FALL 2013
Stat 8003: Statistical Methods I
Lecture 13

Jichun Xie

1 Marginal Effect and Adjusted Effect

Under many circumstances, we are interested in investigating the effect of some variable X on Y. Under linear models, the effect of X on Y can be measured by marginal effect or adjusted effect. We discuss these two types of effects.

To facilitate the discussion, let's assume $\mathbf{Y}$, $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are centered, so that $\bar{Y} = 0$ and $\bar{X}_i = 0$, for $i = 1, \ldots, p$. Then consider the linear model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon,$$

with $\mathtt{E}(\epsilon) = 0$ and $\mathtt{Var}(\epsilon) = \sigma^2$. It can be shown that the least square estimator $\hat{\beta}_0 = 0$. (You need to know why.)

Remark: The centering procedure is very important. Without centering, some of the formulations below need to be altered.

1.1 Marginal Effect

Suppose the independent variable of interest is X_p, and the dependent variable is Y. Consider the following marginal linear model:

$$Y = \alpha_0^* + \alpha_p^* X_p + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \sigma^2)$. There are n *i.i.d.* observations of Y and X_p. Here α_p is called the marginal effect of X_p on Y.

If X_p is a random variable and independent of ϵ. Suppose $\mathtt{Var}(X_p) = \sigma_p^2$. Then

$$\mathtt{Cor}(Y, X_p) = \frac{\mathtt{Cov}(Y, X_p)}{\{\mathtt{Var}(X_p)\mathtt{Var}(Y)\}^{1/2}} = \frac{\alpha_p}{\{\alpha_p^2 + \sigma^2/\sigma_p^2\}^{1/2}}.$$

Note that under this case, $\texttt{Var}(Y) = \alpha_p^2\sigma_p^2 + \sigma^2$. Therefore, if both Y and X_p are standardized (make $\texttt{Var}(Y) = \texttt{Var}(X_p) = 1$), then $\texttt{Cor}(Y, X_p) = \alpha_p$. In other words, α_p measures the marginal correlation between Y and X_p.

Of course, under most circumanstances, we think X_p is fixed. Even though it is not fixed, we can consider all of the inference on the space that conditioning on the value of X_p. So we don't need to consider the variation of X_p.

If we use the least-square-estimates to estimate α_p, then

$$\hat{\alpha}_p = \frac{\sum_{k=1}^n (X_{k,p} - \bar{X}_p)(Y_k - \bar{Y})}{\sum_{k=1}^n (X_{k,p} - \bar{X}_p)^2} = \frac{\sum_{k=1}^n X_{k,p} Y_k}{\sum_{k=1}^n X_{k,p}^2} = (\mathbf{X}_p^{\mathrm{T}} \mathbf{X}_p)^{-1} \mathbf{X}_p^{\mathrm{T}} \mathbf{Y}.$$

If we use matrix formulation, the fitted value $\hat{\mathbf{Y}} = \mathbf{X}_p \hat{\alpha}_p = \mathbf{P}_{X_p} \mathbf{Y}$.

1.2 Adjusted Effect

Suppose there are other independent variables $X_1, \ldots, X_{p-1}$. These variables are not of interest (the effects of these variables on Y are not of interest), shall we include these variables in the linear model?

Suppose the true model is

$$Y = \beta_0 + X_1\beta_1 + \ldots + X_{p-1}\beta_{p-1} + X_p\beta_p + \epsilon, \tag{2}$$

where $\epsilon \sim N(0, \sigma^2)$. There are n *i.i.d.* observations of $Y, X_1, \ldots X_p$.

In this model, β_p is called the adjusted effect of X_p on Y. More specifically, we can say β_p is the effect of X_p on Y, adjusted by $X_1, \ldots, X_{p-1}$.

Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. Based on the least-squares estimation, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, where $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$. (Why?) It can be shown that (you need to know why this is true)

$$\hat{\beta}_p = (\mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{Y}.$$

Here $\mathcal{M} = \{1, \ldots, p-1\}$, and thus $\mathbf{P}_{X_{\mathcal{M}}}$ is the projection matrix projecting to the column space spanned by $\{X_1, \ldots, X_{p-1}\}$.

1.3 Model Mis-specification

Now we discuss the consequences of model mis-specification.

- Case 1: If the true model is (1), what will happen if we assume the model is (2)?
- Case 2: If the true model is (2), what will happen if we assume the model is (1)?

1.3.1 Case 1

It is easy to see that model (2) is a generalization of model (1). Under case 1, $\beta_0 = \alpha_0$, $\beta_1 = \ldots = \beta_{p-1} = 0$, and $\beta_p = \alpha_p$. Then

$$\mathtt{E}(\hat{\beta}_p) = \mathtt{E}\left\{(\mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})(\alpha_0\mathbf{1} + \alpha_p\mathbf{X}_p + \boldsymbol{\epsilon})\right\} = \alpha_p.$$

Under case 1, even though the model is misspecified, $\hat{\beta}_p$ is still unbiased estimator of α_p.

The trouble is, for model (1) we don't know $\beta_1, \ldots, \beta_{p-1}$ are equal to zero in practice. We need to estimate them. That step might bring some extra noise. Now let's compare the variance of $\hat{\alpha}_p$ and $\hat{\beta}_p$.

Given X_p, it is easy to see that $\mathtt{Var}(\hat{\alpha}_p) = \sigma^2(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1} = \sigma^2(|\mathbf{X}_p|_2^2)^{-1}$.

On the other hand, $\mathtt{Var}(\hat{\beta}_p) = \sigma^2\{\mathbf{X}_p^{\mathrm{T}}(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p\}^{-1} = \sigma^2\{|(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p|_2^2\}^{-1}$.

Note that $(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p$ is the prejection of X_p onto the space orthogonal to the column space spanned by $\mathbf{X}_{\mathcal{M}}$. And therefore $|(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p|_2^2 \leq |\mathbf{X}_p|_2^2$. Consequently, $\mathtt{Var}(\hat{\beta}_p) \geq \mathtt{Var}(\hat{\alpha}_p)$. The equal sign can be taken if and only if $\mathbf{X}_p$ is orthogonal to $\mathbf{X}_{\mathcal{M}}$.

To sum up, under case 1, $\mathtt{MSE}(\hat{\beta}_p) = \sigma^2\{|(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p|_2^2\}^{-1}$, and $\mathtt{MSE}(\hat{\alpha}_p) = \sigma^2(|\mathbf{X}_p|_2^2)^{-1}$. The affect of model misspecification can be measured by the difference of MSE:

$$\mathtt{MSE}(\hat{\beta}_p) - \mathtt{MSE}(\hat{\alpha}_p) = \sigma^2\left[\{|(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p|_2^2\}^{-1} - (|\mathbf{X}_p|_2^2)^{-1}\right].$$

1.3.2 Case 2

Under case 2, if we didn't include those important covariates which should be included in the model, then

$$\begin{aligned}\mathtt{E}(\hat{\alpha}_p) = \mathtt{E}\left\{(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}(\beta_0\mathbf{1} + \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right\} \\ = \sum_{i=1}^{p}(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_i\beta_i = \beta_p + \sum_{i\neq p}(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_i\beta_i.\end{aligned}$$

It is easy to see that

$$\mathtt{Bias}(\hat{\alpha}_p) = \sum_{i\neq p}(\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_i\beta_i = (\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_p)^{-1}\mathbf{X}_p^{\mathrm{T}}\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}}.$$

The variances are discussed before, and therefore the affect of model misspecification of case 2 is:

$$\mathtt{MSE}(\hat{\alpha}_p) - \mathtt{MSE}(\hat{\beta}_p) = \boldsymbol{\beta}_{\mathcal{M}}^{\mathrm{T}}\mathbf{X}_{\mathcal{M}}^{\mathrm{T}}\mathbf{P}_{X_p}\mathbf{X}_{\mathcal{M}}\boldsymbol{\beta}_{\mathcal{M}} - \sigma^2\left[\{|(\mathbf{I} - \mathbf{P}_{X_{\mathcal{M}}})\mathbf{X}_p|_2^2\}^{-1} - (|\mathbf{X}_p|_2^2)^{-1}\right].$$

- When $\boldsymbol{\beta}_{\mathcal{M}}$ is large and σ^2 is small, misspecification of case 2 will lead to a huge increase in `MSE`, and therefore harm the estimation.
- However, $\boldsymbol{\beta}_{\mathcal{M}}$ is small and σ^2 is large, misspecification of case 2 might even benifit the fit in terms of decreasing MSE, and therefore benefit the estimation.
- If $\mathbf{X}_{\mathcal{M}}$ is orthogonal to $\mathbf{X}_p$, then the misspecification of case 2 will neither benefit nor harm the estimation.

How do these results help us in real data analysis? Suppose the only variable of interest is X_p. When shall we include other covariates? And when shall we not? Note that it is actually a trade-off between bias and variance.

1.4 An Example

Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The data set contains the following columns:

Variable	Description
M	percentage of males aged 1424 in total state population
So	indicator variable for a southern state
Ed	mean years of schooling of the population aged 25 years or over
Po1	per capita expenditure on police protection in 1960
Po2	per capita expenditure on police protection in 1959
LF	labour force participation rate of civilian urban males in the age-group 14-24
M.F	number of males per 100 females
Pop	state population in 1960 in hundred thousands
NW	percentage of nonwhites in the population
U1	unemployment rate of urban males 1424
U2	unemployment rate of urban males 3539
Wealth	wealth: median value of transferable assets or family income
Ineq	income inequality: percentage of families earning below half the median income
Prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	average time in months served by offenders in state prisons before their first release
Crime	crime rate: number of offenses per 100,000 population in 1960

Figure 1 shows the scatter plot. If the purpose is to only study the adjusted effect of Time on Crime, how shall we choose the variables to be included in the model? For example, you might want to include "So" in the model, since it is correlated with "Crime" but not that correlated with "Time"; on the other hand, you might not want to include "Pop" in the model, since it seems not very correlated with "Crime" but has a strong correlation with "Pop".

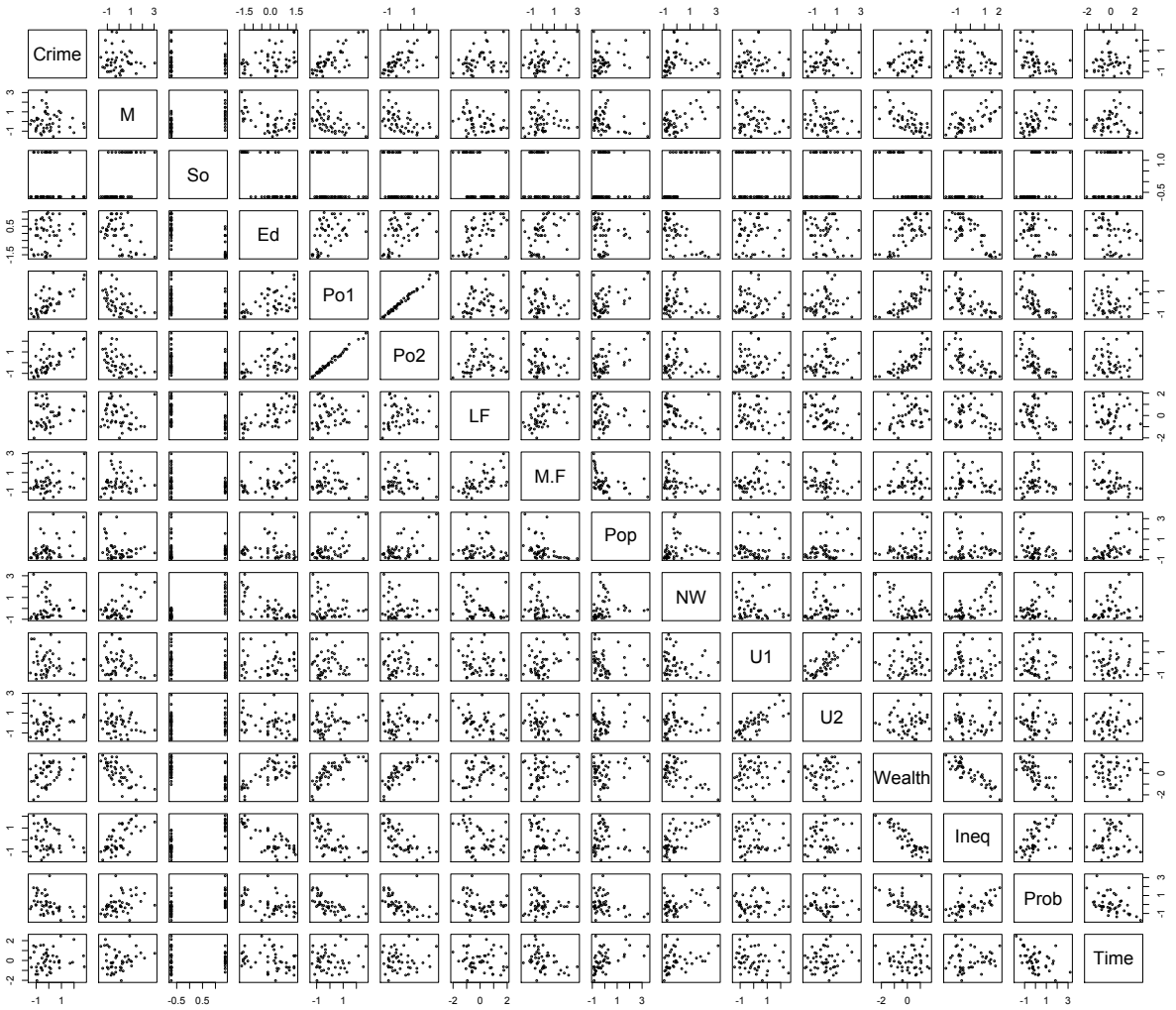

Figure 1: US Crime Scatter Plot

2 EM Algorithm

2.1 A Mixture Example

Suppose we are carrying out a study in which we believe that we have a mixture of two populations and that the outcome from the two populations represent two distributions, say $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$. For example, we might be carrying out a public health study of the size of babies at a hospital in a poor urban Philadelphia neighborhood with the thought that women who are illegal immigrants have poorer nutrition, and thus smaller babies compared to non-immigrants. However, women who are illegal immigrants will generally not reveal that information because of fear of deportation. How to estimate the proportion of the illegal immigrants? And how to estimate the parameters of the new-born baby size distribution?

Consider a two-component mixture Model (Figure 2):

$$\begin{aligned} Y_1 &\sim N(\mu_1, \sigma_1^2) \\ Y_2 &\sim N(\mu_2, \sigma_2^2) \\ Y &= (1-\Delta)\cdot Y_1 + \Delta \cdot Y_2, \end{aligned}$$

where $\Delta \in \{0, 1\}$ with $\mathtt{P}(\Delta = 1) = \pi$. Let $\boldsymbol{\theta}_s = (\mu_s, \sigma_s)$, $s = 1, 2$, and $\boldsymbol{\theta} = (\pi, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. How to estimate $\boldsymbol{\theta}$?

2.1.1 Situation 1: Y and Δ observed

Now let's consider an ideal case: every women in the hospital will reveal their immigration status. How to estimate the unknown parameters?

Note that the log-likelihood function of Y and Δ are

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\Delta}) &= \sum_{k=1}^{n} \log\left[\{(1-\pi)\phi_{\boldsymbol{\theta}_1}(y_k)\}^{1-\Delta_k}\{\pi\phi_{\boldsymbol{\theta}_2}(y_k)\}^{\Delta_k}\right] \\ &= \sum_{k=1}^{n}\{(1-\Delta_k)\log\phi_{\boldsymbol{\theta}_1}(y_k) + \Delta_k \log\phi_{\boldsymbol{\theta}_2}(y_k)\} \\ &\quad + \sum_{k=1}^{n}\{(1-\Delta_k)\log(1-\pi) + \Delta_k \log\pi\} \end{aligned} \tag{3}$$

Take derivative and set to zero, we can get MLE of θ. The solution is actually very straightforward. Let $n_1 = n - \sum_{k=1}^{n}\Delta_k$ and $n_2 = \sum_{k=1}^{n}\Delta_k$.

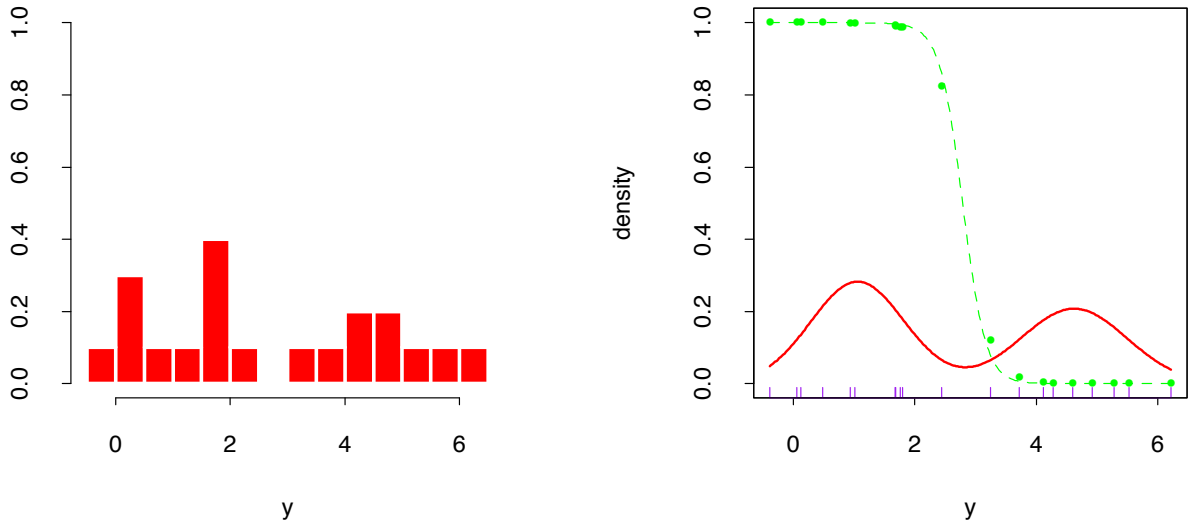

FIGURE 8.5. *Mixture example. (Left panel:) Histogram of data. (Right panel:) Maximum likelihood fit of Gaussian densities (solid red) and responsibility (dotted green) of the left component density for observation y, as a function of y.*

Figure 2: Two-Component Mixture Model

$$\hat{\pi} = \sum_{k=1}^{n} \Delta_k / n$$

$$\hat{\mu}_1 = \frac{\sum_{k=1}^{n}(1-\Delta_k)y_k}{n_1}$$

$$\hat{\sigma}_1 = \frac{1}{n_1}\sum_{k=1}^{n}\{(1-\Delta_k)y_k - \hat{\mu}_1\}^2$$

$$\hat{\mu}_2 = \frac{\sum_{k=1}^{n}\Delta_k y_k}{n_2}$$

$$\hat{\sigma}_1 = \frac{1}{n_2}\sum_{k=1}^{n}\{\Delta_k y_k - \hat{\mu}_2\}^2$$

If Both realizations of Y and Δ are observed, we can easily separate two samples, and obtain MLE based on each sample.

2.1.2 Situation 2: Y observed, but Δ not observed

Now Δ_k, $k = 1, \ldots, n$, are unknown, substituting for each Δ_k in the full log-likelihood (3) by

$$\hat{\gamma}_k(\theta) = \widehat{\mathtt{E}}(\Delta_k \mid \boldsymbol{\theta}, \mathbf{Z}) = \widehat{\mathtt{P}}(\Delta_k = 1 \mid \boldsymbol{\theta}, \mathbf{Z}) = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_k)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_k) + \hat{\pi}\phi_{\hat{\theta}_2}(y_k)},$$

where $\mathbf{Z} = (\mathbf{Y}, \boldsymbol{\Delta})$ are the full data. $\gamma_k(\theta)$ is called the the *responsibility* of model 2 for observation k.

EM Algorithm for Two-component Gaussian Mixtures. (Figure 3)

1. Take initial guesses for the parameters $\hat{\theta}$.
2. *Expectation Step:* compute the responsibilities $\hat{\gamma}_k$.
3. *Maximization Step:* compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{k=1}^n (1-\hat{\gamma}_k) y_k}{\sum_{k=1}^n (1-\hat{\gamma}_k)}, \quad \hat{\sigma}_1^2 = \frac{\sum_{k=1}^n (1-\hat{\gamma}_k)(y_k - \hat{\mu}_1)^2}{\sum_{k=1}^n (1-\hat{\gamma}_k)},$$
$$\hat{\mu}_2 = \frac{\sum_{k=1}^n \hat{\gamma}_k y_k}{\sum_{k=1}^n \hat{\gamma}_k}, \quad \hat{\sigma}_2^2 = \frac{\sum_{k=1}^n \hat{\gamma}_k (y_k - \hat{\mu}_2)^2}{\sum_{k=1}^n \hat{\gamma}_k},$$

   and the mixing probability is $\hat{\pi} = \sum_{k=1}^n \hat{\gamma}_k / n$.

4. Iterate steps 2 and 3 until convergence.

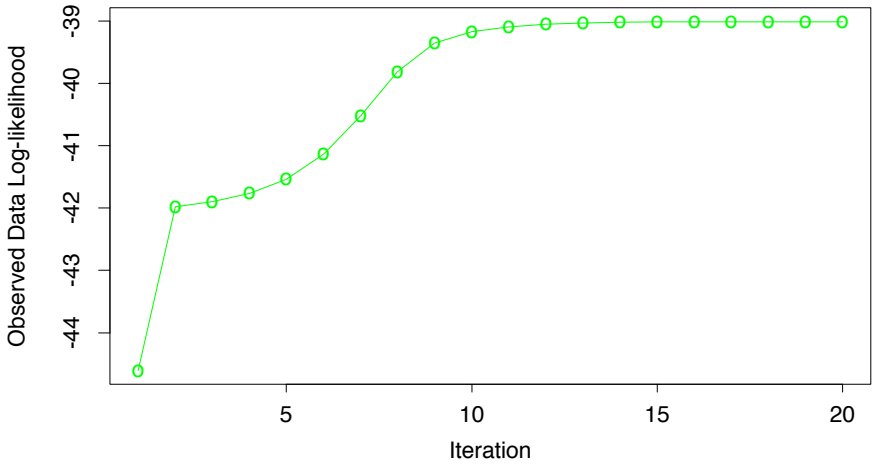

FIGURE 8.6. *EM algorithm: observed data log-likelihood as a function of the iteration number.*

Figure 3: Iterations of EM Algorithm

2.2 The EM Algorithm in General

Let $\mathbf{T} = (\mathbf{Z}, \mathbf{Z}^m)$ to be the compete data.

The EM Algorithm.

1. Start with initial guesses for the parameters $\hat{\theta}^{(0)}$.
2. *Expectation Step:* at the jth step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathtt{E}(l_0(\theta'; \mathbf{T}) \mid \mathbf{Z}, \hat{\theta}^{(j)})$$

   as a function of the dummy argument θ'.
3. *Maximization Step:* determine the new estimate $\hat{\theta}^{(j+1)}$ as the maximizer of $Q(\theta', \hat{\theta}^{(j)})$ over θ'.
4. Iterate steps 2 and 3 until convergence.

It can be proved that the EM algorithm converges to local maximum. We will show the details in Methods III class.

2.2.1 Two-Component Mixture Example

Now let's see how to use the general framework to solve the two-component Gaussian mixture problem.

Note that in (3), the full data is $(\mathbf{y}, \boldsymbol{\Delta})$, where $\mathbf{y}$ is observed and $\boldsymbol{\Delta}$ is missing.

Suppose at the beginning of the jth iteration, the initial value is $\hat{\boldsymbol{\theta}}^{(j-1)}$.

Expectation step. We calcuate $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(j-1)})$ by

$$\begin{aligned} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(j-1)}) &= \mathtt{E}\left\{ l(\boldsymbol{\theta}; \mathbf{y}, \boldsymbol{\Delta}) \mid \mathbf{y}, \hat{\boldsymbol{\theta}}^{(j-1)} \right\} \\ &= \sum_{k=1}^{n} \left\{ (1 - \hat{\gamma}_k^{(j-1)}) \log \phi_{\boldsymbol{\theta}_1}(y_k) + \hat{\gamma}_k^{(j-1)} \log \phi_{\boldsymbol{\theta}_2}(y_k) \right\} \\ &\quad + \sum_{k=1}^{n} \left\{ (1 - \hat{\gamma}_k^{(j-1)}) \log(1 - \pi) + \hat{\gamma}_k^{(j-1)} \log \pi \right\} \end{aligned}$$

Here

$$\hat{\gamma}_k^{(j-1)} = \mathtt{E}(\Delta_k \mid \mathbf{y}, \hat{\boldsymbol{\theta}}^{(j-1)}) = \frac{\hat{\pi}^{(j-1)} \phi_{\hat{\boldsymbol{\theta}}_2^{(j-1)}}(y_k)}{(1 - \hat{\pi}_k^{(j-1)}) \phi_{\hat{\boldsymbol{\theta}}_1^{(j-1)}}(y_k) + \hat{\pi}^{(j-1)} \phi_{\hat{\boldsymbol{\theta}}_2^{(j-1)}}(y_k)}$$

Maximization step. Update $\hat{\boldsymbol{\theta}}^{(j)}$ by maximizing $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(j-1)})$ with respect to $\boldsymbol{\theta}$.

$$\hat{\mu}_1^{(j)} = \frac{\sum_{k=1}^n (1-\hat{\gamma}_k^{(j-1)}) y_k}{\sum_{k=1}^n (1-\hat{\gamma}_k^{(j-1)})}, \quad \hat{\sigma}_1^{2,(j)} = \frac{\sum_{k=1}^n (1-\hat{\gamma}_k^{(j-1)})(y_k - \hat{\mu}_1^{(j-1)})^2}{\sum_{k=1}^n (1-\hat{\gamma}_k^{(j-1)})},$$
$$\hat{\mu}_2^{(j)} = \frac{\sum_{k=1}^n \hat{\gamma}_k^{(j-1)} y_k}{\sum_{k=1}^n \hat{\gamma}_k^{(j-1)}}, \quad \hat{\sigma}_2^{2,(j)} = \frac{\sum_{k=1}^n \hat{\gamma}_k^{(j-1)} (y_k - \hat{\mu}_2^{(j-1)})^2}{\sum_{k=1}^n \hat{\gamma}_k^{(j-1)}},$$

and the mixing probability is $\hat{\pi}^{(j)} = \sum_{k=1}^n \hat{\gamma}_k^{(j-1)}/n$.

2.2.2 Another Example: Positron Emission Tomography

Positron Emission Tomography (PET) is performed by introducing a radioactive tracer into an animal or human subject. Radioactive emissions are then used to assess levels of metabolic activity and blood flow in organs of interest. Positrons emitted by the tracer annihilate with nearby electrons, giving pairs of photons that fly off in opposite directions. Some of these are counted by bands of gamma ray detectors placed around the subjects body, but the others (photons) miss the detectors. The detected counts are used to form an image of the level of metabolic activity in the organs based on the estimate spatial concentration of the isotope.

For a statistical model, the region of interest is divided into n pixels or voxels and it is assumed that the number of emmissions U_{ij} from the jth pixel detected at the ith detector is a Poisson random variable with mean $p_{ij}\lambda_j$; here λ_j is the intensity of emissions from that pixel and p_{ij} is the probability that a single emission is detected at the ith detector: The $\lambda_1, \ldots, \lambda_n$ are unknown and our parameters of interest; the p_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$ depend on the geometry of the detection system, the isotope and other factors, but can be taken to be known. The U_{ij} are unknown but can be plausibly be assumed independent. The counts Y_i $(i = 1, \ldots d)$ at the ith detector are observed and have independent Poisson distributions with mean $\sum_{j=1}^n p_{ij}\lambda_j$.

What are the maximum likelihood estimates for emission intensities $\lambda_1, \ldots, \lambda_n$ based on the observed data $Y_1, \ldots, Y_d$ and the known detection probabilities p_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$?

Consider viewing the U_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, d$ as missing data. The complete data is then

$$(\mathbf{u}, \mathbf{y}) = (\{u_{ij}, i = 1, \ldots, n, j = 1, \ldots, d\}, \{y_1, \ldots, y_n\}),$$

where $y_i = \sum_{j=1}^n u_{ij}$. The complete data log likelihood is

$$l_{\mathbf{u},\mathbf{y}} = \sum_{i=1}^d \sum_{j=1}^n \{u_{ij} \log(p_{ij}\lambda_j) - p_{ij}\lambda_j\}.$$

This is an exponential family. The expected complete data log liklihood, where the missing data follows its conditional distribution given the observed data and the current parameter estimates of $\hat{\lambda}_1^{(k-1)}, \ldots, \hat{\lambda}_n^{(k-1)}$, is

$$Q\left\{(\lambda_1,\ldots,\lambda_n),(\hat{\lambda}_1^{(j-1)},\ldots,\hat{\lambda}_n^{(k-1)})\right\}$$
$$=\sum_{i=1}^{d}\sum_{j=1}^{n}\left\{\mathtt{E}(u_{ij}\mid y_1,\ldots,y_n,\hat{\lambda}_1^{(k-1)},\ldots,\hat{\lambda}_n^{(k-1)})\log(p_{ij}\lambda_j)-p_{ij}\lambda_j\right\}$$

To complete the E step, we need to calculate

$$\mathtt{E}(U_{ij}\mid y_1,\ldots,y_d,\hat{\lambda}_1^{(k-1)},\ldots,\hat{\lambda}_n^{(k-1)}).$$

As $Y_i=\sum_{j=1}^{n}U_{ij}$, the conditional density of U_{ij} given $Y_i=y_i$ is binomial with y_i trials and probability of success

$$\frac{p_{ij}\hat{\lambda}_j^{(k-1)}}{\sum_{h=1}^{n}p_{ih}\lambda_h^{(k-1)}};$$

thus,

$$\mathtt{E}(U_{ij}\mid y_1,\ldots,y_d,\hat{\lambda}_1^{(k-1)},\ldots,\hat{\lambda}_n^{(k-1)})=\frac{p_{ij}\hat{\lambda}_j^{(k-1)}}{\sum_{h=1}^{n}p_{ih}\hat{\lambda}_h^{(k-1)}}.$$

The M step yields

$$\begin{aligned}\hat{\lambda}_j^{(k)}=&\frac{\sum_{i=1}^{d}\mathtt{E}(U_{ij}\mid y_1,\ldots,y_d,\hat{\lambda}_1^{(k-1)},\ldots,\hat{\lambda}_n^{(k-1)})}{\sum_{i=1}^{d}p_{ij}}\\ =&\frac{\sum_{i=1}^{d}(p_{ij}\hat{\lambda}_j^{(k-1)})/\{\sum_{h=1}^{n}p_{ih}\hat{\lambda}_h^{(k-1)}\}}{\sum_{i=1}^{d}p_{ij}}\\ =&\hat{\lambda}_j^{(k-1)}\frac{1}{\sum_{i=1}^{d}p_{ij}}\sum_{i=1}^{d}\frac{y_ip_{ij}}{\sum_{h=1}^{n}\hat{\lambda}_h^{(k-1)}p_{ih}},\quad j=1,\ldots,n.\end{aligned}$$

It has been shown (Vardi *et al.*, 1985, *Journal of the American Statistical Association*) that $\hat{\lambda}_1^{(k)},\ldots,\hat{\lambda}_n^{(k)}$ converges to the maximizer of the likelihood function of $\lambda_1,\ldots,\lambda_n$ given the observed data $y_1,\ldots,y_n$ and the known detection probabilities $p_{ij},i=1,\ldots,n,j=1,\ldots,d$ but there will not be a unique MLE if the number of pixels n is greater than the number of detectors d.

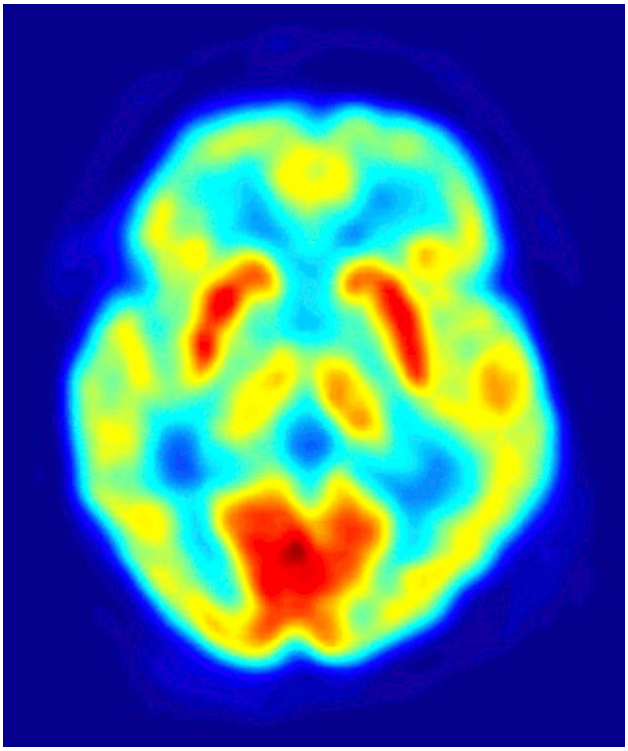

Figure 4: PET Scan of the Human Brain