

22S:166 Computing in Statistics
Introduction to the EM Algorithm
Lecture 11
September 24, 2006

- Brief review of
 - maximum likelihood estimation
 - sufficient statistics
 - maximum likelihood estimation in exponential families
 - expectation and linearity
 - asymptotic variances of MLE's / asymptotic normal approximations to posterior distributions
 - Bayes' theorem for normal mean, variance known

1

2

References on EM
(“Expectation-Maximization”) algorithm

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

Louis, T.A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society B*, 44, 98-130.

Meng, X. and Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80, 267-278.

Meng, X. and Rubin, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899-909.

Tanner, M.A. (1993) *Tools for Statistical Inference, 2nd ed*, Ch. 4. Springer-Verlag.

Wei, G. C. G. and Tanner, M.A. (1990). A Monte Carlo

implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699-704.

3

4

Basic EM framework for missing-data problems

- EM is useful when maximum likelihood calculations would be easy IF we knew some things that we don't know

- We have a model for complete data X with associated p.d.f.

$$f(X|\theta)$$

with unknown parameter θ (probably a vector)

- But we don't observe all of X ; i.e.

$$X = (X_{obs}, X_{mis})$$

- we want to maximize the observed-data likelihood with respect to θ

$$L(\theta|X_{obs}) = \int f(X_{obs}, X_{mis}|\theta) dX_{mis}$$

- * (requires assumption that data are missing at random – that is, probability that a value is missing doesn't depend on the value that would have been observed)

- or if estimation would be easy if we knew values of some nuisance parameters

$$\begin{aligned} L(\theta|X) &= \int f(X, \alpha|\theta) d\alpha \\ &= \int f(X|\alpha, \theta) f(\alpha|\theta) d\alpha \end{aligned}$$

5

6

Over-simplified example of EM: random sample from normal distribution with missing values

- Suppose

- $x_i, i = 1, \dots, n$ are a random sample (i.i.d.) from $N(\mu, \sigma^2)$
- $x_i, i = 1, \dots, m$ are observed, $i = m + 1, \dots, n$ missing
- we want MLE of $\theta = (\mu, \sigma^2)$

- Begin by choosing an initial guess $(\mu^{(0)}, \sigma^{2(0)})$

- E-step for iteration k : compute expectations of sufficient statistics, conditional on observed data and current estimate of parameter values

$$E\left(\sum_i^n x_i | \theta^{(k-1)}, X_{obs}\right) = \sum_{i=1}^m x_i + (n-m)\mu^{(k-1)}$$

$$E\left(\sum_i^n x_i^2 | \theta^{(k-1)}, X_{obs}\right) = \sum_{i=1}^m x_i^2 + (n-m)\left[(\mu^{(k-1)})^2 + (\sigma^{(k-1)})^2\right]$$

7

- key idea of EM:

- compute the expectation of the complete-data log-likelihood, conditional on the current estimate of the parameters

* by “filling in” the functions of X_{mis} that appear linearly in the complete-data log-likelihood

- maximize the resulting log-likelihood to obtain the next estimate of the parameters

- iterate

- EM for exponential families

- compute the expectations of sufficient statistics, conditional on the current estimate of the parameters

- use resulting estimates of sufficient statistics to re-estimate the parameters

- iterate

- M-step for iteration k : maximize the resulting log-likelihood in the usual way

$$\begin{aligned} \mu^{(k)} &= E\left(\sum_i^n x_i | \theta^{(k-1)}, X_{obs}\right) / n \\ (\sigma^{(k)})^2 &= E\left(\sum_i^n x_i^2 | \theta^{(k-1)}, X_{obs}\right) / n - (\mu^{(k)})^2 \end{aligned}$$

8

EM algorithm converges more slowly when the proportion of missing data is higher.

```
[1] 10.897372 1.756017 2.480048 4.424937 3.967961
```

```
[1] "5 observations, 4 observed, 1 missing "
[1] "MLEs: muhat = 4.89, sigmasqhat = 12.98 "
```

```
[1] "Starting values mu(0) = 6, sigmasq(0) = 15 "
```

	ex	exsq	mu	sigmasq
[1,]	25.55838	198.5670	5.111675	13.58418
[2,]	24.67005	187.2804	4.934010	13.11163
[3,]	24.49239	185.0231	4.898477	13.00955
[4,]	24.45685	184.5717	4.891370	12.98883
[5,]	24.44975	184.4814	4.889949	12.98467
[6,]	24.44832	184.4633	4.889665	12.98384
[7,]	24.44804	184.4597	4.889608	12.98367
[8,]	24.44798	184.4590	4.889597	12.98364
[9,]	24.44797	184.4588	4.889594	12.98363
[10,]	24.44797	184.4588	4.889594	12.98363
[11,]	24.44797	184.4588	4.889594	12.98363
[12,]	24.44797	184.4588	4.889594	12.98363
[13,]	24.44797	184.4588	4.889594	12.98363
[14,]	24.44797	184.4588	4.889594	12.98363

More general framework

- EM can be used to
 - maximize likelihoods (frequentist)
 - find modes of posterior distributions (Bayesian)
 - maximize marginal likelihoods (empirical Bayes)

```
[1] 10.897372 1.756017 2.480048 4.424937 3.967961
```

```
[1] "5 observations, 2 observed, 3 missing "
[1] "MLEs: muhat = 6.33, sigmasqhat = 20.89 "
```

```
[1] "Starting values mu(0) = 7.5, sigmasq(0) = 23 "
```

	ex	exsq	mu	sigmasq
[1,]	35.15339	359.5863	7.030678	22.48683
[2,]	33.74542	337.5881	6.749085	21.96748
[3,]	32.90064	324.3892	6.580129	21.57974
[4,]	32.39378	316.4698	6.478755	21.31970
[5,]	32.08966	311.7182	6.417931	21.15381
[6,]	31.90718	308.8673	6.381437	21.05072
[7,]	31.79770	307.1567	6.359540	20.98759
[8,]	31.73201	306.1303	6.346402	20.94925
[9,]	31.69260	305.5145	6.338519	20.92608
[10,]	31.66895	305.1450	6.333789	20.91212
[11,]	31.65476	304.9233	6.330952	20.90372
[12,]	31.64624	304.7903	6.329249	20.89867
[13,]	31.64114	304.7105	6.328227	20.89564
[14,]	31.63807	304.6626	6.327614	20.89382

Example: find posterior mode of normal mean parameter in Bayesian model with normal data and both μ and σ^2 unknown

- from Gelman, Carlin, Stern, and Rubin (section 12.3)
- assume data y_1, y_2, \dots, y_n are independent, identically distributed draws from a normal distribution with unknown population mean μ and unknown population variance σ^2
- need joint prior on both unknown parameters.
- intuitive procedure for specifying a joint prior distribution $p(\mu, \sigma^2)$ if we had prior information on both is:

- assume *a priori* independence
- place an inverse gamma prior on σ^2

$$f(\sigma^2 | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\sigma^2)^{\alpha+1}} \exp\left(-\frac{\beta}{\sigma^2}\right)$$

- Place a normal prior on μ

$$f(\mu | \mu_0, \tau_0^2) = \frac{1}{\sqrt{2\pi}\tau_0} \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right)$$

- Then the joint prior is the product of these two priors

- However, it is *not* a conjugate prior!
- In fact, the marginal posterior distribution $p(\mu|\mathbf{y})$ has no simple conjugate forms.

Example continued

- one "noninformative" prior for σ^2 has $\alpha = 1/2$ and $\beta = 0$; this is equivalent to uniform on $\log \sigma$
- The joint posterior is

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma} \times \exp\left(-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right) \times \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right)$$

- so, ignoring the terms that do not depend on μ or σ^2 , the joint log posterior is as given in GCSR, p.320
- the conditional posterior distribution of $\sigma^2|\mu$ is $p(\sigma^2|\mu, \mathbf{y})$ is Inverse Gamma($\frac{n}{2}, \frac{\sum_{i=1}^n (y_i - \mu)^2}{2}$).

13

```
> Rem2
function(y, mu0, tau0, tol, munit, maxiters)
{
# normpostem
# uses EM algorithm to compute posterior mode of normal mean mu
# Gelman, Carlin, Stern and Rubin Section 9.5

#####
# inputs
#####
# mu0    -- prior mean of mu
# tau0   -- prior variance of mu
# tol    -- maximum difference between successive iterates to
#         consider converged
# munit  -- initial value for mu
# y      -- data (vector)
# maxiters -- maximum number of iterations
#####

# initial setup

n <- length(y)
print(c(mu0, tau0, tol, munit, maxiters, n))
mu <- munit
ybar <- mean(y)
sumsqs <- sum( (y - ybar) ^ 2 )
mu0overtau0 <- mu0 / tau0
tau0inv <- 1 / tau0
nsq <- n^2
nsqybar <- nsq * ybar

absdiff <- tol + 1
iters <- 1
```

15

14

```
while ((absdiff > tol) & (iters < maxiters))
{
sumyminmusq <- sumsqs + n * (ybar - mu[iters]) ^ 2
numer <- mu0overtau0 + nsqybar / sumyminmusq
denom <- tau0inv + nsq / sumyminmusq
mu <- c(mu, numer/denom)
iters <- iters + 1
absdiff <- abs( mu[iters] - mu[iters-1])
}

if (absdiff > tol) {
cat(paste("Failed to converge after", iters, " iters!! \n "))
}

list(mu = mu, iters= iters, absdiff=absdiff)
}

> y <- rnorm(25, 5, 2)
> Rem2( y, mu0 = 0, tau0 = 25, tol = 0.0001, munit = 0, maxiters = 25)
[1] 0.0000 25.0000 0.0001 0.0000 25.0000 25.0000
$mu
[1] 0.000000 4.913130 5.120951 5.121486 5.121487

$iters
[1] 5

$absdiff
[1] 4.7413e-07
```

16

Example: Variance-Components Model

Example: We are manufacturers of a product. We purchase batches of the raw material from which we make our product. We are interested in whether variability between batches of raw material is responsible for variation in the final product yield. We randomly select K batches of raw material and draw n_i samples from each. We measure the product yield from each of the resulting samples.

- model

$$y_{ij} = \alpha_i + e_{ij}, \quad i = 1, \dots, K, j = 1, \dots, n_i$$

$$\alpha_i \sim N(\mu, \sigma_\alpha^2)$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

- We are interested in $\theta = \mu, \sigma_\alpha^2, \sigma_e^2$
- Construct EM algorithm by treating unobserved random parameters $\alpha_1, \dots, \alpha_K$ as “missing data”
- Augmented-data likelihood if α 's were known

$$\prod_i \prod_j \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left[-\frac{1}{2} \frac{(y_{ij} - \alpha_i)^2}{\sigma_e^2}\right] \prod_i \frac{1}{\sqrt{2\pi}\sigma_\alpha} \exp\left[-\frac{1}{2} \frac{(\alpha_i - \mu)^2}{\sigma_\alpha^2}\right]$$

17

or

$$\alpha_i | y_{ij}, \theta \sim N((1 - w_i)\bar{y}_i + w_i\mu, v_i)$$

So at iteration k of EM algorithm

$$T_1^{(k)} = \sum_i [(1 - w_i^{(k-1)})\bar{y}_i + w_i\mu^{(k-1)}]$$

$$T_2^{(k)} = \sum_i [(1 - w_i^{(k-1)})\bar{y}_i + w_i\mu^{(k-1)}]^2 + \sum_i v_i^{(k-1)}$$

$$T_3^{(k)} = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i [w_i^{(k-1)2} (\mu^{(k-1)} - \bar{y}_i)^2 + v_i^{(k-1)}]$$

19

- Augmented-data log-likelihood

$$-\frac{JK}{2} \log \sigma_e^2 - \frac{\sum_i \sum_j (y_{ij} - \alpha_i)^2}{2\sigma_e^2} - \frac{K}{2} \log \sigma_\alpha^2 - \frac{\sum_i (\alpha_i - \mu)^2}{2\sigma_\alpha^2}$$

- Augmented-data sufficient statistics (in which augmented-data log likelihood is linear)

$$T_1 = \sum_i \alpha_i$$

$$T_2 = \sum_i \alpha_i^2$$

$$T_3 = \sum_i \sum_j (y_{ij} - \alpha_i)^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \alpha_i)^2$$

- ML estimates based on augmented data (define M-step of EM)

$$\hat{\mu} = \frac{T_1}{K}$$

$$\hat{\sigma}_\alpha^2 = \frac{T_2}{K} - \hat{\mu}^2$$

$$\hat{\sigma}_e^2 = \frac{T_3}{\sum_i n_i}$$

- For E-step, we need to take expectations of T_1, T_2, T_3 conditional on observed y_{ij} 's and current estimates of θ .

To do this, we need conditional distribution of each α_i given the y_{ij} 's. Use Bayes' theorem!

$$\alpha_i | y_{ij}, \theta \sim N\left(\frac{n_i \sigma_\alpha^2}{n_i \sigma_\alpha^2 + \sigma_e^2} \bar{y}_i + \frac{\sigma_e^2}{n_i \sigma_\alpha^2 + \sigma_e^2} \mu, \frac{\sigma_e^2 \sigma_\alpha^2}{n_i \sigma_\alpha^2 + \sigma_e^2}\right)$$

18

Another example: genetic linkage (Rao, 1973)

- observed data: 197 animals are distributed into 4 categories

$$- Y = (y_1, y_2, y_3, y_4) = (125, 18, 30, 34)$$

$$- \text{cell probabilities } \left(\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right)$$

- observed likelihood proportion to

$$(2 + \theta)^{y_1} (1 - \theta)^{y_2 + y_3} \theta^{y_4}$$

- augmented data: split first cell into 2 cells with probabilities $\frac{1}{2}$ and $\frac{\theta}{4}$

$$- Z = (z_1, z_2, z_3, z_4, z_5)$$

$$- z_1 + z_2 = 125$$

$$- z_3 = y_2$$

$$- z_4 = y_3$$

$$- z_5 = y_4$$

- augmented-data likelihood proportional to

$$\theta^{z_2 + z_5} (1 - \theta)^{z_3 + z_4}$$

20

- E-step: expected log augmented-data likelihood

$$\begin{aligned}
 & - Q(\theta, \theta^k) = \\
 & \quad E[(z_2 + z_5)\log(\theta) + (z_3 + z_4)\log(1 - \theta) | \theta^k, Y] \\
 & - p(Z|\theta^k, Y) = p(z_2|\theta^k, Y) \text{ is Binomial}(125, \frac{\theta^k}{2+\theta^k}) \\
 & - Q(\theta, \theta^k) \text{ simplifies to} \\
 & \quad [E(z_2|\theta^k, Y) + z_5]\log(\theta) + (z_3 + z_4)\log(1 - \theta)
 \end{aligned}$$

- M-step

$$\begin{aligned}
 & \frac{\partial Q(\theta, \theta^k)}{\partial \theta} \Big|_{\hat{\theta}} = 0 \\
 & \frac{E(z_2|\theta^k, Y) + z_5}{\hat{\theta}} - \frac{z_3 + z_4}{1 - \hat{\theta}} = 0 \\
 & \theta^{k+1} = \frac{E(z_2|\theta^k, Y) + z_5}{E(z_2|\theta^k, Y) + z_3 + z_4 + z_5}
 \end{aligned}$$

- mle is $\hat{\theta} = 0.6268$

21

Standard errors for EM algorithm

- based on observed Fisher information matrix

– $I_o(\theta)$ is matrix having (i,j)th element

$$-\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log L(\theta|y)$$

– then estimated variance of mle is $I_o(\theta)^{-1}$

- unlike Newton-Raphson and some other iterative root-finding methods, EM algorithm does not compute observed Fisher information matrix at each iteration

- Meilijson's numerical differentiation method (Meilijson, 1989, JRSSB, pp. 127-138)

– use forward-difference formula with score functions

$$\frac{\partial^2 \log(p(\theta|Y))}{\partial \theta^2} \Big|_{\hat{\theta}} \simeq \frac{1}{\epsilon} [S(Y; \hat{\theta} + \epsilon) - S(Y; \hat{\theta})]$$

where

$$S(Y; \theta_0) = \frac{\partial \log(p(\theta|Y))}{\partial \theta} \Big|_{\theta=\theta_0} = \frac{\partial}{\partial \theta} [Q(\theta, \theta_0)|_{\theta=\theta_0}]$$

– obtain $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p)$

– perturb it by adding $\epsilon > 0$ to 1 component

22

- Louis's method (Louis, 1982, JRSSB, pp. 226-233)

$$\begin{aligned}
 & - \frac{\partial^2 \log(p(\theta|Y))}{\partial \theta^2} = \\
 & - \int \frac{\partial^2 \log(p(\theta|Y, Z))}{\partial \theta^2} p(Z|Y, \theta) dZ - \text{var} \left[\frac{\partial \log(p(\theta|Y, Z))}{\partial \theta} \right]
 \end{aligned}$$

– variance with respect to $p(Z|Y, \theta)$

– right-hand side is sum of conditional variance and variance of conditional expectation

- Louis's method with Monte Carlo evaluation of the final variance

Theorems regarding EM algorithm Reference: Tanner Ch. 4

- Every EM algorithm increases the observed-data likelihood (or observed- data posterior) $p(\theta|Y)$ at each iteration, i.e.

$$p(\theta^{(k+1)}|Y) \geq p(\theta^{(k)}|Y)$$

23

24

- Suppose that for a sequence of EM iterates $\theta^{(k)}$

$$\frac{\partial}{\partial \theta} Q(\theta | \theta^{(k)}) |_{\theta = \theta^{(k+1)}} = 0$$

- $\theta^{(k)}$ s converge to θ^*
- $p(Z|Y, \theta)$ is “sufficiently” smooth

Then the $\theta^{(k)}$ converge to a stationary point; i.e.

$$\frac{\partial}{\partial \theta} \log p(\theta | Y) |_{\theta = \theta^*} = 0$$

- If there are multiple local maxima or saddle points, the algorithm may not converge to the global maximum.
- Start EM from several different initial values in hope of finding all modes.

25

Monte Carlo EM

Suppose it is not feasible to compute the expectation in the E-step analytically.

One solution is to use Monte Carlo integration to compute

$$Q(\theta, \theta^{(k)}) = \int_Z \log p(\theta | Y, Z) p(Z | \theta^{(k)}, Y) dZ$$

- Draw $z_1, z_2, \dots, z_m \sim p(Z | Y, \theta^{(k)})$
- let $\hat{Q}^{(k+1)}(\theta, \theta^{(k)}) = \frac{1}{m} \sum_{j=1}^m \log p(\theta | z_j, Y)$

References:

- Wei and Tanner, 1990, *JASA*
- Chan and Ledolter, 1995, *JASA*

27

Disadvantages of EM

- convergence linear with rate proportional to fraction of information about θ in $p(\theta | Y)$ that is observed
- EM does not yield estimates asymptotically equivalent to ML estimates after single iteration (Newton-Raphson and scoring algorithms do)

26

Standard Errors for Monte Carlo EM

We can approximate the observed information matrix of the observed data at the mle with 2 Monte Carlo integrations:

$$\frac{\partial^2 \log(p(\theta | Y))}{\partial \theta^2} |_{\hat{\theta}} \simeq \frac{1}{m} \sum_{j=1}^m \frac{\partial^2 \log p(\theta | Y, z_j)}{\partial \theta^2} |_{\hat{\theta}} + \frac{1}{m} \sum_{j=1}^m \left(\frac{\partial \log p(\theta | Y, z_j)}{\partial \theta} |_{\hat{\theta}} \right)^2$$

28