Expectation - Maximization Algorithm

Art Dempster - Nan Laird - Donald Rubin '77

Applications:

- Missing data: Y_f full data

"Easy" model for $f(Y_f \mid \underset{\sim}{\theta})$

But complex for $f(Y_o \mid \underset{\sim}{\theta}) = \int f(Y_o, Y_u \mid \underset{\sim}{\theta})\, dY_u$

Each step has 2 sub-steps:

Given value $\underset{\sim}{\theta}_t$ from t^{th} step:

1) Expectation step:

Use $l(Y_f | \underset{\sim}{\theta}) = \log f(Y_f | \underset{\sim}{\theta})$

to work out

$$E\left(l(Y_f | \underset{\sim}{\theta}) \,\middle|\, Y_0, \underset{\sim}{\theta}_t\right)$$

free parameter

parameter for E

Note: This is E of a function of $\underset{\sim}{\theta}$ given a value of $\underset{\sim}{\theta}_t$

$$= Q\left(\underset{\sim}{\theta} \,|\, \underset{\sim}{\theta}_t\right)$$

2) Maximize $Q(\underset{\sim}{\theta}, \underset{\sim}{\theta}_t)$

with respect to $\underset{\sim}{\theta}$

Let $\underset{\sim}{\theta}_{t+1} = \underset{\underset{\sim}{\theta}}{\operatorname{argmax}}\, Q(\underset{\sim}{\theta}, \underset{\sim}{\theta}_t)$

Repeat 2 steps until convergence.

1) $|\underset{\sim}{\theta}_{t+1} - \underset{\sim}{\theta}_t|$ small

2) $|Q(\theta_{t+1}, \theta_t) - Q(\theta_t, \theta_t)|$ small

- On many applications, Y_u is not "missing data" but "random unknown parameters" as in multilevel & longitudinal modeling
- On some MLM packages (e.g. nlme) the E-M algorithm is used "under the hood".
- It's a good method for many complex problems.
- It is generally not easy to apply but once you have an algorithm for a kind of problem then many people use it without knowing.

Example:

WARNING: LONG & TEDIOUS BUT INTERESTING (I THINK) AND GOOD PRACTICE

- Suppose you are sampling from

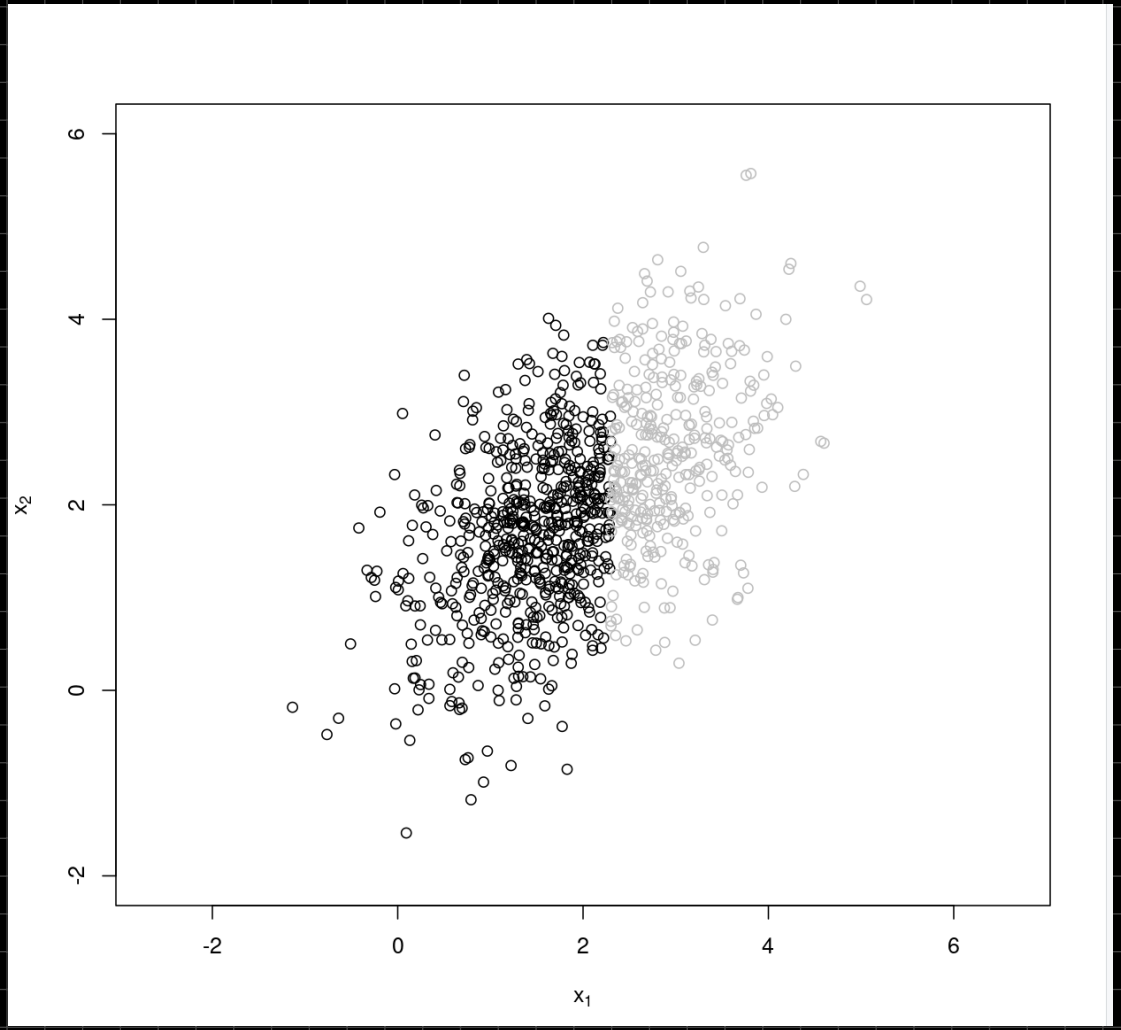
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N_2(\underset{\sim}{\mu}, \Sigma)$$

Simulation using $\underset{\sim}{\mu} = \begin{pmatrix} 2.1 \\ 1.9 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 1.1 & .9 \\ .9 & 1.2 \end{pmatrix}$

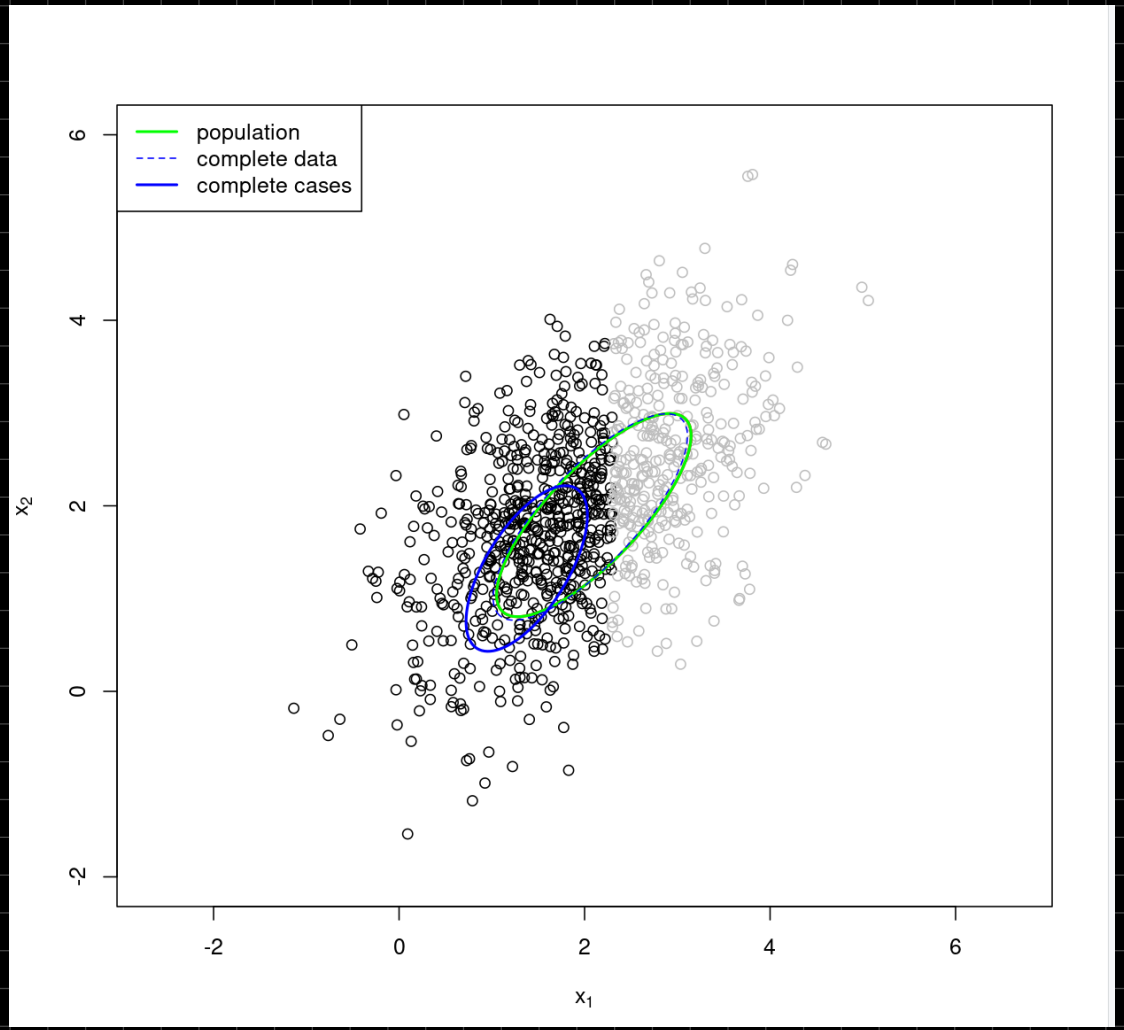
"true" but unknown

$n = 1,000$

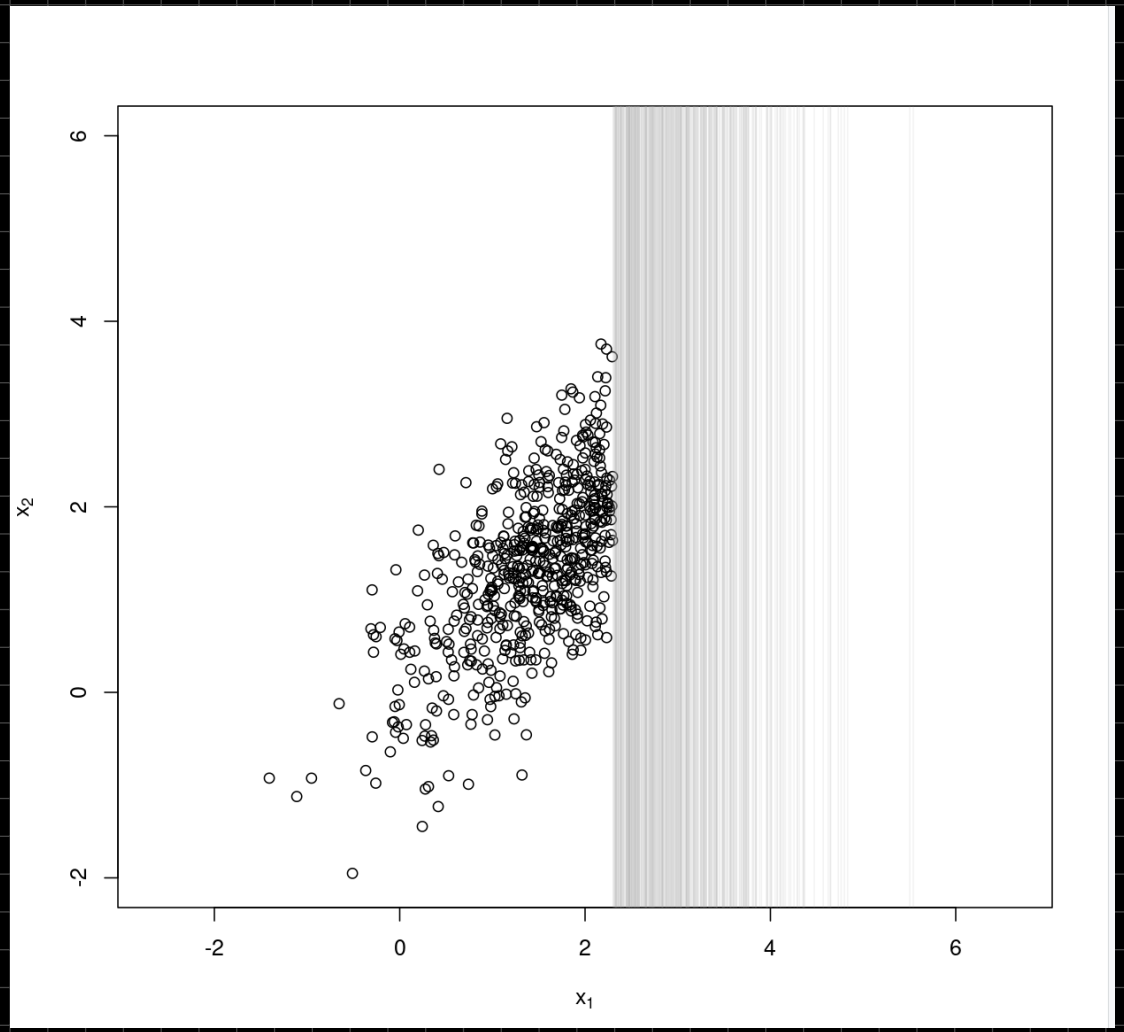
But x_2 not observed if $x_1 > 2.3$.



x_2 missing based on observed value of x_1


population
complete data
complete cases
6
4
2
0
-2
x_2
-2
0
2
4
6
x_1

The data we actually have:



Is there any hope for estimating μ_2?

μ_1 & $\sigma_1^2 = \sigma_{11}$ easy.

What about $\sigma_2^2 = \sigma_{22}$?

and $\rho_{12} = \sigma_{12} / \sqrt{\sigma_{11}\sigma_{22}}$

Data:

$$X = \begin{bmatrix} 2.1 & 3.2 \\ 1.0 & 1.2 \\ \vdots & \vdots \\ 2.2 & 0.9 \\ \hline 3.1 & NA \\ 4.2 & NA \\ \vdots & \vdots \\ 2.4 & NA \end{bmatrix} = \begin{bmatrix} X_c \\ \hline \\ X_m \end{bmatrix}$$

X_c: complete cases

X_m: with missing data

$$= \begin{bmatrix} X_o \\ \\ X_u \end{bmatrix}$$

X_o: observed

X_u: unobserved / missing data or random variable

Since X_u is unknown we think of it as a random vector.

If x_{ij} is missing, we think of x_{ij} as a random variable.

E-M analysis :

Complete data model : (detail to remind some neat methods)

If $\underset{\sim}{X} \sim N_p(\underset{\sim}{\mu}, \Sigma)$

$$f(\underset{\sim}{x} \mid \underset{\sim}{\mu}, \Sigma_i) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu})\right\}$$

Sample of n: $f(\underset{\sim}{x_1}, \underset{\sim}{x_2}, \cdots, \underset{\sim}{x_n} \mid \underset{\sim}{\mu}, \Sigma)$

$$= \frac{1}{(2\pi)^{np/2}} \frac{1}{|\Sigma|^{n/2}} \exp\left\{-\frac{1}{2} \boxed{\sum_{i=1}^{n} (\underset{\sim}{x_i}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x_i}-\underset{\sim}{\mu})}\right\}$$

Playing with $\sum_{i=1}^{n} (\underset{\sim}{x}_i - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu})$

$= \sum_{i=1}^{n} \operatorname{tr}\left[(\underset{\sim}{x}_i - \underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}_i - \underset{\sim}{\mu}) \right]$ (1×1)

$= \sum_{i=1}^{n} \operatorname{tr} \underset{p \times p}{\Sigma^{-1}} \underset{p \times 1}{(\underset{\sim}{x}_i - \underset{\sim}{\mu})} \underset{1 \times p}{(\underset{\sim}{x}_i - \underset{\sim}{\mu})^T}$

$= \operatorname{tr} \Sigma^{-1} \sum_{i=1}^{n} (\underset{\sim}{x}_i - \underset{\sim}{\mu})(\underset{\sim}{x}_i - \underset{\sim}{\mu})^T$ by linearity

$= \operatorname{tr} \Sigma^{-1} \sum_{i=1}^{n} (\underset{\sim}{x}_i \underset{\sim}{x}_i^T - \underset{\sim}{x}_i \underset{\sim}{\mu}^T - \underset{\sim}{\mu} \underset{\sim}{x}_i^T + \underset{\sim}{\mu}\underset{\sim}{\mu}^T)$

TRACE

- tr is sum of diagonal elts
- $\operatorname{tr}(AB) = \operatorname{tr}(BA)$
- $\operatorname{tr}(A) = \operatorname{tr}(A^T)$
- $\operatorname{tr}(A) = a$ if $A = [a]$ is 1×1
- $\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$

$$= \operatorname{tr}\Sigma^{-1}\left(X^T X - n\bar{x}\mu^T - n\mu\bar{x}^T + n\mu\mu^T \right)$$

$$= n\mu^T\Sigma^{-1}\mu + \operatorname{tr}\,\Sigma^{-1}\left(X^T X - \left(\sum_{i=1}^{n} x_i\right)\mu^T - \mu\left(\sum_{i=1}^{n} x_i\right)^T \right)$$

focus on this

split X_c & X_m
complete cases / cases with missing X_2

$$X_c' X_c - \left(\sum_{i \in X_c} x_i\right)\mu^T - \mu\left(\sum_{i \in X_c} x_i\right)^T$$

$$+ X_m' X_m - \sum_{i \in X_m} x_i \mu^T - \mu\left(\sum_{i \in X_m} x_i\right)^T$$

(*)

We need to find $E(* \mid X_{observed}, \mu_t, \Sigma_t)$

Why? Because $\underset{\sim}{l}(X \mid \mu, \Sigma)$ is linear in $(*)$

So $Q(\underset{\sim}{\theta} \mid \underset{\sim}{\theta}_t) = E(l(X \mid \underset{\sim}{\theta}) \mid X_0, \underset{\sim}{\theta}_t)$

$= l(E((*) \mid \underset{\sim}{\theta}) \mid X_0, \underset{\sim}{\theta}_t)$

i.e. $E(l(*)) = l(E(*))$

which works if l is linear in $(*)$
and $(*)$ includes all random variables.

Plan of attack:

note that $\underset{\sim}{\mu}$ is an argument of l
Not a parameter for E

$$X_c' X_c - \left(\sum_{i \in X_c} \underset{\sim}{x}_i\right) \underset{\sim}{\mu}^T - \underset{\sim}{\mu} \left(\sum_{i \in X_c} \underset{\sim}{x}_i\right)^T$$

This part is constant given X_0 ... or is it?

$$+ X_m' X_m - \left[\sum_{i \in X_m} \underset{\sim}{x}_i\right] \underset{\sim}{\mu}^T - \underset{\sim}{\mu} \left[\sum_{i \in X_m} \underset{\sim}{x}_i\right]^T$$

ditto

This part has both constants (x_i's) and random variables (X_2's) given X_0.

We need $E(\quad | X_0, \mu^{(t)}, \Sigma^{(t)})$

Start with:

① $E\left(\sum_{i \in X_m} x_i \mid X_0, \mu^{(t)}, \Sigma^{(t)}\right)$

$$= \sum_{i \in X_m} E\left(x_i \mid X_0, \mu^{(t)}, \Sigma^{(t)}\right)$$

$$= \sum_{i \in X_m} \begin{pmatrix} x_{1i} \\ \mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}}\left(x_{1i} - \mu_1^{(t)}\right) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i \in X_m} x_{1i} \\ n_m \mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}}\left(\sum_{i \in X_m} x_{1i} - n_m \mu_1^{(t)}\right) \end{pmatrix}$$

$$= \left(n_m \mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}} \left(S_{1m} - n_m \mu_1^{(t)} \right) \right)$$

$$S_{1m}$$

② $$E\left(\sum_{i \in X_m} x_{1i} x_{2i} \mid X_0, \mu^{(t)}, \Sigma^{(t)} \right)$$

$$= \sum_{i \in X_m} x_{1i} \, E\left(x_{2i} \mid X_0, \mu^{(t)}, \Sigma^{(t)} \right)$$

$$= \sum_{i \in X_m} x_{1i} \left(\mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}} \left(x_{1i} - \mu_1^{(t)} \right) \right)$$

$$= \mu_2^{(t)} S_{1m} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}} \left(\sum x_{1i}^2 - \mu_1^{(t)} S_{1m} \right)$$

(3) $$E\left(\sum x_{2i}^2 \mid X_0, \mu^{(t)}, \Sigma^{(t)} \right)$$

$$= \sum_{i=1}^{n} \left\{ \left[E\left(x_{2i} \mid X_0, \mu^{(t)}, \Sigma^{(t)} \right) \right]^2 + Var\left(x_{2i} \mid X_0, \mu^{(t)}, \Sigma^{(t)} \right) \right\}$$

$$= \sum_{i=1}^{n} \left\{ \left(\mu_2^{(t)} + \frac{\sigma_{12}^{(t)}}{\sigma_{11}^{(t)}} \left(x_{1i} - \mu_1^{(t)} \right) \right)^2 + \sigma_{22}^{(t)} - \frac{\left(\sigma_{12}^{(t)} \right)^2}{\sigma_{11}^{(t)}} \right\}$$

As we get ready to write an algorithm to do this, we try to identify intermediate quantities to compute.

An important one is the vector

$$\underset{i \in X_m}{E}\left(\underset{\sim}{x}_2 \mid \underset{\sim}{x}_1\right) = \mu_2^t + \frac{\sigma_{12}^t}{\sigma_{11}^t}\left(\underset{\sim}{x}_1 - \mu_1^t\right) = \underset{\sim}{c}_{2\cdot 1}$$

"R notation"

So back to the original

$$X_c' X_c - \left(\sum_{i \in X_c} \tilde{x}_i\right) \tilde{\mu}^T - \tilde{\mu} \left(\sum_{i \in X_c} \tilde{x}_i\right)^T$$

$$+ E\left(X_m' X_m - \sum_{i \in X_m} \tilde{x}_i \tilde{\mu}^T - \tilde{\mu} \left(\sum_{i \in X_m} \tilde{x}_i\right)^T \middle| X_o, \mu^t, \Sigma^t\right)$$

$$= X_c' X_c + S_m - (s_c + s_m) \mu^T - \mu (s_c + s_m)^T$$

where $S_m = E(X_c' X_c \mid X_o, \mu^t, \Sigma^t)$ this should be $X_m' X_m$

$$s_c = \sum_{i \in X_c} \tilde{x}_i \qquad s_m = E\left(\sum_{i \in X_m} \tilde{x}_i\right)$$

$$S'_m = \begin{bmatrix} \sum_{i \in X_m} x_{1i}^2 & \sum_{i \in X_m} x_{1i} e_{2\cdot 1,i} \\ = & \sum_{i \in X_m} e_{2\cdot 1,i}^2 + n_m \left(\sigma_{22}^{(t)} - \frac{(\sigma_{12}^{(t)})^2}{\sigma_{11}^{(t)}} \right) \end{bmatrix}$$

$$\underset{\sim}{S}_m = \sum_{i \in X_m} \begin{pmatrix} x_{1i} \\ \mu_2^t + \frac{\sigma_{12}}{\sigma_{11}}^t (x_{1i} - \mu_1^t) \end{pmatrix}$$

Let $P = X_c^T X_c + S_m$

$$\underset{\sim}{S} = \underset{\sim}{S}_c + \underset{\sim}{S}_m$$

Then $Q(\theta, \theta^t)$ has the form:

$$\ldots \operatorname{tr} \Sigma^{-1}\left(P - \underset{\sim}{S}\underset{\sim}{\mu}^T - \underset{\sim}{\mu}\underset{\sim}{S}^T\right) \ldots$$

In full:

$$Q(\underset{\sim}{\theta}, \underset{\sim}{\theta}^t) = k - \frac{n}{2}\log|\Sigma_1| - \frac{n}{2}\underset{\sim}{\mu}^T\Sigma^{-1}\underset{\sim}{\mu}$$

$$- \frac{1}{2}\operatorname{tr}\Sigma^{-1}\left(P - \underset{\sim}{S}\mu^T - \mu\underset{\sim}{S}^T\right)$$

– This is the log-likelihood for a $N_2(\mu, \Sigma)$ if P is the SSCP matrix and $\underset{\sim}{S}$ is the sum of $\underset{\sim}{x}$'s.

– Which we know is maximized if

$$\underset{\sim}{\mu} = \hat{\underset{\sim}{\mu}} = \frac{1}{n}\underset{\sim}{S}$$

$$\Sigma_1 = \hat{\Sigma} = \left(P - \frac{1}{n}\underset{\sim}{S}\underset{\sim}{S}^T\right)/n$$

So the M-step is easy!

$$\mu^{t+1} = \hat{\mu} = \frac{1}{n}\underset{\sim}{S} \text{ and } \Sigma^{t+1} = \hat{\Sigma}$$