BTRY 4830/6830: Quantitative Genomics and Genetics

Lecture 24: Bayesian inference and inference using MCMC

Jason Mezey
jgm45@cornell.edu
Nov. 18, 2014 (T) 8:40-9:55
Announcements

• Last computer lab this week (!!)

• We will not have any lectures after Thanksgiving (!!!) = this means we have three lectures left: Today, Thurs. and this coming Tues.

• Last office hours are this week (you can meet with us by appointment as always)

• The Final (!!)
  • Take home / Open book (same format as midterm)
  • Available during the day Weds., Dec. 10 and will be due 11:59PM Sat., Dec. 13 (late rules apply)
  • Cumulative (!!) but emphasis on GWAS basics
Summary of lecture 24

- Last lecture, we began our introduction to Bayesian inference
- Today, we will continue our discussion and discuss how to perform Bayesian inference for more complex models using an MCMC algorithm
Review: Intro to Bayesian analysis I

- Remember that in a Bayesian (not frequentist!) framework, our parameter(s) have a probability distribution associated with them that reflects our belief in the values that might be the true value of the parameter.

- Since we are treating the parameter as a random variable, we can consider the joint distribution of the parameter AND a sample \( \mathbf{Y} \) produced under a probability model:

\[
Pr(\theta \cap \mathbf{Y})
\]

- For inference, we are interested in the probability the parameter takes a certain value given a sample:

\[
Pr(\theta | \mathbf{y})
\]

- Using Bayes theorem, we can write:

\[
Pr(\theta | \mathbf{y}) = \frac{Pr(\mathbf{y} | \theta)Pr(\theta)}{Pr(\mathbf{y})}
\]

- Also note that since the sample is fixed (i.e. we are considering a single sample) \( Pr(\mathbf{y}) = c \), we can rewrite this as follows:

\[
Pr(\theta | \mathbf{y}) \propto Pr(\mathbf{y} | \theta)Pr(\theta)
\]
Review: Intro to Bayesian analysis II

- Let’s consider the structure of our main equation in Bayesian statistics:

\[ Pr(\theta|y) \propto Pr(y|\theta)Pr(\theta) \]

- Note that the left hand side is called the posterior probability:

\[ Pr(\theta|y) \]

- The first term of the right hand side is something we have seen before, i.e. the likelihood (!!):

\[ Pr(y|\theta) = L(\theta|y) \]

- The second term of the right hand side is new and is called the prior:

\[ Pr(\theta) \]

- Note that the prior is how we incorporate our assumptions concerning the values the true parameter value may take.

- In a Bayesian framework, we are making two assumptions (unlike a frequentist where we make one assumption: 1. the probability distribution that generated the sample, 2. the probability distribution of the parameter
Review: Intro to Bayesian analysis III

• Up to this point, we have discussed priors in an abstract manner.

• To start making this concept more clear, let’s consider one of our original examples where we are interested in the knowing the mean human height in the US (what are the components of the statistical framework for this example!? Note the basic components are the same in Frequentist / Bayesian!)

• If we assume a normal probability model of human height (what parameter are we interested in inferring in this case and why?) in a Bayesian framework, we will at least need to define a prior:

\[ Pr(\mu) \]

• One possible approach is to make the probability of each possible value of the parameter the same (what distribution are we assuming and what is a problem with this approach), which defines an improper prior:

\[ Pr(\mu) = c \]

• Another possible approach is to incorporate our previous observations that heights are seldom infinite, etc. where one choice for incorporating this observations is my defining a prior that has the same distribution as our probability model, which defines a conjugate prior (which is also a proper prior):

\[ Pr(\mu) \sim N(\kappa, \phi^2) \]
Constructing the posterior probability

- Let’s put this all together for our “heights in the US” example.
- First recall that our assumption is the probability model is normal (so what is the form of the likelihood?):
  \[ Y \sim N(\mu, \sigma^2) \]
- Second, assume a normal prior for the parameter we are interested in:
  \[ Pr(\mu) \sim N(\kappa, \phi^2) \]
- From the Bayesian equation, we can now put this together as follows:
  \[ Pr(\theta|y) \propto Pr(y|\theta)Pr(\theta) \]
  \[ Pr(\mu|y) \propto \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \right) \frac{1}{\sqrt{2\pi\phi^2}} e^{-\frac{(\mu-\kappa)^2}{2\phi^2}} \]
- Note that with a little rearrangement, this can be written in the following form:
  \[ Pr(\mu|y) \sim N\left( \frac{\frac{\kappa}{\sigma^2} + \frac{\sum_{i}^{n} y_i}{\sigma^2}}{\frac{1}{\phi^2} + \frac{n}{\sigma^2}}, \frac{1}{\phi^2 + \frac{n}{\sigma^2}} \right) \]
Bayesian inference: estimation

- Inference in a Bayesian framework differs from a frequentist framework in both estimation and hypothesis testing.

- For example, for estimation in a Bayesian framework, we always construct estimators using the posterior probability distribution, for example:

\[
\hat{\theta} = mean(\theta|y) = \int \theta Pr(\theta|y) d\theta \quad \text{or} \quad \hat{\theta} = median(\theta|y)
\]

- For example, in our “heights in the US” example our estimator is:

\[
\hat{\mu} = median(\mu|y) = mean(\mu|y) = \frac{\left( \frac{k}{\sigma^2} + \frac{n\bar{y}}{\sigma^2} \right)}{\left( \frac{1}{\phi^2} + \frac{n}{\sigma^2} \right)}
\]

- Note 1: again notice that the impact of the prior disappears as the sample size goes to infinite (=same as MLE under this condition):

\[
\frac{\left( \frac{k}{\sigma^2} + \frac{n\bar{y}}{\sigma^2} \right)}{\left( \frac{1}{\phi^2} + \frac{n}{\sigma^2} \right)} \approx \frac{\left( \frac{n\bar{y}}{\sigma^2} \right)}{\left( \frac{n}{\sigma^2} \right)} \approx \bar{y}
\]

- Note 2: estimates in a Bayesian framework can be different than in a likelihood (Frequentist) framework since estimator construction is fundamentally different (!!!)
Bayesian inference: hypothesis testing

- For hypothesis testing in a Bayesian analysis, we use the same null and alternative hypothesis framework:

\[ H_0 : \theta \in \Theta_0 \]

\[ H_A : \theta \in \Theta_A \]

- However, the approach to hypothesis testing is completely different than in a frequentist framework, where we use a Bayes factor to indicate the relative support for one hypothesis versus the other:

\[
Bayes = \frac{\int_{\theta \in \Theta_0} Pr(y|\theta)Pr(\theta) d\theta}{\int_{\theta \in \Theta_A} Pr(y|\theta)Pr(\theta) d\theta}
\]

- Note that a downside to using a Bayes factor to assess hypotheses is that it can be difficult to assign priors for hypotheses that have completely different ranges of support (e.g. the null is a point and alternative is a range of values).

- As a consequence, people often use an alternative “psuedo-Bayesian” approach to hypothesis testing that makes use of credible intervals (which is what we will use in this course).
Bayesian credible intervals (versus frequentist confidence intervals)

- Recall that in a Frequentist framework that we can estimate a confidence interval at some level (say 0.95), which is an interval that will include the value of the parameter 0.95 of the times we performed the experiment an infinite number of times, calculating the confidence interval each time (note: a strange definition...)

- In a Bayesian interval, the parallel concept is a credible interval that has a completely different interpretation: this interval has a given probability of including the parameter value (!!!)

- The definition of a credible interval is as follows:

  \[
  c.i.(\theta) = \int_{-c_\alpha}^{c_\alpha} Pr(\theta|y)d\theta = 1 - \alpha
  \]

- Note that we can assess a null hypothesis using a credible interval by determining if this interval includes the value of the parameter under the null hypothesis (!!)
Bayesian inference: genetic model I

- We are now ready to tackle Bayesian inference for our genetic model (note that we will focus on the linear regression model but we can perform Bayesian inference for any GLM!):

\[
Y = \beta_{\mu} + X_a \beta_a + X_d \beta_d + \epsilon \\
\epsilon \sim N(0, \sigma_{\epsilon}^2)
\]

- Recall for a sample generated under this model, we can write:

\[
y = x\beta + \epsilon \\
\epsilon \sim multiN(0, I\sigma_{\epsilon}^2)
\]

- In this case, we are interested in the following hypotheses:

\[
H_0 : \beta_a = 0 \cap \beta_d = 0 \\
H_A : \beta_a \not= 0 \cup \beta_d \not= 0
\]

- We are therefore interested in the marginal posterior probability of these two parameters
Bayesian inference: genetic model II

- To calculate these probabilities, we need to assign a joint probability distribution for the prior

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) \]

- One possible choice is as follows (are these proper or improper!?):

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) = Pr(\beta_\mu) Pr(\beta_a) Pr(\beta_d) Pr(\sigma^2_\epsilon) \]

\[ Pr(\beta_\mu) = Pr(\beta_a) = Pr(\beta_d) = c \]

\[ Pr(\sigma^2_\epsilon) = c \]

- Under this prior the complete posterior distribution is multivariate normal (!!):

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon | y) \propto Pr(y | \beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) \]

\[ Pr(\theta | y) \propto (\sigma^2_\epsilon)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2_\epsilon} (y - x\beta)^T (y - x\beta)} \]
Bayesian inference: genetic model III

- For the linear model with sample:

\[ y = x\beta + \epsilon \]
\[ \epsilon \sim \text{multiN}(0, I\sigma^2_\epsilon) \]

- The complete posterior probability for the genetic model is:

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon | y) \propto Pr(y | \beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) \]

- With a uniform prior is:

\[ Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon | y) \propto Pr(y | \beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon) \]

- The marginal posterior probability of the parameters we are interested in is:

\[ Pr(\beta_a, \beta_d | y) = \int_0^\infty \int_{-\infty}^\infty Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon | y) d\beta_\mu d\sigma^2_\epsilon \]
Bayesian inference: genetic model IV

- Assuming uniform (improper!) priors, the marginal distribution is:

\[
Pr(\hat{\beta}_a, \hat{\beta}_d|y) = \int_{-\infty}^{\infty} \int_{0}^{\infty} Pr(\beta_\mu, \beta_a, \beta_d, \sigma^2_\epsilon|y) d\beta_\mu d\sigma^2_\epsilon \sim \text{multi-t-distribution}
\]

- With the following parameter values:

\[
mean(Pr(\beta_a, \beta_d|y)) = \begin{bmatrix} \hat{\beta}_a, \hat{\beta}_d \end{bmatrix}^T = C^{-1} [X_a, X_d]^T y \\
C = \begin{bmatrix} X^T_a X_a & X^T_a X_d \\ X^T_d X_a & X^T_d X_d \end{bmatrix}
\]

\[
cov = \frac{(y - [X_a, X_d] \begin{bmatrix} \hat{\beta}_a, \hat{\beta}_d \end{bmatrix}^T)(y - [X_a, X_d] \begin{bmatrix} \hat{\beta}_a, \hat{\beta}_d \end{bmatrix}^T)}{n - 6}
\]

\[
df(\text{multi-t}) = n - 4
\]

- With these estimates (equations) we can now construct a credible interval for our genetic null hypothesis and test a marker for a phenotype association and we can perform a GWAS by doing this for each marker (!!!)
Bayesian inference: genetic model V

\[
Pr(\beta_a, \beta_d | y) \quad \text{and} \quad Pr(\beta_a, \beta_d | y)
\]

Cannot reject H0!

Reject H0!
Bayesian inference for more “complex” posterior distributions

• For a linear regression, with a simple (uniform) prior, we have a simple closed form of the overall posterior

• This is not always (=often not the case), since we may often choose to put together more complex priors with our likelihood or consider a more complicated likelihood equation (e.g. for a logistic regression!)

• To perform hypothesis testing with these more complex cases, we still need to determine the credible interval from the posterior (or marginal) probability distribution so we need to determine the form of this distribution

• To do this we will need an algorithm and we will introduce the Markov chain Monte Carlo (MCMC) algorithm for this purpose
Stochastic and Markov processes

• To introduce the MCMC algorithm for our purpose, we need to consider models from another branch of probability (remember, probability is a field much larger than the components that we use for statistics / inference!): Stochastic processes

• **Stochastic process** - a collection of random vectors (variables) with defined conditional relationships, often indexed by a ordered set $t$

• We will be interested in one particular class of models within this probability sub-field: Markov processes (or more specifically *Markov chains*)

• Our MCMC will be a Markov chain (probability model)
Stochastic and Markov processes II

- A Markov chain can be thought of as a random vector (or more accurately, a set of random vectors), which we will index with $t$

$$X_t, X_{t+1}, X_{t+2}, \ldots, X_{t+k}$$

$$X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-k}$$

- **Markov chain** - a stochastic process that satisfies the Markov property:

$$Pr(X_t | X_{t-1}, X_{t-2}, \ldots, X_{t-k}) = Pr(X_t | X_{t-1})$$

- While we often assume each of the random variables in a Markov chain are in the same class of random variables (e.g. Bernoulli, normal, etc.) we allow the parameters of these random variables to be different, e.g. at time $t$ and $t+1$

- How does this differ from a random vector of an iid sample??
Example of a Markov chain

• As an example, let’s consider a Markov chain where each random variable in the chain has a Bernoulli distribution:

\[ X_1, X_2, X_3, X_4, X_5 \]

\[ X_1 \sim Bern(0.2), X_2 \sim Bern(0.1), X_3 \sim Bern(0.4), X_4 \sim Bern(0.4), X_5 \sim Bern(0.4) \]

• Note that we could draw observations from this Markov chain (since it is just a random vector with a probability distribution!):

\[ 1,0,0,1,1 \]
\[ 0,0,0,0,0 \]
\[ 0,1,0,1,1 \]
\[ 0,1,0,0,0 \]

• How does this differ from an iid random vector?

• Note that for \( t=3-5 \) in this process, the parameters of the Bernoulli distributions are the same (=they do not change over time)

• In our case, we will be interested in Markov chains that “evolve” to such stationary distributions
Stationary distributions and MCMC

• If a Markov chain has certain properties (irreducible and ergodic), we can prove that the chain will evolve (more accurately converge!) to a unique (!!) stationary distribution and will not leave this stationary distribution (where is it often possible to determine the parameters for the stationary distribution!)

• For such Markov chains, if we consider enough iterations $t+k$ (where $k$ may be very large, e.g. infinite), we will reach a point where each following random variable is in the unique stationary distribution:

$$Pr(X_{t+k}) = Pr(X_{t+k+1}) = \ldots$$

• For the purposes of Bayesian inference, we are going to set up a Markov chain that evolves to a unique stationary distribution that is exactly the posterior probability distribution that we are interested in (!!!)

• To use this chain, we will run the Markov chain for enough iterations to reach this stationary distribution and then we will take a sample from this chain to determine (or more accurately approximate) our posterior

• This is Bayesian Markov chain Monte Carlo (MCMC)!
An example of Bayesian MCMC

\[ MCMC = X_{t+k}, X_{t+k+1}, X_{t+k+2}, \ldots, X_{t+k+m} \]

Sample = 0.1, -0.08, -1.4, ..., 0.5

\[ \hat{\theta} = median(Pr(\theta|y)) \approx median(\theta^{[t_ab]}, ..., \theta^{[t_ab+k]}) \]
That’s it for today

- Next lecture, we will provide a brief introduction to pedigree (linkage) analysis and inbred line analysis
- For our final lecture, we will discuss “additive genetic variance” and provide a brief introduction to evolutionary quantitative genetics