Variance Component Estimation with Pedigrees

Partitioning Variance Review

- We previously described a partitioning the total variance of a quantitative trait (σ^2_Y) in an outbred population into variance due to additive genetic effects (σ^2_A), dominance (σ^2_D), epistastis (e.g., for the two locus case we have σ^2_{AA}, σ^2_{AD}, σ^2_{DD}), and environment or residual (σ^2_E).
- The sum of all those components, aside from environmental variance, is generally called the "genetic variance" (σ^2_G)
- We have used the variance decomposition to assess heritability of a trait.

Heritability Review

- Heritability in the broad and narrow sense provide two measures of the importance of genetic factors to a trait,
- The broad-sense heritability, $\frac{\sigma_G^2}{\sigma_Y^2}$, also called the coefficient of genetic determination, expressing the ex- tent to which the phenotype is explained by genotype in a particular population.
- The narrow-sense heritability, $\frac{\sigma_A^2}{\sigma_Y^2}$, typically referred to simply as the "heritability", measures the degree to which, in the given population, the offspring phenotype is explained by the parental phenotypes.

Limitations of other Methods

- In practice, many mapping and heritability studies in outbred populations consider only environmental, additive, and dominance variance or only environmental and additive variance.
- As we have shown, additive variance components can be estimated using the covariances of the trait values for heritability estimations with particular types of relatives that do not have dominance effects.
- Commonly used methods for estimation of components of variance of quantitative traits include parent-offspring regression, correlation and analysis of variance (ANOVA) for trait values of sib and/or half-sib families, as well as MZ and DZ twins
- Such approaches are particularly suited for animal breeding situations but are not ideal for study populations where family designs are often unbalanced and for where the available information for different relationship types may vary.

Variance Components Estimation with Pedigrees

- A more flexible alternative to these types of methods for estimation of variance components is maximum likelihood (ML) (or restricted maximum likelihood [REML]) variance-component estimation.
- In the last 20 years, this methodology has gained significant interest for both variance components estimation as well as for the mapping of quantitative traits.
- For the price of assuming a particular distribution, generally multivariate normal, for the phenotype, the method allows one to partition the variance into its basic genetic and non-genetic components, using a sample of individuals of known relationship.

Variance Components Estimation with Pedigrees

- The ML/REML analysis can use information from all types of relative pairs in the data, without concern for balanced numbers of families of restricted relationship types
- As a result, the information inherently available in the data is used more efficiently with ML/REML than in methods like ANOVA, regression, and relative-pair trait correlation methods.
- We will focus on estimating variance component with pedigrees using ML/REML

ACDE Trait Model: fixed effects and random effects

- The ACDE model decomposes the total variance of the phenotype into four components due to:
  - Additive genetic
  - Common shared environment
  - Dominance genetic effects
  - unique Environment effects.
- Hence, the term "ACDE" model

ACDE Trait Model: fixed effects and random effects

- Consider a study that contains measurements for a quantitative trait of interest for individuals sampled from F families.
- Assume the quantitative trait follows the following ACDE model such that for individual j in family i we have the following:

$$y_{ij} = \mathbf{x}_{ij}^T \beta + A_{ij} + D_{ij} + C_i + \varepsilon_{ij}$$

where

- y_{ij} is the trait value for the jth individual from family i
- $\mathbf{x}_{ij}$ is a vector of covariate values for the individual (such as age, sex, etc.), and β is a vector of fixed effects
- $A_{ij} \sim N(0, \sigma_A^2)$ is the additive genetic random effect , $D_{ij} \sim N(0, \sigma_D^2)$ is the dominance genetic random effect, $C_i \sim N(0, \sigma_C^2)$ is the family effect for family i, and $\varepsilon_{ij} \sim N(0, \sigma_E^2)$ is the residual (unique environment) for individual j from family i.

ACDE Trait Model

$$y_{ij} = \mathbf{x}_{ij}^T \beta + A_{ij} + D_{ij} + C_i + \varepsilon_{ij}$$

- Note that A_{ij} is the sum of the additive effects across all loci that influence the trait for individual j from family i. Similarly D_{ij} is the sum of all dominance effects across all loci that influence the trait for the individual.
- What is the the expected value of y_{ij}, i.e., $\mu_{ij} = E(y_{ij})$?
- What is the variance of y_{ij}, i.e, $\sigma_y^2 = E(y_{ij} - \mu_{ij})^2$?

ACDE Trait Model Mean and Variance

$$y_{ij} = \mathbf{x}_{ij}^T \beta + A_{ij} + D_{ij} + C_i + \varepsilon_{ij}$$

- $\mu_{ij} = \mathbf{x}_{ij}^T \beta$ under the ACDE model assumptions.
- The four 'error" components are assumed to be mutually independent so that the total variance of the trait is the sum of the variance components

$$\sigma_y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_C^2 + \sigma_E^2$$

- So the distribution of y_{ij} is normal with mean μ_{ij} and variance σ_y^2, i.e.,

$$y_{ij} \sim N(\mu_{ij}, \sigma_y^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left\{ -\frac{(y_{ij} - \mu_{ij})^2}{2\sigma_y^2} \right\}$$

ACDE Trait Model Covariances

- We have the means and variances for each of the individuals in the study.
- Individuals in the study are related, so we also need to model the covariances between relatives.
- Now consider two outbred individuals j and k. What are the covariances for the ACDE components for these two individuals for each of the four variance components in the ACDE trait model?

ACDE Trait Model Covariances

- Let's consider the covariance matrix of the shared family component C.
- Let $\mathbf{C} = (C_j, C_k)^T$ be the random family and/or shared environment effects vector for individuals j and k.
- If individuals j and k are from the same family, then covariance matrix of the shared family components for these two individuals is

$$Cov(\mathbf{C}) = \sigma_c^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \tag{1}$$

  If individuals j and k are from different families, then covariance matrix of the shared family effect is

$$Cov(\mathbf{C}) = \sigma_c^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{2}$$

- What is the covariance matrix for the additive effects for individuals j and k?

ACDE Trait Model

- Let $\mathbf{A} = (A_j, A_k)^T$ be the random additive effects vector for individuals j and k. The covariance matrix of the additive effects is

$$Cov(\mathbf{A}) = \sigma_A^2 \begin{pmatrix} 1 & 2\theta_{jk} \\ 2\theta_{jk} & 1 \end{pmatrix} \tag{3}$$

  where θ_{jk} is the kinship coefficient for individuals j and k
- In general, if there are n individuals in a study, then the covariance matrix of the additive effects is

$$Cov(\mathbf{A}) = 2\mathbf{\Theta}\sigma_A^2 = \sigma_A^2 \begin{pmatrix} 1 & 2\theta_{12} & \dots & 2\theta_{1n} \\ 2\theta_{12} & 1 & \dots & 2\theta_{2n} \\ \vdots & \dots & \dots & \vdots \\ 2\theta_{1n} & 2\theta_{2n} & \dots & 1 \end{pmatrix}, \tag{4}$$

ACDE Trait Model

- The dominance effects covariance matrix with n individuals in a study is

$$Cov(\mathbf{D}) = \sigma_D^2 \boldsymbol{\Delta_7} = \sigma_D^2 \begin{pmatrix} 1 & \Delta_7^{(12)} & \dots & \Delta_7^{(1n)} \\ \Delta_7^{(12)} & 1 & \dots & \Delta_7^{(2n)} \\ \vdots & \dots & \dots & \vdots \\ \Delta_7^{(1n)} & \Delta_7^{(2n)} & \dots & 1 \end{pmatrix}, \quad (5)$$

where $\Delta_7^{(kj)}$ is Jacquard's identity state 7 for individuals k and j, which is the probability that j and k share two alleles IBD.

ACDE Trait Model

- If there are n individuals in a study, then the covariance matrix of the unique environmental effects (or residuals) is $\sigma_E^2 \mathbf{I}$ where $\mathbf{I}$ is the identity matrix:

$$\sigma_E^2 \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \tag{6}$$

ACDE Trait Model

- So for a vector of trait values for n individuals from F families is $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, we have that the $Cov(\mathbf{y}) = \mathbf{\Omega}$ where

$$\mathbf{\Omega} = 2\mathbf{\Theta}\sigma_A^2 + \sigma_D^2\mathbf{\Delta_7} + \sigma_C^2\mathbf{\Phi_C} + \sigma_E^2\mathbf{I}$$

  where $\mathbf{\Phi_C}$ has (j,k)th entry equal to 1 if j and k are from the same family, and 0 otherwise.

Multivariate Normal Distribution of Trait

- When considering the distribution of all of the trait values of the n sample individuals, $\mathbf{y}$ is often molded as a multivariate normal distribution and the log likelihood function for $\mathbf{y}$ is

$$l(\mathbf{y}|\mathbf{V},\beta) = -\frac{n}{2}\ \text{ln}\ 2\pi - \frac{1}{2}\ \text{ln}\ |\mathbf{\Omega}|$$

$$-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{\Omega}^{-1}(\mathbf{y}-\mathbf{X}\beta)$$

  where $\mathbf{V} = [\sigma^2_A, \sigma^2_D, \sigma^2_C, \sigma^2_E]$, is a vector of variance component parameters, β is a vector of fixed effects, $\mathbf{\Omega}$ is the covariance matrix $Cov(\mathbf{y})$ and is a function of $\mathbf{V}$ and $|\mathbf{\Omega}|$ is the determinant of $\mathbf{\Omega}$.

- Obtaining maximum likelihood estimates for the fixed effects and the variance components is not trivial!

Maximum Likelihood (ML) Estimation

- Maximum-likelihood estimation of variance components does not, in general, take into account the loss in degrees of freedom that results from estimation of the fixed effects, and, as a result, ML estimators tend to be biased.
- In particular, estimates of the variance components are generally downwardly biased, with the bias increasing as the number of fixed effects increases.
- If the sample size is small, this bias can become quite substantial.

Restricted (or Residual) Maximum Likelihood (REML) Estimation

- An alternative to ML estimation is REML estimation (Searle et al. 1992), which essentially maximizes only that portion of the likelihood that depends on the variance components and not on the fixed effects.
- Hence, bias of this type is re- moved by REML in a manner analogous to the removal of bias in a variance estimator by dividing by the degrees of freedom rather than by dividing by the sample size.
- REML, instead of using the data vector y directly, is based on a linear transformation of the data, where the transformation is chosen in such a way that the fixed effects are eliminated from the model. Given the mixed

REML

- Consider the mixed model

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{g} + \varepsilon$$

  where $\mathbf{g} = A + D$ is the polygenic effect of the trait containing both the additive and dominance effects of all of the loci that influence the trait. We could easily include a shared environmental effect (C) for families as well), but assume, without loss of generality, that the family effect (C) is negligible for this particular trait.
- Consider a matrix $\mathbf{K}$ such that $\mathbf{KX} = 0$. Applying this transformation to the above mixed model equation results in what equation?

REML

$$\mathbf{Ky} = \mathbf{Kg} + \mathbf{K}\varepsilon$$

- If $\mathbf{y}$ is multivariate normal with mean $\mathbf{X}\beta$ and variance $\mathbf{\Omega}$, then $\mathbf{Ky}$ is also multivariate normal.
- What is the mean and variance of $\mathbf{Ky}$?

REML

- If matrix $\mathbf{K}$ has the property that $\mathbf{KX} = 0$. Applying this transformation to the above mixed model equation results in

$$\mathbf{Ky} = \mathbf{Kg} + \mathbf{K}\varepsilon$$

- The mean of $\mathbf{Ky}$ is $\mathbf{0}$ and variance is $\mathbf{K\Omega K^T}$
- REML then proceeds as ML, but with the transformed data vector and covariance matrix
- Although REML requires one to compute the matrix $\mathbf{K}$, this matrix can be formulated only in terms of $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{\Omega}$.