

Example 1: Normal Model

Suppose that we observe random variables Y_i , $i = 1, \dots, n$, and we assume that

$$Y_i | \mu \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, 1), \quad i = 1, \dots, n \quad (1)$$

The notation of (1) indicates that conditional on a quantity μ , the Y_i are independent and identically distributed as normal with mean μ and variance 1. This defines (for a given μ) a joint distribution for Y_1, Y_2, \dots, Y_n , so we can speak of a joint probability density associated with any vector of possible values for Y_1, \dots, Y_n :

$$f(y_1, \dots, y_n | \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y_i - \mu)^2\right), \quad (2)$$

where the product form follows from the i.i.d. assumption, which can be simplified to:

$$= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^2\right]. \quad (3)$$

We will suppose that μ is not known but is assumed to have a distribution with density $p(\mu)$. For concreteness, suppose $p(\mu)$ is a normal density with mean a and variance v . It will simplify some of the formulas if we work instead with the inverse of the variance $\tau := 1/v$. This is called the *precision*. Then

$$p(\mu) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\left(-\frac{\tau}{2}(\mu - a)^2\right). \quad (4)$$

If we regard $p(\mu)$ as a marginal density for μ , and $f(y_1, \dots, y_n | \mu)$ as a conditional density for the data, this specifies a joint distribution:

$$p(\mu, y_1, \dots, y_n) = p(\mu)f(y_1, \dots, y_n | \mu),$$

which in turn defines the conditional distribution:

$$p(\mu | y_1, \dots, y_n) = \frac{p(\mu, y_1, \dots, y_n)}{p(y_1, \dots, y_n)} = \frac{p(\mu)f(y_1, \dots, y_n | \mu)}{p(y_1, \dots, y_n)}. \quad (5)$$

Here we are using $p(\cdot)$ to refer to a number of different density functions; which one is meant can be inferred by examining its arguments.

Notice that the denominator of (5) is a constant (in μ), and can be obtained by integrating the

numerator with respect to μ :

$$p(y_1, \dots, y_n) = \int p(\mu, y_1, \dots, y_n) d\mu = \int p(\mu) f(y_1, \dots, y_n | \mu) d\mu$$

Intuitively, *the denominator is whatever constant c that makes*

$$p(\mu | y_1, \dots, y_n) = \frac{1}{c} \cdot p(\mu) f(y_1, \dots, y_n | \mu)$$

a proper density function, in the sense of integrating to 1. Because c is a multiplicative factor which does not depend on μ , we can simplify the notation by using “ \propto ” to mean “proportional to,” and write

$$p(\mu | y_1, \dots, y_n) \propto p(\mu) f(y_1, \dots, y_n | \mu)$$

With some algebra, this can be shown to be proportional to:

$$\exp\left(-\frac{1}{2}(\tau + n)(\mu - \mu^*)^2\right), \tag{6}$$

where

$$\mu^* := \frac{\tau}{\tau + n} a + \frac{n}{\tau + n} \left(\frac{1}{n} \sum_{i=1}^n y_i\right)$$

The expression in (6) is proportional to (equal to a constant times) the normal density for μ with mean μ^* and precision $(\tau + n)$. We conclude that:

$$\mu | Y_1 = y_1, \dots, Y_n = y_n \quad \sim \quad \mathcal{N}(\mu^*, (\tau + n)^{-1}).$$

Notice that μ^* , the conditional mean of μ , is a weighted average of a and the sample mean $\sum_{i=1}^n y_i$, with the weight depending on τ and n .

Suppose we desire to provide a *point estimate* d of μ , that minimizes the quadratic loss $(\mu - d)^2$. Since μ is regarded as having a distribution rather than being fixed, we could choose d to minimize

$$\mathbb{E}((\mu - d)^2 | y_1, \dots, y_n),$$

where the expectation is with respect to the conditional distribution of μ . The best choice will generally depend on (y_1, \dots, y_n) , so the optimal policy can be regarded as a function $d(y_1, \dots, y_n)$. Solving for the optimum:

$$d(y_1, \dots, y_n) = \mathbb{E}(\mu | y_1, \dots, y_n) = \mu^*.$$

Notice that as $\tau \rightarrow 0$ we get the “usual” estimator \bar{y} .

This notion of optimality applies *if* the mean μ does in fact have the specified normal distribution. What if instead μ is fixed, but unknown and allowed to take any value in \mathbb{R} ? In later lectures, we will study optimality in this case and show that in a certain sense, the procedure just described

still does pretty well.

The marginal density $p(\mu)$ is usually called a *prior* density, and $p(\mu|y_1, \dots, y_n)$ is called a *posterior* density. So far we have not really discussed how to interpret or choose a prior, but we will have more to say about this in later lectures. Equation (5) is a special case of Bayes' Theorem; hence the term Bayesian analysis.

Example 2: Asset Allocation with Estimation Risk

This example is based on Barberis (2000). Consider an investor trying to decide how much of her wealth to allocate among two assets whose returns are uncertain, say a stock index and a bond index. Denote by y_{t1} the log return on the stock index, and y_{t2} the log return on the bond index, at time t .

At time T , the investor has wealth $W_T = 1$ and must choose how much to place in each asset. Let $a \in [0, 1]$ denote the allocation to the stock index (we are ruling out short sales), and for simplicity suppose that the investor has a buy-and-hold strategy: she chooses the portfolio allocation a at time T , and holds on to it for H periods. Wealth in period $T + H$ is thus given by

$$W_{T+H} = a \exp(y_{T+1,1} + \dots + y_{T+H,1}) + (1 - a) \exp(y_{T+1,2} + \dots + y_{T+H,2}).$$

If the investor has isoelastic utility:

$$U(W) = \frac{W^{1-\gamma}}{1-\gamma},$$

then it seems reasonable for her to choose a in order to maximize expected utility

$$\mathbb{E} \left[\frac{(a \exp(y_{T+1,1} + \dots + y_{T+H,1}) + (1 - a) \exp(y_{T+1,2} + \dots + y_{T+H,2}))^{1-\gamma}}{1-\gamma} \right], \quad (7)$$

where the expectation is with respect to the distribution of future returns. We will discuss expected utility theory in more detail in lecture note 3, but for now assume that this is a reasonable thing to do. This raises the question, which distribution to use?

We could assume that the vector of returns $y_t := (y_{t1}, y_{t2})'$ has a multivariate normal distribution

$$y_t | \mu, \Sigma \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma). \quad (8)$$

Assuming that μ and Σ are given, the investor could proceed by evaluating (7) for different values of a , and choose the allocation that gives the highest expected utility. What happens if, instead, there is uncertainty (“estimation risk”) concerning μ and Σ ?

If the investor has access to data on earlier returns, y_1, \dots, y_T , this could be useful for predicting future returns. By analogy with the previous section, we could choose a prior $p(\mu, \Sigma)$ for the

parameters, and use Bayes' Theorem to obtain the posterior $p(\mu, \Sigma | y_1, \dots, y_T)$. (In later lectures we will derive an explicit expression for the posterior density, but for now assume that this can be done.) However, we are no longer directly concerned with the parameters, but instead would like to place a distribution on future returns y_{T+1}, \dots, y_{T+H} that incorporates uncertainty about the parameters. Notice that

$$p(y_{T+1}, \dots, y_{T+H} | y_1, \dots, y_T) = \int p(y_{T+1}, \dots, y_{T+H}, \mu, \Sigma | y_1, \dots, y_T) d(\mu, \Sigma). \quad (9)$$

This is called a *predictive density*. We can decompose the density inside the integral, so this can be written

$$= \int p(y_{T+1}, \dots, y_{T+H} | \mu, \Sigma, y_1, \dots, y_T) p(\mu, \Sigma | y_1, \dots, y_T) d(\mu, \Sigma).$$

A further simplification is possible, since by (8) the distribution of y_t conditional on μ, Σ is independent of past returns:

$$= \int p(y_{T+1}, \dots, y_{T+H} | \mu, \Sigma) p(\mu, \Sigma | y_1, \dots, y_T) d(\mu, \Sigma).$$

The first term inside the integral is a product of multivariate normal densities, and the second term is the posterior density for the parameters.

It is important to recognize that the distribution of y_{T+1}, \dots, y_{T+H} given μ, Σ , is independent of y_1, \dots, y_T , but it is not independent of the past data if we remove the conditioning on the (unknown) μ, Σ . From the perspective of the investor, who does not know μ, Σ , the return process can display persistence, even though the "true" process does not. If this were not the case, there would be no value to observing the historical return data y_1, \dots, y_T for making forecasts about future returns.

Treatment Assignment

Many studies have examined the impact of job-training and other social programs on economic outcomes of individuals. Dehejia (2005) considers data from a social experiment, the GAIN program in California. This was a randomized experiment comparing standard AFDC (welfare) to a treatment which consisted of education, job training, and job search activities. Dehejia argues that instead of focusing on simply estimating an average treatment effect, it may be useful to consider the problem of a social planner, who can choose to assign individuals to either of the two options based on the individual's background characteristics.

Let $T_i = 1$ denote that individual i received the GAIN program and let $T_i = 0$ denote receipt of the standard AFDC program. The outcome of interest is individual earnings in various quarters after the program. Since many welfare recipients had zero earnings, Dehejia used a Tobit model.

A simplified version of Dehejia's model is:

$$Y_i = \max\{0, \alpha'_1 X_i + \alpha_2 T_i + \alpha'_3 X_i \cdot T_i + \epsilon_i\},$$

where the ϵ_i are IID $N(0, \sigma^2)$. We assume that individuals in the data are a random sample from some population, so that (X_i, T_i) are IID. Dehejia estimated this model using the n experimental subjects, and then produced predictive distributions for a hypothetical $(n + 1)$ th subject to assess different ways of assigning treatments.

To put this in a similar framework as the previous examples, let the parameter vector be $\theta = (\alpha_1, \alpha_2, \alpha_3, \sigma^2)$. For a given individual i , the probability of observing zero earnings is:

$$\begin{aligned} Pr(Y_i = 0 | X_i = x_i, T_i = t_i, \theta) &= Pr(\alpha'_1 x_i + \alpha_2 t_i + \alpha'_3 x_i \cdot t_i + \epsilon_i \leq 0 | x_i, t_i, \theta) \\ &= Pr\left(\frac{\epsilon_i}{\sigma} \leq -\frac{\alpha'_1 x_i + \alpha_2 t_i + \alpha'_3 x_i \cdot t_i}{\sigma} \mid x_i, t_i, \theta\right) \\ &= \Phi\left(\frac{\alpha'_1 x_i + \alpha_2 t_i + \alpha'_3 x_i \cdot t_i}{\sigma}\right), \end{aligned}$$

where Φ is the standard normal CDF.

If $Y_i > 0$, its density is

$$\phi(y_i | \alpha'_1 x_i + \alpha_2 t_i + \alpha'_3 x_i \cdot t_i, \sigma^2),$$

where $\phi(y | \mu, \sigma^2)$ is the PDF function of a normal random variable with mean μ and variance σ^2 .

So we can write the single-observation conditional likelihood function as:

$$f(y_i | x_i, t_i, \theta) = \Phi\left(\frac{\alpha'_1 x_i + \alpha_2 t_i + \alpha'_3 x_i \cdot t_i}{\sigma}\right)^{1(y_i=0)} \times \phi(y_i | \alpha'_1 x_i + \alpha_2 t_i + \alpha'_3 x_i \cdot t_i, \sigma^2)^{1(y_i>0)}.$$

The joint (conditional) likelihood based on the n observations in the experimental data can be written as

$$f(y_1, \dots, y_n | x_1, \dots, x_n, t_1, \dots, t_n, \theta) = \prod_{i=1}^n f(y_i | x_i, t_i, \theta).$$

If we have a prior distribution for the parameters θ , say $p(\theta)$, then the posterior can be viewed as:

$$p(\theta | y_1, \dots, y_n, x_1, \dots, x_n, t_1, \dots, t_n) \propto p(\theta) f(y_1, \dots, y_n | x_1, \dots, x_n, t_1, \dots, t_n, \theta).$$

Now suppose that a social planner wants to assign a new individual $(n + 1)$ to one of the two treatments, based on observing their covariates X_{n+1} . For simplicity, suppose that the social planner's objective is to choose treatment t to maximize expected income $E[Y_{n+1} | X_{n+1}, T_{n+1} = t]$.

If the social planner knew θ , then she could compare

$$E[Y_{n+1} | X_{n+1}, T_{n+1} = 1, \theta] \quad \text{vs.} \quad E[Y_{n+1} | X_{n+1}, T_{n+1} = 0, \theta]$$

and pick whichever treatment gave higher expected earnings.

If θ is not known, but the social planner has access to the data ($i = 1, \dots, n$), then the planner could form a predictive distribution for Y_{n+1} given X_{n+1} and the past data.

Exercise: show that

$$f(y_{n+1}|x_{n+1}, t_{n+1}, y_1, \dots, y_n, x_1, \dots, x_n, t_1, \dots, t_n) = \int f(y_{n+1}|x_{n+1}, t_{n+1}, \theta)p(\theta|y_1, \dots, y_n, x_1, \dots, x_n, t_1, \dots, t_n).$$

With this predictive distribution, the planner can form two expectations for Y_{n+1} (corresponding to the two possible values of T_{n+1}), and assign the individual to the treatment that gives higher expected earnings (with respect to the predictive distribution). Note that this predictive distribution will depend on the prior $p(\theta)$, because the posterior distribution $p(\theta|y_1, \dots, y_n, x_1, \dots, x_n, t_1, \dots, t_n)$ depends on $p(\theta)$.