E. Slud 10/15/09

Stat 705, HW 13 Due Fri. 10/30/09

This problem describes and asks you to implement the EM algorithm for a balanced two-way mixed-effects normal linear model with block fixed effects and cluster random effects, and to compare the results of your ML estimation via EM with the *nlm*-based iteration which can be made much simpler in this problem.

Data & Problem Setting. Suppose that an array $\mathbf{X} = \{X_{ij},\ 1 \le i \le B,\ 1 \le j \le M\}$ of random variables is to be observed, quantitative responses in an experiment where i indexes a fixed number B of *blocks*, and j indexes 'clusters' the number M of which is large and growing. The distributional assumption is

$$X_{ij} \;=\; \alpha_i \,+\, U_j \,+\, \epsilon_{ij} \tag{1}$$

for all indices i, j, where the errors $\mathbf{U}$ and ϵ are independent, and

$$U_j \overset{iid}{\sim} \mathcal{N}(0,\, \sigma_u)^2 \quad , \qquad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0,\, \sigma^2)$$

Here the α_i are regarded as unknown constant 'fixed block effect' parameters, which the U_j are unobserved 'random cluster effect' random variables. (If the values U_j were also treated as parameters, their number M would grow proportionately to the total dataset size MB, which would invalidate most of our large-sample MLE theory.) Thus the unknown parameters, of dimension $B + 2$, are

$$\vartheta \;\equiv\; (\underline{\alpha},\, \sigma_u^2,\, \sigma^2) \;\in\; \Theta \;\equiv\; \mathbf{R}^B \times \mathbf{R}^+ \times \mathbf{R}^+$$

Problem (to be handed in): Simulate 10 datasets of the type just described, with $B = 10, M = 40$, and $\sigma_u^2 = 2, \sigma^2 = 1$, and α_i generated (once, kept the same for all 10 simulations) as *iid* Uniform$(8, 12)$ random variates. For each of the datasets, find the maximum likelihood estimators $\hat{\vartheta}$ in two ways, by numerical iteration using *nlm* on the Likelihood, and by the EM algorithm obtained by treating $\mathbf{U} = \{U_j,\ j = 1, \ldots, M\}$ as *missing data* (Little and Rubin, **Statistical Analysis with Missing Data**, 2nd ed. 2002). **Make sure that the ML estimators you obtain for both methods are the same, and verify that each EM iteration increases the log-likelihood expression (3).**

The computing formulas for both the maximum likelihood calculation and the EM iteration are derived on the following pages. The ML simplifies in this problem because the estimators $\hat{\alpha}_i = \bar{X}_{i\cdot}$ are actually given in closed form and can be substituted into the log-likelihood, leaving only a 2-dimensional numerical maximization over σ^2 and σ_u^2.

Maximum Likelihood Calculation

The ML estimators can in fact be found in closed form in this problem, as we will show, but some care is needed because the data X_{ij} are independent only in clusters corresponding to the columns $\mathbf{X}^{(j)} = (X_{ij})_{i=1}^B \in \mathbf{R}^B$ with multivariate normal distribution expressed in terms of the corresponding column of the ϵ_{ij} array as

$$\mathbf{X}^{(j)} = \underline{\alpha} + U_j \, \mathbf{1}_B + \epsilon^{(j)} \overset{iid}{\sim} \mathcal{N}(\underline{\alpha}, \, \sigma^2 \, I_B + \sigma_u^2 \, \mathbf{1}_B^{\otimes 2}) \tag{2}$$

where I_B denotes the $B \times B$ identiy matrix and $\mathbf{1}_B$ the B-dimensional column vector of 1's. Now for future reference define

$$\bar{x} \equiv \frac{1}{B} \sum_{i=1}^{B} x_i \quad , \quad \bar{\alpha} \equiv \frac{1}{B} \sum_{i=1}^{B} \alpha_i \quad , \qquad \gamma \equiv \frac{B\sigma_u^2}{B\sigma_u^2 + \sigma^2}$$

In terms of these notations, we can write the density corresponding to (2) as

$$f_{\mathbf{X}^{(j)}}(\mathbf{x}) = (2\pi)^{-B/2} \, \sigma^{-B+1} \, (\sigma^2 + B\sigma_u^2)^{-1/2} \, \exp\left\{ -\frac{1}{2\sigma^2} \left(\|\mathbf{x} - \underline{\alpha}\|^2 - B \, \gamma \, (\bar{\mathbf{x}} - \bar{\alpha})^2 \right) \right\}$$

Then it is not hard to check after partial differentiation with respect to α_i and, separately, with respect to σ_u^2 and to σ^2, that the observed-data log-likelihood $logLik(\vartheta, \mathbf{X})$ (after subtraction of a constant not involving ϑ) is $=$

$$-\frac{M(B-1)}{2} \log \sigma^2 - \frac{M}{2} \log(\sigma^2 + B\sigma_u^2) - \frac{1}{2\sigma^2} \sum_{i,j} (X_{ij} - \alpha_i)^2 + \frac{B\gamma}{2\sigma^2} \sum_j (\bar{X}_{\cdot j} - \bar{\alpha})^2 \tag{3}$$

and is maximized uniquely over $\vartheta = \hat{\vartheta}$ precisely when

$$\hat{\alpha}_i = \bar{X}_{i\cdot} \quad , \qquad B \, \hat{\sigma}_u^2 + \hat{\sigma}^2 = \frac{1}{M} \, SSBC \quad , \qquad \hat{\sigma}^2 = \frac{SSW - SSBC}{M(B-1)} \tag{4}$$

Here we have made use of the standard notations

$$\bar{X}_{i\cdot} = \frac{1}{M} \sum_{j=1}^{M} X_{ij} \quad , \quad \bar{X}_{\cdot j} = \frac{1}{B} \sum_{i=1}^{B} X_{ij} \quad , \quad \bar{X}_{\cdot\cdot} = \frac{1}{BM} \sum_{i=1}^{B} \sum_{j=1}^{M} X_{ij}$$

and, with *SSW* standing for 'within-block' and *SSBC* for 'between-cluster' sums of squares,

$$SSW = \sum_{i=1}^{B} \sum_{j=1}^{M} (X_{ij} - \bar{X}_{i\cdot})^2 \quad , \qquad SSBC = B \sum_{j=1}^{M} (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2$$

Note that the MLE equations closely resemble the Method-of-Moment or REML estimators you would have learned in an Analysis of Variance course: indeed, those estimators are just as in (4) with M replaced by $M-1$ in the denominators in the second and third equations.

Calculations of Explicit E and M steps in the EM algorithm

To obtain the EM iteration explicitly, we begin with the conditional distribution of the missing data given the observed data. As argued in class, first the independence of $(U_j,\, \mathbf{X}^{(j)})$ for $j = 1, \ldots, M$, and the 'sufficiency' of $\bar{X}_{\cdot j}$ for U_j regarded as a parameter (with ϑ known and fixed), implies that

$$f_{U_j|\mathbf{X}}(\mathbf{u}|\mathbf{X}) \;=\; f_{U_j|\mathbf{X}^{(j)}}(\mathbf{u}|\mathbf{X}^{(j)}) \;=\; f_{U_j|\bar{X}_{\cdot j}}(\mathbf{u}|x) \;\sim\; \mathcal{N}(\gamma(x-\bar{\alpha}),\, \frac{\gamma}{B}\sigma^2) \qquad (5)$$

where the last normal-density relation of (5) follows from the fact that U_j and $\bar{X}_{\cdot j}$ are jointly Gaussian and that the two random variables $U_j - \gamma(\bar{X}_{\cdot j} - \bar{\alpha})$ and $\bar{X}_{\cdot j}$ are uncorrelated.

Next, if a parameter $\vartheta^* = (\underline{\alpha}^*,\, (\sigma_u^*)^2,\, (\sigma^*)^2)$ value is fixed with respect to which conditional expectations are calculated, then by (5),

$$E_{\vartheta^*}(U_j|\mathbf{X}) \;=\; \gamma^*\,(\bar{X}_{\cdot j} - \bar{\alpha}^*) \;\;,\quad E_{\vartheta^*}(U_j^2\,|\mathbf{X}) \;=\; (\gamma^*)^2\,(\bar{X}_{\cdot j} - \bar{\alpha}^*)^2 + \frac{\gamma^*}{B}(\sigma^*)^2 \quad (6)$$

Now the joint log-likelihood of the observed and missing data has the form

$$-\frac{M}{2}\log\sigma_u^2 \;-\; \frac{MB}{2}\,\log\sigma^2 \;-\; \frac{1}{2}\,(\frac{1}{\sigma_u^2} \;+\; \frac{B}{\sigma^2})\sum_{j=1}^{M} U_j^2$$

$$+\frac{B}{\sigma^2}\sum_{j=1}^{M} U_j(\bar{X}_{\cdot j} - \bar{\alpha}) \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{B}\sum_{j=1}^{M}(X_{ij} - \alpha_i)^2$$

from which we calculate conditional expectation given $\mathbf{X}$ under parameters ϑ^*, with the explicit E-step result

$$-\frac{M}{2}\log\sigma_u^2 \;-\; \frac{MB}{2}\,\log\sigma^2 \;-\; \frac{1}{2}\,(\frac{1}{\sigma_u^2} \;+\; \frac{B}{\sigma^2})\sum_{j=1}^{M}\Big((\gamma^*)^2\,(\bar{X}_{\cdot j} - \bar{\alpha}^*)^2 + \frac{\gamma^*}{B}\,(\sigma^*)^2\Big)$$

$$+\frac{\gamma^* B}{\sigma^2}\sum_{j=1}^{M}(\bar{X}_{\cdot j} - \bar{\alpha}^*)\,(\bar{X}_{\cdot j} - \bar{\alpha}) \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^{B}\sum_{j=1}^{M}(X_{ij} - \alpha_i)^2 \qquad (7)$$

The M-step consists in finding explicit formulas for the maximizer with respect to $\vartheta = (\underline{\alpha}, \sigma_u^2, \sigma^2)$ of expression (7). First, it is easy to see that if in each iteration the parameter $\bar{\alpha}_i^*$ is taken equal to $\bar{X}_{i\cdot}$, then the maximizer $\tilde{\alpha}_i$ of (7) has the same value ! Thus there is no loss of generality (and considerable simplification) in taking the EM estimator of $\alpha_i = \bar{X}_{i\cdot}$ in every iteration. As a result, the quantity (7) to maximize with respect to (σ_u^2, σ^2) takes the form

$$-\frac{M}{2}\log\sigma_u^2 - \frac{MB}{2}\log\sigma^2 - \frac{SSW}{2\sigma^2} - M\frac{\gamma^*}{2}\,(\frac{1}{B\sigma_u^2} + \frac{1}{\sigma^2})\,(\sigma^*)^2$$

$$-\frac{SSBC}{2}\Big((\gamma^*)^2\,(\frac{1}{B\sigma_u^2} + \frac{1}{\sigma^2}) - \frac{2\,\gamma^*}{\sigma^2}\Big) \qquad (8)$$

where *SSW, SSBC* are as defined within our Maximum Likelihood derivation.

Maximizing (8) respectively with respect to σ_u^2 and σ^2 yields the two unique solutions

$$\tilde{\sigma}_u^2 = \frac{1}{B}\,\gamma^*\,(\sigma^*)^2 + \frac{SSBC}{BM}\,(\gamma^*)^2 \qquad (9)$$

$$\tilde{\sigma}^2 = \frac{1}{MB}\Big(SSW + \gamma^*\,SSBC\,(\gamma^* - 2)\Big) + \frac{1}{B}\,\gamma^*(\sigma^*)^2 \qquad (10)$$

Thus, if at the k'th EM iteration we have parameter values $\alpha_i^{(k)} = \bar{X}_{i\cdot}$ and $(\sigma_u^{(k)})^2$, $(\sigma^{(k)})^2$, then at the $(k+1)$'st iteration the EM estimates are obtained from equations (9) and (10) as

$$\alpha_i^{(k+1)} = \bar{X}_{i\cdot} \qquad (11)$$

$$(\sigma_u^{(k+1)})^2 = \frac{(\sigma_u^{(k)})^2}{B\,(\sigma_u^{(k)})^2 + (\sigma^{(k)})^2}\Big((\sigma^{(k)})^2 + \frac{SSBC\,\cdot\,B\,(\sigma_u^{(k)})^2}{M\,\cdot\,(B(\sigma_u^{(k)})^2 + (\sigma^{(k)})^2)}\Big) \qquad (12)$$

$$(\sigma^{(k+1)})^2 = (\sigma_u^{(k+1)})^2 + \frac{1}{MB}\Big(SSW - \frac{2\,B\,(\sigma_u^{(k)})^2}{B(\sigma_u^{(k)})^2 + (\sigma^{(k)})^2}\cdot SSBC\Big) \qquad (13)$$