G0A63a: Optimization and Numerical Methods

Geert Molenberghs, UHasselt & KULeuven

Francis Tuerlinckx, KULeuven

KATHOLIEKE UNIVERSITEIT LEUVEN

Course notes by Geert Molenberghs,
Geert Verbeke, and Francis Tuerlinckx

Academic year 2011–2012

universiteit hasselt

Chapter 9

Expectation-Maximization Algorithm

- Examples
- Missing data
- E- and M-step
- Acceleration
- Precision estimation
- Examples
- Exercises

9.1 Example: Multinomial Sampling

- A classical example from Dempster, Laird and Rubin (1977)

- Data and model:

Z_{11}	Z_{12}	Z_2	Z_3	Z_4
$\frac{1}{2}$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$

Y_1	Y_2	Y_3	Y_4
$\frac{1}{2}+\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$
125	18	20	34

9.3 Incomplete Data

- Subject i at occasion (time) $j = 1, \ldots, n_i$

- **Measurement** Y_{ij}

- **Dropout indicator**

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$$

- Group Y_{ij} into a vector

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})' = (\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m)$$

$$\begin{cases} \boldsymbol{Y}_i^o & \text{contains } Y_{ij} \text{ for which } R_{ij} = 1, \\ \boldsymbol{Y}_i^m & \text{contains } Y_{ij} \text{ for which } R_{ij} = 0. \end{cases}$$

- Group R_{ij} into a vector $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{in_i})'$

- D_i: time of dropout: $D_i = 1 + \Sigma_{j=1}^{n_i} R_{ij}$

9.4 Factorizing the Distribution

Consider the distribution of the full data:

$$f(Y_i, D_i|\boldsymbol{\theta}, \boldsymbol{\psi})$$

- $\boldsymbol{\theta}$ parametrizes the measurement distribution

- $\boldsymbol{\psi}$ parametrizes the missingness process

- Most models are based on the following factorization:

$$f(Y_i, D_i|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(Y_i|\boldsymbol{\theta}) f(D_i|Y_i, \boldsymbol{\psi})$$

  - ▷ the first factor is the marginal density of the measurement process

  - ▷ the second factor is the density of the missingness process, given the outcomes

9.5 Missing Data Processes Taxonomy

$$f(D_i|\boldsymbol{Y}_i, \boldsymbol{\psi}) = f(D_i|\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{\psi}).$$

- **Missing Completely At Random (MCAR):** Missingness is independent of the measurements:
$$f(D_i|\boldsymbol{\psi}).$$

- **Missing At Random (MAR):** Missingness is independent of the unobserved (missing) measurements, possibly depending on the observed measurements:
$$f(D_i|\boldsymbol{Y}_i^o, \boldsymbol{\psi}).$$

- **Missing Not At Random (MNAR):** Missingness depends on the missing values

Above terminology is independent of the statistical framework chosen to analyse the data

This is to be contrasted with the terms ignorable *and* nonignorable *missing data*

9.6 Ignorability

Let us decide to use likelihood based estimation

The full data likelihood contribution for subject i:

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi} | \boldsymbol{Y}_i, D_i) \propto f(\boldsymbol{Y}_i, D_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

Base inference on the observed data:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} | \boldsymbol{Y}_i, D_i) \propto f(\boldsymbol{Y}_i^o, D_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

with

$$\begin{aligned} f(\boldsymbol{Y}_i^o, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{Y}_i, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\boldsymbol{Y}_i^m \\ &= \int f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{\theta}) f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m, \boldsymbol{\psi}) d\boldsymbol{Y}_i^m. \end{aligned}$$

Under a MAR process:

$$\begin{aligned} f(\boldsymbol{Y}_i^o, D_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\boldsymbol{Y}_i^o, \boldsymbol{Y}_i^m | \boldsymbol{\theta}) f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{\psi}) d\boldsymbol{Y}_i^m \\ &= f(\boldsymbol{Y}_i^o | \boldsymbol{\theta}) f(D_i | \boldsymbol{Y}_i^o, \boldsymbol{\psi}) \end{aligned}$$

The likelihood factorizes into two components

9.8 General Theory of EM

- For subject i is the complete data vector $\boldsymbol{Z}_i$ and the observed data vector $\boldsymbol{Y}_i$

- Examples:
  - ▷ $\boldsymbol{Y}_i$ is an incomplete version of $\boldsymbol{Z}_i$ in a repeated measures study with missing data
  - ▷ $\boldsymbol{Y}_i$ is a partially classified version of $\boldsymbol{Z}_i$, such as in the original example
  - ▷ $\boldsymbol{Y}_i$ is a grouped version of $\boldsymbol{Z}_i$ (e.g., the number of cigarettes smoked per day is reported as number of packs per day)
  - ▷ $\boldsymbol{Z}_i$ is $\boldsymbol{Y}_i$, augmented with random effects and/or latent variables/classes, e.g., group membership

- In general, $\boldsymbol{Y}_i$ is called a coarse version of $\boldsymbol{Z}_i$, and the process leading to $\boldsymbol{Y}_i$ is called the coarsening process

- One of the most common versions is missingness

- Without loss of generality, the arguments are developed for the missing data setting

- The observed log-likelihood can be maximized by iteratively maximizing the complete data log-likelihood, using the EM algorithm

- The condition for the EM algorithm to be valid, in its basic form, is ignorability and hence MAR

9.9 The EM Algorithm

- **Initial step:** Compute starting values

- **E step:** Compute the expected augmented-data log-likelihood; this step often reduces to simple sufficient statistics

- **M step:** Maximize this likelihood

- Iterate until convergence

9.10 The Initial Step

Let $\boldsymbol{\theta}^{(0)}$ be an initial parameter vector, which can be found from e.g.

- A complete case analysis, an available case analysis, a simple method of imputation, when data are incomplete

- An arbitrary split for misclassified data

- A two-stage analysis in a mixed model where the random effects are considered missing

- ...

9.11 The E Step

- Given current values $\boldsymbol{\theta}^{(t)}$ for the parameters, the E step computes the objective function:
$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E\left[\ell(\boldsymbol{\theta}|\boldsymbol{Z})|\boldsymbol{Y},\boldsymbol{\theta}^{(t)}\right]$$

- For incomplete data, this comes down to calculating the expected value of the observed data loglikelihood, given the observed data and the current parameters

- In other words, the functions through which the incomplete data enter the complete data likelihood are replaced by their expectations, given the observed portion of the data and the current version of the parameters

- Precisely, for missing data,
$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \int \ell(\boldsymbol{\theta},\boldsymbol{Y})f(\boldsymbol{Y}^m|\boldsymbol{Y}^o,\boldsymbol{\theta}^{(t)})d\boldsymbol{Y}^m \\ &= E\left[\ell(\boldsymbol{\theta}|\boldsymbol{Y})|\boldsymbol{Y}^o,\boldsymbol{\theta}^{(t)}\right] \end{aligned}$$
i.e. substituting the expected value of $\boldsymbol{Y}^m$, given $\boldsymbol{Y}^o$ and $\boldsymbol{\theta}^{(t)}$

- In case the log-likelihood is linear in sufficient statistics, the E step reduces to calculating the expected values of the sufficient statistics

- This is the case, for example, for the exponential family

- For multinomial data, the sufficient statistics are linear functions of the data. The E step comes down to the expectation of the incomplete data, given the complete ones and the current values of the parameter vector

9.12 The M Step

- The M step determines $\boldsymbol{\theta}^{(t+1)}$, the parameter vector maximizing the objective function (i.e., the current value of the expected complete-data likelihood)

- Formally, $\boldsymbol{\theta}^{(t+1)}$ satisfies

$$Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \qquad \text{for all } \boldsymbol{\theta}$$

- One can, but does not have to, continue until full convergence in the M step

- It can be shown that the likelihood increases at every step

- Since the log-likelihood is bounded from above, convergence is forced to apply

9.13 Example: Multinomial Sampling

- Data and model:

Z_{11}	Z_{12}	Z_2	Z_3	Z_4
$\frac{1}{2}$	$\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$

Y_1	Y_2	Y_3	Y_4
$\frac{1}{2}+\frac{1}{4}\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}\theta$
125	18	20	34

- Three ways of analysis:
  - ▷ Direct likelihood and non-iterative solution
  - ▷ Direct likelihood and iterative solution
  - ▷ EM algorithm

9.13.1 Likelihood for Complete and Observed Data

- Log-likelihood for complete data:

$$\begin{aligned}\ell_c(\theta) &= \sum_{j=1}^{5} \ln[\pi_j^c(\theta)] \\ &= Z_{11}(125;\theta)\ln\left(\frac{1}{2}\right) + Z_{12}(125;\theta)\ln\left(\frac{1}{4}\theta\right) \\ &\quad +18\ln\left(\frac{1}{4}(1-\theta)\right) + 20\ln\left(\frac{1}{4}(1-\theta)\right) \\ &\quad +34\ln\left(\frac{1}{4}\theta\right)\end{aligned}$$

- Log-likelihood for observed data:

$$\begin{aligned}\ell(\theta) &= \sum_{j=1}^{4} \ln[\pi_j(\theta)] \\ &= 125\ln\left(\frac{1}{2}+\frac{1}{4}\theta\right) + 18\ln\left(\frac{1}{4}(1-\theta)\right) \\ &\quad +20\ln\left(\frac{1}{4}(1-\theta)\right) + 34\ln\left(\frac{1}{4}\theta\right)\end{aligned}$$

9.13.2 Non-iterative Solution

- Direct calculation of the first derivative of the observed-data log-likelihood yields:

$$4 \cdot S(\theta) = \frac{y_1}{2+\theta} - \frac{y_2}{1-\theta} - \frac{y_3}{1-\theta} + \frac{y_4}{\theta} = 0$$

- Rewriting this equation produces a quadratic equation:

$$-197 \cdot \theta^2 + 15 \cdot \theta + 68 = 0$$

- There are two solutions

$$0.6268 \qquad \text{and} \qquad -0.5507$$

- Obviously:

$$\boxed{\widehat{\theta} = 0.626821497871}$$

9.13.3 Iterative Solution of the Observed Likelihood

- Define the matrix that connects the observed to the complete data:

$$C = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

- Thus: $\boldsymbol{\pi}(\theta) = C\boldsymbol{\pi}^c(\theta)$

- Write

$$\boldsymbol{\pi}^c(\theta) = \begin{pmatrix} 0.50 \\ 0 \\ 0.25 \\ 0.25 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0.25 \\ -0.25 \\ -0.25 \\ 0.25 \end{pmatrix} \theta = \boldsymbol{X}_0 + \boldsymbol{X}_1\theta$$

- The score function is

$$S(\theta) = \boldsymbol{X}_1' C' (C\text{cov}(\boldsymbol{Z})C')^{-}(\boldsymbol{Y} - nC\boldsymbol{\pi}^c)$$

  and the second derivative is

$$H(\theta) = n\boldsymbol{X}_1' C' (C\text{cov}(\boldsymbol{Z})C')^{-} C\boldsymbol{X}_1$$

  with updating algorithm

$$\theta^{(t+1)} = \theta^{(t)} + S(\theta^{(t)})/H(\theta^{(t)}$$

- Obviously, at maximum, $W(\theta)$ can be used to estimate standard errors

9.13.4 The EM Algorithm

- The likelihood for the complete data is

$$\begin{aligned}
\ell_c(\theta) &= Z_{11}(125;\theta)\ln\left(\frac{1}{2}\right) + Z_{12}(125;\theta)\ln\left(\frac{1}{4}\theta\right) \\
&\quad +18\ln\left(\frac{1}{4}(1-\theta)\right) + 20\ln\left(\frac{1}{4}(1-\theta)\right) \\
&\quad +34\ln\left(\frac{1}{4}\theta\right)
\end{aligned}$$

- Hence, the objective function is

$$\begin{aligned}
Q(\theta|\theta^{(t)}) &= Z_{11}(125;\theta^{(t)})\ln\left(\frac{1}{2}\right) \\
&\quad +Z_{12}(125;\theta^{(t)})\ln\left(\frac{1}{4}\theta\right) \\
&\quad +18\ln\left(\frac{1}{4}(1-\theta)\right) + 20\ln\left(\frac{1}{4}(1-\theta)\right) \\
&\quad +34\ln\left(\frac{1}{4}\theta\right)
\end{aligned}$$

The E Step

The E step requires the calculation of $Z_{11}(125;\theta^{(t)})$ and $Z_{12}(125;\theta^{(t)})$

$$Z_{11}(125;\theta^{(t)}) \;=\; 125 \cdot \frac{2}{2+\theta^{(t)}}$$

$$Z_{12}(125;\theta^{(t)}) \;=\; 125 \cdot \frac{\theta^{(t)}}{2+\theta^{(t)}}$$

The M Step

- The objective function is:

$$4 \cdot S_c(\theta) = \frac{Z_{12}}{\theta} - \frac{Z_2}{1-\theta} - \frac{Z_3}{1-\theta} + \frac{Z_4}{\theta} = 0$$

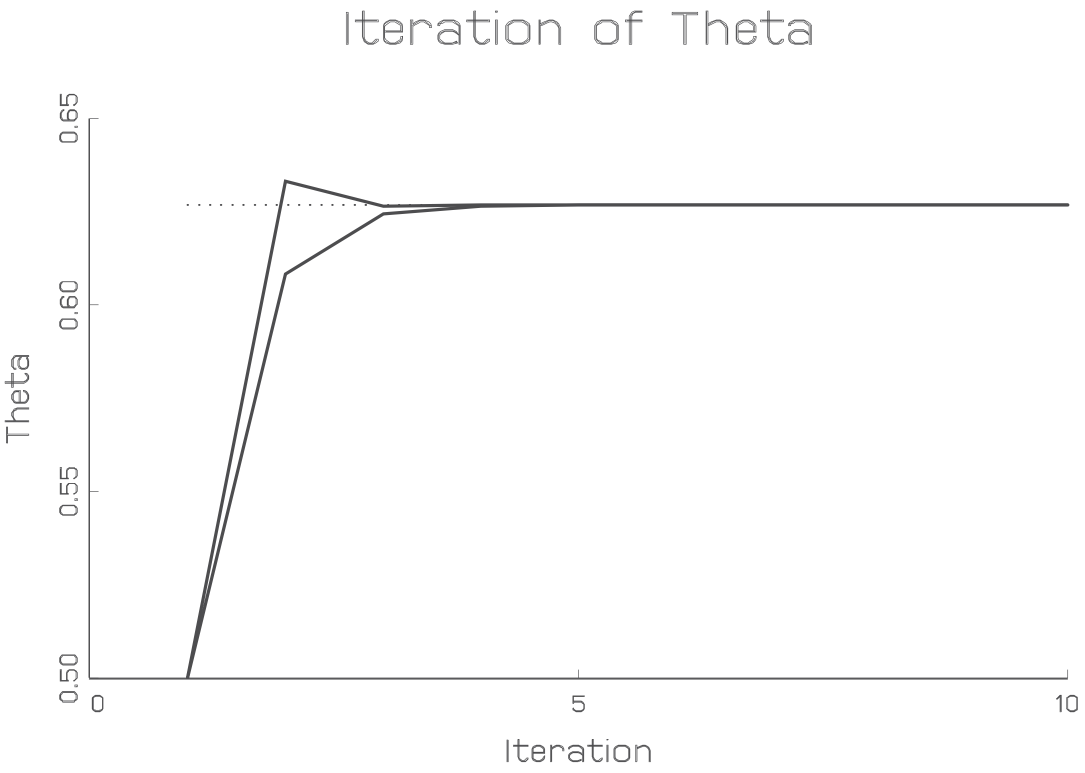
- Rewriting this equations produces a linear equation:

$$Z_{12}^{(t)} + Z_4 = \theta[Z_{12}^{(t)} + Z_2 + Z_3 + Z_4]$$

leading to the solution

$$\theta^{(t+1)} = \frac{Z_{12}^{(t)} + Z_4}{Z_{12}^{(t)} + Z_2 + Z_3 + Z_4}$$

Iteration History for Multinomial Data



	Newton-Raphson		EM	
t	$\theta^{(t)}$	rate	$\theta^{(t)}$	rate
1	0.500000000000	0.0506	0.500000000000	0.1464
2	0.633248730964	0.0447	0.608247422680	0.1346
3	0.626534069270	0.0449	0.624321050369	0.1330
4	0.626834428416	0.0449	0.626488879080	0.1328
5	0.626820916320	0.0449	0.626777322347	0.1327
6	0.626821524027	0.0449	0.626815632110	0.1327
7	0.626821496695	0.0449	0.626820719019	0.1327
8	0.626821497924	0.0453	0.626821394456	0.1327

9.15 Advantages and Disadvantages

Advantages

- The EM algorithm is guaranteed to convergence to a (local) maximum

- Often, the computations in both the E step as well as in the M step are much simpler than in the corresponding direct-likelihood method, such as Newton-Raphson or Fisher scoring

- Solutions are guaranteed to be valid at the augmented data level:
  - ▷ E.g., The complete-data cell probabilities (and hence predicted counts) are non-negative
  - ▷ E.g., variance components in random-effects models are non-negative (one has to reflect carefully upon the issue whether this is really what one wants)

Disadvantages

- Slow convergence (linear or superlinear)
- No automatic provision of precision estimates
- Both of these disadvantages are linked, and can be overcome with the same family of modifications