Linear Mixed Models (a.k.a. Variance Component Models).

General background reading:
P. Sham pages 198-220, 261-267
K. Lange Chapter 8 (pages 139-165)

Acknowledgements: thanks to Eric for his slides on QTL association

Appropriate Study Design

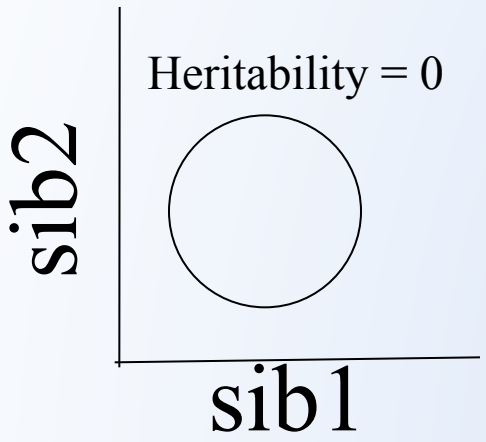
- All combinations of families and individuals can be used, from one very large family to many “unrelated” individuals.
- Quantitative traits that can be transformed to normality.
- When using families, random ascertainment or ascertainment through explicitly defined probands are best.

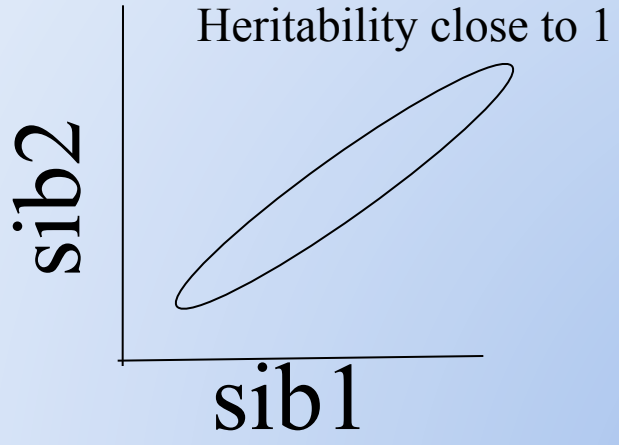
Outline

- Background:
  - polygenic model,
  - linear mixed model,
  - IBD
  - kinship
- Heritability
- Association

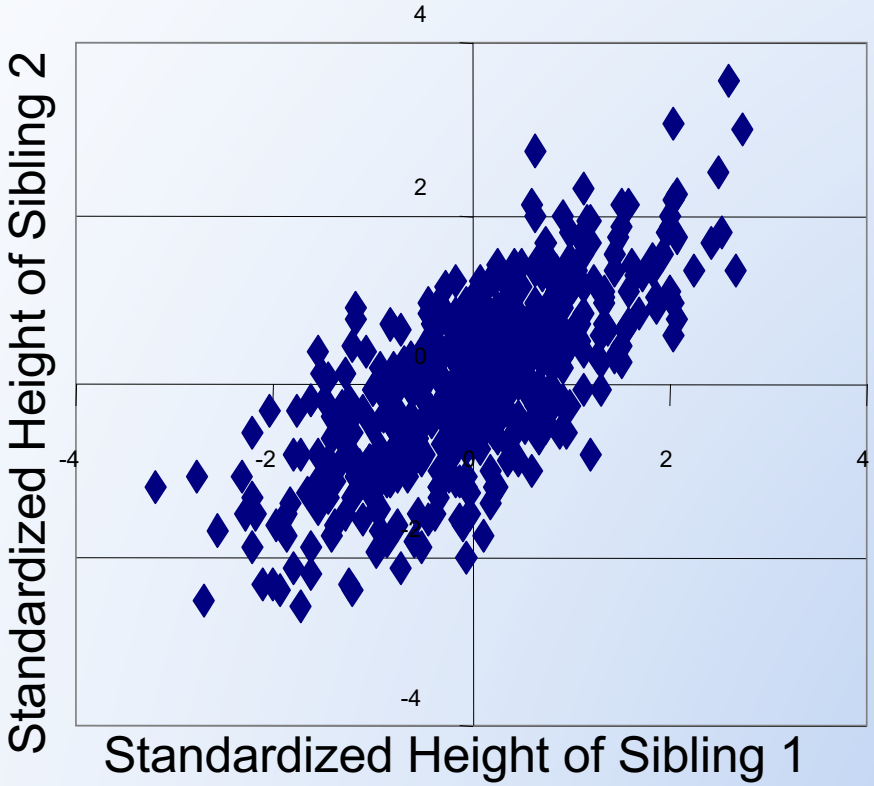
Fisher's Polygenic Inheritance Model Explains Genetic Effects on Quantitative Traits

- Trait values are determined by many genes of small additive effect each acting independently.
- Trait values in a population follow a normal distribution. Location and scale depend on the trait mean and variance.
- The bivariate distributions of relative pairs are 2-dimensional normal distributions – ellipsoidal level curves whose shape depends on the heritability of the trait.

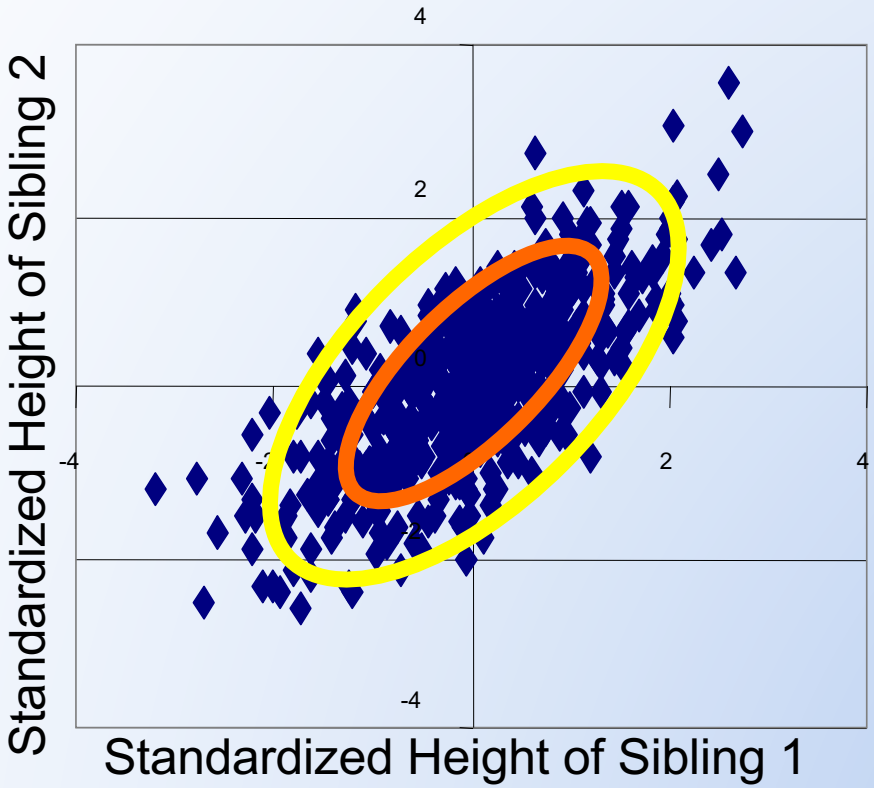




Bivariate Distribution Example: Sibling Height Values



Bivariate Distribution Example: Sibling Height Values



Linear Mixed Model and Family Data

- Simple example: $Y_i = \mu + \beta x_{1i} + A_i + e_i$ where μ is the population mean, x_{1i} is the number of minor alleles of snp1, e_i is a random effect that corresponds to individual variation and and A_i is a random effect that corresponds to familial correlated variation.
- So for two related individuals, i and j,

$$\text{var}(e_i) = \sigma_e^2,\ \text{cov}(e_i, e_j) = 0$$

$$\text{var}(A_i) = \sigma_A^2,\ \text{cov}(A_i, A_j) \neq 0$$

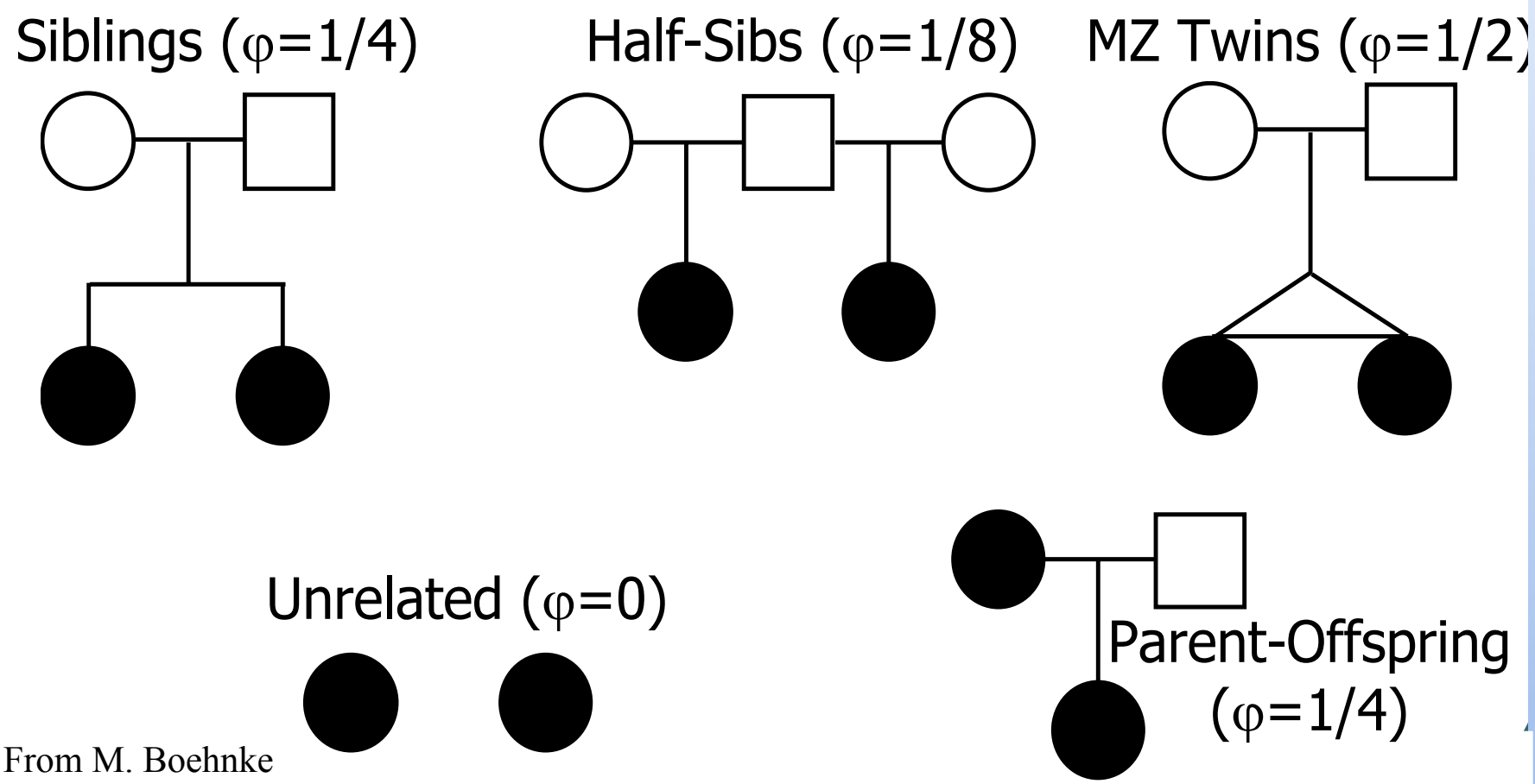
Estimating Covariance of Traits among Relatives

- As written we would need to estimate the covariation for each relative pair but that doesn't work.
- However, we can come up with a sensible and powerful model because we expect two closely related family member's values to be more similar than two randomly selected, unrelated individual's values.

Quantify the Extent of Relationship by Calculating Kinship Coefficients

- Identity by Descent: Two genes are IBD if they are copies of the same ancestral gene.
- The global kinship coefficient, Φ_{ij}, is the probability that two genes, one chosen randomly from individual i and one from j are identical by descent.
- We can use the pedigree structure to calculate Φ_{ij} (the theoretical global kinship) or use markers dispersed about the chromosomes (the estimated global kinship = genetic relationship matrix).
- The kinship coefficient captures degree of relatedness in a way that can be used to parsimoniously model the covariation in quantitative traits of two relatives.

Examples of Kinships Based on Pedigree Information:



From M. Boehnke

Returning to our Simple Linear Mixed Model Example:

- Simple example: $Y_i = \mu + \beta x_{1i} + A_i + e_i$ where μ is the population mean, x_{1i} is the number of minor alleles of snp1, e_i is a random effect that corresponds to individual variation and and A_i is a random effect that corresponds to familial correlated variation.
- Now we have that for related individuals, i and j,

$$\text{var}(e_i) = \sigma_e^2,\ \text{cov}(e_i, e_j) = 0$$

$$\text{var}(A_i) = \sigma_A^2,\ \text{cov}(A_i, A_j) = 2\Phi_{ij}\sigma_A^2$$

$$\text{var}(Y_i) = \sigma_A^2 + \sigma_e^2, \text{cov}(Y_i, Y_j) = 2\Phi_{ij}\sigma_A^2$$

“Modern” Measures of the Global Kinship

- Use GWAS data to estimate the global kinship (e.g. Mendel ped gwas option) and then use this global kinship to estimate heritability. Examples of two ways to estimate Φ_{ij} from GWAS data: GRM and methods of moments (MoM).
- Why estimate global kinships using GWAS data?
- To determine to what degree the SNPS in the GWAS capture the variability in the trait (SNP heritability). Using distantly related relatives, “unrelateds,” reduces the confounding of genetics and common environment.
- Pedigree structures can be inaccurate.

Genetic Relationship Matrix (GRM)

- The idea: Two individuals are more related if they have the same alleles at many SNPs along their chromosomes.
- The implementation:
  - Let SNP k have MAF p_k. Let X_{ik} denote the number of minor alleles at SNP k for person i so X_{ik} can be 0, 1, or 2. The expected value of X_{ik} is $2p_k$ and the variance of X_{ik} is $2p_k(1-p_k)$.
  - The contribution SNP k makes to the correlation of individual i and j's values is: $$\frac{(X_{ki}-2p_k)(X_{kj}-2p_k)}{2p_k(1-p_k)}$$
  - Averaging over all the S SNPs $$\Phi^*_{ij}=\frac{1}{2S}\sum_{k=1}^{S}\frac{(X_{ik}-2p_k)(X_{jk}-2p_k)}{2p_k(1-p_k)}$$
  - Have a Φ^*_{ij} for each set of individuals (including when i = j). Store in a matrix Φ^*.
  - Problem: rare variants can distort this estimate.

The Method of Moment Estimates

- Rare variants have less influence.

$$\tilde{\Phi}_{ij} = \frac{e_{ij} - \sum_{k=1}^{S}\left[p_k^2 + \left(1 - p_k^2\right)\right]}{S - \sum_{k=1}^{S}\left[p_k^2 + \left(1 - p_k^2\right)\right]}$$

$$e_{ij} = \frac{1}{4}\sum_{k=1}^{S}\left[X_{ik}X_{jk} + \left(2 - X_{ik}\right)\left(2 - X_{jk}\right)\right]$$

- The GRM centers and scales each genotype (which up-weights rare variants) whereas the MoM centers and scales the aggregate.

Using the Linear Mixed Model in Genetics: Heritability Estimation

- Heritability h^2 is the proportion of the variation in the trait that is attributable to genetics. Under the simplest assumptions,

$$h^2 = \sigma^2_A / (\sigma^2_A + \sigma^2_e)$$

- Heritability is highly dependent on the conditions under which it is calculated making comparing heritability estimates calculated in different ways, in different populations, almost impossible.
- Heritability estimates can be improved by allowing for other sources of covariation.

Why Estimate Heritability?

REASONS:

- To address the question: Is the trait “genetic”?
- With a study design, is there enough power to make it feasible that one can find genes that influence the trait values?
- How much do the putative predictors change the heritability of the trait?

Should we still use the Old Way of Calculating Heritability (Conditional on the Pedigree Structure)?

- Provides an estimate of heritability before genotyping or sequencing.
- Provides another bound for the true heritability: GWAS based methods provide an underestimate, pedigree based method provides an overestimate.

Gene Mapping using Linear Mixed Models

The Principles

- At various locations along the genome:
  - Add the effects of a gene region (called a quantitative trait locus or QTL) to the linear mixed model and test whether it improves the model fit.
  - The locations that most improve the model fit are the most likely QTLs.

Association with LMM

- Association is simply a statistical statement about the co-occurrence of alleles.
- Example: Allele A_1 from a marker is associated with allele t from a gene increasing trait values, if people with t (and thus high trait values) also have A_1 more often than would be predicted from the individual frequencies of t and A_1 in the population.

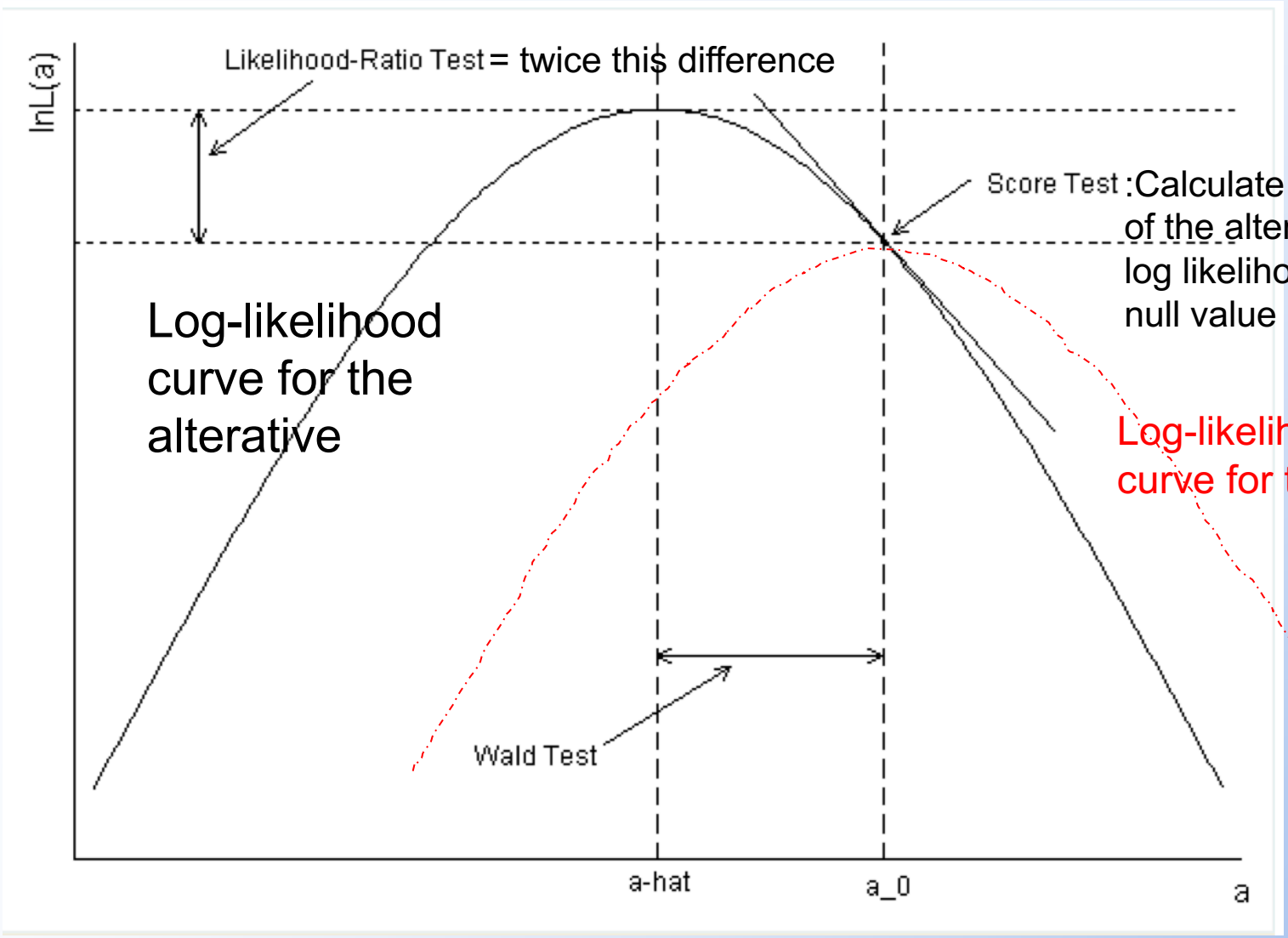
We treat genotypes as Fixed Effects in the LMM

- Suggested by Boerwinkle *et al.* (1986) and others and referred to as the "Measured Genotype Approach."
- This method uses the allele counts at a marker as covariates, i.e., (additional) predictors, for each individual's trait value. X now represents genotypes as well as traditional predictors.
- $$Y_i = \mu + \beta^T X_i + A_i + e_i$$
- More than one genotype, gene x gene interactions and gene x environment interactions can be included in $\beta^T X_i$

Inference

- Just like the linear model – we can test hypotheses using an LRT, a Wald test, or a score test.

The Three Methods of Inference



Adapted From: Fox, J. (1997) Applied regression analysis, linear models, and related methods. Thousand Oaks, CA: Sage Publications.

HELP!!! - Need Fast, Low-Memory use Methods

- What if have 1M+ markers and 100K+ individuals?
- Using score tests isn't sufficient to make computation feasible.
- Traditional methods (including score tests) require inversion of a large matrix. Ω is n x n where n is the number of individuals.
- $\Omega = 2\sigma_A^2\Phi^* + \sigma_E^2\mathrm{I}$.
- Decomposing the matrix can make it easier to invert but still requires too much memory.
- Finding solutions are very active research areas for computational statisticians.
- Example: Very clever solutions employed in BOLT-LMM (P-R Loh et al. 2015 Nature Gen.) which uses Monte-Carlo sampling to estimate the effects in order to circumvent

General Conclusions:

- Variance component models are useful because they are flexible.
  - Arbitrary family structures can be used.
  - Multiple traits can be analyzed together.
  - They can be extended to model more complicated situations than were shown today.
  - Can estimate heritability, test for linkage or test for association.
- Variance components (LMM) underlie a number of methods for gene mapping for common and rare variants.

Final Comments on LMMs:

- Like all models, LMMs are cartoons of reality, making simplifying assumptions that need to be understood and checked.
- There are a number of LMM programs specifically designed for genetic data, some easily handle pedigrees, some not.
  - Examples: EMMAX, FamSKAT, Fast-LMM, GCTA, GEMMA, Mendel, MONSTER, SOLAR.
- There are Alternative Methods of Testing for Linkage or Association for Quantitative Traits in Families.
  - e.g. Family Based Association Testing; Mendel's Gamete Competition; PseudoMarker
- Active area of research: develop accurate, powerful and fast rare variant-quantitative traits association tests in families.

Inversion and Decomposition

collects the corresponding trait means into a vector $\boldsymbol{v}$ and the corresponding covariances into a matrix $\boldsymbol{\Omega}$ and represents the loglikelihood of a pedigree as

$$L = -\frac{1}{2}\ln\det\boldsymbol{\Omega} - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{v})^t\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{v}), \quad (1)$$

where det denotes the determinant function and the covariance matrix is typically parametrized as

$$\boldsymbol{\Omega} = 2\sigma_a^2\boldsymbol{\Phi} + \sigma_d^2\boldsymbol{\Delta}_7 + \sigma_h^2\boldsymbol{H} + \sigma_e^2\boldsymbol{I}. \quad (2)$$

$$S(\boldsymbol{\theta}) = dL(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}\nabla L(\boldsymbol{\theta}) \approx dL(\boldsymbol{\theta})[-d^2L(\boldsymbol{\theta})]^{-1}\nabla L(\boldsymbol{\theta})$$

From: Fast Genome-Wide QTL Association Mapping on Pedigree and Population Data, H. Zhou, J. Blangero,T. D. Dyer K. K. Chan, K. Lange E. M. Sobel (2017) Genetic Epidemiology 41:174-186

One must compute the quantities

$$\sum_{i=1}^{n} \nabla_{\beta} L_i(\boldsymbol{\theta}) = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i \\ \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i \end{pmatrix}$$

$$\sum_{i=1}^{n} \mathrm{E}[-d_{\beta}^2 L_i(\boldsymbol{\theta})] = \begin{pmatrix} \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i & \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i \\ \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i & \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i \end{pmatrix}.$$

At the maximum likelihood estimates under the null model, the partial score vector $\sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i$ vanishes. Hence, the score statistic for testing a SNP can be expressed as

$$S = \boldsymbol{R}^t \left[Q - \boldsymbol{W}^t \left(\sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{N}_i \right)^{-1} \boldsymbol{W} \right]^{-1} \boldsymbol{R},$$

where

$$Q = \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i, \quad \boldsymbol{R} = \sum_{i=1}^{n} \boldsymbol{a}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{r}_i,$$

$$\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{N}_i^t \boldsymbol{\Omega}_i^{-1} \boldsymbol{a}_i.$$