Gov 2002 - Section 1: Introduction to Bayesian Statistics

Connor Jerzak

September 15, 2016

Harvard University, Department of Government

Why Use Bayesian Statistics?

Setup and Motivation for Bayesian Statistics

- Let D denote the data, and θ denote an arbitrary parameter vector.

$$
\begin{aligned}
\Pr(\theta|D) &= \frac{\Pr(D,\theta)}{\Pr(D)}; \\
&= \frac{\Pr(D|\theta)\cdot\Pr(\theta)}{\Pr(D)}; \\
&= \frac{\Pr(D|\theta)\cdot\Pr(\theta)}{\int_\theta \Pr(D|\theta)d\theta}.
\end{aligned}
\tag{1}
$$

The Prior. $\Pr(\theta)$ corresponds to initial beliefs about θ. The properties of $\Pr(\theta)$ have long been at the center of disputes about statistics. Historical trajectory: flat priors $\rightarrow$ uninformative priors $\rightarrow$ weakly informative priors $\rightarrow$ informative priors.

The Likelihood. $\Pr(D|\theta)$. This object is at the central of Frequentist statistics. In Frequentist statistics, we find the θ^* that maximizes the probability of having observed the data. The problem with this is that this optimization procedure ignores information, and is often overly sensitive to noise in the data.

The probability of the data. $\Pr(D)$ corresponds to the probability of the data, D. Notice that $\Pr(D) = \int_\theta \Pr(D|\theta)d\theta$. In words, $\Pr(D)$ corresponds to the overall probability of observing B, without regard to the selection of θ.

The Posterior. $\Pr(\theta|D)$ corresponds to the posterior distribution. It captures the new belief about θ having accounted for the data.

Applications of Bayesian Statistics

1. **Incorporate prior information into parameter estimates**.

2. **Obtain a distribution over $\hat{\theta}$, not only a point estimate**.

3. **Regularizing statistical models**. In MLE, we find $\hat{\theta}$ by selecting the value for $\hat{\theta}$ that maximizes the probability of seeing the observed data. This technique is prone to *overfitting*.

- *On overfitting*. When overfitting occurs, our model will likely be sensitive to small perturbations in our dataset: due to chance alone, one covariate might be highly correlated with our outcome, generating extreme estimates of $\hat{\theta}$. Bayesian regression models can be less sensitive to this kind of overfitting, since our point estimate for $\hat{\theta}$ will be shrunk towards the prior mean (often 0).

Gibbs Sampling in Theory

Why use a Gibbs Sampler

- In Bayesian statistics, we want to obtain $\Pr(\theta|D)$, or $\mathbb{E}[\theta|D]$. For example, θ might be the weights in a Bayesian neural network, and I want to find the posterior mean for θ given my dataset.
- If I knew the posterior analytically, I could directly use this in my calculations. But usually, the posterior is too complicated to obtain using analytical methods.
- As a result, Monte Carlo (i.e. simulation-based) methods must be used. With these methods, we are going to use random numbers to get samples from $\Pr(\theta|D)$. When I have these samples in hand, I can take the mean (to approximate $\mathbb{E}[\theta|D]$), or could measure any other property of $\mathbb{E}[\theta|D]$. The theory of Monte Carlo methods states that my numerical estimate of $\mathbb{E}[\theta|D]$ will be within an ϵ of the true expectation if my algorithm has converged.

Why does Gibbs sampling work?

1. Recall that we want to generate random draws from $\Pr(\theta|D)$.
2. Recall also that $\Pr(\theta|D) \propto \Pr(D, \theta) = \Pr(D|\theta) \cdot \Pr(\theta)$.
3. The theory behind Gibbs sampling states that we can generate draws from $\Pr(\theta|D)$ if we can sample from $\Pr(\theta|D)$ and $\Pr(D)$. Gibbs sampling thus most applicable when $\Pr(D, \theta)$ is of an unknown form, but the conditional distributions are easier to sample from.
4. "Gibbs sampling can be thought of as a practical implementation of the fact that knowledge of the conditional distributions is sufficient to determine a joint distribution" (Casella & George, 1992).

The Gibbs sampler algorithm

Let Γ denote a p-dimensional random variable.

$$\Gamma = (\Gamma_1, \Gamma_2, ..., \Gamma_p)^T.$$

1. Initialize $\Gamma_p^{(0)} \sim g_p(x)$ for $p = 1, 2, ..., P$.
2. For iteration $i = 1, 2, ..., n$:

2.1 For component $p = 1, 2, ..., P$:

$$\Gamma_p^{(i)} \sim \Pr(\Gamma_p^{(i)} | \Gamma_{-p}^{(i-1)}).$$

- This works because $\Pr(\Gamma_1, \Gamma_2, ..., \Gamma_P)$ is equivalent to $\Pr(\Gamma_1 | \Gamma_2, ..., \Gamma_P) \cdot \Pr(\Gamma_2 | \Gamma_3, ..., \Gamma_P) \cdot [\cdots] \cdot \Pr(\Gamma_P)$.

Extensions of the Gibbs sampler

- *Blocked Gibbs sampler*. In a blocked Gibbs sampler, we sample from the joint distribution of two or more variables conditional on all other variables. This is a powerful technique used in some hidden Markov models, as well as in applications of the latent Dirichlet allocation.

A Gibbs Sampler for Bayesian Regression

A Gibbs Sampler for Bayesian Regression

- Assume
$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
where
$$\beta_0 \sim N(0, 1/\tau_{\beta_0})$$
$$\beta_1 \sim N(0, 1/\tau_{\beta_1})$$
$$\tau \sim \text{Gamma}(\alpha, \beta)$$
- Notice that $\alpha, \beta, \tau_{\beta_0}$, and τ_{β_1} are fixed hyperparameters.

A Gibbs Sampler for Bayesian Regression, Continued

- From the Normal-Normal conjugacy, (as well as Gamma-gamma conjugacy), we know

$$f(\beta_0|\beta_1,\tau,Y_1,...,Y_n) \sim N\left(\frac{\tau}{n\tau+\tau_{\beta_0}}\sum_{i=1}^{N}(y_i-\beta_1 x_i), \quad \frac{1}{n\tau+\tau_{\beta_0}}\right)$$

$$f(\beta_1|\beta_0,\tau,Y_1,...,Y_n) \sim N\left(\frac{\tau\sum_{i=1}^{n}(Y_i-\beta_0)x_i}{\tau\sum_{i=1}^{n}(x_i^2)+\tau_{\beta_1}}, \quad \frac{1}{\tau\sum_{i=1}^{n}(x_i^2)+\tau_{\beta_1}}\right)$$

$$f(\tau|\beta_0,\beta_1,Y_1,....,Y_n) \sim$$

$$\mathrm{Gamma}\left(\alpha+\frac{n}{2},\beta+\frac{1}{2}\cdot\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 x_i)^2\right)$$

```
#hyperparameters
alpha <- 0.001
beta <- 0.001

#book keeping
n_iter <- 5000
n <- length(y) #keep track of the number of observations

#starting values for the parameters of interest
beta0 <- mean(y)#starting values for beta0
beta1 <- 0#starting value for beta1
tau <- 1
```

```
result <- matrix(nrow = niter, ncol = 3)#matrix to store the results
for (i in 1:niter) #this loop contains the Gibbs sampler itself
{
  #conditional draw of beta0
  beta0 <- rnorm(1, mean = (tau/(n * tau + tau.beta0)) *
                                 sum(y - beta1 * x),
                  sd = 1/sqrt(n * tau + tau.beta0))

  #conditional draw of beta1
  beta1 <- rnorm(1, mean = (tau * sum((y - a) * x)) /
                                 (tau * sum(x^2) + tau.beta1),
                  sd = 1/sqrt(tau * sum(x^2) + tau.beta1))

  #conditional draw of tau
  tau <- rgamma(1, shape = alpha + n/2,
                 rate = beta + 0.5 * sum((y - beta0 - beta1 * x)^2))
result[i, ] <- c(a, beta1, tau)
}
```

Convergence metrics

Convergence metrics

- **Gelman-Rubin**. "Uses parallel chains with dispersed initial values to test whether they all converge to the same target distribution. Failure could indicate the presence of a multi-mode posterior distribution (different chains converge to different local modes) or the need to run a longer chain (burn-in is yet to be completed)." This metric uses the fact that the pooled variance should be equal to the marginal posterior variance of θ under convergence.
- **Geweke**. "Tests whether the mean estimates have converged by comparing means from the early and latter part of the Markov chain."
- **Autocorrelation metrics**. "Autocorrelation metrics capture dependency among Markov chain samples." It is common practice to present autocorrelation plots when evaluating whether a sampler has reached a stationary distribution.

The Upshot

The Upshot

- Bayesian statistics depends on finding some joint distribution, $\Pr(\theta, D)$.
- It is often impossible to find $\Pr(\theta, D)$ analytically, so Monte Carlo methods are usually used.
- The Gibbs sampler is a common MCMC algorithm, and provides an efficient approximation of the joint distribution, provided we are able to sample from each of the conditional distributions.
- There are various methods for evaluating whether an MCMC algorithm has reached its stationary distribution.

Acknowledgements

- These slides build from material found in Geman & Geman, 1984; Gelfand & Smith, 1990; Gelman & Rubin (1992); Pursue University materials, and other sources.

Appendix

The Metropolis-Hastings Algorithm

- *Goal*: Construct a dependent Markov Chain that converges to a stationary distribution matching the posterior of interest. In a Markov chain, we let $t(x, y)$ denote the probability of transitioning ("jumping") from state x to state y. So, we want the Markov Chain defined by $t(x, y)$ to have some $f(x)$ as its stationary distribution.

Requirements

- $f(y) = \sum_x f(x)t(x,y)$ or, less generally, $f(y)t(y,x) = f(x)t(x,y)$. This condition is described as the "detailed balance" or "time reversibility" condition.
- We want to find some probability distribution $j(x,y)$ such that, given x, we "jump" to y with probability $\alpha(x,y)$, and stay put at x with probability $1 - \alpha(x,y)$.

More details

- There are two possibilities.

1. $y = x$. In this case, $f(y)t(y,x) = f(x)t(x,y)$ holds trivially because $f(x)t(x,x) = f(x)t(x,x)$.

2. $y \neq x$. In this case, let $t(x,y) = j(x,y)\alpha(x,y)$. Detailed balance gives:

$$
\begin{aligned}
&f(y) \cdot \mathsf{Pr}(\text{Transition from } y \text{ to } x) \cdot \mathsf{Pr}(\text{Accept jump from } y \text{ to } x) = \\
&\qquad f(x) \cdot \mathsf{Pr}(\text{Transition from } x \text{ to } y) \cdot \mathsf{Pr}(\text{Accept jump from } x \text{ to } y)
\end{aligned} \tag{2}
$$

More details, Continued

$$
\begin{gathered}
f(x)t(y,x)\alpha(y,x) = f(x)t(x,y)\alpha(x,y); \\
\Longrightarrow \\
\frac{\alpha(x,y)}{\alpha(y,x)} = \frac{f(y)j(y,x)}{f(x)j(x,y)},
\end{gathered} \tag{3}
$$

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