36-705: Intermediate Statistics **Fall 2017**

Lecture 31: November 17

Lecturer: Siva Balakrishnan

31.1 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:

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

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 $\lambda = (n/\sigma^2) \sum_{i=1}^d y_i^2$.

You will do this more carefully in your HW (using χ^2 tail bounds) but for now, let us 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

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 $\to 0$ as $d \to \infty$.

31.2 Computation and MCMC

The basic idea in Bayesian inference is that we have a prior $\pi(\theta)$ and we observe data $(X_1, \ldots, X_n)$. We then compute the posterior distribution:

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

and use it in various ways (point estimation - take posterior mean, "confidence" intervals - use posterior credible intervals).

Suppose we consider point estimation: we want to compute the mean of the posterior:

$$\widehat{\theta} = \int \theta \ \pi(\theta|X_1, \ldots, X_n)d\theta.$$

The main difficulty here is that the normalizing constant:

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

is difficult to compute because the parameter space could be large/high-dimensional so we do not really "know" the posterior.

One simple idea is that if we could sample $(\theta_1, \ldots, \theta_m)$ from the posterior then we could try to approximate:

$$\widehat{\theta} \approx \frac{1}{m}\sum_{i=1}^{m} \theta_i.$$

Of course, sampling from the posterior could be as hard as computing the distribution. Markov Chain Monte Carlo is a way to sample from the posterior without computing the normalizing constant.

31.3 Monte Carlo Integration

Suppose I want to (approximately) compute an expectation:

$$\mu = \mathbb{E}_{X\sim P}[f(X)],$$

but cannot do so analytically. However, suppose that I can sample $X_1, \ldots, X_n \sim P$, then I could approximate this expectation as:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

This is the idea of Monte Carlo Integration.

A first question is how accurate is this approximation? We know by the CLT that $\widehat{\mu} - \mu$ will be close to normally distributed so it only remains to calculate the variance.

$$\mathrm{Var}(\widehat{\mu}) = \frac{1}{n}\mathrm{Var}(f(X)),$$

which is for instance small if f is bounded, i.e if $|f| \leq M$, in which case:

$$\mathrm{Var}(f(X)) \leq M^2,$$

and the variance is roughly $1/n$. So if we take enough samples, then our estimate will be quite good. Furthermore, we can estimate this variance in the usual way:

$$\widehat{\mathrm{Var}}(\widehat{\mu}) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2.$$

Example 1: Suppose we want to compute the standard Gaussian CDF $\Phi(x)$ at some point x. This is given by:

$$\mu = \Phi(x) = \int_{-\infty}^{x} f(u)du,$$

where $f(x)$ is the standard Gaussian pdf. We can re-write this as an expectation:

$$\mu = \int_u h(u)f(u)du,$$

where $h(x) = \mathbb{I}(u \leq x)$. In order to use Monte Carlo integration, we would just draw many samples from a standard Gaussian, and then use:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leq x).$$

31.4 Importance Weighting

A next scenario is one where we cannot directly sample from the distribution P but still want to approximate:

$$\mu = \mathbb{E}_{X\sim P}[f(X)].$$

Let us suppose that we can instead sample from some distribution Q. Let us suppose that these two distributions have densities p, q and further that we can evaluate ratios of the form:

$$p(X)/q(X),$$

for any sampled value X. We can see that:

$$\mu = \mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim Q}\left[\frac{p(X)}{q(X)}f(X)\right] = \mathbb{E}_{X\sim Q}[w(X)f(X)],$$

at least provided that the weights are never infinite. This means that we can use Monte Carlo integration, given samples $X_1, \ldots, X_n$,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i).$$

The variance of this estimate depends on both the variance of w and the variance of f.

An important point is that even when we can sample from P it might be desirable to use importance sampling, as it might have much smaller variance. Let us consider a simple example:

Example 2: Suppose we want to estimate

$$\mu = \mathbb{P}(Z > 3) = 0.0013,$$

where $Z \sim N(0, 1)$. We can write this as an expectation as before, and then use Monte Carlo. Using $n = 100$, we find (from simulating many times) that $\mathbb{E}(\widehat{\mu}) = .0015$ and $\mathrm{Var}(\widehat{\mu}) = .0039$. An important observation is that most of the samples are wasted since very few samples are in the right tail (i.e. bigger than 3).

Suppose we instead used importance sampling, and sampled from Q which is $N(4, 1)$ (a shifted normal). In this case we find that $\mathbb{E}(\widehat{\mu}) = .0011$ and $\mathrm{Var}(\widehat{\mu}) = .0002$. Importance sampling reduces the variance by a factor of 20.

The rough guideline is that the variance of importance sampling is smallest if we sample proportional to $p|f|$, i.e. we sample more where both p and $|f|$ are large.

31.5 Markov Chain Monte Carlo

In the Bayesian problem we began with we cannot really use either Monte Carlo or importance weighting. We instead use MCMC.

First, the big picture: In Markov Chain Monte Carlo (MCMC) the goal is to design a Markov Chain, whose limiting distribution is the distribution P under which we want to compute an integral. We then sample from the Markov Chain, and then use the law of large numbers (adapted to Markov Chains) to argue that our estimate is good.

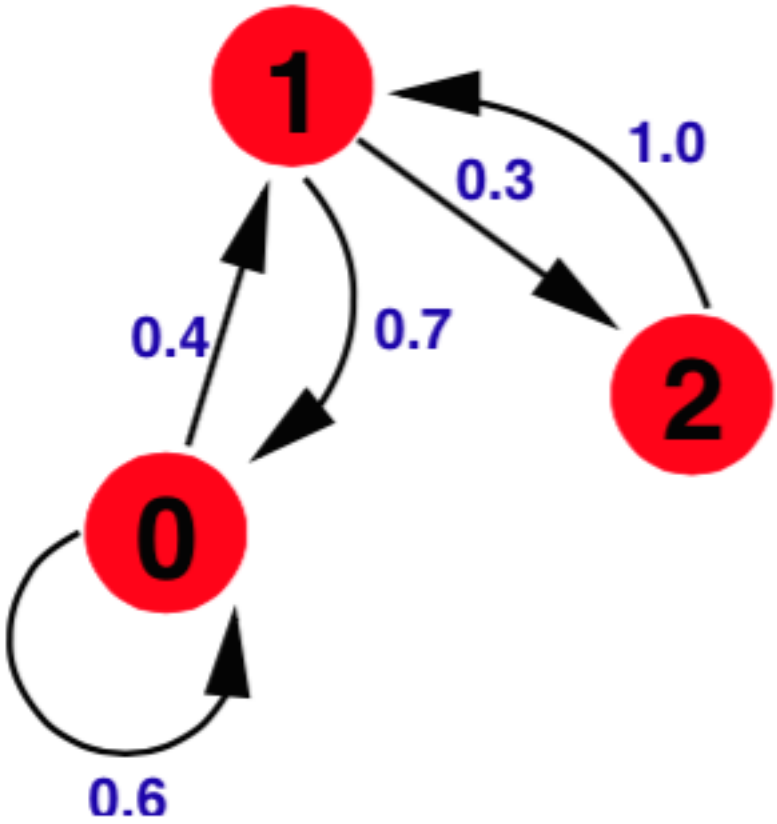
The details are a bit involved. We will do this at a very high-level.

First, what is a Markov Chain? A Markov Chain is a collection of random variables $\{X_1, \ldots, X_n\}$ that forms a graphical model: $X_1 \to X_2 \to \ldots \to X_n$.

In order to specify the joint distribution of a Markov Chain, we need to specify $p(X_1)$ and $p(X_{i+1}|X_i)$. To do this conveniently the focus is usually on what are called time-homogenous Markov Chains, i.e. $p(X_{i+1}|X_i) = T(X_i, X_{i+1})$, for a function T that does not depend on i. This function T is called the transition matrix or transition kernel of the Markov Chain.

Let us consider a simple example: suppose all the variables are discrete and take values $\{0, 1, 2\}$.

Consider a diagram of the form:



This diagram is specifying the transition matrix for us. Particularly, it says that the probability: $P(X_{i+1} = 2|X_i = 1) = 0.3$.

The next thing, we need to know is that Markov Chains almost always have a limiting distribution. The limiting distribution roughly, is the distribution of the random variable X_n for n large. We denote it by π, i.e.

$$\pi(x) = P(\lim_{n\to\infty} X_n = x).$$

An important aspect of Markov Chains is that they forget their initial state (i.e. they are ergodic/they mix), so π does not depend on $p(X_1)$.

Markov Chains also have a LLN associated with them: If $\{X_1, \ldots, X_n\}$ is a Markov Chain with limiting distribution π, then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \int_x \pi(x) f(x) dx = \mathbb{E}_\pi[f(x)].$$

This is a pretty magical property. The Markov Chain samples are actually dependent, i.e. X_2 depends on X_1, and X_3 depends on X_2 and X_1, and so on. The LLN says that even though there are these dependencies, they are weak (and get much weaker as you get further away in the chain), so samples from a Markov Chain behave quite similarly to i.i.d. samples from the limiting distribution.

With all of these preliminaries in place only one thing remains: construct a Markov Chain with a particular limiting distribution (think, the posterior distribution).

31.6 Computing the limiting distribution

Suppose I specified a Markov chain and promised you that it had a well-defined (unique) limiting distribution. Then one basic question is how can I analytically compute what the limiting distribution is?

There are several ways to do this. One that we will use today is to check what are called *detailed balance* conditions, i.e. any distribution π (i.e. $\pi \geq 0$ and $\int \pi(x)dx = 1$), that satisfies:

$$\pi(x)T(x, y) = \pi(y)T(y, x),$$

for every x, y, where T is the transition kernel of the Markov chain, is the limiting distribution of the Markov chain. As a quick note: the reverse implication is not necessarily true, i.e. the limiting distribution of a Markov chain does not need to satisfy detailed balance (when it does, such Markov chains are called reversible Markov chains).

To summarize, if we construct a Markov chain, and there is some distribution π that satisfies the detailed balance conditions then that distribution is the limiting distribution of the Markov chain.

31.7 The Metropolis-Hastings algorithm

Recall, that our broad goal is to draw samples from a distribution f (say).

Choose X_0 arbitrarily. For each subsequent index i we follow the algorithm given below:

1. Sample a proposal $y \sim q(y|X_i = x)$ from a "proposal distribution" q.

2. Evaluate the ratio:

$$r = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}.$$

3. Accept the new sample Y with probability r, and reject it otherwise. Alternatively, think of sampling $u \sim U[0,1]$ and accept if $u \leq r$ and reject otherwise.

Some basic intuition: We will understand formally why this works, but for now consider the case when the proposal is symmetric, i.e. $q(y|x) = q(x|y)$. In this case, we accept a new sample with probability:

$$r = \min\left\{\frac{f(y)}{f(x)}, 1\right\}.$$

Now, let us think about what it means to sample from f, roughly we want to draw more samples where f is high, and fewer samples where f is low. Our rule above basically says, always accept a sample if the density is higher at the proposed point (like hill-climbing) and if the density is lower at the proposed point you accept it with a smaller probability. This sampling rule is effectively biased to accept samples from regions where the density is high.

Three tasks remain: we need to decide how to choose a proposal distribution, we need to show that this algorithm does what we set out to, i.e. roughly generates samples from f, and we need to understand why this is useful in sampling from the posterior distribution for example.

31.7.1 Choosing a proposal distribution

This one is mostly an art, i.e. you try to pick a proposal distribution that somehow approximates the shape of the distribution you care about (f).

Often what we do is to choose:

$$q(Y|X = x) \sim N(x, \sigma^2),$$

so we sample a proposal around our current data point, and try to tune the tuning parameter σ (by trying to maintain a reasonable acceptance ratio while still enforcing that we explore most of the space).

31.7.2 Sampling posteriors

Our goal in the beginning of last lecture was to sample from the posterior distribution:

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

More generally, suppose we have a distribution that we know up to the normalizing constant, i.e. we can compute $g(x)$ but we want to sample from f which is given by:

$$f(x) = \frac{g(x)}{\int g(x)dx},$$

and the denominator can be difficult to compute.

The key point: The Metropolis Hastings algorithm, only interacts with f through ratios of the form

$$\frac{f(x)}{f(y)} = \frac{g(x)}{g(y)},$$

which are easy to compute. When you take the ratio, the normalizing constant disappears.

31.7.3 Limiting distributions

Now, we go back to the Metropolis Hastings algorithm. We need to show that this algorithm is constructing a Markov chain and the limiting distribution of this Markov chain is f.

The first part is easy: each subsequent sample X_{i+1} only depends on X_i and q and does not depend on any of the prior $X_1,\ldots,X_{i-1}$ (conditional on X_i) so the samples $X_1,\ldots,X_n$ form a Markov chain.

Let us first understand the transition probabilities of our Markov chain. In order to transition from x to y we need to sample y from the proposal and then need to accept this proposal. This happens with probability:

$$T(x,y) = q(y|x)\min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)},1\right\}.$$

Using this we can see that detailed balance is satisfied and f is the limiting distribution if:

$$f(x)q(y|x)\min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)},1\right\} \stackrel{?}{=} f(y)q(x|y)\min\left\{\frac{f(x)q(y|x)}{f(y)q(x|y)},1\right\}.$$

This is easy to check by some case analysis. For instance, suppose $f(x)q(y|x) \geq f(y)q(x|y)$, then this reduces to:

$$f(x)q(y|x)\frac{f(y)q(x|y)}{f(x)q(y|x)} \stackrel{?}{=} f(y)q(x|y),$$

which is clearly true. We can similarly check this is true in the case when $f(x)q(y|x) < f(y)q(x|y)$. From this we can conclude that f is the limiting distribution of the Markov chain we have constructed.

31.7.4 Some caution

While MCMC is a really nice trick in order to generate samples from something close to the posterior, there is an important caveat that I have ignored. For a Markov chain, the limiting distribution is its "asymptotic distribution", i.e. it is the distribution you are getting samples from asymptotically (as $n \to \infty$).

The hope is usually that for small (finite) values of n the distribution is close to the limiting distribution. This is called mixing or rapid mixing. Unfortunately, however, in many cases we do not know if the Markov chain mixes rapidly (this depends in a complicated fashion on the proposal and the unknown density f).

To a large extent, MCMC is a sensible heuristic, and some caution/care is required in applying it to difficult problems.