

Statistical Analysis with **Missing** Data

Module 8

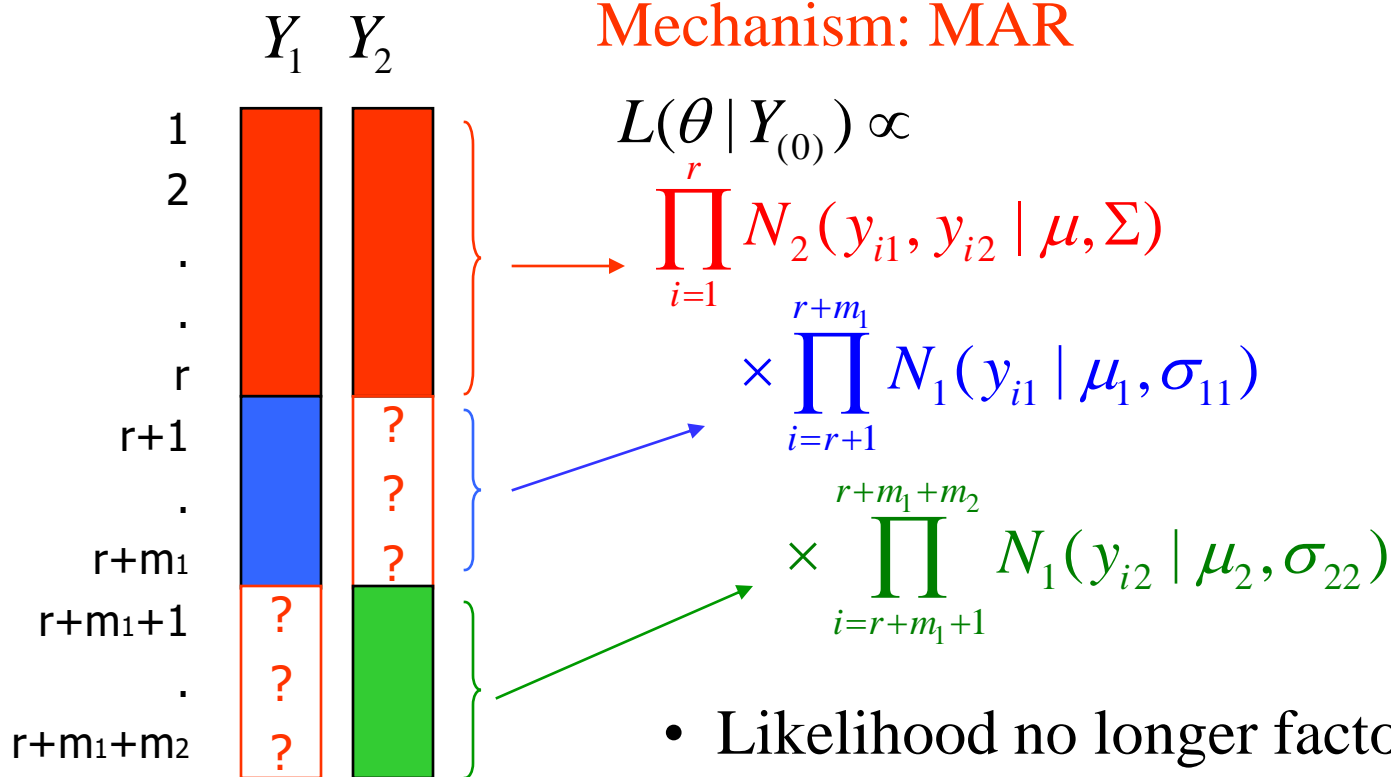
Maximum Likelihood: EM algorithm and
extensions



ML for General Bivariate Pattern

Model $(y_{i1}, y_{i2}) \sim N_2(\mu, \Sigma)$

Mechanism: MAR



Numerical Maximization

- Standard numerical methods (Newton-Raphson, Scoring) can be used to maximize the likelihood:

$S(\theta | Y_{(0)})$ = score function

$I(\theta | Y_{(0)})$ = observed information matrix

$J(\theta | Y_{(0)})$ = expected information matrix

Newton - Raphson: $\theta^{(t+1)} = \theta^{(t)} + \left[I^{-1}(\theta^{(t)} | Y_{obs}) \right] S(\theta^{(t)} | Y_{obs})$

Scoring: $\theta^{(t+1)} = \theta^{(t)} + \left[J^{-1}(\theta^{(t)} | Y_{obs}) \right] S(\theta^{(t)} | Y_{obs})$

- A popular alternative is the *Expectation-Maximization (EM)* algorithm:

- tuned to missing-data problem
- link with imputation

(Dempster, Laird and Rubin 1977; Little and Rubin 2019, chapter 8)

EM algorithm

Choose starting values $\theta^{(0)}$

Current estimate at iteration t : $\theta^{(t)}$

Iteration $t+1$

E(xpectation) Step:

Compute $Q(\theta, \theta^{(t)}) \equiv E\left[\log p(Y_{(0)}, \mathbf{Y}_{(1)} \mid \theta) \mid Y_{(0)}, \theta^{(t)}\right]$

M(aximization) Step:

Choose $\theta^{(t+1)}$ to maximize $Q(\theta, \theta^{(t)})$ with respect to θ

EM algorithm

Key idea: each iteration increases $\log L(\theta | Y_{(0)})$

$$\log p(Y_{(0)}, \mathbf{Y}_{(1)} | \theta) = \log p(Y_{(0)} | \theta) + \log p(\mathbf{Y}_{(1)} | Y_{(0)}, \theta)$$

$$\begin{aligned} Q(\theta, \theta^{(t)}) &\equiv E \left[\log p(Y_{(0)}, \mathbf{Y}_{(1)} | \theta) | Y_{(0)}, \theta^{(t)} \right] \\ &= \log p(Y_{(0)} | \theta) + H(\theta, \theta^{(t)}), \text{ where} \end{aligned}$$

$$H(\theta, \theta^{(t)}) = E \left[\log p(\mathbf{Y}_{(1)} | Y_{(0)}, \theta) | Y_{(0)}, \theta^{(t)} \right]$$

$$\log p(Y_{(0)} | \theta) = \ell(\theta | Y_{(0)}) = Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)})$$

$H(\theta, \theta^{(t)})$ is maximized at $\theta = \theta^{(t)}$ (Jensen)

So $Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta^{(t)}, \theta^{(t)})$, $H(\theta^{(t+1)}, \theta^{(t)}) \leq H(\theta^{(t)}, \theta^{(t)})$

Hence $\ell(\theta^{(t+1)} | Y_{(0)}) \geq \ell(\theta^{(t)} | Y_{(0)})$

EM for Exponential Family Models

- Many complete-data models belong to the exponential family, with log-likelihoods of the form:

$$\ell(\theta | Y) = \log b(Y) + \sum_{j=1}^d \theta_j s_j(Y) - a(\theta)$$

$s(Y) = (s_1(Y), \dots, s_d(Y))$ = complete-data sufficient statistics

$\theta = (\theta_1, \dots, \theta_d)$ = natural parameters

Examples: Normal, Poisson, Multinomial, Gamma, Exponential, Generalized Linear Models with canonical links

EM for these models has a particularly simple form...

EM for Exponential Family Models

$$\ell(\theta | Y) = \log b(Y) + \sum_{j=1}^d \theta_j s_j(Y) - a(\theta)$$

Iteration t of EM:

E-step: $s_j^{(t+1)} = E(s_j(Y) | Y_{(0)}, \theta = \theta^{(t)})$

“Imputes the complete-data sufficient statistics by expectation given data and current parameter estimates”

M-step:

$\theta^{(t+1)}$ to maximize $Q(\theta | \theta^{(t)}) = \ell(\theta | s_j(Y) = s_j^{(t+1)}, j = 1, \dots, d)$

“Complete-data maximization with complete-data sufficient statistics replaced by estimates from E-step”

Theory shows EM iterates increase the likelihood and (under conditions) converge to ML estimate of θ

E-Step for Bivariate Normal Data

$$s_1 = \sum_{i=1}^n y_{i1}, s_2 = \sum_{i=1}^n y_{i2}, s_3 = \sum_{i=1}^n y_{i1}^2, s_4 = \sum_{i=1}^n y_{i2}^2, s_5 = \sum_{i=1}^n y_{i1} y_{i2}$$

$$s_j^{(t)} = E(s_j | Y_{(0)}, \theta^{(t)}), \theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$$

$$s_1^{(t)} = \sum_{i=1}^r y_{i1} + \sum_{i=r+1}^{r+m_1} y_{i1} + \sum_{i=r+m_1+1}^n \hat{y}_{i1}^{(t)}$$

$$s_3^{(t)} = \sum_{i=1}^r y_{i1}^2 + \sum_{i=r+1}^{r+m_1} y_{i1}^2 + \sum_{i=r+m_1+1}^n \left(\hat{y}_{i1}^{(t)2} + \hat{\sigma}_{11.2}^{(t)} \right)$$

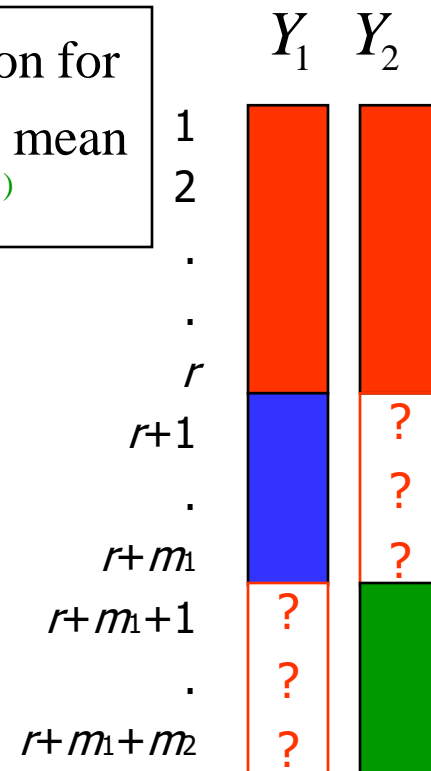
$$s_5^{(t)} = \sum_{i=1}^r y_{i1} y_{i2} + \sum_{i=r+1}^{r+m_1} y_{i1} \hat{y}_{i2}^{(t)} + \sum_{i=r+m_1+1}^n \hat{y}_{i1}^{(t)} y_{i2}$$

$$\hat{y}_{i1}^{(t)} = \hat{\beta}_{10.2}^{(t)} + \hat{\beta}_{12.2}^{(t)} y_{i2} \quad \hat{y}_{i2}^{(t)} = \hat{\beta}_{20.1}^{(t)} + \hat{\beta}_{21.1}^{(t)} y_{i1}$$

Similar expressions for $s_2^{(t)}, s_4^{(t)}$

Correction for
imputing mean

$$\hat{y}_{i1}^{(t)}$$



M-Step for Bivariate Normal Data

- M-Step as for complete data, with complete-data sufficient statistics estimated from the E-Step:

$$\mu_1^{(t+1)} = \frac{1}{n} s_1^{(t)}$$

$$\sigma_{11}^{(t+1)} = \frac{1}{n} \left(s_3^{(t)} - \left(s_1^{(t)} \right)^2 / n \right)$$

$$\sigma_{12}^{(t+1)} = \frac{1}{n} \left(s_5^{(t)} - s_1^{(t)} s_2^{(t)} / n \right), \text{ etc.}$$

- In practice working means can be included in these calculations to reduce rounding errors.

Multivariate Normal EM

- M-Step: as for complete data, using estimates of complete-data (cd) sufficient statistics
- E-Step: given current parameter estimates $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ enter loop over cases; for case i :
 - (a) Sweep for regression coefficients, res cov matrix of missing variables on observed variables, as functions of $\theta^{(t)}$;
 - (b) fill in missing values y_{ij} with estimates $\hat{y}_{ij}^{(t)}$ computed using regression equation with coefficients from (a);
 - (c) add case i to vector of running means, sum of squares and cross products (sscp) matrix;
 - (d) add residual covariance matrix of missing variables given observed variables in case i to sscp matrix;
- recompute cd sufficient statistics from mean, sscp matrix

Multivariate Normal EM ctd.

- Computations are greatly simplified using the SWEEP operator (LR section 7.4.3)
- Standard errors are harder to compute; one alternative is to calculate bootstrap se's using bootstrap samples
- ML estimates of functions of (μ, Σ) are found by evaluating them at the ML estimates. Hence in particular can get ML estimates of correlations, regression coefficients from incomplete data

EM for multinomial data

- EM for multinomial data (that is, partially classified contingency tables) is very straightforward:
 - E-Step: Allocate supplemental margins into the full table using current estimates of cell probabilities
 - M-Step: Re-estimate cell probabilities as proportions based on filled-in data
- Loglinear models: E-Step the same, M-Step replaced by max subject to constraints of model

Example: 2x2 Table

		Y_2	
		1	2
Y_1	1	100	50
	2	75	75

		Y_2	
		1	2
Y_1	1	30	
	2	60	

		Y_2	
		1	2
Y_1	1	28	60
	2		

Model: fully-classified counts multinomial with probabilities

$$\theta = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$$

$$\theta_{jk} = \Pr(Y_1 = j, Y_2 = k)$$

Mechanism: MAR

EM for 2x2 table

Initial estimates $\{\theta_{jk}^{(0)}\}$ from complete cases:

		Y_2	
		1	2
Y_1	1	$100/300 = 1/3$	$50/300 = 1/6$
	2	$75/300 = 1/4$	$75/300 = 1/4$

First E-Step: allocate supplemental margins into table

$100 + 20 + 16$	$50 + 10 + 24$	30
$75 + 30 + 12$	$75 + 30 + 36$	60

28 60

=

136	84
117	141

478

EM for 2x2 table ctd.

First M-Step: New estimates $\{\theta_{jk}^{(1)}\}$ from filled-in data

$$\begin{bmatrix} 136/478 & 84/478 \\ 117/478 & 141/478 \end{bmatrix} = \begin{bmatrix} .2845 & .1757 \\ .2448 & .2950 \end{bmatrix}$$

Second E-Step: reallocate margins using new probabilities:

$$\begin{array}{cc} 100 + 18.6 + 15.1 & 50 + 11.4 + 20.6 & 30 \\ 75 + 27.2 + 12.9 & 75 + 32.8 + 39.4 & 60 \end{array} = \begin{bmatrix} 133.7 & 82.0 \\ 115.1 & 147.2 \end{bmatrix}$$

478

28 60

Successive EM iterates

	Iteration				
	0	1	2	3	4
θ_{11}	0.33	0.28	0.28	0.28	0.28
θ_{12}	0.17	0.17	0.17	0.17	0.17
θ_{21}	0.25	0.24	0.24	0.24	0.24
θ_{22}	0.25	0.31	0.31	0.31	0.31

EM Pros and Cons

Pros

Often easy to program
Output has useful statistical interpretation
No need to compute and invert information matrices
Stable convergence: likelihood increases

Cons

Slow convergence
Convergence can be hard to establish
Standard errors not an output

EM with parameter constraints

Often models are fitted that place constraints on the model parameters

- loglinear models in contingency tables
- models for repeated measures with restricted covariance structures, e.g. compound symmetry:

$$\Sigma(\theta_1, \theta_2) = \begin{pmatrix} \theta_1 & \theta_2 & \dots & \theta_2 \\ \theta_2 & \theta_1 & \theta_2 & \theta_2 \\ \dots & \theta_2 & \dots & \theta_2 \\ \theta_2 & \theta_2 & \theta_2 & \theta_1 \end{pmatrix}$$

EM with parameter constraints

- A useful feature of EM is that parameter constraints do not change the E-Step, which is the missing-data part of the problem. Specifically, EM is as follows:
- E-Step is unaffected
- M-Step maximizes expected complete-data loglikelihood subject to the model constraints
 - If this step yields explicit estimates, this is an easy modification of EM for the unconstrained model
 - If iterative, standard software may be available. For example, loglinear models in contingency tables.

Rate of convergence of EM

$$i_{\text{obs}} = i_{\text{com}} - i_{\text{mis}} \quad (\text{LR Eq 8.27})$$

observed information = complete info - missing info

$$i_{\text{obs}} = I(\theta | Y_{\text{obs}}) |_{\theta=\theta^*} = \text{observed information}$$

$$\theta^* = \text{converged value of } \theta$$

$$i_{\text{com}} = -D^{20} Q(\theta | \theta) |_{\theta=\theta^*} = \text{complete information}$$

$$i_{\text{mis}} = -D^{20} H(\theta | \theta) |_{\theta=\theta^*} = \text{missing information}$$

Means:
differentiate
first argument
of Q twice

"Fraction of missing information": $DM = i_{\text{mis}} i_{\text{com}}^{-1}$

(DM is a matrix when θ is a vector)

Standard errors and EM

- Some approaches to computing standard errors are:
 - Compute and invert information matrix at end
 - *Supplemented* EM algorithm: computes information matrix using EM iterates
 - Apply to Bootstrap samples and build bootstrap distribution of estimates
 - Bayes' simulation methods (more below)

Speeding EM by clever choices of missing data

Speed of convergence: often

$$|\theta^{(t+1)} - \theta^{(t)}| \approx \lambda |\theta^{(t)} - \theta^*|$$

λ = largest eigenvalue of DM

MORE MISSING DATA = SLOWER CONVERGENCE

- Sometimes the speed can be increased by clever choice of missing data.

Example 1: Multivariate T model.

- Robust estimation for symmetrically distributed continuous data (See LR Section 12.2.2)
- One approach is to replace normal by a distribution with longer than normal tails, such as the t distribution.
- Suppose x_i are a simple random sample from the multivariate t distribution with mean μ , scale matrix Ψ and known degrees of freedom:

$$x_i \mid \theta \sim_{\text{iid}} t_K(\mu, \Psi, \nu), i = 1, \dots, n$$

Note: if $x_i \mid \mathbf{q}_i, \theta \sim_{\text{iid}} N_K(\mu, \Psi / \mathbf{q}_i)$, $\mathbf{q}_i \sim \chi_\nu^2 / \nu$

then marginally, $x_i \mid \theta \sim_{\text{iid}} t_K(\mu, \Psi, \nu)$

Construct an EM algorithm based on this property...

EM for t model

- In LR Section 12.2.2 it is shown that ML estimates can be obtained for this model by introducing fake missing data $Q = (q_1, \dots, q_n)$, where

$$x_i \mid q_i, \theta \sim_{\text{iid}} N_K(\mu, \Psi / q_i), \quad q_i \sim \chi_\nu^2 / \nu$$

$$\text{E-step of EM: } w_i^{(t)} = E(q_i \mid x_i, \mu^{(t)}, \Psi^{(t)}) = \frac{\nu + K}{\nu + d_i^{(t)}}$$

$$d_i^{(t)} = (x_i - \mu^{(t)})^T (\Psi^{(t)})^{-1} (x_i - \mu^{(t)})$$

= measure of how far x_i is from $\mu^{(t)}$

M-step of EM: compute $\mu^{(t+1)}, \Psi^{(t+1)}$ by weighted least squares:

$$\mu^{(t+1)} = \sum_{i=1}^n w_i^{(t)} x_i / \sum_{i=1}^n w_i^{(t)}; \quad \Psi^{(t+1)} = \sum_{i=1}^n w_i^{(t)} (x_i - \mu^{(t+1)}) (x_i - \mu^{(t+1)})^T / n$$

Note: cases far from mean receive low weight

Speeding EM for t model

Suppose we augment the data as $(x_i, q_i(a))$

$$x_i \mid q_i(a), \theta \sim_{\text{iid}} N_K \left(\mu, |\Psi|^{-a} \Psi / q_i(a) \right), q_i(a) \sim |\Psi|^{-a} \chi_\nu^2 / \nu$$

Note: still leads to required t model for x_i

$a = 0$ yields previous algorithm


fraction of missing information depends on choice of a

Choosing $a = a_{\text{opt}} = 1 / (\nu + K)$ minimizes this fraction,

hence speeds up rate of convergence. New M-step is

$$\mu^{(t+1)} = \sum_{i=1}^n w_i^{(t)} x_i / \sum_{i=1}^n w_i^{(t)};$$

replacing n by this yields
dramatic increases in speed!

$$\Psi^{(t+1)} = \sum_{i=1}^n w_i^{(t)} \left(x_i - \mu^{(t+1)} \right) \left(x_i - \mu^{(t+1)} \right)^T / \sum_{i=1}^n w_i^{(t)}$$


PX-EM

- See LR Section 8.5.3 for extensions of this idea: parameter expanded EM
- Adding parameters that can't be estimated can reduce the fraction of missing information, and hence speed EM
- General principles for how to do this are not clear – all we have are special cases

Beyond EM

- Generalized EM (GEM) algorithms replace maximization in M Step by a step that increases the Q function; examples are:

ECM (LR Section 8.5.1) Conditional maximization replaces full maximization when the latter is iterative

ECME (LR Section 8.5.2) Some conditional M steps max full likelihood rather than Q function

- Hybrid methods (LR Section 8.6) combine EM and Newton or other steps

Extensions of EM when the M-Step is Hard

- Recall that the basic EM algorithm has two steps:
- E-Step: compute the Q function, the expected value of the complete-data loglikelihood given the observed data and current parameter estimates
- M-Step: maximize the Q function to obtain new estimate
- When the M-Step of EM does not have an explicit solution and itself requires iteration, the EM algorithm involves a double iteration.

GEM algorithms

- Computation time can be reduced by modifying the M-Step to merely increase, rather than maximize, the value of the Q function. The resulting algorithm is a generalized EM (GEM) algorithm.
- We consider here three GEM algorithms, the EM gradient algorithm, the ECM algorithm, and a variant of ECM called the ECME algorithm. GEM algorithms do not always share the convergence properties of EM, but these algorithms do.

The EM gradient algorithm

- If the M-Step is not explicit, one might apply a Newton algorithm for the M-Step. Rather than iterating the Newton algorithm to convergence, the EM gradient algorithm takes one step of a Newton algorithm in the M-Step.
- Since such a step is not guaranteed to increase the Q function, a fractional step might be taken to ensure that the likelihood is increased.
- The resulting algorithm is a GEM algorithm with a single iterative loop that has similar properties to EM. (Jamshidian and Jennrich 1993, Lange 1995a,b).

The ECM algorithm

- Expectation -- Conditional Maximization (ECM) algorithm
- Breaks M-Step into S conditional maximization (CM) steps that are easier to carry out than joint maximization
- CM step s maximizes Q -function with respect to θ fixing some vector function $g_s(\theta)$ at its current values.
- The ECM algorithm is a generalized EM (GEM), meaning each CM step increases (but does not maximize) the Q function.

The ECM algorithm

- A common of ECM writes $\theta = (\theta_1, \dots, \theta_J)$ where θ_j is a subset of the parameters, and sets $g_s(\theta)$ to be the set of all the components of θ except θ_j . The M-Step is then replaced by the following J fractional CM-Steps:

CM $(t + 1 / J)$: Max $Q(\theta_1, \theta_2^{(t)}, \dots, \theta_J^{(t)} \mid \theta^{(t)})$ wrt $\theta_1 \rightarrow \theta_1^{(t+1)}$

CM $(t + 2 / J)$: Max $Q(\theta_1^{(t+1)}, \theta_2, \theta_3^{(t)}, \dots, \theta_J^{(t)} \mid \theta^{(t)})$ wrt $\theta_2 \rightarrow \theta_2^{(t+1)}$

...

CM $(t + J / J)$: Max $Q(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \dots, \theta_{J-1}^{(t+1)}, \theta_J \mid \theta^{(t)})$ wrt $\theta_J \rightarrow \theta_J^{(t+1)}$

E: compute $Q(\theta, \theta^{(t+1)})$

Example 2. ECM for multivariate repeated-measures model

- A flexible model for repeated-measures data assumes n independent observations from the k -variate normal distribution: $Y_i \sim N_k(X_i\beta, \Sigma), i = 1, \dots, n$
- Where X_i is a known $(k \times p)$ design matrix, β is a $(p \times 1)$ matrix of regression coefficients and Σ is an unstructured covariance matrix.
- With complete data, explicit ML estimates of the parameters are not available, but can be found iteratively by cycling between estimation of β given Σ and estimation of Σ given β .

Multivariate normal example

- That is, given current estimates $\beta^{(t)}, \Sigma^{(t)}$

CM $(t + 1 / 2)$: Given $\Sigma = \Sigma^{(t)}$, compute

$$\beta^{(t+1)} = \left(\sum_{i=1}^n X_i^T (\Sigma^{(t)})^{-1} X_i \right) \left(\sum_{i=1}^n X_i^T (\Sigma^{(t)})^{-1} Y_i \right)$$

CM $(t + 2 / 2)$: Given $\beta = \beta^{(t+1)}$, compute

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i \beta^{(t+1)}) (Y_i - X_i \beta^{(t+1)})^T$$

- This is called the *Iterative Conditional Modes* algorithm

ECM for normal model

- Suppose now that Y_i have missing components and the mechanism is MAR.
- E-Step of EM is similar to that for the multivariate normal model with constant mean μ , except that the mean $\mu^{(t)}$ involved in predictions of missing components of Y_i and $Y_i Y_i^T$ is replaced by the current estimate $X_i \beta^{(t)}$ of the mean for that case.
- M-step for $\beta^{(t+1)}, \Sigma^{(t+1)}$ can be replaced by the two CM steps given above, with $\theta_1 = \beta$, $\theta_2 = \Sigma$, and Y_i and $Y_i Y_i^T$ replaced by their conditional expectations given the observed components $Y_{\text{obs},i}$ and current parameter estimates $\theta^{(t)}$.

Example 3. Loglinear models for contingency tables with supplemental margins

- E-Step: classifies supplemental margins into the full table using the current estimates of the cell probabilities.
- M-Step: ML applied to the filled in data-set. For some models this requires an iterative algorithm such as iterative proportional fitting (IPF):
 - fitted marginal counts that are sufficient under the model are adjusted sequentially to match the observed marginal counts.
- E.g. model with no three-way interactions $\{AB, BC, CA\}$ for a three-way contingency table with factors A (levels $i = 1, \dots, I$), B (levels $j = 1, \dots, J$), and C (levels $k = 1, \dots, K$).
- IPF: given current estimates, compute new estimates in three steps:

ECM for 3 way contingency table

$$\text{CM Step } (t + 1 / 3): \theta_{ijk}^{(t+1/3)} = \theta_{ij(k)}^{(t)} \frac{y_{ij+}^{(t)}}{N} \text{ for all } i, j, k,$$

$$\text{where } \theta_{ij(k)}^{(t)} = \theta_{ijk}^{(t)} / \theta_{ij+}^{(t)}$$

$$\text{CM Step } (t + 2 / 3): \theta_{ijk}^{(t+2/3)} = \theta_{i(j)k}^{(t+1/3)} \frac{y_{i+k}^{(t)}}{N} \text{ for all } i, j, k.$$

$$\text{CM Step } (t + 3 / 3): \theta_{ijk}^{(t+3/3)} = \theta_{(i)jk}^{(t+2/3)} \frac{y_{+jk}^{(t)}}{N} \text{ for all } i, j, k.$$

Iterating these steps yields the M-Step of EM

ECM does just one iteration, then goes to next E-Step

This is faster!

Convergence of ECM

- The key property for the ECM algorithm to have the convergence properties of EM is that the set of constraint functions is *space-filling*, in the sense that the set of constrained maximizations provides for maximization over the entire unrestricted parameter space of rather than over a restricted subspace.
- Technical details are provided in Meng and Rubin (1993).

ECME algorithm

- In some cases particular CM steps in the ECM algorithm are difficult, perhaps requiring iterative algorithms.
- It may be nearly as simple to carry out a constrained maximization of the actual loglikelihood function based on the observed data, rather than over the Q function. This can be advantageous, since it is more directly tied to the ultimate objective and hence tends to speed the convergence.
- The ECME (Expectation - Conditional Maximization Either) algorithm is a variant of ECM where one or more of the CM steps is a conditional maximization of the actual observed data likelihood rather than a conditional maximization of the Q function (Liu and Rubin 1994).

Example 4 ECME for T Model with Unknown Degrees of Freedom

Consider n observations $x_i \sim t_K(\mu, \Psi, \nu)$, $i = 1, \dots, n$
from the K – variate t distribution with mean μ , scale Ψ and df ν

Before we assumed ν is known

With sufficient data, we can treat ν as an additional parameter
to be estimated -- an example of *adaptive* robust estimation

Augment data as before to $(x_i, q_i(a))$

$$x_i \mid q_i, \theta \sim_{iid} N_K(\mu, |\Psi|^{-a} \times \Psi / q_i(a)),$$

$a = 1 / (\nu + K)$ to optimize speed

E-Step: as before, except that ν is set to current estimate $\nu^{(t)}$

Example 4 CM steps

CM(t+1/2): Compute $\mu^{(t+1)}$, $\Psi^{(t+1)}$ by weighted least squares:

$$\mu^{(t+1)} = \sum_{i=1}^n w_i^{(t+1)} x_i / \sum_{i=1}^n w_i^{(t+1)};$$

$$\Psi^{(t+1)} = \sum_{i=1}^n w_i^{(t+1)} (x_i - \mu^{(t+1)}) (x_i - \mu^{(t+1)})^T / \sum_{i=1}^n w_i^{(t+1)}$$

CM(t+2/2): Compute $\nu^{(t+1)}$ to maximize $Q(\mu^{(t+1)}, \Psi^{(t+1)}, \nu^{(t+1)} | \mu^{(t)}, \Psi^{(t)}, \nu^{(t)})$ wrt ν

In fact this is same as a full M – Step in this case,

since CM(t+1/2) does not involve the current estimate of ν

But since CM(t+2/2) is hard, ECME replaces it by

CM*(t+2/2): $\nu^{(t+1)}$ to maximize $\ell(\mu^{(t+1)}, \Psi^{(t+1)}, \nu | x_{(0)})$ wrt ν

Note that Q
is replaced by
real objective ℓ

**REPLACES HARD CM STEP BY HARD CM STEP
THAT SPEEDS THE ALGORITHM UP!**

CM*(t+2/2) is not that hard since it only optimizes wrt a scalar parameter ν

Summary

- Numerous ways of extending and speeding EM
- For more information see Lange, K. (2004) *Optimization* (Springer)
- I feel that to some extent these algorithms have been superceded by algorithms that yields Bayes posterior distributions
 - Discuss Bayes and multiple imputation next

EM Algorithm

Missing Covariates

Factor Models

Mixed Effect Models

Detection Limits