Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.5201.03

Lecture 26: Bayesian analysis II (MCMC)

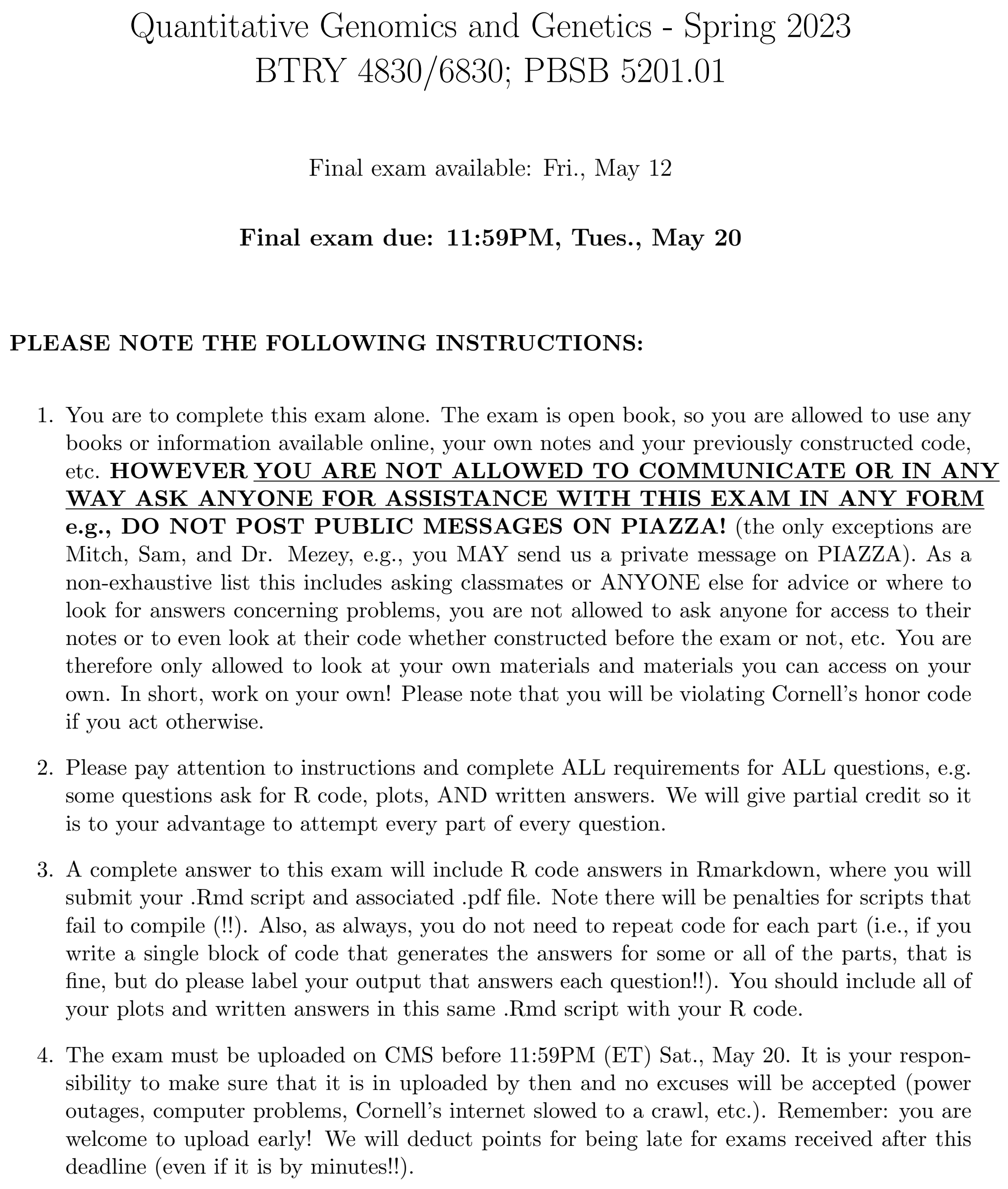
Jason Mezey

May 4, 2023 (Th) 8:05-9:20

Announcements

- PLEASE NOTE (!!): Tues, May 9 (last lecture) I will lecture from Ithaca (regular classroom) with no classroom in NYC (please join by zoom)!
- Additional (optional!) lecture decks and videos are available at the end of the class site: (#1: haplotypes and haplotype testing; #2 epistasis and eQTL; #3 alternative tests in GWAS - including permutation!)
- We only have 1 computer labs left (!!)
  - Thurs / Fri (May 4 / 5) MCMC algorithm for Bayesian inference
  - NO COMPUTER LAB next week of class Thurs / Fri (May 11 / 12)
- Last office hours tomorrow (!!) Fri, (May 5) 12:30-2:30 - DIFFERENT ZOOM LINK = see Piazza message!
- Final Exam
  - Available Fri., May 12 and due 11:59pm Sat., May 20
  - Content: you will need to perform a GWAS using a linear regression with and without covariates AND a logistic regression with and without covariates (and we will give you the covariates already coded)!

Final - instructions



Quantitative Genomics and Genetics - Spring 2023
BTRY 4830/6830; PBSB 5201.01

Final exam available: Fri., May 12

Final exam due: 11:59PM, Tues., May 20

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

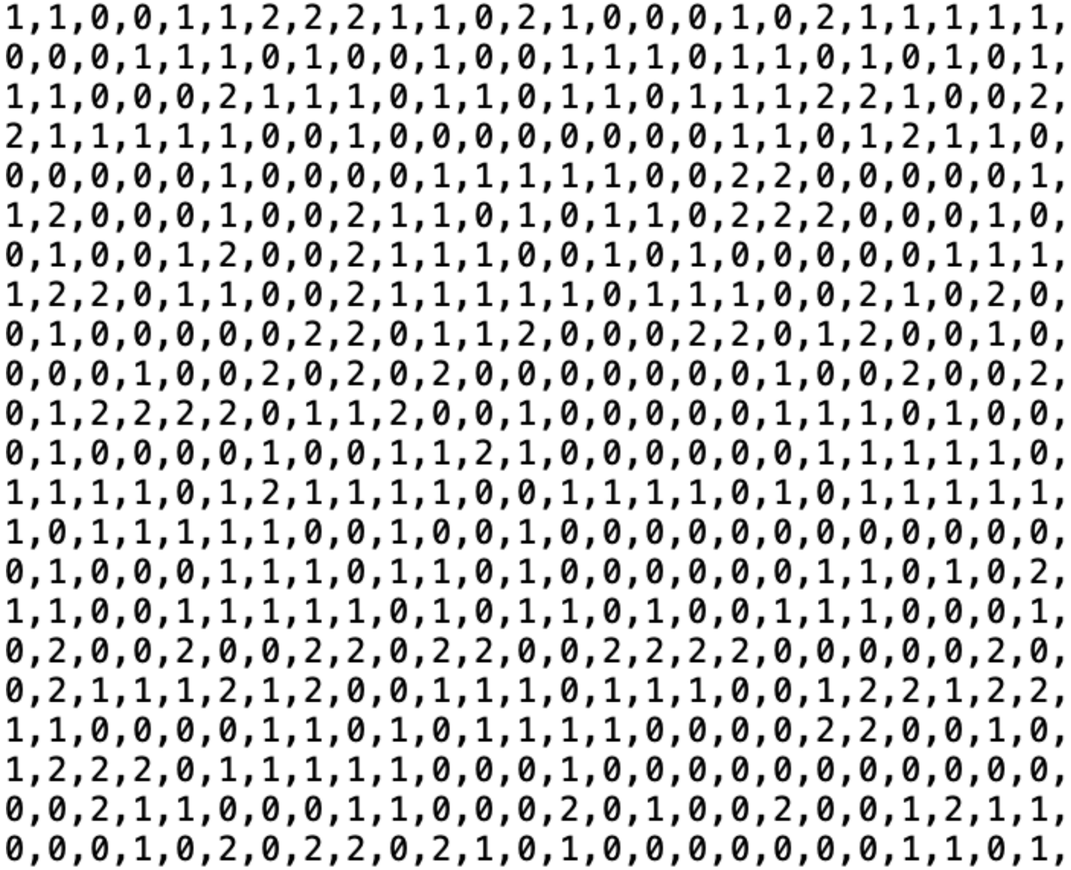
1. You are to complete this exam alone. The exam is open book, so you are allowed to use any books or information available online, your own notes and your previously constructed code, etc. **HOWEVER YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM e.g., DO NOT POST PUBLIC MESSAGES ON PIAZZA!** (the only exceptions are Mitch, Sam, and Dr. Mezey, e.g., you MAY send us a private message on PIAZZA). As a non-exhaustive list this includes asking classmates or ANYONE else for advice or where to look for answers concerning problems, you are not allowed to ask anyone for access to their notes or to even look at their code whether constructed before the exam or not, etc. You are therefore only allowed to look at your own materials and materials you can access on your own. In short, work on your own! Please note that you will be violating Cornell's honor code if you act otherwise.

2. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for R code, plots, AND written answers. We will give partial credit so it is to your advantage to attempt every part of every question.

3. A complete answer to this exam will include R code answers in Rmarkdown, where you will submit your .Rmd script and associated .pdf file. Note there will be penalties for scripts that fail to compile (!!). Also, as always, you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!). You should include all of your plots and written answers in this same .Rmd script with your R code.

4. The exam must be uploaded on CMS before 11:59PM (ET) Sat., May 20. It is your responsibility to make sure that it is in uploaded by then and no excuses will be accepted (power outages, computer problems, Cornell's internet slowed to a crawl, etc.). Remember: you are welcome to upload early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).

Final - genotype data!



```
1,1,0,0,1,1,2,2,2,1,1,0,2,1,0,0,0,1,0,2,1,1,1,1,1,
0,0,0,1,1,1,0,1,0,0,1,0,0,1,1,1,0,1,1,0,1,0,1,0,1,
1,1,0,0,0,2,1,1,1,0,1,1,0,1,1,0,1,1,1,2,2,1,0,0,2,
2,1,1,1,1,1,0,0,1,0,0,0,0,0,0,0,1,1,0,1,2,1,1,0,
0,0,0,0,0,1,0,0,0,0,1,1,1,1,1,0,0,2,2,0,0,0,0,0,1,
1,2,0,0,0,1,0,0,2,1,1,0,1,0,1,1,0,2,2,2,0,0,0,1,0,
0,1,0,0,1,2,0,0,2,1,1,1,0,0,1,0,1,0,0,0,0,0,1,1,1,
1,2,2,0,1,1,0,0,2,1,1,1,1,1,0,1,1,1,0,0,2,1,0,2,0,
0,1,0,0,0,0,0,2,2,0,1,1,2,0,0,0,2,2,0,1,2,0,0,1,0,
0,0,0,1,0,0,2,0,2,0,2,0,0,0,0,0,0,0,1,0,0,2,0,0,2,
0,1,2,2,2,2,0,1,1,2,0,0,1,0,0,0,0,0,1,1,1,0,1,0,0,
0,1,0,0,0,0,1,0,0,1,1,2,1,0,0,0,0,0,0,1,1,1,1,1,0,
1,1,1,1,0,1,2,1,1,1,1,0,0,1,1,1,1,0,1,0,1,1,1,1,1,
1,0,1,1,1,1,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,
0,1,0,0,0,1,1,1,0,1,1,0,1,0,0,0,0,0,0,1,1,0,1,0,2,
1,1,0,0,1,1,1,1,1,0,1,0,1,1,0,1,0,0,1,1,1,0,0,0,1,
0,2,0,0,2,0,0,2,2,0,2,2,0,0,2,2,2,2,0,0,0,0,0,2,0,
0,2,1,1,1,2,1,2,0,0,1,1,1,0,1,1,1,0,0,1,2,2,1,2,2,
1,1,0,0,0,0,1,1,0,1,0,1,1,1,1,0,0,0,0,2,2,0,0,1,0,
1,2,2,2,0,1,1,1,1,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,
0,0,2,1,1,0,0,0,1,1,0,0,0,2,0,1,0,0,2,0,0,1,2,1,1,
0,0,0,1,0,2,0,2,2,0,2,1,0,1,0,0,0,0,0,0,0,1,1,0,1,
```

Summary of lecture 26: Introduction to pedigree and inbred line analysis

- Today, we will complete our discuss of MCMC for Bayesian statistics
- We will also (briefly) discuss pedigree analysis and inbred line analysis (!!)

Review: Intro to Bayesian analysis

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

$$Pr(\theta|\mathbf{y}) \propto Pr(\mathbf{y}|\theta)Pr(\theta)$$

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

$$Pr(\theta|\mathbf{y})$$

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

$$Pr(\mathbf{y}|\theta) = L(\theta|\mathbf{y})$$

- 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: Bayesian inference: estimation I

- 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:

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

- Estimates in a Bayesian framework can be different than in a likelihood (Frequentist) framework since estimator construction is fundamentally different (!!)

Review: Bayesian credible 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:

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

- 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 (!!)

Review: Bayesian inference: genetic model 1

- 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:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon$$

$$\epsilon \sim multiN(0, \mathbf{I}\sigma_\epsilon^2)$$

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

$$H_0 : \beta_a = 0 \cap \beta_d = 0 \qquad H_A : \beta_a \neq 0 \cup \beta_d \neq 0$$

- We are therefore interested in the *marginal posterior probability* of these two parameters

Review: Bayesian inference: genetic model II

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

$$Pr(\beta_a, \beta_d | \mathbf{y}) = \int_{-\infty}^{\infty} \int_{0}^{\infty} Pr(\beta_\mu, \beta_a, \beta_d, \sigma_\epsilon^2 | \mathbf{y}) d\beta_\mu d\sigma_\epsilon^2 \sim multi\text{-}t\text{-}distribution$$

- With the following parameter values:

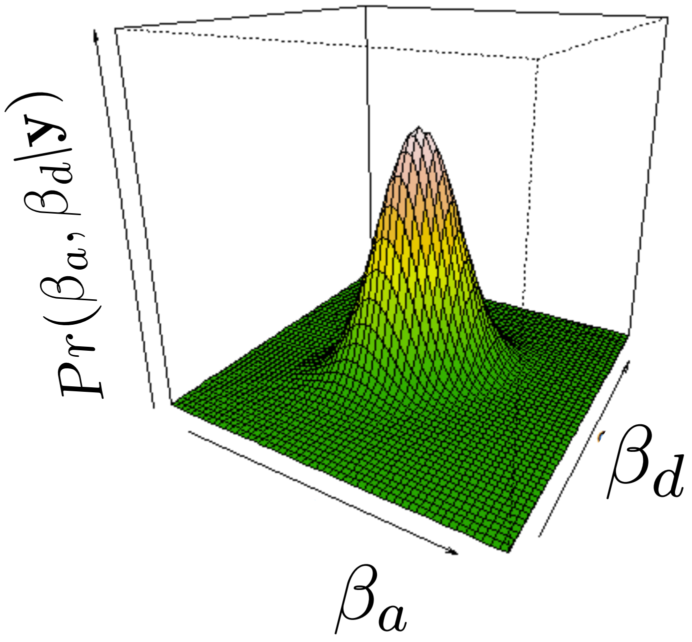
$$mean(Pr(\beta_a, \beta_d | \mathbf{y})) = \left[\hat{\beta}_a, \hat{\beta}_d\right]^{\mathrm{T}} = \mathbf{C}^{-1} \left[\mathbf{X}_a, \mathbf{X}_d\right]^{\mathrm{T}} \mathbf{y} \qquad \mathbf{C} = \begin{bmatrix} \mathbf{X}_a^{\mathrm{T}}\mathbf{X}_a & \mathbf{X}_a^{\mathrm{T}}\mathbf{X}_d \\ \mathbf{X}_d^{\mathrm{T}}\mathbf{X}_a & \mathbf{X}_d^{\mathrm{T}}\mathbf{X}_d \end{bmatrix}$$

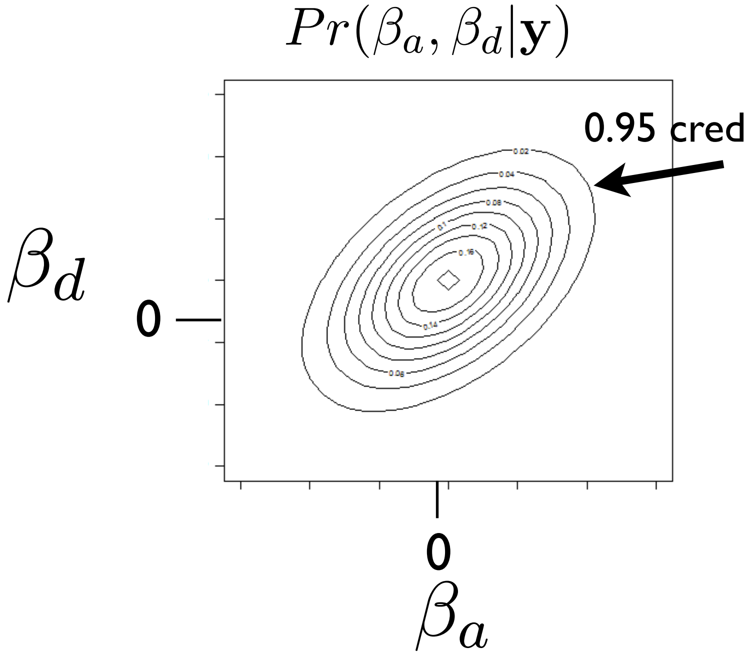
$$cov = \frac{(\mathbf{y} - [\mathbf{X}_a, \mathbf{X}_d] \left[\hat{\beta}_a, \hat{\beta}_d\right]^{\mathrm{T}})^{\mathrm{T}} (\mathbf{y} - [\mathbf{X}_a, \mathbf{X}_d] \left[\hat{\beta}_a, \hat{\beta}_d\right]^{\mathrm{T}})}{n-6} \mathbf{C}^{-1}$$

$$df(multi{-}t) = n - 4$$

- 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 (!!)

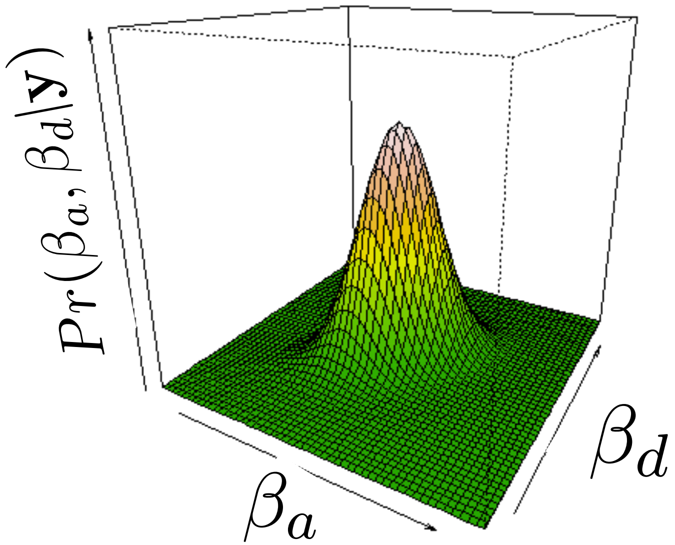
Review: Bayesian inference: genetic model III

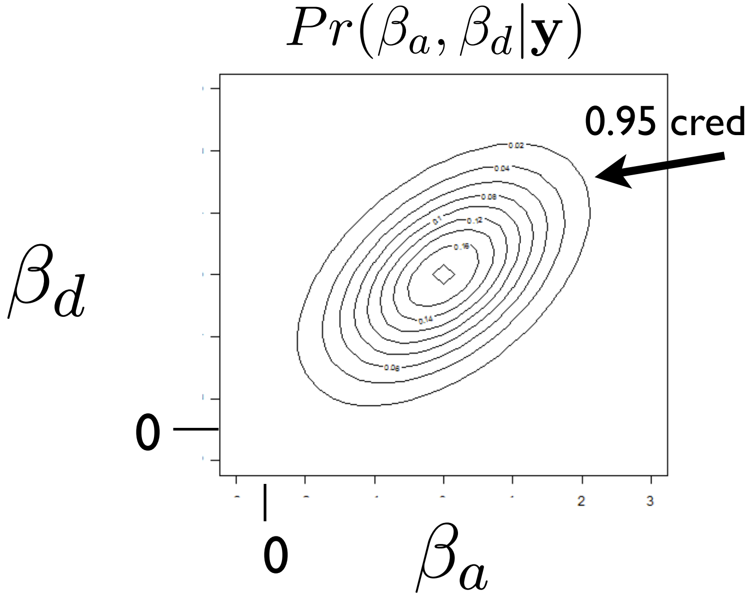




0.95 credible interval

Cannot reject H0!





0.95 credible interval

Reject H0!

Review: 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

Review: Stochastic 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** (intuitive def) - a collection of random vectors (variables) with defined conditional relationships, often indexed by an 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)

Review: Markov processes

- 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*:

$$X_t, X_{t+1}, X_{t+2}, ..., X_{t+k}$$

$$X_t, X_{t-1}, X_{t-2}, ..., X_{t-k}$$

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

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

- 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!?

Review: 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:

$$X_1, X_2 ..., X_{1001}, X_{1002}$$

$$X_1 \sim Bern(0.2), X_2 \sim Bern(0.21), ..., X_{1001} \sim Bern(0.4), X_{1002} \sim Bern(0.4)$$

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

1,0,...,1,1 0,0,...,0,0

0,1,...,1,1 0,1,...,0,0

- How does this differ from an iid random vector?

- Note that for *t* late 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:

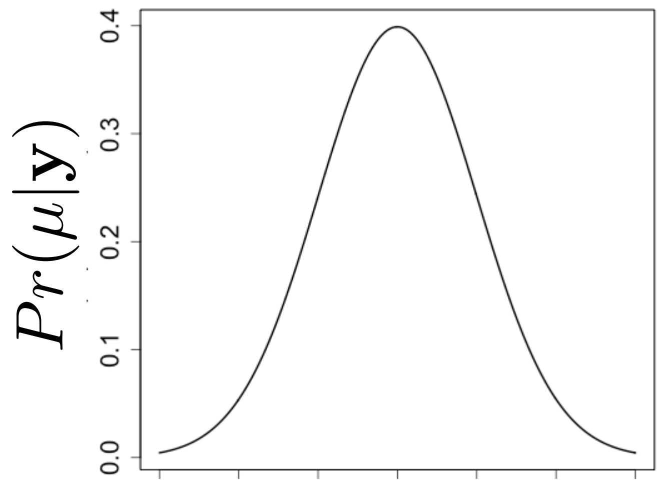
$$Pr(X_{t+k}) = Pr(X_{t+k+1}) = ...$$

- 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

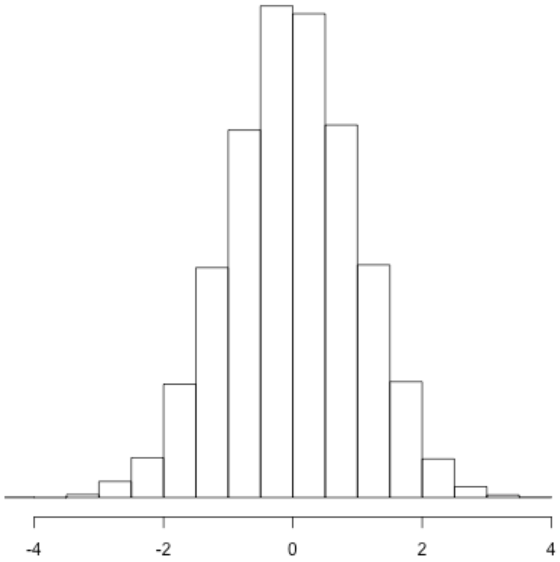
- This is Bayesian Markov chain Monte Carlo (MCMC)!

An example of Bayesian MCMC



$$MCMC = X_{t+k}, X_{t+k+1}, X_{t+k+2}, ..., X_{t+k+m}$$

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



$$\hat{\theta} = median(Pr(\theta|\mathbf{y}) \simeq median(\theta^{[t_{ab}]}, ..., \theta^{[t_{ab}+k]})$$

Constructing an MCMC

- Instructions for constructing an MCMC using Metropolis-Hastings approach:

1. Choose $\theta^{[0]}$, where $Pr(\theta^{[0]}|\mathbf{y}) > 0$.

2. Sample a *proposal* parameter value θ^* from a jumping distribution $J(\theta^*|\theta^{[t]})$, where $t = 0$ or any subsequent iteration.

3. Calculate $r = \frac{Pr(\theta^*|\mathbf{y})J(\theta^{[t]}|(\theta^*)}{Pr(\theta^{[t]}|\mathbf{y})J(\theta^*|\theta^{[t]})}$.

4. Set $\theta^{[t+1]} = \theta^*$ with $Pr(\theta^{[t+1]} = \theta^*) = min(r, 1)$ and $\theta^{[t+1]} = \theta^{[t]}$ with $Pr(\theta^{[t+1]} = \theta^{[t]}) = 1 - min(r, 1)$.

- Running the MCMC algorithm:

1. Set up the Metropolis-Hastings algorithm.

2. Initialize the values for $\theta^{[0]}$.

3. Iterate the algorithm for $t >> 0$, such that we are past t_{ab}, which is the iteration after the 'burn-in' phase, where the realizations of $\theta^{[t]}$ start to behave as though they are sampled from the stationary distribution of the Metropolis-Hastings Markov chain (we will discuss how many iterations are necessary for a burn-in below).

4. Sample the chain for a set of iterations after the burn-in and use these to approximate the posterior distribution and perform Bayesian inference.

Constructing an MCMC for genetic analysis

- For a given marker part of our GWAS, we define our glm (which gives us our likelihood) and our prior (which we provide!), and our goal is then to construct an MCMC with a stationary distribution (which we will sample to get the posterior "histogram":

$$\theta^{[t]} = \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \\ \sigma^2_\epsilon \end{bmatrix}^{[t]}, \quad \theta^{[t+1]} = \begin{bmatrix} \beta_\mu \\ \beta_a \\ \beta_d \\ \sigma^2_\epsilon \end{bmatrix}^{[t+1]}, \ldots$$

- One approach is setting up a Metropolis-Hastings algorithm by defining a jumping distribution

- Another approach is to use a special case of the Metropolis-Hastings algorithm called the Gibbs sampler (requires no rejections!), which samples each parameter from the conditional posterior distributions (which requires you derive these relationships = not always possible!)

$$Pr(\beta_\mu | \beta_a, \beta_d, \sigma^2_\epsilon, \mathbf{y})$$

$$Pr(\beta_a | \beta_\mu, \beta_d, \sigma^2_\epsilon, \mathbf{y})$$

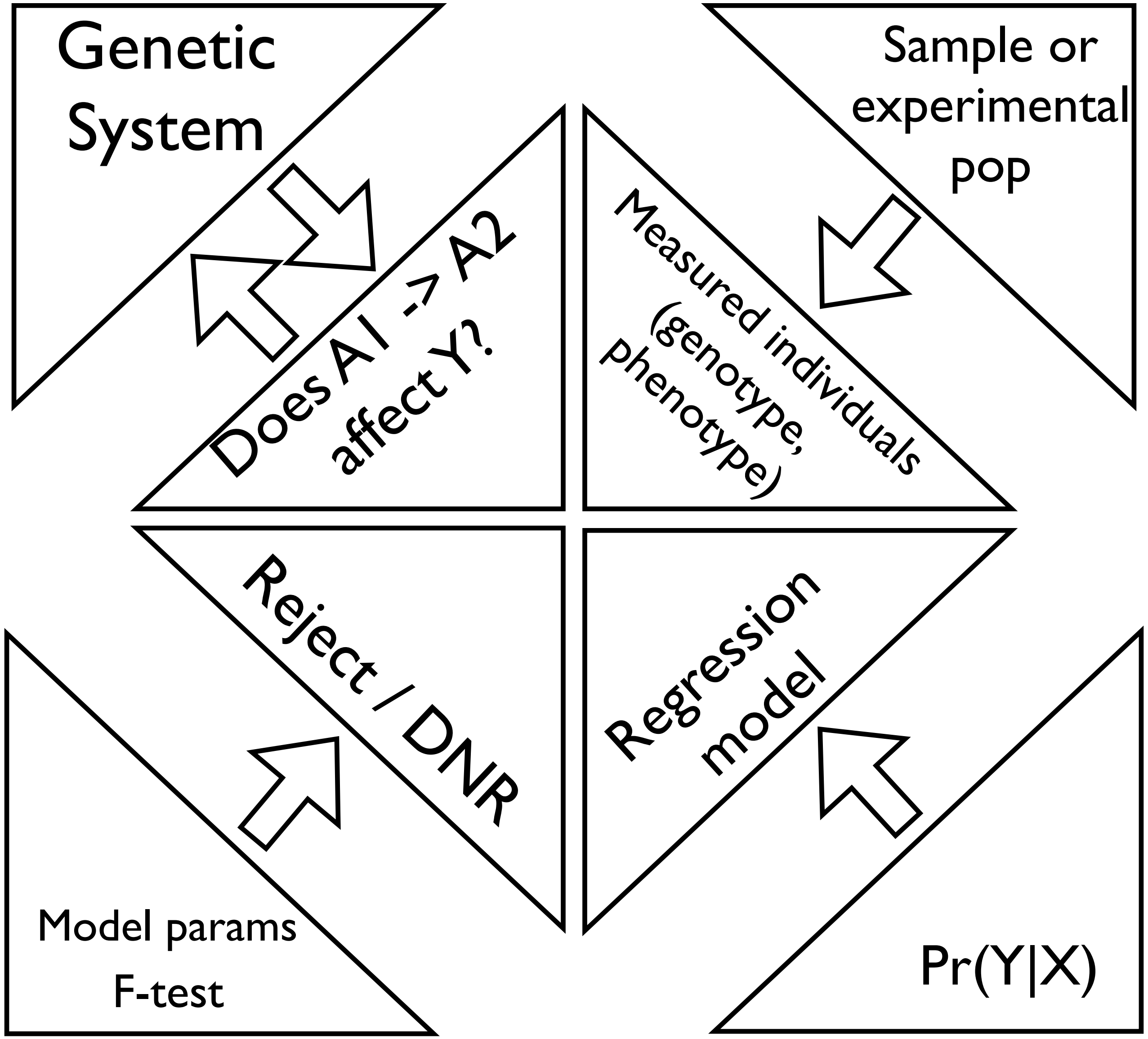
$$Pr(\beta_d | \beta_\mu, \beta_a, \sigma^2_\epsilon, \mathbf{y})$$

$$Pr(\sigma^2_\epsilon | \beta_\mu, \beta_a, \beta_d, \mathbf{y})$$

Importance for MCMC

- Constructing MCMC for Bayesian inference is extremely practical
- The constraint is they are computationally intensive
- This is one reason for the surge in the practical use of Bayesian data analysis is when computers increased in speed
- This is definitely the case where the number of Bayesian MCMC approaches in genetic analysis has steadily increased over the last decade or so
- One issue is that, even with a fast computer, MCMC algorithms can be inefficient (they take a long time to converge, they do not sample modes of a complex posterior efficiently, etc.)
- There are therefore other algorithm approaches to Bayesian genetic inference, e.g. variational Bayes

Conceptual Overview



Genetic System

Sample or experimental pop

Does A1 -> A2 affect Y?

Measured individuals (genotype, phenotype)

Reject / DNR

Regression model

Model params

F-test

Pr(Y|X)

GWAS definitions I

- **Association analysis** - any analysis involving a statistical assessment of a relation between genotype and phenotype, e.g. a hypothesis test involving a multiple regression model
- **Mapping analysis** - an association analysis
- **Linkage disequilibrium (LD) mapping** - an association analysis
- **Segregating** - any locus where there is more than one allele in the population
- **Genetic marker** - any segregating polymorphism we have measured in a GWAS, i.e. SNPs genotyped in a GWAS
- **Tag SNP** - a SNP correlated with a causal polymorphism
- **Locus** or **Genetic Locus** - a position in the genome (which may refer to a single polymorphism or an entire genomic segment, e.g. that contains the coding region of a gene

GWAS definitions II

- **Mendelian trait** - any phenotype largely affected by one or at most two loci where environment does not have a large effect on the phenotype
- **Complex trait** - any phenotype affected by more than one or two loci and/or where environmental effects account for most of the variation we observe in a population
- **Quantitative trait** - a complex trait

That’s it for today

- Next week (last lecture!): introduction to pedigree, inbred line and evolutionary quantitative genomics - and continuing thoughts!