Quantitative Genomics and Genetics
BTRY 4830/6830; PBSB.5201.01

Lecture 19: Minimal GWAS analysis steps and Brief introduction to the mixed model (and EM algorithm)

Jason Mezey
jgm45@cornell.edu
April 21, 2016 (Th) 8:40-9:55
Announcements

• Project is due the (11:59PM) last day of class!

• Final will be on DURING THE **FIRST** WEEK OF EXAMS: available Mon., May 16 and due Thurs., May 19
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!)
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!?)
Minimal GWAS III: check data

- Look at the files (!!) using a text editor (if they are too large to do this - you will need another approach)

- Make sure you can identify: phenotypes, genotypes, covariates, and that you know what all other information indicates, i.e. indicators of the structure of the data, missing data, information that is not useful, etc. (also make sure you do not have any strange formatting, etc. in your file that will mess up your analysis!)

- Make sure you understand how phenotypes are coded and what they represent (how are they collected? are they the same phenotype?) and the structure of the genotype data (are they SNPs? are there three states for each?) - ideally talk to your collaborator about this (!!)
Minimal GWAS IV: phenotype data

- Plot your phenotype data (histogram!)
- Check for odd phenotypes or outliers (remove if applicable)
- Make sure it conforms to a distribution that you expect and can model (!!) - this will determine which analysis techniques you can use
  - e.g. if the data is continuous, is it approximately normal (or can be transformed to normal?)
  - e.g. if it has two states, make sure you have coded the two states appropriately and know what they represent (are there enough in each category to do an analysis?)
  - e.g. what if your phenotype does not conform to either?
Minimal GWAS V: genotype data

- Make sure you know how many states you have for your genotypes and that they are coded appropriately

- Filter your genotypes (fuzzy rules!):
  - Remove individuals with >10% missing data across all genotypes (also remove individuals without phenotypes!)
  - Remove genotypes with >5% missing data across the entire individual
  - Remove genotypes with MAF < 5%
  - Remove individuals that fail a test of Hardy-Weinberg equilibrium (where appropriate!)
  - Remove individuals that fail transmission, sex chromosome test, etc.

- Perform a Principal Component Analysis (PCA) to check for clustering of n individuals (population structure!) or outliers, i.e. use the covariance matrix among individuals after scaling genotypes (by mean and sd) and look at the loadings of each individual on the PCs (you may have to “thin” the data!)
Minimal GWAS VI: GWAS analysis

- Perform an association analysis considering the association of each marker one at a time (always do this not matter how complicated your experimental design!)

- Apply as many individual analyses as you find informative (i.e. perform individual GWAS each with a different statistical analysis technique), e.g. trying different sets of covariates, different types of tests (see next lecture!), etc.

- **CHECK QQ PLOTS FOR EACH INDIVIDUAL GWAS ANALYSIS** and use this information to indicate if your analysis can be interpreted as indicating the positions of causal polymorphisms (if not, try more analyses, different filtering, etc. = experience is key!)

- For significant markers (multiple test correction!) do a “local” Manhattan plot and visualize the LD among the markers (r^2 or D’ if possible but just a correlation of you Xa can work) to determine if anything might be amiss

- Compare significant “hits” among different analyses (what might be causing the differences if there are any?)
Minimal GWAS VII: present results

• 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’ve run my initial analyses using several tests and produced the following (now what!?):
Comparing results of multiple analyses of the same GWAS data II

• 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).
Conceptual Overview

Genetic System

Does A1 -> A2 affect Y?

Sample or experimental pop

Measured individuals (genotype, phenotype)

Pr(Y|X)

Regression model

Reject / DNR

Model params
F-test
(Brief) introduction to mixed models I

- 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:

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

- Recall our linear regression model has the following structure:

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

- 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?

\[
\begin{align*}
y_1 &= \beta_\mu + X_{1,a} \beta_a + X_{1,d} \beta_d + \alpha_1 \\
y_2 &= \beta_\mu + X_{2,a} \beta_a + X_{2,d} \beta_d + \alpha_2
\end{align*}
\]
Introduction to mixed models IV

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

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

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

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

- 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 (!!)
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 I

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

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

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

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

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

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

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

\[
V = \sigma_a^2 A + \sigma_e^2 I
\]
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: \(\beta^{[0]} = [\beta_{[0]}^{[0]}, \beta_{a}^{[0]}, \beta_{d}^{[0]}], \sigma_{a}^{2,[0]}, \sigma_{\epsilon}^{2,[0]}\).
   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 \([t]\):
   \[
   a^{[t]} = \left( Z^{T}Z + A^{-1}\frac{\sigma_{\epsilon}^{2,[t-1]}}{\sigma_{a}^{2,[t-1]}} \right)^{-1} Z^{T}(y - x\beta^{[t-1]})
   \]
   \[
   V_{a}^{[t]} = \left( Z^{T}Z + A^{-1}\frac{\sigma_{\epsilon}^{2,[t-1]}}{\sigma_{a}^{2,[t-1]}} \right)^{-1} \sigma_{\epsilon}^{2,[t-1]}
   \]

3. Calculate the maximization step for \([t]\):
   \[
   \beta^{[t]} = (x^{T}x)^{-1}x^{T}(y - Za^{[t]})
   \]
   \[
   \sigma_{a}^{2,[t]} = \frac{1}{n} \left[ a^{[t]} A^{-1} a^{[t]} + tr(A^{-1}V_{a}^{[t]}) \right]
   \]
   \[
   \sigma_{\epsilon}^{2,[t]} = -\frac{1}{n} \left[ y - x\beta^{[t]} - Za^{[t]} \right]^{T} \left[ y - x\beta^{[t]} - Za^{[t]} \right] + tr(Z^{T}ZV_{a}^{[t]})
   \]
   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: 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

- Next lecture: we will discuss epistasis (!!)
- We will also introduce alternative testing options in a GWAS