Lecture 16: Hierarchical models and miscellanea

It is often easier to model a practical situation by thinking of things in a hierarchy.

Example 4.4.1 (binomial-Poisson hierarchy)

- An insect lays many eggs, each surviving with probability p.
- On the average, how many eggs will survive?
- Let Y be the number of eggs and X be the number of survivors; both are random variables.
- We can model this situation by first modeling the distribution of Y; given Y, we then model the distribution of $X|Y$.
- We can then obtain the joint distribution of (X, Y) and marginal distributions of X and Y.
- We can model the number of eggs by a Poisson distribution, i.e., $Y \sim Poisson(\lambda)$, where $\lambda > 0$ is the average of eggs.
- Given Y, we can model the number of survivors as $X|Y \sim binomial(p, Y)$.

- What is the marginal distribution of X?
  For $x = 0, 1, 2, \ldots,$

$$\begin{aligned} P(X=x) &= \sum_{y=x}^{\infty} P(X=x, Y=y) \qquad\qquad x \le y \\ &= \sum_{y=x}^{\infty} P(X=x|Y=y)P(Y=y) \\ &= \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \frac{e^{-\lambda}\lambda^y}{y!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{[(1-p)\lambda]^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{[(1-p)\lambda]^{t}}{t!} \qquad\qquad t = y-x \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} e^{(1-p)\lambda} = \frac{e^{-p\lambda}(\lambda p)^x}{x!} \end{aligned}$$

The distribution of X is $Poisson(p\lambda)$!

- Then, $E(X) = p\lambda$, which can also be obtained using

$$E(X) = E[E(X|Y)] = E(pY) = pE(Y) = p\lambda$$

without using the marginal distribution of X.

- If we begin with our model by saying that $X \sim Poisson(\theta)$, then $\theta = p\lambda$ with Y playing no role at all. Introducing Y in the hierarchy was mainly aid our understanding of the model.

Example (binomial-binomial hierarchy)

A very similar hierarchical model can be described as follows.

- A market survey is conducted to study whether a new product is preferred over the product currently available in the market (old product).
- The survey is conducted by mail. Questionnaires are sent along with the sample products (both new and old) to N customers randomly selected from a population.
- Each customer is asked to fill out the questionnaire and return it, with response 1 (new is better than old) or 0 (otherwise).

- Let X be the number of ones in the returned questionnaires. What is the distribution of X?
- If every customer returns the questionnaire, then (from elementary probability) $X \sim binomial(p, N)$ (assuming that the population is large enough so that customers respond independently), where $p \in (0,1)$ is the overall rate of customers who prefer the new product.
- Some customers, however, do not return the questionnaires. Let Y be the number of customers who respond.
- If customers respond independently with the same probability $\pi \in (0,1)$, then $Y \sim binomial(\pi, N)$.
- Given $Y = y$ (an integer between 0 and N), $X|Y = y \sim (p, y)$ if $y \geq 1$ and the point mass at 0 if $y = 0$.
- For $x = 0, 1, ..., N$,

$$P(X = x) = \sum_{k=x}^{N} P(X = x, Y = k) = \sum_{k=x}^{N} P(X = x|Y = k)P(Y = k)$$

$$
\begin{aligned}
&= \sum_{k=x}^{N} \binom{k}{x} p^x (1-p)^{k-x} \binom{N}{k} \pi^k (1-\pi)^{N-k} \\
&= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \sum_{k=x}^{N} \binom{N-x}{k-x} \left(\frac{\pi - \pi p}{1-\pi p} \right)^{k-x} \left(\frac{1-\pi}{1-\pi p} \right)^{N-k} \\
&= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x}
\end{aligned}
$$

It turns out that the marginal distribution of X is the *binomial*$(\pi p, N)$ distribution.

- Hierarchical models can have more than two stages.
- The advantage is that complicated processes may be modeled by a sequence of relatively simple models placed in a hierarchy.
- Conditional distributions play a central role.
- The random variables in hierarchical models may be all discrete (as in our previous examples), all continuous, or some discrete and some continuous.

More general joint pdf's and conditional distributions

So far we have considered only the situation where all random variables are continuous or all are discrete.

What if X is a continuous random variable and Y is a discrete random variable on $\mathscr{Y}$?

We can define the joint "pdf" to be a function $f(x,y)$ satisfying

$$P(X \leq x, Y = y) = \int_{-\infty}^{x} f(t,y)dt, \qquad x \in \mathscr{R},\ y \in \mathscr{Y}$$

Then the marginal pdf of X and pmf of Y are respectively

$$f_X(x) = \sum_{y \in \mathscr{Y}} f(x,y) \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

The conditional pdf of $X|Y$ or pmf of $Y|X$ can be defined as before.

Similarly we can deal with the situation where X is a continuous random vector and Y is a discrete random vector.

If X or Y is neither discrete nor continuous, we can still define conditional distributions. But it involves higher level mathematics.

Example.

Suppose that X is a continuous random variable and Y is a discrete random variable with $\mathscr{Y} = \{0, 1\}$, and the joint pdf is

$$f(x,y) = \begin{cases} \frac{\alpha}{\sqrt{2\pi}} e^{-x^2/2} & y = 0, \\ \frac{1-\alpha}{2} e^{-|x|} & y = 1, \end{cases} \qquad x \in \mathscr{R}$$

where $0 < \alpha < 1$. Then

$$f_X(x) = f(x,0) + f(x,1) = \frac{\alpha}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1-\alpha}{2} e^{-|x|} \qquad x \in \mathscr{R}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \begin{cases} \alpha & y = 0 \\ 1 - \alpha & y = 1 \end{cases}$$

$$f(y|x) = \frac{f(x,y)}{f_X(x)} = \begin{cases} \frac{\alpha}{\sqrt{2\pi}} e^{-x^2/2} \Big/ \left(\frac{\alpha}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1-\alpha}{2} e^{-|x|} \right) & y = 0 \\ \frac{1-\alpha}{2} e^{-|x|} \Big/ \left(\frac{\alpha}{\sqrt{2\pi}} e^{-x^2/2} + \frac{1-\alpha}{2} e^{-|x|} \right) & y = 1 \end{cases}$$

$$f(x|y) = \frac{f(x,y)}{f_Y(y)} = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} & y = 0 \\ \frac{1}{2} e^{-|x|} & y = 1 \end{cases} \qquad x \in \mathscr{R}$$

Mixture distributions

In the previous example, the pdf f_X is referred to as a mixture distribution (or pdf), since it is a convex combination of two distributions (pdf's).

Example 4.4.5 (three-stage hierarchy)

Consider a generalization of Example 4.4.1, where instead of one mother insect there are a large number of mothers and one mother is chosen at random from a (possibly continuous) distribution G.

The following three-stage hierarchy may be more appropriate:

number of survivors X from Y eggs	$X\|Y \sim binomial(p, Y)$
number of eggs Y from a given mother Z	$Y\|Z \sim Poisson(Z)$
a particular mother Z	$Z \sim G$

Suppose that $Z \sim exponential(\beta)$. Then

$$E(X) = E[E(X|Y)] = pE(Y) = pE[E(Y|Z)] = pE(Z) = p\beta$$

The three-stage model can be thought of as a two-stage model by combining the last two stages; we just need to obtain the marginal distribution of Y.

$$\begin{aligned}
P(Y = y) &= \int_0^\infty f(y|z) f_Z(z) dz = \int_0^\infty \frac{e^{-z} z^y}{y!} \frac{1}{\beta} e^{-z/\beta} dz \\
&= \frac{1}{\beta y!} \int_0^\infty z^y e^{-z(1+\beta^{-1})} dz = \frac{1}{\beta y!} \Gamma(y+1) \frac{1}{(1+\beta^{-1})^{y+1}} \\
&= \frac{1}{\beta} \frac{\beta^{y+1}}{(1+\beta)^{y+1}} = \frac{1}{1+\beta} \left(1 - \frac{1}{1+\beta}\right)^y
\end{aligned}$$

which is the geometric distribution with probability $(1+\beta)^{-1}$, i.e.,

number of survivors X from Y eggs	$X|Y \sim binomial(p, Y)$
number of eggs Y	$Y \sim geometric((1+\beta)^{-1})$

However, the three-stage model is easier to understand.

Calculation of mean and variance

Aside from the advantage in understanding models, hierarchical models can often make the calculation of mean and variance easier.

Let X be a random variable having a pdf that is the noncentral chi-square pdf with noncentrality parameter λ and degrees of freedom n:

$$f_X(x) = \sum_{k=1}^{\infty} \frac{x^{n/2+k-1} e^{-x/2}}{\Gamma(n/2+k) 2^{n/2+k}} \frac{(\lambda/2)^k e^{-\lambda/2}}{k!} \qquad x > 0$$

Calculating $E(X)$ is directly using this pdf not easy.

However, the pdf is the marginal pdf of X in the following two-stage model:

$$\begin{aligned} X|K &\sim \text{chi square with degrees of freedom } n+2K \\ K &\sim Poisson(\lambda/2) \end{aligned}$$

Then

$$\begin{aligned} E(X) &= E[E(X|K)] = E(n+2K) = n+2E(K) = n+\lambda \\ \mathrm{Var}(X) &= E[\mathrm{Var}(X|K)] + \mathrm{Var}(E(X|K)) = E(2n+4K) + \mathrm{Var}(n+2K) \\ &= 2n+4E(K) + 4\mathrm{Var}(K) = 2n+4\lambda \end{aligned}$$

Example 4.4.6.

Suppose that

$$X|P \sim binomial(n, P), \qquad P \sim beta(\alpha, \beta)$$

How to calculate $E(X)$ and $\mathrm{Var}(X)$?

$$E(X) = E[E(X|P)] = E(nP) = nE(P) = \frac{n\alpha}{\alpha+\beta}$$

Using the variance formula derived previously,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(E(X|P)) + E[\mathrm{Var}(X|P)] \\
&= \mathrm{Var}(nP) + E[nP(1-P)] \\
&= n^2\mathrm{Var}(P) + nE(P) - nE(P^2) \\
&= \frac{n^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} + \frac{n\alpha}{\alpha+\beta} - \frac{n\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\
&= \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}$$

Theorem 4.7.9 (Covariance inequality)

Let X be a random variable and g and h be nondecreasing functions such that $E[g(X)]$, $E[h(X)]$, and $E[g(X)h(X)]$ exist. Then

$$\mathrm{Cov}(g(X), h(X)) \geq 0.$$

Proof.

Let $Z = h(X)$, $\mu_z = E[h(X)]$, and $h^{-1}(t) = \inf\{x : h(x) \geq t\}$.

Then $h(x) \geq \mu_z$ iff $x \geq h^{-1}(\mu_z)$, and

$$\begin{aligned}
& \mathrm{Cov}(g(X), h(X)) \\
= & \, E[\{g(X) - E[g(X)]\}(Z - \mu_z)] = E[g(X)(Z - \mu_z)] \\
= & \, E[g(X)(Z - \mu_z)I(\{Z < \mu_z\})] + E[g(X)(Z - \mu_z)I(\{Z \geq \mu_z\})] \\
\geq & \, E[g(h^{-1}(\mu_z))(Z - \mu_z)I(\{Z < \mu_z\})] \\
& + E[g(h^{-1}(\mu_z))(Z - \mu_z)I(\{Z \geq \mu_z\})] \\
= & \, E[g(h^{-1}(\mu_z))(Z - \mu_z)] = g(h^{-1}(\mu_z))E(Z - \mu_z) \\
= & \, 0
\end{aligned}$$

Best prediction

Let X be a random variable on with $E(X^2) < \infty$

We want to predict the future value of X by constructing a $g(Y)$, where Y is another random variable currently observed.

We now show that $E(X|Y)$ is the best predictor of X in the sense that

$$E[X - E(X|Y)]^2 = \min_{g:E[g(Y)]^2<\infty} E[X - g(Y)]^2.$$

$$\begin{aligned}
E[X-g(Y)]^2 &= E[X - E(X|Y) + E(X|Y) - g(Y)]^2 \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{[X - E(X|Y)][E(X|Y) - g(Y)]\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\big\{E\{[X - E(X|Y)][E(X|Y) - g(Y)]|Y\}\big\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{[E(X|Y) - g(Y)]E[X - E(X|Y)|Y]\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\geq E[X - E(X|Y)]^2
\end{aligned}$$

In some applications, the concept of independence of X and Y is not enough.

There are situations in which X and Y are not independent, but if some information regarding another random vector Z is given, then X and Y are conditionally independent.

Definition (conditional Independence)

Let X, Y, and Z be random vectors. X and Y are conditionally independent given Z iff

$$F(x,y|z) = F(x|z)F(y|z) \qquad \text{for all } x,\ y,\ z$$

where $F(x,y|z)$ is the cdf of $(X,Y)|Z=z$, $F(x|z)$ is the cdf of $X|Z=z$, and $F(y|z)$ is the cdf of $Y|Z=z$.

Example

Let Z, U, and V be independent random variables, $X = Z+U$, and $Y = Z+V$.

Then X and Y are not independent because of the common term Z.

For example, if $E(XY)$ exists, then

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{Cov}(Z+U, Z+V) \\ &= \mathrm{Cov}(Z,Z)+\mathrm{Cov}(Z,U)+\mathrm{Cov}(Z,V)+\mathrm{Cov}(U,V) \\ &= \mathrm{Var}(Z) > 0\end{aligned}$$

But conditioned on $Z = z$, X and Y are independent because $X = z + U$ and $Y = z + V$ and U and V are independent.

A rough proof is the following.

$$\begin{aligned}F(x-z, y-z|z) &= F(u,v|z) = F(u,v) = F(u)F(v) \\ &= F(u|z)F(v|z) = F(x-z|z)F(y-z|z)\end{aligned}$$

This holds for all x, y, and z, and hence

$$F(x,y|z) = F(x|z)F(y|z)$$