COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Recall: Markov Chains

- A Markov chain is a sequence of random variables $\{X^{(1)}, X^{(2)}, \dots\}$ such that the mechanism that generates $X^{(s+1)}$ depends on the value of $X^{(s)}$ but not on $\{X^{(s-1)}, x^{(s-2)}, \dots, X^{(1)}\}$.
- For a Markov chain "given the present, the future is independent of the past."
- Both the Gibbs sampler and the Metropolis algorithm are ways of generating Markov chains that approximate a target probability distribution $f(x)$.
- In Bayesian analysis $f(x)$ is typically the posterior distribution we want to explore, but the algorithms we considered can be used more broadly.

Recall: Gibbs sampler

Let $f(x) = f(u, v)$. For example, we would have $u = \theta$, $v = \sigma^2$ and $f(u, v) = f(\theta, \sigma^2|\boldsymbol{x})$ in the Normal model.
Recall the Gibbs sampler: Given $(u^{(s)}, v^{(s)})$, a new value is generated as follows:

1. Update u: sample $u^{(s+1)} \sim f(u|\ v^{(s)})$
2. Update v: sample $v^{(s+1)} \sim f(v|\ u^{(s+1)})$

Alternatively, you can update the variables by first sampling v and then u. If you have more than two variables, they can be sampled sequentially in any order, or even randomly.

Recall: Metropolis algorithm

The Metropolis algorithm proposes changes to $X = (U, V)$ and then accepts or rejects those changes based on $f(x)$. An alternative way to implement the Metropolis algorithm is to propose and then accept or reject changes to one element at a time:

1. Update u:

a) Sample $u^* \sim J_u(u|u^{(s)})$

b) Compute $r = \frac{f(u^*, v^{(s)})}{f(u^{(s)}, v^{(s)})}$

c) Set $u^{(s+1)} = u^*$ or $u^{(s)}$ with probs. min(1, r) and max (0, $1 - r$).

2. Update v:

a) Sample $v^* \sim J_v(v|v^{(s)})$

b) Compute $r = \frac{f(u^{(s+1)}, v^*)}{f(u^{(s+1)}, v^{(s)})}$

c) Set $v^{(s+1)} = v^*$ or $v^{(s)}$ with probs. min(1, r) and max (0, $1 - r$).

Metropolis - Hastings

The Metropolis-Hastings algorithm generalizes both approaches by allowing arbitrary proposal distributions.

1. Update u:

a) Sample $u^* \sim J_u(u|u^{(s)}, v^{(s)})$

b) Compute $r = \frac{f(u^*,v^{(s)})}{f(u^{(s)},v^{(s)})} \times \frac{J_u(u^{(s)}|u^*, v^{(s)})}{J_u(u^*|u^{(s)}, v^{(s)})}$

c) Set $\mathrm{u}^{(\mathrm{s}+1)} = u^*$ or $u^{(\mathrm{s})}$ with probs. min(1, r) and max (0, 1 – r).

2. Update v:

a) Sample $v^* \sim J_v(v|\, u^{(s+1)}, v^{(s)})$

b) Compute $r = \frac{f(u^{(s+1)},v^*)}{f(u^{(s+1)},v^{(s)})} \times \frac{J_v(v^{(s)}|u^{(s+1)}, v^*)}{J_v(v^*|u^{(s+1)}, v^{(s)})}$

c) Set $v^{(s+1)} = \mathrm{v}^*$ or $\mathrm{v}^{(\mathrm{s})}$ with probs. min(1, r) and max (0, 1 – r).

Properties

- In the Metropolis-Hastings algorithm the proposal distributions J_u and J_v are not required to be symmetric.
- The Metropolis-Hastings algorithm looks a lot like the Metropolis algorithm, except that the acceptance ratio contains an extra factor, the ratio of the probability of generating the current value from the proposed to the probability of generating the proposed from the current. This can be viewed as a "correction factor:" If a value u^* is much more likely to be proposed than the current value $u^{(s)}$, then we must down-weight the probability of accepting u^* accordingly, otherwise the value u^* will be overrepresented in our sequence.
- That the Metropolis algorithm is a special case of the Metropolis-Hastings algorithm is easy to see: If J_u is symmetric, meaning that $J(u_a|u_b, v) = J(u_b|u_a, v)$ for all possible u_a, u_b and v, then the correction factor in the Metropolis-Hastings acceptance ratio is equal to 1.

Connection with Gibbs sampler

That the Gibbs sampler is a type of Metropolis - Hastings algorithm is almost as easy to see. In the Gibbs sampler the proposal distribution for *U* is the full conditional distribution of *U* given *V* = *v*. That is, $J_u\left(u^*\middle|u^{(s)}, v^{(s)}\right) = f\left(u^*\middle|v^{(s)}\right)$. Then:

$$r = \frac{f\left(u^*,v^{(s)}\right)}{f\left(u^{(s)},v^{(s)}\right)} \times \frac{J_u\left(u^{(s)}\middle|u^*, v^{(s)}\right)}{J_u\left(u^*\middle|u^{(s)}, v^{(s)}\right)} = \frac{f\left(u^*,v^{(s)}\right)}{f\left(u^{(s)},v^{(s)}\right)} \times \frac{f\left(u^{(s)}\middle|v^{(s)}\right)}{f\left(u^*\middle|v^{(s)}\right)}$$

$$= \frac{f\left(u^*|v^{(s)}\right)f\left(v^{(s)}\right)}{f(u^{(s)},|v^{(s)})f(v^{(s)})} \times \frac{f\left(u^{(s)}\middle|v^{(s)}\right)}{f\left(u^*\middle|v^{(s)}\right)} = \frac{f\left(v^{(s)}\right)}{f(v^{(s)})} = 1$$

So, if we propose a value from the full conditional distribution the accep-tance probability is 1, and the algorithm is equivalent to the Gibbs sampler.

More general Metropolis-Hastings

A more general form of the Metropolis-Hastings algorithm is as follows: Given a current value $x^{(s)}$ of X,

1. Generate x* from $J_s\left(x^*|x^{(s)}\right)$
2. Compute the acceptance ratio
$$r = \frac{f(x^*)}{f\left(x^{(s)}\right)} \times \frac{J_s\left(x^{(s)}|x^*\right)}{J_s\left(x^*|x^{(s)}\right)}$$
3. Sample u ~ U(0, 1). If u < r set $x^{(s+1)}$ =x*, else set $x^{(s+1)}$ = $x^{(s)}$

Note that the proposal distribution may also depend on the iteration number s. The primary restriction we place on $J_s\left(x^*|x^{(s)}\right)$ is that it does not depend on values in the sequence previous to $x^{(s)}$.

Choice of the proposal distribution

We want to choose J_s so that the Markov chain can converge to the target distribution $f(x)$. For example, we want to make sure that every value of x such that $f(x) > 0$ will eventually be proposed (and so accepted some fraction of the time), regardless of where we start the Markov chain. For example, if the target distribution is normal, but the proposal exponential, the chain will never explore the negative half of the real line. In this case, the Metropolis-Hastings produces a Markov chain, but the chain will generate only positive numbers. This type of Markov chain is called *reducible*.

Definition: Given a measure φ, the Markov chain $\{X_n\}$ with transition kernel $K(x, y) = f(X^{(s+1)} = y \mid X^{(s)} = x)$ is φ-irreducible if, for every $A \in \mathcal{B}(\mathcal{X})$ with $\varphi(\mathrm{A}) > 0$, there exists n such that $K^n(x, A) > 0$.

Ergodic Theorem

For a sequence $X_1, X_2, \ldots$ define the partial sums

$$S_n(h) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

If $\{X_n\}$ is a Markov chain with a σ-finite invariant measure π, then the following two statements are equivalent:

(a) If $f, g \in L^1(\pi)$ with $\int g(x)d\pi(x) \neq 0$, then

$$\lim_{n\to\infty} \frac{S_n(f)}{S_n(g)} = \frac{\int f(x)d\pi(x)}{\int g(x)d\pi(x)}$$

(b) The Markov chain $\{X_n\}$ is Harris-recurrent.

Note: the ergodic theorem is most used from (b) to (a) with the special choice $g(x) = 1$ and π is called the stationary distribution.

The fine print

- A σ-finite measure π is invariant for the transition kernel $K(\cdot, \cdot)$ if

$$\pi(B) = \int K(x, B)\pi(dx)\,, \forall\, B \in \mathcal{B}(\mathcal{X})$$

- For a set A and a sequence $\{X_n\}$ define the number of passages of $\{X_n\}$ in A as

$$\eta_A = \sum_{n=1}^{\infty} I_A(X_n)$$

- A set A is Harris recurrent if $P_x(\eta_A = \infty) = 1$ for all $x \in A$. The chain $\{X_n\}$ is Harris recurrent if there exists a measure ψ such that $\{X_n\}$ is ψ-irreducible and for every set A with $\psi(A) > 0$, the set A is Harris recurrent.

Back to Metropolis-Hastings

- In most problems it is not too hard to construct Metropolis-Hastings algorithms that generate Markov chains that satisfy the ergodic theorem.
- For example, if $f(x)$ is continuous, then using a normal proposal centered around the current value guarantees that $K(A|x) > 0$ for every x and A with $f(A) > 0$.
- How do we prove that the Metropolis-Hastings chain converges to $f(x)$?
- All that needs to be checked is that the target density $f(x)$ is the invariant measure of the chain. Then the ergodic theorem guarantees the convergence.

The fine print, part 2

- A Markov chain $\{X_n\}$ is recurrent if

(i) There exists a measure ψ such that $\{X_n\}$ is ψ-irreducible

(ii) For every set $A \in \mathcal{B}(\mathcal{X})$ with $\psi(A) > 0$, we have $\mathrm{E}_x(\eta_A) = \infty$ for every $x \in A$.

- Every ψ-irreducible chain is either recurrent or transient.
- When there exists an invariant probability measure for a ψ-irreducible chain, the chain is called positive. Recurrent chains that do not allow for a finite invariant measure are called null recurrent.
- If $\{X_n\}$ is a recurrent chain, there exists an invariant σ-finite invariant measure which is unique up to a multiplicative constant. Therefore, the chain must converge to its unique invariant distribution if it satisfies the ergodic theorem.

Sketch of proof Metropolis-Hastings works

- Detailed balance condition: A Markov chain with transition kernel K satisfies the detailed balance condition if there exists a function f satisfying

$$K(y,x)f(y) = K(x,y)f(x)$$

for every x, y. While this condition is not necessary for f to be a stationary measure, it provides a sufficient condition.

- The transition kernel of the Metropolis-Hastings is

$$K(x,y) = \min\{r,1\}J(y|x) + (1 - r(x))\delta_x(y)$$

where $r(x) = \int \min\{r,1\}J(y|x)dy$ and $\delta_x(y)$ is Dirac mass at x.

- Convince yourself that the above kernel satisfies the detailed balance condition.

Extensions: Combining Metropolis and Gibbs

- In complex models it is often the case that conditional distributions are available for some parameters but not for others.
- In these situations, we can combine Gibbs and Metropolis-type proposal distributions to generate a Markov chain to approximate the joint posterior distribution of all the parameters.
- For the parameters for which we know the full conditional distribution we use Gibbs steps, and for the parameters we do not know their conditional distributions we use Metropolis steps with a carefully chosen proposal distribution.

Example: regression with correlated errors

The ordinary regression model is:

$$\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2 \sim N_n(\boldsymbol{X\beta},\sigma^2\boldsymbol{I})$$

What if the data suggests that the errors are autocorrelated? This means we must replace the covariance matrix $\boldsymbol{\sigma^2 I}$ in the ordinary regression model with a matrix **Σ** that can represent correlation between sequential observations. One simple, popular class of covariance matrices for temporally correlated data are those having first-order autoregressive structure:

$$\Sigma = \sigma^2 C_p = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix}$$

Bayesian analysis

Having observed ***y*** and ***X*** the parameters to estimate in this model include **β**, σ^2 and ρ. Use the same priors as we did in the ordinary regression. Then it can be shown that

$$\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2,\rho \sim N(\boldsymbol{\beta}_n,\boldsymbol{\Sigma}_n)$$

where

$$\boldsymbol{\Sigma}_n = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{\boldsymbol{X}'\boldsymbol{C}_{\boldsymbol{p}}^{-\mathbf{1}}\boldsymbol{X}}{\sigma^2}\right)^{-1}$$

$$\boldsymbol{\beta}_n = \boldsymbol{\Sigma}_n\left(\boldsymbol{\Sigma}_0^{-1}\beta_0 + \frac{X'\boldsymbol{C}_{\boldsymbol{p}}^{-\mathbf{1}}y}{\sigma^2}\right)$$

and

$$\sigma^2|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\rho \sim \mathrm{IG}\left(\frac{\nu_0+n}{2},\frac{\nu_0\sigma_0^2+SSE_\rho(\boldsymbol{\beta})}{2}\right)$$

where $SSE_\rho(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X\beta})'\,\boldsymbol{C}_{\boldsymbol{p}}^{-\mathbf{1}}(\boldsymbol{y}-\boldsymbol{X\beta})$

Posterior of ρ

Unfortunately, the full conditional distribution for ρ will be nonstandard for most prior distributions, suggesting that the Gibbs sampler is not applicable here and we may have to use a Metropolis algorithm. Recall that in this algorithm we are allowed to use different proposal distributions at each step. We can iteratively update β, σ^2 and ρ at different steps, making proposals with full conditional distributions for β and σ^2 (Gibbs proposals) and a symmetric proposal distribution for ρ (a Metropolis proposal). A Metropolis-Hastings algorithm to generate a new set of parameter value for ρ is as follows:

a) Propose $\rho^* \sim U(\rho^{(s)}-\delta, \rho^{(s)}+\delta)$

b) Compute the acceptance ratio

$$r = \frac{f(y|X,\beta,\sigma^2,\rho^*)\pi(\rho^*)}{f\left(y\middle|X,\beta,\sigma^2,\rho^{(s)}\right)\pi(\rho^{(s)})}$$

c) set $\rho^{(s+1)} = \rho^*$ with prob. r, otherwise set $\rho^{(s+1)} = \rho^{(s)}$.

Further extensions: Variable Dimension Models

- In general, a variable dimension model is, to quote Peter Green, a *"model where one of the things you do not know is the number of things you do not know"*.
- This setting is closely associated with *model selection,* a collection of statistical procedures that are used at the early state of a statistical analysis, namely, when the model to be used is not yet fully determined.
- Another example is mixture models in which we don’t know the number of components.
- A Bayesian variable dimension model is defined as a collection of models where

$$M_k = \{f(\cdot\,|\theta_k) : \theta_k \in \Theta_k\}, k = 1, \dots, K$$

each with prior $\pi_k(\theta_k)$ and prior on the indices k.

Reversible jump algorithms

Note that, at this stage, regular Gibbs sampling is impossible when considering distributions of the form above: if one conditions on k, then $\theta_k \in \Theta_k$, and if one conditions on θ_k, then k cannot move. Therefore, a standard Gibbs sampler cannot provide moves between models Θ_k without further modification of the setting.

Solution: Green (1995) reversible jump algorithm.