LECTURE 20

§22. Simple single-parameter models.

Next we will consider some simple single-parameter models. Let us first assume that $y = (x_1, \dots, x_n)$ is a sample from a normal distribution with unknown mean θ and known variance σ^2. The likelihood is then

$$p(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\theta)^2} \propto e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2} \propto e^{-\frac{n}{2\sigma^2}(\theta-\overline{X})^2},$$

where as usual

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Since the likelihood has the form

$$p(x_1, x_2, ..., x_n|\theta) \propto e^{-\frac{n}{2\sigma^2}(\overline{X}-\theta)^2} \propto \mathcal{N}\left(\overline{X}/\theta, \frac{\sigma^2}{n}\right)$$

by replacing σ^2/n with τ_0^2, and $\overline{X}$ with μ_0, we find a conjugate prior

$$p(\theta) \propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2},$$

which is $\mathcal{N}(\mu_0, \tau_0^2)$.

With this prior the posterior becomes

$$p(\theta|x_1, x_2, ..., x_n) \propto p(\theta)\, p(x_1, x_2, ..., x_n|\theta) \propto e^{-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2} e^{-\frac{n}{2\sigma_0^2}(\theta-\overline{X})^2} \propto$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)\left(\theta^2 - 2\,\frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{X}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\,\theta\right)\right\} \propto \exp\left\{-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right\} \propto \mathcal{N}\left(\theta/\mu_n, \tau_n^2\right),$$

where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{X}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \tau_n^2 = \left(\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}\right)^{-1} = \frac{1}{\dfrac{1}{\tau_0^2} + \dfrac{n}{\sigma^2}}.$$

Thus, the posterior distribution is $N(\mu_n, \tau_n^2)$. The inverse of variance is called precision. We see that posterior precision = prior precision + data precision, where the prior precision is $1/\tau_0^2$ and data precision n/σ^2 (the inverse of the variance of the sample mean).

The posterior mean is a weighted average of the prior mean μ_0 and sample mean $\overline{X}$ where the weights are the corresponding precisions. When $n \to \infty$ (or when $\tau_0^2 \to \infty$), the role of the prior information vanishes. Thus, for large values of n, approximately $\theta|x_1, x_2, ..., x_n \sim N(\overline{X}, \sigma^2/n)$.

The Poisson distribution is often used to model rare incidents, such as traffic accidents or rare diseases. For a vector $y = (y_1, \ldots, y_n)$ of iid observation, the likelihood is

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!} e^{-\theta} \propto \theta^{\sum y_i} e^{-n\theta}.$$

Given that the prior distribution is $\Gamma(\alpha, \beta)$, the posterior

$$p(\theta|y) \propto p(\theta)\, p(y|\theta) \propto \theta^{\alpha-1}\, e^{-\beta\theta}\, \theta^{\sum y_i} e^{-n\theta} \propto \theta^{\alpha+\sum y_i - 1} e^{-(\beta+n)\theta}$$

is $\Gamma(\alpha + \sum y_i, \beta + n)$.

The negative binomial distribution. When the prior and posterior distributions can be written in closed form, the marginal likelihood $p(y)$ can be computed using the formula

$$p(y) = \frac{p(y|\theta)\, p(\theta)}{p(\theta|y)}.$$

For example, if y is a single observation from $Poi(\theta)$, then

$$p(y) = \frac{\frac{\theta^y}{y!} e^{-\theta} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\, \theta^{\alpha-1}\, e^{-\beta\theta}}{\frac{(\beta+1)^{\alpha+y}}{\Gamma(\alpha+y)} \theta^{\alpha+y-1}\, e^{-(\beta+1)\theta}} = \binom{\alpha+y-1}{y} \left(\frac{\beta}{\beta+1}\right)^{\alpha} \left(\frac{1}{\beta+1}\right)^{y},$$

which is Neg-Bin(α, β), the negative binomial distribution.

On the other hand,

$$p(y) = \int p(y|\theta)\, p(\theta)\, d\theta = \int Poi(y|\theta)\, \Gamma(\theta|\alpha, \beta)\, d\theta,$$

implying that the negative binomial distribution is compound distribution where the Poisson distribution is compounded using the Gamma distribution as a weight distribution.

§23. SIMPLE MULTIPARAMETER MODELS

Next we consider simple models having more than one parameter. The natural extension of the previous lecture in which the variance σ^2 was considered known is to consider the more realistic case in which the variance is not known. In reality, we typically do not know σ^2 any more than we know μ, and thus we have two quantities of interest that we should be updating with new information.

A full probability model for μ and σ^2 would look like (posterior distribution):

$$p(\mu, \sigma^2/x_1, x_2, ..., x_n) = \frac{f(x_1, x_2, ..., x_n/\mu, \sigma^2)p(\mu, \sigma^2)}{\int f(x_1, x_2, ..., x_n/\mu, \sigma^2)p(\mu, \sigma^2)d\mu\, d\sigma^2} \propto f(x_1, x_2, ..., x_n/\mu, \sigma^2)p(\mu, \sigma^2).$$

This model is similar to the one in the previous lecture, we have now explicitly noted that σ^2 is also unknown quantity, by including it in the prior distribution. Therefore, we now need to specify a joint prior for both μ and σ^2, and not just a prior for μ. We assume μ and σ^2 are independent. There is no reason the two parameters need be related, then we can consider

$$p(\mu, \sigma^2) = p(\mu)\, p(\sigma^2)$$

and establish separate prior for each.

In the previous lecture we established the prior for μ to be $N(\mu_0, \tau_0^2)$, where μ_0 was the prior mean and τ_0^2 was a measure of uncertainty we had in μ.

Recal from Central Limit Theorem that sample mean $\bar{X} \sim N(\mu, \sigma^2/n)$, thus σ^2 and τ_0^2 are related: $\dfrac{\sigma^2}{n}$ should be an estimate for σ_0^2, and so treating σ^2 as fixed yields an updated τ_0^2 that depends heavily on the new sample data.

Let us assume that $x = (x_1, ..., x_n)$ is a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown.

How do we specify a prior distribution for μ and σ^2 in a more general case? There are several ways to do this in the normal distribution problem, but two of the most common approaches lead not the same prior distribution. One approach is to assign a uniform prior the real line for μ and the same uniform prior for $\ln(\sigma^2)$. We assign a uniform prior on $\ln(\sigma^2)$, because σ^2 is a nonnegative quantity, and the transformation to $\ln(\sigma^2)$ stretches this new parameter across the real line. If we transform the uniform prior distribution on $\ln(\sigma^2)$ into a density for σ^2, we obtain $p(\sigma^2) \propto \dfrac{1}{\sigma^2}$. Thus our joint prior distribution is

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

A second way to obtain this prior is to give μ and σ^2 proper prior distributions (not uniform over the real line, which is improper). If we continue with the assumption that $\mu \sim N(\mu_0, \sigma_0^2)$, we can choose values of μ_0 and σ_0^2 that yield the flat distribution. We can also choose a relatively noninformative prior for σ^2 by first noting that variance parameters follow an inverse gamma distribution and then choosing values for the inverse gamma distribution that produce a noninformative prior. IF $\sigma^2 \sim IG(\alpha, \beta)$, the density function appears as:

$$f(x, \alpha, \beta) \propto (\sigma^2)^{-(\alpha+1)} \exp -\beta/x.$$

In the limit, if we let parameters α and β approach 0, a noninformative prior distribution is obtained $\dfrac{1}{\sigma^2}$. Strictly speaking, however if α and β are 0, the distribution is improper, but we can let both parameters approach 0. We can then use this as our prior distribution for σ^2 (that is, $\sigma^2 \sim IG(0, 0)$; $p(\sigma^2) \propto \dfrac{1}{\sigma^2}$).

If the joint prior is $p(\mu, \sigma^2) \propto 1/\sigma^2$, or equivalently $p(\mu, log(\sigma^2)) \propto 1$, the posterior is

$$p(\mu, \sigma^2 | x_1, x_2, ..., x_n) \propto \frac{1}{\sigma^2} \times \frac{1}{(\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$= \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\,\sigma^2} \left[\sum_{i=1}^{n} (x_i - \overline{X})^2 + n(\overline{X} - \mu)^2\right]\right)$$

$$= \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\,\sigma^2} [(n-1)s^2 + n(\overline{X} - \mu)^2]\right)$$

where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{X})^2$ is the sample variance.

The marginal posterior of σ^2 is obtained by integrating μ out:

$$p(\sigma^2 | y) \propto \int_{-\infty}^{\infty} \frac{1}{(\sigma^2)^{n/2+1}} \exp\left(-\frac{1}{2\,\sigma^2} [(n-1)s^2 + n(\overline{y} - \mu)^2]\right) d\mu.$$

The integral of the factor $\exp\left(-\frac{1}{2\,\sigma^2}n(\overline{y}-\mu)^2\right)$ is a simple normal integral, so

$$
\begin{aligned}
p(\sigma^2|y) &\propto \frac{1}{(\sigma^2)^{n/2+1}}\,\exp\left(-\frac{1}{2\,\sigma^2}(n-1)s^2\right)\,\sqrt{2\pi\sigma^2/n} \\
&\propto \frac{1}{(\sigma^2)^{(n+1)/2}}\,\exp\left(-\frac{(n-1)s^2}{2\,\sigma^2}\right).
\end{aligned}
$$

This is a scaled inverse-χ^2 -density:

$$\sigma^2|y \sim Inv-\chi^2(n-1,s^2).$$

Thus, $\{(n-1)s^2/\sigma^2|y\} \sim \chi^2_{n-1}$. This is analogous with the corresponding sampling theory result. However, in sampling theory, s^2 is considered random, while here σ^2 is random. By making the substitution

$$z=\frac{A}{\sigma^2},\quad \text{where}\quad A=(n-1)s^2+n(\overline{y}-\mu)^2,$$

we obtain the marginal density of μ:

$$
\begin{aligned}
p(\mu|y) &\propto \int_0^\infty \frac{1}{(\sigma^2)^{n/2+1}}\,\exp\left(-\frac{1}{2\,\sigma^2}[(n-1)s^2+n(\overline{y}-\mu)^2]\right)\,d\sigma^2 \\
&\propto A^{-n/2}\int_0^\infty z^{n/2-1}\exp(-z)\,dz \propto \left[1+\frac{n(\mu-\overline{y})^2}{(n-1)s^2}\right]^{-n/2}.
\end{aligned}
$$

This is the $t_{n-1}(\overline{y},s^2/n)$ density. Thus, $\{(\mu\overline{y})/(s/\sqrt{n})|y\} \sim t_{n-1}$. This is again analogous to the sampling theory result. It can also be shown (exercise) that the density of a new observation $\tilde{y}$ is $t_{n-1}(\overline{y},s^2(1+1/n))$. The posterior can be simulated using $p(\sigma^2|y)$ and $p(\mu|\sigma^2,y)=N(\mu|\overline{y},\sigma^2/n)$.