Statistical Genetics, Fall 2025

Class notes 9: Linear Mixed Models for heritability estimation

Reminder: *The mystery of missing heritability* has been on the statistical genetics community's mind for well over a decade. Leading theories on the explanations (obviously, the truth may be a mixture of all of them):

1. Instead of relatively small number of variants with big effect, the heritability is due to a large/huge number of variants with tiny effects
   ⇒ Even with huge sample size, there is not enough *power* to identify these effects and declare them significant

2. Rare variants with big effect are the main cause of heritability
   ⇒ Each "tall family" or family with cancer is has it for a different genetic reason
   Note that such variants are not even measured in traditional GWAS, which considers only *common* variants. Also, studies of unrelated individuals by definition have no power to identify such variants

3. Additive model is wrong, this can have several consequences:
   - Twin estimates may be (badly) inflated due to assuming additive model
   - GWAS which tests one site at a time might miss strong interactions

4. Epigenetics: effects which are heritable and affect the genome, but are not expressed in the sequence. This includes methylation as well as more mysterious effects.

The Linear Mixed Model (LMM) approach

To possibly deal with the first issue above (large number of small effects), we take this to extreme, and assume all variants have a tiny effect, giving the following model:

$$Y_i = \sum_{j=1}^{M} Z_{ij} b_j + \epsilon_i \ , \quad i = 1, \ldots, n,$$

where $Z_{n \times M}$ is the genotype matrix.

In the standard view, this model has M parameters (say $M = 10^6$), so it is not really that useful. But we can adopt the random effects approach, which now considers b_j as random variables, say assuming: $b_j \sim N(0, \sigma_b^2)$, $\forall j$, and that the b_j's are independent between them. Then, assuming $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, we get:

$$Y|Z \sim N\left(0, ZZ^t\sigma_b^2 + I_n\sigma_\epsilon^2\right),$$

note that by defining the coefficients as random variables, we have moved them from the mean part of Y into the covariance matrix. ZZ^t is an $n \times n$ matrix, where $(ZZ^t)_{ij} = \sum_{k=1}^M Z_{ik}Z_{jk}$. It is common to denote:

$$G = \frac{ZZ^t}{M} \ , \ \sigma_g^2 = M\sigma_b^2 \ \Longrightarrow Y|Z \sim N(0, G\sigma_g^2 + I\sigma_\epsilon^2).$$

Note that this is a problem with two parameters only: $\sigma_g^2, \sigma_\epsilon^2$, and if we assume that the phenotype is standardized to have norm 1, only one parameter, σ_g^2, with $\sigma_\epsilon^2 = 1 - \sigma_g^2$, and resulting heritability $H^2 = \sigma_g^2$.

A slightly less naive view considers the following additional aspects:

1. The problem may also have fixed effects $X_{n\times p}$ like age, sex, nutrition, or known SNPs with big effects that don't fit this model. The more realistic model is then:

$$Y|Z, X \sim N(X\beta, G\sigma_g^2 + I_n\sigma_\epsilon^2), \tag{1}$$

   with $p + 2$ parameters.

2. The assumption that all coefficients are tiny and of the same magnitude may be too naive, because:

   - It may be more realistic to assume that a small fraction of SNPs have an effect, it will still be a large number and the effects can still be small. This can be expressed as a mixture distribution of zero and a normal (spike and slab)
   - Properties of the SNPs, such as which genomic region they reside in, and what is the function of that region, are likely to be related to their effect size

   So we may want to assume $b_j \sim F$ not normal (such as spike and slab). It is important to note, however, that as long as M is big and this mixture distribution is not too crazy, we can still use the CLT to assume

$$Y|Z \mathrel{\dot\sim} N(0, G\sigma_g^2 + I\sigma_\epsilon^2).$$

3. It is possible to assume that the SNPs are divided into multiple groups, each one with its own variance parameter

The model (1) is known as the LMM as it has both fixed effects β and random effects b. The $p + 2$ parameters are:

- p fixed effects β.
- Two random effects (variance components) parameters $\sigma_g^2, \sigma_\epsilon^2$.

These are typically estimated with maximum likelihood (more accurately, restricted maximum likelihood (REML), which gives less biased estimates of the variance components).

What do the entries in G look like? We typically assume that columns of Z are *standardized*, so approximately (why not exactly?) $\sum_k Z_{ik} \approx 0$, $\sum_k Z_{ik}^2 \approx M$. What about $G_{ij} = \sum_k Z_{ik}Z_{jk}/M$ for $j \neq k$? It is usually assumed that the individuals in the study are *unrelated*, meaning that after standardization we expect that $Z_{ik}Z_{jk}$ be centered around 0. Hence once we sum over all pairs we typically get

$$G_{ij} = \frac{1}{M}\sum_k Z_{ik}Z_{jk} \approx 0 \ , \ \ \forall i \neq j,$$

are very small. However they are not exactly 0, expressing small variations in degree of similarity between pairs of individuals. This is the key in still being able to estimate the variance components — the off-diagonal elements in G being non-zero.

The famous paper of Yang et al. (2010) applies this approach to height GWAS data and demonstrates that the REML estimate gives $\hat{\sigma}_g^2 \approx 0.5$, so finding most of the 80% heritability we are looking for!

However we can do better from a statistical perspective: A common approach is to assume the *Spike and Slab* model, where instead of $b_j \sim N(0, \sigma_b^2)$ iid,we assume $b_j \sim F$ iid, with the same mean and variance, but high probability of being zero, and low probability of being normal with bigger variance:

$$F = (1-\pi)0 + \pi N(0, \sigma_b^2/\pi).$$

An MCEM approach for the spike and slab model

This section follows Golan and Rosset (2011).

Under the spike and slab model, we can observe the following points:

- The basic model (1) is still a decent approximation, as long as π is not too tiny, due to the CLT.
- However, we can also propose a more accurate and explicit model, by considering the identities of the SNPs with non-zero effects as missing data vector $I \in \{0,1\}^M$. Given I, we can write the *exact* likelihood of Y given the complete data:

$$Y|I \sim N\left(X\beta, G_I(\sigma_g^2) + I_n(1-\sigma_g^2)\right) \ , \ \ G_I = \frac{Z_I Z_I^T}{|I|}. \tag{2}$$

- Consequently, we can write the complete data log-likelihood:

$$\ell(\sigma_g^2, \pi; Y, I) = \log\left[\pi^{|I|}(1-\pi)^{M-|I|} \times \mathbb{P}(Y|I; \sigma_g^2)\right]. \tag{3}$$

This last observation leads us to consider an EM approach to estimating the parameters of interest. For simplicity, assume $p = 0$, so we are only estimating the heritability σ_g^2, and the slab probability π. Note that under the iid model, $I_i \sim Ber(\pi)$ iid.

E-step:

$$\ell_r^E(\sigma_b^2, \pi) = \sum_{i=0}^{2^M-1} \mathbb{P}\left(I = i | Y, \left(\sigma_g^2\right)^{(r)}, \pi^{(r)}\right) \ell\left(\sigma_g^2, \pi; Y, I = i\right).$$

And using Bayes theorem:

$$\mathbb{P}\left(I = i | Y, \left(\sigma_g^2\right)^{(r)}, \pi^{(r)}\right) = \frac{\mathbb{P}\left(I = i, Y\right)}{\mathbb{P}\left(Y\right)} = \frac{\binom{M}{|i|}\left(\pi^{(r)}\right)^{|i|}\left((1-\pi)^{(r)}\right)^{M-|i|}\mathbb{P}(Y|I = i; \left(\sigma_g^2\right)^{(r)})}{\sum_j \binom{M}{|j|}\left(\pi^{(r)}\right)^{|j|}\left((1-\pi)^{(r)}\right)^{M-|j|}\mathbb{P}(Y|I = j; \left(\sigma_g^2\right)^{(r)})}.$$

M-step is trivial given this function — simply optimize it numerically relative to the two parameters.

Major problem: the E-step requires summing over 2^M configurations, with $M \approx 10^6$ in typical GWAS!

Solution: Instead of summing, use a stochastic sampling approach like Markov Chain Monte Carlo (MCMC) to sample vectors I according to $\mathbb{P}\left(I = i | Y, \left(\sigma_g^2\right)^{(r)}, \pi^{(r)}\right)$.