So far in this course, we have discussed several different methods for estimation. In frequentist statistics, we covered MoM estimators and MLEs, as well as desirable properties estimators may satisfy (such as unbiasedness, consistency, low MSE, etc.). In Bayesian statistics, the goal was to find the posterior distribution, then extract any quantities of interest from it (such as the posterior mean, the MAP estimator, the posterior variance, etc.)

Another approach to estimation uses decision theory.

1 General set-up of decision theory

The main idea of statistical decision theory is to make decisions about random/uncertain phenomena. We encode the information we know/don't know about a problem in a probability distribution F, which is assumed to come from a class of possible distributions $\mathcal{F}$. We can make a "decision" in the problem based on data or an estimate of interest, and F describes the uncertainties in the data or estimate. The "decision" is also known as the "action" a, and the set of all possible actions is denoted by $\mathcal{A}$.

Definition 1.1. Loss Function. A loss function

$$L(F, a) : (\mathcal{F} \times \mathcal{A}) \to [0, \infty)$$

quantifies the consequences of taking an action a when the true state of nature is F.

Large values indicate worse outcomes, so we should make a decision that minimizes the loss.

2 Set-up of decision theory for estimation

In this class, we will focus on decision theory for estimation. In this context, the decision we make is about a "good" value of a parameter θ. That is, the action we take is the estimate

$$a = \hat{\theta}(X),$$

so that $\mathcal{A}$ is a subset of $\mathbb{R}$, $\mathbb{R}^d$, or $\mathbb{Z}$ (for a discrete outcome).

Common loss functions include squared error loss, linear loss, absolute error loss, L^p loss, and zero-one loss, but they are chosen depending on the particular problem.

3 Risk

Notice that the loss function depends on the unknown parameter θ, as well as the random data used in $\hat{\theta}$. To make a decision then, we need to summarize the loss. We might do that by looking at the expected loss.

There are several notions of expected loss we could use:

(a) The **(frequentist) risk** is

$$R(\theta, \hat{\theta}) = \mathbb{E}_\theta \left(L(\theta, \hat{\theta}) \right) = \int L(\theta, \hat{\theta}(x)) f(x; \theta) dx.$$

This averages over the different possible realizations x, and θ is treated as fixed. After calculating this, in general it will be a function of the unknown value θ (it is possible for it to turn into a constant though). From a Bayesian perspective, we can also write that risk as the expected loss conditional on θ:

$$\mathbb{E}(L(\theta, \hat{\theta})|\theta).$$

(b) The **posterior risk** is a Bayesian concept and is defined as

$$r(\hat{\theta}|X) = \mathbb{E}(L(\theta, \hat{\theta}|X)) = \int L(\theta, \hat{\theta}(X)) f(\theta|X) d\theta.$$

This averages over the possible values of θ after conditioning on X. After calculating this, in general it will be a function of X.

(c) The **Bayes risk** is another Bayesian concept and is defined as

$$\begin{aligned} r(f, \hat{\theta}) &= \mathbb{E}(L(\theta, \hat{\theta}) \\ &= \mathbb{E} \left(\mathbb{E}(L(\theta, \hat{\theta})|\theta) \right) = \mathbb{E}(R(\theta, \hat{\theta}) \end{aligned}$$

This averages over both the possible realizations of X as well as the possible values for θ. Note that the averaging can be done in "either order", although typically the second formula will be easiest to use. After calculating the Bayes risk, it will not be a function of the data X or the parameter θ.

4 Comparing Estimators with Risk

4.1 Admissibility

Any of these risks quantify the quality of an estimator $\hat{\theta}$ on average.

Admissibility is a frequentist concept that helps to decide which estimator $\hat{\theta}$ should be chosen. Since it is frequentist, it uses the frequentist risk of course.

Definition 4.1. Inadmissible An estimator $\hat{\theta}$ is inadmissible if there exists another estimator $\tilde{\theta}$ such that

$$R(\theta, \tilde{\theta}) \leq R(\theta, \hat{\theta}) \text{ for all } \theta$$
$$R(\theta, \tilde{\theta}) < R(\theta, \hat{\theta}) \text{ for at least one } \theta.$$

Otherwise, $\hat{\theta}$ is called admissible.

Note that if an estimator is admissible, we shouldn't necessarily be impressed with that. It means that the estimator isn't uniformally worse than other estimators.

4.2 Bayes Risk & Maximum Risk

Admissibility isn't enough to choose among estimators, since we can have multiple admissible estimators that have different performances for different θ. To make a decision, we need a way to summarize the risks. Then we can try to minimize them.

Here are two main ways:

(a) **Bayes risk**: As we saw previously, the Bayes risk does not depend on the parameter θ or the data X. Therefore, it is already a one-number summary. However, it does depend on what prior $f(\theta)$ we use. Minimizing the Bayes risk gives rise to Bayes rules/estimators.

(b) **Maximum risk:** One way to get a one-number summary from the frequentist risk is to take the supremum:

$$\bar{R}(\hat{\theta}) = \sup_{\theta} R(\theta, \hat{\theta}).$$

This measures an estimator's "worst-case" risk over all possible values of θ. Minimizing this maximum risk gives rise to minimax rules/estimators.

5 Bayes Rule & Minimax Estimators

An estimator that minimizes the Bayes risk is called a Bayes rule.

Definition 5.1. Bayes rule. The estimator $\hat{\theta}$ is a Bayes rule, or Bayes estimator, (under a particular prior f and loss function) if

$$r(f, \hat{\theta}) = \inf_{\tilde{\theta}} r(f, \tilde{\theta}).$$

.

Note that we can get a Bayes rule by minimizing the posterior risk! This is tyically the easiest way.

Why does this work? By definition, the Bayes risk is

$$r(f, \hat{\theta}) = \mathbb{E}\left(r(\hat{\theta}|X)\right) = \int r(\hat{\theta}|x) f(x) dx,$$

where $r(\hat{\theta}|x)$ is the posterior risk for a particular x. By a property of integrals, the value of $\hat{\theta}$ that minimizes the integral is the same value of $\hat{\theta}$ that minimizes $r(\hat{\theta}|x)$.

Caution: In the above formula, the integrating is being done with respect to X (although we don't actually have to integrate, as already noted). But when we were calculating Bayes risk, we said it was typically easiest to integrate with respect to θ rather than X. To conclude, the formula that *integrates with respect to* X is typically easiest for *minimizing the Bayes risk*, while the formula that *integrates with respect to* θ is typically easiest for *calculating the Bayes risk*.

An estimator that minimizes the maximum risk is called a minimax rule (or minimax estimator).

Definition 5.2. Minimax Rule. An estimator $\hat{\theta}$ is the minimax rule (or minimax estimator) if

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

In general, it can be difficult to find minimax rules. But there are various properties of rules that imply an estimator is minimax. If we can't prove that an estimator is a minimax estimator, it's often still helpful to bound its minimax risk (so that we know what its "worst-case" behavior would be).

6 Connections between admissibility and the different estimators

There are many properties that can be used to determine different estimators, their risks, and their admissibility. Please see the lecture notes for this.

7 Stein's Paradox

TBD

Problem 1. Calculating Risks

(Example on pg. 5 of lecture notes)

Suppose $\bar{X}_n \sim N(\theta, \frac{1}{n})$, and we are estimating θ under squared error loss. Consider $\hat{\theta}_c(X) = c\bar{X}_n$, where c is chosen ahead of time (non-random). Calculate the (a) frequentist risk, (b) posterior risk, and (c) Bayes risk. Assume for (b) and (c) that the prior is $\theta \sim N(0, b^2)$.

Solution

Useful fact from Lecture 5 notes on Bayesian statistics that we will use: Suppose $Y_1, \ldots, Y_m|\theta \overset{\text{iid}}{\sim} N(\theta, \sigma^2)$ and $\theta \sim N(a, b^2)$, where σ^2, a, and b are known. Then $\theta|Y_1, \ldots, Y_m$ is normally distributed with mean and variance

$$\mu = \frac{mb^2\bar{Y} + \sigma^2 a}{mb^2 + \sigma^2}$$
$$\tau^2 = \frac{\sigma^2 b^2}{mb^2 + \sigma^2}.$$

(a) The risk is

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}_\theta\left((\theta - \hat{\theta})^2\right) \\ &= \mathbb{E}_\theta\left((\theta - c\bar{X})^2\right) \\ \mathbb{E}_\theta\left(\theta^2 - 2c\theta\bar{X} + c^2\bar{X}^2\right) & \\ &= \theta^2 - 2c\theta\mathbb{E}(\bar{X}) + c^2\mathbb{E}(\bar{X}^2). \end{aligned}$$

We have

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \theta \\ \mathbb{E}(\bar{X}^2) &= \text{Var}(\bar{X}) + \left(\mathbb{E}(\bar{X})\right)^2 \\ &= \frac{1}{n} + \theta^2. \end{aligned}$$

Plugging this in, the risk is

$$\begin{aligned} R(\theta, \hat{\theta}) &= \theta^2 - 2c\theta^2 + c^2\left(\frac{1}{n} + \theta^2\right) \\ &= \frac{c^2}{n} + (1 - 2c + c^2)\theta^2. \end{aligned}$$

(b) The posterior risk is

$$\begin{aligned} r(\hat{\theta}|\bar{X}) &= \mathbb{E}\left((\theta-\hat{\theta})^2|\bar{X}\right) \\ &= \mathbb{E}\left((\theta-c\bar{X})^2|\bar{X}\right) \\ &= \mathbb{E}\left(\theta^2-2c\theta\bar{X}+c^2\bar{X}^2|\bar{X}\right) \\ &= \mathbb{E}(\theta^2|\bar{X})-2c\bar{X}\mathbb{E}(\theta|\bar{X})+c^2\bar{X}^2. \end{aligned}$$

Now we will need the posterior distribution of $\theta|\bar{X}$. We can use the fact from above. Since our likelihood is already in terms of $\bar{X}$, $m=1$. Additionally, $\sigma^2=\frac{1}{n}$ and $a=0$. Thus,

$$\begin{aligned} \mu &= \frac{b^2\bar{X}+\frac{1}{n}\cdot 0}{b^2+\frac{1}{n}} = \mathbb{E}(\theta|\bar{X}) \\ \tau^2 &= \frac{\frac{1}{n}b^2}{b^2+\frac{1}{n}} = \mathrm{Var}(\theta|\bar{X}), \end{aligned}$$

and

$$\mathbb{E}(\theta^2|\bar{X}) = \tau^2+\mu^2.$$

Plugging all of this in, we have

$$\begin{aligned} r(\hat{\theta}|\bar{X}) &= \frac{\frac{1}{n}b^2}{b^2+\frac{1}{n}} + \left(\frac{b^2\bar{X}}{b^2+\frac{1}{n}}\right)^2 - 2c\bar{X}\left(\frac{b^2\bar{X}}{b^2+\frac{1}{n}}\right) + c^2\bar{X}^2 \\ &= \frac{b^2}{b^2+\frac{1}{n}}\left(\frac{1}{n}+\left(\frac{b^2}{b^2+\frac{1}{n}}-2c+\left(\frac{b^2+\frac{1}{n}}{b^2}\right)c^2\right)\bar{X}^2\right). \end{aligned}$$

(c) The Bayes risk is

$$\begin{aligned} r(f,\hat{\theta}) &= \mathbb{E}\left(R(\theta,\hat{\theta})\right) \text{ (where } f \text{ is the prior density)} \\ &= \mathbb{E}\left(\frac{c^2}{n}+\theta^2(1-2c+c^2)\right) \\ &= \frac{c^2}{n}+(1-2c+c^2)\mathbb{E}(\theta^2) \\ &= \frac{c^2}{n}+(1-2c+c^2)b^2, \end{aligned}$$

where $\mathbb{E}(\theta^2)$ was calculated using the fact that

$$\begin{aligned} \mathbb{E}(\theta^2) &= \mathrm{Var}(\theta)+(\mathbb{E}(\theta))^2 \\ &= b^2+0^2 = b^2. \end{aligned}$$

Problem 2. Calculating a Bayes rule

(Follow-up to previous problem)

For the same set-up as the above problem, find the Bayes rule for the absolute error loss. Is this Bayes rule the same or different as the Bayes rule for the mean squared error loss, and why?

Solution

We want to find an estimator $\hat{\theta}$ that minimizes the Bayes risk. It is sufficient to find an estimator $\hat{\theta}$ that minimizes the posterior risk

$$r(\hat{\theta}|X) = \mathbb{E}\left(|\theta - \hat{\theta}|\middle|X\right).$$

In general, it can be shown that the value c that minimizes

$$\mathbb{E}\left(|Y - c|\right)$$

is median(Y).

Thus, the value that minimizes risk is the posterior median. From Problem 1., we found that the posterior was normal with mean $\mu = \frac{b^2\bar{X}}{b^2+\frac{1}{n}}$. Since in a normal distribution the mean equals the median, this implies the posterior median is μ.

In conclusion, the Bayes rule with the absolute error loss is μ. That is the same as the Bayes rule with the mean squared error loss. In general, Bayes rules with different loss functions will not be the same, but here it worked out that way because we had the normal distribution.