

Stat 111 Spring 2011 Week 11: Random Effects Models and the EM algorithm

1. Random Effects Models

Often a study involves multiple measurements, with several different x values, on each of k individuals (e.g., mice) who are selected to be representatives of some larger population. To account for individual differences, one could add $k - 1$ indicator variables to a multiple regression model and fit $k - 1$ additional β 's. This would allow estimates of characteristics specific to the individuals who ended up in the study. If we aren't interested in these particular individuals, we could instead just estimate the variance τ^2 of individual effects in the population, and use this variance estimate to adjust the other β estimates for the differences among individuals. This saves $k - 2$ degrees of freedom, and explicitly models the notion that these individuals were randomly selected from some larger population.

$$Y = X\beta + \mathbf{M}\alpha + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2\mathbf{I}); \quad \alpha \sim N_k(0, \tau^2\mathbf{I}).$$

The $k \times 1$ vector of α 's represents the individual differences from the overall mean. The $n \times k$ matrix \mathbf{M} multiplies α to give an $n \times 1$ vector that matches the appropriate element of α to each of the n measurements. That is, $\mathbf{M}_{ij} = 1$ if measurement i is on individual j , and 0 otherwise.

- a) Write out the joint density function for Y and α , given β , σ^2 and τ^2 . Integrate out α to find the marginal likelihood function for β , σ^2 and τ^2 . Explain the idea behind "Restricted Maximum Likelihood Estimation" (REML) for the variance components, and demonstrate for the simpler cases where you have data $Y_i | \mu, \sigma^2 \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, or $Y_i | \beta, \sigma^2 \stackrel{\text{indep}}{\sim} N(X_i'\beta, \sigma^2)$.
 - b) Fit a random effects model for data on the time it takes different chimps to learn different words in American sign language. Show how the random effects model is a compromise between the model that assumes no chimp differences, and the model where each chimp is modeled using a "fixed effect" (our usual indicator variable method for fitting different group means).
 - c) Show how the significance of a main effect variable may change considerably when using random effects as opposed to fixed effects.
2. Describe the EM algorithm and prove that applying the E and M steps will increase the log-likelihood function (except at the mode). Demonstrate this for the introductory example in the Dempster, Laird and Rubin EM paper, and for a simple version of the Normal hierarchical model:

$$\text{Level-1: } Y_i | \theta_i \stackrel{\text{indep}}{\sim} N(\theta_i, V), \quad i = 1, \dots, k.$$

$$\text{Level-2: } \theta_i | A \stackrel{\text{i.i.d.}}{\sim} N(\mu, A).$$

For example, Level-1 might represent the variability in average test scores for samples of n students from each of k schools, and level-2 the variability in the overall mean scores for the schools. For now, assume μ and V are known. The goal is to find the maximum likelihood estimate of A , the variance of the level-2 means, which may be thought of as "missing data."

3. Poisson-Gamma Hierarchical Model

Suppose you have Poisson counts for k different individuals. For example, the numbers of turnovers committed by each of k basketball players in the NCAA tournament. Consider the following model:

$$\begin{aligned}\text{Level-1: } & Y_i | \theta_i \sim \text{Poisson}(m_i \theta_i), \quad i = 1, \dots, k. \\ \text{Level-2: } & \theta_i | \mu \sim \text{Gamma}(\alpha, \alpha/\mu).\end{aligned}$$

The m_i 's represent different "exposures", which might be minutes played in the basketball example. For now, treat α as known.

- Derive the marginal probability function $P(Y_i = y_i | \mu)$, and write out the marginal likelihood function for μ .
- Find the conditional density $f_{\theta|y,\mu}(\theta_i | \mu, y_i)$, and use this to implement an EM algorithm to find the MLE of μ . Demonstrate for real or simulated data.
- Consider the generalization where μ might differ for different individuals (e.g., players of different positions) depending on covariates X_i .

Problems to turn in.

- For NFL data going back to 2003, the logistic regression fit for estimating the probability p_i of a home win ($y_i = 0$ or 1) from the Vegas point spread (x_i) is close to

$$\text{logit}(p_i) = 0.00 + 0.14x_i.$$

- The odds of the home team winning is $p_i/(1 - p_i)$. Based on the fitted model, how do the odds change if x_i increases by 1, or if x_i decreases by 1?
 - Find the estimated odds and the estimated probability of a home win corresponding to $x_i = -6$, $x_i = 0$, and $x_i = 6$.
- You could define a logistic regression model with no intercept to get the model

$$Y_i | x_i \sim \text{Bin}(1, p_i); \quad \text{logit}(p_i) = x_i \beta.$$

In this case, x_i and β are both scalars. Suppose you have $n = 10$ values with $x' = (-9, -7, -5, -3, -1, 1, 3, 5, 7, 9)$.

- Work out the Newton Raphson method for finding the MLE for β .
- Explain why the algorithm will not converge if you observe $y' = (0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$ (arranged to match the corresponding x_i 's).
- Suppose you observe $y' = (0, 0, 0, 0, 1, 1, 1, 0, 1, 1)$. Make a graph of the likelihood function for β , and show that your algorithm converges to the mode. Give your sequence of estimates β_t and the corresponding log-likelihood values for the first six iterations, starting from $\beta_{(0)} = 0$. You may use any program you like.
- Find the p -value for testing $H_o : \beta = 0$ vs. $H_a : \beta \neq 0$.
- Using a Normal approximation to the sampling distribution of $\hat{\beta}$, find a 90% CI for β . (Note - the CI does include the value of β I used to generate these y 's.)