (see next slide).
- EM transfers optimization from ℓ to the surrogate function G, which is more convenient to maximize.

The nature of EM

- One-dimensional illustration of EM algorithm as a minorization or optimization transfer strategy.
- Each E step forms a minorizing function G, and each M step maximizes it to provide an uphill step.



Proof

- We have

$$f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta}) = \frac{f(\boldsymbol{y}, \boldsymbol{z}; \boldsymbol{\theta})}{f(\boldsymbol{y}; \boldsymbol{\theta})} = \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{y}; \boldsymbol{\theta})} \Rightarrow f(\boldsymbol{y}; \boldsymbol{\theta}) = \frac{f(\boldsymbol{x}; \boldsymbol{\theta})}{f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta})}$$

- Consequently,

$$\ell(\boldsymbol{\theta}) = \log f(\boldsymbol{y}; \boldsymbol{\theta}) = \underbrace{\log f(\boldsymbol{x}; \boldsymbol{\theta})}_{\ell_c(\boldsymbol{\theta})} - \log f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta})$$

- Taking expectations on both sides wrt the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and using $\boldsymbol{\theta}^{(t)}$ for $\boldsymbol{\theta}$:

$$\ell(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\log f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta}) \mid \boldsymbol{y}]}_{H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})} \tag{2}$$

Proof: $\theta^{(t)}$ is a maximizer of $H(\theta, \theta^{(t)})$

- Now, for all $\boldsymbol{\theta} \in \Theta$,

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\log \frac{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta})}{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(t)})} \mid \boldsymbol{y} \right] \tag{3a}$$

$$\leq \log \underbrace{\mathbb{E}_{\boldsymbol{\theta}^{(t)}} \left[\frac{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta})}{f(\boldsymbol{Z} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(t)})} \mid \boldsymbol{y} \right]}_{\int \frac{f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta})}{f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(t)})} f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta}^{(t)}) d\boldsymbol{z}} (*) \tag{3b}$$

$$\leq \log \underbrace{\int f(\boldsymbol{z} \mid \boldsymbol{y}; \boldsymbol{\theta}) d\boldsymbol{z}}_{1} = 0 \tag{3c}$$

(*): from the concavity of the log and Jensen's inequality.

- Hence, $\boldsymbol{\theta}^{(t)}$ is a maximizer of $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$

Proof: $\ell(\cdot)$ dominates $G(\cdot, \theta^{(t)})$

Hence, for all $\boldsymbol{\theta} \in \Theta$,

$$H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) \geq H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$$

$$Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta})$$

$$\ell(\boldsymbol{\theta}) \geq \underbrace{Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) + \ell(\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})}_{G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})}.$$

Proof: G is tangent to ℓ at $\boldsymbol{\theta}^{(t)}$

- As $\boldsymbol{\theta}^{(t)}$ maximizes $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) - \ell(\boldsymbol{\theta})$, we have

$$H'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \ell'(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = 0,$$

  so

$$Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \ell'(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}.$$

- Consequently, as $G(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) + \text{cst}$,

$$G'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = Q'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} = \ell'(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}.$$

Proof: monotonicity

- From (2),

$$\ell(\boldsymbol{\theta}^{(t+1)}) - \ell(\boldsymbol{\theta}^{(t)}) = \underbrace{Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})}_{A} - \left[\underbrace{H(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})}_{B} \right]$$

- $A \geq 0$ because $\boldsymbol{\theta}^{(t+1)}$ is a maximizer of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, and $B \leq 0$ because, from (3), $\boldsymbol{\theta}^{(t)}$ is a maximizer of $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- Hence,

$$\boxed{\ell(\boldsymbol{\theta}^{(t+1)}) \geq \ell(\boldsymbol{\theta}^{(t)})}$$

Overview

Overview

Monte Carlo EM (MCEM)

- Sometimes, the conditional expectation of $\ell_c(\boldsymbol{\theta})$ given $\boldsymbol{y}$ cannot be easily computed analytically in the E step.
- Approach: randomly generate sets of missing values according to the conditional distribution $f(\mathbf{z}|\mathbf{y};\boldsymbol{\theta}^{(t)})$, and replace the expectation by an average over generated data sets.

Monte Carlo EM (MCEM)

- Replace the t-th E step with
  1. Draw missing datasets $\mathbf{Z}_1^{(t)}, \ldots, \mathbf{Z}_{m^{(t)}}^{(t)}$ i.i.d. from $f(\mathbf{z}|\mathbf{y}; \boldsymbol{\theta}^{(t)})$. Each $\mathbf{Z}_j^{(t)}$ is a vector of all the missing values needed to complete the observed dataset, so $\mathbf{X}_j^{(t)} = (\mathbf{y}, \mathbf{Z}_j^{(t)})$ denotes a completed dataset where the missing values have been replaced by $\mathbf{Z}_j^{(t)}$.
  2. Calculate

$$\widehat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f(\mathbf{X}_j^{(t)}; \boldsymbol{\theta}).$$

- Then $\widehat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ is a Monte Carlo estimate of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- The M step is modified to maximize $\widehat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.

Remarks

- It is advised to increase $m^{(t)}$ as iterations progress to reduce the Monte Carlo variability of $\widehat{Q}$.
- MCEM will not converge in the same sense as ordinary EM, rather values of $\boldsymbol{\theta}^{(t)}$ will bounce around the true maximum, with a precision that depends on $m^{(t)}$.

Overview

Generalized EM (GEM) algorithm

- In the original EM algorithm, $\boldsymbol{\theta}^{(t+1)}$ is a maximizer of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$, i.e.,

$$Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$$

  for all $\boldsymbol{\theta}$.
- However, to ensure convergence, we only need that

$$Q(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})$$

- Any algorithm that chooses $\boldsymbol{\theta}^{(t+1)}$ at each iteration to guarantee the above condition (without maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$) is called a Generalized EM (GEM) algorithm.

EM gradient algorithm

- Replace the M step with a single step of Newton's method, thereby approximating the maximum without actually solving for it exactly.
- Instead of maximizing, choose:

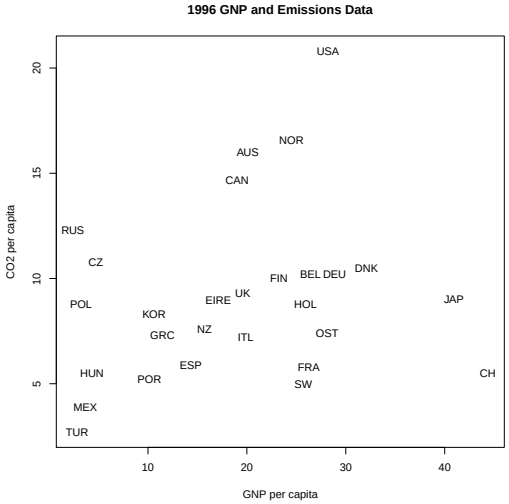
$$\begin{aligned}\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \mathbf{Q}''(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mathbf{Q}'(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\ &= \boldsymbol{\theta}^{(t)} - \mathbf{Q}''(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \ell'(\boldsymbol{\theta}^{(t)})\end{aligned}$$

- Ascent is ensured for canonical parameters in exponential families. Backtracking can ensure ascent in other cases; inflating steps can speed up convergence.

Overview

Overview

Introductory example



Introductory example (continued)

- The data in the previous slide do not show any clear linear trend.
- However, there seem to be several groups for which a linear model would be a reasonable approximation.
- How to identify those groups and the corresponding linear models?

Model

- Model: the response variable Y depends on the input variable X in different ways, depending on a latent variable Z. (Beware: we have switched back to the classical notation for regression models!)
- This model is called mixture of regressions or switching regressions. It has been widely studied in the econometrics literature.
- Model:

$$Y = \begin{cases} \beta_1^T X + \epsilon_1, \ \epsilon_1 \sim \mathcal{N}(0, \sigma_1) & \text{if } Z = 1, \\ \vdots \\ \beta_K^T X + \epsilon_K, \ \epsilon_K \sim \mathcal{N}(0, \sigma_K) & \text{if } Z = K, \end{cases}$$

with $X = (1, X_1, \ldots, X_p)$ and $Z \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$, so

$$f(y|X = x) = \sum_{k=1}^{K} \pi_k \phi(y; \beta^T x, \sigma_k)$$

Observed and complete-data likelihoods

- Observed-data likelihood:

$$L(\theta) = \prod_{i=1}^{n} f(y_i|x_i;\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi(y_i; \beta_k^T x_i, \sigma_k)$$

- Complete-data likelihood:

$$\begin{aligned} L_c(\theta) &= \prod_{i=1}^{n} f(y_i, z_i|x_i;\theta) = \prod_{i=1}^{n} f(y_i|x_i, z_i;\theta) p(z_i \ |\pi) \\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \phi(y_i; \beta_k^T x_i, \sigma_k)^{z_{ik}} \pi_k^{z_{ik}}, \end{aligned}$$

with $z_{ik} = 1$ if $z_i = k$ and $z_{ik} = 0$ otherwise.

Derivation of function Q

- Complete-data log-likelihood:

$$\ell_c(\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \phi(y_i; \beta_k^T x_i, \sigma_k) + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \log \pi_k$$

- It is linear in the z_{ik}. Consequently, the Q function is simply

$$Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \log \phi(y_i; \beta_k^T x_i, \sigma_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \log \pi_k$$

with $z_{ik}^{(t)} = \mathbb{E}_{\theta^{(t)}}[Z_{ik}|y_i] = \mathbb{P}_{\theta^{(t)}}[Z_i = k|y_i]$.

EM algorithm

- E-step: compute

$$z_{ik}^{(t)} = \mathbb{P}_{\theta^{(t)}}[Z_i = k | y_i]$$
$$= \frac{\phi(y_i; \beta_k^{(t)T} x_i, \sigma_k^{(t)}) \pi_k^{(t)}}{\sum_{\ell=1}^{K} \phi(y_i; \beta_\ell^{(t)T} x_i, \sigma_\ell^{(t)}) \pi_\ell^{(t)}}$$

- M-step: Maximize $Q(\theta, \theta^{(t)})$. As before, we get

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{n},$$

with $N_k^{(t)} = \sum_{i=1}^{N} z_{ik}^{(t)}$.

M-step: update of the β_k and σ_k

- In $Q(\theta, \theta^{(t)})$, the term depending on β_k is

$$SS_k = \sum_{i=1}^{n} z_{ik}^{(t)} (y_i - \beta_k^T x_i)^2.$$

- Minimizing SS_k w.r.t. β_k is a weighted least-squares (WLS) problem. In matrix form,

$$SS_k = (\boldsymbol{y} - \boldsymbol{X}\beta_k)^T \boldsymbol{W}_k (\boldsymbol{y} - \boldsymbol{X}\beta_k)$$

with $\boldsymbol{W}_k = \text{diag}(z_{1k}^{(t)}, \ldots, z_{nk}^{(t)})$.

M-step: update of the β_k and σ_k (continued)

- The solution is the WLS estimate of β_k:

$$\beta_k^{(t+1)} = (\boldsymbol{X}^T \boldsymbol{W}_k \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W}_k \boldsymbol{y}$$

- The value of σ_k maximizing $Q(\theta, \theta^{(t)})$ is the weighted average of the residuals,

$$\begin{aligned}\sigma_k^{2(t+1)} &= \frac{1}{N_k^{(t)}} \sum_{i=1}^{n} z_{ik}^{(t)} (y_i - \beta_k^{(t+1)T} x_i)^2 \\ &= \frac{1}{N_k^{(t)}} (\boldsymbol{y} - \boldsymbol{X}\beta_k^{(t+1)})^T \boldsymbol{W}_k (\boldsymbol{y} - \boldsymbol{X}\beta_k^{(t+1)})\end{aligned}$$

Mixture of regressions using `flexmix`

```
library(mixtools); library(flexmix)
data("CO2data")
fit1 <- flexmix(CO2 ~ GNP,data = CO2data,k = 1)
loglik_best <- -Inf
for(i in 1:10){
  fit2 <- flexmix(CO2 ~ GNP,data = CO2data,k = 2)
  if(logLik(fit2)>loglik_best){
    loglik_best<-logLik(fit2)
    fit2best<-fit2
  }
}
fit2<-fit2best
AIC(fit1, fit2)
> AIC(fit1, fit2)
     df      AIC
fit1  3 161.9675
fit2  7 147.9675
```

Best solution in 10 runs

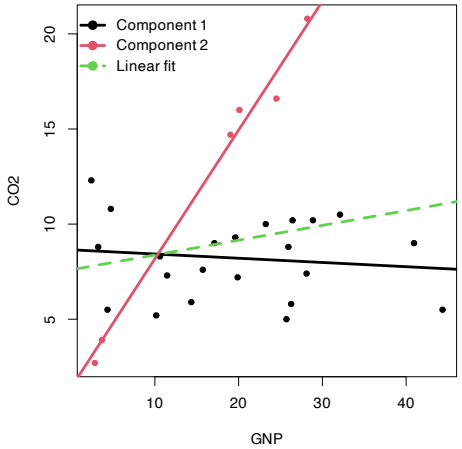
```
par1<-parameters(fit1)
par <- parameters(fit2)
cl <- clusters(fit2)

plot(CO2data$GNP, CO2data$CO2,col = cl,pch = 19,
     cex = 0.8,xlab = "CO2",ylab = "GNP")

abline(a = par["coef.(Intercept)", 1],
          b = par["coef.GNP", 1],col = 1,lwd = 3)

abline(a = par["coef.(Intercept)", 2],
          b = par["coef.GNP", 2],col = 2, lwd = 3)

abline(a = par1["coef.(Intercept)", 1],
           b = par1["coef.GNP", 1],
           col = 3,lwd = 3, lty=2)
```



Overview

Making the mixing proportions predictor-dependent

- An interesting extension of the previous model is to assume the proportions π_k to be partially explained by a vector of concomitant variables W.
- If $W = X$, we can approximate the regression function by different linear functions in different regions of the predictor space.
- In ML, this method is referred to as the mixture of experts methods.
- A useful parametric form for π_k that ensures $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$ is the multinomial logit model

$$\pi_k(w, \alpha) = \frac{\exp(\alpha_k^T w)}{\sum_{\ell=1}^{K} \exp(\alpha_\ell^T w)}$$

with $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\alpha_1 = 0$.

EM algorithm

- The Q function is the same as before, except that the π_k now depend on the w_i and parameter α:

$$Q(\theta, \theta^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \log \phi(y_i; \beta_k^T x_i, \sigma_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \log \pi_k(w_i, \alpha)$$

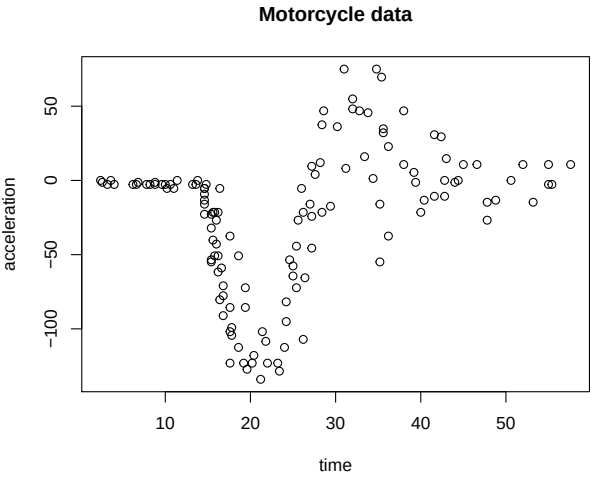
- In the M-step, the update formula for β_k and σ_k are unchanged.
- No closed-form expression for

$$\widehat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik}^{(t)} \log \pi_k(w_i, \alpha)$$

- GEM algorithm: perform a single step of Newton's algorithm. Backtracking can be used to ensure ascent.

Example: motorcycle data

```
library("MASS")
x<-mcycle$times
y<-mcycle$accel
plot(x,y)
```



Mixture of experts using `flexmix`

```
library(flexmix)

K<-5
res <- flexmix(y ~ x,k=K,model=FLXMRglm(family="gaussian"),
                        concomitant=FLXPmultinom(formula=~x))

beta <- parameters(res)[1:2,]
alpha <- res@concomitant@coef
```

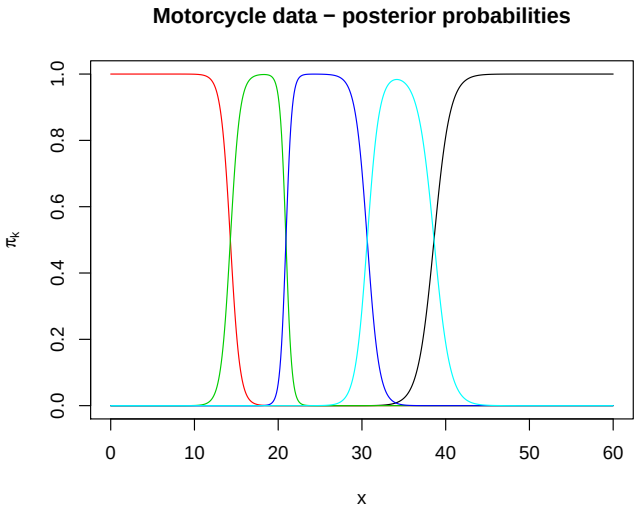
Plotting the posterior probabilities

```
xt <- seq(0,60,0.1)
Nt <- length(xt)
plot(x,y)

pit <- matrix(0,Nt,K)
for(k in 1:K) pit[,k] <- exp(alpha[1,k]+alpha[2,k]*xt)
pit <- pit/rowSums(pit)

plot(xt,pit[,1],type="l",col=1)
for(k in 2:K) lines(xt,pit[,k],col=k)
```

Posterior probabilities



Plotting the predictions

```
yhat <- rep(0,Nt)
for(k in 1:K) yhat <- yhat+pit[,k]*(beta[1,k]+beta[2,k]*xt)

plot(x,y,main="Motorcycle data",xlab="time",ylab="acceleration")

for(k in 1:K) abline(beta[1:2,k],lty=2)
lines(xt,yhat,col='red',lwd=2)
```

Regression lines and predictions