Biostat 276: Summary of April lectures

1 Markov Chains on Continuous State Space

A *Markov Chain* is a sequence of random variables $X_0, X_1, X_2, \ldots$ taking values in a space $\mathcal{X}$. The main property of the chain is that the *past is conditionally independent of the future given the present.* If $\mathcal{X}$ is a discrete space, the sequence of values is governed by a probability transition matrix $P_{x,y} = P(X_k = y | X_{k-1} = x)$. More generally, we are interested in continuous state spaces in which case the chain is governed by a transition kernel K which has two main properties: (i) $K(x, \cdot)$ is a probability measure for every value of $x \in \mathcal{X}$ and (ii) $K(\cdot, A)$ is measurable for every set $A \subset \mathcal{X}$. The meaning of these two properties is simple: (i) says given that the chain is currently at x, K defines a probability density to say where the chain will move at the next step; (ii) says that we can always evaluate the probability that the chain will jump into some set A from all possible values x.

If the kernel K is well-behaved, then the Markov chain will have a *stationary distribution* π. This means that if we simulate X_0 from π and then run one step of the chain starting at X_0, then the resulting value X_1 will be distributed according to π. Notationally, we write $\pi(y) = \int_{\mathcal{X}} K(x, y)\pi(y)dy, \forall y \in \mathcal{X}$. A stronger requirement is that the chain be *reversible* with respect to π, that is

$$\pi(x)K(x, y) = \pi(y)K(y, x).$$

The main use of Markov chains in simulation settings is that (for well-behaved kernels K), they have a unique stationary distribution that coincides with the *limiting distribution* of the chain. Computationally, this means that if we run a Markov chain long enough that it produces simulations from a distribution that is *independent* of the initial value of the chain. In a Bayesian setting, if we can define Markov chains that have stationary distribution $p(\theta|y)$ then we can get approximate simulations from this posterior distribution by running the chain for a long time.

The main requirement for the chain to reach its stationary distribution in the limit is that it is *irreducible* and *aperiodic*. Irreducibility is defined as:

$$\forall x, y \in \mathcal{X}, \exists n < \infty, \text{ such that } K^n(x, y) > 0.$$

In other words, the chain can jump from anywhere to anywhere in a finite number of steps. If it is possible to jump from anywhere to anywhere in one step, the chain is said to be *strongly irreducible*, and these sorts of chains tend to have the fastest convergence properties. Aperiodicity means that there exist no subsets of the state space that can only be visited periodically.

2 Gibbs sampler and Metropolis-Hastings algorithm

We have defined two Markov chains that have stationary distribution equal to the posterior distribution $p(\theta|y)$. The first of these, the *Gibbs sampler*, is useful when $\theta = (\phi, \psi)$ (and extends in same way to three or more components).

1. Initialize $\phi^{(0)}$.

2. For `i in 1:M`:

   - Simulate $\psi^{(i)}$ from $p(\psi|\phi^{(i-1)}, y)$.
   - Simulate $\phi^{(i)}$ from $p(\phi|\psi^{(i)}, y)$.

The reason this is such a useful algorithm is that the *full conditional distributions* $p(\phi|\psi, y)$ and $p(\psi|\phi, y)$ are often *available* (that is easy to simulate from) even though the joint posterior distribution is complicated. This becomes especially true when θ has many components.

The second of these algorithms, the *Metropolis-Hastings algorithm* is quite similar to the Accept-Reject algorithm:

1. Initialize $\theta^{(0)}$.

2. For `i in 1:M`:

   - Propose a candidate θ^* using a proposal density $q(\theta^*|\theta^{(i-1)})$.
   - Set

   $$\theta^{(i)} = \begin{cases} \theta^* & \text{w.p. } \alpha \\ \theta^{(i-1)} & \text{w.p. } 1-\alpha \end{cases}$$

   where

   $$\alpha = \min\left(\frac{p(\theta^*|y)}{p(\theta^{(i-1)}|y)} \frac{q(\theta^{(i-1)}|\theta^*)}{q(\theta^*|\theta^{(i-1)})}, 1\right).$$

3 MCMC examples

3.1 Poisson changepoint model

Consider the model

$$Y_i \sim \begin{cases} \mathrm{P}(\lambda), & i = 1, \ldots, m \\ \mathrm{P}(\phi), & i = m+1, \ldots, n \end{cases}$$

λ, ϕ and m are unknown parameters which we assume to have independent prior distributions $\mathrm{G}(\alpha, \beta), \mathrm{G}(\gamma, \delta)$, and discrete uniform respectively. Direct computations are difficult for this model because of the unknown m, but full conditional distributions are straightforward:

$$\begin{aligned} \lambda|\phi, m, y &\sim G(\alpha + \sum_{i=1}^{m} Y_i, \beta + m) \\ \phi|\lambda, m, y &\sim G(\gamma + \sum_{i=m+1}^{n} Y_i, \delta + n - m) \\ Prob(m = k) &\propto \lambda^{\alpha + \sum_{i=1}^{m} Y_i - 1} e^{-(\beta+m)} \phi^{\gamma + \sum_{i=m+1}^{n} Y_i - 1} e^{-(\delta+n-m)} \end{aligned}$$

where the proportionality constant is resolved in the last equation by dividing over the sum over the values $k = 1, \dots, n$.

This defines a simple Gibbs sampler which I implemented in C code and discussed in lecture.

3.2 Bayesian regression

Suppose we have n observations $Y = (Y_1, \dots, Y_n)^T$ with n by p covariate matrix X. In general, these observations might have variance-covariance matrix Σ. A Bayesian model might be:

$$\begin{aligned} Y|\beta, \Sigma &\sim N_n(X\beta, \Sigma) \\ \beta &\sim N_p(b, C) \\ \Sigma^{-1} &\sim W(\nu, (\nu\Lambda)^{-1}) \end{aligned}$$

where b is the prior mean and C the prior variance-covariance matrix for the regression parameters β, and Λ is our prior guess for Σ with degrees of freedom ν. We assume these four hyper-parameters are known for now.

Conditional on Σ, Y and β are jointly normal with mean vector $\begin{pmatrix} Xb \\ b \end{pmatrix}$. and variance-covariance matrix $\begin{pmatrix} \Sigma + XCX^T & XC \\ CX^T & C \end{pmatrix}$.

Thus, β given Y (and Σ) is also normal with mean $\hat{\beta}$ and variance-covariance V_β. To find these, we can use the usual rules for conditional distribution of partitions of multivariate normals. We have that $V_\beta = C - CX^T(\Sigma + XCX^T)^{-1}XC$. and $\hat{\beta} = b + CX^T(\Sigma + XCX^T)^{-1}(Y - Xb)$. These can be rewritten in more familiar form using matrix identities such as

$$(A + BCB^T)^{-1} = A^{-1} - A^{-1}B(C^{-1} + B^TA^{-1}B)^{-1}B^TA^{-1}$$

for conformable A, B, and C. Doing so we find that $V_\beta = (X^T\Sigma^{-1}X + C^{-1})^{-1}$ and $\hat{\beta} = V_\beta(X^T\Sigma^{-1}Y + C^{-1}b)$. These quantities can be interpreted as precision weighted averages of prior and data estimates.

By writing out the likelihood times the prior we find that the full conditional distribution of Σ^{-1} is also Wishart with updated parameters $\nu + 1$ and $(\nu\Lambda + (Y - X\beta)(Y - X\beta)^T)^{-1}$. This suggests a Gibbs sampler alternating between

$$\begin{aligned} \beta|\Sigma, Y &\sim N(\hat{\beta}, V_\beta) \\ \Sigma^{-1}|\beta, Y &\sim W(\nu + 1, (\nu\Lambda + (Y - X\beta)(Y - X\beta)^T)^{-1}). \end{aligned}$$

To make this more useful in practice, we extend in two ways. First, we consider $Y = (Y_1, \dots, Y_n)$ where each Y_i is a k-vector of measurements (e.g. longitudinal) on the ith subject. Then we might have the model:

$$Y_i|\beta, \Sigma \sim N_k(X_i\beta, \Sigma)$$

$$
\begin{aligned}
\beta &\sim N_p(b, C) \\
\Sigma^{-1} &\sim W(\nu, (\nu\Lambda)^{-1})
\end{aligned}
$$

and the Gibbs sampler works like

$$
\begin{aligned}
\beta|\Sigma, Y &\sim N(\hat{\beta}, V_\beta) \\
\Sigma^{-1}|\beta, Y &\sim W(\nu + n, (\nu\Lambda + \sum_{i=1}^{n}(Y_i - X_i\beta)(Y_i - X_i\beta)^T)^{-1}).
\end{aligned}
$$

where now $V_\beta = \left(\sum_{i=1}^{n} X_i^T \Sigma^{-1} X_i + C^{-1}\right)^{-1}$ and $\hat{\beta} = V_\beta \left(\sum_{i=1}^{n}(X_i^T \Sigma^{-1} Y_i) + C^{-1} b\right)$.

3.3 Scale mixtures of normals

Robust extensions of the normal distribution can be defined through scale mixtures of normals. The basic setup is:

$$
\begin{aligned}
Y_i|\mu_i, \sigma^2, \lambda_i &\sim N(\mu_i, \sigma^2/\lambda_i) \\
\lambda_i &\sim g(\alpha)
\end{aligned}
$$

where $g(\alpha)$ is some distribution to mix over. For instance, if we choose $\lambda_i \sim G(\nu/2, \nu/2)$ then $Y_i|\mu_i, \sigma^2$ has a student-t distribution with ν degrees of freedom. Choosing $\lambda_i \sim Exp(2)$ gives a double exponential marginal for Y_i. Other choices are discussed in the references listed on the webpage.