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!

Notation

$\mathbf{X}$: Observed variables.

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

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

In Bayesian settings, $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{Y}$ often refer to sets of parameters, rather than data.

Suppose we seek to maximize $L(\boldsymbol{\theta}|\mathbf{x})$ with respect to $\boldsymbol{\theta}$.

Define $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ to be the expectation of the joint log likelihood for the complete data, conditional on the observed data $\mathbf{X} = \mathbf{x}$. Namely

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& \mathrm{E}\left\{ \log L(\boldsymbol{\theta}|\mathbf{Y}) \;\middle|\; \mathbf{x}, \boldsymbol{\theta}^{(t)} \right\} && (49) \\ =& \mathrm{E}\left\{ \log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) \;\middle|\; \mathbf{x}, \boldsymbol{\theta}^{(t)} \right\} && (50) \\ =& \int \left[\log f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) \right] f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) \, d\mathbf{z} && (51) \end{aligned}$$

where (51) emphasizes that $\mathbf{Z}$ is the only random part of $\mathbf{Y}$ once we are given $\mathbf{X} = \mathbf{x}$.

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.**

Trival example:

$Y_1, Y_2 \sim$ i.i.d. $\text{Exp}(\theta)$ **with** $y_1 = 5$ **observed but** y_2 **missing.**

The complete data log likelihood function is

$$\log\{L(\theta|\mathbf{y})\} = \log\{f_{\mathbf{Y}}(\mathbf{y}|\theta)\} = 2\log\{\theta\} - \theta y_1 - \theta y_2. \tag{52}$$

Thus

$$Q(\theta|\theta^{(t)}) = 2\log\{\theta\} - 5\theta - \theta/\theta^{(t)} \tag{53}$$

since $\text{E}\{Y_2|y_1, \theta^{(t)}\} = \text{E}\{Y_2|\theta^{(t)}\} = 1/\theta^{(t)}$ **follows from independence.**

The maximizer of $Q(\theta|\theta^{(t)})$ **is the root of** $2/\theta - 5 - 1/\theta^{(t)} = 0$**. Thus** $\theta^{(t+1)} = \frac{2\theta^{(t)}}{5\theta^{(t)}+1}$**. Converges quickly to** $\hat{\theta} = 0.2$**.**

The E step and M step do not need to be re-derived at each iteration

This example is not realistic. Easy analytic solution. Taking the required expectation is tricker in real applications because one needs to know the conditional distribution of the complete data given the missing data.

Example: Peppered moths

Wing color determined by a single gene with three possible alleles which we denote C, I, and T. C is dominant to I, and T is recessive to I. Thus genotypes CC, CI, and CT result in the *carbonaria* phenotype having solid black coloring. The TT genotype results in the *typica* phenotype with light-colored wings. The II and IT genotypes produce an intermediate *insularia* phenotype with mottled wings.

There are six possible genotypes, but only three phenotypes are measurable in field work.

In UK and USA, *carbonaria* nearly replaced the paler phenotypes in areas affected by coal-fired industries. This change in allele frequencies in the population is cited as an instance when we may observe micro-evolution occurring on a human time scale. The theory (supported by experiments) is that "differential predation by birds on moths that are variously conspicuous against backgrounds of different reflectance" induces selectivity that favors *carbonaria* in times/regions where sooty, polluted conditions reduce the reflectance of the surface of tree bark on which the moths rest. When improved environmental standards reduced pollution, prevalence of the lighter-colored phenotypes increased and *carbonaria* prevalence plummeted.

How do we estimate allele frequencies from phenotype counts?

Under Hardy-Weinberg, if the allele frequencies in the population are p_{C}, p_{I}, and p_{T}, then the genotype frequencies should be p_{C}^2, $2p_{\mathrm{C}}p_{\mathrm{I}}$, $2p_{\mathrm{C}}p_{\mathrm{T}}$, p_{I}^2, $2p_{\mathrm{I}}p_{\mathrm{T}}$, and p_{T}^2, for genotypes CC, CI, CT, II, IT, and TT, respectively.

Capture n moths, of which there are n_{C}, n_{I}, and n_{T} of the *carbonaria*, *insularia*, and *typica* phenotypes, respectively. Thus $n = n_{\mathrm{C}} + n_{\mathrm{I}} + n_{\mathrm{T}}$. If we knew the genotype of each moth rather than merely its phenotype, we could generate genotype counts n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, and n_{TT}. Allele frequencies could then be easily tabulated: each moth with genotype CI contributes one C allele and one I allele, whereas a II moth contributes two I alleles.

Observed data: $\mathbf{x} = (n_{\text{C}}, n_{\text{I}}, n_{\text{T}})$

Complete data: $\mathbf{y} = (n_{\text{CC}}, n_{\text{CI}}, n_{\text{CT}}, n_{\text{II}}, n_{\text{IT}}, n_{\text{TT}})$

Complete data log likelihood is multinomial:

$$\begin{aligned}\log\{f_{\mathbf{Y}}(\mathbf{y}|\mathbf{p})\} = n_{\text{CC}} \log\{p_{\text{C}}^2\} + n_{\text{CI}} \log\{2p_{\text{C}}p_{\text{I}}\} + n_{\text{CT}} \log\{2p_{\text{C}}p_{\text{T}}\} \\ + n_{\text{II}} \log\{p_{\text{I}}^2\} + n_{\text{IT}} \log\{2p_{\text{I}}p_{\text{T}}\} + n_{\text{TT}} \log\{p_{\text{T}}^2\} \\ + \log \binom{n}{n_{\text{CC}}\ n_{\text{CI}}\ n_{\text{CT}}\ n_{\text{II}}\ n_{\text{IT}}\ n_{\text{TT}}}. \quad (54)\end{aligned}$$

Many-to-one mapping from Y to X: $(n_{\text{C}}, n_{\text{I}}, n_{\text{T}}) = (n_{\text{CC}} + n_{\text{CI}} + n_{\text{CT}}, n_{\text{II}} + n_{\text{IT}}, n_{\text{TT}})$.

E-Step:

Conditional on n_{C} and a parameter vector $\mathbf{p}^{(t)} = (p_{\mathrm{C}}^{(t)}, p_{\mathrm{I}}^{(t)})$, the latent counts for the three *carbonaria* genotypes have a three-cell multinomial distribution with count parameter n_{C} and cell probabilities proportional to $\left(p_{\mathrm{C}}^{(t)}\right)^2$, $2p_{\mathrm{C}}^{(t)}p_{\mathrm{I}}^{(t)}$, and $2p_{\mathrm{C}}^{(t)}p_{\mathrm{T}}^{(t)}$.

Same reasoning for other genotypes.

Thus

$$\mathrm{E}\{N_{\mathrm{CC}}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{\mathrm{CC}}^{(t)} = \frac{n_{\mathrm{C}}(p_{\mathrm{C}}^{(t)})^2}{(p_{\mathrm{C}}^{(t)})^2 + 2p_{\mathrm{C}}^{(t)}p_{\mathrm{I}}^{(t)} + 2p_{\mathrm{C}}^{(t)}p_{\mathrm{T}}^{(t)}} \tag{55}$$

and so forth.

Thus

$$\begin{aligned} Q(\mathbf{p}|\mathbf{p}^{(t)}) = n_{\mathrm{CC}}^{(t)} \log\{p_{\mathrm{C}}^2\} + n_{\mathrm{CI}}^{(t)} \log\{2p_{\mathrm{C}}p_{\mathrm{I}}\} + n_{\mathrm{CT}}^{(t)} \log\{2p_{\mathrm{C}}p_{\mathrm{T}}\} \\ + n_{\mathrm{II}}^{(t)} \log\{p_{\mathrm{I}}^2\} + n_{\mathrm{IT}}^{(t)} \log\{2p_{\mathrm{I}}p_{\mathrm{T}}\} + n_{\mathrm{TT}} \log\{p_{\mathrm{T}}^2\} + k(n_{\mathrm{C}}, n_{\mathrm{I}}, n_{\mathrm{T}}, \mathbf{p}^{(t)}). \end{aligned} \tag{56}$$

M-Step:

Differentiating with respect to p_{C} and p_{I} yields

$$\frac{dQ(\mathbf{p}|\mathbf{p}^{(t)})}{dp_{\mathrm{C}}} = \frac{2n_{\mathrm{CC}}^{(t)} + n_{\mathrm{CI}}^{(t)} + n_{\mathrm{CT}}^{(t)}}{p_{\mathrm{C}}} - \frac{2n_{\mathrm{TT}}^{(t)} + n_{\mathrm{CT}}^{(t)} + n_{\mathrm{IT}}^{(t)}}{1 - p_{\mathrm{C}} - p_{\mathrm{I}}} \tag{57}$$

$$\frac{dQ(\mathbf{p}|\mathbf{p}^{(t)})}{dp_{\mathrm{I}}} = \frac{2n_{\mathrm{II}}^{(t)} + n_{\mathrm{IT}}^{(t)} + n_{\mathrm{CI}}^{(t)}}{p_{\mathrm{I}}} - \frac{2n_{\mathrm{TT}}^{(t)} + n_{\mathrm{CT}}^{(t)} + n_{\mathrm{IT}}^{(t)}}{1 - p_{\mathrm{C}} - p_{\mathrm{I}}}. \tag{58}$$

Setting these derivatives equal to zero and solving for p_{C} and p_{I} completes the M step, yielding

$$p_{\mathrm{C}}^{(t+1)} = \frac{2n_{\mathrm{CC}}^{(t)} + n_{\mathrm{CI}}^{(t)} + n_{\mathrm{CT}}^{(t)}}{2n} \tag{59}$$

$$p_{\mathrm{I}}^{(t+1)} = \frac{2n_{\mathrm{II}}^{(t)} + n_{\mathrm{IT}}^{(t)} + n_{\mathrm{CI}}^{(t)}}{2n} \tag{60}$$

$$p_{\mathrm{T}}^{(t+1)} = \frac{2n_{\mathrm{TT}}^{(t)} + n_{\mathrm{CT}}^{(t)} + n_{\mathrm{IT}}^{(t)}}{2n}, \tag{61}$$

Observed data: $n_{\mathrm{C}} = 85$**,** $n_{\mathrm{I}} = 196$**, and** $n_{\mathrm{T}} = 341$

Table 5: EM results for peppered moth example. $R^{(t)}$ is the relative convergence criterion; $D_{\mathrm{C}}^{(t)}$, and $D_{\mathrm{I}}^{(t)}$ are ratios of consecutive errors.

t	$p_{\mathrm{C}}^{(t)}$	$p_{\mathrm{I}}^{(t)}$	$R^{(t)}$	$D_{\mathrm{C}}^{(t)}$	$D_{\mathrm{I}}^{(t)}$
0	0.333333	0.333333			
1	0.081994	0.237406	5.7×10^{-1}	0.0425	0.337
2	0.071249	0.197870	1.6×10^{-1}	0.0369	0.188
3	0.070852	0.190360	3.6×10^{-2}	0.0367	0.178
4	0.070837	0.189023	6.6×10^{-3}	0.0367	0.176
5	0.070837	0.188787	1.2×10^{-3}	0.0367	0.176
6	0.070837	0.188745	2.1×10^{-4}	0.0367	0.176
7	0.070837	0.188738	3.6×10^{-5}	0.0367	0.176
8	0.070837	0.188737	6.4×10^{-6}	0.0367	0.176

The nature of EM

Ascent: **Each M-step increases the log likelihood.**

Convergence: **is linear (slow!). Rate is inversely related to the proportion of missing data.**

Optimization transfer:

$$l(\boldsymbol{\theta}|\mathbf{x}) \geq Q(\boldsymbol{\theta}|\theta^{(t)}) + l(\boldsymbol{\theta}^{(t)}|\mathbf{x}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}). \tag{62}$$

The last two terms in $G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ are constant with respect to $\boldsymbol{\theta}$, 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. EM transfers optimization from l to the surrogate function G, which is more convenient to maximize.

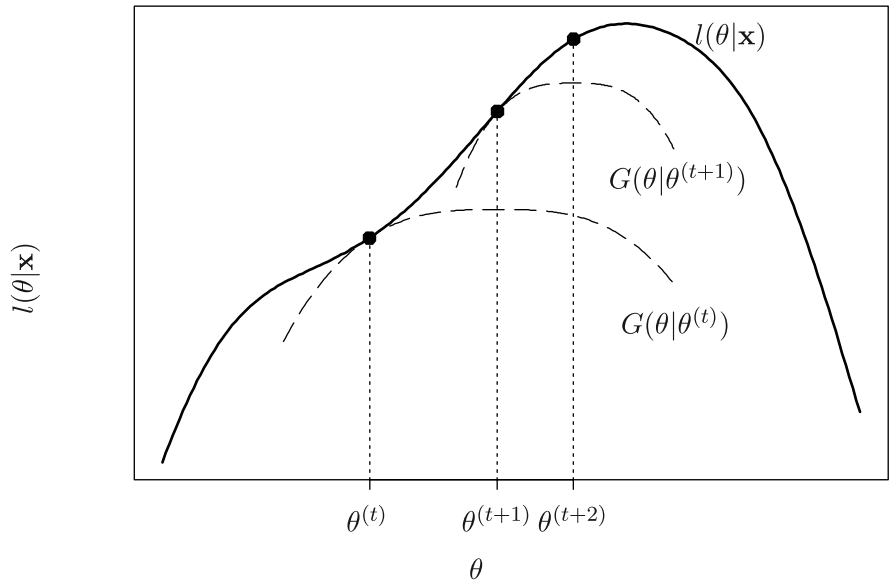

Figure 21: One-dimensional illustration of EM algorithm as a minorization or optimization transfer strategy.

Each E step forms a minorizing function G, and each M step maximizes it to provide an uphill step.

EM in exponential families

1. **E step: Compute the expected values of the sufficient statistics for the complete data, given the observed data and using the current parameter guesses, $\boldsymbol{\theta}^{(t)}$. Let $\mathbf{s}^{(t)} = \mathrm{E}\{\mathbf{s}(\mathbf{Y})|\mathbf{x}, \boldsymbol{\theta}^{(t)}\} = \int \mathbf{s}(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})\, d\mathbf{z}$.**
2. **M step: Set $\boldsymbol{\theta}^{(t+1)}$ to the value which makes the unconditional expectation of the sufficient statistics for the complete data equal to $\mathbf{s}^{(t)}$. In other words, $\boldsymbol{\theta}^{(t+1)}$ solves $\mathrm{E}\{\mathbf{s}(\mathbf{Y})|\boldsymbol{\theta})\} = \mathbf{s}^{(t)}$.**
3. **Return to the E step unless a convergence criterion has been met.**

The missing information principle

Can show

$$\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{63}$$

where

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\left\{\log f_{\mathbf{Z}|\mathbf{X}}(\mathbf{Z}|\mathbf{x},\boldsymbol{\theta}) \,\middle|\, \mathbf{x}, \boldsymbol{\theta}^{(t)}\right\}. \tag{64}$$

Taking second order partial derivatives of (63) and negating both sides yields

$$-\mathrm{l}''(\boldsymbol{\theta}|\mathbf{x}) = -\,\mathrm{Q}''(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}} + \mathrm{H}''(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}} \tag{65}$$

where the primes on Q'' and H'' denote derivatives with respect to the first argument, $\boldsymbol{\theta}$.

Rewrite (65) as

$$\hat{\mathrm{i}}_{\mathbf{X}}(\boldsymbol{\theta}) = \hat{\mathrm{i}}_{\mathbf{Y}}(\boldsymbol{\theta}) - \hat{\mathrm{i}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta}) \tag{66}$$

where $\hat{\mathrm{i}}_{\mathbf{X}}(\boldsymbol{\theta}) = -\mathrm{l}''(\boldsymbol{\theta}|\mathbf{x})$ is the observed information, and $\hat{\mathrm{i}}_{\mathbf{Y}}(\boldsymbol{\theta})$ and $\hat{\mathrm{i}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta})$ will be called the complete information and the missing information, respectively.

The observed information equals the complete information minus the missing information.

Obtaining variance estimates

1. **Bootstrap**
2. **SEM algorithm (next)**

Less common methods include Louis's method (using missing information principle), empirical information, and numerical differentiation.

Supplemented EM (SEM) algorithm

Let Ψ denotes the EM mapping, having fixed point $\hat{\boldsymbol{\theta}}$ and Jacobian matrix $\boldsymbol{\Psi}'(\boldsymbol{\theta})$ with (i,j)th element equaling $\frac{d\Psi_i(\boldsymbol{\theta})}{d\theta_j}$. It can be shown that

$$\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})^T = \hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\hat{\boldsymbol{\theta}})\hat{\mathbf{i}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1} \tag{67}$$

Further use of the missing information principle leads to

$$\widehat{\mathrm{var}}\{\hat{\boldsymbol{\theta}}\} = \hat{\mathbf{i}}_{\mathbf{Y}}(\hat{\boldsymbol{\theta}})^{-1}\left(\mathbf{I} + \boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})^T(\mathbf{I} - \boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})^T)^{-1}\right). \tag{68}$$

SEM is numerically stable and requires little extra work.

SEM algorithm:

1. **Run the EM algorithm to convergence, finding $\hat{\boldsymbol{\theta}}$.**
2. **Restart the algorithm from some $\boldsymbol{\theta}^{(0)}$ near to $\hat{\boldsymbol{\theta}}$. For $t = 0, 1, 2, \ldots$**
   (a) **Take a standard E step and M step to produce $\boldsymbol{\theta}^{(t+1)}$ from $\boldsymbol{\theta}^{(t)}$.**
   (b) **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**

$$r_{ij}^{(t)} = \frac{\Psi_i(\boldsymbol{\theta}^{(t)}(j)) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j} \tag{69}$$

   for $i = 1, \ldots, p$. (Recall that $\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}$.)
   (c) **Stop when all $r_{ij}^{(t)}$ have converged**
3. **The (i, j)th element of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ equals $\lim_{t\to\infty} r_{ij}^{(t)}$. Plug the final estimate of $\boldsymbol{\Psi}'(\hat{\boldsymbol{\theta}})$ into (68) to get the variance.**

Facilitating the E-step

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 $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})$. Each $\mathbf{Z}_j^{(t)}$ is a vector of all the missing values needed to complete the observed dataset, so $\mathbf{Y}_j = (\mathbf{x}, \mathbf{Z}_j)$ denotes a completed dataset where the missing values have been replaced by $\mathbf{Z}_j$.**
2. **Calculate $\hat{Q}^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f_{\mathbf{Y}}(\mathbf{Y}_j^{(t)}|\boldsymbol{\theta})$.**

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 $m^{(t)}$.

Facilitating the M-step

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.

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

$$\mathbf{g}_s(\boldsymbol{\theta}) = \mathbf{g}_s(\boldsymbol{\theta}^{(t+(s-1)/S)}) \tag{70}$$

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.

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.

(Sketched) Example:

Multivariate regression (Meng and Rubin, 1993)

Let $\mathbf{U}_1, \ldots, \mathbf{U}_n$ be independent where

$$\mathbf{U}_i \sim N_d\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}\right) \tag{71}$$

for $\boldsymbol{\mu}_i = \mathbf{V}_i\boldsymbol{\beta}$, where the $\mathbf{V}_i$ are known $d \times p$ design matrices, and $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are unknown parameters.

To estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ when some elements of some of the $\mathbf{U}_i$ are missing, we can partition each CM cycle into two stages: one for $\boldsymbol{\beta}$ and one for $\boldsymbol{\Sigma}$.

The algebra is messy, but the idea turns out to be simple.

- **E-step: Find the expected values of the complete data sufficient statistics conditional on the observed data and current guesses $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\Sigma}^{(t)}$. (The sufficient stats are $\sum_{i=1}^{n} U_{ij}$ for $j = 1, \ldots, d$ and $\sum_{i=1}^{n} U_{ij}U_{ik}$ for $j, k = 1, \ldots, d$.)**
- **CM 1: $\boldsymbol{\beta}^{(t+1/2)}$ is the weighted least squares estimate given $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(t)}$, computed using the sufficient stats from the E-step.**
- **CM 2: $\boldsymbol{\Sigma}^{(t+2/2)}$ is the sample covariance matrix of the completed data, using $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t+1/2)}$.**
- **Return to the E-step**

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:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mathbf{Q}'(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \tag{72}$$

$$= \boldsymbol{\theta}^{(t)} - \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})^{-1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \mathbf{l}'(\boldsymbol{\theta}^{(t)}|\mathbf{x}) \tag{73}$$

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

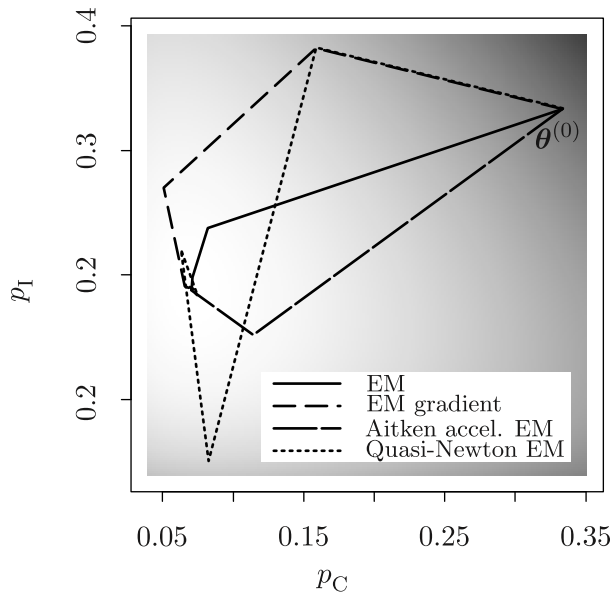

Figure 22: Steps taken by the EM gradient algorithm (long dashes). Ordinary EM steps are shown with the solid line. Steps from two methods from later sections (Aitken and quasi-Newton acceleration) are also shown, as indicated in the key. The observed data log likelihood is shown with the grey scale, with light shading corresponding to high likelihood. All algorithms were started from $p_{\mathrm{C}} = p_{\mathrm{I}} = 1/3$.

EM Acceleration Methods

Standard EM is reliable but slow. Acceleration methods use the EM setup to motivate particular forms of Newton-like steps.

Aitken acceleration: Equivalent to applying the Newton-Raphson method to find a zero of $\Psi(\boldsymbol{\theta}) - \boldsymbol{\theta}$ where Ψ is the mapping defined by the ordinary EM algorithm producing $\boldsymbol{\theta}^{(t+1)} = \Psi(\boldsymbol{\theta}^{(t)})$.

Quasi-Newton acceleration: Take

$$\mathbf{M}^{(t)} = \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \mathbf{B}^{(t)}$$

where $\mathbf{B}^{(t)}$ is crafted to make $\mathbf{M}^{(t)}$ approximate $\mathrm{l}''(\boldsymbol{\theta}^{(t)}|\mathrm{x})$. A potentially superior strategy approximates $(\mathrm{l}'')^{-1}$ instead of l''.

A final example: the Baum-Welch algorithm

A *hidden Markov model* can be used to describe the joint probability of a sequence of unobserved ('hidden') discrete state variables, $\mathbf{H} = (H_0, \ldots, H_n)$, and a sequence of corresponding observed variables $\mathbf{O} = (O_0, \ldots, O_n)$ for which O_i is dependent on H_i for each i.

We say that H_i emits O_i.

Let the (discrete) state spaces for elements of H and O be $\mathcal{H}$ and $\mathcal{E}$, respectively.

Such models are extremely useful in statistical genetics, signal processing, and environmental time series. For example, the hidden states can be "weather patterns" and the observed variables can be measurable weather events like rain, snow, smog, etc.

The Baum-Welch algorithm fits such models. It is an EM algorithm.

The parameters of a HMM are $\boldsymbol{\theta} = (\boldsymbol{\pi}, \mathbf{P}, \mathbf{E})$, where $\boldsymbol{\pi}$ is a vector of initial state probabilities, P is a matrix of transition probabilities, E is a matrix of emission probabilities.

For an observed sequence o, define the *forward variables* **to be**

$$\alpha(i,h) = P[\mathbf{O}_{\leq i} = \mathbf{o}_{\leq i},\ H_i = h] \tag{74}$$

and the *backward variables* **to be**

$$\beta(i,h) = P[\mathbf{O}_{>i} = \mathbf{o}_{>i} | H_i = h] \tag{75}$$

for $i = 1, \ldots, n$ and each $h \in \mathcal{H}$. Dependence on $\boldsymbol{\theta}$ is suppressed.

There are efficient recursive formulas to calculate the $\alpha(i,h)$ and $\beta(i,h)$.

The E-step boils down to:

Let $N(h)$ denote the number of times $H_0 = h$, let $N(h, h^*)$ denote the number of transitions from h to h^*, and let $N(h, o)$ denote the number of emissions of o when the underlying state is h. Then:

$$\mathrm{E}\{N(h)\} = \frac{\alpha(0,h)\beta(0,h)}{P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]} \tag{76}$$

$$\mathrm{E}\{N(h,h^*)\} = \sum_{i=0}^{n-1} \frac{\alpha(i,h)p(h,h^*)e(h^*,o_{i+1})\beta(i+1,h^*)}{P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]} \tag{77}$$

$$\mathrm{E}\{N(h,o)\} = \sum_{i:\ O_i = o} \frac{\alpha(i,h)\beta(i,h)}{P[\mathbf{O} = \mathbf{o}|\boldsymbol{\theta}]}. \tag{78}$$

The M-step boils down to:

$$\pi(h)^{(t+1)} = \frac{\mathrm{E}\{N(h)|\boldsymbol{\theta}^{(t)}\}}{\sum_{h^* \in \mathcal{H}} \mathrm{E}\{N(h^*)|\boldsymbol{\theta}^{(t)}\}} \tag{79}$$

$$p(h, h^*)^{(t+1)} = \frac{\mathrm{E}\{N(h, h^*)|\boldsymbol{\theta}^{(t)}\}}{\sum_{h^{**} \in \mathcal{H}} \mathrm{E}\{N(h, h^{**})|\boldsymbol{\theta}^{(t)}\}} \tag{80}$$

$$e(h, o)^{(t+1)} = \frac{\mathrm{E}\{N(h, o)|\boldsymbol{\theta}^{(t)}\}}{\sum_{o^* \in \mathcal{E}} \mathrm{E}\{N(h, o^*)|\boldsymbol{\theta}^{(t)}\}}. \tag{81}$$

(Oversimplified) HMM data:

A region is hypothesized to cycle between two weather patterns: "tending-wet" and "tending-dry". The probability of measurable precipitation is w for a day during a tending-wet cycle and d for a day in the tending-dry cycle. Cycles switch from dry to wet or vice versa with probability s.

Observed Data: was there measurable precipitation on each of 200 consecutive days?

Latent Data: which pattern was extant on each day.

Goal: Estimate w, d, and s.

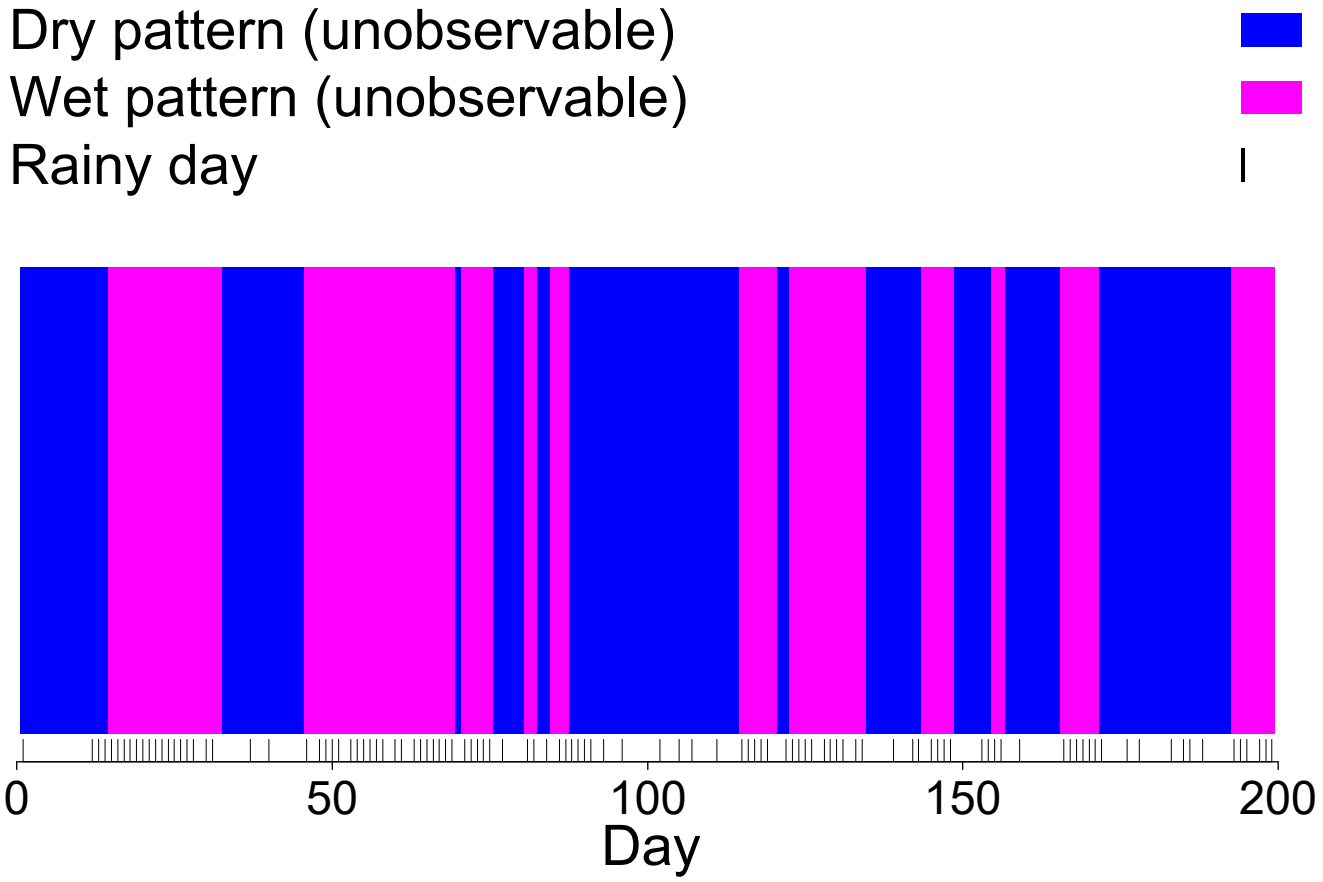

Figure 23: Simple data for a meteorological hidden Markov model.

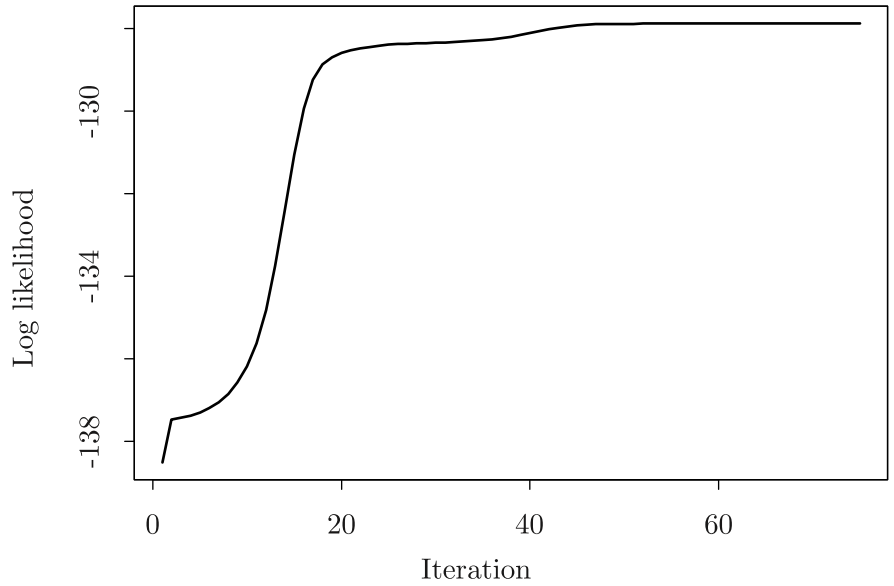

Figure 24: Log likelihood for parameters of hidden Markov model, against Baum-Welch iteration.

As an EM algorithm, the Baum-Welch approach reliably ascends as iterations progress.

Iteration, t	$d^{(t)}$	$w^{(t)}$	$s^{(t)}$
0	0.4500	0.5500	0.4000
1	0.4941	0.6042	0.3983
5	0.4632	0.6365	0.3872
10	0.3853	0.7127	0.3467
20	0.2217	0.8658	0.1750
30	0.2349	0.8528	0.1444
40	0.2475	0.8497	0.1314
50	0.2595	0.8468	0.1190
75	0.2608	0.8460	0.1174
100	0.2608	0.8460	0.1174

Table 6: Selected iterations during Baum-Welch search for MLEs for hidden Markov model for rain data. The data were simulated with $d = 0.25$, $w = 0.85$, and $s = 0.10$.