October 16, 2015

1 Data

The data set contains six measurements made on 100 genuine and 100 counterfeit old Swiss 1000 franc bank notes.

- ***Status*** — the status of the banknote: genuine or counterfeit
- ***Length*** — Length of bill (mm)
- ***Left*** — Width of left edge (mm)
- ***Right*** — Width of right edge (mm)
- ***Bottom*** — Bottom margin width (mm)
- ***Top*** — Top margin width (mm)
- ***Diagonal*** — Length of diagonal (mm)

2 Mclust

```
# incude packages
library(mclust)
data(banknote)
# Ellipsoidal, varying volume, shape, and orientation
system.time(res <- Mclust(banknote[, -1], 2, "VVV"))
##    user  system elapsed
##   0.025   0.001   0.027
# assess the classification
table(res$cl, banknote[, 1])
##
##     counterfeit genuine
##   1           0      99
##   2         100       1
# mclust correctly classifies all but one of the observations
```

3 Methodology

3.1 Introduction

Traditional MBC has computational problems when the number of variables (p) is very large. Memory contraints are encountered. The proposed method attempts to lessen this problem instead of fitting a single EM algorithm on all of the variables at once, by working dividing the variables into batches. Separate EM models are fitted to each of the batches, the starting values for each batch have been determined by the previous batches. MBCbig P is approximating the problem recursively. The advantage of this approach is that is a lot more memory efficient. Don't need to try and store these really big objects.

3.2 Notation

η mixing proportions — population MLE probabilities of belonging to group g
y is the data
z are the latent variables, indicating group membership
θ is the set of parameters that are relevent
μ is the vector of means
Σ is the covariance matrix
g is the number of groups
n is the number of data points

3.3 Definition

Initilise data $p = 4$. Let $a = (1, 2)$ & $b = (3, 4)$

Specify the data. Data had 4 variables 1:4. Group a is for variable 1 & 2 and group b is for variable 3 & 4

$$f(y_i) = \sum_{g=1}^{2} \eta_g f(y_i|\mu_g, \Sigma_g)$$

$$f\left[(y_a, y_b)^T\right] = \sum_{g=1}^{2} \eta_g f\left[(y_a, y_b)^T \,|\, (\mu_{ga}, \mu_{gb})^T\right] \begin{pmatrix} \Sigma_{gaa} & \Sigma_{gab} \\ \Sigma_{gba} & \Sigma_{gbb} \end{pmatrix}$$

First is the standard form of the function, second is the form of the alternative proposed, where there are parameters for the two groups: two sets of data, two means, and the covariance matrix will be 2x2, with variances and covariances

$$\text{Also} \quad \Lambda_g = \Sigma_g^{-1} = \begin{pmatrix} \Lambda_{ga} & \Lambda_{gab} \\ \Lambda_{gba} & \Lambda_{gb} \end{pmatrix} \tag{1}$$

Using Λ to represent the inverse of the covariance matrix

$$\text{Note} \quad \Lambda_{gba} = \Lambda_{gab}^T \tag{2}$$

Just one of those matrix results from the Matrix Cookbook etc.

$$\text{We have} \quad f_g\left[(y_a, y_b)^T\right] = f_g \underbrace{[y_a|y_b]}_{N\left(\mu_{ga|b}, \Lambda_{gaa}^{-1}\right)} \quad f_g \underbrace{[y_b]}_{N\left(\mu_b, \Sigma_{gbb}\right)} \tag{3}$$

This is a joint probability distribution property for random variables can be written as the product of the product of the conditional distribution and the marginal distribution. Both of these are normally distributed.

$$\text{Where} \quad \mu_{ga|b} = \mu_{ga} - \Lambda_{ga}^{-1}\Lambda_{gab}\left(y_b - \mu_{gb}\right)$$
$$\Lambda_g^{-1} = \Sigma_{ga} - \Sigma_{gab}\Sigma_{gb}^{-1}\Sigma_{gba}$$

These are results from working with block matrices. Definition of the conditional mean and conditional variance.

3.4 Batch A

Say that the data is divided into three batches A, B & C

```
# define the batches
BatchA <- banknote[, 2:3]
BatchB <- banknote[, 4:5]
BatchC <- banknote[, 6:7]
```

So for batch A of the variables we have:

$$L = p\left(y_A|\eta,\theta\right) = \prod_{i=1}^{N}\sum_{g=1}^{2}\eta_g p\left(y_{iA}|\theta_g\right), \text{where} \quad \theta_g = \{\mu_g, \Sigma_g\} \tag{4}$$

We have a density function $p\left(y_A|\eta,\theta\right)$ that is governed by the set of parameters(e.g. mean and covariances for Gaussians). We also have a data set of size N, supposedly drawn from this distribution.Assume that these data vectors are independent and identically distributed with distribution p. Therefore, the resulting density for the samples

$$\text{Accounting for missing } z_i\text{'s} \quad (z_i = (z_{i1}, z_{i2}) \,\&\, z_{ig} = \{1, 0\}) \tag{5}$$

Optimizing the likelihood function is analytically intractable. The likelihood function can be simplified by assuming the existence of and latent parameters. Here this parameter is defined z The zs are treated as the 'missing' data and are indicator variables indicating the group membership of each of the N observations i.e. if G is the total number of groups present where z_{ig} = {1 if observation i belongs to group g, 0 otherwise}

$$\begin{aligned} L_c\left(\theta,\eta|y_a z\right) &= p\left(y_a z|\eta,\theta\right) \\ &= \prod_{i=1}^{N}\prod_{g=1}^{2}\left[\eta_g p\left(y_{iA}|\theta_g\right)\right]^{z_{ig}} \end{aligned}$$

Likelihood of the parameters given the data, or just the likelihood function. The likelihood is thought of as a function of the parameterswhere the data y is fixed. In the maximum likelihood problem, our goal is to find the parameters that maximizes L.

Initialize

$\eta^{(0)}$ & $\theta^{(0)}$ Let $t = 0$

Need to specify starting values for the mixing proportions and for the parameters values

E-Step: Compute

$$\hat{z_{ig}^{(t)}} = \frac{\eta_{g1}^{(t)} p\left(y_{iA}|\theta_g^{(t)_g}\right)}{p\left(y_i|\theta_G\right)} \text{ Bayes rule}$$
$$= \frac{\eta_g^{(t)} p\left(y_{iA}|\theta_g^{(t)}\right)}{\sum_{g=1}^{G} \eta_{g1}^{(t)} p\left(y_{iA}|\theta_g^{(t)}\right)} \quad i = 1, \ldots, N \ \& \ g = 1, 2$$

here the expected value of the missing data, conditional on the observed data and all current parameter estimates, is computed. This is just the probability of of one group over the total probability.

M-Step: Maximize

$$l_c\left(\theta, \eta|y, z\right) = \sum_{i=1}^{N}\sum_{g=1}^{2} \hat{z}_{ig}^{(t)} \left[\log \eta_g + \log p\left(y_{iA}|\theta_g\right)\right] \tag{6}$$

here the expected complete data log likelihood is maximized with respect to the parameters. η and θ are unrelated they can be worked out separately. as going to be using multivariate normal distribution the these parameters that need to be estimated are μ Σ

$$\frac{\partial}{\partial \eta_g}\left[\sum_{i=1}^{N}\sum_{g=1}^{G} log\left(\eta_g\right) z_{ig} + \lambda\left(\sum_g \eta_g - 1\right)\right] = 0 \tag{7}$$

$$\sum_{i=1}^{N} \frac{1}{\eta_g} z_{ig} + \lambda = 0 \quad \text{summing both sides, } \lambda = -N \tag{8}$$

$$\text{to yield} \quad \eta_g^{(t+1)} = \frac{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)}}{N}$$

$$p(y|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} exp(-\frac{1}{2}(y_i - \mu)'\Sigma^{-1}(y_i - \mu)) \tag{9}$$

Need to redo the indices and the subscripts in this section. Need to show some of the matrix operations that are needed.

$$\sum_{i=1}^{N}\sum_{g=1}^{G}\left(\frac{1}{2}\log\left(\Sigma\right) - \frac{1}{2}\left(y_i - \mu_i\ \right)^T \Sigma^{-1}\left(y_i - \mu_i\right)\right) z_i \tag{10}$$

Take the derivative with respect to μ and set equal to zero get

$$\sum_{i=1}^{N} \Sigma^{-1}\left(y_i - \mu\right) zi = 0 \tag{11}$$

$$\text{solve for } \mu \quad \mu_g^{(t+1)} = \frac{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)} y_{iA}}{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)}} \tag{12}$$

Want to find Σ

$$\sum_{i=1}^{N}\left[\frac{1}{2}\log\left(\Sigma^{-1}\right)\sum_{g=1}^{G} z_{ig} - \frac{1}{2}\sum_{g=1}^{G} z_{ig} tr\left(\Sigma^{-1}\left(y_i - \mu\right)\left(y_i - \mu_g\right)^T\right)\right] \tag{13}$$

Take derivative with respect to Σ_g

$$\frac{1}{2}\sum_{n=1}^{N} z_{ig}\left(2\Sigma - diag\left(\Sigma\right)\right) - \frac{1}{2}\sum_{i=1}^{N} z_{ig}\left(2\left(y_i-\mu\right)\left(y_i-\mu_g\right)^T - diag\left(y_i-\mu\right)\left(y_i-\mu_g\right)^T\right) \quad (14)$$

$$= \frac{1}{2}\sum_{i=1}^{N} z_{ig}\left(2\Sigma - \left(y_i-\mu\right)\left(y_i-\mu_g\right)^T - diag\left(\Sigma - \left(y_i-\mu\right)\left(y_i-\mu_g\right)^T\right)\right) \quad (15)$$

Setting the derivative equal to zero implies

$$\sum_{i-1}^{N} z_{ig}\left(\Sigma - \left(y_i-\mu\right)\left(y_i-\mu_g\right)^T\right) = 0 \quad (16)$$

$$\Sigma_g^{(t+1)} = \frac{\sum_{i=1}^{N}\hat{z}_{ig}^{(t)}\left[y_{iA}-\mu_g^{(t+1)}\right]\left[y_{iA}-\mu_g^{(t+1)}\right]^T}{\sum_{i=1}^{n}\hat{z}_{ig}^{(t)}} \quad (17)$$

Compute the parameter estimates by differenting with respect to each of the parameters set to zero and solve. The estimates of the new parameters are in terms of the old parameters. These equations for the parameters perform both the expectation step and the maximization step simultaneously. The algorithm proceeds by using the newly derived parameters as the guess for the next iteration.

It is computationally cheaper to express in terms of the conditional expectations of the sufficent

statistics T_{i1}, T_{i2} and T_{i3} given by

$$S_{g1}^{(t)} = \sum_{i=1}^{n}\hat{z}_{ig}^{(t)}$$

$$S_{g2}^{(t)} = \sum_{i=1}^{n}\hat{z}_{ig}^{(t)} y_{iA}$$

both of which are used in $S_{g3}^{(t)} = \sum_{i=1}^{n}\hat{z}_{ig}^{(t)} y_{iA} y_{iA}^T$

Compute sufficient statistics e.g. $\Pr(x\|t,\theta) = \Pr(x\|t)$.

$$\text{where}\sum_{g}^{(t+1)} = \frac{\left\{S_{g3}^{(t)} - S_{g1}^{(t)-1} S_{g2}^{(t)} S_{gs}^{(t)^T}\right\}}{S_{g1}^{t}} \qquad g = 1,2 \quad (18)$$

Compute Σ from the sufficient statistics. This speeds up the calculations as these numbers can be substituted instead of working out the formula again.

A "convergence" we have $\eta_g^{(A)}$, μ_g^A & $\Sigma_g^{(A)}$ & $p\left(y_{iA}|\theta_g^{(A)}\right)$ for g=1,2

at the end of Batch A have estimates for the parameters for Batch A and the likihood for batch A

3.5 Batch B

For Batch B of the variables we have have:

Batch B is very similar to Batch A, but now everything for B is going to be conditional on A

$$\begin{aligned} L_c\left(\theta,\eta|\left(y_A,y_B\right),z\right) &= p\left(y_A,y_B,z|\eta,\theta\right) \\ &= \prod_{i=1}^{N}\prod_{g=1}^{2}\left[\eta_g p\left(y_{iA},y_{iB}|\theta_g\right)\right]^{zig} \\ &= \prod_{i=1}^{N}\prod_{g=1}^{2}\left[\eta_g p\left(y_{iB}|y_{iA},\theta_g\right)p\left(y_{iA}|\theta_g\right)\right]^{zig} \end{aligned}$$

complete data log likelihood

$$\text{Note that now} \quad \theta_g = \left(\theta_{gA},\theta_{gB}\right) = \left\{\begin{pmatrix}\mu_{gA}\\ \mu_{gB}\end{pmatrix}\begin{pmatrix}\Sigma_{gA} & \Sigma_{gBA}\\ \Sigma_{gAB} & \Sigma_{gB}\end{pmatrix}\right\}$$

Initialize

Set $\eta_g^{(0)} = \eta_g^A$, g=1,. . .,G,let t=0

for this batch will need to actually compute estimates for the parameters rather than just specifying the initiliasation values. Going to be using the information on for Batch A for Batch B which hopefully will act as very good starting values

$$\mu_g^{(0)} = \left(\mu_{gA},\mu_{gB}\right)^{(0)} = \left(\mu_{gA}^{(A)},y_{gB}\right)$$

$$\Sigma_g^{(0)} = \begin{pmatrix}\Sigma_{gA}^{(A)} & \text{Cov}\left(y_{gA},y_{gB}\right)\\ \text{Cov}\left(y_{gA},y_{gB}\right) & \text{Cov}\left(y_{gB}\right)\end{pmatrix}$$

E-Step: Compute

$$\begin{aligned} \hat{z}_{ig}^{(t)} &= \frac{\eta_g^{(t)} p\left(y_{iA},y_{iB}|\theta_g^{(t)}\right)}{\sum_{g=1}^{G}\eta_g^{(t)} p\left(y_{iA},y_{iB}|\theta_g^{(t)}\right)} \text{i=1, . . ., N g=1,2} \\ &= \frac{\eta_g^{(t)} p\left(y_{iB}|y_{iA},\theta_g^{(t)}\right)p\left(y_{iA}|\theta_g^{(A)}\right)}{\sum_{g=1}^{G}\eta_g^{(t)} p\left(y_{iB}|y_{iA},\theta_g^{(t)}\right)p\left(y_{iA}|\theta_g^{(A)}\right)} \end{aligned}$$

Again probability over total probability, second equation is rewriting using probability rules

$$\text{where} \quad p\left(y_{iB}|y_{iA},\theta_g^{(t)}\right) \sim \text{MVN}\left(\mu_{gB|A},\Lambda_{gB}^{-1}\right)$$

Going to assume that the distribution is normally with those parameters

$$\begin{aligned} \mu_{gB|A} &= \mu_{gB} - \Lambda_{gB}^{-1}\Lambda_{gBA}\left(y_{iA}-\mu_{gA}\right) \\ \Lambda_{gB}^{-1} &= \Sigma_{gB} - \Sigma_{gBA}\Sigma_{gA}^{-1}\Sigma_{gAB} \quad \text{(use plug estimates)} \end{aligned}$$

Need to compute the covariances between Batch A and Batch B, could just work out the covariance of the batch and use this as an estimate

M-Step Maximize

$$l_c\left(\theta,\eta|\left(y_A,y_B\right),z\right)=\sum_{i=1}^{N}\sum_{g=1}^{2}\hat{z}_{ig}^{(t)}\left[\log\eta_g+\log p\left(\left(y_{iA},y_{iB}\right)|\theta_g\right)\right]$$

$$\text{to yield}\quad \eta_g^{(t+1)}=\frac{\sum_{i=1}^{N}\hat{z}_{ig}^{(t)}}{N}$$

$$\mu_{gB}^{(t+1)}=\frac{\sum_{i=1}^{N}\hat{z}_i g^{(t)} y_{iB}}{\sum_{i=1}^{N}\hat{z}_{ig}^{(t)}}$$

$$\Sigma_{gB}^{(t+1)}=\frac{\left\{S_{g3}^{(t)}-S_{g1}^{(t)-1}S_{g2}^{(t)}S_{g2}^{(t)T}\right\}}{S_{g1}^{(t)}}\qquad \text{g=1,2}$$

where

$$S_{g1}^{(t)}=\sum_{i=1}^{n}\hat{z}_{ig}^{(t)}$$

$$S_{g2}^{(t)}=\sum_{i=1}^{n}\hat{z}_{ig}^{(t)}y_{iB}$$

$$S_{g3}^{(t)}=\sum_{i=1}^{n}\hat{z}_{ig}^{(t)}y_{iB}y_{iB}^{T}$$

Want to compute new parameters for this Batch. Will be interesting to see how much the parameters are changing over the batches. Again look at the sufficnient statistics in order to calculate Σ efficiently

Convergence

$$\eta_g^{(B)},\mu_g^{(B)},\Sigma_g^{(B)}\ \ \&\ \ p\left(y_{iB}|y_{iA},\theta_g^{(B)}\right)\qquad g=1,2$$

After this step will have estimates for the parameters from batch B, the probability will depend on Batch A

3.6 Batch C

Batch C again very similar just the probabilities depend on Batch A and Batch B the data and the parameter estimates

$$L_c\theta,\eta|\left(y_1,y_B,y_C\right)=p\left(y_A,y_B,y_C,z|\theta,\eta\right)$$
$$=\prod_{i=1}^{n}\prod_{g=1}^{2}\left[\eta_g p\left(y_{iA},y_{iB},y_{iC}|\theta_g\right)\right]^{z_{ig}}$$

$$\text{Note that now}\quad \theta_g=\left\{\begin{pmatrix}\mu_{gAB}\\ \mu_{gC}\end{pmatrix}\begin{pmatrix}\Sigma_{gAB} & \Sigma_{gCAB}\\ \Sigma_{gABC} & \Sigma_{gC}\end{pmatrix}\right\}$$

Initialize

$$\text{Set}\quad \eta_g=\eta_g^{(B)}\quad g=1,\ldots,G\quad t=0$$

$$\mu_g^{(0)}=\left(\mu_g^{(B)},\bar{y}_{gC}\right)$$

$$\Sigma_g^{(0)}=\begin{bmatrix}\Sigma_{gAB}^{(B)} & Cov\left(y_{gAB},y_{gC}\right)\\ Cov\left(y_{gAB},y_{gC}\right) & Cov\left(y_{gC}\right)\end{bmatrix}$$

E-Step compute

$$\hat{z}_{ig}^{(t)}=\frac{\eta_g^{(t)}p\left(y_{iA},y_{iB},y_{iC}|\theta_g^{(t)}\right)}{\sum_{g=1}^{G}\eta_g^{(t)}p\left(y_{iA},y_{iB},y_{iC}|\theta_g^{(t)}\right)}$$

$$\approx\frac{\eta_g^{(t)}p\left(y_{iC}|y_{iA},y_{iB},\theta_g^{(t)}\right)p\left(y_{iB}|y_{iA},\theta_g^{(B)}\right)p\left(y_{iA}|\theta_g^{(A)}\right)}{\sum_{g=1}^{G}\eta_g^{(t)}p\left(y_{iC}|y_{iA},y_{iB},\theta_g^{(t)}\right)p\left(y_{iB}|y_{iA},\theta_g^{(B)}\right)p\left(y_{iA}|\theta_g^{(A)}\right)}$$

where:

$$p\left(y_{iC}|y_{iA},y_{iB},\theta_g^{(t)}\right)\sim MVN\left(\mu_{gC|B},\Lambda_{gC}^{-1}\right)$$

where

$$\mu_{gC|B}=\mu_{gC}-\Lambda_{gC}^{-1}\Lambda_{gCB}\left(y_{iB}-\mu_{gB}\right)$$
$$\Lambda_{gC}^{-1}=\Sigma_{gC}-\Sigma_{gCB}\Sigma_{gB}^{-1}\Sigma_{gBC}\quad\text{use plug in estimates}$$

Need to compute the covariances between Batch B and Batch C

M-Step: Maximize

$$l_c\left(\theta,\eta|\left(y_A,y_B,y_C\right),z\right)$$

to yield

$$\eta_g^{(t+1)} = \frac{\sum \hat{z}_{ig}^{(t)}}{N}$$
$$\mu_{gC}^{(t+1)} = \frac{\sum \hat{z}_{ig}^{(t)} y_{iC}}{\sum_{i=1}^{N} \hat{z}_{ig}^{(t)}}$$
$$\Sigma_{gC}^{(t+1)} = \frac{\left\{ S_{g3}^{(t)} - S_{g1}^{(t)^{-1}} S_{g2}^{(t)^T} S_{g2} \right\}}{S_{g1}^{(t)}}$$

where

$$S_{g1}^{(t)} = \sum \hat{z}_{ig}^{(t)}$$
$$S_{g2}^{(t)} = \sum \hat{z}_{ig}^{(t)} y_{iC}$$
$$S_{g3}^{(t)} = \sum \hat{z}_{ig}^{(t)} y_{iC} y_{iC}^T$$

Compute

$$p\left(y_{iC} | y_{iA}, y_{iB}, \theta_g^{(C)}\right) \dot{\sim} MVN\left(\mu_{gC|B}, \Lambda_{gC}^{-1}\right)$$

Convergence

$$\eta_g^{(C)}, \mu_g^{(C)}, \Sigma_g^{(C)} \quad \& \quad p\left(y_{iC} | y_{iA}, Y_{iB}, \theta_g^{(C)}\right)$$

Will have estimates for the parameters from this Batch. The probability for this Batch depends on both of the other batches