Estimating *m* and *u* Probabilities Using EM

- Based on Winkler 1988 "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671.
- Uses the identity

$$P(\gamma)=P(\gamma|M)P(M)+P(\gamma|U)P(U)$$

- Imposes conditional independence

Estimating *m* and *u* Probabilities Using EM: Algorithm I

- Select blocking variables that give file sizes for the α and β files that are feasible (this depends on the size of your computer). There are *N* elements in $\alpha \times \beta$.
- For each matching variable, choose an initial m_k and u_k, often 0.9 and 0.1 respectively. Note that they do not have to sum to one.

Estimating *m* and *u* Probabilities Using EM: Algorithm II

- Set up the complete data model:
  - Parameters: *m, u, p*, where the scalar *p* is the proportion of matches in α x β and *m* and *u* are the (*k* x 1) vectors of unknown probabilities. An initial value for *p* is also required.
  - r_j is an element of α x β; γ^j is its associated agreement vector
  - Either r_j is an element of *M* or r_j is an element of *U.* Let $g_j = (1,0)$ when r_j is an element of *M* and $g_j = (0,1)$ when r_j is an element of *U.*
  - Complete data g = (g_j , γ^j)

Complete Data Likelihood Function

$$\ln f\left(x \mid m, u, p\right) = const. + \sum_{j=1}^{n} g_j \bullet \left(\ln P\left(\lambda^j \mid M\right), \ln P\left(\lambda^j \mid U\right)\right)$$

$$+ \sum_{j=1}^{n} g_j \bullet \left(\ln p, \ln\left(1 - p\right)\right)$$

E-step

- Replace g_j with its expectation (P($M|\gamma^j$), P($U|\gamma^j$))

$$P\left(M \mid \gamma^j\right) = \frac{\hat{p}\prod_{k=1}^{K}\left(\hat{m}_k\right)^{\gamma_k^j}\left(1-\hat{m}_k\right)^{1-\gamma_k^j}}{\hat{p}\prod_{k=1}^{K}\left(\hat{m}_k\right)^{\gamma_k^j}\left(1-\hat{m}_k\right)^{1-\gamma_k^j} + \left(1-\hat{p}\right)\prod_{k=1}^{K}\left(\hat{u}_k\right)^{\gamma_k^j}\left(1-\hat{u}_k\right)^{1-\gamma_k^j}}$$

$$P\left(U \mid \gamma^j\right) = \frac{\left(1-\hat{p}\right)\prod_{k=1}^{K}\left(\hat{u}_k\right)^{\gamma_k^j}\left(1-\hat{u}_k\right)^{1-\gamma_k^j}}{\hat{p}\prod_{k=1}^{K}\left(\hat{m}_k\right)^{\gamma_k^j}\left(1-\hat{m}_k\right)^{1-\gamma_k^j} + \left(1-\hat{p}\right)\prod_{k=1}^{K}\left(\hat{u}_k\right)^{\gamma_k^j}\left(1-\hat{u}_k\right)^{1-\gamma_k^j}}$$

M-step

- Maximize the complete data likelihood function

$$\hat{m}_k = \frac{\sum_{j=1}^{N} P\left(M \mid \gamma^j\right)\gamma_k^j}{\sum_{j=1}^{N} P\left(M \mid \gamma^j\right)}$$

$$\hat{u}_k = \frac{\sum_{j=1}^{N} P\left(U \mid \gamma^j\right)\gamma_k^j}{\sum_{j=1}^{N} P\left(U \mid \gamma^j\right)}$$

$$\hat{p} = \frac{\sum_{j=1}^{N} P\left(M \mid \gamma^j\right)}{N}$$

Convergence

- Alternate E and M steps
- Compute the change in the complete data likelihood function
- Stop when the change in the complete data likelihood function is small