



STK4051/STK9051

Computational Statistics

Riccardo De Bin

`debin@math.uio.no`

Outline of the lecture

- EM Optimization Methods
 - Missing Data, Marginalization, and Notation
 - The EM Algorithm
 - Convergence
 - Usage in Exponential Families
 - Variance Estimation
 - EM Variants
 - Improving the E Step
 - Improving the M Step
 - Acceleration Methods

EM Optimization Methods: Introduction

The **Expectation-Maximization (EM)** algorithm:

- **iterative** optimization strategy;
- motivated by a notion of **missingness**;
- based on consideration of the **conditional distribution** of what is missing given what is observed;
- popular method, 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;
- “missing data” **may not truly be missing**:
 - ▶ conceptual ploy to exploit the EM.

Missing Data, Marginalization, and Notation

Notation:

- $\mathbf{Y} = (\mathbf{X}, \mathbf{Z})$ are the **complete data**:
 - ▶ $\mathbf{X}$ are **observed**;
 - ▶ $\mathbf{Z}$ are **missing**;
- we observe $\mathbf{X} = M(\mathbf{Y})$,
 - ▶ $M(\cdot)$ is a “**many-to-fewer**” mapping;
- density of complete data: $f_Y(\mathbf{y}|\boldsymbol{\theta})$, which is **available**;
- density of observed data: $f_X(\mathbf{x}|\boldsymbol{\theta}) = \int_{\mathbf{y}: M(\mathbf{y})=\mathbf{x}} f_Y(\mathbf{y}|\boldsymbol{\theta}) d\mathbf{y}$;
- **conditional density** of the missing, give the observed
$$f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = f_Y(\mathbf{y}|\boldsymbol{\theta})/f_X(\mathbf{x}|\boldsymbol{\theta});$$

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

The EM Algorithm: Introduction

Main idea: Iterate between

- estimate $\mathbf{Z}$ given $\mathbf{X}, \boldsymbol{\theta}$ (E step)
- estimate $\boldsymbol{\theta}$ given $(\mathbf{X}, \mathbf{Z})$ (M step)

Practically a bit more complicated:

- we **cannot** directly maximize $\log L(\boldsymbol{\theta}|\mathbf{Y})$ as if we had the complete data;
- we need to **approximate** it;
- with $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, the **expectation of the joint log likelihood for the complete data, conditional on the observed data $\mathbf{X} = \mathbf{x}$.**

The EM Algorithm: Introduction

In formula:

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

where $\mathbf{Z}$ is the **only random** part of $\mathbf{Y}$ once we are given $\mathbf{X} = \mathbf{x}$.

The **algorithm** then becomes:

1. **E step**: Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$;
2. **M step**: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ w.r.t. $\boldsymbol{\theta}$ to obtain $\boldsymbol{\theta}^{(t+1)}$;
3. return to the E step unless a stopping criterion has been met.

The EM Algorithm: Trivial example

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

- **complete** data log-likelihood function is
$$\log\{L(\theta|\mathbf{y})\} = \log\{f_{\mathbf{Y}}(\mathbf{y}|\theta)\} = 2\log\{\theta\} - \theta y_1 - \theta y_2;$$
- $Q(\theta|\theta^{(t)}) = 2\log\{\theta\} - 5\theta - \theta/\theta^{(t)},$
 - ▶ since $E\{Y_2|y_1, \theta^{(t)}\} = E\{Y_2|\theta^{(t)}\} = 1/\theta^{(t)}$ as $Y_1 \perp\!\!\!\perp Y_2$;
- the maximizer of $Q(\theta|\theta^{(t)})$ is the root of $2/\theta - 5 - 1/\theta^{(t)} = 0$;
- $\theta^{(t+1)} = 2\theta^{(t)}/(5\theta^{(t)} + 1);$
- the algorithm quickly converges to $\hat{\theta} = 0.2$.

Here:

- **Same** E step and M step at each iteration;
- this example is **not realistic** (easy analytic solution);
- the E step is **trickier** in real applications (need $f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$).

The EM Algorithm: Example (Peppered moths)

Moths are insects of the same order of butterflies;

- interesting story about evolution and pollution;
- wing color determined by a **single gene**;
- with three possible alleles (C, I, and T);
- C is **dominant** to I, and T is **recessive** to I:
 - ▶ TT genotype results in **light-coloured** wings;
 - ▶ CC, CI, and CT result in **black** colouring;
 - ▶ II and IT genotypes produce an **intermediate** case with mottled wings;
- 6 possible genotypes, only **3 observable** phenotypes.

The EM Algorithm: Example (Peppered moths)

How do we estimate **allele frequencies** from phenotype counts?

- Assume frequencies p_C, p_I, p_T , with $p_C + p_I + p_T = 1$;
- Under Hardy-Weinberg, if the allele frequencies are

Colour	Probability
light-coloured (TT)	p_T^2
intermediate (II, IT)	$p_I^2 + 2p_I p_T$
black (CC, CI, CT)	$p_C^2 + 2p_C p_I + 2p_C p_T$

If we had the **complete data** ($n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT}$):

- easy solution, multinomial distribution;
- MLE based on the frequencies;

Instead, we have **observed data** (n_T, n_I, n_C) at phenotype level.

The EM Algorithm: Example (Peppered moths)

- Observed data: $\mathbf{x} = (n_C, n_I, n_T)$;
- Complete data: $\mathbf{y} = (n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}, n_{TT})$
- the **complete data log-likelihood** (multinomial),

$$\begin{aligned} \log\{f_{\mathbf{Y}}(\mathbf{y}|\mathbf{p})\} &= n_{CC} \log\{p_C^2\} + n_{CI} \log\{2p_C p_I\} + n_{CT} \log\{2p_C p_T\} \\ &\quad + n_{II} \log\{p_I^2\} + n_{IT} \log\{2p_I p_T\} + n_{TT} \log\{p_T^2\} \\ &\quad + \log \binom{n}{n_{CC} \quad n_{CI} \quad n_{CT} \quad n_{II} \quad n_{IT} \quad n_{TT}}. \end{aligned}$$

- the **many-to-one mapping** from $\mathbf{Y}$ to $\mathbf{X}$, ($M(\mathbf{Y})$),
 $(n_C, n_I, n_T) = (n_{CC} + n_{CI} + n_{CT}, n_{II} + n_{IT}, n_{TT})$.

The EM Algorithm: Example (Peppered moths)

E Step:

- **conditional** on n_C and a parameter vector $\mathbf{p}^{(t)} = (p_C^{(t)}, p_I^{(t)})$, the latent counts n_{CC} , n_{CI} , and n_{CT} have distribution Multinomial(n_C , $(p_C^{(t)})^2$, $2p_C^{(t)} p_I^{(t)}$, $2p_C^{(t)} p_T^{(t)}$)
- same reasoning for n_{II} and n_{IT} ;
- $n_{TT} = n_T$.

Thus:

$$E\{N_{CC}|n_C, n_I, n_T, \mathbf{p}^{(t)}\} = n_{CC}^{(t)} = \frac{n_C (p_C^{(t)})^2}{(p_C^{(t)})^2 + 2p_C^{(t)} p_I^{(t)} + 2p_C^{(t)} p_T^{(t)}};$$

and so on. Substituting each term,

$$Q(\mathbf{p}|\mathbf{p}^{(t)}) = n_{CC}^{(t)} \log\{p_C^2\} + n_{CI}^{(t)} \log\{2p_C p_I\} + n_{CT}^{(t)} \log\{2p_C p_T\} \\ + n_{II}^{(t)} \log\{p_I^2\} + n_{IT}^{(t)} \log\{2p_I p_T\} + n_{TT} \log\{p_T^2\} + k(n_C, n_I, n_T, \mathbf{p}^{(t)}),$$

where $k(n_C, n_I, n_T, \mathbf{p}^{(t)})$ **does not depend** on p .

The EM Algorithm: Example (Peppered moths)

M Step:

Differentiating with respect to p_C and p_I yields

$$\frac{dQ(\mathbf{p}|\mathbf{p}^{(t)})}{dp_C} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I};$$
$$\frac{dQ(\mathbf{p}|\mathbf{p}^{(t)})}{dp_I} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_I} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{1 - p_C - p_I}.$$

Setting these **derivatives equal to zero** and solving for p_C and p_I ,

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n};$$
$$p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n};$$
$$p_T^{(t+1)} = \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{2n}.$$

The EM Algorithm: Example (Peppered moths)

With observed data $n_C = 85$, $n_I = 196$, and $n_T = 341$,

t	$p_C^{(t)}$	$p_I^{(t)}$	$R^{(t)}$	$D_C^{(t)}$	$D_I^{(t)}$
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

Here:

- $R^{(t)}$ is the **relative convergence criterion**;
- $D_C^{(t)}$, and $D_I^{(t)}$ are **ratios of consecutive errors**.

The EM Algorithm: Convergence

We show two **properties**:

- **Ascent**: Each M step increases the log likelihood;
- **Order of Convergence**: linear (slow!),
 - ▶ with a rate inversely related to the proportion of missing data.

The EM Algorithm: Convergence

To show that each iteration increases the log-likelihood $l(\boldsymbol{\theta}|\mathbf{x})$,

$$f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) = \frac{f_Y(\mathbf{y}|\boldsymbol{\theta})}{f_X(\mathbf{x}|\boldsymbol{\theta})}$$

$$\Downarrow$$

$$\log f_X(\mathbf{x}|\boldsymbol{\theta}) = \log f_Y(\mathbf{y}|\boldsymbol{\theta}) - \log f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$$

$$\Downarrow$$

$$E[\log f_X(\mathbf{x}|\boldsymbol{\theta})] = E[\log f_Y(\mathbf{y}|\boldsymbol{\theta})] - E[\log f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})]$$

where the expectation with respect to $\mathbf{Z} | (\mathbf{x}, \boldsymbol{\theta}^{(t)})$, so

$$\log f_X(\mathbf{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \underbrace{E[\log f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) | \mathbf{x}, \boldsymbol{\theta}^{(t)}]}_{H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}$$

The EM Algorithm: Convergence

We need to prove that

$$\begin{aligned} \log f_X(\mathbf{x}|\boldsymbol{\theta}^{(t+1)}) - \log f_X(\mathbf{x}|\boldsymbol{\theta}^{(t)}) = \\ Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - [H(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})] \geq 0. \end{aligned}$$

Because $\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$,

- $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \geq 0$

The second term is **negative** as $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is max when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$,

$$\begin{aligned} H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ = E \left\{ \log f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) - \log f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) \mid \mathbf{x}, \boldsymbol{\theta}^{(t)} \right\} \end{aligned}$$

$$= \int -\log \left[\frac{f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})}{f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)})} \right] f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}$$

$$\begin{aligned} \text{Jensen's inequality} \\ \geq -\log \left[\int f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}) d\mathbf{z} \right] = 0 \end{aligned}$$

The EM Algorithm: Order of Convergence

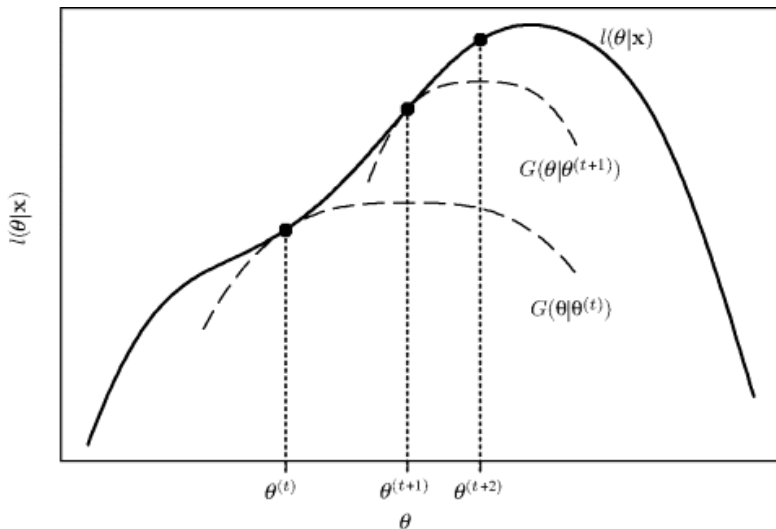
The EM algorithm

- defines a mapping $\boldsymbol{\theta}^{(t+1)} = \Psi(\boldsymbol{\theta}^{(t)})$;
- when it converges, it **converge to a fixed point** $\hat{\boldsymbol{\theta}} = \Psi(\hat{\boldsymbol{\theta}})$;
- **Taylor expansion**:

$$\begin{aligned}\boldsymbol{\varepsilon}^{(t+1)} &\equiv \boldsymbol{\theta}^{(t+1)} - \hat{\boldsymbol{\theta}} \\ &= \Psi(\boldsymbol{\theta}^{(t)}) - \Psi(\hat{\boldsymbol{\theta}}) \\ &\approx \Psi(\boldsymbol{\theta}^{(t)}) - [\Psi(\boldsymbol{\theta}^{(t)}) + \Psi'(\boldsymbol{\theta}^{(t)})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{(t)})] \\ &= \Psi'(\boldsymbol{\theta}^{(t)})(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}}) \\ &= \Psi'(\boldsymbol{\theta}^{(t)})\boldsymbol{\varepsilon}^{(t)}\end{aligned}$$

- convergence order β if $\lim_{t \rightarrow \infty} \frac{|\boldsymbol{\varepsilon}^{(t+1)}|}{|\boldsymbol{\varepsilon}^{(t)}|^\beta} = \rho$
- $p = 1$: $\lim_{t \rightarrow \infty} \frac{|\boldsymbol{\varepsilon}^{(t+1)}|}{|\boldsymbol{\varepsilon}^{(t)}|} = \Psi'(\hat{\boldsymbol{\theta}})$, **linear convergence**
- $p > 1$: **still linear** if $-\ell''(\hat{\boldsymbol{\theta}}|\mathbf{x})$ is positive definite

The EM Algorithm: Convergence



(Givens & Hoeting, 2013, figure 4.1)

The EM Algorithm: Usage in Exponential Families

If the complete data distribution belongs to the **exponential family**,

$$f(\mathbf{y}|\boldsymbol{\theta}) = c_1(\mathbf{y})c_2(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \mathbf{s}(\mathbf{y})\},$$

with $\boldsymbol{\theta}$ the **natural parameters** and $\mathbf{s}(\mathbf{y})$ the **sufficient statistics**.

The **E step** finds

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = k + \log c_2(\boldsymbol{\theta}) + \int \boldsymbol{\theta}^T \mathbf{s}(\mathbf{y}) f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}$$

where k is a quantity that does not depend on $\boldsymbol{\theta}$.

To carry out the **M step**, set the gradient of $dQ(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})/d\boldsymbol{\theta} = 0$,

$$\frac{-c_2'(\boldsymbol{\theta})}{c_2(\boldsymbol{\theta})} = \int \mathbf{s}(\mathbf{y}) f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}.$$

The EM Algorithm: Usage in Exponential Families

Knowing that

$$c_2'(\boldsymbol{\theta}) = -c_2(\boldsymbol{\theta})E\{s(\mathbf{y})|\boldsymbol{\theta}\},$$

the **M step** is completed by setting $\boldsymbol{\theta}^{(t+1)}$ equal to the $\boldsymbol{\theta}$ that solves

$$E\{s(\mathbf{y})|\boldsymbol{\theta}\} = \int s(\mathbf{y})f_{Z|X}(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}^{(t)}) d\mathbf{z}.$$

Aside from replacing $\boldsymbol{\theta}^{(t)}$ with $\boldsymbol{\theta}^{(t+1)}$,

- the form of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is **unchanged** for the next E step;
- the next M step solves the **same** optimization problem.

The EM Algorithm: Usage 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, $\theta^{(t)}$,

$$s^{(t)} = E\{s(\mathbf{Y})|\mathbf{x}, \theta^{(t)}\} = \int s(\mathbf{y}) f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \theta^{(t)}) d\mathbf{z};$$

2. **M step:** Set $\theta^{(t+1)}$ to the value which makes the **unconditional expectation** of the sufficient statistics for the complete data equal to $s^{(t)}$. In other words,

$$\text{find } \theta^{(t+1)} \text{ that solves } E\{s(\mathbf{Y})|\theta\} = s^{(t)};$$

3. Return to the E step unless a **convergence criterion** is met.

The EM Algorithm: Usage in Exponential Families

Consider again the **peppered moths example**:

- the **multinomial distribution** is in the exponential family;
- the sufficient statistics are the (first five) **genotype counts**;
- the natural parameters are the corresponding **log probabilities**.

Consider p_C :

- the **conditional expectations** for the E step are $s_{CC}^{(t)} = n_{CC}^{(t)}$, $s_{CI}^{(t)} = n_{CI}^{(t)}$, $s_{CT}^{(t)} = n_{CT}^{(t)}$;
- the **unconditional expectations** of the sufficient statistics are np_C^2 , $2np_Cp_I$, and $2np_Cp_T$.
- finding p_C s.t. $np_C^2 + 2np_Cp_I + 2np_Cp_T = n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}$ (M step) reduces to the update of slide 12.

EM updates for p_I and p_T are found **analogously**.

The EM Algorithm: Variance Estimation

Among the methods to **estimate the variance**, we only consider

- SEM algorithm
- bootstrapping

as they are **easy to implement** and provide **fast and reliable** results.

Other approaches include Louis's method, empirical information, and numerical differentiation:

- **less common** in practice;
- may be **useful in particular settings**.

The EM Algorithm: Variance Estimation

The missing information principle

Consider negating both side of the second order partial derivatives of $\log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. One obtains

$$-\ell''(\boldsymbol{\theta}|\mathbf{x}) = -\mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}} + \mathbf{H}''(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}}$$

where $\mathbf{Q}''$ and $\mathbf{H}''$ denote derivatives w.r.t. $\boldsymbol{\theta}$.

In other terms,

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

where:

- $\hat{\mathbf{i}}_{\mathbf{X}}(\boldsymbol{\theta}) = -\ell''(\boldsymbol{\theta}|\mathbf{x})$ is the **observed** information;
- $\hat{\mathbf{i}}_{\mathbf{Y}}(\boldsymbol{\theta})$ is the **complete** information;
- $\hat{\mathbf{i}}_{\mathbf{Z}|\mathbf{X}}(\boldsymbol{\theta})$ is the **missing** information.

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

The EM Algorithm: Variance Estimation

Supplemented EM (SEM) algorithm

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

It can be shown that

$$\Psi'(\hat{\theta})^T = \hat{\mathbf{i}}_{Z|X}(\hat{\theta})\hat{\mathbf{i}}_Y(\hat{\theta})^{-1}.$$

Rewriting the missing information principle as

$$\hat{\mathbf{i}}_X(\theta) = \left[I - \hat{\mathbf{i}}_{Z|X}(\theta)\hat{\mathbf{i}}_Y(\theta)^{-1} \right] \hat{\mathbf{i}}_Y(\theta),$$

inverting $\hat{\mathbf{i}}_X(\theta)$ leads to

$$\widehat{\text{var}}\{\hat{\theta}\} = \hat{\mathbf{i}}_Y(\hat{\theta})^{-1} \left(\mathbf{I} + \Psi'(\hat{\theta})^T (\mathbf{I} - \Psi'(\hat{\theta})^T)^{-1} \right).$$

SEM is numerically stable and requires little extra work.

The EM Algorithm: Variance Estimation

SEM algorithm:

1. run the EM algorithm to **convergence**, finding $\hat{\theta}$;
2. **restart** the algorithm from some $\theta^{(0)}$ (better near to $\hat{\theta}$).
For $t = 0, 1, 2, \dots$
 - 2.1 take a **standard E step and M step** to get $\theta^{(t+1)}$ from $\theta^{(t)}$;
 - 2.2 For $j = 1, \dots, p$, define
 - ▶ $\theta^{(t)}(j) = (\hat{\theta}_1, \dots, \hat{\theta}_{j-1}, \theta_j^{(t)}, \hat{\theta}_{j+1}, \dots, \hat{\theta}_p)$;
 - ▶ $r_{ij}^{(t)} = \frac{\Psi_i(\theta^{(t)}(j)) - \hat{\theta}_i}{\theta_j^{(t)} - \hat{\theta}_j}$ for $i = 1, \dots, p$. (Recall that $\Psi(\hat{\theta}) = \hat{\theta}$);

stop when all $r_{ij}^{(t)}$ have converged;
3. The (i, j) th element of $\Psi'(\hat{\theta})$ equals $\lim_{t \rightarrow \infty} r_{ij}^{(t)}$. Plug-in the final estimate of $\Psi'(\hat{\theta})$ to get the **variance**.

The EM Algorithm: Variance Estimation

Bootstrapping:

Basic idea (more on bootstrap in the next lecture):

1. apply EM to the sample $x_1, \dots, x_n$ to find $\hat{\theta}_0$;
2. for $b = 1, \dots, B$:
 - 2.1 sample **with replacement** from $x_1, \dots, x_n$;
 - 2.2 Get $\hat{\theta}_b$ by applying the **same EM** to these pseudo-data;
3. compute the **sample variance** of the B estimates.

Note:

- **basic** bootstrap implementation;
- there is an EM **loop** in every iteration (nested loops);
- may be **computationally burdensome** when the solution of each EM problem is slow.

EM Variants: Improving the E Step

When computing $Q(\theta|\theta^{(t)})$ is difficult, it can be approximated via **Monte Carlo** (more on that in future lectures):

1. **Draw** missing data $\mathbf{Z}_1^{(t)}, \dots, \mathbf{Z}_{m^{(t)}}^{(t)}$ **i.i.d.** from $f_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}, \theta^{(t)})$; each $\mathbf{Z}_j^{(t)}$ is a vector of **all the missing values** needed to complete the observed dataset, and $\mathbf{Y}_j = (\mathbf{x}, \mathbf{Z}_j)$;
2. calculate $\hat{Q}^{(t+1)}(\theta|\theta^{(t)}) = \frac{1}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \log f_{\mathbf{Y}}(\mathbf{Y}_j^{(t)}|\theta)$.

Notes:

- The M step is **modified** to maximize $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$;
- increase $m^{(t)}$ to **reduce the MC variability** of $\hat{Q}$;
- MCEM will not converge in the **same sense** as ordinary EM,
 - ▶ values of $\theta^{(t)}$ will **bounce around the true maximum**;
 - ▶ with a **precision** that depends on $m^{(t)}$.

EM Variants: Improving the M Step

ECM Algorithm: Replace the M step with a series of **simpler conditional maximization (CM)** steps,

- the collection of CM steps after the t -th E step is a **CM cycle**;
- the t -th E step + the t -th CM cycle $\rightarrow$ t -th iteration of ECM;
- let S denote the **total number of CM steps** in each CM cycle.

The t -th iteration of ECM is then:

- for $s = 1, \dots, S$, maximize $Q(\theta|\theta^{(t)})$ **subject to a constraint**,

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

where $\theta^{(t+(s-1)/S)}$ is from the $(s-1)$ -th step;

- set $\theta^{(t+1)} = \theta^{(t+S/S)}$ and **proceed to the E step**.

Note:

- An effective ECM algorithm has **cleverly chosen constraints**.

EM Variants: Improving the M Step

Partition θ into S subvectors, $\theta = (\theta_1, \dots, \theta_S)$. Possible choices for the CM cycle:

- Iterated Conditional Modes / Gauss-Seidel
 - ▶ in the s th CM step, maximize Q w.r.t. θ_s while holding all other components of θ fixed,
 $g_s(\theta) = (\theta_1, \dots, \theta_{s-1}, \theta_{s+1}, \dots, \theta_S)$;
- Book's choice 2:
 - ▶ at the s th CM step, maximize Q with respect to all other components of θ while holding θ_s fixed,
 $g_s(\theta) = \theta_s$;
- additional systems of constraints can be imagined;
- a variant of ECM inserts an E step between each pair of CM steps, thereby updating Q at every stage of the CM cycle.

EM Variants: Improving the M Step

Example: Multivariate regression (Meng & Rubin, 1993)

Let $\mathbf{U}_1, \dots, \mathbf{U}_n$ be independent where $\mathbf{U}_i \sim N_d(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, with

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

EM Variants: Improving the M Step

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 $\beta^{(t)}, \Sigma^{(t)}$,
 - ▶ the sufficient stats are $\sum_{i=1}^n U_{ij}$ for $j = 1, \dots, d$ and $\sum_{i=1}^n U_{ij}U_{ik}$ for $j, k = 1, \dots, d$;
- **CM 1:** $\beta^{(t+1/2)}$ is the **weighted least squares estimate** given $\Sigma = \Sigma^{(t)}$, computed using the sufficient stats from the E step;
- **CM 2:** $\Sigma^{(t+2/2)}$ is the **sample covariance matrix** of the completed data, using $\beta = \beta^{(t+1/2)}$;
- Return to the E step

EM Variants: Improving the M Step

EM gradient algorithm: replace the M step with a **single step of Newton's method**,

- approximate the maximum **without solving** for it exactly,
- in formula,

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

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

EM Variants: Acceleration Methods

Motivation and basic idea:

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

Quasi-Newton acceleration:

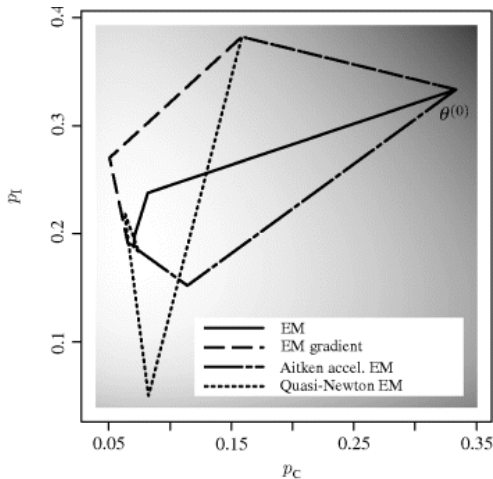
- $\mathbf{M}^{(t)} = \mathbf{Q}''(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} - \mathbf{B}^{(t)}$
 - ▶ $\mathbf{B}^{(t)}$ is crafted to make $\mathbf{M}^{(t)}$ approximate $\mathbf{l}''(\boldsymbol{\theta}^{(t)}|\mathbf{x})$;
 - ▶ potentially better to approximate $(\ell'')^{-1}$ instead of ℓ'' .

EM Variants: Acceleration Methods

Aitken acceleration:

- $\theta^{(t+1)} = \theta^{(t)} - \ell''(\theta^{(t)}|\mathbf{x})^{-1} \hat{\mathbf{i}}_{\mathbf{Y}}(\theta)(\theta_{EM}^{(t+1)} - \theta^{(t)})$
- equivalent to applying the **Newton-Raphson method** to find a zero of $\Psi(\theta) - \theta$,
 - Ψ is the mapping defined by the ordinary EM algorithm;
- potential **numerical instabilities** and **convergence failures**;
- relies on $\ell''(\theta^{(t)}|\mathbf{x})$;
- **quadratic convergence ratio**, but precise **only if close** to $\hat{\theta}$,
- apply only **after initial steps** with ordinary EM.

EM Variants: Acceleration Methods



(Givens & Hoeting, 2013, figure 4.3)

References |

- GIVENS, G. H. & HOETING, J. A. (2013). *Computational Statistics*. Wiley.
- MENG, X.-L. & RUBIN, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika* **80**, 267–278.