Computational statistics

EM algorithm

Thierry Denœux

February-March 2017

EM Algorithm

- An iterative optimization strategy motivated by a notion of missingness and by consideration of the conditional distribution of what is missing given what is observed.
- Can be very simple to implement. Can reliably find an optimum through stable, uphill steps.
- Difficult likelihoods often arise when data are missing. EM simplifies such problems. In fact, the 'missing data' may not truly be missing: they may be only a conceptual ploy to exploit the EM simplification!

Overview

Overview

Notation

$\mathbf{Y}$: Observed variables.

$\mathbf{Z}$: Missing 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.

Notation

- 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

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

where the last equation emphasizes that $\mathbf{Z}$ 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. Return to the E step unless a stopping criterion has been met; e.g.,

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

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.

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

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

  where $\phi(\cdot; \mu, \sigma)$ is the normal pdf, $c = (2a)^{-1}$, 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:

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

Derivation of function Q

- Complete-data log-likelihood:

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

- 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

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

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

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

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

Remark

- As mentioned before, the EM algorithm finds only a local maximum of $\ell(\theta)$.
- It is easy to find a global maximum: if μ is equal to some y_i and $\sigma = 0$, then $\phi(y_i; \mu, \sigma) = \infty$ and, consequently, $\ell(\theta) = +\infty$.
- We are not interested in these global maxima, because they correspond to degenerate solutions!

Bayesian posterior mode

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

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

- 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

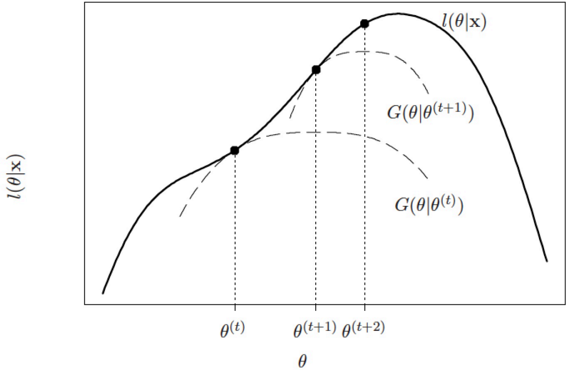
Why does it work?

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

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

- The last two terms in $G(\theta, \theta^{(t)})$ are constant with respect to θ, so Q and G are maximized at the same θ.
- Further, G is tangent to ℓ at $\theta^{(t)}$, and lies everywhere below ℓ. We say that G is a minorizing function for ℓ.
- 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

$$f(z|y;\theta) = \frac{f(x;\theta)}{f(y;\theta)} \Rightarrow f(y;\theta) = \frac{f(x;\theta)}{f(z|y;\theta)}$$

- Consequently,

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

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

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

Proof - the minorizing function

- Now, for all $\theta \in \Theta$,

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

$$\leq \log \mathbb{E}_{\theta^{(t)}}\left[\frac{f(Z|y;\theta)}{f(Z|y;\theta^{(t)})}|y\right] \; (*) \tag{3b}$$

$$= \log \int f(z|y;\theta)dz = 0 \tag{3c}$$

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

- Hence, for all $\theta \in \Theta$,

$$H(\theta, \theta^{(t)}) \leq H(\theta^{(t)}, \theta^{(t)}) = Q(\theta^{(t)}, \theta^{(t)}) - \ell(\theta^{(t)}), \text{ or}$$

$$\ell(\theta) \geq Q(\theta, \theta^{(t)}) + \ell(\theta^{(t)}) - Q(\theta^{(t)}, \theta^{(t)}) = G(\theta, \theta^{(t)}) \tag{4}$$

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

- From (4), $\ell(\theta^{(t)}) = G(\theta^{(t)}, \theta^{(t)})$.
- Now, we can rewrite (4) as

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

Consequently, $\theta^{(t)}$ maximizes $Q(\theta, \theta^{(t)}) - \ell(\theta)$, hence

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

and

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

Proof - monotonicity

- From (2),

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

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

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

Overview

Overview

Monte Carlo EM (MCEM)

- Replace the tth 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 $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f(\mathbf{X}_j^{(t)}; \boldsymbol{\theta})$.
- Then $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ is a Monte Carlo estimate of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- The M step is modified to maximize $\hat{Q}^{(t+1)}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$.
- Increase $m^{(t)}$ as iterations progress to reduce the Monte Carlo variability of $\hat{Q}$. 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:

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

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

ECM algorithm

- Replaces the M step with a series of computationally simpler conditional maximization (CM) steps.
- Call the collection of simpler CM steps after the tth E step a CM cycle. Thus, the tth iteration of ECM is comprised of the tth E step and the tth CM cycle.
- Let S denote the total number of CM steps in each CM cycle.

ECM algorithm (continued)

- For $s = 1, \ldots, S$, the sth CM step in the tth cycle requires the maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ subject to (or conditional on) a constraint, say
$$\mathbf{g}_s(\boldsymbol{\theta}) = \mathbf{g}_s(\boldsymbol{\theta}^{(t+(s-1)/S)})$$
where $\boldsymbol{\theta}^{(t+(s-1)/S)}$ is the maximizer found in the $(s-1)$th CM step of the current cycle.
- When the entire cycle of S steps of CM has been completed, we set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t+S/S)}$ and proceed to the E step for the $(t+1)$th iteration.
- ECM is a GEM algorithm, since each CM step increases Q.
- The art of constructing an effective ECM algorithm lies in choosing the constraints cleverly.

Choice 1: Iterated Conditional Modes / Gauss-Seidel

- Partition $\boldsymbol{\theta}$ into S subvectors, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_S)$.
- In the sth CM step, maximize Q with respect to $\boldsymbol{\theta}_s$ while holding all other components of $\boldsymbol{\theta}$ fixed.
- This amounts to the constraint induced by the function

$$g_s(\boldsymbol{\theta}) = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{s-1}, \boldsymbol{\theta}_{s+1}, \ldots, \boldsymbol{\theta}_S).$$

Choice 2

- At the sth CM step, maximize Q with respect to all other components of $\boldsymbol{\theta}$ while holding $\boldsymbol{\theta}_s$ fixed.
- Then $g_s(\boldsymbol{\theta}) = \boldsymbol{\theta}_s$.
- Additional systems of constraints can be imagined, depending on the particular problem context.
- A variant of ECM inserts an E step between each pair of CM steps, thereby updating Q at every stage of the CM cycle.

Overview

Variance of the MLE

- Let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$.
- As $n \to \infty$, the limiting distribution of $\widehat{\boldsymbol{\theta}}$ is $\mathcal{N}(\boldsymbol{\theta}^*, I(\boldsymbol{\theta}^*)^{-1})$, where $\boldsymbol{\theta}^*$ is the true value of $\boldsymbol{\theta}$, and

$$I(\boldsymbol{\theta}) = \mathbb{E}[\ell'(\boldsymbol{\theta})\ell'(\boldsymbol{\theta})^T] = -\mathbb{E}[\ell''(\boldsymbol{\theta})]$$

  is the expected Fisher information matrix (the second equality holds under some regularity conditions).
- $I(\boldsymbol{\theta}^*)$ can be estimated by $I(\widehat{\boldsymbol{\theta}})$, or by $-\ell''(\widehat{\boldsymbol{\theta}}) = I_{obs}(\widehat{\boldsymbol{\theta}})$ (observed information matrix).
- Standard error estimates can be obtained by computing the square roots of the diagonal elements of $I_{obs}(\widehat{\boldsymbol{\theta}})^{-1}$.

Obtaining variance estimates

- The EM algorithms allows us to estimate $\widehat{\boldsymbol{\theta}}$, but it does not directly provide an estimate of $I(\boldsymbol{\theta}^*)$.
- Direct computation of $I(\widehat{\boldsymbol{\theta}})$ or $I_{obs}(\widehat{\boldsymbol{\theta}})$ is often difficult.
- Main methods:
  1. Louis' method
  2. Supplement EM (SEM) algorithm
  3. Bootstrap

Overview

Missing information principle

- We have seen that
$$f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}) = \frac{f(\boldsymbol{x};\boldsymbol{\theta})}{f(\boldsymbol{y};\boldsymbol{\theta})},$$
from which we get
$$\ell(\boldsymbol{\theta}) = \ell_c(\boldsymbol{\theta}) - \log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta}).$$
- Differentiating twice and negating both sides, then taking expectations over the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{y}$,
$$\underbrace{-\ell''(\boldsymbol{\theta})}_{\hat{\imath}_{\mathbf{Y}}(\boldsymbol{\theta})} = \underbrace{\mathbb{E}\left[-\ell_c''(\boldsymbol{\theta})|\boldsymbol{y}\right]}_{\hat{\imath}_{\mathbf{X}}(\boldsymbol{\theta})} - \underbrace{\mathbb{E}\left[-\frac{\partial^2 \log f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|\boldsymbol{y}\right]}_{\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})}$$
where
  - $\hat{\imath}_{\mathbf{Y}}(\boldsymbol{\theta})$ is the observed information,
  - $\hat{\imath}_{\mathbf{X}}(\boldsymbol{\theta})$ is the complete information, and
  - $\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is the missing information.

Louis' method

- Computing $\hat{\imath}_{\mathbf{X}}(\boldsymbol{\theta})$ and $\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is sometimes easier than computing $-\ell''(\boldsymbol{\theta})$ directly
- We can show that

$$\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \mathrm{Var}[S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})],$$

  where the variance is taken w.r.t. $\boldsymbol{Z}|\boldsymbol{y}$, and

$$S_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{z}|\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

  is the conditional score.
- As the expected score is zero at $\widehat{\boldsymbol{\theta}}$, we have

$$\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}}) = \int S_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}}) S_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}})^T f(\boldsymbol{z}|\boldsymbol{y};\widehat{\boldsymbol{\theta}}) d\boldsymbol{z}$$

Monte Carlo approximation

- When they cannot be computed analytically, $\hat{\imath}_{\mathbf{X}}(\boldsymbol{\theta})$ and $\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ can sometimes be approximated by Monte Carlo simulation.
- Method: generate simulated datasets $\boldsymbol{x}_j = (\boldsymbol{y}, \boldsymbol{z}_j)$, $j = 1, \ldots, N$, where $\boldsymbol{y}$ is the observed dataset, and the $\boldsymbol{z}_j$ are imputed missing datasets drawn from $f(\boldsymbol{z}|\boldsymbol{y}; \boldsymbol{\theta})$
- Then,

$$\hat{\imath}_{\mathbf{X}}(\boldsymbol{\theta}) \approx \frac{1}{N} \sum_{j=1}^{N} -\frac{\partial^2 \log f(\boldsymbol{x}_j; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$$

and $\hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\boldsymbol{\theta})$ is approximated by the sample variance of the values

$$\frac{\partial f(\boldsymbol{z}_j|\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Overview

EM mapping

- Let $\boldsymbol{\Psi}$ denotes the EM mapping, defined by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\Psi}(\boldsymbol{\theta}^{(t)})$$

- From the convergence of EM, $\widehat{\boldsymbol{\theta}}$ is a fixed point:

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}).$$

- The Jacobian matrix of $\boldsymbol{\Psi}$ is the $p \times p$ matrix

$$\boldsymbol{\Psi}'(\boldsymbol{\theta}) = \left(\frac{\partial \Psi_i(\boldsymbol{\theta})}{\partial \theta_j} \right).$$

- It can be shown that

$$\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})^T = \hat{\boldsymbol{\imath}}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\boldsymbol{\theta}}) \hat{\boldsymbol{\imath}}_{\mathbf{X}}(\widehat{\boldsymbol{\theta}})^{-1}$$

Using $\mathbf{\Psi}'(\theta)$ for variance estimation

- From the missing information principle,

$$\begin{aligned}\hat{\imath}_{\mathbf{Y}}(\widehat{\theta}) &= \hat{\imath}_{\mathbf{X}}(\widehat{\theta}) - \hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\theta}) \\ &= \left[\mathbf{I} - \hat{\imath}_{\mathbf{Z}|\mathbf{Y}}(\widehat{\theta})\hat{\imath}_{\mathbf{X}}(\widehat{\theta})^{-1}\right]\hat{\imath}_{\mathbf{X}}(\widehat{\theta}) \\ &= \left[\mathbf{I} - \mathbf{\Psi}'(\widehat{\theta})^T\right]\hat{\imath}_{\mathbf{X}}(\widehat{\theta}).\end{aligned}$$

  Hence,

$$\hat{\imath}_{\mathbf{Y}}(\widehat{\theta})^{-1} = \hat{\imath}_{\mathbf{X}}(\widehat{\theta})^{-1}\left[\mathbf{I} - \mathbf{\Psi}'(\widehat{\theta})^T\right]^{-1}$$

- From the equality

$$(\mathbf{I}-\mathbf{P})^{-1} = (\mathbf{I}-\mathbf{P}+\mathbf{P})(\mathbf{I}-\mathbf{P})^{-1} = \mathbf{I} + \mathbf{P}(\mathbf{I}-\mathbf{P})^{-1},$$

  we get

$$\hat{\imath}_{\mathbf{Y}}(\widehat{\theta})^{-1} = \hat{\imath}_{\mathbf{X}}(\widehat{\theta})^{-1}\left\{\mathbf{I} + \mathbf{\Psi}'(\widehat{\theta})^T\left[\mathbf{I} - \mathbf{\Psi}'(\widehat{\theta})^T\right]^{-1}\right\}. \tag{5}$$

Estimation of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$

- Ler r_{ij} be the element (i,j) of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$. By definition,

$$
\begin{aligned}
r_{ij} &= \frac{\partial \Psi_i(\widehat{\boldsymbol{\theta}})}{\partial \theta_j} \\
&= \lim_{\theta_j \to \widehat{\theta}_j} \frac{\Psi_i(\widehat{\theta}_1, \ldots, \widehat{\theta}_{j-1}, \theta_j, \widehat{\theta}_{j+1}, \ldots, \widehat{\theta}_p) - \Psi_i(\widehat{\boldsymbol{\theta}})}{\theta_j - \widehat{\theta}_j} \\
&= \lim_{t \to \infty} \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \Psi_i(\widehat{\boldsymbol{\theta}})}{\theta_j^{(t)} - \widehat{\theta}_j} = \lim_{t \to \infty} r_{ij}^{(t)}
\end{aligned}
$$

  where $\boldsymbol{\theta}^{(t)}(j) = (\widehat{\theta}_1, \ldots, \widehat{\theta}_{j-1}, \theta_j^{(t)}, \widehat{\theta}_{j+1}, \ldots, \widehat{\theta}_p)$, and $(\theta_j^{(t)})$, $t = 1, 2, \ldots$ is a sequence of values converging to $\widehat{\theta}_j$.

- Method: compute the $r_{ij}^{(t)}$, $t = 1, 2, \ldots$ until they stabilize to some values. Then compute $\hat{\boldsymbol{\imath}}_{\boldsymbol{Y}}(\widehat{\boldsymbol{\theta}})^{-1}$ using (5).

SEM algorithm

1. Run the EM algorithm to convergence, finding $\widehat{\boldsymbol{\theta}}$.
2. Restart the algorithm from some $\boldsymbol{\theta}^{(0)}$ near $\widehat{\boldsymbol{\theta}}$. For $t = 0, 1, 2, \ldots$
   1. Take a standard E step and M step to produce $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$.
   2. For $j = 1, \ldots, p$, define $\boldsymbol{\theta}^{(t)}(j) = (\hat{\theta}_1, \ldots, \hat{\theta}_{j-1}, \theta_j^{(t)}, \hat{\theta}_{j+1}, \ldots, \hat{\theta}_p)$ and

      $$r_{ij}^{(t)} = \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j}$$

      for $i = 1, \ldots, p$. (Recall that $\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\theta}}$.)
   3. Stop when all $r_{ij}^{(t)}$ have converged
3. The (i, j)th element of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$ equals $\lim_{t\to\infty} r_{ij}^{(t)}$. Use the final estimate of $\boldsymbol{\Psi}'(\widehat{\boldsymbol{\theta}})$ to get the variance.
4. SEM is numerically stable and requires little extra work.

Overview

Principle

- Consider the case of iid data $\boldsymbol{y} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$
- If we knew the distribution of the $\boldsymbol{W}_i$, we could
  - generate many samples $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$,
  - compute the ML estimate $\widehat{\boldsymbol{\theta}}_j$ of $\boldsymbol{\theta}$ from each sample $\boldsymbol{y}_j$, and
  - estimate the variance of $\widehat{\boldsymbol{\theta}}$ by the sample variance of the estimates $\widehat{\boldsymbol{\theta}}_1, \ldots, \widehat{\boldsymbol{\theta}}_N$.
- Bootstrap principle: use the empirical distribution in place of the true distribution of the $\boldsymbol{W}_i$

Algorithm

1. Calculate $\widehat{\boldsymbol{\theta}}_{EM}$ using a suitable EM approach applied to $\boldsymbol{y} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$. Let $j = 1$ and set $\widehat{\boldsymbol{\theta}}_j^* = \widehat{\boldsymbol{\theta}}_{EM}$.
2. Increment j. Sample pseudo-data $\boldsymbol{y}_j^* = (\boldsymbol{w}_{j1}^*, \ldots, \boldsymbol{w}_{jn}^*)$ at random from $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)$ with replacement.
3. Calculate $\widehat{\boldsymbol{\theta}}_j^*$ by applying the same EM approach to the pseudo-data $\boldsymbol{y}_j^*$
4. Stop if $j = B$ (typically, $B \geq 1000$); otherwise return to step 2.

The collection of parameter estimates $\widehat{\boldsymbol{\theta}}_1^*, \ldots, \widehat{\boldsymbol{\theta}}_B^*$ can be used to estimate the variance of $\widehat{\boldsymbol{\theta}}$,

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{B} \sum_{j=1}^{B} (\widehat{\boldsymbol{\theta}}_j^* - \overline{\widehat{\boldsymbol{\theta}}^*})(\widehat{\boldsymbol{\theta}}_j^* - \overline{\widehat{\boldsymbol{\theta}}^*})^T,$$

where $\overline{\widehat{\boldsymbol{\theta}}^*}$ is the sample mean of $\widehat{\boldsymbol{\theta}}_1^*, \ldots, \widehat{\boldsymbol{\theta}}_B^*$.

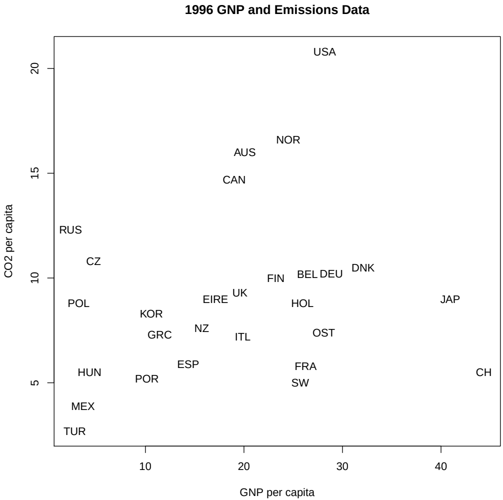
Pros and cons of the bootstrap

1. Advantages:
   - The method is very general, complex analytical derivations are avoided.
   - Allows the estimation of other aspects of the sampling distribution of $\widehat{\boldsymbol{\theta}}$, such as expectation (bias), quantiles, etc.
2. Drawback: bootstrap embeds the EM loop in a second loop of B iterations. May be computationally burdensome when the EM algorithm is slow (because, e.g., of a high proportion of missing data, or high dimensionality).

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 $X = (1, X_1, \ldots, X_p)$, so

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

Observed and complete-data likelihoods

- Observed-data likelihood:

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

- Complete-data likelihood:

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

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 $\boldsymbol{W}_k = \mathrm{diag}(z_{i1}^{(t)}, \ldots, z_{iK}^{(t)})$.

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

- The solution is the WLS estimate of β_k:

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

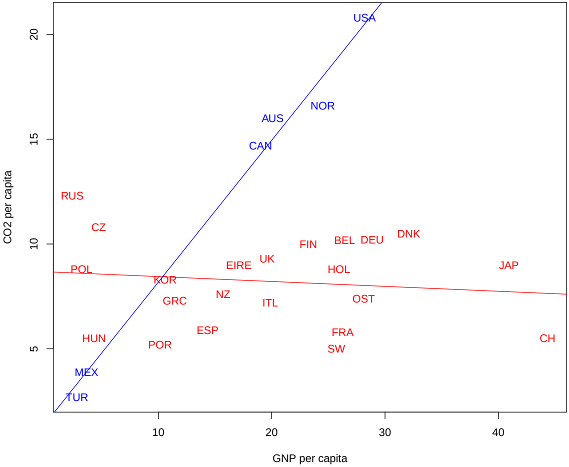
- The value of σ^k minimizing $Q(\theta, \theta^{(t)})$ is the weighted average of the residuals,

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

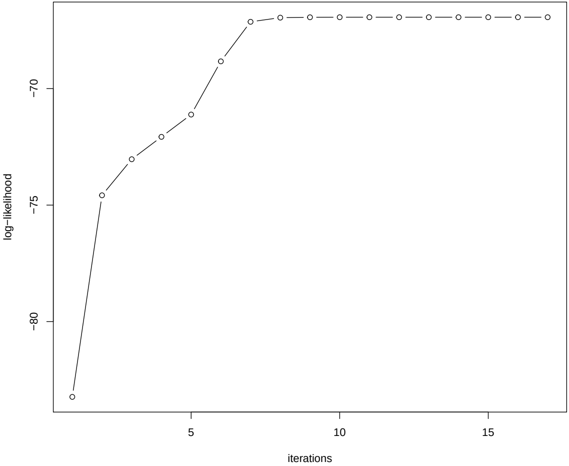
Mixture of regressions using mixtools

```
library(mixtools)
data(CO2data)
attach(CO2data)

CO2reg <- regmixEM(CO2, GNP)
summary(CO2reg)

ii1<-CO2reg$posterior>0.5
ii2<-CO2reg$posterior<=0.5
text(GNP[ii1],CO2[ii1],country[ii1],col='red')
text(GNP[Cii2],CO2[ii2],country[ii2],col='blue')
abline(CO2reg$beta[,1],col='red')
abline(CO2reg$beta[,2],col='blue')
```

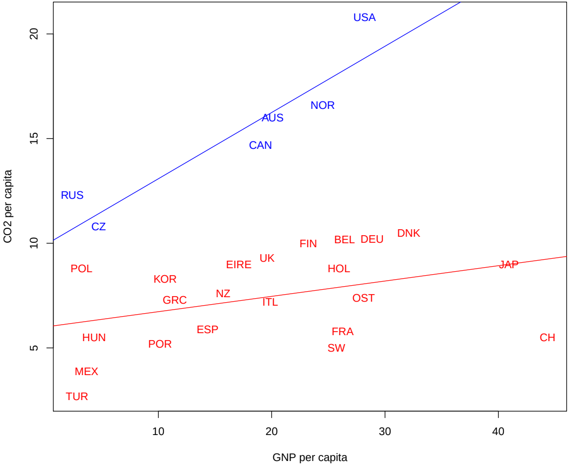
Best solution in 10 runs



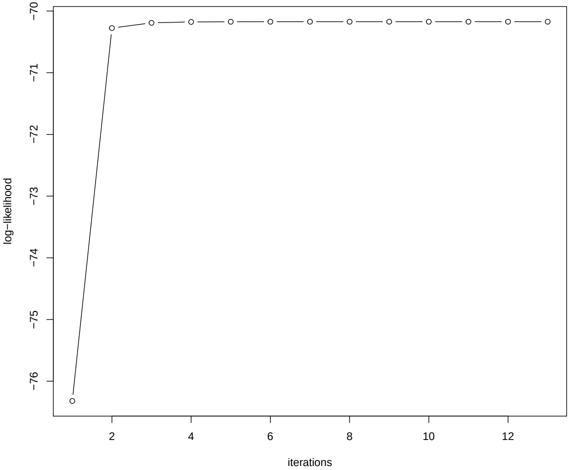
Increase of log-likelihood



Another solution (with lower log-likelihood)



Increase of log-likelihood



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.
- The last term of $Q(\theta, \theta^{(t)})$ can be maximized w.r.t. α using an iterative algorithm, such as the Newton-Raphson procedure. (See remark on next slide)

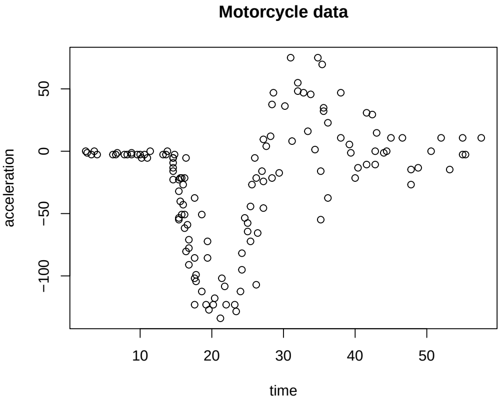
Generalized EM algorithm

- To ensure convergence of EM, we only need to increase (but not necessarily maximize) $Q(\theta, \theta^{(t)})$ at each step.
- Any algorithm that chooses $\theta^{(t+1)}$ at each iteration to guarantee the above condition (without maximizing $Q(\theta, \theta^{(t)})$) is called a Generalized EM (GEM) algorithm.
- Here, we can perform a single step of the Newton-Raphson algorithm to maximize

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

with respect to α.
- Backtracking can be used to ensure ascent.

Example: motorcycle data




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

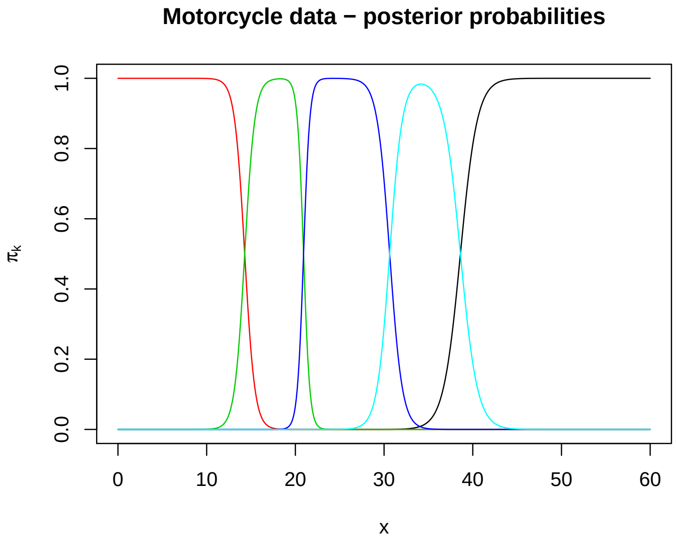
Mixture of experts using flexmix

```
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