7. Polygenic Models

§7.1. The Variance Component Model

The nature of polygenic models

A polygenic model concerns with quantitative traits which are determined by a large number of loci acting independently and additively:

\[ X_i = \sum_k X_{ik}^k. \]

\( X_i \): the total genetic trait value of person \( i \);
\( X_{ik}^k \): the contribution of locus \( k \) to \( X_i \). \( X_{ik}^k \)'s are independent.

If the number of \( X_{ik}^k \) is large and the \( X_{ik}^k \)'s are comparable, \( X_i \) can be assumed to follow a normal distribution by the central limit theorem.
Let $\mathbf{X} = (X_1, \ldots, X_n)^T$ be the vector of trait values for a pedigree with $n$ persons.

$$\text{Cov}(X_i, X_j) = \text{Cov} \left( \sum_k X^k_i, \sum_k X^k_j \right)$$

$$= \sum_k \text{Cov}(X^k_i, X^k_j)$$

$$= \sum_k \left[ 2\Phi_{ij} \sigma^2_{ka} + \Delta_{ij} \sigma^2_{kd} \right]$$

$$= 2\Phi_{ij} \sum_k \sigma^2_{ka} + \Delta_{ij} \sum_k \sigma^2_{kd}$$

$$= 2\Phi_{ij} \sigma^2_a + \Delta_{ij} \sigma^2_d.$$

Let $\Sigma = \text{Var}(\mathbf{X})$. Assume $E \mathbf{X} = 0$. Then

$$\Sigma = 2\sigma^2_a \Phi + \sigma^2_d \Delta \gamma, \quad \mathbf{X} \sim N(0, \Sigma).$$

- **Covariates and other non-genetic variance components**

  Environmental effects and effects of other covariates can be incorporated into the model.
through a random variable $Z$ which is independent with the genetic trait variable $X$:

$$Y_i = X_i + Z_i.$$ 

Let $Z = (Z_1, \ldots, Z_n)^T$. It can be assumed that

$$Z \sim N(\nu, \Upsilon).$$

It can be formulated that

$$\nu = A\mu,$$

$$\Upsilon = \sigma_h^2 H + \sigma_e^2 I,$$

where $A$ is an $m \times p$ design matrix consisting of values of covariates of the $n$ individuals, $H = (h_{ij})$ is a household indicator matrix with

$$h_{ij} = \begin{cases} 1 & i \text{ and } j \text{ are in the same household,} \\ 0 & \text{otherwise,} \end{cases}$$

and $\sigma_e^2$ accounts for the variance caused by independent random effects including measuring errors.
The general form of variance component model

Combining the assumptions on $X$ and $Z$,

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Omega}),$$

where

$$\boldsymbol{\Omega} = \Sigma_X + \Upsilon = 2\sigma_a^2 \Phi + \sigma_d^2 \Delta_7 + \sigma_h^2 \mathit{H} + \sigma_e^2 \mathit{I}.$$

In general, assume that

$$\boldsymbol{\Omega} = \sum_{k=1}^{r} \sigma_k^2 \mathit{\Gamma}_k,$$

where $\mathit{\Gamma}_k$’s are known nonnegative definite and $\mathit{\Gamma}_r = \mathit{I}$.

Let $\gamma = (\boldsymbol{\mu}^T, \boldsymbol{\sigma}^2)^T$ where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_r^2)^T$.

The log likelihood function is given by

$$L(\gamma) = -\frac{1}{2} \ln \det \boldsymbol{\Omega} - \frac{1}{2} (\mathbf{y} - A\boldsymbol{\mu})^T \boldsymbol{\Omega}^{-1} (\mathbf{y} - A\boldsymbol{\mu}).$$
§7.2. The Estimation of Variance Component Model

• Some preliminaries

Denote

\[ B = (b_{ij}), \quad \frac{\partial}{\partial \theta} B = \left( \frac{\partial b_{ij}}{\partial \theta} \right), \]

\[ \frac{\partial b}{\partial \mu} = \left( \frac{\partial b}{\partial \mu_1}, \ldots, \frac{\partial b}{\partial \mu_m} \right)^T. \]

1. The determinant of \( B \) can be computed by

\[ \det B = \sum_j b_{ij} B_{ij}, \]

where \( B_{ij} \) is the cofactor corresponding to \( b_{ij} \).

2. The inverse of \( B \) can be computed by

\[ B^{-1} = \frac{1}{\det B} (c_{ij}), \]

where \( c_{ij} = B_{ji} \).
3. The trace of $B$ is defined as $\text{tr } B = \sum_i b_{ii}$. We have $\text{tr } BC = \text{tr } CB$.

4. Some formulas:

$$E[BY] = BE[Y],$$
$$EY^TBY = \text{tr } [B\text{Var}(Y)] + [EY]^T B [EY],$$
$$\frac{\partial}{\partial \theta} \text{tr } B = \text{tr } \left( \frac{\partial}{\partial \theta} B \right),$$
$$\frac{\partial}{\partial \theta} (BC) = \left( \frac{\partial}{\partial \theta} B \right) C + B \left( \frac{\partial}{\partial \theta} C \right),$$
$$\frac{\partial}{\partial \mu} a^T \mu = a,$$
$$\frac{\partial}{\partial \mu} \mu^T A \mu = 2A \mu,$$
$$\frac{\partial}{\partial \theta} B^{-1} = -B^{-1} \left( \frac{\partial}{\partial \theta} B \right) B^{-1},$$
$$\frac{\partial}{\partial \theta} \ln \det B = \text{tr } (B^{-1} \frac{\partial}{\partial \theta} B).$$
• The scoring method

Let

\[ U(\gamma) = \frac{\partial L}{\partial \gamma}, \quad J(\gamma) = -E \left[ \frac{\partial^2 L}{\partial \gamma \partial \gamma^T} \right] \]

The scoring method solves iteratively

\[ J(\gamma^{\text{OLD}})[\gamma^{\text{NEW}} - \gamma^{\text{OLD}}] = U(\gamma^{\text{OLD}}). \]

Calculation of the derivatives:

\[
\frac{\partial L}{\partial \mu} = -\frac{1}{2} \frac{\partial}{\partial \mu}(y - A\mu)^T \Omega^{-1} (y - A\mu) \\
= A^T \Omega^{-1} (y - A\mu),
\]

\[
\frac{\partial L}{\partial \sigma_k^2} = -\frac{1}{2} \frac{\partial}{\partial \sigma_k^2} \ln \det \Omega - \frac{1}{2} (y - A\mu)^T \frac{\partial}{\partial \sigma_k^2} \Omega^{-1} (y - A\mu) \\
= -\frac{1}{2} \text{tr} (\Omega^{-1} \Gamma_k) + \frac{1}{2} (y - A\mu)^T \Omega^{-1} \Gamma_k \Omega^{-1} (y - A\mu).
\]
\[
\begin{align*}
\frac{\partial^2 L}{\partial \mu \partial \mu^T} &= -A^T \Omega^{-1} A, \\
\frac{\partial^2 L}{\partial \sigma^2_k \partial \mu} &= -A^T \Omega^{-1} \Gamma_k \Omega^{-1} (y - A\mu) \\
\frac{\partial^2 L}{\partial \sigma^2_k \partial \sigma^2_l} &= \frac{1}{2} \text{tr} \left( \Omega^{-1} \Gamma_k \Omega^{-1} \Gamma_l \right) \\
&\quad - (y - A\mu)^T \Omega^{-1} \Gamma_k \Omega^{-1} \Gamma_l \Omega^{-1} (y - A\mu).
\end{align*}
\]

Denote
\[
U_{\sigma^2}(\gamma) = \frac{\partial L}{\partial \sigma^2}, \quad J_{\sigma^2}(\gamma) = \left( \frac{\partial^2 L}{\partial \sigma^2_k \partial \sigma^2_l} \right).
\]

- **The algorithm**

Set starting values: \( \gamma^{\text{old}} = (\mu^{\text{old}T}, \sigma^{2\text{old}T})^T \).

(a) Compute
\[
\begin{align*}
\mu^{\text{new}} &= [A^T \Omega^{-1} (\gamma^{\text{old}}) A]^{-1} A^T \Omega^{-1} (\gamma^{\text{old}}) y, \\
\sigma^{2\text{new}} &= [J_{\sigma^2}(\gamma^{\text{old}})]^{-1} U_{\sigma^2}(\gamma^{\text{old}}).
\end{align*}
\]

(b) Check convergence. If converged, stop, otherwise, let
\[
\gamma^{\text{old}} = (\mu^{\text{new}T}, \sigma^{2\text{new}T})^T,
\]

go to (a).
§7.3. The Variance Component Model for Multivariate Traits

Let \( \mathbf{X} = (X_1, \ldots, X_n)^T \), \( \mathbf{Y} = (Y_1, \ldots, Y_n)^T \) be vectors of two genetic trait values of \( n \) members of a non-inbred pedigree. Suppose \( \mathbf{X} \) and \( \mathbf{Y} \) are determined by the same loci. \( \mathbf{X} \) and \( \mathbf{Y} \) have the following variance and covariance structure:

\[
\begin{align*}
\text{Var}(\mathbf{X}) &= 2\sigma_{ax}^2 \Phi + \sigma_{dx}^2 \Delta_7, \\
\text{Var}(\mathbf{Y}) &= 2\sigma_{ay}^2 \Phi + \sigma_{dy}^2 \Delta_7, \\
\text{Cov}(\mathbf{X}, \mathbf{Y}) &= 2\sigma_{axy} \Phi + \sigma_{dxy} \Delta_7,
\end{align*}
\]

where

\[
\begin{align*}
\sigma_{axy} &= \frac{1}{2}(\sigma_{az}^2 - \sigma_{ax}^2 - \sigma_{ay}^2), \\
\sigma_{dxy} &= \frac{1}{2}(\sigma_{dz}^2 - \sigma_{dx}^2 - \sigma_{dy}^2).
\end{align*}
\]

Here \( \sigma_{az}^2 \) and \( \sigma_{dz}^2 \) are the additive and dominant genetic variances of the trait \( X + Y \).
It can be written that

$$\text{Var} \left( \begin{pmatrix} X \\ Y \end{pmatrix} \right) = 2 \begin{pmatrix} \sigma_{ax}^2 & \sigma_{axy} \\ \sigma_{axy} & \sigma_{ay}^2 \end{pmatrix} \otimes \Phi + \begin{pmatrix} \sigma_{dx}^2 & \sigma_{dxy} \\ \sigma_{dxy} & \sigma_{dy}^2 \end{pmatrix} \otimes \Delta_7.$$

Environmental and other effects can be incorporated similarly. For instance, the variance components of household effect and random error effect can be specified as

$$\begin{pmatrix} \sigma_{hx}^2 & \sigma_{hxy} \\ \sigma_{hxy} & \sigma_{hy}^2 \end{pmatrix} \otimes H, \quad \begin{pmatrix} \sigma_{ex}^2 & \sigma_{exy} \\ \sigma_{exy} & \sigma_{ey}^2 \end{pmatrix} \otimes I.$$

If still let \( \mathbf{X} \) and \( \mathbf{Y} \) denote the vectors with other effects incorporated, then, in general, \( (\mathbf{X}^T, \mathbf{Y}^T)^T \) can be modeled as a multivariate normal vector with mean vector

$$\boldsymbol{\nu} = \begin{pmatrix} A \\ B \end{pmatrix} \boldsymbol{\mu},$$

10
and variance covariance matrix

$$\Omega = \sum_{k=1}^{r} \Sigma_k \otimes \Gamma_k,$$

where $\Sigma_k$’s are unknown $2 \times 2$ variance matrices and $\Gamma_k$’s are $n \times n$ non-negative definite matrices with $\Gamma_r = I$.

The estimation of the model can be done in exactly the same way as in the one-trait case except the increased complexity.

§7.4. Applications of the Variance Component Model

- Application to Gc measure genotype data

The trait: Plasma concentration of the Gc protein.

Covariates: Gc genotypes (1/1, 1/2, 2/2),
Gender (male, female), Age.

Data:
Measurements of the above variables on 31 identical twins, 13 fraternal twins, and 45 unrelated individuals.

Model:
y: vector of plasma concentrations in Gc protein.
Let $x_{li} = 1$ if the $i$th individual has the Gc genotype with $l$ 2-alleles, otherwise, $x_{li} = 0$, $l = 1, 2$.

$x_l = (x_{l1}, \ldots, x_{ln})^T$, $l = 1, 2$.

$g$: gender vector; $a$: age vector.

$A = (1, x_1, x_2, g, a)$.

$\mu = (\beta_0, \beta_1, \beta_2, \beta_g, \beta_a)^T$.

$\nu = A\mu$. 
\[ \Phi = \begin{pmatrix} \Phi_I & 0 & 0 \\ 0 & \Phi_F & 0 \\ 0 & 0 & \frac{1}{2} I \end{pmatrix}, \]

\[ \Delta_7 = \begin{pmatrix} \Delta_7I & 0 & 0 \\ 0 & \Delta_7F & 0 \\ 0 & 0 & I \end{pmatrix}. \]

\[ \Phi_I = \text{Diag}(\begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}), \]

\[ \Phi_F = \text{Diag}(\begin{pmatrix} 1/2 & 1/4 \\ 1/4 & 1/2 \end{pmatrix}). \]

\[ \Delta_7I = \text{Diag}(\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}), \]

\[ \Delta_7F = \text{Diag}(\begin{pmatrix} 1 & 1/4 \\ 1/4 & 1 \end{pmatrix}). \]

\[ \Omega = 2\sigma_a^2 \Phi + \sigma_d^2 \Delta_7 + \sigma_e^2 I. \]

\[ \mathbf{y} \sim N(\nu, \Omega). \]
If \( i \) and \( j \) are identical twins, then \( \Omega_{ij} = \sigma_a^2 + \sigma_d^2 \). If \( i \) and \( j \) are fraternal twins, then \( \Omega_{ij} = \sigma_a^2/2 + \sigma_d^2/4 \).

Reparameterization:

\[
\sigma_T^2 = \sigma_a^2 + \sigma_d^2 + \sigma_e^2,
\]

\[
\rho_I = (\sigma_a^2 + \sigma_d^2) / \sigma_T^2,
\]

\[
\rho_F = (\sigma_a^2/2 + \sigma_d^2/4) / \sigma_T^2.
\]

\[
\Omega = \sigma_T^2 \begin{pmatrix}
\Omega_I & 0 & 0 \\
0 & \Omega_F & 0 \\
0 & 0 & I
\end{pmatrix},
\]

\[
\Omega_I = \text{Diag}(\begin{pmatrix} 1 & \rho_I \\ \rho_I & 1 \end{pmatrix}),
\]

\[
\Omega_F = \text{Diag}(\begin{pmatrix} 1 & \rho_F \\ \rho_F & 1 \end{pmatrix}).
\]

Note: \( \sigma_d^2 = 0 \iff \rho_F = \rho_I/2 \).
Hypothesis testing:

<table>
<thead>
<tr>
<th>Model</th>
<th>Loglikelihood</th>
<th>No. of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full model</td>
<td>−217.610</td>
<td>8</td>
</tr>
<tr>
<td>$\rho_F = \rho_I/2$</td>
<td>−217.695</td>
<td>7</td>
</tr>
<tr>
<td>$\beta_1 = \beta_2 = 0$</td>
<td>−230.252</td>
<td>6</td>
</tr>
</tbody>
</table>

Conclusion:
1. No evidence against $\sigma^2_d = 0$;
2. Concentration of plasma in Gc protein is a highly heritable trait.
3. Gc locus has a major impact on the trait.

- Left and right hand finger ridge counts — A bivariate trait example

The trait: Left and right hand ridge counts.

Covariate: Gender
**Data:** Measurements of the above variables on 48 nuclear families and 18 pairs of identical twins.

**Model:**
\( \mathbf{y} = (\mathbf{y}_l^T, \mathbf{y}_r^T)^T \) : vector of left and right hand ridge counts.
\( \mathbf{x} \), vector of indicators of gender: 1 for female, 0 for male.
\( X = (1, \mathbf{x}), \quad A = I \otimes X \).
\( \beta = (\beta_{l0}, \beta_{l1}, \beta_{r0}, \beta_{r1})^T \).
If denote the mean values by \( \mu_{ml}, \mu_{fl}, \mu_{mr}, \mu_{fr} \), then \( \mu_{ml} = \beta_{l0}, \mu_{fl} = \beta_{l0} + \beta_{l1}, \mu_{mr} = \beta_{r0}, \mu_{fr} = \beta_{r0} + \beta_{r1} \).
Let \( \nu = A\beta \).
Assume \( \sigma^2_{dl} = \sigma^2_{dr} = 0 \).
\( \Omega = 2 \begin{pmatrix} \sigma^2_{al} & \sigma^2_{alr} \\ \sigma^2_{alr} & \sigma^2_{ar} \end{pmatrix} \otimes \Phi + \begin{pmatrix} \sigma^2_{el} & \sigma^2_{elr} \\ \sigma^2_{elr} & \sigma^2_{er} \end{pmatrix} \otimes I \).
\[ y \sim N(\nu, \Omega). \]

**MLE:**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_{ml} )</td>
<td>66.6 ( \pm ) 2.8</td>
<td>( \rho_{alr} )</td>
<td>0.992 ( \pm ) 0.008</td>
</tr>
<tr>
<td>( \mu_{fl} )</td>
<td>59.0 ( \pm ) 2.8</td>
<td>( \sigma_{ar}^2 )</td>
<td>657.5 ( \pm ) 69.2</td>
</tr>
<tr>
<td>( \mu_{mr} )</td>
<td>68.9 ( \pm ) 2.9</td>
<td>( \sigma_{el}^2 )</td>
<td>30.3 ( \pm ) 7.5</td>
</tr>
<tr>
<td>( \mu_{fr} )</td>
<td>62.7 ( \pm ) 2.8</td>
<td>( \rho_{elr} )</td>
<td>( -0.146 \pm 0.178 )</td>
</tr>
<tr>
<td>( \sigma_{al}^2 )</td>
<td>638.8 ( \pm ) 65.7</td>
<td>( \sigma_{er}^2 )</td>
<td>35.6 ( \pm ) 9.9</td>
</tr>
</tbody>
</table>

**Conclusions:**

1. Ridge counts higher for males than females and for right than left hands.
2. Left and right hand ridge counts are highly heritable traits.
3. Left and right hand ridge counts are determined by the same set of genes.
4. Environmental effects on two hands are independent.
• QTL mapping

Data:

\[ y = (y_1, \ldots, y_n)^T \]: Quantitative trait values observed on members of pedigrees.

\[ M = \{M_1, \ldots, M_n\} \]: Phenotypical or genotypical information of certain markers for the members of the pedigrees.

\[ X, \text{ a } n \times p \text{ matrix of values of other covariates.} \]

Conditional Model:
Consider one major QTL at a fixed position \( d \) and other minor QTL. Suppose dominant effects of the QTL are zero. Denote the additive effect by \( \sigma_*^2 \).
\( y|\mathbf{M}, \mathbf{x} \sim N(\mathbf{\nu}, \hat{\Omega}), \)
\( \mathbf{\nu} = X\beta, \)
\( \hat{\Omega} = \sigma^2_* \hat{\Pi}(d) + 2\sigma^2_a \Phi + \sigma^2_e I. \)

\( \hat{\Pi}_{ii} = 1, \hat{\Pi}_{ij} = E[\pi_{ij}|\mathbf{M}, d], \)

where \( \pi_{ij} \) is the proportion of ibd alleles between individuals \( i \) and \( j \).

**Likelihood ratio statistics:**

Likelihood function: \( L(\beta, \sigma^2_a, \sigma^2_e, \sigma^2_*, d). \)

Likelihood ratio statistic at \( d \):

\[
T(d) = 2 \ln \frac{\max L(\beta, \sigma^2_a, \sigma^2_e, \sigma^2_*, d)}{\max L(\beta, \sigma^2_a, \sigma^2_e, \sigma^2_* = 0)}.
\]
Mapping procedure
Compute $T(d)$ at each $d$ and then find $T = \max_d T(d)$.
If $T$ is significant, $\hat{d} = \arg\max_d T(d)$ is claimed as the position of the QTL.

Remark:
The model can be extended to multiple major QTL straightforwardly.