Statistical Genetics, Spring 2022

Class notes 8: Heritability estimation and Linear Mixed Models

Notations:

- $Y \in \mathbb{R}^n$ – continuous pheontype (e.g. height)
- $X \in \mathbb{R}^{n \times p}$ vector of *fixed effect* covariates (p typically small), with coefficients $\beta \in \mathbb{R}^p$.
- $Z \in \mathbb{R}^{n \times M}$ vector of *random effect* coefficients (q typically huge), with coefficients $b \in m\mathbb{R}^M$, typically assume $b_j \sim N(0, \sigma_b^2)$ iid
- $\epsilon \sim N(0\sigma_\epsilon^2 I_n)$ vector of iid errors

with $Y = X\beta + Zb + \epsilon$, $G = ZZ^t$ and $\sigma_g^2 = M\sigma_b^2$, this gives us the well known LMM:

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

with $p + 2$ parameters. We note that it is also common to assume $Var(Y) = 1$ and therefore $\sigma_\epsilon^2 = 1 - \sigma_g^2$ and we reduce to $p + 1$ parameters with one variance component σ_g^2.

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 (specifically, subtle stratification like living in the same city). 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:

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

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 $\mathbb{P}\left(I = i|Y, \left(\sigma_g^2\right)^{(r)}, \pi^{(r)}\right)$.

Basic binormal identities

Assume

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(0, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right).$$

Then:

$$Y|X = x \sim N\left(\rho\frac{\sigma_Y}{\sigma_X}x, (1-\rho^2)\sigma_Y^2\right).$$

Heritability estimations and LMMs for disease studies

Assume we have a disease with prevalence K in the population ($\mathbb{P}(Y=1)=K$). Going back to basic definitions, we can adopt the *liability threshold* model, assuming an underlying continuous pheontype:

$$L = G + E \ , \ Y = \mathbb{I}\{L > t\}.$$

Since the *liability* L is unobserved, we can safely assume $\sigma_L^2 = \sigma_g^2 + \sigma_e^2 = 1$, and even $L \sim N(0,1) \ \Rightarrow$ $t = Z_{1-K}$

Now we can investigate the connections between the observed Y and unobserved L:

$$\begin{aligned}
\mathbb{E}(L|Y=1) &= \mathbb{E}(L|L>t) = \frac{\int_t^\infty u\phi(u)du}{1-\Phi(t)} = \frac{1}{K}\int_t^\infty (2\pi)^{-1/2}\exp\{-0.5\cdot u^2\}udu = \\
&= \frac{1}{K}\cdot\left[-(2\pi)^{-1/2}\exp\{-0.5\cdot u^2\}|_{u=t}^\infty\right] = \frac{\phi(t)}{1-\Phi(t)} \\
\mathbb{E}(L^2|Y=1) &= \mathbb{E}(L^2|L>t) = \frac{\int_t^\infty u^2\phi(u)du}{1-\Phi(t)} = \frac{1}{K}\left[-\phi(u)u|_{u=t}^\infty + \int_t^\infty \phi(u)du\right] = \\
&= \frac{t\phi(t) + (1-\Phi(t))}{1-\Phi(t)} = 1 + \frac{t\phi(t)}{1-\Phi(t)}. \\
Cov(L,Y) &= \mathbb{E}(YL) - \mathbb{E}(Y)\mathbb{E}(L) = \mathbb{P}(Y=1)\mathbb{E}(L|Y=1) = K\frac{\phi(t)}{K} = \phi(t). \\
Cor(L,Y) &= \frac{\phi(Z_{1-K})}{\sqrt{K(1-K)}}.
\end{aligned}$$

In twins studies, we can still use MZ and DZ twins to estimate the *observed-level heritability* using the same formula:

$$\hat{H}_{ad,obs} = 2(r(MZ) - r(DZ)).$$

The question is, what is the connection between $H_{ad,obs}$ and $\sigma_g^2 = H_{ad,liability}$?
This turns out to be a difficult problem, which we will get back to later. For now, let's deal with an easier problem.

Assume we observe G, and therefore can estimate $Cor(Y,G)$ — can we use this to estimate $\sigma_g^2 = Cov(L,G)$? The famous result which addresses that is by Dempster and Lerner (1950), for which we need to derive some additional formulas:

$$\mathbb{E}(G|L>t) = \sigma_g^2\frac{\phi(t)}{K} \ \ \Rightarrow \ \ Cor(Y,G) = \sigma_g\frac{\phi(t)K}{K\sqrt{K(1-K)}} = \frac{\sigma_g\phi(t)}{\sqrt{K(1-K)}}.$$

Therefore, we can conclude:

$$H_l^2 = \sigma_g^2 = Cor^2(Y, G) \cdot \frac{K(1-K)}{\phi(t)^2}.$$

For rare diseases (K small), we have $\phi(Z_{1-K}) \approx K$, and therefore $H_l^2 >> Cor^2(Y, G) = H_{obs}^2$, in other words: the heritability is much bigger on the unobserved liability scale than on the observed scale.

Thus, a methodology which arises;

1. Somehow estimate "heritability"$= Cor^2(Y, G)$ on the observed scale
2. Perform the transformation to transform it to the liability scale

This approach was adopted by Lee et al. (2011), who used the LMM approach, meaning they assumed (for absolutely no good reason) the model (1) for the 0/1 phenotype:

$$\frac{Y}{\sqrt{K(1-K)}} | Z, X \dot{\sim} N(X\beta, G\sigma_g^2 + I_n(1-\sigma_g^2)),$$

then in that model, estimated $\widehat{H_{obs}^2} = \hat{\sigma}_g^2$, and applied the Depmster-Lerner correction:

$$\widehat{H_l^2} = \widehat{H_{obs}^2} \cdot \frac{K(1-K)}{\phi(t)^2}.$$

Surprisingly(?), this approach generally gives pretty good (unbiased) estimates of H_l^2 in simulations.

An equally unsubstantiated approach uses twins to estimate H_{obs}^2 and applies the same correction.

These are both fundamentally unsubstantiated because there is no reason to assume additivity between G and E on the observed scale, which is fundamental to the formulas we derived. A second (major?) concern for the Lee et al. method is that normality is impossible here by definition, since $Y \in \{0, 1\}$.

Dealing with case-control sampling

Assume now that in addition, we perform case-control sampling. That is now, we assume:

- $L = G + E$ as before
- Y is not taken randomly from the population (where $\mathbb{P}(Y = 1) = K$), but from a "rebalanced" population, where $\mathbb{P}(Y = 1) = P >> K$.

In this setting, many interesting and surprising changes occur. We will discuss them the next time.