Summer School in Statistics for
Astronomers & Physicists
June 15–17, 2005

Center for Astrostatistics

Pennsylvania State University

**Computational algorithms for astrostatistics:**

# EM Algorithm

Thriyambakam Krishnan
Systat Software Asia–Pacific Limited
Bangalore, India

# Introduction to EM

- EM (Expectation–Maximization) algorithm
- computing maximum likelihood estimates
- "incomplete data problems"—nasty (examples)
- "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 the 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

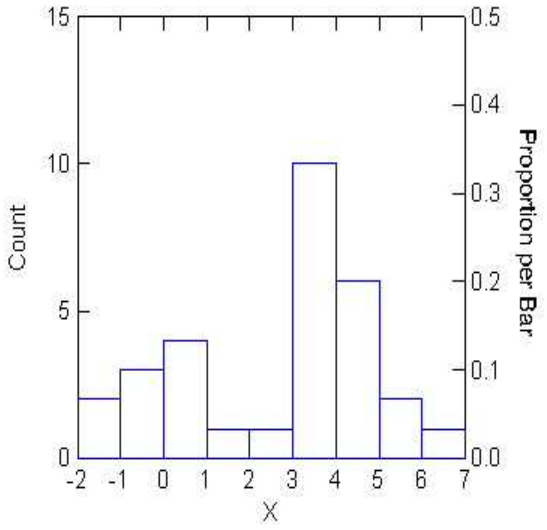
# EXAMPLES OF EM ALGORITHM

- Normal mixtures
- Missing data from bivariate normal
- Image Restoration: Tomography

# **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 $p, \mu$



**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:

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

$p, \mu$ unknown, $0 < p < 1$
Sample $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ from $f(y;p,\mu)$
To find MLE of $p, \mu$
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)
- $\mu$ 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 $\pi_i$ of case $i$ from group 1
- This is exactly like posterior prob in discriminant analysis
- **M-Step**: Mean of $\pi_i$ is an estimate of $p$ for group 1

  Weighted mean of $y_i$ with weights $\pi_i$ is estimate of $\mu$
- 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 $\mu$ (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 $p^{(0)}, \mu^{(0)}$
$k^{\text{th}}$ 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 | $p$ | $\mu$ |
|---|---|---|
| 0 | 0.6 | 3.5 |
| 1 | 0.68 | 4.1 |
| 2 | 0.67 | 4.15 |
| 3 | 0.67 | 4.15 |

# 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 |
| $\infty$ | 13 | 14.6152 | 40 | 20.8851 | 26.7540 | 189.881 |

## Example 2: Bivariate Normal Data with Missing Values: Continued: Problem Formulation

$\boldsymbol{W} = (W_1, W_2)^T,\ \boldsymbol{W} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$\boldsymbol{\Psi} = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T$$

Data on the $i$th variate $W_i$ missing in $m_i$ of the units $(i = 1, 2)$
$\boldsymbol{w}_j = (w_{1j}, w_{2j})^T\,(j = 1, \ldots, m)$: fully observed data points $m$=$n - m_1 - m_2$
$\boldsymbol{w}_{2j}\,(j = m+1, \ldots,\ m+m_1)$: $m_1$ observations with the values of the first variate $w_{1j}$ missing
$\boldsymbol{w}_{1j}\,(j = m + m_1 + 1, \ldots, n)$: $m_2$ observations with the values of the second variate $w_{2j}$ missing

$$W_i \sim N(\mu_i, \sigma_{ii}), i = 1, 2$$

Observed data:

$$\boldsymbol{y} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_m^T, \boldsymbol{v}^T)^T$$

$\boldsymbol{v} = (w_{2,m+1}, \ldots, w_{2,m+m_1}, w_{1,m+m_1+1}, \ldots, w_{1,n})^T$

$$\log L(\boldsymbol{\Psi}) = -n \log(2\pi)$$

$$-\tfrac{1}{2} m \log \mid \boldsymbol{\Sigma} \mid -\tfrac{1}{2} \sum_{j=1}^{m} (\boldsymbol{w}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}_j - \boldsymbol{\mu})$$

$$-\tfrac{1}{2} \sum_{i=1}^{2} m_i \log \sigma_{ii}$$

$$-\tfrac{1}{2} \{\sigma_{11}^{-1} \sum_{j=m+m_1+1}^{n} (w_{1j} - \mu_1)^2 + \sigma_{22}^{-1} \sum_{j=m+1}^{m+m_1} (w_{2j} - \mu_2)^2\}$$

Complete data:

$$\boldsymbol{x} = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_n^T)^T,$$

Missing-data vector $\boldsymbol{z}$ is

$$\boldsymbol{z} = (w_{1,m+1}, \ldots, w_{1,m+m_1}, w_{2,m+m_1+1}, \ldots, w_{2,n})^T$$

$$\log L_c(\boldsymbol{\Psi}) = -n \log(2\pi) - \tfrac{1}{2} n \log \mid \boldsymbol{\Sigma} \mid$$

$$-\tfrac{1}{2} \sum_{j=1}^{n} (\boldsymbol{w}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}_j - \boldsymbol{\mu})$$

$$= -n \log(2\pi) - \tfrac{1}{2} n \log \xi$$

$$-\tfrac{1}{2}\xi^{-1}[\sigma_{22}T_{11} + \sigma_{11}T_{22} - 2\sigma_{12}T_{12}$$

$$-2\{T_1(\mu_1\sigma_{22} - \mu_2\sigma_{12}) + T_2(\mu_2\sigma_{11} - \mu_1\sigma_{12})\}$$

$$+n(\mu_1^2\sigma_{22} + \mu_2^2\sigma_{11} - 2\mu_1\mu_2\sigma_{12})]$$

$$T_i = \sum_{j=1}^{n} w_{ij} \quad (i = 1, 2)$$

$$T_{hi} = \sum_{j=1}^{n} w_{hj} w_{ij} \quad (h, i = 1, 2)$$

$$\xi = \sigma_{11}\sigma_{22}(1 - \rho^2)$$

$$\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{\frac{1}{2}}$$

$$\boldsymbol{T} = (T_1, T_2, T_{11}, T_{12}, T_{22})^T$$

$$\hat{\bar{\mu}}_i = T_i/n, \ (i = 1, 2)$$

$$\hat{\bar{\sigma}}_{hi} = (T_{hi} - n^{-1} T_h T_i)/n \quad (h, i = 1, 2)$$

$$Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) = E_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi}) \mid \boldsymbol{y}\}$$

$$E_{\boldsymbol{\Psi}^{(k)}}(W_{1j} \mid w_{2j})$$

$$E_{\boldsymbol{\Psi}^{(k)}}(W_{1j}^2 \mid w_{2j})$$

$$j = m+1, \ldots, m+m_1$$

$$E_{\boldsymbol{\Psi}^{(k)}}(W_{2j} \mid w_{1j})$$

$$E_{\boldsymbol{\Psi}^{(k)}}(W_{2j}^2 \mid w_{1j})$$

$$j = m+m_1+1, \ldots, n$$

$$E_{\boldsymbol{\Psi}^{(k)}}(W_{2j} \mid w_{1j}) = w_{2j}^{(k)}$$

$$w_{2j}^{(k)} = \mu_2^{(k)} + (\sigma_{12}^{(k)}/\sigma_{11}^{(k)})(w_{1j} - \mu_1^{(k)})$$

$$E_{\mathbf{\Psi}^{(k)}}(W_{2j}^2 \mid w_{1j}) = w_{2j}^{(k)^2} + \sigma_{22.1}^{(k)}$$

$j = m + m_1 + 1, \ldots, n$

Similarly $E_{\mathbf{\Psi}^{(k)}}(W_{1j} \mid w_{2j})$; $E_{\mathbf{\Psi}^{(k)}}(W_{1j}^2 \mid w_{2j})$

$$\mu_i^{(k+1)} = T_i^{(k)}/n \quad (i = 1, 2)$$

$$\sigma_{hi}^{(k+1)} = (T_{hi}^{(k)} - n^{-1} T_h^{(k)} T_i^{(k)})/n, \ (h, i = 1, 2)$$

## Example 3: Image Restoration: Tomography

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 $g$

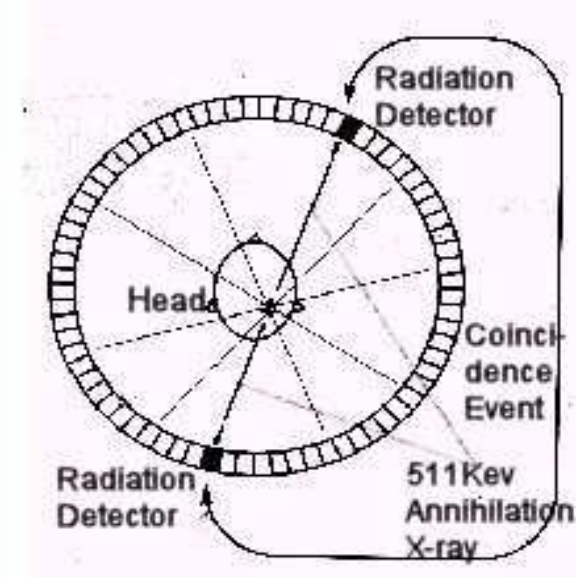
Image analysis: $g_c$ true distorted image
$g$: recorded blurred image
$g_c$, $g$: grey-level intensities
function $h(x,y)$, which is assumed to be a bounded nonnegative function on $D_{g_c} \times D_g$: characterizes the blurring mechanism

Examples: image reconstruction in PET/SPECT
traditional statistical estimation problems—grouped and truncated data

# Introduction to Positron Emission Tomography

In both PET (Positron Emission Tomography) and SPECT (Single-Photon Emission Computerized Tomography) a radioactive tracer is introduced into the organ under study of the human or animal patient. The radioisotype is incorporated into a molecule that is absorbed into the tissue of the organ and resides there in concentrations that indicate levels of metabolic activity and blood flow. As the isotope decays, it emits either single photons (SPECT) or positrons (PET), which are counted by bands of gamma detectors strategically placed around the patient's body. The method of collection of the data counts differs quite considerably for PET and SPECT because of the fundamental differences in their underlying physical processes. SPECT isotopes are gamma emitters that tend to have long half lives. PET isotopes emit positrons which annihilate with nearby electrons in the tissue to generate pairs of photons that fly off on paths almost $180^o$ apart. In both imaging modalities, not all emissions are counted because the photons may travel along directions that do not cross the detectors or because they may be attenuated by the body's tissues.

With PET and SPECT, the aim is to estimate the spatial distribution of the isotope concentration in the organ on the basis of the projected counts recorded at the detectors. In a statistical framework, it is assumed that the emissions occur according to a spatial Poisson point process in the region under study with an unknown intensity function, which is usually referred to as the emission density. In order to estimate the latter, the process is discretized as follows. The space over which the reconstruction is required is finely divided into a number $n$ of rectangular pixels (or voxels in three-dimensions), and it is assumed that the (unknown) emission density is a constant $\lambda_i$ for the $i^{\text{th}}$ pixel ($i = 1, \ldots, n$). Let $y_j$ denote the number of counts recorded by the $j^{\text{th}}$ detector ($j = 1, \ldots, d$), where $d$ denotes the number of detectors. As it is common to work with arrays of $128 \times 128$ or $256 \times 256$ square pixels, the detectors move within the plane of measurement during a scan in order to record more observations than parameters. They count for only a short time at each position.


Radiation
Detector
Head
Coinci-
dence
Event
Radiation
Detector
511Kev
Annihilation
X-ray

## Analysis of PET Data

PET: positron emission tomography
radioactive tracer introduced into the organ
number $n$ of rectangular pixels (or voxels)
(unknown) emission density: $\lambda_i$ for $i^{\text{th}}$ pixel $(i = 1, \ldots, n)$
$y_j$:number of counts recorded by $j^{\text{th}}$ detector $(j = 1, \ldots, d)$
reconstruction aims to infer

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)^T$$

from $\boldsymbol{y} = (y_1, \ldots, y_d)^T$

Poisson regression model for counts

Given $\boldsymbol{\lambda}$, $y_1, \ldots, y_d$, are conditionally independent according to a Poisson distribution

$$Y_j \sim P(\mu_j)$$

$\mu_j$ of $Y_j$ is modeled as

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

$p_{ij}$: conditional probability that a photon/positron is counted by the $j^{\text{th}}$ detector given that it was emitted from within the $i^{\text{th}}$ pixel

Complete-data vector:$\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})^T$
vector $\boldsymbol{z}$: unobservable data counts $z_{ij}$
$z_{ij}$: number of photons/positrons emitted within pixel $i$ and recorded at the $j^{\text{th}}$ detector ($i = 1, \ldots, n$; $j = 1, \ldots, d$)
Assumption: given $\boldsymbol{\lambda}$, $\{Z_{ij}\}$ conditionally independent

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

Since

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

these assumptions for $\{z_{ij}\}$ imply incomplete data model

Complete-data log likelihood:

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

E-Step: $Z_{ij}|\boldsymbol{y}, \boldsymbol{\lambda}^{(k)}$ ∼Binomial with parameters $y_j$ and probability

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

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

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

M-step:

$$\lambda_i^{(k+1)} = q_i^{-1} \sum_{j=1}^{d} p_{ij} E_{\boldsymbol{\lambda}_i^{(k)}}(Z_{ij} \mid \boldsymbol{y})$$

$$= \lambda_i^{(k)} q_i^{-1} \sum_{j=1}^{d} \{y_j p_{ij} / \sum_{h=1}^{n} \lambda_h^{(k)} p_{hj}\} \quad (i = 1, \ldots, n)$$

$$q_i = \sum_{j=1}^{d} p_{ij}$$

probability that an emission from the $i^{\text{th}}$ pixel is recorded by one of the $d$ detectors $(i = 1, \ldots, n)$

# 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 $\boldsymbol{x}$ density $f(\boldsymbol{x}|\boldsymbol{\theta})$; sample space $\mathcal{X}$

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

$\boldsymbol{S}(\boldsymbol{\theta})$: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

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

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

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

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

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

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

**E–Step**: Compute

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

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

**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

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

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

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

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

## Other Aspects of EM

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

**References**:

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

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

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

McLachlan, G.J., and Krishnan, T. (1997). *The EM Algorithm and Extensions*. New York: John Wiley & Sons.