Stat 5421 Lecture Notes
Bayesian Inference
Charles J. Geyer
February 29, 2016

1 Introduction

1.1 Philosophy

The first (and only) postulate of Bayesian inference is

$$\textit{probability is the right way to describe uncertainty.} \tag{1}$$

Once you buy this, the rest is obvious. If a parameter is unknown, that means you are uncertain about what its value is, hence the right way to describe that uncertainty is a probability distribution. Thus we consider the parameter θ a random variable that has a distribution characterized by its PDF f.

In this Bayesians differ from "frequentists." The scare quotes are because being a "frequentist" doesn't have anything to do with the frequentist philosophy of probability but rather means one thinks statistical inference should be based on sampling distributions. "Samplingdistributionist" would be a better name than "frequentist" if English made compound words that way.

1.2 Bayes' Rule

Like frequentists, Bayesians also have the distribution of the data given the parameters. Bayesians write $f(x \mid \theta)$ where frequentists write $f_\theta(x)$. This emphasizes their fundamental philosophical disagreement: Bayesians think θ is a random variable, so $f(x \mid \theta)$ is the conditional distribution of x given θ, whereas frequentists deny that θ is a random variable, so $f_\theta(x)$ is the distribution of x, a different distribution for each different θ. But

$$f(x \mid \theta) = f_\theta(x), \qquad \text{for all } x \text{ and } \theta$$

so they are talking about the same thing in different language. Thought of as a function of θ rather than x, this same object is also called the likelihood

$$L_x(\theta) = f(x \mid \theta) = f_\theta(x), \qquad \text{for all } x \text{ and } \theta. \tag{2}$$

It's still the same thing in different language.

In the Bayesian setup, it is natural to think of the joint distribution of the data and parameter

$$\begin{aligned} \text{joint} &= \text{conditional} \times \text{marginal} \\ f(x, \theta) &= f(x \mid \theta) f(\theta) \end{aligned}$$

and of the other conditional

$$\begin{aligned} \text{conditional} &= \frac{\text{joint}}{\text{marginal}} \\ f(\theta \mid x) &= \frac{f(x, \theta)}{f(x)} \\ &= \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta) \, d\theta} \end{aligned}$$

There is nothing controversial here. This is just the way conditional probability works. However, in this context,

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta) \, d\theta} \tag{3}$$

is often called *Bayes' rule* or *Bayes' theorem* because it was formulated by Thomas Bayes before conditional probability was formalized. It appeared in a paper published posthumously in 1763.

We have above what mathematicians call "abuse of notation" in using the same letter f for five different functions

$$\begin{gathered} f(\theta) \\ f(x) \\ f(x \mid \theta) \\ f(\theta \mid x) \\ f(x, \theta) \end{gathered}$$

(two marginals, two conditionals, one joint). To be pedantically correct, we should distinguish them by different letters or different decoration. The three that appear in Bayes' rule are called

- the *prior distribution* of the parameter $f(\theta)$, which expresses your uncertainty about the value of the parameter before you have seen the data,

- the *posterior distribution* of the parameter $f(\theta \mid x)$, which expresses your uncertainty about the value of the parameter after you have seen the data,
- the *data model*, also called the *likelihood*, $f(x \mid \theta)$ (see equation (2)), which the Bayesian shares with the frequentist.

Decorating the prior and posterior, Bayes' rule becomes

$$f_{\text{posterior}}(\theta \mid x) = \frac{f(x \mid \theta) f_{\text{prior}}(\theta)}{\int f(x \mid \theta) f_{\text{prior}}(\theta)\, d\theta}$$

It is not clear (to me) that this pedantry actually helps understanding.

1.3 Unnormalized Densities

We can also express Bayes' rule in words as follows. Express Bayes' rule as

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)} \tag{4}$$

and recall that one way of expressing the philosophical disagreement between Bayesians and frequentists is that

- the frequentist thinks x is random but θ is not random,
- and the Bayesian thinks θ is random but x is not random after the data have been seen. (Before the data were seen they were random, described by $f(x \mid \theta)$, but afterwards they are just the numbers they turned out to be and no more random than any other numbers, and similarly for categorical data. If we observe $x = 2.7$, there is nothing random about 2.7).

So the Bayesian thinks of x as a constant, hence of $f(x)$ in (4) as a constant. If we leave it out, we still have an unnormalized posterior distribution.

An *unnormalized* PDF is a function h that is nonnegative and integrates to some number that is not zero and not infinity. The relation between h and the corresponding PDF f is

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

and this normalization operation is just what Bayes' rule does, just what the operation of deriving a conditional from a joint does. Since the unnormalized

density h determines the normalized density f, it makes sense to talk about the unnormalized density as characterizing the distribution.

With that bit of jargon explained, we can return to (4) and drop the "constant" $f(x)$ obtaining

$$\text{unnormalized posterior} = \text{likelihood} \times \text{prior}$$
$$h(\theta \mid x) = f(x \mid \theta) f(\theta)$$

(recall that $f(x \mid \theta)$ is also the likelihood, see equation (2)) and we also realize that it does no harm if the prior is unnormalized here (we just get a different unnormalized posterior)

$$\text{unnormalized posterior} = \text{likelihood} \times \text{unnormalized prior}$$
$$h(\theta \mid x) = f(x \mid \theta) h(\theta)$$

1.4 Improper Priors

One more bit of arcana and we are done with the theory of Bayesian inference. We also realize that it does no harm, except perhaps philosophically, if the unnormalized prior doesn't even integrate, in which case it is called an *improper prior*. So long as applying Bayes' rule is concerned, it still works so long as the posterior is proper. If the integral is finite in

$$f(\theta \mid x) = \frac{f(x \mid \theta) h(\theta)}{\int f(x \mid \theta) h(\theta) \, d\theta} \tag{5}$$

it doesn't matter whether

$$\int h(\theta) \, d\theta \tag{6}$$

is finite or infinite. We can still interpret (5) as a posterior distribution. When we use (5) when (6) is infinite, we say h is an *improper prior*.

But this is philosophical abuse of probability theory. An improper prior does not characterize a *probability distribution* of the parameter before the data are seen (as a normalized prior does). This is because a function h that doesn't integrate to something finite isn't an unnormalized density. It doesn't characterize any probability distribution. And this has nothing to do with Bayesian inference, except that only Bayesians are crazy enough to use "improper" unnormalized densities.

And not all Bayesians are so crazy. Some Bayesians say improper priors are complete nonsense. Others blithely use them. They say improper priors are a harmless approximation to proper priors.

But the technical literature says otherwise.

- If you don't check whether the posterior integrates when using an improper prior, you can get nonsense, and it will be embarrassing when someone else discovers it. This happened to your humble author (Geyer, 1992, see "note added in proof" — actually my mistake was forgetting the Jacobian in a change-of-parameter rather than forgetting to check integrability, but it came to the same thing and was just as embarrassing). It has also happened to others (who shall remain nameless).

- Some researchers have tried to make a virtue out of errors and tried to show that, at least sometimes, some sense can be made of improper posteriors, but only in very special situations (Hobert and Casella, 1998; Gelfand and Sahub, 1999). Everybody agrees that improper posteriors are mostly nonsense.

- Even when the improper prior yields a proper posterior, technical issues arise.

  - Using an improper prior isn't really using probability theory (an improper prior doesn't characterize a probability distribution), hence if you naively try to reason probabilistically, you may get paradoxical results (Dawid, Stone, and Zidek, 1973).
  - More technically, some improper priors lead to inadmissible inferences (Eaton, 1992; Eaton, Hobert, Jones, and Lai, 2008; Shea and Jones, 2014) or to strongly inconsistent inferences (Eaton and Sudderth, 1999), two kinds of technical problems that I don't even want to try to describe, but which are really bad, and, moreover, figuring out whether an improper prior falls into either of these "really bad" classes is really hard (a PhD thesis worth or even harder).

With all of these problems (there are alligators in that swamp), improper priors are best avoided unless you know what you are doing (and few do). At the very least, academic weasel wording should be employed to hint at lack of guarantees that using an improper prior makes sense. For example, pedantically, (5) when h is improper is called the *formal Bayes' rule* ("formal" in the sense of having the form but not the content).

However, many enthusiastic and naive Bayesians use improper priors unthinkingly.

A technical solution to all issues with improper priors is to use finitely additive proper priors that mimic improper priors in most (all?) cases where improper priors do behave well (Sudderth, 1980), but since very few (not even a handful) of statisticians understand finitely additive probability theory, this proposal has not gotten much traction.

2 Monte Carlo

2.1 Ordinary Monte Carlo

Bayesian inference isn't always easy. The integral in Bayes' rule isn't always easy to do. In mathematical statistics (Stat 4102, 5102, 8102, 8111) one learns to do Bayesian inference in the few simple problems where it can be done by hand. Otherwise, nowadays, one does Bayesian inference by "computationally intensive methods," also called brute force and ignorance.

If one can simulate a probability distribution, one can answer any question at all about it by averaging over the simulations. Suppose X_1, X_2, ... are independent and identically distributed (IID) from some distribution. Then any parameter

$$\theta = E\{g(X)\} \tag{7}$$

can be estimated by an average over the simulations

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i). \tag{8}$$

The law of large numbers (LLN) guarantees that $\hat{\theta}_n$ can be gotten arbitrarily close to θ if only one does enough simulation (if only the Monte Carlo sample size n is large enough, if only one runs the computer long enough).

If $g(X)$ has finite variance, say

$$\sigma^2 = \operatorname{var}\{g(X)\}, \tag{9}$$

then the central limit theorem (CLT) tells us about the distribution of the Monte Carlo error

$$\hat{\theta}_n - \theta \approx \text{Normal}\left(0, \frac{\sigma^2}{n}\right). \tag{10}$$

Since this obeys the "square root law" (the asymptotic standard deviation is $\sigma/\sqrt{n}$), the precision one gets is limited. If a minute of computing time yields about three-significant-figure accuracy, then six-significant-figure accuracy

(1000 times the precision) would require 1000^2 times the time (almost two years of computing time).

If the parameter in question is a probability, then the theory is the same because probability is just expectation of indicator functions

$$\Pr(X \in A) = E\{I_A(X)\}$$

where I_A is the *indicator function* of the event A, defined by

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$$

So much for the theory of computation by IID computer simulation. It is just elementary statistics. The only difference is not in the math, but in the application, applying it to computer simulation rather than data about the real world.

The methodology described in this section — computation by averaging over IID simulations is called ordinary Monte Carlo (OMC) or (somewhat facetiously) good old-fashioned Monte Carlo (GOFMC).

The only problem with OMC is its limited applicability. There are zillions of different methods for simulating univariate distributions (Devroye, 1986). There are very few methods for simulating multivariate distributions. Uniform on a box can be easily simulated because the components of the random vector are independent. Multivariate normal can be easily simulated because they can are linear functions of normal random vectors with independent components. Uniform on a (hyper)sphere can be easily simulated by taking an isotropic multivariate normal, normalizing it to length one, and then multiplying it by the radius of the (hyper)sphere. Any multivariate distribution that factors as the product of univariate conditionals and marginals

$$f(x_1, \ldots, x_n) = f(x_1 \mid x_2, \ldots, x_n) f(x_2 \mid x_3, \ldots, x_n) \cdots f(x_{n-1} \mid x_n) f(x_n)$$

can be easily simulated by going right to left along the equation (simulate x_n, then simulate x_{n-1} from its conditional distribution given the simulated value of x_n, then simulate x_{n-2} from its conditional distribution given the variables already simulated, and so forth). The multinomial distribution factors as a product of binomials and the Dirichlet distribution factors as a product of betas, so these distributions are easily simulated. And that is about the end of the story for multivariate OMC. That leaves lots and lots of distributions having no OMC method available. And that includes most non-toy Bayesian posterior distributions.

2.2 Markov Chain Monte Carlo

2.2.1 Introduction

Markov chain Monte Carlo (MCMC) is the same, only different. For any multivariate distribution, a Markov chain that simulates it (in fact an infinite variety of Markov chains) are described by the the Metropolis-Hastings-Green algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, Teller, 1953; Hastings, 1970; Green, 1995) and its special case the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990).

What is a Markov chain? It is a sequence of random variables or random vectors X_1, X_2, ... having the property that the past and future are conditionally independent given the present: for any time t, the sets of variables or vectors $\{X_1, \ldots, X_{t-1}\}$ and $\{X_{t+1}, X_{t+2}, \ldots\}$ are conditionally independent given X_t. This means the Markov chain has an even simpler factorization than the one discussed in the preceding section

$$f(x_1, \ldots, x_n) = f(x_1) \prod_{t=2}^{n} f(x_t \mid x_{t-1})$$

When all of the conditional distributions on the right-hand side are the same we say the Markov chain has *stationary transition probabilities*, but this usually goes without saying. Authoritative books on Markov chain theory (Nummelin, 1984; Meyn and Tweedie, 2009) discuss only Markov chains with stationary transition probabilities and call them Markov chains (without qualification). The term MCMC always implies stationary transition probabilities unless explicitly stated otherwise.

It is easy to simulate a Markov chain. Simulate x_1, then simulate x_2 from its conditional distribution given the simulated value of x_1, and so forth, at time t simulate x_t from its conditional distribution given x_{t-1}.

Almost any method of simulation whatsoever, any computer program that looks like

```
Initialize x
repeat {
    Generate pseudorandom change to x
    Output x
}
```

Simulates a Markov chain so long as x denotes the entire state of the computer program (comprising all its variables). It simulates a Markov

chain with stationary transition probabilities so long as the code is not self-modifying (only x changes, the code that changes x remains the same). (We are treating pseudorandom as really random and not considering internals of pseudorandom number generators as part of the state.)

This shows us that MCMC is a very general simulation method. It can do any distribution. And almost any simulation method, when thought about the right way, is a special case of MCMC.

2.2.2 Equilibrium

Given specified transition probabilities (that is, given $f(x_t \mid x_{t-1})$, which is the same for all t, by stationary transition probabilities) we say an initial distribution $f(x_1)$ is *invariant* if X_2 has the same (marginal) distribution as X_1, in which case X_t will have the same (marginal) distribution for all t. Other names for an invariant distribution are *stationary distribution* and *equilibrium distribution*.

A Markov chain need not have an equilibrium distribution (consider a Markov chain with integer-valued state satisfying $X_{t+1} = X_t + 1$). But the Metropolis-Hastings-Green algorithm always describes a Markov chain having a specified equilibrium distribution (the distribution you want to sample).

If a Markov chain has a unique equilibrium distribution, then the LLN for Markov chains applies. If $X_1, X_2, \ldots$ are a Markov chain having a unique equilibrium distribution, and we want to calculate an expectation (7), where now the expectation is with respect to that equilibrium distribution, and we use the same estimator (8) that we used in OMC except now on the Markov chain, this is MCMC. We still have that (8) converges to (7) as the Monte Carlo sample size n goes to infinity.

Under some conditions, the CLT for Markov chains holds, and we have (10) holding but with σ^2 not being the population variance (9). The σ^2 in the Markov chain case is explained without unnecessary technicality by Geyer (2011, Section 1.8). Conditions for the Markov chain CLT are given by Chan and Geyer (1994), Roberts and Rosenthal (1997), Jones (2004), and Roberts and Rosenthal (2004), but are not easy reading. We will consider them beyond the scope of this course.

Suffice it to say that for nice enough Markov chains both the LLN and CLT hold, in which case the theory of MCMC is just like the theory of OMC except the formula for the variance in the CLT is different. This does not matter. It only matters that we know how to estimate the variance, and we will explain only one such method, the *method of batch means*. Geyer (2011,

Section 1.10) explains more about this method and also explains two other methods.

A section of the Markov chain $X_{k+1}, \ldots, X_{k+b}$ is called a *batch* and the length b is called the *batch length*. If the batch is long enough it has all of the properties of the whole chain. In particular, its average, the corresponding *batch mean*

$$\frac{1}{b} \sum_{i=k+1}^{b} g(X_i)$$

obeys the LLN and the CLT just like the whole chain. The only difference between the batch and the whole chain is that the lengths are different (b and n, respectively) and, consequently, the variances in their CLT's are different σ^2/b and σ^2/n. Also, if b is large enough, different batches will be nearly independent.

Hence the method of batch means. Choose b so that n is a multiple of b, and b is large but small enough so that n/b is also large. We abbreviate this $1 \ll b \ll n$. Then calculate the batch means for consecutive batches of length b

$$\hat{\theta}_{n,k} = \frac{1}{b} \sum_{i=1}^{b} g(X_{(k-1)b+i}), \qquad k = 1, \ldots, n/b.$$

For sufficiently large b, these are asymptotically IID with mean θ, the quantity to be estimated, (Geyer, 1992, Section 3.2), so the usual z confidence interval or (if n/b is not large) t confidence interval will cover θ with the stated coverage probability.

It is important that the batch length b be as large as possible. It is nice, but not necessary if one uses t confidence intervals rather than z confidence intervals to characterize Monte Carlo error, to have n/b large too. But it is far more important that b be large than n/b be large. Of course, if n is really large, one can have both. If not, make b large and n/b moderate.

References

Chan, K. S. and Geyer, C. J. (1994). Discussion of the paper by Tierney. *Annals of Statistics*, **22**, 1747–1758.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **35**, 189–233.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York. This authoritative reference is unfortunately out of print. The author has put it on-line at `http://www.nrbook.com/devroye/`.

Eaton, M. L. (1992). A statistical diptych: Admissible inferences–recurrence of symmetric Markov Chains. *Annals of Statistics*, **20**, 1147–1179.

Eaton, M. L., Hobert, J. P., Jones, G. L., and Lai, W.-L. (2008). Evaluation of formal posterior distributions via Markov chain arguments. *Annals of Statistics*, **36**, 2423–2452.

Eaton, M. L., and Sudderth, W. D. (1999). Consistency and strong inconsistency of group-invariant predictive inferences. *Bernoulli*, **5**, 833–854.

Gelfand, A. E., and Sahub, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, **94**, 247–253.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, edited by Brooks, S., Gelman, A., Jones, G., and Meng, X.-L., pp. 3–48. Chapman & Hall/CRC, Boca Raton, FL.

Geyer, C. J., and Johnson, L. T. (2015). R package `mcmc` (Markov Chain Monte Carlo), version version 0.9-4. `http://cran.r-project.org/package=mcmc`. To install use the R function `install.packages` or use the package management menu on the R app.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hobert, J. P., and Casella, G. (1998). Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics*, **7**, 42–60.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, **1**, 299–320.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, second edition. Cambridge: Cambridge University Press.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge: Cambridge University Press.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electronic Communications in Probability*, **2**, 13–25.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys*, **1**, 20–71.

Shea, B. P., and Jones, G. L. (2014). Evaluating default priors with a generalization of Eaton's Markov chain *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, **50**, 1069–1091.

Sudderth, W. D. (1980). Finitely additive priors, coherence and the marginalization paradox. *Journal of the Royal Statistical Society, Series B*, **42**, 339–341.