STAT 801: Mathematical Statistics

Bayesian estimation

In this section I will focus on the problem of estimation of a 1 dimensional parameter, θ. Earlier we discussed comparing estimators in terms of Mean Squared Error. In the language of decision theory Mean Squared Error corresponds to using

$$L(d,\theta) = (d-\theta)^2$$

which is called squared error loss. The multivariate version would be

$$L(d,\theta) = ||d-\theta||^2$$

or possibly the more general formula

$$L(d,\theta) = (d-\theta)^T \mathbf{Q}(d-\theta)$$

for some positive definite symmetric matrix $\mathbf{Q}$. The risk function of a procedure (estimator) $\hat\theta$ is

$$R_{\hat\theta}(\theta) = E_\theta[(\hat\theta-\theta)^2].$$

Now consider prior with density $\pi(\theta)$. The Bayes risk of $\hat\theta$ is

$$\begin{aligned} r_\pi &= \int R_{\hat\theta}(\theta)\pi(\theta)d\theta \\ &= \int\int (\hat\theta(x)-\theta)^2 f(x;\theta)\pi(\theta)dxd\theta \end{aligned}$$

For a Bayesian the problem is then to choose $\hat\theta$ to minimize r_π? This problem will turn out to be analogous to the calculations I made when I minimized $\beta+\lambda\alpha$ in hypothesis testing. First recognize that $f(x;\theta)\pi(\theta)$ is really a joint density

$$\int\int f(x;\theta)\pi(\theta)dxd\theta = 1$$

For this joint density: conditional density of X given θ is just the model $f(x;\theta)$. This justifies the standard notation $f(x|\theta)$ for $f(;\theta)$¿ Now I will compute r_π a different way by factoring the joint density a different way:

$$f(x|\theta)\pi(\theta) = \pi(\theta|x)f(x)$$

where now $f(x)$ is the marginal density of x and $\pi(\theta|x)$ denotes the conditional density of θ given X. We call $\pi(\theta|x)$ the **posterior density** of θ given the data $X=x$. This posterior density may be found via Bayes' theorem (which is why this is Bayesian statistics):

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\phi)\pi(\phi)d\phi}$$

With this notation we can write

$$r_\pi(\hat\theta) = \int \left[\int (\hat\theta(x)-\theta)^2 \pi(\theta|x) d\theta\right] f(x)dx$$

[REMEMBER the meta-theorem: when you see a double integral it is always written in the wrong order. Change the order of integration to learn something useful.] Notice that by writing the integral in this order you see that you can choose $\hat\theta(x)$ separately for each x to minimize the quantity in square brackets (as in the NP lemma).

The quantity in square brackets is a quadratic function of $\hat\theta(x)$; it is minimized by

$$\hat\theta(x) = \int \theta\pi(\theta|x)d\theta$$

which is

$$E(\theta|X)$$

and is called the **posterior expected mean** of θ.

Example: estimating normal mean μ.

Imagine, for example that μ is the true speed of sound.

I think this is around 330 metres per second and am pretty sure that I am within 30 metres per second of the truth with that guess. I might summarize my opinion by saying that I think μ has a normal distribution with mean ν =330 and standard deviation $\tau = 10$. That is, I take a prior density π for μ to be $N(\nu, \tau^2)$.

Before I make any measurements my best guess of μ minimizes

$$\int (\hat\mu - \mu)^2 \frac{1}{\tau\sqrt{2\pi}} \exp\{-(\mu-\nu)^2/(2\tau^2)\} d\mu$$

This quantity is minimized by the prior mean of μ, namely,

$$\hat\mu = E_\pi(\mu) = \int \mu\pi(\mu)d\mu = \nu\,.$$

Now collect 25 measurements of the speed of sound. Assume: the relationship between the measurements and μ is that the measurements are unbiased and that the standard deviation of the measurement errors is $\sigma = 15$ which I assume that we know. So model is: given μ, $X_1,\ldots,X_n$ are iid $N(\mu,\sigma^2)$ variables.

The joint density of the data and μ is then

$$(2\pi)^{-n/1}\sigma^{-n} \exp\{-\sum (X_i-\mu)^2/(2\sigma^2)\} \times (2\pi)^{-1/2}\tau^{-1}\exp\{-(\mu-\nu)^2/\tau^2\}.$$

Thus $(X_1,\ldots,X_n,\mu) \sim MVN$. Conditional distribution of θ given $X_1,\ldots,X_n$ is normal. We can now use standard MVN formulas to calculate conditional means and variances.

Alternatively: the exponent in joint density has the form

$$-\frac{1}{2}\left[\mu^2/\gamma^2 - 2\mu\psi/\gamma^2\right]$$

plus terms not involving μ where

$$\frac{1}{\gamma^2} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)$$

and

$$\frac{\psi}{\gamma^2} = \frac{\sum X_i}{\sigma^2} + \frac{\nu}{\tau^2}$$

So: the conditional distribution of μ given the data is $N(\psi, \gamma^2)$. In other words the posterior mean of μ is

$$\frac{\frac{n}{\sigma^2}\bar{X} + \frac{1}{\tau^2}\nu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

which is a weighted average of the prior mean ν and the sample mean $\bar{X}$.

Notice: the weight on the data is large when n is large or σ is small (precise measurements) and small when τ is small (precise prior opinion).

Improper priors: When the density does not integrate to 1 we can still follow the machinery of Bayes' formula to derive a posterior.
Example: $N(\mu, \sigma^2)$; consider prior density

$$\pi(\mu) \equiv 1.$$

This "density" integrates to ∞; using Bayes' theorem to compute the posterior would give

$$\pi(\mu|X) = \frac{(2\pi)^{-n/2}\sigma^{-n} \exp\{-\sum(X_i - \mu)^2/(2\sigma^2)\}}{\int (2\pi)^{-n/2}\sigma^{-n} \exp\{-\sum(X_i - \xi)^2/(2\sigma^2)\} d\xi}$$

It is easy to see that this cancels to the limit of the case previously done when $\tau \to \infty$ giving a $N(\bar{X}, \sigma^2/n)$ density. That is, the Bayes estimate of μ for this improper prior is $\bar{X}$.

Admissibility: Bayes procedures corresponding to proper priors are admissible. It follows that for each $w \in (0, 1)$ and each real ν the estimate

$$w\bar{X} + (1 - w)\nu$$

is admissible. That this is also true for $w = 1$, that is, that $\bar{X}$ is admissible is much harder to prove.
Minimax estimation: The risk function of $\bar{X}$ is simply σ^2/n. That is, the risk function is constant since it does not depend on μ. Were $\bar{X}$ Bayes for a proper prior this would prove that $\bar{X}$ is minimax. In fact this is also true but hard to prove.
Example: Given p, X has a Binomial(n, p) distribution.

Give p a Beta(α, β) prior density

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1}$$

The joint "density" of X and p is

$$\binom{n}{X} p^X (1-p)^{n-X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\,;$$

posterior density of p given X is of the form

$$cp^{X+\alpha-1}(1-p)^{n-X+\beta-1}$$

for a suitable normalizing constant c.

This is Beta$(X+\alpha, n-X+\beta)$ density. Mean of Beta(α,β) distribution is $\alpha/(\alpha+\beta)$.

So Bayes estimate of p is

$$\frac{X+\alpha}{n+\alpha+\beta} = w\hat{p} + (1-w)\frac{\alpha}{\alpha+\beta}$$

where $\hat{p} = X/n$ is the usual mle.

Notice: again weighted average of prior mean and mle.

Notice: prior is proper for $\alpha > 0$ and $\beta > 0$.

To get $w = 1$ take $\alpha = \beta = 0$; use improper prior

$$\frac{1}{p(1-p)}$$

Again: each $w\hat{p} + (1-w)p_o$ is admissible for $w \in (0,1)$.

Again: it is true that $\hat{p}$ is admissible but our theorem is not adequate to prove this fact.

The risk function of $w\hat{p} + (1-w)p_0$ is

$$R(p) = E[(w\hat{p} + (1-w)p_0 - p)^2]$$

which is

$$w^2\text{Var}(\hat{p}) + (wp + (1-w)p - p)^2 = w^2 p(1-p)/n + (1-w)^2(p-p_0)^2.$$

Risk function constant if coefficients of p^2 and p in risk are 0.

Coefficient of p^2 is

$$-w^2/n + (1-w)^2$$

so $w = n^{1/2}/(1+n^{1/2})$.

Coefficient of p is then

$$w^2/n - 2p_0(1-w)^2$$

which vanishes if $2p_0 = 1$ or $p_0 = 1/2$.

Working backwards: to get these values for w and p_0 require $\alpha = \beta$. Moreover

$$w^2/(1-w)^2 = n$$

gives

$$n/(\alpha+\beta) = \sqrt{n}$$

or $\alpha = \beta = \sqrt{n}/2$. Minimax estimate of p is

$$\frac{\sqrt{n}}{1+\sqrt{n}}\hat{p} + \frac{1}{1+\sqrt{n}}\frac{1}{2}$$

Example: $X_1, \ldots, X_n$ iid $MVN(\mu, \Sigma)$ with Σ known.

Take improper prior for μ which is constant.

Posterior of μ given X is then $MVN(\bar{X}, \Sigma/n)$.

Multivariate estimation: common to extend the notion of squared error loss by defining

$$L(\hat{\theta}, \theta) = \sum (\hat{\theta}_i - \theta_i)^2 = (\hat{\theta} - \theta)^t (\hat{\theta} - \theta)\,.$$

For this loss risk is sum of MSEs of individual components.

Bayes estimate is again posterior mean. Thus $\bar{X}$ is Bayes for an improper prior in this problem.

It turns out that $\bar{X}$ is minimax; its risk function is the constant $trace(\Sigma)/n$.

If the dimension p of θ is 1 or 2 then $\bar{X}$ is also admissible but if $p \geq 3$ then it is inadmissible.

Fact first demonstrated by James and Stein who produced an estimate which is better, in terms of this risk function, for every μ.

So-called **James Stein** estimator is essentially never used.