Decision Theory: Perils of the Likelihood Principle
10716, Spring 2020
Pradeep Ravikumar (amending notes from Larry Wasserman)

1 Introduction: Bayesian, Frequentist, Likelihood Principle

Consider an estimator $\delta : \mathcal{X} \mapsto \Theta$. From a decision theoretic perspective, is said to be "frequentist" if it is designed to minimize the "frequentist" risk $R(\theta^*, \delta)$, ideally in some global sense as θ^* ranges over Θ (for instance, in a minimax sense).

Looking at the frequentist risk $R(\theta^*, \delta) = \mathbb{E}_{X \sim P(\cdot;\theta^*)} L(\theta^*, \delta(X))$, it cares about the performance not only on observed data X, but over all possible datasets, and thus might lead to estimators that do not satisfy the "conditonality principle" introduced earlier.

An estimator is said to be Bayesian if it specifically aims to minimize $r(\pi, \delta) = \int_{\theta^*} R(\theta^*, \delta)\pi(\theta^*)d\theta^*$. The beauty of the Bayesian estimator is that it does satisfy the conditionality principle due to the result that:

$$\delta_{\text{Bayes}}(X) \in \arg\inf_{a \in \mathcal{A}} \int_{\theta^*} L(\theta^*, a)\pi(\theta^*|x)d\theta^*,$$

where $\pi(\theta^*|X)$ is the posterior distribution over Θ given the observations X. Given a pre-specified prior, it can be seen that it only depends on the data via its likelihood function $L(\theta^*) = p(X|\theta^*)$, and thus also satisfies the likelihood principle.

Now consider $R(\theta^*, \delta) = \mathbb{E}_{X \sim P(\cdot;\theta^*)} L(\theta^*, \delta(X))$. In machine learning practice, we evaluate the loss of an estimator over multiple datasets, and care about the overall performance across datasets. So for instance, given N parameters $\{\theta_i^*\}_{i=1}^N$, and N datasets $X^{(i)}$, we might care about the performance:

$$\frac{1}{N} \sum_{i=1}^{N} L(\theta_i^*, \delta(X^{(i)})).$$

Here, the average is over multiple datasets from different parameters in contrast the typical frequentist risk of averaging over multiple datasets drawm from a single parameter. Interestingly, this latter notion dates back to the earliest frequentist ideas of Neyman and Pearson (Neyman 1967).

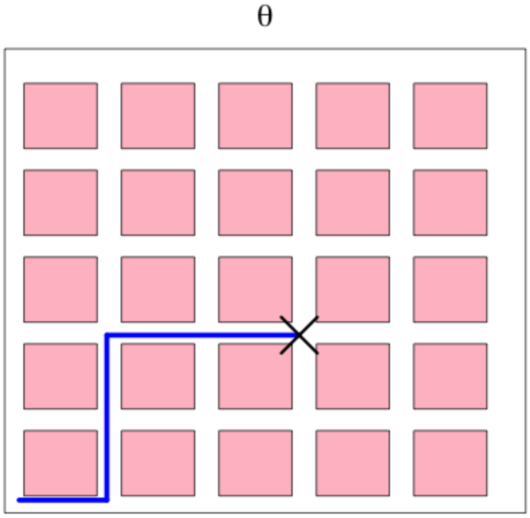
But overall, frequentist estimators care about performance over multiple datasets, and might not satisfy the likelihood principle. This additional flexibility might seem like it could lead to bad and even irrational estimators, such as in the examples in previous lecture. In this lecture, we will see that this additional flexibility can also help us.

2 Stone's Paradox

We will first discuss a famous example from Stone (1970, 1976, 1982). In technical jargon, he shows that "a finitely additive measure on the free group with two generators is nonconglomerable." In English: even for a simple problem with a discrete parameters space, flat priors (and consequently, likelihood principle) can lead to surprises. Fortunately, we don't need to know anything about free groups to understand this example.

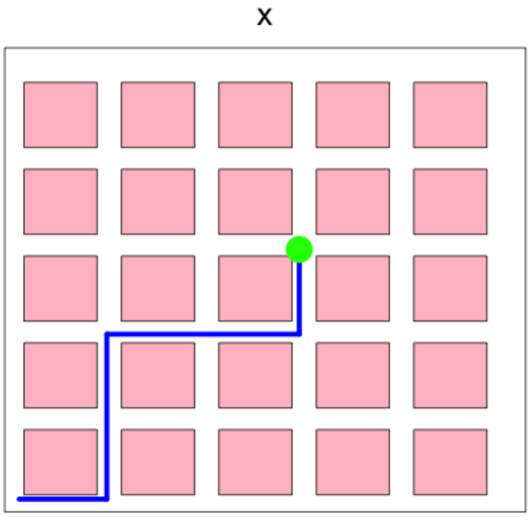
2.1 Hunting For a Treasure In Flatland

Suppose that I can wander randomly in a two dimensional grid-world. I can only move in four directions: North, South, West, East. I wander around for a while then I stop and bury a treasure. Call this path θ^*. Here is an example:



My Path. The X is the treasure.

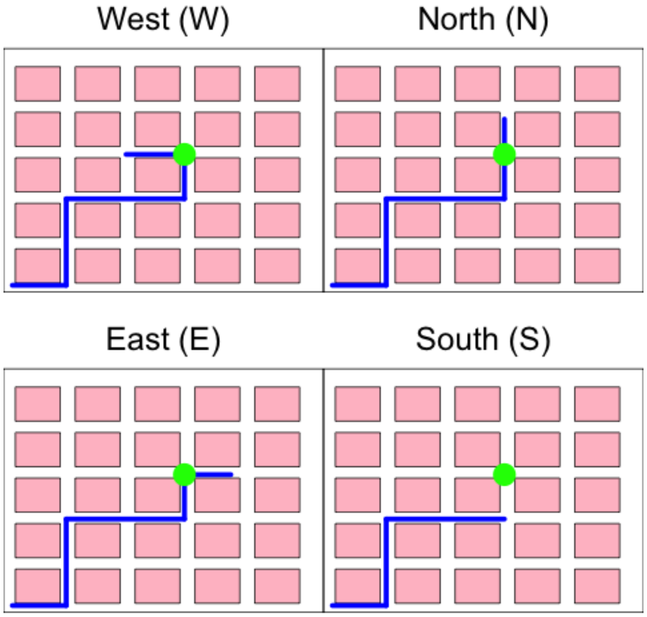
Now I take one more random step. Each direction has equal probability. Call this final path x. So it might look like this:



x

One more random step.

There are only four possible paths that could have yielded x, namely the paths that end one step away from x:



West (W)

North (N)

East (E)

South (S)

Let us call these four paths N, S, W, E. The likelihood is the same for each of these. That is, $p(x|\theta) = 1/4$ for $\theta \in \{N, S, W, E\}$. It can be seen that the likelihood function has no information at all about the four possible parameter values. Note that the likelihood principle says that the observed likelihood contains all the useful information in the data.

Now suppose two people, Bob (a Bayesian) and Carla (a classical statistician) want to find the treasure. Suppose Bob uses a flat prior (so that he is again following the likelihood principle). Since the likelihood is also flat, his posterior is

$$P(\theta = N|x) = P(\theta = S|x) = P(\theta = W|x) = P(\theta = E|x) = \frac{1}{4}.$$

His Bayesian risk of any path decision $\widehat{\theta} \in \{N, S, W, E\}$, with respect to the zero-one loss, is then 3/4.

Carla (the classical statistician) on the other hand is not bound by the likelihood principle. She reasons: for any treasure path θ^*, out of the four extensions to the observed x, only one would reduce the length of the path. That is, there is a 3/4 chance that the path of x would be longer than that of θ^*. Since South is the only estimate such that the observed x is longer, she decides to output South. It can be seen that her risk is now only 1/4, since she will be wrong only 1/4 of the times that the extension x is a reduction of θ^*.

Bayesian Bob trying to follow this line of thought could be led further astray. Suppose he says: given x, the *posterior* probability of a θ^* being longer than x is 3/4 i.e. $P(\theta^* \in \{N, W, E\}|x) = 3/4$. So he decides to output one of these paths at random. Even if gets zero loss if any of these three choices were right, his risk would still be 1/4.

Here is quote from Stone (1976): (changing his B and C to Bob and Carla):

" ... it is clear that when Bob and Carla repeatedly engage in this treasure hunt, Bob will find that his posterior probability assignment becomes increasingly discrepant with his proportion of wins and that Carla is, somehow, doing better than [s]he ought. However, there is no message ... that will allow Bob to escape from his Promethean situation; he cannot learn from his experience because each hunt is independent of the other."

2.2 More Trouble For Bob

Let A^*_θ be the event that the final step reduced the length of the treasure path. Using the posterior above, we see that Bob finds that $P(A^*_\theta|x) = 3/4$ for each x. Since this holds for each x, Bob deduces that $P(A^*_\theta) = 3/4$. On the other hand, Bob notes that $P(A^*_\theta|\theta^*) = 1/4$ for every θ^*. Hence, $P(A^*_\theta) = 1/4$.

Bob has just proved that $3/4 = 1/4$.

2.3 The Four Sided Die

Here is another description of the problem. Consider a four sided die whose sides are labeled with the symbols $\{a, b, a^{-1}, b^{-1}\}$. We roll the die several times and we record the label on the lowermost face (there is a no uppermost face on a four-sided die). A typical outcome might look like this string of symbols:

$$a \;\; a\; b\; a^{-1}\; b\; b^{-1}\; b\; a\; a^{-1}\; b$$

Now we apply an annihilation rule. If a and a^{-1} appear next to each other, we eliminate these two symbols. Similarly, if b and b^{-1} appear next to each other, we eliminate those two symbols. So the sequence above gets reduced to:

$$a \;\; a\; b\; a^{-1}\; b\; b$$

Let us denote the resulting string of symbols, after removing annihilations, by θ. Now we toss the die one more time. We add this last symbol to θ and we apply the annihilation rule once more. This results in a string which we will denote by x.

You get to see x and you want to infer θ.

Having observed x, there are four possible values of θ and each has the same likelihood. For example, suppose $x = (a, a)$. Then θ has to be one of the following:

$$(a), \quad (a\,a\,a), \quad (a\,a\,b^{-1}), \quad (a\,a\,b)$$

The likelihood function is constant over these four values.

Suppose we use a flat prior on θ. Then the posterior is uniform on these four possibilities.

But fix any θ. Now note that x is smaller than θ only if the last symbol of θ is annihilated, which occurs with probability 1/4. So a smart frequentist choice for an estimate of θ is the only one that is smaller than x.

2.4 Likelihood Principle

In these examples, the likelihood does not distinguish the four possible parameter values. Whereas, for instance, in the path example, the direction of the path from the current position — which does not affect the likelihood — has lots of information. This suggests that both MLE and Bayesian inference might not always be the best decision theoretic strategy.

Technically the problem here is that there is group structure and the group is not amenable. Hidden beneath this seemingly simple example is some rather deep group theory.

3 Partial Observations

The example in question is from Robins and Ritov (1997). A simplified version appeared in Wasserman (2004) and Robins and Wasserman (2000). The example is related to ideas from the foundations of survey sampling (Basu 1969, Godambe and Thompson 1976) and also to ancillarity paradoxes (Brown 1990, Foster and George 1996).

3.1 The Model

Here is (a version of) the example. Consider iid random variables

$$(X_1, Y_1, R_1), \ldots, (X_n, Y_n, R_n).$$

The random variables take values as follows:

$$X_i \in [0,1]^d, \quad Y_i \in \{0,1\}, \quad R_i \in \{0,1\}.$$

Think of d as being very, very large. For example, $d = 100,000$ and $n = 1,000$.

The idea is this: we observe X_i. Then we flip a biased coin R_i. If $R_i = 1$ then you get to see Y_i. If $R_i = 0$ then you don't get to see Y_i. The goal is to estimate

$$\psi = P(Y_i = 1).$$

Here are the details. The distribution takes the form

$$p(x, y, r) = p_X(x) p_{Y|X}(y|x) p_{R|X}(r|x).$$

Note that Y and R are independent, given X. For simplicity, we will take $p(x)$ to be uniform on $[0,1]^d$. Next, let

$$\theta(x) \equiv p_{Y|X}(1|x) = P(Y = 1|X = x)$$

where $\theta(x)$ is a function. That is, $\theta : [0,1]^d \to [0,1]$. Of course,

$$p_{Y|X}(0|x) = P(Y = 0|X = x) = 1 - \theta(x).$$

Similarly, let

$$\pi(x) \equiv p_{R|X}(1|x) = P(R = 1|X = x)$$

where $\pi(x)$ is a function. That is, $\pi : [0,1]^d \to [0,1]$. Of course,

$$p_{R|X}(0|x) = P(R = 0|X = x) = 1 - \pi(x).$$

The function π is **known.** We construct it. Remember that $\pi(x) = P(R = 1|X = x)$ is the probability that we get to observe Y given that $X = x$. Think of Y as something that is expensive to measure. We don't always want to measure it. So we make a random decision about whether to measure it. And we let the probability of measuring Y be a function $\pi(x)$ of x. And we get to construct this function.

Let $\delta > 0$ be a known, small, positive number. We will assume that

$$\pi(x) \geq \delta$$

for all x.

The only thing in the the model we don't know is the function $\theta(x)$. Again, we will assume that

$$\delta \leq \theta(x) \leq 1 - \delta.$$

Let Θ denote all measurable functions on $[0,1]^d$ that satisfy the above conditions. The parameter space is the set of functions Θ.

Let $\mathcal{P}$ be the set of joint distributions of the form

$$p(x)\,\pi(x)^r(1-\pi(x))^{1-r}\,\theta(x)^y(1-\theta(x))^{1-y}$$

where $p(x) = 1$, and $\pi(\cdot)$ and $\theta(\cdot)$ satisfy the conditions above. So far, we are considering the sub-model $\mathcal{P}_\pi$ in which π is known.

The parameter of interest is $\psi = P(Y = 1)$. We can write this as

$$\psi = P(Y=1) = \int_{[0,1]^d} P(Y=1|X=x)p(x)dx = \int_{[0,1]^d} \theta(x)dx.$$

Hence, ψ is a function of θ. If we know $\theta(\cdot)$ then we can compute ψ.

3.2 Likelihood, Bayesian Analysis

Now consider the likelihood function. The likelihood for one observation takes the form $p(x)p(r|x)p(y|x)^r$. The reason for having r in the exponent is that, if $r = 0$, then y is not observed so the $p(y|x)$ gets left out. The likelihood for n observations is

$$\prod_{i=1}^n p(X_i)p(R_i|X_i)p(Y_i|X_i)^{R_i} = \prod_i p(X_i)\pi(X_i)^{R_i}(1-\pi(X_i))^{1-R_i}\,\theta(X_i)^{Y_iR_i}(1-\theta(X_i))^{(1-Y_i)R_i}.$$

But remember that the only unknown is $\theta(x)$, and that $\pi(x)$, and hence $\pi(X_i)^{R_i}(1-\pi(X_i))^{1-R_i}$ is known. So, the likelihood is

$$\mathcal{L}(\theta) \propto \prod_i \theta(X_i)^{Y_iR_i}(1-\theta(X_i))^{(1-Y_i)R_i}.$$

Now comes the interesting part. The likelihood has essentially no information in it.

To see that the likelihood has no information, consider a simpler case where $\theta(x)$ is a function on $[0, 1]$. Now discretize the interval into many small bins. Let B be the number of bins. We can then replace the function θ with a high-dimensional vector $\theta = (\theta_1, \ldots, \theta_B)$. With $n < B$, most bins are empty. The data contain no information for most of the θ_j's.

Indeed, we have the following theorem from Robins and Ritov (1997):

Theorem. (Robins and Ritov 1997). Any estimator that is not a function of $\pi(\cdot)$ cannot be uniformly consistent.

This means that, at no finite sample size, will an estimator $\widehat{\psi}$ that is not a function of π be close to ψ for all distributions in $\mathcal{P}$. In fact, the theorem holds for a neighborhood around every pair (π, θ). But when π is known and is used in the estimator (as we will see in the frequentist estimator in the next section) we can have uniform consistency.

For a Bayesian analysis, we can combine the likelihood above with a prior W on Θ, to creates a posterior distribution on Θ which we will denote by W_n. Since the parameter of interest ψ is a function of θ, the posterior W_n for θ defines a posterior distribution for ψ. But note that this Bayesian analysis will also ignore π since the $\pi(X_i)'s$ are just constants in the likelihood.

A Bayesian criticism might be that this is because we have not enforced any smoothness (perhaps via the prior) on $\theta(x)$. Without smoothness, knowing $\theta(x)$ does not tell you anything about $\theta(x + \epsilon)$ (assuming the prior W does not depend on π). This is true and better inferences would obtain if we used a prior that enforced smoothness. But this argument falls apart when d is large. When d is large, forcing $\theta(x)$ to be smooth does not help unless you make it very, very, very smooth. The larger d is, the more smoothness you need to get borrowing of information across different values of $\theta(x)$. But this introduces a huge bias which precludes uniform consistency.

3.3 Frequentist Analysis

The usual frequentist estimator is the Horwitz-Thompson estimator

$$\widehat{\psi} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i R_i}{\pi(X_i)}.$$

It is easy to verify that $\widehat{\psi}$ is unbiased and consistent. Furthermore, $\widehat{\psi} - \psi = O_P(n^{-\frac{1}{2}})$. In fact, let us define

$$I_n = [\widehat{\psi} - \epsilon_n,\ \widehat{\psi} + \epsilon_n]$$

where

$$\epsilon_n = \sqrt{\frac{1}{2n\delta^2}\log\left(\frac{2}{\alpha}\right)}.$$

It follows from Hoeffding's inequality that

$$\sup_{P\in\mathcal{P}_\pi} P(\psi \in I_n) \geq 1-\alpha$$

Thus we have a finite sample, $1-\alpha$ confidence interval with length $O(1/\sqrt{n})$.

4 Freedman's Theorem

We now review an interesting result by David Freedman (Annals of Mathematical Statistics, Volume 36, Number 2 (1965), 454-456).

The result says that, "almost all" Bayesian posterior distributions are inconsistent, in a sense we'll make precise below. The math is uncontroversial but, as you might imagine, the intepretation of the result is likely to be controversial.

The paper is very short, barely more than two pages. This summary will be even shorter (with slightly different notation.)

Let $X_1,\ldots,X_n$ be an iid sample from a distribution P on the natural numbers $I=\{1,2,3,\ldots,\}$. Let $\mathcal{P}$ be the set of all such distributions. We endow $\mathcal{P}$ with the weak* topology. Hence, $P_n \to P$ iff $P_n(i)\to P(i)$ for all i.

Let μ denote a prior distribution on $\mathcal{P}$. (More precisely, a prior on an appropriate σ-field.) Let Π be all priors. Again, we endow the set with the weak* topology. Thus $\mu_n \to \mu$ iff $\int f d\mu_n \to \int f d\mu$ for all bounded, continuous, real functions f.

Let μ_n be the posterior corresponding to the prior μ after n observations. We will say that the pair (P,μ) is consistent if

$$P^\infty\left(\lim_{n\to\infty}\mu_n = \delta_P\right) = 1$$

where P^∞ is the product measure corresponding to P and δ_P is a point mass at P.

Now we need to recall some topology. A set is nowhere dense if its closure has an empty interior. A set is meager if it is a countable union of nowehere dense sets. Meager sets are small; think of a meager set as the topological version of a null set in measure theory.

Freedman's theorem is: the sets of consistent pairs (P,μ) is meager.

This means that, in a topological sense, consistency is rare for Bayesian procedures. From this result, it can also be shown that most pairs of priors lead to inferences that disagree. (The agreeing pairs are meager.) Or as Freedman says in his paper:

“ ... it is easy to prove that for essentially any pair of Bayesians, each thinks the other is crazy.”

5 References

Basu, D. (1969). Role of the Sufficiency and Likelihood Principles in Sample Survey Theory. *Sankya*, 31, 441-454.

Brown, L.D. (1990). An ancillarity paradox which appears in multiple linear regression. *The Annals of Statistics*, 18, 471-493.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society. Series B,* 195-233.

Foster, D.P. and George, E.I. (1996). A simple ancillarity paradox. *Scandinavian journal of statistics*, 233-242.

Godambe, V. P., and Thompson, M. E. (1976), Philosophy of Survey-Sampling Practice. In *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science*, eds. W.L.Harper and A.Hooker, Dordrecht: Reidel.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91, 1343-1370.

Robins, J.M. and Ritov, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models. *Statistics in Medicine*, 16, 285–319.

Robins, J. and Wasserman, L. (2000). Conditioning, likelihood, and coherence: a review of some foundational concepts. *Journal of the American Statistical Association*, 95, 1340-1346.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J.M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99, 439-456.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models. *Journal of the American Statistical Association*, 1096-1120.

Sims, Christopher. On An Example of Larry Wasserman. Available at: http://www.princeton.edu/ sims/.

Stone, M. (1970). Necessary and sufficient condition for convergence in probability to invariant posterior distributions. *The Annals of Mathematical Statistics*, 41, 1349-1353,

Stone, M. (1976). Strong inconsistency from uniform priors. *Journal of the American Statistical Association*, 71, 114-116.

Stone, M. (1982). Review and analysis of some inconsistencies related to improper priors and finite additivity. *Studies in Logic and the Foundations of Mathematics*, 104, 413-426.

Wasserman, L. (2004). *All of Statistics: a Concise Course in Statistical Inference.* Springer Verlag.