Lecture 24: (Brief) Introduction to Bayesian Inference

Jason Mezey
jgm45@cornell.edu
May 5, 2020 (Th) 8:40-9:55
Announcements

• The FINAL EXAM (!!!)
  • Same format as midterm (i.e., take home, open book, no restrictions on material you may access BUT ONCE THE EXAM STARTS YOU MAY NOT ASK ANYONE ABOUT ANYTHING THAT COULD RELATE TO THE EXAM (!!!!)

• Timing: Available Evening May 16 (!!) (Sat.) and will be due 11:59PM May 20 (Weds.)

• If you prepare, the exam should take 8-12 hours (i.e., allocate about 1 day if you are well prepared)

• You will have to do a logistic regression analysis of GWAS!
Summary of lecture 24

• Today we will complete our discussion of mixed models
• Discuss where to go for more depth on some other GWAS subjects
• …and we will begin our (very brief) Introduction to Bayesian Statistics
Review: Intro to mixed models

- Recall our linear regression model has the following structure:
  \[ y_i = \beta_\mu + X_{i,a} \beta_a + X_{i,d} \beta_d + \epsilon_i \quad \epsilon_i \sim N(0, \sigma_\epsilon^2) \]

- For example, for \( n=2 \):
  \[
  \begin{align*}
  y_1 &= \beta_\mu + X_{1,a} \beta_a + X_{1,d} \beta_d + \epsilon_1 \\
  y_2 &= \beta_\mu + X_{2,a} \beta_a + X_{2,d} \beta_d + \epsilon_2
  \end{align*}
  \]

- What if we introduced a correlation?
  \[
  \begin{align*}
  y_1 &= \beta_\mu + X_{1,a} \beta_a + X_{1,d} \beta_d + a_1 \\
  y_2 &= \beta_\mu + X_{2,a} \beta_a + X_{2,d} \beta_d + a_2
  \end{align*}
  \]
Review: mixed model formalism

- The formal structure of a mixed model is as follows:

\[ y = X\beta + Za + \epsilon \]

\[ \epsilon \sim multiN(0, I\sigma^2_\epsilon) \quad a \sim multiN(0, A\sigma^2_a) \]

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  \vdots \\
  y_n
\end{bmatrix} =
\begin{bmatrix}
  1 & X_{i,a} & X_{i,d} \\
  1 & X_{i,a} & X_{i,d} \\
  1 & X_{i,a} & X_{i,d} \\
  \vdots & \vdots & \vdots \\
  1 & X_{i,a} & X_{i,d}
\end{bmatrix}
\begin{bmatrix}
  \beta_\mu \\
  \beta_\alpha \\
  \beta_d
\end{bmatrix} +
\begin{bmatrix}
  1 & 0 & 0 & 0 & 0 \\
  0 & 1 & 0 & 0 & 0 \\
  0 & 0 & 1 & 0 & 0 \\
  \vdots & \vdots & \vdots & \vdots & \vdots \\
  0 & \ldots & \ldots & \ldots & 1
\end{bmatrix}
\begin{bmatrix}
  a_1 \\
  a_2 \\
  a_3 \\
  \vdots \\
  a_n
\end{bmatrix} +
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

- Note that \( X \) is called the “design” matrix (as with a GLM), \( Z \) is called the “incidence” matrix, the \( a \) is the vector of random effects and note that the \( A \) matrix determines the correlation among the ai values where the structure of \( A \) is provided from external information (!!)
Review: mixed model inference

• We perform inference (estimation and hypothesis testing) for the mixed model just as we would for a GLM (!!)

• Note that in some applications, people might be interested in estimating the variance components $\sigma^2_\epsilon, \sigma^2_a$ but for GWAS, we are generally interested in regression parameters for our genotype (as before!): $\beta_a, \beta_d$

• For a GWAS, we will therefore determine the MLE of the genotype association parameters and use a LRT for the hypothesis test, where we will compare a null and alternative model (what is the difference between these models?)
Review: mixed model likelihood

- To estimate parameters, we will use the MLE, so we are concerned with the form of the likelihood equation

\[ L(\beta, \sigma_a^2, \sigma_\epsilon^2 | y) = \int_{-\infty}^{\infty} Pr(y | \beta, a, \sigma_\epsilon^2) Pr(a | A\sigma_a^2) da \]

\[ L(\beta, \sigma_a^2, \sigma_\epsilon^2 | y) = \left| I \sigma_\epsilon^2 \right|^{\frac{-1}{2}} e^{-\frac{1}{2\sigma_\epsilon^2} [y - X\beta - Z\hat{a}]^T [y - X\beta - Z\hat{a}]} \left| A \sigma_a^2 \right|^{\frac{-1}{2}} e^{-\frac{1}{2\sigma_a^2} a^T A^{-1} a} \]

\[ l(\beta, \sigma_a^2, \sigma_\epsilon^2 | y) \propto -\frac{n}{2} \ln \sigma_a^2 - \frac{n}{2} \ln \sigma_\epsilon^2 - \frac{1}{\sigma_\epsilon^2} [y - X\beta - Z\hat{a}]^T [y - X\beta - Z\hat{a}] + \frac{1}{2\sigma_a^2} a^T A^{-1} a \]

- Unfortunately, there is no closed form for the MLE since they have the following form:

\[ MLE(\hat{\beta}) = (X\hat{V}^{-1} X^T)^{-1} X^T \hat{V}^{-1} y \]

\[ MLE(\hat{V}) = f(X, \hat{V}, y, A) \]

\[ V = \sigma_a^2 A + \sigma_\epsilon^2 I \]
where the values of parameters are maximized. In the mixed model framework, the 'miss-
ing' data is the vector \( \mathbf{a} \) and the parameters are \( \beta_0, \beta_1, \beta_2, \sigma_0, \sigma_1, \sigma_2 \). With these defined, we can write out the steps of the EM algorithm as follows:

1. At step \([t]\) for \( t = 0 \), assign values to the parameters: \( \beta[0] = [\beta_0[0], \beta_1[0], \beta_2[0]], \sigma_0[0], \sigma_1[0], \sigma_2[0] \).

   These need to be selected such that they are possible values of the parameters (e.g. no negative values for the variance parameters).

2. Calculate the expectation step for \([t]\):

   \[
   \mathbf{a}[t] = \left( \mathbf{Z}^T \mathbf{Z} + \mathbf{A}^{-1} \frac{\sigma_2[0,t-1]}{\sigma_2[0,t-1]} \right)^{-1} \mathbf{Z}^T (\mathbf{y} - \mathbf{x}\beta[t-1])
   \]

   \[
   \mathbf{V}_a[t] = \left( \mathbf{Z}^T \mathbf{Z} + \mathbf{A}^{-1} \frac{\sigma_2[0,t-1]}{\sigma_2[0,t-1]} \right)^{-1} \sigma_2[0,t-1]
   \]

3. Calculate the maximization step for \([t]\):

   \[
   \beta[t] = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T (\mathbf{y} - \mathbf{Z}\mathbf{a}[t])
   \]

   \[
   \sigma_a^2[t] = \frac{1}{n} \left[ \mathbf{a}[t] \mathbf{A}^{-1} \mathbf{a}[t] + tr(\mathbf{A}^{-1} \mathbf{V}_a[t]) \right]
   \]

   \[
   \sigma_\epsilon^2[t] = -\frac{1}{n} \left[ \mathbf{y} - \mathbf{x}\beta[t] - \mathbf{Z}\mathbf{a}[t] \right]^T \left[ \mathbf{y} - \mathbf{x}\beta[t] - \mathbf{Z}\mathbf{a}[t] \right] + tr(\mathbf{Z}^T \mathbf{Z} \mathbf{V}_a[t])
   \]

   where \( tr \) is a trace function, which is equal to the sum of the diagonal elements of a matrix.

4. Iterate steps 2, 3 until \((\beta[t], \sigma_a^2[t], \sigma_\epsilon^2[t]) \approx (\beta[t+1], \sigma_a^2[t+1], \sigma_\epsilon^2[t+1])\) (or alternatively \( lnL[t] \approx lnL[t+1] \)).
Mixed models: p-values

- For hypothesis testing, we will calculate a LRT:

\[
LRT = -2 \ln \Lambda = 2l(\hat{\theta}_1|y) - 2l(\hat{\theta}_0|y)
\]

- To do this, run the EM algorithm twice, once for the null hypothesis (again what is this?) and once for the alternative (i.e. all parameters unrestricted) and then substitute the parameter values into the log-likelihood equations and calculate the LRT

- The LRT is then distributed (asymptotically) as a Chi-Square distribution with two degrees of freedom (as before!)
Mixed models inference in practice

• In general, a mixed model is an advanced methodology for GWAS analysis but is proving to be an extremely useful technique for covariate modeling

• There is software for performing a mixed model analysis (e.g. R-package: lrgpr, EMMAX, FAST-LMM, TASSEL, etc.)

• Mastering mixed models will take more time than we have to devote to the subject in this class, but what we have covered provides a foundation for understanding the topic
Construction of $A$ matrix 1

- The matrix $A$ is an $n \times n$ covariance matrix (what is the form of a covariance matrix?)

- Where does $A$ come from? This depends on the modeling application...

- In GWAS, the random effect is usually used to account for population structure OR relatedness among individuals
  - For relatedness, we use estimates of identity by descent, which can be estimated from a pedigree or genotype data
  - For population structure, a matrix is constructed from the covariance (or similarity) among individuals based on their genotypes
Construction of $\mathbf{A}$ matrix II

$\text{Data} = \begin{bmatrix} z_{11} & \cdots & z_{1k} & y_{11} & \cdots & y_{1m} & x_{11} & \cdots & x_{1N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n1} & \cdots & z_{nk} & y_{n1} & \cdots & y_{nm} & x_{11} & \cdots & x_{nN} \end{bmatrix}$

- Calculate the $n \times n$ ($n=$sample size) covariance matrix for the individuals in your sample across all genotypes - this is a reasonable $\mathbf{A}$ matrix!

- There is software for calculating $\mathbf{A}$ and for performing a mixed model analysis (e.g. R-package: lrgpr, EMMAX, FAST-LMM, TASSEL, etc.)

- Mastering mixed models will take more time than we have to devote to the subject in this class, but what we have covered provides a foundation for understanding the topic
Topics that we don’t have time to cover (but 2019 lectures available!) - will briefly mention last lecture…

- Alternative tests in GWAS (2019 Lecture 19)
- Haplotype testing (2019 Lecture 19)
- Multiple regression analysis / epistasis (2019 Lecture 21)
- Multivariate regression analysis / eQTL (2019 Lecture 21)
- Basics of linkage analysis (2019 Lecture 24)
- Basics of inbred line analysis (2019 Lecture 25)
- Basics of evolutionary quantitative genetics (2019 Lecture 25)
Introduction to Bayesian analysis

- Up to this point, we have considered statistical analysis (and inference) using a Frequentist formalism.

- There is an alternative formalism called Bayesian that we will now introduce in a very brief manner.

- Note that there is an important conceptual split between statisticians who consider themselves Frequentist of Bayesian but for GWAS analysis (and for most applications where we are concerned with analyzing data) we do not have a preference, i.e. we only care about getting the “right” biological answer so any (or both) frameworks that get us to this goal are useful.

- In GWAS (and mapping) analysis, you will see both frequentist (i.e. the framework we have built up to this point!) and Bayesian approaches applied.
Introduction to Bayesian analysis II

- In both frequentist and Bayesian analyses, we have the same probabilistic framework (sample spaces, random variables, probability models, etc.) and when assuming our probability model falls in a family of parameterized distributions, we assume that a single fixed parameter value(s) describes the true model that produced our sample.

- However, in a Bayesian framework, we now allow the parameter to have its own probability distribution (we DO NOT do this in a frequentist analysis), such that we treat it as a random variable.

- This may seem strange - how can we consider a parameter to have a probability distribution if it is fixed?

- However, we can if we have some prior assumptions about what values the parameter value will take for our system compared to others and we can make this prior assumption rigorous by assuming there is a probability distribution associated with the parameter.

- It turns out, this assumption produces major differences between the two analysis procedures (in how they consider probability, how they perform inference, etc.)
Introduction to Bayesian analysis III

• To introduce Bayesian statistics, we need to begin by introducing Bayes theorem

• Consider a set of events (remember events!?) \( A = A_1...A_k \) of a sample space \( \Omega \) (where \( k \) may be infinite), which form a partition of the sample space, i.e. \( \bigcup_i^k A_i = \Omega \) and \( A_i \cap A_j = \emptyset \) for all \( i \neq j \)

• For another event \( B \subset \Omega \) (which may be \( \Omega \) itself) define the Law of total probability:

\[
Pr(B) = \sum_{i=1}^{k} Pr(B \cap A_i) = \sum_{i=1}^{k} Pr(B|A_i)Pr(A_i)
\]

• Now we can state Bayes theorem:

\[
Pr(A_i|B) = \frac{Pr(A_i \cap B)}{Pr(B)} = \frac{Pr(B|A_i)Pr(A_i)}{Pr(B)} = \frac{Pr(B|A_i)Pr(A)}{\sum_{i=1}^{k} Pr(B|A_i)Pr(A_i)}
\]
Introduction to Bayesian analysis IV

• 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 $Y$ produced under a probability model:

$$Pr(\theta \cap Y)$$

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

$$Pr(\theta | y)$$

• Using Bayes theorem, we can write:

$$Pr(\theta | y) = \frac{Pr(y|\theta)Pr(\theta)}{Pr(y)}$$

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

$$Pr(\theta | y) \propto Pr(y|\theta)Pr(\theta)$$
Introduction to Bayesian analysis

- 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
Probability in a Bayesian framework

• By allowing for the parameter to have an prior probability distribution, we produce a change in how we consider probability in a Bayesian versus Frequentist perspective

• For example, consider a coin flip, with Bern(p)
  
  • In a Frequentist framework, we consider a conception of probability that we use for inference to reflect the outcomes as if we flipped the coin an infinite number of times, i.e. if we flipped the coin 100 times and it was “heads” each time, we do not use this information to change how we consider a new experiment with this same coin if we flipped it again
  
  • In a Bayesian framework, we consider a conception of probability can incorporate previous observations, i.e. if we flipped a coin 100 times and it was “heads” each time, we might want to incorporate this information in to our inferences from a new experiment with this same coin if we flipped it again

• Note that this philosophic distinction is very deep (=we have only scratched the surface with this one example)
Debating the Frequentist versus Bayesian frameworks

• Frequentists often argue that because they “do not” take previous experience into account when performing their inference concerning the value of a parameter, such that they do not introduce biases into their inference framework

• In response, Bayesians often argue:
  • Previous experience is used to specify the probability model in the first place
  • By not incorporating previous experience in the inference procedure, prior assumptions are still being used (which can introduce logical inconsistencies!)
  • The idea of considering an infinite number of observations is not particular realistic (and can be a non-sensical abstraction for the real world)
  • The impact of prior assumptions in Bayesian inference disappear as the sample size goes to infinite

• Again, note that we have only scratched the surface of this debate!
That’s it for today

• See you Thurs.!