Bayesian Inference

Giselle Montamat

Harvard University

Spring 2020

A 1-slide summary of Bayesian inference

1. Beliefs about an unknown parameter θ:

   Prior distribution: $\theta \sim \pi$

2. Update beliefs using Baye's rule:

   $$\textbf{Posterior distribution:}\ \theta \sim f(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{f(D)} = \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta)d\theta}$$

Boils down to calculating (or approximating) such posterior distribution

- If $\pi(\theta)$ is a conjugate prior for $f(D|\theta)$, can find posterior analytically.
- Otherwise, find approximation with simulation method: MCMC (Markov Chain Monte Carlo).
  (Importance weighting is another simulation method that allows to get an estimate of posterior expectation of a function $h(\theta)$).

Conjugate priors

If the posterior distribution $f(\theta|D)$ is in the same family of distributions as the prior distribution $\pi(\theta)$, the prior and posterior are then called *conjugate distributions*, and the prior is called a *conjugate prior* for the likelihood function. In these cases, we can analytically derive the posterior density $f(\theta|D)$.

For example:

$$\text{Likelihood: } D|\theta \sim N(\theta, \Sigma)$$

$$\text{Prior: } \theta \sim N(\mu, \Omega)$$

$$\Rightarrow \text{Posterior: } \theta|D = d \sim N(\mu + \Omega(\Sigma + \Omega)^{-1}(d - \mu), \Omega - \Omega(\Omega + \Sigma)^{-1}\Omega)$$

Conjugate priors

Exercise: Find the posterior distribution in the following cases:

1. Likelihood: $X_1, ..., X_n \sim Be(\theta)$, Prior: $\theta \sim Beta(\alpha, \beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
2. Likelihood: $X \sim Bin(j, \theta)$, Prior: $\theta \sim Beta(\alpha, \beta)$
3. Likelihood: $X_1, ..., X_n \sim P(X_i = j) = \theta_j$ where $X_i \in \{1, 2, ..., k\}$ and $\sum_j \theta_j = 1$, Prior: $\theta \sim Dir(\alpha_1, ..., \alpha_k) = const \cdot \theta_1^{\alpha_1-1} \cdot \theta_2^{\alpha_2-1} \cdot ... \cdot \theta_k^{\alpha_k-1}$
4. Likelihood: $X_1, ..., X_n \sim exp(\alpha) = \alpha e^{-\alpha X_i}$, Prior: $\alpha \sim \Gamma(a, b) = const \cdot \alpha^{a-1} \cdot e^{-b\alpha}$
5. Likelihood: $X_1, ..., X_n \sim Pareto(\underline{x}, \alpha) = \alpha\frac{\underline{x}^{\alpha}}{x^{\alpha+1}}$, Prior: $\alpha \sim \Gamma(a, b)$

For more cases, see Wikipedia page on Conjugate Prior.

MCMC

Idea: simulation-based technique for generating draws from a target density $\phi(\theta)$.

In particular, for our Bayesian framework:

$$\text{Target density: } \underbrace{\pi(\theta|D)}_{\equiv\phi(\theta)} = \underbrace{\frac{1}{f(D)}}_{\equiv[\int\tilde{\phi}(\theta)d\theta]^{-1}>0} \underbrace{f(D|\theta)\pi(\theta)}_{\equiv\tilde{\phi}(\theta)}$$

-It is difficult to compute $f(D)$ numerically if dimension of θ is high, thus difficult to compute $\phi(\theta)$ numerically.

-Method allows us to just use only $\tilde{\phi}(\theta)$ to make draws from $\phi(\theta)$.

-How? Construct a Markov Chain on parameter space Θ, $\{\theta^j\}_{j=1}^{\infty}$ whose stationary distribution is $\phi(\theta)$. Starting from θ^0, simulate a long Markov Chain $\theta^1, ..., \theta^J$. After chain has reached it stationary distribution, any further draws will be distributed according to $\phi(\theta)$.

MCMC

Makov Chain: a stochastic process that satisfies the Markov property:

$$\{\theta^j\}_{j=1}^{\infty} \text{ where } \theta^j \in \Theta$$

$$\theta^{j+1}|(\theta^j, \theta^{j-1}, ...) \sim \theta^{j+1}|\theta^j$$

Can be understood as a rule for (stochastically) stepping through elements of Θ, where the next step is determined only by the current position:

$$\theta^j : \text{ current position}$$

$$\zeta : \text{proposal for next step drawn from a distribution } \zeta \sim p(\zeta|\theta^j)$$

$$\alpha(\zeta|\theta^j) : \text{ acceptance probability of the proposal}$$

$$\theta^{j+1} : \text{ next step}$$

So, if the chain is at θ^j, the next step will be:

$$\theta^{j+1} = \begin{cases} \zeta \text{ with prob } \alpha(\zeta|\theta^j) \\ \theta^j \text{ with prob } 1 - \alpha(\zeta|\theta^j) \end{cases}$$

MCMC

Such a Markov Chain of random draws θ^j converges to a stationary distribution $\phi(\theta)$ as $j \to \infty$ (under mild conditions). In other words, both θ^j and θ^{j+1} are distributed according to $\phi(\theta)$ as $j \to \infty$ (this is why you have a "burnt-in period": you throw away the first k draws of your simulated Markov Chain because you still haven't reached convergence).

So, basically, need to find the acceptance probability $\alpha(\zeta|\theta)$ (given a proposal density $p(\zeta|\theta)$) such that his holds.

Metropolis-Hastings is one possible algorithm that achieves this.

MCMC: Metropolis-Hastings

Given $\phi(\theta)$ and $p(\zeta|\theta)$, suggested acceptance probability:

$$\alpha(\zeta|\theta^j) = min\left\{1, \frac{\phi(\zeta)p(\theta^j|\zeta)}{\phi(\theta^j)p(\zeta|\theta^j)}\right\} = min\left\{1, \frac{\tilde{\phi}(\zeta)p(\theta^j|\zeta)}{\tilde{\phi}(\theta^j)p(\zeta|\theta^j)}\right\}$$

Note 1: $\frac{\phi(\zeta)}{\phi(\theta^j)} = \frac{\tilde{\phi}(\zeta)}{\tilde{\phi}(\theta^j)}$, so only need to work with $\tilde{\phi}$!

Note 2:

- If $\frac{\tilde{\phi}(\zeta)p(\theta^j|\zeta)}{\tilde{\phi}(\theta^j)p(\zeta|\theta^j)} \geq 1$, $\alpha(\zeta|\theta^j) = 1$ so always accept proposal.
- If $\frac{\tilde{\phi}(\zeta)p(\theta^j|\zeta)}{\tilde{\phi}(\theta^j)p(\zeta|\theta^j)} < 1$, $0 < \alpha(\zeta|\theta^j) < 1$ so accept with some probability.

Note 3: If $p(\theta^j|\zeta) = p(\zeta|\theta^j)$ we are always accepting proposal if this leads to an increase in the posterior density and we are sometimes accepting the proposal if it leads to a decrease. This is *not* trying to maximize the posterior density, but rather trying to stay within regions of values of θ for which the posterior is high and sometimes visit those regions for which it is low.

MCMC: Metropolis-Hastings

In particular, for our Bayesian framework:

$$\phi(\theta) = f(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{f(D)}$$

$$\phi(\zeta) = f(\zeta|D) = \frac{f(D|\zeta)\pi(\zeta)}{f(D)}$$

$$\alpha(\zeta|\theta^j) = min\left\{1, \frac{f(\zeta|D)p(\theta^j|\zeta)}{f(\theta^j|D)p(\zeta|\theta^j)}\right\} = min\left\{1, \frac{f(D|\zeta)\pi(\zeta)p(\theta^j|\zeta)}{f(D|\theta^j)\pi(\theta^j)p(\zeta|\theta^j)}\right\}$$

MCMC: Metropolis-Hastings

Algorithm:

1. Pick θ^0
2. For each $j = 0, ..., J-1$
    1. Draw $\zeta \sim p(\zeta|\theta^j)$
    2. Draw $u \sim U[0,1]$
    3. $\theta^{j+1} = \begin{cases} \zeta \;\text{ if }\; u \leq \frac{f(D|\zeta)\pi(\zeta)p(\theta^j|\zeta)}{f(D|\theta^j)\pi(\theta^j)p(\zeta|\theta^j)} \equiv \rho^j \;\text{ (i.e., with prob }\; \rho^j) \\ \theta^j \;\text{ if }\; u > \frac{f(D|\zeta)\pi(\zeta)p(\theta^j|\zeta)}{f(D|\theta^j)\pi(\theta^j)p(\zeta|\theta^j)} \equiv \rho^j \;\text{ (i.e., with prob }\; 1-\rho^j) \end{cases}$
3. Drop the first k values of your chain and use the empirical distribution of $\theta^{k+1}, ..., \theta^J$ as the estimate for the posterior distribution $f(\theta|D)$.

Note: If $u \sim U[0,1]$:

$$P(u \leq \rho^j) = \rho^j \;\text{ if }\; \rho^j \leq 1$$

$$P(u \leq \rho^j) = 1 \;\text{ if }\; \rho^j > 1$$

MCMC: Metropolis-Hastings

So the three ingredients that we need for the algorithm are: the likelihood, the prior and the proposal density.

To complete our algorithm, need to pick the proposal density $p(.|.)$. One alternative is to pick a symmetric proposal, which implies:

$$p(\zeta|\theta^j) = p(\theta^j|\zeta)$$

For example, the Random Walk Metropolis-Hastings suggests a normal density.

Note: $p(\zeta|\theta^j) = p(\theta^j|\zeta) = \tilde{p}(\zeta - \theta^j)$ when $p(.|.)$ is a multivariate normal density.

MCMC: Metropolis-Hastings

Exercise: Problem Set 7, Exercise 2 asked you to implement the Random Walk Metropolis-Hastings algorithm.

$$\text{Likelihood: } f(D|\theta): \begin{bmatrix} D_1 \\ D_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \theta_1\theta_2 \\ \theta_1/\theta_2 \end{bmatrix}, I_{2\times 2} \right)$$

$$\text{Prior: } \pi(\theta): \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N(0, 10 \times I_{2\times 2})$$

$$\text{Proposal: } f(\zeta|\theta): \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} | \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \frac{1}{4} \times I_{2\times 2} \right)$$

Remember the functional form for a multivariate normal density, where z and γ are $k \times 1$:

$$f(z|\gamma) = (2\pi)^{-\frac{k}{2}} det(\Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(z-\gamma)'\Sigma(z-\gamma)}$$

MCMC: Metropolis-Hastings

Algorithm:

1. Pick θ^0
2. For each $j = 0, ..., J-1$
    1. Draw $\zeta \sim p(\zeta|\theta^j)$
    2. Draw $u \sim U[0,1]$
    3. $\theta^{j+1} = \begin{cases} \zeta & \text{if } u \leq \frac{f(D|\zeta)\pi(\zeta)}{f(D|\theta^j)\pi(\theta^j)} \\ \theta^j & \text{if } u > \frac{f(D|\zeta)\pi(\zeta)}{f(D|\theta^j)\pi(\theta^j)} \end{cases}$

$$f(D|\theta) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} - \begin{bmatrix} \theta_1\theta_2 \\ \theta_1/\theta_2 \end{bmatrix}\right)'\left(\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} - \begin{bmatrix} \theta_1\theta_2 \\ \theta_1/\theta_2 \end{bmatrix}\right)}$$

$$f(D|\zeta) = \frac{1}{2\pi} e^{-\frac{1}{2}\left(\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} - \begin{bmatrix} \zeta_1\zeta_2 \\ \zeta_1/\zeta_2 \end{bmatrix}\right)'\left(\begin{bmatrix} D_1 \\ D_2 \end{bmatrix} - \begin{bmatrix} \zeta_1\zeta_2 \\ \zeta_1/\zeta_2 \end{bmatrix}\right)}$$

$$\pi(\theta) \sim N(0, 10 \times I_{2\times 2}) \qquad \pi(\zeta) \sim N(0, 10 \times I_{2\times 2})$$

Importance weighting

A simpler simulation method that allows to simulate posterior expectation (rather than overal posterior distribution):

- Want to compute $E_{f(\theta|D)}[r(\theta)|D]$ (posterior expectation of $r(\theta)$)
- Can be expressed:
  $E_{f(\theta|D)}[r(\theta)|D] = \int r(\theta) f(\theta|D) d\theta = \int r(\theta) \frac{f(D|\theta)\pi(\theta)}{\int f(D|\theta)\pi(\theta) d\theta} d\theta = \frac{\int r(\theta) f(D|\theta)\pi(\theta) d\theta}{\int f(D|\theta)\pi(\theta) d\theta}$
- Method suggests a way of taking draws from numerator and from denominator (separately). For a continuous posterior distribution, propose a $h(\theta)$ that is a continuous density that's everywhere positive:

  Numerator: $\int \frac{r(\theta) f(D|\theta)\pi(\theta)}{h(\theta)} h(\theta) d\theta = E_{h(\theta)}\left[\frac{r(\theta) f(D|\theta)\pi(\theta)}{h(\theta)}\right]$

  Denominator: $\int \frac{f(D|\theta)\pi(\theta)}{h(\theta)} h(\theta) d\theta = E_{h(\theta)}\left[\frac{f(D|\theta)\pi(\theta)}{h(\theta)}\right]$

  You “re-weight” using $\frac{1}{h(\theta)}$.

Importance weighting

- Simulation: take an iid sample of draws from $h(\theta)$ and estimate expectations with sample means:

  Numerator: $\frac{1}{J}\sum_{j=1}^{J}\left[\frac{r(\theta_j)f(D|\theta_j)\pi(\theta_j)}{h(\theta_j)}\right]$

  Denominator: $\frac{1}{J}\sum_{j=1}^{J}\left[\frac{f(D|\theta_j)\pi(\theta_j)}{h(\theta_j)}\right]$

- So take as consistent estimate of posterior expectation:

$$\frac{\frac{1}{J}\sum_{j=1}^{J}\left[\frac{r(\theta_j)f(D|\theta_j)\pi(\theta_j)}{h(\theta_j)}\right]}{\frac{1}{J}\sum_{j=1}^{J}\left[\frac{f(D|\theta_j)\pi(\theta_j)}{h(\theta_j)}\right]}$$

Importance weighting

- While consistent estimates, computational performance very sensitive to choice of $h(\theta)$. In particular, need many draws J to attain convergence if $h(\theta)$ is very different from $f(D|\theta)\pi(\theta)$:
  - I will get many θ_j draws of values for which $h(\theta)$ has high density; these are weighted low (since $h(\theta_j)$ large) and also $f(D|\theta_j)\pi(\theta_j)$ is small.
  - I will get few draws θ_j of values for which $h(\theta)$ has low density; these are highly weighted (since $h(\theta_j)$ is small) and also $f(D|\theta_j)\pi(\theta_j)$ is large.

So we get many small values and a few very large values of $\frac{f(D|\theta_j)\pi(\theta_j)}{h(\theta_j)}$. High variance means I need more simulations to achieve convergence to expectation. Numerical performance depends on proposal $h(\theta)$.

A few final comments

- In MCMC, can sample within chain $\theta^{k+1}, ..., \theta^J$ to reduce serial correlation.
- Convergence quite sensitive to proposal $h(\theta)$ in importance weighting: need many simulations if $h(\theta)$ very different from posterior density.

  With MCMC, a "bad" proposal $p(\zeta|\theta^j)$ can also lead to slow convergence. Rule of thumb: tune it so that you're not always accepting or always rejecting. But! Pset exercise was an example in which, because the posterior density was bimodal, you wanted a proposal that would explore space by proposing extreme moves (high variance); your rate of acceptance was "lower" than rule of thumb but you achieved convergence and didn't get stuck in one of the two regions.

Estimators based on posterior distribution

We found/simulated the posterior distribution, which summarizes everything we know about θ after seeing the data. Now we can use summary statistics of this posterior to derive estimators and do inference.

- **Mode of posterior distribution ("maximum a posteriori")**

$$\hat{\theta} = \underset{\gamma}{arg\,max}\ \ f(\gamma|D)$$

  Note: $\underset{\gamma}{arg\,max}\ \ f(\gamma|D) = \underset{\gamma}{arg\,max}\ \ f(D|\gamma)\pi(\gamma)$. If flat prior ($\pi(\theta) = 1$) this leads to the MLE. A non-flat prior will weight the different regions of the parameter space (more intuitions to come in exercises of next section).

- **Mean of posterior distribution:**

$$\hat{\theta} = E_{f(\theta|D)}[\theta|D = d]$$

  Next section: we'll discuss how the posterior mean can be shown to be the Bayesian estimator based on quadratic square loss, i.e, the decision function that is the best according to the Bayesian criteria in decision theory, and how this relates to a variance-bias trade-off.

Choosing the prior

We discussed in lecture criteria for choosing a prior: comes out naturally based on specific context, subjective, consensus/knowledge, mathematical convenience (conjugate priors), default rule (flat prior), use data to learn about it (Empirical and Hierarchical Bayes).

- **Empirical Bayes**: goes one step further and estimates the prior using data; specifically, assume model for prior ($\pi(.|\gamma)$) and estimate the "hyperparameter" (γ) by maximizing marginal likelihood ($f(D|\gamma)$) (aka, do MLE). Then, compute the posterior distribution and, for example, its mean,c by plugging-in this estimated hyperparameter into the prior.

$$\theta \sim \pi(.|\gamma)$$

$$f(D|\gamma) = \int f(D|\theta)\pi(\theta|\gamma)d\theta$$

$$\hat{\gamma} = arg\max_{\gamma} \; f(D|\gamma)$$

$$f(\theta|D) = \frac{f(D|\theta)\pi(\theta|\hat{\gamma})}{\int f(D|\theta)\pi(\theta|\hat{\gamma})d\theta}$$

$$\hat{\theta}_{EB} = E_{f(\theta|D;\hat{\gamma})}[\theta|D=d] = \int \theta \frac{f(D|\theta)\pi(\theta|\hat{\gamma})}{\int f(D|\theta)\pi(\theta|\hat{\gamma})d\theta} d\theta$$

Choosing the prior

- **Hierarchical Bayes**: goes one step further and imposes a prior on the hyperparameter (called a "hyperprior"); then compute the posterior density of θ and hyperparameter γ as usual.

$$\theta \sim \pi(.|\gamma)$$

$$\gamma \sim \tilde{\pi}$$

$$f(\theta, \gamma|D) = \frac{f(D|\gamma)\pi(\theta|\gamma)\tilde{\pi}(\gamma)}{\int\int f(D|\gamma)\pi(\theta|\gamma)\tilde{\pi}(\gamma) d\theta d\gamma}$$