36-705: Intermediate Statistics **Fall 2019**

Lecture 33: November 22

Lecturer: Larry Wasserman

Today we will begin a new topic of Bayesian Inference. Siva will return to wrap up the last bit on non-parametric regression.

33.1 Bayesian Inference

We have already talked about the mechanics of Bayesian inference when we discussed constructing point estimates by treating the parameters as random with a prior, and the computing and summarizing the posterior. What we really did not talk about yet, and will spend a small amount of time talking about now is the philosophy of Bayesian inference.

The philosophical distinction between Bayes and frequentists is deep. We have so far followed the frequentist framework, where, to us a probability is representing some type of long run frequency, i.e. when we say the probability that our estimator is close to some unknown "true" parameter with probability at least $1 - \delta$ we are really imagining repeating this (or some other) experiment many many times and then our guarantees will be correct for at least $1 - \delta$ of these experiments. Similarly, with confidence intervals, we imagine many people across the world construct confidence intervals and our guarantee is that 95% of those intervals would trap the true parameter, i.e. **the goal of frequentist inference is to create procedures with long run guarantees.**

Moreover, the guarantees should be uniform over θ if possible. For example, a confidence interval traps the true value of θ with probability $1 - \alpha$, no matter what the true value of θ is. **In frequentist inference, procedures are random while parameters are fixed, unknown quantities.**

In the *Bayesian approach*, probability is regarded as a measure of **subjective degree of belief**. One can view the Bayesian approach as a way to manipulate beliefs. Beliefs are then assumed to follow the rules of normal probabilities by a notion called *coherence*. In this framework, everything, including parameters, is regarded as random. These procedures do not have to satisfy frequency guarantees.

Here is a table from Larry:

A summary of the main ideas is in Table 33.1.

	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
θ	random variable	fixed
X	random variable	random variable
Use Bayes' theorem?	Yes. To update beliefs.	Yes, if it leads to procedure with good frequentist behavior. Otherwise no.

Table 33.1: Bayesian versus Frequentist Inference

33.2 The mechanics of Bayesian Inference

Roughly, the setup in Bayesian Inference is exactly the same as in frequentist inference: we begin by specifying a statistical model, i.e. a collection of distributions $\{P_\theta : \theta \in \Theta\}$.

The main distinction is that we now treat the parameter θ as random, and encode our "prior beliefs" about the value of the parameter in a distribution π.

We assume that the observed data, is from the *conditional distribution*, conditional on some realization of the random parameter, i.e. the setup is:

$$\theta \sim \pi$$
$$\{X_1, \ldots, X_n\}|\theta \sim P_\theta.$$

We do not observe θ but can compute our "posterior belief" using Bayes' rule, i.e.:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)}{\int_\theta \mathcal{L}(\theta; X_1, \ldots, X_n)\pi(\theta)},$$

i.e. while the frequentist treats the likelihood as just a function of θ, the Bayesian (weights and) normalizes the likelihood and interprets it as a distribution over Θ.

We have seen examples of this whole thing before, except rather than treat the posterior as an object of interest, we used it to obtain point estimates.

33.3 The goals of Bayesian inference

In frequentist inference the goal was somehow part of the definition: create procedures that have good frequency properties.

In Bayesian inference the goal was not very clearly articulated, i.e. when is a Bayesian analysis considered successful. There are two camps here:

1. **Pure Bayesian viewpoint:** Once you write down a prior that captures your prior belief and compute the posterior, you are essentially done, i.e. you have succeeded.
2. **The frequentist viewpoint:** You are successful if you are successful in the frequentist viewpoint, i.e. treat the parameter as fixed, and define success as your posterior concentrating around the true parameter (i.e. some equivalent of consistency) and confidence intervals computed using the posterior have frequentist coverage guarantees.

33.4 Bayesian confidence sets and Frequentist guarantees

Once we have a posterior distribution, we can construct what are called credible sets: they are the Bayesian analogue of confidence sets but are quite different.

A $1 - \alpha$ credible set/interval is simply any set C_α to which the posterior assigns $1 - \alpha$ mass, i.e.

$$\int_{C_\alpha} \pi(\theta | X_1, \ldots, X_n) d\theta = 1 - \alpha.$$

Once again notice that the thing that is random is θ, the data is conditioned on (i.e. fixed). We could write this as:

$$\mathbb{P}_{\theta \sim \pi(\theta | X_1, \ldots, X_n)}(\theta \in C_\alpha | X_1, \ldots, X_n) = 1 - \alpha.$$

The set C_α is fixed (i.e. not random) here, unlike in a frequentist confidence interval. These intervals do not typically have frequency guarantees, and we will see examples of this.

If one is interested in the frequency properties of Bayesian inference then one might also be interested in some notion of frequentist consistency and rates of convergence. The typical way to formulate frequentist consistency is via something called *posterior contraction*, i.e. in the frequentist setup (where θ^* is fixed, unknown) we want that our posterior concentrates around the true value of the parameter (consistency) and does so quickly (rates of convergence).

Formally, consistency says that for any fixed $\epsilon > 0$,

$$\pi(\{\theta : \|\theta - \theta^*\| \geq \epsilon\} | X_1, \ldots, X_n) \to 0,$$

when $X_1, \ldots, X_n \sim P_{\theta^*}$. We would also say that the rate of convergence is ϵ_n if for some $\delta_n \to 0$ we have that,

$$\pi(\{\theta : \|\theta - \theta^*\| \leq \epsilon_n\} | X_1, \ldots, X_n) \geq 1 - \delta_n,$$

as $n \to \infty$ (again, $X_1, \ldots, X_n \sim P_{\theta^*}$).

33.5 Bernstein-von Mises theorem

At a high-level the Bernstein-von Mises theorem guarantees us that in fixed-dimensional problems, under the assumption that the prior is continuous, and (strictly) positive in a neighborhood around θ^*, the posterior is close to a Gaussian, i.e.

$$\left\| \pi(\theta | X_1, \ldots, X_n) - N\left(\widehat{\theta}_n, \frac{I(\widehat{\theta}_n)^{-1}}{n}\right) \right\|_{\mathrm{TV}} \to 0,$$

where $\widehat{\theta}_n$ is the MLE and the distance between the two distributions is the total-variation distance, i.e. for two distributions with densities p, q:

$$\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \int |p(x) - q(x)| dx.$$

We might discuss this further at some point but for now you should take away that the posterior is very close to a Gaussian centered at the MLE with rapidly shrinking variance.

One immediate consequence of the BvM result is that credible intervals will be roughly identical to the usual Wald interval (based on the MLE) as $n \to \infty$. The key take away: in fixed dimension, large sample-size problems, under some conditions Bayesian procedures will have strong frequency guarantees.

The condition that "(strictly) positive in a neighborhood around θ^*" is extremely strong in high-dimensions or in a non-parametric problem. In general, when the parameter space is large you should be suspicious of this assumption.

33.6 Where do priors come from?

The "correct" answer is that the prior should truly be an encoding of your prior beliefs. Often we choose priors by convenience (recall our examples from the minimax lecture). Some might argue that in many cases the priors do not matter (see below) but this is only rigorously true in low-dimensional, parametric problems.

Some might also choose priors based on the data, this is known as *empirical Bayes*. This is often a good idea but some would consider it to not strictly adhere to the Bayesian philosophy.

Some have argued for what are called non-informative priors, i.e. priors that somehow capture complete ignorance about the parameter. The natural first attempt would be to say that we take $\pi(\theta) \propto 1$, however this has some drawbacks. If we have no information about the parameter θ then presumably we should also have no information about some transformation about the parameter, i.e. say θ^2. However, if you transform the flat prior to a prior on θ^2 it will not be flat. A prior that is in fact invariant under transformations is called Jeffreys prior where we choose $\pi(\theta) \propto \sqrt{I(\theta)}$, where $I(\theta)$ is the Fisher information for the model under consideration. You will verify this in your HW.

There are also so-called hierarchical priors, i.e. we can parameterize the prior, treat these parameters as random variables and then place priors on those parameters as well. There is some folklore intuition that results are less sensitive to the parameters of the higher-level priors but this is somewhat difficult to make precise.

Finally, perhaps all of this is missing the point? There is a sense in which many pragmatic Bayesians believe that the prior is not the important piece of Bayesian inference, it is the complicated averaging that happens in Bayesian inference. Again this is difficult to make precise. There are frequentist versions of the model averaging idea (look up exponentially weighted aggregation or mirror averaging) that lead to aggregated models with great properties.

33.7 Priors = Regularizers?

A slightly different viewpoint that is often articulated is that one can view priors as regularizers.

Most regularized frequentist estimators, for instance estimating Bernoulli probabilities with "Laplace smoothing" (i.e. adding psuedo-counts) is just the posterior mean with a Beta prior. The LASSO regression estimator or Ridge regression estimator are just the posterior mode with either a Laplace prior or a Gaussian prior (respectively). Relatedly, one can derive model complexity regularizers with appropriate model complexity dependent priors.

The argument is that "many sensible frequentists procedures (ones with strong guarantees) are just posterior summaries with particular priors."

This argument should be taken with a grain of salt: lets focus on the LASSO, which is the posterior mode with a Laplace prior. As we have discussed previously the LASSO has some very desirable properties (high-dimensional prediction consistency) and so one might wonder does the (full) posterior have nice properties in a high-dimensional setting?

The answer turns out to be no: *once you leave the realm of the Bernstein-von Mises theorem (fixed d, growing n) things can break down.* In particular, in the high-dimensional regression problem the posterior itself does not meaningfully concentrate and sampling from the posterior will lead to completely meaningless inference (from a frequentist point of view). Essentially only the posterior mode has nice properties, the rest of the posterior is useless.

33.8 Failure of credible intervals

We have said many times now that things can break down in high-dimensions. This is particularly alarming if we treat credible intervals as confidence intervals. They are not valid confidence intervals. Let us verify this in a simple example.

Suppose we are in the Gaussian sequence model, i.e. we observe,

$$y_i = \theta_i + \epsilon_i, \quad i \in \{1, \ldots, d\},$$

and $\epsilon_i \sim N(0, \sigma^2/n)$.

We choose a flat prior (although this is not really crucial it makes the calculations simpler), i.e. $\pi(\theta_1, \ldots, \theta_d) \propto 1$. This is an example of something called an *improper prior*, i.e. it is not really a valid distribution. We can still use the usual mechanics to obtain a valid posterior.

Our goal is to construct a confidence interval for the parameter $\mu = \sum_{i=1}^d \theta_i^2$. Since the prior is flat the posterior is easy to compute and in particular, the posterior factorizes over the parameters (since the prior is flat and the likelihood factorizes) and we have:

$$\pi(\theta_i | y_1, \ldots, y_d) \overset{d}{=} N(y_i, \sigma^2/n).$$

The posterior for $\mu | y_1, \ldots, y_d$ is σ^2/n times a non-central χ^2 distribution, with d degrees of freedom and non-centrality parameter $\lambda = (n/\sigma^2) \sum_{i=1}^d y_i^2$.

Observe that, the mean of the posterior for μ is at $\sum_{i=1}^d (y_i^2 + \sigma^2/n)$, and the variance of the posterior for μ is $4\sigma^2 (\sum_{i=1}^d y_i^2)/n + 2\sigma^4 d/n^2$.

If we were to examine frequentist properties, we would fix a θ, and then mean of the posterior in expectation would be at $\mathbb{E}\left[\sum_{i=1}^d (y_i^2 + \sigma^2/n)\right]$, i.e. at $\mu + 2\sigma^2 d/n$, while the standard deviation of the posterior would be on the order of $\sigma\sqrt{\mu/n} + \sigma^2\sqrt{d}/n$. So the posterior is centered at the wrong point, and its spread is quite small. Using these two facts along with Chebyshev's inequality, you can see that a posterior credible interval will have coverage that $\rightarrow 0$ as $d \rightarrow \infty$.