Lecture 21: Introduction to Mixed Models
Announcements

• Midterm grading in process
• Project is posted
  • We post slides and videos today
  • We will start answering questions on Piazza today
• Reminder: office hours today
You have now reached a stage when you are ready to perform a real GWAS data on your own (please note that there is more to learn and analyzing GWAS well requires that you jump in and analyze!!)

Our final concept to allow you to do this are minimal GWAS steps, i.e. a list of analyses you should always do when analyzing GWAS data (you now know how to do most of these, a few you will have to do additional work to figure out)

While these minimal steps are fuzzy (=they do not apply in every situation!) they provide a good guide to how you should think about analyzing your GWAS data (in fact, no matter how experienced you become, you will always consider these steps!)
Review: Minimal GWAS II

- The minimal steps are as follows:
  - Make sure you understand the data and are clear on the components of the data
  - Check the phenotype data
  - Check and filter the genotype data
  - Perform a GWAS analysis and diagnostics
  - Present your final analysis and consider other evidence

- Note I: the software PLINK (google it!) is a very useful tool for some (but not all) of these steps (but you can do everything in R!)

- Note II: GWAS analysis is not “do this and you are done” - it requires that you consider the output of each step (does it make sense? what does it mean in this case?) and that you use this information to iteratively change your analysis / try different approaches to get to your goal (what is this goal!?)
Review: Minimal GWAS III

• List ALL of the steps (methods!) you have taken to analyze the data such that someone could replicate what you did from your description (!!), i.e. what data did you remove? what intermediate analyses did you do? how did you analyze the data? if you used software what settings did you use?

• Plot a Manhattan and QQ plot (at least!)

• Present your hits (many ways to do this)

• Consider other information available from other sources (databases, literature) to try to determine more about the possible causal locus, i.e. are there good candidate loci, control regions, known genome structure, gene expression or other types of data, pathway information, etc.
Comparing results of multiple analyses of the same GWAS data I

• I’ve run my initial analyses using several tests and produced the following (now what!?):
The best case is that the same markers (SNPs) pass a multiple test correction regardless of the testing approach used, i.e. the result is robust to testing approach.

In cases where this does not happen (most) it becomes helpful to understand why test results could be different:

- Are tests capturing additive vs. dominance effects?
- Are tests less powerful than others or depend on certain assumptions being true? Are they handling missing data in different ways?
- Are particular covariates altering the results if included/excluded? Why might this be?
- Does it depend on how you partition the data (e.g. batch effects)?

This can help narrow down the set of tests you feel are the most informative. In general, a good publishing strategy is limiting yourself to one or two tests that both give you significant results that you believe!
Comparing results of multiple analyses of the same GWAS data III

- Beyond limiting the number of tests applied, there is no perfect way to put results together and/or get your paper past review.

- For your own work (i.e. getting the right answer = locus) it is often helpful to compile a ranked list of the tests you find most useful and check the following:

  - Which ones rank highly on both?

  - Are there any that are picking out different SNPs but these SNPs are in LD and therefore are likely tagging the same causal effect (for humans, Haploview may be a helpful tool for this purpose).

http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview
Comparing results of multiple analyses of the same GWAS data IV

- Overall the most convincing approaches will have components of the following: 1. A known mapped locus should be identifiable with the approach, 2. The hits identify loci / genomic positions that are stable as you add more data, 3. The hits identify loci / genomic positions that can be replicated in an independent GWAS experiment (that you conduct or someone else conducts).
Review: Modeling covariates

• Say you have GWAS data (a phenotype and genotypes) and your GWAS data also includes information on a number of covariates, e.g. male / female, several different ancestral groups (different populations!!), other risk factors, etc.

• First, you need to figure out how to code the $X_z$ in each case for each of these, which may be simple (male / female) but more complex with others (where how to code them involves fuzzy rules, i.e. it depends on your context!!)

• Second, you will need to figure out which to include in your analysis (again, fuzzy rules!) but a good rule is if the parameter estimate associated with the covariate is large (=significant individual p-value) you should include it!

• There are many ways to figure out how to include covariates (again a topic in itself!!)
People geographically separate through migration and then the set of alleles present in the population evolves (=changes) over time.
“Population structure” or “stratification” is a case where a sample includes groups of people that fit into two or more different ancestry groups (fuzzy def!)

Population structure is often a major issue in GWAS where it can cause lots of false positives if it is not accounted for in your model

Intuitively, you can model population structure as a covariate if you know:

- How many populations are represented in your sample
- Which individual in your sample belongs to which population

QQ plots are good for determining whether there may be population structure

“Clustering” techniques are good for detecting population structure and determining which individual is in which population (=ancestry group)

Mixed models provide an excellent covariate approach to account for population structure
A *mixed model* describes a class of models that have played an important role in early quantitative genetic (and other types) of statistical analysis before genomics (if you are interested, look up variance component estimation).

These models are now used extensively in GWAS analysis as a tool for model covariates (often population structure!).

These models considered effects as either “fixed” (they types of regression coefficients we have discussed in the class) and “random” (which just indicates a different model assumption) where the appropriateness of modeling covariates as fixed or random depends on the context (fuzzy rules!).

These models have logistic forms but we will introduce mixed models using linear mixed models (“simpler”).
Introduction to mixed models II

- Recall that for a linear regression of sample size n, we model the distributions of n total $y_i$ phenotypes using a linear regression model with normal error:

$$y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i \quad \epsilon_i \sim N(0, \sigma_\epsilon^2)$$

- A reminder about how to think about / visualize multivariate (bivariate) normal distributions and marginal normal distributions:

- We can therefore consider the entire sample of $y_i$ and their associated error in an equivalent multivariate setting:

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \quad \epsilon \sim \text{multi}N(\mathbf{0}, \mathbf{I}\sigma_\epsilon^2)$$
Introduction to mixed models III

- 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 \( n=2 \):
  \[ y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + \epsilon_1 \]
  \[ y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + \epsilon_2 \]

- What if we introduced a correlation?
  \[ y_1 = \beta_\mu + X_{1,a}\beta_a + X_{1,d}\beta_d + \alpha_1 \]
  \[ y_2 = \beta_\mu + X_{2,a}\beta_a + X_{2,d}\beta_d + \alpha_2 \]
Introduction to mixed models IV

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

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

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

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

• 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 \( a_i \) values where the structure of \( A \) is provided from external information (!!!)
Introduction to mixed models V

• 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 population structure, a matrix is constructed from the covariance (or similarity) among individuals based on their genotypes

  • For relatedness, we use estimates of identity by descent, which can be estimated from a pedigree or genotype data
Introduction to mixed models VI

• 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 $\sigma_e^2, \sigma_a^2$ 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?)
Mixed models: inference 1

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

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

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

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

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

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

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

\[
V = \sigma_a^2 A + \sigma_\epsilon^2 I
\]
Mixed models: inference II

• We therefore need an algorithm to find the MLE for the mixed model

• We will introduce the EM (Expectation-Maximization) algorithm for this purpose, which is an algorithm with good theoretical and practical properties, e.g. hill-climbing algorithm, guaranteed to converge to a (local) maximum, it is a stable algorithm, etc.

• We do not have time to introduce these properties in detail so we will just show the steps / equations you need to implement this algorithm (such that you can implement it yourself = see computer lab next week where this will be an optional assignment!)
Mixed models: inference III

1. At step $[t]$ for $t = 0$, assign values to the parameters: $\beta^{[0]} = [\beta^{[0]}_{\mu}, \beta^{[0]}_a, \beta^{[0]}_d, \sigma^{2,[0]}_a, \sigma^{2,[0]}_\epsilon]$. 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 $[t]$:

   \[ a^{[t]} = \left( Z^T Z + A^{-1} \frac{\sigma^{2,[t-1]}_\epsilon}{\sigma^{2,[t-1]}_a} \right)^{-1} Z^T (y - x \beta^{[t-1]}) \]

   \[ V_a^{[t]} = \left( Z^T Z + A^{-1} \frac{\sigma^{2,[t-1]}_\epsilon}{\sigma^{2,[t-1]}_a} \right)^{-1} \sigma^{2,[t-1]}_\epsilon \]

3. Calculate the maximization step for $[t]$:

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

   \[ \sigma^{2,[t]}_a = \frac{1}{n} \left[ a^{[t]} A^{-1} a^{[t]} + tr(A^{-1} V_a^{[t]}) \right] \]

   \[ \sigma^{2,[t]}_\epsilon = -\frac{1}{n} \left[ y - x \beta^{[t]} - Z a^{[t]} \right]^T \left[ y - x \beta^{[t]} - Z a^{[t]} \right] + tr(Z^T Z V_a^{[t]}) \]

   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 \((\beta^{[t]}, \sigma^{2,[t]}_a, \sigma^{2,[t]}_\epsilon) \approx (\beta^{[t+1]}, \sigma^{2,[t+1]}_a, \sigma^{2,[t+1]}_\epsilon)\) (or alternatively \(lnL^{[t]} \approx lnL^{[t+1]}\)).
Mixed models: inference IV

• 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!)
• 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
That’s it for today

- See you on Tues.!