The 3rd IIA-PennState Astrostatistics School

19-27th July, 2010

EM Algorithm

T.Krishnan

Strand Life Sciences, Bangalore

EM Algorithm known to astronomers as **Richardson-Lucy** Deconvolution or Richardson Lucy Algorithm

E: Expectation;M: Maximization

EM Algorithm

• generic procedure for computing maximum Likelihood estimates (MLE) in awkward problems

• iterative procedure with E and M steps in each iteration cycle

W.H.Richardson (1972): Baysian-based iterative method of image restoration. *Journal of Optical Society of America*, **62**, 55–59.

L.B.Lucy (1974): An iterative technique for the rectification of observed distributions. *Astronomical Journal*, **79**, 745–754

Examples of Astronomy applications

• image restoration

• classification, say of galaxies, gamma-ray bursts (GRB), etc.

Example of Image Restoration

J.Núñez and J.Llacer (1998): Bayesian image reconstruction with space-invariant noise suppression. *Astronomy and Astrophysics Supplement Series*, **131**, 167–180.

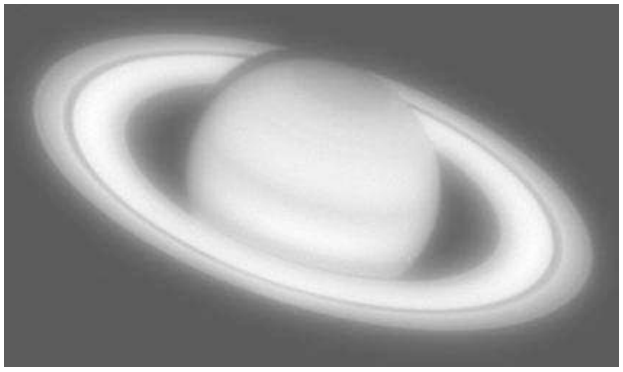


Figure 8: Raw image of planet Saturn obtained with the WF/PC camera of the Hubble Space Telescope.

Figure 1

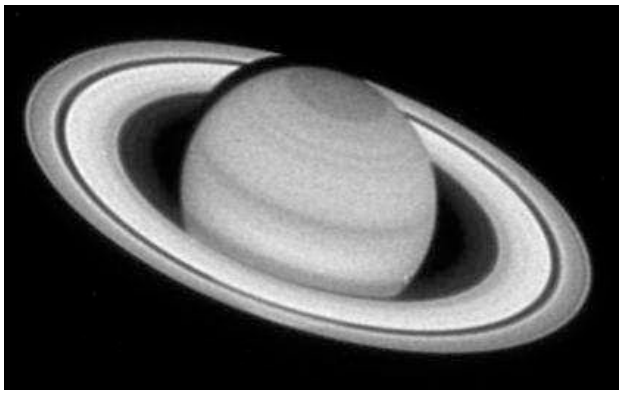


Figure 9: Reconstruction of the image of Saturn using the Richardson-Lucy algorithm.

Figure 2

Example of GRB Classification

L.Hováth, L.G.Balázs, Z.Bagoly, F.Ryde, and A.Mészáros (2006): A new definition of the intermediate group of gamma-ray bursts. *Astronomy & Astrophysics*.

Table 4. Results of the EM algorithm in the $\{\log T_{90}; \log H_{321}\}$ plane. $k = 2 \quad L_{max} = 920$

l	p_l	a_x	a_y	σ_x	σ_y	r
1	0.276	-0.251	0.544	0.531	0.256	0.016
2	0.725	1.479	0.132	0.479	0.287	0.123

Table 5. Results of the EM algorithm. $k = 3 \quad L_{max} = 980$

l	p_l	a_x	a_y	σ_x	σ_y	r
1	0.233	-0.354	0.560	0.486	0.237	0.082
2	0.154	0.722	0.057	0.480	0.432	-0.356
3	0.613	1.588	0.174	0.404	0.249	-0.048

Table 6. Results of the EM algorithm. $k = 4 \quad L_{max} = 982$

l	p_l	a_x	a_y	σ_x	σ_y	r
1	0.234	-0.354	0.559	0.485	0.238	0.078
2	0.148	0.704	0.062	0.447	0.432	-0.335
3	0.333	1.580	0.115	0.403	0.268	-0.141
4	0.284	1.600	0.236	0.400	0.214	0.064

three groups. We see that the mean hardness of

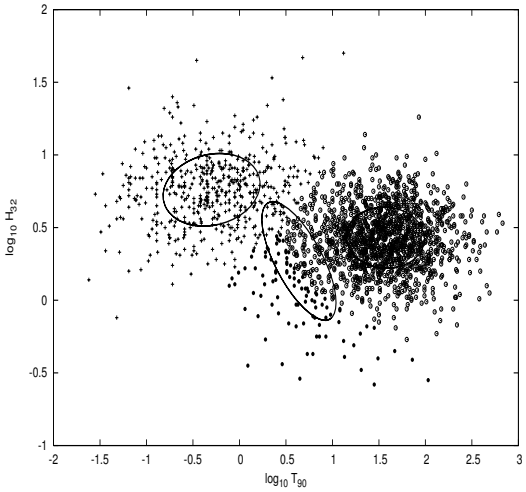


Fig. 1. Distribution of $N = 1956$ GRBs in the $\{\log T_{90}; \log H_{32}\}$ plane. The 1σ ellipses of the three Gaussian distributions are also shown, which were obtained in the ML procedure. The different symbols (crosses, filled circles and open circles) mark bursts belonging to the short, intermediate and long classes, respectively.

Figure 3

A Bit of EM Algorithm History

• EM as a general method of ML estimation introduced by Dempster-Laird-Rubin in 1977

A.P.Demspter, N.M.Laird, and D.B.Rubin (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, **B 39**, 1–38.

• EM is a synthesis of innumerable similar algorithms like Richardson-Lucy

• Shepp and Vardi applied EM to image reconstruction—medical image—Positron Emission Tomography (PET)

L.A.Shepp and Y.Vardi (1982): Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, **1**, 113–122.

• Shepp-Vardi algorithm is identical to algorithms independently obtained by Richardson and Lucy in astronomy context

Linear Inverse Problems

"Incomplete-data problems" form a special case of a more general class of problems called "linear inverse problems".

Linear Inverse Problems with positivity restrictions
statistical estimation problems from incomplete data
solve the equation

$$g(y) = \int_{D_{gc}} h(x,y) g_c(x) dx$$

D_{g_c}, D_g: Domains of the nonnegative real-valued functions g_c and g

Image analysis: g_c true distorted image
g: recorded blurred image
g_c, g: grey-level intensities
function $h(x,y)$, which is assumed to be a bounded nonnegative function on $D_{g_c} \times D_g$: characterizes the blurring mechanism

Examples: image reconstruction in PET/SPECT

traditional statistical estimation problems—grouped and truncated data

About EM

- a method for computing MLE
- useful in many situations where direct maximization methods are tedious or impossible

Examples of these situations

- Missing data
- Incomplete data
- Censored observations
- Difficult distributions
- Unsupervised data
- Blurred images
- .....

Introduction to EM

- EM (Expectation–Maximization) algorithm
- computing maximum likelihood estimates
- "incomplete data problems"—nasty
- "complete data problem"—easier MLE
- "missing values" or "augmented data"
- "statistically tuned" optimization method
- finding the marginal posterior mode

Informal Description of EM

- formulate 'nice' complete-data problem
- write down log-likelihood of complete-data problem
- start with some initial estimates of parameters
- **E-Step**: compute conditional expectation of log-likelihood of complete data problem given actual data, at current parameter values
- **M-Step**: recompute parameter estimates using the simpler MLE for complete data problem
- repeat E- and M-steps until convergence

EXAMPLES OF EM ALGORITHM

- Normal mixtures (Cluster Analysis; Classification)
- Missing data from bivariate normal
- Image Restoration: Tomography
- Hidden Markov models
- Neural Networks
- .....

Image restoration problem same in Astronomy and Medical Imaging

Model for Image Restoration

Vector of emission densities (gray levels) (parameters to be estimated) at n pixels (locations) of true image: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$
Vector of the observations at d positions of device $\boldsymbol{y} = (y_1, \ldots, y_d)^T$
Poisson model for counts
• Given $\boldsymbol{\lambda}$, $y_1, \ldots, y_d$, are conditionally independent Poisson

$$Y_j \sim P(\mu_j), \quad \mu_j = \sum_{i=1}^{n} \lambda_i p_{ij} \quad (j = 1, \ldots, d),$$

• p_{ij}: conditional probability that photon/positron is counted by jth detector given that it was emitted from ith pixel (in **PET**; known detector design parameters);
• p_{ij}: known point spread function (fraction of light from location j observed at position i) (**Image Processing**)

Heuristic Solution for Image Restoration:

Let z_{ij}:number of photons in pixel (i, j) in a two-dimensional image (e.g., CCD)

$$(i = 1, \ldots, n;\ j = 1, \ldots, d)$$

$$Z_{ij} \sim P(\lambda_i p_{ij}) \quad (i = 1, \ldots, d;\ j = 1, \ldots, n).$$

$$y_j = \sum_{i=1}^{n} z_{ij}, \quad (j = 1, \ldots, d),$$

$$\lambda_i p_{ij} / \sum_{h=1}^{n} \lambda_h p_{hj} \qquad (i = 1, \ldots, n;\ j = 1, \ldots, d)$$

is proportion of y_j emitted by i.

If we know $\boldsymbol{\lambda}$, Z_{ij} estimated by

$$y_j \lambda_i p_{ij} / \sum_{h=1}^{n} \lambda_h p_{hj} \qquad (E)$$

Then λ_i is estimated by

$$\sum_{i=1}^{d} z_{ij}$$

Hence the following iteration:

$$\lambda_i^{(k+1)} =$$

$$\lambda_i^{(k)} \sum_{j=1}^{d} \{ y_j p_{ij} / \sum_{h=1}^{n} \lambda_h^{(k)} p_{hj} \}$$

$$(i = 1, \ldots, n)$$

- iteration converges to MLE under Poisson Model
- this is Richardson-Lucy algorithm

Above Heuristic Solution as EM Algorithm

Given problem: an incomplete-data problem
Consider z_{ij} as missing data
consistent with $y_j = \sum_{i=1}^{n} z_{ij}$
Regard this as complete data
Complete-data log-likelihood:

$$\log L_c(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \sum_{j=1}^{d} \{-\lambda_i p_{ij} + z_{ij} \log(\lambda_i p_{ij}) - \log z_{ij}!\}$$

leading to complete-data MLE of λ_i as

$$\frac{\sum_{i=1}^{d} z_{ij}}{\sum_{i=1}^{d} p_{ij}} \qquad (M)$$

We exploit (E) and (M) in an iterative scheme

Let Z_{ij} be the random variable corresponding to observation z_{ij}. Given $\boldsymbol{y}$ and $\boldsymbol{\lambda}^{(k)}$

$$Z_{ij} \sim \text{Binomial}(y_j, \lambda_i^{(k)} p_{ij} / \sum_{h=1}^{n} \lambda_h^{(k)} p_{hj})$$

$$\boldsymbol{E}_{\boldsymbol{\lambda}^{(k)}}(Z_{ij} \mid \boldsymbol{y}) = z_{ij}^{(k)},$$

where

$$z_{ij}^{(k)} = y_j \lambda_i^{(k)} p_{ij} / \sum_{h=1}^{n} \lambda_h^{(k)} p_{hj} \qquad (\mathbf{E-Step})$$

Replace z_{ij} by $z_{ij}^{(k)}$ in (M) on the $(k+1)^{\text{st}}$ iteration **(M-Step)**

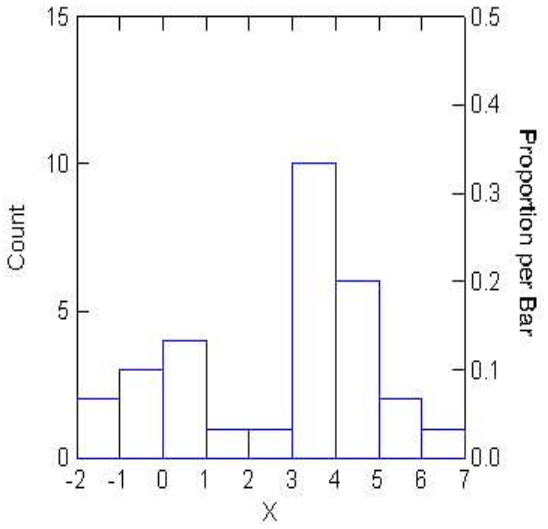
$$\begin{aligned} \lambda_i^{(k+1)} &= q_i^{-1} \sum_{j=1}^{d} p_{ij} E_{\boldsymbol{\lambda}_i^{(k)}}(Z_{ij} \mid \boldsymbol{y}) \\ &= \lambda_i^{(k)} q_i^{-1} \sum_{j=1}^{d} \{ y_j p_{ij} / \sum_{h=1}^{n} \lambda_h^{(k)} p_{hj} \} \end{aligned}$$

(i=1, . . . , n) where $q_i = \sum_{j=1}^{d} p_{ij}$.

Normal Mixtures:

Data:
3.54 3.90 3.93 5.19 3.58 4.60 3.85 4.69 4.29
4.067 3.77 3.45 5.36 2.62 4.80 4.65 3.65 3.67
6.23 3.35 1.58 -0.19 -1.89 0.08 0.34 0.90 -0.03
0.55 -0.57 -1.20

Histogram of 30 observations
Suspected to be from a mixture of two normals
Let us model as a mixture of $\mathcal{N}(0,1), \mathcal{N}(\mu, 4)$
Mixture proportions $1-p, p, 0 < p < 1$
MLE of two parameters p, μ



Mixture Density:

$$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$$

$$\phi(y-\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\mu)^2}$$

$\mathcal{N}(0,1)$ and $\mathcal{N}(\mu,1)$ densities
Mixture of these two normal densities:

$$f(y;p,\mu) = \{p\phi(y-\mu) + (1-p)\phi(y)\}$$

p, μ unknown, $0 < p < 1$
Sample $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ from $f(y;p,\mu)$
To find MLE of p, μ
Mixture resolution;
Unsupervised learning;
Cluster Analysis

Maximizing log-Likelihood:

Likelihood:

$$L_{\boldsymbol{y}}(p,\mu) = \prod_{i=1}^{n}[p\phi(y_i-\mu)+(1-p)\phi(y_i)]$$

To maximize

1. Find $\ell(p,\mu) = \log L_y(p,\mu)$
2. Find $\dot{\ell}(p,\mu) = (\frac{\partial \ell(p,\mu)}{\partial p}, \frac{\partial \ell(p,\mu)}{\partial \mu})$
3. Solve $\dot{\ell}(p,\mu) = 0$
4. Find

$$\ddot{\ell}(p,\mu) = \begin{bmatrix} \frac{\partial^2 \ell(p,\mu)}{\partial p^2} & \frac{\partial^2 \ell(p,\mu)}{\partial p \partial \mu} \\ \frac{\partial^2 \ell(p,\mu)}{\partial p \partial \mu} & \frac{\partial^2 \ell(p,\mu)}{\partial \mu^2} \end{bmatrix} = -\boldsymbol{I}(p,\mu;\boldsymbol{y})$$

called **Observed Information Matrix**
Newton-Raphson: Iterate:

$$(p^{(k+1)},\mu^{(k+1)}) = \boldsymbol{I}^{-1}(p,\mu;\boldsymbol{y})\dot{\ell}(p^{(k)},\mu^{(k)})^T$$

Fisher's Scoring Method: replace

$\boldsymbol{I}$ by $\mathcal{I}(p,\mu) = E(-\boldsymbol{I}(p,\mu;\boldsymbol{y}))$

called the **Expected Information Matrix.**

Both are possible, but messy.

Heuristic Description of EM for this Problem:

- Consider the corresponding supervised estimation problem
- Supervised data identifies group of each case
- If model is correct, one group has mean 0 (group 0) and other group has mean $\mu \neq 0$ (group 1)
- μ is estimated by sample mean of group 1
- p can be estimated by the proportion in group 1
- But we do not have supervised data
- Make an initial guess of parameters, say $\mu = 2, p = 0.75$
- **E-Step**: Using this find prob say π_i of case i from group 1
- This is exactly like posterior prob in discriminant analysis
- **M-Step**: Mean of π_i is an estimate of p for group 1

  Weighted mean of y_i with weights π_i is estimate of μ
- Iterate E-and M-steps until convergence
- Convergence test by say, successive parameter values
- This is EM algorithm

Incomplete and Complete Data:

Two groups:
Group 1 with mean μ (proportion p)
Group 0 with mean 0 (proportion $1-p$)
Pretend for each i, we know the group, say $z_i = 1$ or 0
Supervised Learning Problem (Discriminant Analysis)
$\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$ i.i.d. with

$$P(Z_i = 0) = 1 - p; P(Z_i = 1) = p$$

$$(Y_i|Z_i = 0) \sim \mathcal{N}(0,1), (Y_i|Z_i = 1) \sim \mathcal{N}(\mu, 1)$$

Then $(Z_i, Y_i), i = 1, 2, \ldots, n$ called **Complete Data**
$(Y_i), i = 1, 2, \ldots, n$ called **Incomplete Data**

Complete Data Problem Solution:

Complete Data Likelihood:

$$L_{\boldsymbol{z},\boldsymbol{y}}(p,\mu) = \prod_{i=1}^{n} p^{z_i}\phi(y_i-\mu)^{z_i}(1-p)^{1-z_i}\phi(y_i)^{1-z_i}$$

$$= \text{constant} + p^{\sum z_i}(1-p)^{n-\sum z_i}\prod_{i=1}^{n}\phi(y_i-\mu)^{z_i}$$

$$\ell_{\boldsymbol{z},\boldsymbol{y}}(p,\mu) = \log L_{\boldsymbol{z},\boldsymbol{y}}(p,\mu) = \text{constant}$$

$$+\log p\sum_{i=1}^{n} z_i + \log(1-p)(n-\sum_{i=1}^{n} z_i) - \frac{1}{2}\sum_{i=1}^{n} z_i(y_i-\mu)^2 \qquad (A)$$

$$\dot{\ell} = 0 \Longrightarrow$$

$$\widehat{p} = \frac{\sum_{i=1}^{n} z_i}{n};\ \widehat{\mu} = \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i} \qquad (B)$$

MLE for Complete Data Problem simple

EM exploits this simplicity in an iterative process

E-Step:

For iteration, initial values $p^{(0)}, \mu^{(0)}$
k^{th} iteration values $p^{(k)}, \mu^{(k)}$
Find surrogate for $\ell_{\boldsymbol{z},\boldsymbol{y}}(p, \mu)$ by taking

$$E_{p^{(k)},\mu^{(k)}}(\ell_{\boldsymbol{z},\boldsymbol{y}}(p, \mu)|\boldsymbol{Y} = \boldsymbol{y})$$

$$= \log p \sum_{i=1}^{n} z_i^{(k+1)} + \log(1-p) \sum_{i=1}^{n} (n - z_i^{(k+1)})$$

$$-\frac{1}{2} \sum_{i=1}^{n} z_i^{(k+1)} (y_i - \mu)^2 \qquad (C)$$

where

$$z_i^{(k+1)} = E_{p^{(k)},\mu^{(k)}}(Z_i|Y_i = y_i)$$

$$= P_{p^{(k)},\mu^{(k)}}(Z_i = 1|Y_i = y_i)$$

M-Step:

Equation (C) same form as (A); hence MLE same form as (B)

$$p^{(k+1)} = \frac{\sum_{i=1}^{n} z_i^{(k+1)}}{n}; \mu^{(k+1)} = \frac{\sum_{i=1}^{n} z_i^{(k+1)} y_i}{\sum_{i=1}^{n} z_i^{(k+1)}}$$

$$z_i^{(k+1)} = E(Z_i|Y_i = y_i) = P(Z_i = 1|Y_i = y_i)$$

$$= \frac{p^{(k)}\phi(y_i - \mu^{(k)})}{p^{(k)}\phi(y_i - \mu^{(k)}) + (1 - p^{(k)})\phi(y_i)}$$

which is just the posterior probability (as in Discriminant Analysis)

Iterate E- and M-steps until convergence

Results of EM Algorithm (starting $p = 0.6; \mu = 3.5$):

Iteration	p	μ
0	0.6	3.5
1	0.68	4.1
2	0.67	4.15
3	0.67	4.15

Cluster Analysis using EM algorithm for Normal Mixtures will be discussed in the **Cluster Analysis** lecture.

Example 2: Bivariate Normal Data with Missing Values: Computations

Variate 1:	8	11	16	18	6	4	20	25	9	13
Variate 2:	10	14	16	15	20	4	18	22	?	?

Results of the EM Algorithm for Example 2.1 (Missing Data on One Variate).

Iteration	$\mu_1^{(k)}$	$\mu_2^{(k)}$	$\sigma_{11}^{(k)}$	$\sigma_{12}^{(k)}$	$\sigma_{22}^{(k)}$	$-2\log L(\boldsymbol{\theta}^{(k)})$
1	13	14.8750	40	32.3750	28.8593	1019.64
2	13	14.5528	40	21.2385	24.5787	210.930
3	13	14.5992	40	20.9241	26.2865	193.331
4	13	14.6116	40	20.8931	26.6607	190.550
5	13	14.6144	40	20.8869	26.7355	190.014
6	13	14.6150	40	20.8855	26.7503	189.908
7	13	14.6151	40	20.8852	26.7533	189.886
8	13	14.6152	40	20.8851	26.7538	189.882
9	13	14.6152	40	20.8851	26.7539	189.881
10	13	14.6152	40	20.8851	26.7540	189.881
11	13	14.6152	40	20.8851	26.7540	189.881
12	13	14.6152	40	20.8851	26.7540	189.881
∞	13	14.6152	40	20.8851	26.7540	189.881

THEORY AND METHODOLOGY OF EM

- Incomplete-data problems
- E- and M-steps
- Convergence of EM
- Rate of convergence of EM
- Standard error computation in EM

Incomplete-Data Problems

Incomplete-data problem; incomplete-data likelihood L

Missing or latent or augmented data; missing data (conditional) distribution

Complete-data problem; complete-data likelihood

variety of statistical data models, including mixtures, convolutions, random effects, grouping, censoring, truncated and missing observations

observed data $\boldsymbol{y}$; density $g(\boldsymbol{y}|\boldsymbol{\theta})$; sample space $\mathcal{Y}$; objective is to maximize $\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) = \log(g(\boldsymbol{y}|\boldsymbol{\theta}))$
Complete data $\mathbf{x}$ density $f(\mathbf{x}|\boldsymbol{\theta})$; sample space $\mathcal{X}$

$$g(\boldsymbol{y}|\boldsymbol{\theta}) = \int_{\boldsymbol{y}=\boldsymbol{y}(\mathbf{x})} f(\mathbf{x}|\boldsymbol{\theta})dx$$

$\boldsymbol{S}(\boldsymbol{\theta})$:gradient vector (Fisher score vector)

$\boldsymbol{H}(\boldsymbol{\theta})$: Hessian matrix of $\ell_{\boldsymbol{y}}(\boldsymbol{\theta})$

$I(\boldsymbol{\theta}) = -\boldsymbol{H}(\boldsymbol{\theta})$: observed information matrix

expected value of $I(\boldsymbol{\theta})$= $\mathcal{I}(\boldsymbol{\theta})$: expected information matrix

$\boldsymbol{S}(\boldsymbol{\theta}) = \boldsymbol{O}$: likelihood equations

$-\boldsymbol{H}^{-1}$: estimate of asymptotic covariance matrix

$\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$: estimate of the asymptotic covariance matrix

E- and M-Steps

$$\ell_{\boldsymbol{y}}(\boldsymbol{\theta}) = \log(g(\boldsymbol{y}|\boldsymbol{\theta}))$$

$$\ell_{\mathbf{x}}(\boldsymbol{\theta}) = \log(f(\mathbf{x}|\boldsymbol{\theta}))$$

$$\ell_{\mathbf{x}|\boldsymbol{y}}(\boldsymbol{\theta}) = \log(k(\mathbf{x}|\boldsymbol{y},\boldsymbol{\theta}))$$

$$k(\mathbf{x}|\boldsymbol{y},\boldsymbol{\theta}) = f(\mathbf{x}|\boldsymbol{y},\boldsymbol{\theta})/g(\boldsymbol{y}|\boldsymbol{\theta})$$

$$\ell_{\mathbf{x}}(\boldsymbol{\theta}) = \ell_{\boldsymbol{y}}(\boldsymbol{\theta}) + \ell_{\mathbf{x}|\boldsymbol{y}}(\boldsymbol{\theta})$$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \ell_{\boldsymbol{y}}(\boldsymbol{\theta}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}')$$

E–Step: Compute

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = E(\log(f(\mathbf{x}|\boldsymbol{\theta}))$$

where the expectation is taken with respect to $k(\mathbf{x}|\boldsymbol{y},\boldsymbol{\theta}^{(k)})$

M–Step: Maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ as a function of $\boldsymbol{\theta}$, to obtain $\boldsymbol{\theta}^{(k+1)}$

Appealing properties:

1. It is numerically stable with each EM iteration increasing the likelihood.

2. Under fairly general conditions, it has reliable global convergence properties.

3. It is easily implemented, analytically and computationally.

4. It can be used to provide estimates of 'missing data'.

Drawbacks:

1. It does not provide a natural covariance estimator for the MLE.

2. It is sometimes very slow to converge.

Standard Errors of EM Estimates

1. No natural way to compute covariance matrix
2. Augment EM computation with standard error computation
3. Exploit EM computations
4. Known methods based on observed information matrix, the expected information matrix or on resampling methods

numerically differentiate $\dot{\ell}_{\boldsymbol{y}}$ to obtain the Hessian. In a EM-aided differentiation approach, Meilijson suggests perturbation of the incomplete-data score vector to compute the observed information matrix.

Meng and Rubin: **Supplemented EM** (**SEM**) algorithm numerical techniques are used to compute the derivative of the EM operator $\boldsymbol{M}$ and using this together with the complete-data observed information matrix in the equation

$$\boldsymbol{H} = \ddot{Q}(\boldsymbol{I} - \dot{M})$$

the incomplete-data observed information matrix is computed.

Jamshidian and Jennrich: approximately obtains observed information matrix by numerical differentiation and suggest various alternatives to the SEM algorithm

Oakes' formula

$$\frac{\partial^2 \ell_{\mathbf{x}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^2} = \{\frac{\partial^2 Q(\boldsymbol{\theta}'|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'^2} + \frac{\partial^2 Q(\boldsymbol{\theta}'|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}' \partial \boldsymbol{\theta}}\}_{\boldsymbol{\theta}'=\boldsymbol{\theta}},$$

which is valid for all $\boldsymbol{\theta}'$. By evaluating the right-hand side at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we get the observed information matrix.

Other Aspects of EM

- Acceleration methods
- Monte Carlo versions
- To compute Bayesian Posterior mode
- Connections to MCMC

More References:

Lange, K. (1999): *Numerical Analysis for Statisticians*. New York: Springer-Verlag.

McLachlan, G.J., and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Second Edition. New York: John Wiley & Sons.