Summer Institute in Statistical Genetics

Module4: Mixed Models in Quantitative Genetics

Week 1 – Session2
Thursday9 – Friday 10 February 2017

Instructors
Professor Jian Yang – University of Queensland
Professor Bruce Walsh – University of Arizona

The University Of Queensland AUSTRALIA | IMB
Institute for Molecular Bioscience

SYLLABUS
MIXED MODELS IN QUANTITATIVE GENETICS

INSTRUCTORS:

Jian Yang, QBI, UQ
jian.yang@imb.uq.edu.au

Bruce Walsh, Department of Ecology & Evolutionary Biology, University of Arizona
jbwalsh@u.arizona.edu

LW = Lynch & Walsh: Genetics and Analysis of Quantitative Traits (book)

WL = Walsh & Lynch: Evolution and Selection of Quantitative Traits (website)
http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html

LECTURE SCHEDULE

Thursday, 9 Feb 2017

8:30 10:00 am 1. Introduction to matrix algebra (Walsh)
Background reading: LW, Chapter 8
Additional reading: LW Appendix 3; WL Appendix 5

10:00 10:30 am Break

10:30 12:00 2. The General Linear Model (Walsh)
Background reading: LW Chapter 8
Additional reading: LW Appendices 3, 4; WL Appendices 2, 3

12:00 1:30 pm Lunch

1:30 3:00 pm 3. Overview of the mixed model (Yang)
Additional reading: < LW Chapters 26, 27?>

3:00 3:30 pm Break

3:30 5:00 pm 4. Application: Association mapping (Yang)
Additional reading:

Friday, 10 Feb 2017

8:30 10:00 am 5. Application: BLUP and BLUP breeding values (Walsh)
Additional reading WL Chapter 19

10:00	10:30 am	Break
10:30	12:00	6. Application: Genomic Prediction (Yang) Additional reading:
12:00	1:30 pm	Lunch
1:30	3:00 pm	7. Application: Associative effects (Walsh) Additional reading: WL Chapter 20
3:00	3:30 pm	Break
3:30	5:00 pm	8. Random Regressions (Walsh)

Lecture 1: Intro/refresher in Matrix Algebra

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
9 – 10 Feb 2017

Topics

- Definitions, dimensionality, addition, subtraction
- Matrix multiplication
- Inverses, solving systems of equations
- Quadratic products and covariances
- The multivariate normal distribution
- Eigenstructure
- Basic matrix calculations in R
- The Singular Value Decompositon (SVD)

Matrices: An array of elements

Vectors: A matrix with either one row or one column.

Usually written in bold lowercase, e.g. **a, b, c**

$$\mathbf{a} = \begin{pmatrix} 12 \\ 13 \\ 47 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 2 & 0 & 5 & 21 \end{pmatrix}$$

Column vector (3 x 1)

Row vector (1 x 4)

Dimensionality of a matrix: r x c (rows x columns)
think of Railroad Car

General Matrices

Usually written in bold uppercase, e.g. **A, C, D**

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 0 & 1 \\ 3 & 4 \\ 2 & 9 \end{pmatrix}$$

(3 x 3)

Square matrix

(3 x 2)

Dimensionality of a matrix: r x c (rows x columns)
think of Railroad Car

A matrix is defined by a list of its elements.
B has ij-th element B_{ij} -- the element in row i and column j

Addition and Subtraction of Matrices

If two matrices have the same dimension (both are r x c), then matrix addition and subtraction simply follows by adding (or subtracting) on an element by element basis

Matrix addition: $(A+B)_{ij} = A_{ij} + B_{ij}$

Matrix subtraction: $(A-B)_{ij} = A_{ij} - B_{ij}$

Examples:

$$\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} 4 & 2 \\ 3 & 3 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}$$

Partitioned Matrices

It will often prove useful to divide (or partition) the elements of a matrix into a matrix whose elements are itself matrices.

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \left(\begin{array}{c|cc} 3 & 1 & 2 \\ \hline 2 & 5 & 4 \\ 1 & 1 & 2 \end{array}\right) = \begin{pmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{d} & \mathbf{B} \end{pmatrix}$$

$$\mathbf{a} = (3), \quad \mathbf{b} = (1 \quad 2), \quad \mathbf{d} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 5 & 4 \\ 1 & 2 \end{pmatrix}$$

One useful partition is to write the matrix as either a row vector of column vectors or a column vector of row vectors

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{pmatrix}$$

A column vector whose elements are row vectors

$$\mathbf{r}_1 = (3 \quad 1 \quad 2)$$
$$\mathbf{r}_2 = (2 \quad 5 \quad 4)$$
$$\mathbf{r}_3 = (1 \quad 1 \quad 2)$$

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = (\mathbf{c}_1 \quad \mathbf{c}_2 \quad \mathbf{c}_3)$$

A row vector whose elements are column vectors

$$\mathbf{c}_1 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{c}_2 = \begin{pmatrix} 1 \\ 5 \\ 1 \end{pmatrix}, \quad \mathbf{c}_3 = \begin{pmatrix} 2 \\ 4 \\ 2 \end{pmatrix}$$

Towards Matrix Multiplication: dot products

The dot (or inner) product of two vectors (both of length n) is defined as follows:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

Example:

$$\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = (4 \quad 5 \quad 7 \quad 9)$$

a ·b = 1*4 + 2*5 + 3*7 + 4*9 = 60

Matrices are compact ways to write systems of equations

$$
\begin{aligned}
5x_1 + 6x_2 + 4x_3 &= 6 \\
7x_1 - 3x_2 + 5x_3 &= -9 \\
-x_1 - x_2 + 6x_3 &= 12
\end{aligned}
$$

$$
\begin{pmatrix} 5 & 6 & 4 \\ 7 & -3 & 5 \\ -1 & -1 & 6 \end{pmatrix}
\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}
=
\begin{pmatrix} 6 \\ -9 \\ 12 \end{pmatrix}
$$

$$
\mathbf{Ax} = \mathbf{c}, \qquad \text{or} \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{c}
$$

The least-squares solution for the linear model

$$y = \mu + \beta_1 z_1 + \cdots \beta_n z_n$$

yields the following system of equations for the β_i

$$
\begin{aligned}
\sigma(y, z_1) &= \beta_1\sigma^2(z_1) + \beta_2\sigma(z_1, z_2) + \cdots + \beta_n\sigma(z_1, z_n) \\
\sigma(y, z_2) &= \beta_1\sigma(z_1, z_2) + \beta_2\sigma^2(z_2) + \cdots + \beta_n\sigma(z_2, z_n) \\
&\vdots \quad\quad \vdots \quad\quad \vdots \quad\quad \ddots \quad\quad \vdots \\
\sigma(y, z_n) &= \beta_1\sigma(z_1, z_n) + \beta_2\sigma(z_2, z_n) + \cdots + \beta_n\sigma^2(z_n)
\end{aligned}
$$

This can be more compactly written in matrix form as

$$
\begin{pmatrix}
\sigma^2(z_1) & \sigma(z_1, z_2) & \dots & \sigma(z_1, z_n) \\
\sigma(z_1, z_2) & \sigma^2(z_2) & \dots & \sigma(z_2, z_n) \\
\vdots & \vdots & \ddots & \vdots \\
\sigma(z_1, z_n) & \sigma(z_2, z_n) & \dots & \sigma^2(z_n)
\end{pmatrix}
\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}
=
\begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \\ \vdots \\ \sigma(y, z_n) \end{pmatrix}
$$

X^TX $\quad\quad$ β $\quad\quad$ X^Ty

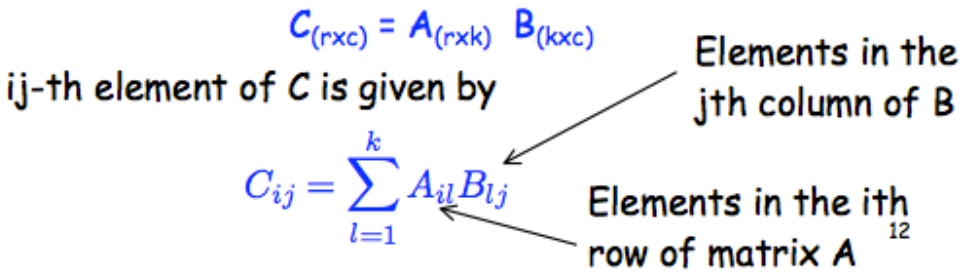
or, $\beta = (X^TX)^{-1} X^Ty$

Matrix Multiplication:

The order in which matrices are multiplied affects the matrix product, e.g. $AB \neq BA$

For the product of two matrices to exist, the matrices must conform. For AB, the number of columns of A must equal the number of rows of B.

The matrix C = AB has the same number of rows as A and the same number of columns as B.



Outer indices given dimensions of resulting matrix, with r rows (A) and c columns (B)

$$C_{(rxc)} = A_{(rxk)}\ B_{(kxc)}$$

Inner indices must match columns of A = rows of B

Example: Is the product ABCD defined? If so, what is its dimensionality? Suppose

$$A_{3x5}\ B_{5x9}\ C_{9x6}\ D_{6x23}$$

Yes, defined, as inner indices match. Result is a 3 x 23 matrix (3 rows, 23 columns)

More formally, consider the product L = MN

Express the matrix M as a column vector of row vectors

$$\mathbf{M} = \begin{pmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \vdots \\ \mathbf{m_r} \end{pmatrix} \quad \text{where} \quad \mathbf{m_i} = (M_{i1} \quad M_{i2} \quad \cdots \quad M_{ic})$$

Likewise express N as a row vector of column vectors

$$\mathbf{N} = (\mathbf{n_1} \quad \mathbf{n_2} \quad \cdots \quad \mathbf{n_b}) \quad \text{where} \quad \mathbf{n_j} = \begin{pmatrix} N_{1j} \\ N_{2j} \\ \vdots \\ N_{cj} \end{pmatrix}$$

The ij-th element of L is the inner product of M's row i with N's column j

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1 \cdot n_1} & \mathbf{m_1 \cdot n_2} & \cdots & \mathbf{m_1 \cdot n_b} \\ \mathbf{m_2 \cdot n_1} & \mathbf{m_2 \cdot n_2} & \cdots & \mathbf{m_2 \cdot n_b} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m_r \cdot n_1} & \mathbf{m_r \cdot n_2} & \cdots & \mathbf{m_r \cdot n_b} \end{pmatrix}$$

Example

$$\mathbf{AB} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

Likewise

$$\mathbf{BA} = \begin{pmatrix} ae+cf & eb+df \\ ga+ch & gd+dh \end{pmatrix}$$

ORDER of multiplication matters! Indeed, consider $C_{3x5}\, D_{5x5}$ which gives a 3 x 5 matrix, versus $D_{5x5}\, C_{3x5}$, which is not defined.

Matrix multiplication in R

```
> A<-matrix(c(1,2,3,4),nrow=2)
> B<-matrix(c(4,5,6,7),nrow=2)
> A
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> B
     [,1] [,2]
[1,]    4    6
[2,]    5    7
> A %*% B
     [,1] [,2]
[1,]   19   27
[2,]   28   40
```

R fills in the matrix from the list c by filling in as columns, here with 2 rows (nrow=2)

Entering A or B displays what was entered (always a good thing to check)

The command %*% is the R code for the multiplication of two matrices

On your own: What is the matrix resulting from BA? What is A if nrow=1 or nrow=4 is used?

The Transpose of a Matrix

The transpose of a matrix exchanges the rows and columns, $A^T_{ij} = A_{ji}$

Useful identities

$(AB)^T = B^T A^T$

$(ABC)^T = C^T B^T A^T$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

Inner product = $a^T b = a^T_{(1 \times n)}\, b_{(n \times 1)}$

Indices match, matrices conform

Dimension of resulting product is 1 X 1 (i.e. a scalar)

$$(a_1 \quad \cdots \quad a_n) \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

Note that $b^T a = (b^T a)^T = a^T b$

Outer product = $ab^T = a_{(n \times 1)} b^T_{(1 \times n)}$

Resulting product is an n x n matrix

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix}$$

$$= \begin{pmatrix} a_1b_1 & a_1b_2 & \dots & a_1b_n \\ a_2b_1 & a_2b_2 & \dots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \dots & a_nb_{bn} \end{pmatrix}$$

R code for transposition

```
> t(A)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
```

t(A) = transpose of A

```
> a<-matrix(c(1,2,3),nrow=3)
> a
     [,1]
[1,]    1
[2,]    2
[3,]    3
> t(a) %*% a
     [,1]
[1,]   14
> a %*% t(a)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    4    6
[3,]    3    6    9
```

Enter the column vector a

Compute inner product $\mathbf{a}^T\mathbf{a}$

Compute outer product $\mathbf{a}\mathbf{a}^T$

Solving equations

- The identity matrix I
  - Serves the same role as 1 in scalar algebra, e.g., a*1=1*a =a, with AI=IA= A
- The inverse matrix A^{-1} (IF it exists)
  - Defined by $A A^{-1} = I$, $A^{-1}A = I$
  - Serves the same role as scalar division
    - To solve ax = c, multiply both sides by (1/a) to give:
    - (1/a)*ax = (1/a)c or (1/a)*a*x = 1*x = x,
    - Hence x = (1/a)c
    - To solve Ax = c, $A^{-1}Ax = A^{-1} c$
    - Or $A^{-1}Ax = Ix = x = A^{-1} c$

The Identity Matrix, I

The identity matrix serves the role of the number 1 in matrix multiplication: AI =A, IA = A

I is a square diagonal matrix, with all diagonal elements being one, all off-diagonal elements zero.

$$I_{ij} = \begin{array}{l} 1 \text{ for } i = j \\ 0 \text{ otherwise} \end{array}$$

$$\mathbf{I}_{3x3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The Identity Matrix in R

diag(k), where k is an integer, return the k x k I matix

```
> I<-diag(4)
> I
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
> I2 <-diag(2)
> I2
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```


The Inverse Matrix, A^{-1}

For a square matrix A, define its Inverse A^{-1}, as the matrix satisfying

$$\mathbf{A^{-1}A = AA^{-1} = I}$$

For $\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ $\mathbf{A}^{-1} = \frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$

If this quantity (the determinant) is zero, the inverse does not exist.

If det(A) is not zero, A^{-1} exists and A is said to be non-singular. If det(A) = 0, A is singular, and no *unique* inverse exists (generalized inverses do)

Generalized inverses, and their uses in solving systems of equations, are discussed in Appendix 3 of Lynch & Walsh

A^- is the typical notation to denote the G-inverse of a matrix

When a G-inverse is used, *provided* the system is consistent, then some of the variables have a family of solutions (e.g., $x_1 = 2$, but $x_2 + x_3 = 6$)

Inversion in R

solve(A) computes A^{-1}

det(A) computes determinant of A

Using A entered earlier

```
> A
     [,1] [,2]
[1,]    1    3
[2,]    2    4
```

Compute A^{-1}

```
> solve(A)
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
```

Showing that $A^{-1} A = I$

```
> solve(A) %*% A
     [,1]          [,2]
[1,]    1 -8.881784e-16
[2,]    0  1.000000e+00
```

Computing determinant of A

```
> det(A)
[1] -2
```

Homework

Put the following system of equations in matrix form, and solve using R

$$
\begin{aligned}
3x_1 + 4x_2 + 4x_3 + 6x_4 &= -10 \\
9x_1 + 2x_2 - x_3 - 6x_4 &= 20 \\
x_1 + x_2 + x_3 - 10x_4 &= 2 \\
2x_1 + 9x_2 + 2x_3 - x_4 &= -10
\end{aligned}
$$

Example: solve the OLS for β in $y = \alpha + \beta_1 z_1 + \beta_2 z_2 + e$

$$\beta = \mathbf{V}^{-1}\,\mathbf{c} \qquad \mathbf{c} = \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma^2(z_1) & \sigma(z_1, z_2) \\ \sigma(z_1, z_2) & \sigma^2(z_2) \end{pmatrix}$$

It is more compact to use $\sigma(z_1, z_2) = \rho_{12}\,\sigma(z_1)\sigma(z_2)$

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1-\rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix}$$

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1-\rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix} \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \end{pmatrix}$$

$$\beta_1 = \frac{1}{1-\rho_{12}^2}\left[\frac{\sigma(y,z_1)}{\sigma^2(z_1)} - \rho_{12}\frac{\sigma(y,z_2)}{\sigma(z_1)\sigma(z_2)}\right]$$

$$\beta_2 = \frac{1}{1-\rho_{12}^2}\left[\frac{\sigma(y,z_2)}{\sigma^2(z_2)} - \rho_{12}\frac{\sigma(y,z_1)}{\sigma(z_1)\sigma(z_2)}\right]$$

If $\rho_{12} = 0$, these reduce to the two univariate slopes,

$$\beta_1 = \frac{\sigma(y,z_1)}{\sigma^2(z_1)} \quad \text{and} \quad \beta_2 = \frac{\sigma(y,z_2)}{\sigma^2(z_2)}$$

Likewise, if $\rho_{12} = 1$, this reduces to a univariate regression,

Useful identities

$(A^T)^{-1} = (A^{-1})^T$

$(AB)^{-1} = B^{-1} A^{-1}$

For a diagonal matrix **D**, then det (**D**), which is also denoted by |**D**|, = product of the diagonal elements

Also, the determinant of any square matrix **A**, det(**A**), is simply the product of the eigenvalues λ of **A**, which statisfy

$$\mathbf{Ae} = \lambda\mathbf{e}$$

If A is n x n, solutions to λ are an n-degree polynomial. **e** is the eigenvector associated with λ. If any of the roots to the equation are zero, A^{-1} is not defined. In this case, for some linear combination **b**, we have **Ab** = 0.

Variance-Covariance matrix

- A very important square matrix is the variance-covariance matrix **V** associated with a vector **x** of random variables.
- $V_{ij} = Cov(x_i,x_j)$, so that the i-th diagonal element of **V** is the variance of x_i, and off-diagonal elements are covariances
- **V** is a symmetric, square matrix

The trace

The trace, tr(**A**) or trace(**A**), of a square matrix A is simply the sum of its diagonal elements

The importance of the trace is that it equals the sum of the eigenvalues of **A**, $tr(\mathbf{A}) = \Sigma \lambda_i$

For a covariance matrix **V**, tr(V) measures the total amount of variation in the variables

λ_i / tr(**V**) is the fraction of the total variation in x contained in the linear combination $\mathbf{e}_i^T\mathbf{x}$, where $\mathbf{e}_i$, the i-th principal component of **V** is also the i-th eigenvector of **V** ($\mathbf{V}\mathbf{e}_i = \lambda_i \mathbf{e}_i$)

Eigenstructure in R

`eigen(A)` returns the eigenvalues and vectors of A

```
> V<-matrix(c(10,-5,10,-5,20,0,10,0,30),nrow=3)
> V
     [,1] [,2] [,3]
[1,]   10   -5   10
[2,]   -5   20    0
[3,]   10    0   30
> eigen(V)
$values
[1] 34.410103 21.117310  4.472587

$vectors
           [,1]       [,2]       [,3]
[1,]  0.3996151  0.2117936  0.8918807
[2,] -0.1386580 -0.9477830  0.2871955
[3,]  0.9061356 -0.2384340 -0.3493816
```

PC 1

Trace = 60

PC 1 accounts for 34.4/60 = 57% of all the variation

$0.400^* x_1 - 0.139^* x_2 + 0.906^* x_3$

Quadratic and Bilinear Forms

Quadratic product: for $A_{n \times n}$ and $x_{n \times 1}$

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$$ Scalar (1 x 1)

Bilinear Form (generalization of quadratic product)
for $A_{m \times n}$, $a_{n \times 1}$, $b_{m \times 1}$ their bilinear form is $b^T_{1 \times m} A_{m \times n} a_{n \times 1}$

$$\mathbf{b}^T\mathbf{A}\mathbf{a} = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}b_i a_j$$

Note that $b^T A\, a = a^T A^T b$

Covariance Matrices for Transformed Variables

What is the variance of the linear combination, $c_1x_1 + c_2x_2 + \ldots + c_nx_n$? (note this is a scalar)

$$\begin{aligned}\sigma^2(\mathbf{c}^T\mathbf{x}) &= \sigma^2\left(\sum_{i=1}^{n} c_i x_i\right) = \sigma\left(\sum_{i=1}^{n} c_i x_i, \sum_{j=1}^{n} c_j x_j\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sigma(c_i x_i, c_j x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sigma(x_i, x_j) \\ &= \mathbf{c}^T\mathbf{V}\mathbf{c}\end{aligned}$$

Likewise, the covariance between two linear combinations can be expressed as a bilinear form,

$$\sigma(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{x}) = \mathbf{a}^T\mathbf{V}\,\mathbf{b}$$

Example: Suppose the variances of x_1, x_2, and x_3 are 10, 20, and 30. x_1 and x_2 have a covariance of -5, x_1 and x_3 of 10, while x_2 and x_3 are uncorrelated.

What are the variances of the indices $y_1 = x_1-2x_2+5x_3$ and $y_2 = 6x_2-4x_3$?

$$\mathbf{V} = \begin{pmatrix} 10 & -5 & 10 \\ -5 & 20 & 0 \\ 10 & 0 & 30 \end{pmatrix}, \quad \mathbf{c}_1 = \begin{pmatrix} 1 \\ -2 \\ 5 \end{pmatrix}, \quad \mathbf{c}_2 = \begin{pmatrix} 0 \\ 6 \\ -4 \end{pmatrix}$$

$\text{Var}(y_1) = \text{Var}(c_1^T x) = c_1^T \text{Var}(x)\, c_1 = 960$

$\text{Var}(y_2) = \text{Var}(c_2^T x) = c_2^T \text{Var}(x)\, c_2 = 1200$

$\text{Cov}(y_1,y_2) = \text{Cov}(c_1^T x, c_2^T x) = c_1^T \text{Var}(x)\, c_2 = -910$

Homework: use R to compute the above values

The Multivariate Normal Distribution (MVN)

Consider the pdf for n independent normal random variables, the ith of which has mean μ_i and variance σ^2_i

$$p(\mathbf{x}) = \prod_{i=1}^{n} (2\pi)^{-1/2} \sigma_i^{-1} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

$$= (2\pi)^{-n/2} \left(\prod_{i=1}^{n} \sigma_i\right)^{-1} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right)$$

This can be expressed more compactly in matrix form

Define the covariance matrix V for the vector x of the n normal random variable by

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_n^2 \end{pmatrix} \qquad |\mathbf{V}| = \prod_{i=1}^{n} \sigma_i^2$$

Define the mean vector μ by gives

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

$$\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Hence in matrix from the MVN pdf becomes

$$p(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right]$$

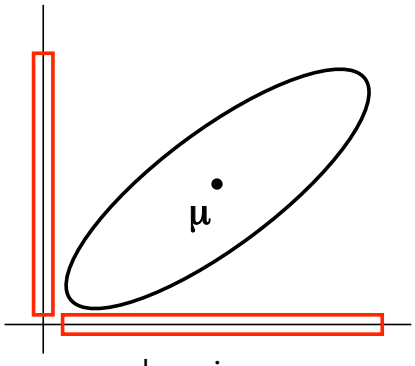
Notice this holds for any vector μ and symmetric positive -definite matrix V, as | V | > 0.

The multivariate normal

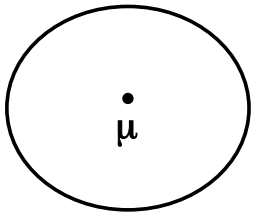
- Just as a univariate normal is defined by its mean and spread, a multivariate normal is defined by its mean vector $\boldsymbol{\mu}$ (also called the centroid) and variance -covariance matrix $\mathbf{V}$

Vector of means $\boldsymbol{\mu}$ determines location

Spread (geometry) about $\boldsymbol{\mu}$ determined by $\mathbf{V}$



x_1, x_2 equal variances, positively correlated

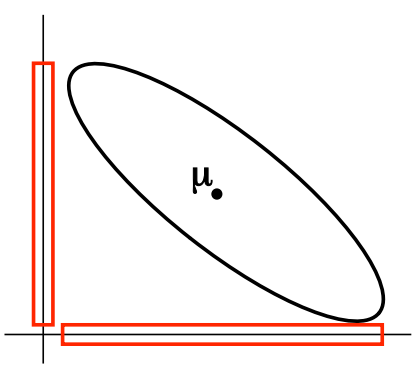


x_1, x_2 equal variances, uncorrelated

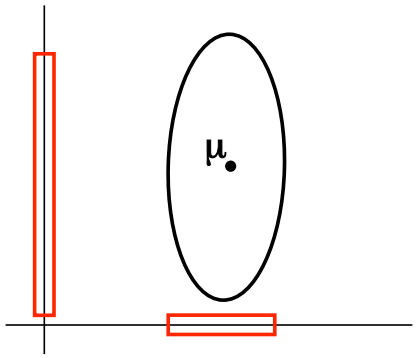
Eigenstructure (the eigenvectors and their corresponding eigenvalues) determines the geometry of $\mathbf{V}$.

Vector of means μ determines location

Spread (geometry) about μ determined by $\mathbf{V}$



x_1, x_2 equal variances, negatively correlated

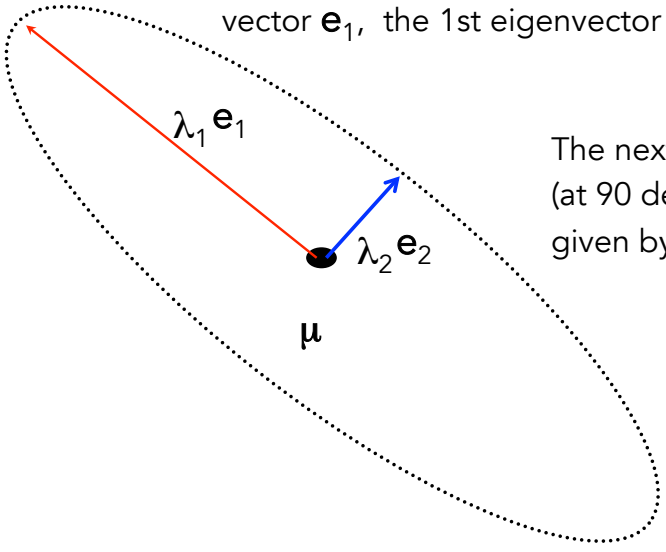


$Var(x_1) < Var(x_2)$, uncorrelated

Positive tilt = positive correlations
Negative tilt = negative correlation
No tilt = uncorrelated

Eigenstructure of V

The direction of the largest axis of variation is given by the unit-length vector $\mathbf{e}_1$, the 1st eigenvector of V.



The next largest axis of orthogonal (at 90 degrees from) $\mathbf{e}_1$, is given by $\mathbf{e}_2$, the 2nd eigenvector

Principal components

- The principal components (or PCs) of a covariance matrix define the axes of variation.
  - PC1 is the direction (linear combination $\mathbf{c}^T\mathbf{x}$) that explains the most variation.
  - PC2 is the next largest direction (at 90degree from PC1), and so on
- PC_i = ith eigenvector of V
- Fraction of variation accounted for by PCi = λ_i / trace(V)
- If V has a few large eigenvalues, most of the variation is distributed along a few linear combinations (axis of variation)
- The singular value decomposition is the generalization of this idea to nonsquare matrices

Properties of the MVN - I

1) If **x** is MVN, any subset of the variables in **x** is also MVN

2) If **x** is MVN, any linear combination of the elements of **x** is also MVN. If **x** ~ MVN(μ,V)

$$\text{for} \quad \mathbf{y} = \mathbf{x} + \mathbf{a}, \qquad \mathbf{y} \text{ is } \mathrm{MVN}_n(\boldsymbol{\mu} + \mathbf{a}, \mathbf{V})$$

$$\text{for} \quad y = \mathbf{a}^T\mathbf{x} = \sum_{k=1}^{n} a_i x_i, \qquad y \text{ is } \mathrm{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\mathbf{V}\,\mathbf{a})$$

$$\text{for} \quad \mathbf{y} = \mathbf{A}\mathbf{x}, \qquad \mathbf{y} \text{ is } \mathrm{MVN}_m\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}^T\mathbf{V}\mathbf{A}\right)$$

Properties of the MVN - II

3) Conditional distributions are also MVN. Partition x into two components, x_1 (m dimensional column vector) and x_2 (n-m dimensional column vector)

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V_{x_1 x_1}} & \mathbf{V_{x_1 x_2}} \\ \mathbf{V^T_{x_1 x_2}} & \mathbf{V_{x_2 x_2}} \end{pmatrix}$$

x_1 | x_2 is MVN with m-dimensional mean vector

$$\boldsymbol{\mu}_{\mathbf{x_1}|\mathbf{x_2}} = \boldsymbol{\mu_1} + \mathbf{V_{x_1 x_2}} \mathbf{V^{-1}_{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu_2})$$

and m x m covariance matrix

$$\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}} = \mathbf{V_{x_1 x_1}} - \mathbf{V_{x_1 x_2}} \mathbf{V^{-1}_{x_2 x_2}} \mathbf{V^T_{x_1 x_2}}$$

Properties of the MVN - III

4) If x is MVN, the regression of any subset of x on another subset is linear and homoscedastic

$$\mathbf{x_1} = \boldsymbol{\mu}_{\mathbf{x_1}|\mathbf{x_2}} + \mathbf{e}$$
$$= \boldsymbol{\mu_1} + \mathbf{V_{x_1 x_2}} \mathbf{V^{-1}_{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e}$$

Where e is MVN with mean vector 0 and variance-covariance matrix $\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}}$

$$\boldsymbol{\mu_1} + \mathbf{V_{x_1 x_2}} \mathbf{V}^{-1}_{\mathbf{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e}$$

The regression is linear because it is a linear function of x_2

The regression is homoscedastic because the variance-covariance matrix for e does not depend on the value of the x's

$$\mathbf{V_{x_1|x_2}} = \mathbf{V_{x_1 x_1}} - \mathbf{V_{x_1 x_2}} \mathbf{V}^{-1}_{\mathbf{x_2 x_2}} \mathbf{V}^{T}_{\mathbf{x_1 x_2}}$$

All these matrices are constant, and hence the same for any value of x

Example: Regression of Offspring value on Parental values

Assume the vector of offspring value and the values of both its parents is MVN. Then from the correlations among (outbred) relatives,

$$\begin{pmatrix} z_o \\ z_s \\ z_d \end{pmatrix} \sim \text{MVN} \left[\begin{pmatrix} \mu_o \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_z^2 \begin{pmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{pmatrix} \right]$$

Let $\mathbf{x_1} = (z_o), \quad \mathbf{x_2} = \begin{pmatrix} z_s \\ z_d \end{pmatrix}$

$$\mathbf{V_{x_1,x_1}} = \sigma_z^2, \quad \mathbf{V_{x_1,x_2}} = \frac{h^2 \sigma_z^2}{2} (1 \quad 1), \quad \mathbf{V_{x_2,x_2}} = \sigma_z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \boldsymbol{\mu_1} + \mathbf{V_{x_1 x_2}} \mathbf{V}^{-1}_{\mathbf{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e}$$

Regression of Offspring value on Parental values (cont.)

$$= \boldsymbol{\mu_1} + \mathbf{V_{x_1 x_2}} \mathbf{V^{-1}_{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e}$$

$$\mathbf{V}_{\mathbf{x}_1,\mathbf{x}_1} = \sigma_z^2, \quad \mathbf{V}_{\mathbf{x}_1,\mathbf{x}_2} = \frac{h^2\sigma_z^2}{2}(1 \quad 1), \quad \mathbf{V}_{\mathbf{x}_2,\mathbf{x}_2} = \sigma_z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Hence,

$$\begin{aligned} z_o &= \mu_o + \frac{h^2\sigma_z^2}{2}(1 \quad 1)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_s - \mu_s \\ z_d - \mu_d \end{pmatrix} + e \\ &= \mu_o + \frac{h^2}{2}(z_s - \mu_s) + \frac{h^2}{2}(z_d - \mu_d) + e \end{aligned}$$

Where e is normal with mean zero and variance

$$\mathbf{V_{x_1|x_2}} = \mathbf{V_{x_1 x_1}} - \mathbf{V_{x_1 x_2}} \mathbf{V^{-1}_{x_2 x_2}} \mathbf{V^{T}_{x_1 x_2}}$$

$$\begin{aligned} \sigma_e^2 &= \sigma_z^2 - \frac{h^2\sigma_z^2}{2}(1 \quad 1)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{h^2\sigma_z^2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \sigma_z^2 \left(1 - \frac{h^4}{2}\right) \end{aligned}$$

Hence, the regression of offspring trait value given the trait values of its parents is

$z_o = \mu_o + h^2/2(z_s - \mu_s) + h^2/2(z_d - \mu_d) + e$

where the residual e is normal with mean zero and $Var(e) = \sigma_z^2(1-h^4/2)$

Similar logic gives the regression of offspring breeding value on parental breeding value as

$A_o = \mu_o + (A_s - \mu_s)/2 + (A_d - \mu_d)/2 + e$
$= A_s/2 + A_d/2 + e$

where the residual e is normal with mean zero and $Var(e) = \sigma_A^2/2$

The Singular-Value Decomposition (SVD)

An $n \times p$ matrix $\mathbf{A}$ can always be decomposed as the product of three matrices: an $n \times p$ diagonal matrix $\boldsymbol{\Lambda}$ and two unitary matrices, $\mathbf{U}$ which is $n \times n$ and $\mathbf{V}$ which is $p \times p$. The resulting **singular value decomposition** (**SVD**) of $\mathbf{A}$ is given by

$$\mathbf{A}_{n\times p} = \mathbf{U}_{n\times n}\boldsymbol{\Lambda}_{n\times p}\mathbf{V}^T_{p\times p} \tag{39.16a}$$

We have indicated the dimensionality of each matrix to allow the reader to verify that each matrix multiplication conforms. The diagonal elements $\lambda_1, \cdots, \lambda_s$ of $\boldsymbol{\Lambda}$ correspond to the **singular values** of $\mathbf{A}$ and are ordered by decreasing magnitude. Returning to the unitary matrices $\mathbf{U}$ and $\mathbf{V}$, we can write each as a row vector of column vectors,

$$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_i, \cdots \mathbf{u}_n), \qquad \mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_i, \cdots \mathbf{v}_p) \tag{39.16b}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are n and p-dimensional column vectors (often called the **left** and **right singular vectors**, respectively). Since both $\mathbf{U}$ and $\mathbf{V}$ are unitary, by definition (Appendix 4) each column vector has length one and are mutually orthogonal (i.e., if $i \neq j$, $\mathbf{u}_i\mathbf{u}_j^T = \mathbf{v}_i\mathbf{v}_j^T = 0$). Since $\boldsymbol{\Lambda}$ is diagonal, it immediately follows from matrix multiplication that we can write any element in $\mathbf{A}$ as

$$A_{ij} = \sum_{k=1}^{s} \lambda_k \, u_{ik} \, v_{kj} \tag{39.16c}$$

where λ_k is the kth singular value and $s \leq \min(p, n)$ is the number of non-zero singular values.

The importance of the singular value decomposition in the analysis of G×E arises from the **Eckart-Young theorem** (1938), which relates the best approximation of a matrix by some lower-rank (say k) matrix with the SVD. Define as our measure of goodness of fit between a matrix $\mathbf{A}$ and a lower rank approximation $\widehat{\mathbf{A}}$ as the sum of squared differences over all elements,

$$\sum_{ij}(A_{ij} - \hat{A}_{ij})^2$$

Eckart and Young show that the best fitting approximation $\widehat{\mathbf{A}}$ of rank $m < s$ is given from the first m terms of the singular value decomposition (the **rank-m SVD**),

$$\hat{A}_{ij} = \sum_{k=1}^{m} \lambda_k \, u_{ik} \, v_{kj} \tag{39.17a}$$

For example, the best rank-2 approximation for the G×E interaction is given by

$$GE_{ij} \simeq \lambda_1 \, u_{i1} \, v_{j1} + \lambda_2 \, u_{i2} \, v_{j2} \tag{39.17b}$$

where λ_i is the ith singular value of the **GE** matrix, $\mathbf{u}$ and $\mathbf{v}$ are the associated singular vectors (see Example 39.3). The fraction of total variation of a matrix accounted for by taking the first m terms in its SVD is

$$\sum_{k=1}^{m} \lambda_k^2 / \sum_{ij} A_{ij}^2 = \frac{\lambda_1^2 + \cdots + \lambda_m^2}{\lambda_1^2 + \cdots + \lambda_s^2}$$

A data set for soybeans grown in New York (Gauch 1992) gives the GE matrix as

$$\mathbf{GE} = \begin{pmatrix} 57 & 176 & -233 \\ -36 & -196 & 233 \\ -45 & -324 & 369 \\ -66 & 178 & -112 \\ 89 & 165 & -254 \end{pmatrix}$$

Where GE_{ij} = value for Genotype i in envir. j

In **R**, the compact SVD (Equation 39.16d) of a matrix X is given by **svd(X)**, returning the SVD of $\mathbf{GE}$ as

$$\begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 131.36 & 0 \\ 0 & 0 & 0.53 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

The first singular value accounts for $746.10^2/(743.26^2 + 131.36^2 + 0.53^2) = 97.0\%$ of the total variation of $\mathbf{GE}$, while the second singular value accounts for 3.0%, so that together they account for essentially all of the total variation. The rank-1 SVD approximation of $\mathbf{GE}$ is given by setting all of the diagonal elements of $\boldsymbol{\Lambda}$ except the first entry to zero,

$$\mathbf{GE}_1 = \begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

Similarly, the rank-2 SVD is given by setting all but the first two singular values to zero,

$$\mathbf{GE}_2 = \begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 131.36 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

For example, the rank-1 SVD approximation for GE_{32} is $g_{31}\lambda_1 e_{12}$ = 746.10*(-0.66)*0.64 = -315

While the rank-2 SVD approximation is $g_{31}\lambda_2 e_{12} + g_{32}\lambda_2 e_{22}$ = 746.10*(-0.66)*0.64 + 131.36* 0.12*(-0.51) = -323

Actual value is -324

Generally, the rank-2 SVD approximation for GE_{ij} is $g_{i1}\lambda_1 e_{1j} + g_{i2}\lambda_2 e_{2j}$

Additional R matrix commands

Operator or Function	Description
A * B	Element-wise multiplication
A %*% B	Matrix multiplication
A %o% B	Outer product. **AB'**
crossprod(A,B) **crossprod(A)**	**A'B** and **A'A** respectively.
t(A)	Transpose
diag(x)	Creates diagonal matrix with elements of **x** in the principal diagonal
diag(A)	Returns a vector containing the elements of the principal diagonal
diag(k)	If k is a scalar, this creates a k x k identity matrix. Go figure.
solve(A, b)	Returns vector **x** in the equation **b = Ax** (i.e., $A^{-1}b$)
solve(A)	Inverse of **A** where A is a square matrix.
ginv(A)	Moore-Penrose Generalized Inverse of **A**. ginv(A) requires loading the **MASS** package.
y<-eigen(A)	**y$val** are the eigenvalues of **A** **y$vec** are the eigenvectors of **A**
y<-svd(A)	Single value decomposition of **A**. **y$d** = vector containing the singular values of **A** **y$u** = matrix with columns contain the left singular vectors of **A** **y$v** = matrix with columns contain the right singular vectors of **A**

Additional R matrix commands (cont)

R <- chol(A)	Choleski factorization of **A**. Returns the upper triangular factor, such that **R'R = A**.
y <- qr(A)	QR decomposition of **A**. **y$qr** has an upper triangle that contains the decomposition and a lower triangle that contains information on the Q decomposition. **y$rank** is the rank of A. **y$qraux** a vector which contains additional information on Q. **y$pivot** contains information on the pivoting strategy used.
cbind(A,B,...)	Combine matrices(vectors) horizontally. Returns a matrix.
rbind(A,B,...)	Combine matrices(vectors) vertically. Returns a matrix.
rowMeans(A)	Returns vector of row means.
rowSums(A)	Returns vector of row sums.
colMeans(A)	Returns vector of column means.
colSums(A)	Returns vector of coumn means.

Additional references

- Lynch & Walsh Chapter 8 (intro to matrices)
- Online notes:
  - Appendix 4 (Matrix geometry)
  - Appendix 5 (Matrix derivatives)

Lecture 2:
Linear and Mixed Models

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
9 – 10 Feb 2017

Quick Review of the Major Points

The general linear model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

- $\mathbf{y}$ = vector of observed dependent values
- $\mathbf{X}$ = Design matrix: observations of the variables in the assumed linear model
- $\boldsymbol{\beta}$ = vector of unknown parameters to estimate
- $\mathbf{e}$ = vector of residuals (deviation from model fit), $\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$

$y = X\beta + e$

Solution to β depends on the ***covariance structure*** (= covariance matrix) of the vector **e** of residuals

Ordinary least squares (OLS)

- OLS: $\mathbf{e} \sim MVN(\mathbf{0}, \sigma^2 \mathbf{I})$
- Residuals are homoscedastic and uncorrelated, so that we can write the cov matrix of **e** as $Cov(e) = \sigma^2 I$
- the OLS estimate, $OLS(\beta) = (X^T X)^{-1} X^T y$

Generalized least squares (GLS)

- GLS: $e \sim MVN(\mathbf{0}, \mathbf{V})$
- Residuals are heteroscedastic and/or dependent,
- $GLS(\beta) = (X^T V^{-1} X)^{-1} V^{-1} X^T y$

BLUE

- Both the OLS and GLS solutions are also called the Best Linear Unbiased Estimator (or BLUE for short)
- Whether the OLS or GLS form is used depends on the assumed covariance structure for the residuals
  - Special case of $Var(\mathbf{e}) = \sigma_e^2 \mathbf{I}$ -- OLS
  - All others, i.e., $Var(\mathbf{e}) = \mathbf{R}$ -- GLS

Linear Models

One tries to explain a dependent variable y as a linear function of a number of independent (or predictor) variables.

A multiple regression is a typical linear model,

$$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_x + e$$

Here e is the residual, or deviation between the true value observed and the value predicted by the linear model.

The (partial) regression coefficients are interpreted as follows: a unit change in x_i while holding all other variables constant results in a change of β_i in y

Linear Models

As with a univariate regression (y = a + bx + e), the model parameters are typically chosen by least squares, wherein they are chosen to minimize the sum of squared residuals, $\Sigma\ e_i^2$

This unweighted sum of squared residuals assumes an OLS error structure, so all residuals are equally weighted (homoscedastic) and uncorrelated

If the residuals differ in variances and/or some are correlated (GLS conditions), then we need to minimize the weighted sum $\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e}$, which removes correlations and gives all residuals equal variance.

Linear Models in Matrix Form

Suppose we have 3 variables in a multiple regression, with four (y,x) vectors of observations.

$$y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

In matrix form, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

The design matrix X. Details of both the experimental design and the observed values of the predictor variables all reside solely in X

Rank of the design matrix

- With n observations and p unknowns, X is an n x p matrix, so that X^TX is p x p
- Thus, at most X can provide unique estimates for up to p < n parameters
- The rank of X is the number of independent rows of X. If X is of full rank, then rank = p
- A parameter is said to be estimable if we can provide a unique estimate of it. If the rank of X is k < p, then exactly k parameters are estimable (some as linear combinations, e.g. β_1-3β_3 = 4)
- if $\det(X^TX) = 0$, then X is not of full rank
- Number of nonzero eigenvalues of X^TX gives the rank of X.

Experimental design and X

- The structure of X determines not only which parameters are estimable, but also the expected sample variances, as $Var(\beta) = k (X^TX)^{-1}$
- Experimental design determines the structure of X before an experiment (of course, missing data almost always means the final X is different form the proposed X)
- Different criteria used for an optimal design. Let $V = (X^TX)^{-1}$. The idea is to chose a design for X given the constraints of the experiment that:
  - A-optimality: minimizes tr(V)
  - D-optimality: minimizes det(V)
  - E-optimality: minimizes leading eigenvalue of V

Ordinary Least Squares (OLS)

When the covariance structure of the residuals has a certain form, we solve for the vector β using OLS

If residuals follow a MVN distribution, OLS = ML solution

If the residuals are homoscedastic and uncorrelated, $\sigma^2(e_i) = \sigma_e^2$, $\sigma(e_i,e_j) = 0$. Hence, each residual is equally weighted,

Sum of squared residuals can be written as

$$\sum_{i=1}^{n} \widehat{e}_i^2 = \widehat{\mathbf{e}}^T\widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

Predicted value of the y's

Ordinary Least Squares (OLS)

$$\sum_{i=1}^{n} \widehat{e}_i^2 = \widehat{\mathbf{e}}^T \widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Taking (matrix) derivatives shows this is minimized by

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This is the OLS estimate of the vector β

The variance-covariance estimate for the sample estimates is

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$$

The ij-th element gives the covariance between the estimates of β_i and β_j.

Sample Variances/Covariances

The residual variance can be estimated as

$$\widehat{\sigma_e^2} = \frac{1}{n - \operatorname{rank}(X)} \sum_{i=1}^{n} \widehat{e}_i^2$$

The estimated residual variance can be substituted into

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$$

To give an approximation for the sampling variance and covariances of our estimates.

Confidence intervals follow since the vector of estimates ~ MVN($\boldsymbol{\beta}$, $\mathbf{V}_{\boldsymbol{\beta}}$)

Example: Regression Through the Origin

$$y_i = \beta x_i + e_i$$

Here $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ $\boldsymbol{\beta} = (\beta)$

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} x_i^2 \qquad \mathbf{X}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i$$

$$\beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \frac{\sum x_i y_i}{\sum x_i^2} \qquad \sigma^2(b) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma_e^2 = \frac{\sigma_e^2}{\sum x_i^2}$$

$$\sigma^2(\beta) = \frac{1}{n-1}\frac{\sum(y_i - \beta x_i)^2}{\sum x_i^2} \qquad \sigma_e^2 = \frac{1}{n-1}\sum(y_i - \beta x_i)^2$$

Polynomial Regressions

GLM can easily handle any function of the observed predictor variables, provided the parameters to estimate are still linear, e.g. $Y = \alpha + \beta_1 f(x) + \beta_2 g(x) + \cdots + e$

Quadratic regression:

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Interaction Effects

Interaction terms (e.g. sex x age) are handled similarly

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + e_i$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}$$

With x_1 held constant, a unit change in x_2 changes y by $\beta_2 + \beta_3 x_1$ (i.e., the slope in x_2 depends on the current value of x_1)

Likewise, a unit change in x_1 changes y by $\beta_1 + \beta_3 x_2$

The GLM lets you build your own model!

- Suppose you want a quadratic regression forced through the origin where the slope of the quadratic term can vary over the sexes (pollen vs. seed parents)
- $Y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 s_i x_i^2$
- s_i is an indicator (0/1) variable for the sex (0 = male, 1 = female).
  - Male slope = β_2,
  - Female slope = $\beta_2 + \beta_3$

Generalized Least Squares (GLS)

Suppose the residuals no longer have the same variance (i.e., display heteroscedasticity). Clearly we do not wish to minimize the *unweighted* sum of squared residuals, because those residuals with smaller variance should receive more weight.

Likewise in the event the residuals are correlated, we also wish to take this into account (i.e., perform a suitable transformation to remove the correlations) before minimizing the sum of squares.

Either of the above settings leads to a GLS solution in place of an OLS solution.

In the GLS setting, the covariance matrix for the vector e of residuals is written as R where $R_{ij} = \sigma(e_i, e_j)$

The linear model becomes $y = X\beta + e$, $cov(e) = R$

The GLS solution for β is

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}$$

The variance-covariance of the estimated model parameters is given by

$$\mathbf{V_b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\sigma_e^2$$

Model diagnostics

- It's all about the residuals
- Plot the residuals
  - Quick and easy screen for outliers
  - Plot y or yhat on e
- Test for normality among estimated residuals
  - Q-Q plot
  - Wilk-Shapiro test
  - If non-normal, try transformations, such as log

OLS, GLS summary

	OLS	GLS
Assumed distribution of residuals	$\mathbf{e} \sim (\mathbf{0}, \sigma_e^2 \mathbf{I})$	$\mathbf{e} \sim (\mathbf{0}, \mathbf{V})$
Least-squares estimator of $\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
$\mathrm{Var}(\widehat{\boldsymbol{\beta}})$	$(\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
Predicted values, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
$\mathrm{Var}(\widehat{\mathbf{y}})$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_e^2$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

Fixed vs. Random Effects

In linear models are are trying to accomplish two goals: estimation the values of model parameters and estimate any appropriate variances.

For example, in the simplest regression model, $y = \alpha + \beta x + e$, we estimate the values for α and β and also the variance of e. We, of course, can also estimate the $e_i = y_i - (\alpha + \beta x_i)$

Note that α/β are ***fixed constants*** are we trying to estimate (fixed factors or fixed effects), while the e_i values are drawn from some probability distribution (typically Normal with mean 0, variance σ^2_e). The e_i are random effects.

This distinction between fixed and random effects is extremely important in terms of how we analyzed a model. If a parameter is a fixed constant we wish to estimate, it is a fixed effect. If a parameter is drawn from some probability distribution and we are trying to make inferences on either the distribution and/or specific realizations from this distribution, it is a random effect.

We generally speak of estimating fixed factors (BLUE) and predicting random effects (BLUP -- best linear unbiased Predictor)

"Mixed" models (MM) contain both fixed and random factors

$$y = Xb + Zu + e, \quad u \sim MVN(0,R), \; e \sim MVN(0,\sigma^2_e I)$$

Key: need to specify covariance structures for MM

Random effects models

- It is often useful to treat certain effects as random, as opposed to fixed
  - Suppose we have k effects. If we treat these as fixed, we lose k degrees of freedom
  - If we assume each of the k realizations are drawn from a normal with mean zero and unknown variance, only one degree of freedom lost --- that for estimating the variance
    - We can then predict the values of the k realizations

Environmental effects

- Consider yield data measured over several years in a series of plots.
- Standard to treat year-to-year variation at a specific site as being random effects
- Often the plot effects (mean value over years) are also treated as random.
- For example, consider plants group in growing region i, location j within that region, and year (season) k for that location-region effect
  - $E = R_i + L_{ij} + e_{ijk}$
  - Typically R can be a fixed effect, while L and e are random effects, $L_{ik} \sim N(0,\sigma^2_L)$ and $e_{ikj} \sim N(0,\sigma^2_e)$

Random models

- With a random model, one is assuming that all "levels" of a factor are not observed. Rather, some subset of values are drawn from some underlying distribution
  - For example, year to year variation in rainfall at a location. Each year is a random sample from the long-term distribution of rainfall values
  - Typically, assume a functional form for this underlying distribution (e.g., normal with mean 0) and then use observations to estimate the distribution parameters (here, the variance)

Random models (cont)

- Key feature:
  - Only one degree of freedom used (estimate of the variance)
  - Using the fixed effects and the estimated underlying distribution parameters, one then predicts the actual realizations of the individual values (i.e., the year effects)
  - Assumption: the covariance structure among the individual realizations of the realized effects. If only a variance is assume, this implies they are independent. If they are assumed to be correlated, this structure must be estimated.

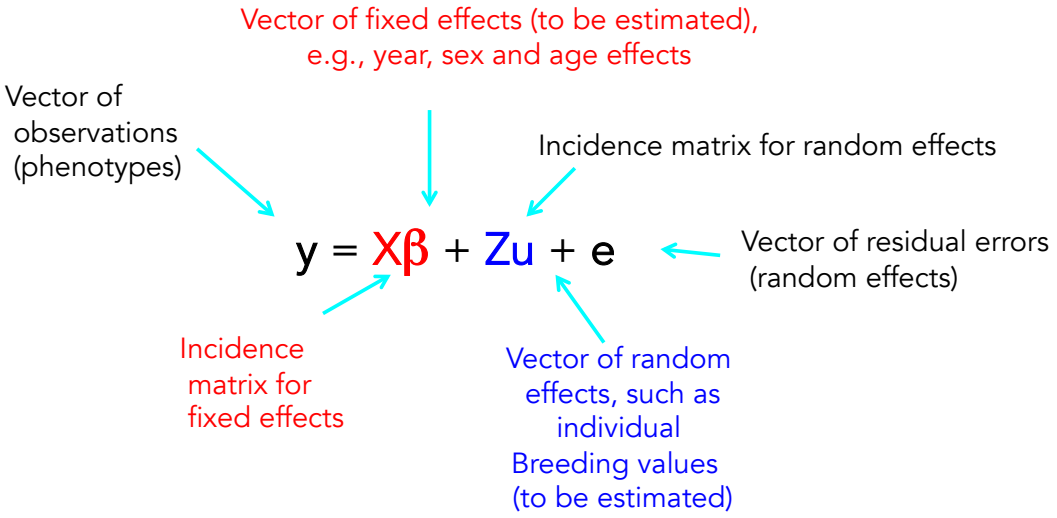
Random models

- Let's go back to treating yearly effects as random
- If assume these are uncorrelated, only use one degree of freedom, but makes assumptions about covariance structure
  - Standard: Uncorrelated
  - Option: some sort of autocorrelation process, say with a yearly decay of r (must also be estimated)
- Conversely, could all be treated as fixed, but would use k degrees of freedom for k years, but no assumptions on their relationships (covariance structure)

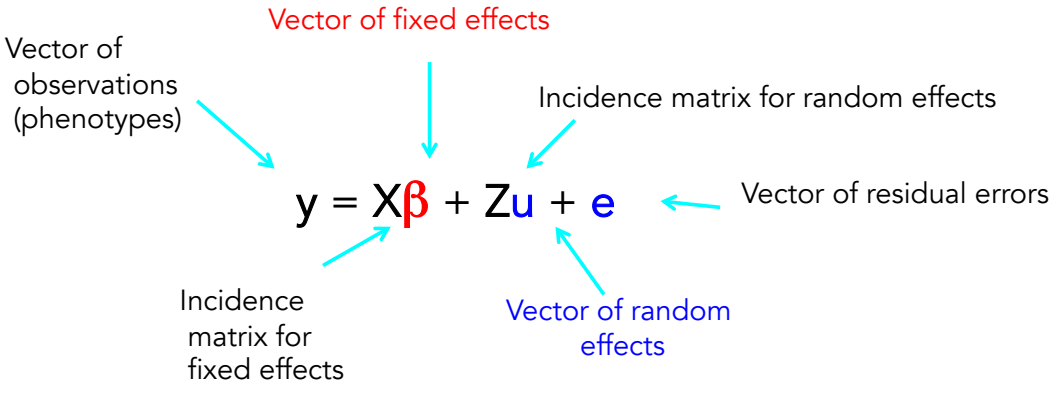
Identifiability

- Recall that a fixed effect is said to be estimable if we can obtain a unique estimate for it (either because X is of full rank or when using a generalized inverse it returns a unique estimate)
  - Lack of estimable arises because the experiment design confounds effects
- The analogous term for random models is identifiability
  - The variance components have unique estimates

The general mixed model



The general mixed model



Observe **y, X, Z.**

Estimate fixed effects **β**

Estimate random effects **u, e**

Means & Variances for $y = X\beta + Zu + e$

Means: $E(u) = E(e) = 0$, $E(y) = X\beta$

Variances:

Let R be the covariance matrix for the residuals. We typically assume $R = \sigma^2_e * I$

Let G be the covariance matrix for the vector **u** of random effects

The covariance matrix for y becomes

$V = ZGZ^T + R$

Hence, $y \sim MVN(X\beta, V)$

Mean $X\beta$ due to fixed effects
Variance V due to random effects

Estimating fixed Effects & Predicting Random Effects

For a mixed model, we observe **y**, X, and Z

β, u, R, and G are generally unknown

Two complementary estimation issues

(i) Estimation of **β** and **u**

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$ Estimation of fixed effects

BLUE = Best Linear Unbiased Estimator

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\beta}\right)$$ Prediction of random effects

BLUP = Best Linear Unbiased Predictor

Recall $V = ZGZ^T + R$

Different statistical models

- GLM = general linear model
  - OLS ordinary least squares: e ~ MVN(0,cI)
  - GLS generalized least squares: e ~ MVN(0,R)
- Mixed models
  - Both fixed and random effects (beyond the residual)
- Mixture models
  - A weighted mixture of distributions
- Generalized linear models
  - Nonlinear functions, non-normality

Mixture models

- Under a mixture model, an observation potentially comes from one of several different distributions, so that the density function is $\pi_1\phi_1 + \pi_2\phi_2 + \pi_3\phi_3$
  - The mixture proportions π_i sum to one
  - The ϕ_i represent different distribution, e.g., normal with mean μ_i and variance σ^2
- Mixture models come up in QTL mapping -- an individual could have QTL genotype QQ, Qq, or qq
  - See Lynch & Walsh Chapter 13
- They also come up in codon models of evolution, were a site may be neutral, deleterious, or advantageous, each with a different distribution of selection coefficients
  - See Walsh & Lynch (volume 2A website), Chapters 10,11

Generalized linear models

The **Generalized Linear Model** (note the **ized** ending) takes this a step further by assuming for some monotonic function g, that

$$E[y_i] = g\left(\mu + \sum_{k=1}^{n} \beta_k x_{ik}\right) \tag{2}$$

In particular, taking the inverse g^{-1} of the function g returns a linear model, with

$$g^{-1}(E[y_i]) = \mu + \sum_{k=1}^{n} \beta_k x_{ik} \tag{3}$$

The function f with the property that expresses the expected value of the response variable as a linear function of the predictor variables, i.e.,

$$f(E[y_i]) = \mu + \sum_{k=1}^{n} \beta_k x_{ik}$$

is called the **link function** of the particular generalized linear model.

Typically assume non-normal distribution for residuals, e.g., Poisson, binomial, gamma, etc

Lecture #3

GREML: estimation of genetic variance in unrelated individuals

Jian Yang
Institute for Molecular Bioscience
The University of Queensland

Keywords

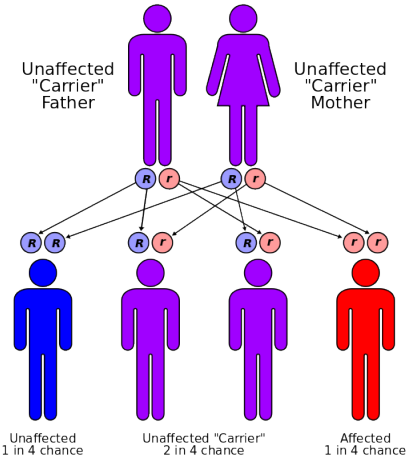
Genetic variance (V_G): the amount of phenotypic variance (V_P) in a population attributable to genetic factors.

What do we need to estimate V_G in **unrelated individuals**?

Why is **mixed linear model** (MLM) useful in this case?

Mendelian traits

Cystic fibrosis



Complex traits

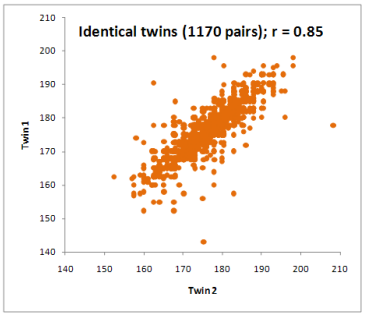
Human height

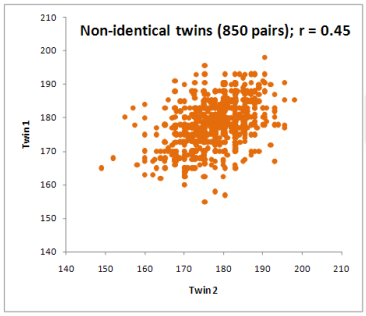
Body mass index

Schizophrenia

Heritability

Resemblance between twins for human height





Heritability = ~80-90%

Resemblance between relatives for body mass index (BMI)

Relatedness	Correlation
Full-sibs	0.36
Father-son	0.28

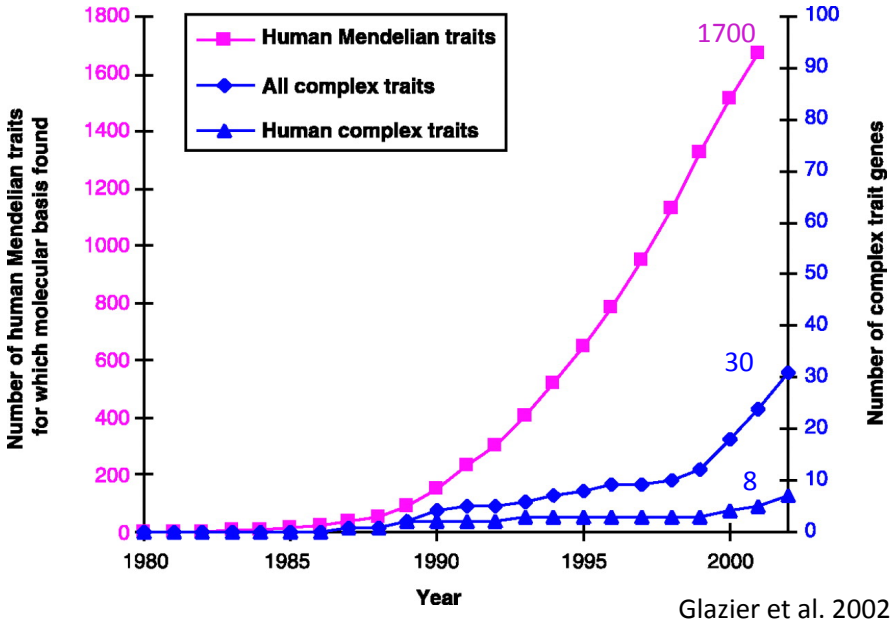
Heritability = ~40-60%

Risk of schizophrenia (%)

Heritability = ~70-80%

	Population	1st degree relative	MZ twin
Schizophrenia	1	10	50

Identifying genes underlying complex traits

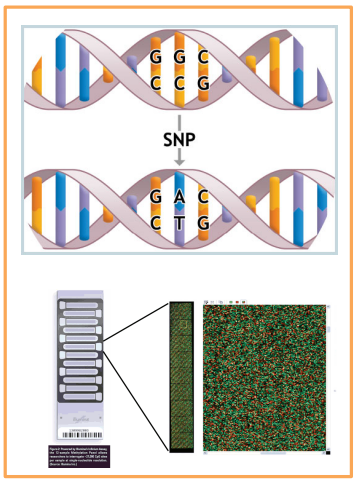


Glazier et al. 2002 Science

8 "known genes" for human complex traits before 2002

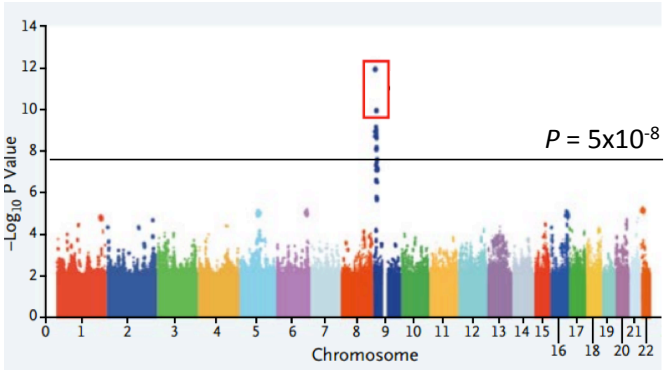
Genome-wide Association Study (GWAS)

High-throughput genotyping technology



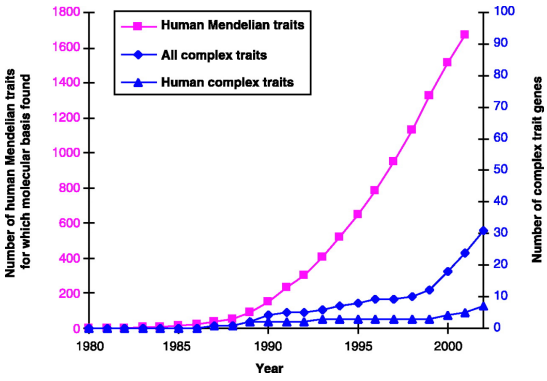
Simple methodology

Testing correlation between x and y
x = 0, 1 and 2 (GG, AG and AA)
y = trait (e.g. height)
$y = b_0 + x_1 b_1 + e$

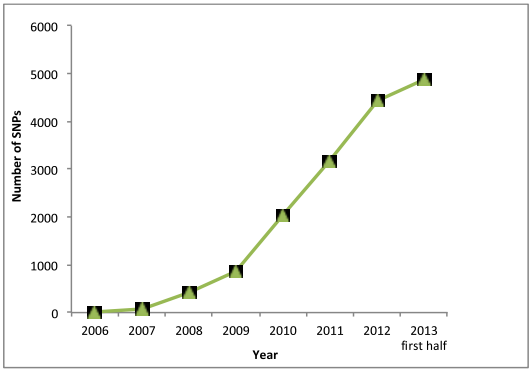


An explosion of "gene" discoveries

Prior to GWAS



GWAS



Glazier et al. 2002 Science

1000s genetic variants associated with 100s traits / diseases

The missing heritability problem

Height:

- 40 genes
- ~5% of variance explained
- heritability = ~80%

Disease	Number of loci	Percent of Heritability Measure Explained	Heritability Measure
Age-related macular degeneration	5	50%	Sibling recurrence risk
Crohn's disease	32	20%	Genetic risk (liability)
Systemic lupus erythematosus	6	15%	Sibling recurrence risk
Type 2 diabetes	18	6%	Sibling recurrence risk
HDL cholesterol	7	5.2%	Phenotypic variance
Height	40	5%	Phenotypic variance
Early onset myocardial infarction	9	2.8%	Phenotypic variance
Fasting glucose	4	1.5%	Phenotypic variance

Manolio et al. 2009 Nature

To recap the previous lectures

- Linear regression model

$y = b_0 + x_1 b_1 + x_2 b_2 + \ldots + x_p b_p + e$

y = phenotype

x_i = independent variable

$y \sim N(b_0 + x_1 b_1 + x_2 b_2 + \ldots + x_p b_p, \sigma^2_e)$

b_0 = mean term

$b_1 \ldots b_p$ = effect sizes (regression coefficients)

e = residual, $e \sim N(0, \sigma^2_e)$

Linear regression model

- In matrix form

$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$

$\mathbf{y} = \{y_j\}_{n \times 1}$; $\mathbf{X} = \{X_{ij}\}_{n \times p}$; $\mathbf{b} = \{b_i\}_{p \times 1}$; $\mathbf{e} = \{e_j\}_{n \times 1}$

- Estimation

$\mathbf{b}\text{-hat} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

$\text{var}(\mathbf{b}\text{-hat}) = \sigma^2_e(\mathbf{X}^T\mathbf{X})^{-1}$

No unique solution if p > n.

Mixed linear model (MLM)

- $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$

Fixed effects: **b** (special case: **X** = **1** and **b** = b_0)
Random effects: **u** = $\{u_i\}$, **u** ~ N(**0**, $\sigma^2_u\mathbf{A}$)
A = correlation matrix between u_i and u_j
E(**y**) = **Xb**
var(**y**) = **V** = $\mathbf{ZAZ}^T\sigma^2_u + \mathbf{I}\sigma^2_e$

If random effects are independent, then
var(**y**) = **V** = $\mathbf{ZZ}^T\sigma^2_u + \mathbf{I}\sigma^2_e$

Parameter estimation

- Estimation of variance components (σ^2_u)
  logL = $-1/2(\log|\mathbf{V}| + \log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}^T\mathbf{P}y$
  $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$

- Prediction of random effects (**u**)
  u-hat = σ^2_u-hat $\mathbf{Z}^T\mathbf{Py}$

- Estimation of fixed effects (**b**)
  b-hat = $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ (Generalized least squares)
  Linear model: **b**-hat = $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (Least squares)

GREML: fitting all SNPs in a MLM

- $\mathbf{y} = \mathbf{Zu} + \mathbf{e}$
  $\mathbf{Z} = \{z_{ij}\}_{n \times m}$, z_{ij} = standardized SNP genotype
  $\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}\sigma^2_u)$
  $\text{var}(\mathbf{y}) = \mathbf{ZZ}^T\sigma^2_u + \mathbf{I}\sigma^2_e$
  variance explained = $m\sigma^2_u / (m\sigma^2_u + \sigma^2_e)$

- An equivalent model if we let $\mathbf{g} = \mathbf{Zu}$
  $\mathbf{y} = \mathbf{g} + \mathbf{e}$
  $\text{var}(\mathbf{y}) = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$
  $\mathbf{g} \sim \mathbf{N(0, A}\sigma^2_g)$, $\mathbf{A} = \mathbf{ZZ}^T/m$ (genetic relationship matrix)
  variance explained = $\sigma^2_g / (\sigma^2_g + \sigma^2_e)$

Reconciling family studies and GWAS

Family studies: comparing phenotypic similarity to family relatedness
– GREML: ***comparing phenotypic similarity to genetic similarity (estimated from SNPs) in unrelated individuals***

GWAS: testing a SNP at a time in unrelated samples
– GREML: ***Estimating the contribution from all SNPs together***

ANALYSIS

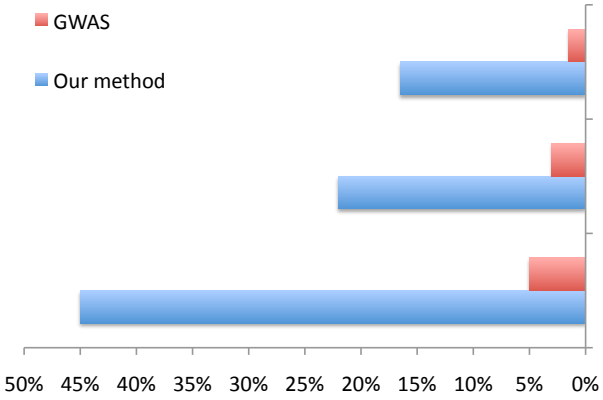
nature genetics

~50% of variation explained by all SNPs for height vs. ~10% from GWAS

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

Height is not the only example



Yang et al. 2011 Nat Genet

Lee et al. 2012 Nat Genet

Yang et al. 2010 Nat Genet

Nature vs. nurture – genetics of intelligence

Table 1 Estimates of variance explained by all SNPs

	g_c	g_f
N	3254	3181
h^2 (s.e.)	0.40 (0.11)	0.51 (0.11)
P-value	5.7×10^{-5}	1.2×10^{-7}

Davis et al. 2011 Mol Psychiatry

Deary et al. 2012 Nature

Genome partitioning

- Single component MLM

  $\mathbf{y} = \mathbf{g} + \mathbf{e}$ (or $\mathbf{y} = \mathbf{Zu} + \mathbf{e}$)

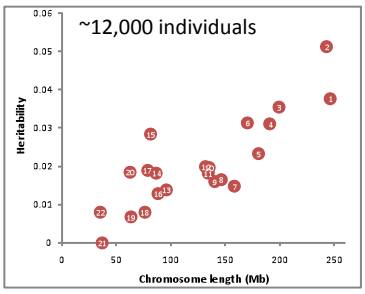
- Multi-component MLM

  $\mathbf{y} = \mathbf{g}_1 + \mathbf{g}_2 + \ldots + \mathbf{g}_{22} + \mathbf{e}$

  $\text{var}(\mathbf{y}) = \mathbf{A}_1\sigma^2_{g1} + \mathbf{A}_2\sigma^2_{g2} + \ldots + \mathbf{A}_{22}\sigma^2_{g22} + \mathbf{I}\sigma^2_e$

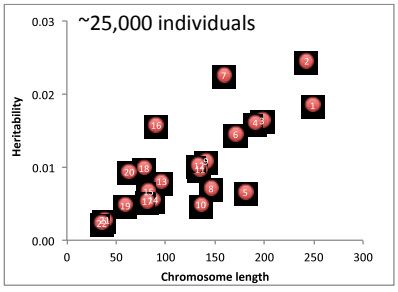
Partitioning genetic variance into chromosomes

Height



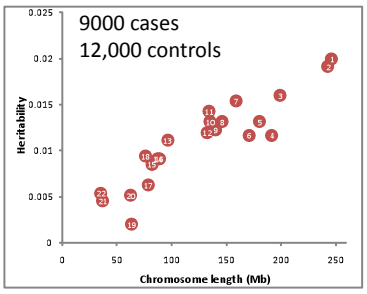
Yang et al. 2011 Nat Genet

BMI



Yang et al. unpublished

Schizophrenia



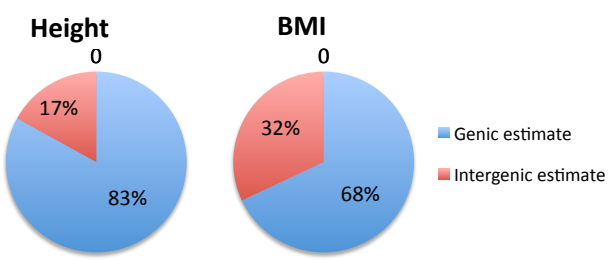
Lee et al. 2012 Nat Genet

Important implications:

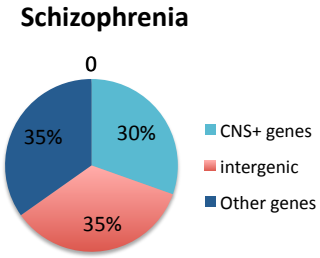
Gave confidence to continue with the GWAS paradigm

More genes for complex traits can be found with larger sample sizes

Partitioning the genetic variance based on functional annotation



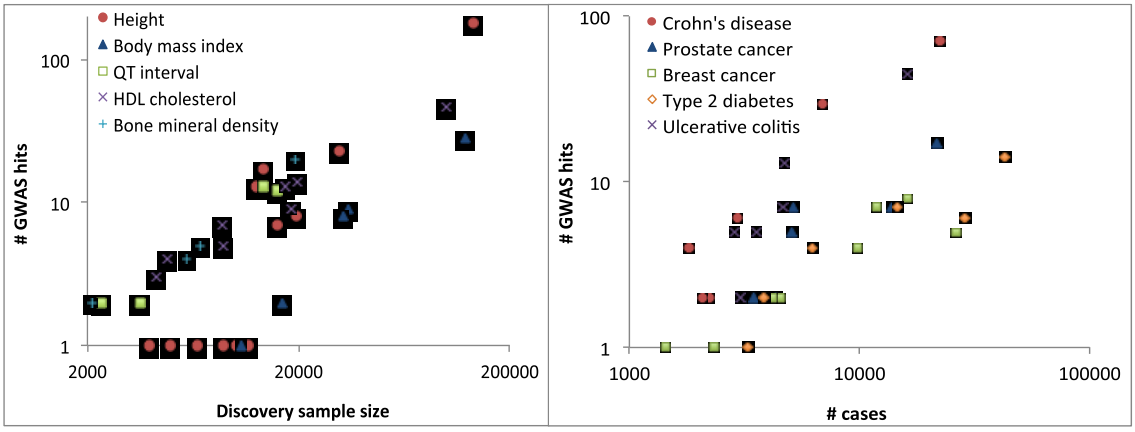
Yang et al. 2011 Nat Genet



Lee et al. 2012 Nat Genet

Genetic signals are enriched in or close to functional genes

GWAS discovery vs. sample size



Visscher et al. 2012 Am J Hum Genet

Estimation of dominance variance in unrelated individuals

- A + D model

$y = b_0 + x_a*b_1 + x_d*b_2 + e$

$b_1 = a; b_2 = d$

Genotype	E(y)	x_a	x_d
AA	mean	0	0
AG	mean + a + d	1	1
GG	mean + 2a	2	0

- Additive model

$y = b_0 + x_a*b_1 + e$

$b_1 = a + (1 – 2p)d$

Lynch and Walsh 1996

Estimation of dominance variance in unrelated individuals

- Dominance model

Genotype	x_a	x_d	x_d'
AA	0	0	0
AG	1	1	2p
GG	2	0	(4p – 2)

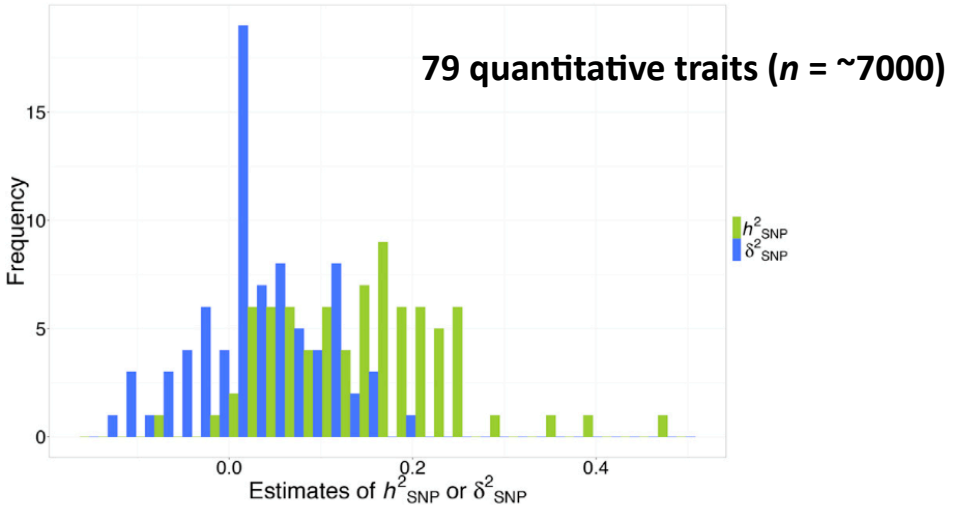
$\mathbf{y} = \mathbf{g_a} + \mathbf{g_d} + \mathbf{e}$

$var(\mathbf{y}) = \mathbf{A}_a\sigma^2_a + \mathbf{A}_d\sigma^2_d + \mathbf{I}\sigma^2_e$

$\mathbf{A}_a$ = additive GRM; $\mathbf{A}_d$ = dominance GRM

$\mathbf{A}_d = \mathbf{Z}_d^T\mathbf{Z}_d / m$, where $\mathbf{Z}_d$ = standardised x'_d matrix

Estimating dominance variation in unrelated individuals



Mean h^2_{SNP} = 0.15
Mean δ^2_{SNP} = 0.03

Zhu et al. 2015 Am J Hum Genet

Bivariate GREML analysis to estimate genetic correlation in unrelated individuals

$\mathbf{y}_1 = \mathbf{X}_1\mathbf{b}_1 + \mathbf{g}_1 + \mathbf{e}_1$
$\mathbf{y}_2 = \mathbf{X}_2\mathbf{b}_2 + \mathbf{g}_2 + \mathbf{e}_2$

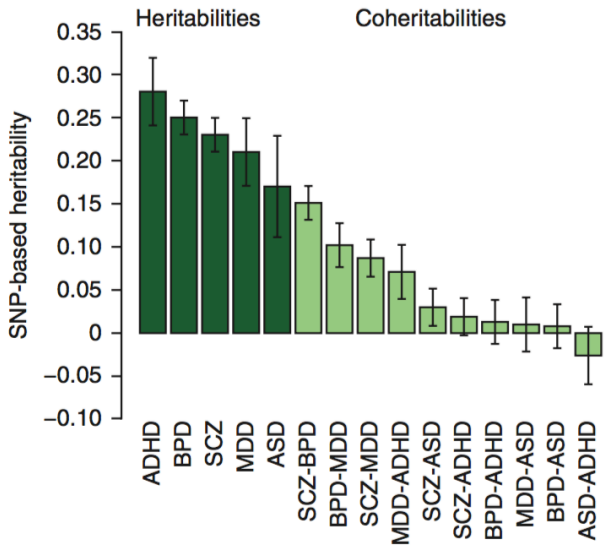
$$\mathbf{V} = \mathrm{var}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} A_1\sigma^2_{g1} + I\sigma^2_{e1} & A_{12}\sigma_{g1g2} + I\sigma_{e1e2} \\ A_{12}\sigma_{g1g2} + I\sigma_{e1e2} & A_2\sigma^2_{g2} + I\sigma^2_{e2} \end{bmatrix}$$

For traits measures on different samples

$$\mathbf{V} = \mathrm{var}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} A_1\sigma^2_{g1} + I\sigma^2_{e1} & A_{12}\sigma_{g1g2} \\ A_{12}\sigma_{g1g2} & A_2\sigma^2_{g2} + I\sigma^2_{e2} \end{bmatrix}$$

Lee et al. 2012 Bioinformatics

Estimating genetic correlation between traits measured on different samples



Lee et al. 2013 Nat Genet

Questions and discussion

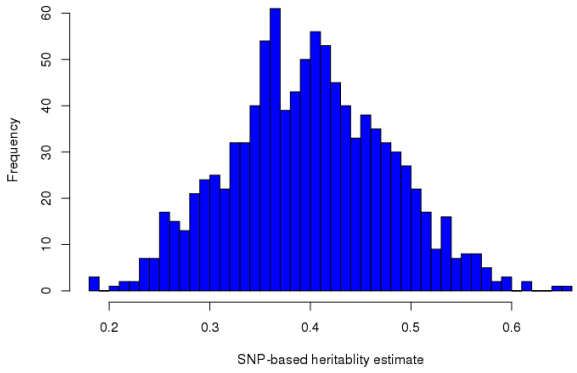
Lecture #4

Application of GREML and related methods to GWAS data with GCTA

Jian Yang
Institute for Molecular Bioscience
The University of Queensland

SE of GREML estimate

- SE = sd of estimate



- Sampling variance of GREML estimate

= SE^2

= ~ 2 / [N * var(GRM)] Visscher et al. 2014 PLoS Genet

- var(GRM) is proportional to 1 / M_e where M_e is the effective number of independent markers.

A frequently asked question:

how many individuals are required to run a GCTA-GREML analysis

- SE^2 = ~ 2 / [N^2 * var(GRM)]

- For analysis in unrelated individuals with HapMap3 SNP (~1M), var(GRM) = ~2e-5, so SE = ~ 316 / N.

- For the analysis with whole-genome sequence data, var(GRM) = ~4.5e-6, so SE = ~667 / N.

SE of the estimate from bivariate GCTA-GREML

- Depending on more parameters (estimates)
- For traits measured on the same sample

$$\mathrm{var}(\hat{r}_G) \approx \frac{(1-r_G r_P)^2 + (r_G - r_P)^2}{h^2_{G1} h^2_{G2} N^2 \,\mathrm{var}(A_{ij})} \qquad (7)$$

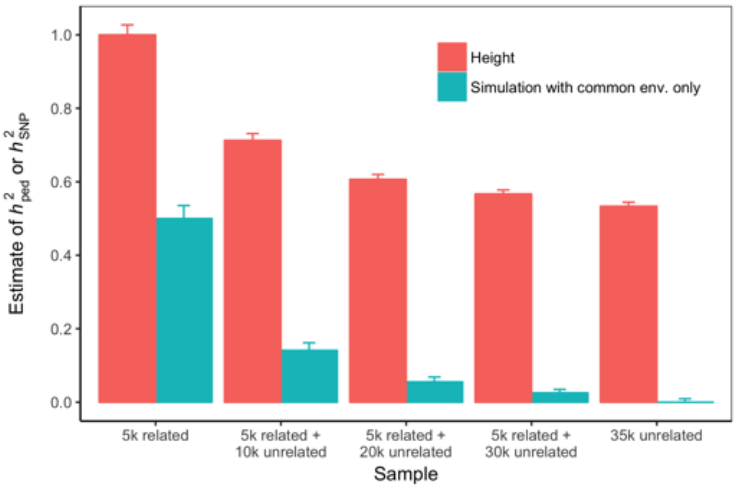
- For traits measured on different samples

$$\mathrm{var}(\hat{r}_G) \approx \frac{r^2_G (N_1^2 h^4_{G1} + N_2^2 h^4_{G2}) + 2h^2_{G1} h^2_{G2} N_1 N_2}{2h^4_{G1} h^4_{G2} N_1^2 N_2^2 \,\mathrm{var}(A_{ij})} \qquad (10)$$

Visscher et al. 2014 PLoS Genet

GCTA-GREML analysis in family data

- The confounding of real genetic effects with common environmental effects shared between relatives



Yang et al. 2017 under review

GCTA-GREML analysis in family data

- Solution #1: remove close relatives
- Solution #2: estimate SNP-based and pedigree-based heritability simultaneously

$\mathbf{y} = \mathbf{g_1} + \mathbf{g_2} + \mathbf{e}$

$\text{var}(\mathbf{y}) = \mathbf{A}_1\sigma^2_1 + \mathbf{A}_2\sigma^2_2 + \mathbf{I}\sigma^2_e$

$\mathbf{A}_1$ = GRM

$\mathbf{A}_2$ = GRM with large relatedness values only

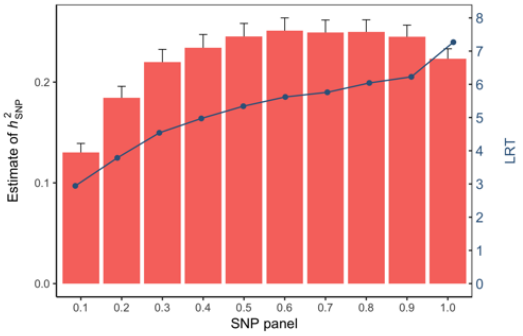
σ^2_1 / σ^2_P = SNP-based heritability

$(\sigma^2_1 + \sigma^2_2) / \sigma^2_P$ = pedigree-based heritability

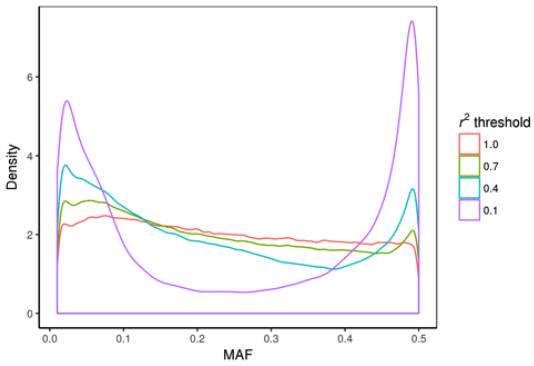
Zaitlen et al. 2013 PLoS Genet

SNPs need to be pruned for LD?

- Estimate increases with the decrease of LD pruning (PLINK threshold but LRT does not



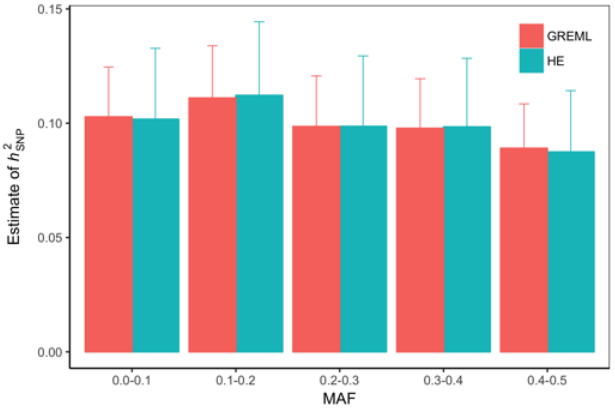
- LD pruning changes the MAF spectrum (cautious about the interpretation of the estimate)



Yang et al. 2017 under review

Large sample size

- Computational challenge when n > 100,000
- Really necessary to run a GREML analysis with n > 100K?



- HE regression?

$y_i y_j \sim b_0 + b_1 * A_{ij} + e$

$b_0 = V_g$ if y_i and y_j are standardised

Yang et al. 2017 under review

Demo

- Simulating phenotypes based on a real GWAS data set in GCTA
- Creating the genetic relatedness matrix using all SNPs (by-product: PCA analysis).
- GCTA-GREML analysis to estimate the SNP-based heritability
- Bivariate GREML analysis to estimate the genetic correlation between traits

Script 1

```
# Randomly sample 5 SNPs as causal variants
bim = read.table("test.bim", colClasses=c(rep("character",6)))
qtl = sample(bim$V2, 5)
write.table(qtl, "test.qtl", row.names=F, col.names=F, sep="\t", quote=F)

# Generate phenotype
gcta64 --bfile test --simu-qt --simu-causal-loci test.qtl --simu-hsq 0.1 --out test

# Compute GRM
gcta64 --bfile test --make-grm --out test --thread-num 30

# REML analysis
gcta64 --grm test --pheno test.phen --reml --out test --thread-num 30

# PCA analysis - by product
gcta64 --grm test --pca --out test --thread-num 30

pc = read.table("test.eigenvec")
plot(pc$V3, pc$V4, xlab="PC1", ylab="PC2", col="red")
```

Script 2

```
# Simulate two traits
bim = read.table("test.bim", colClasses=c(rep("character",6)))
qtl_comm = sample(bim$V2, 5)
tmp = bim$V2[which(is.na(match(bim$V2, qtl_comm)))]
qtl1 = c(qtl_comm, sample(tmp,5))
qtl2 = c(qtl_comm, sample(tmp,5))
write.table(qtl1, "test.qtl1", row.names=F, col.names=F, sep="\t", quote=F)
write.table(qtl2, "test.qtl2", row.names=F, col.names=F, sep="\t", quote=F)

gcta64 --bfile test --simu-qt --simu-causal-loci test.qtl1 --simu-hsq 0.1 --out test_tr1
gcta64 --bfile test --simu-qt --simu-causal-loci test.qtl2 --simu-hsq 0.2 --out test_tr2

tr1=read.table("test_tr1.phen")
tr2=read.table("test_tr2.phen")
tr2=tr2[match(tr1$V2, tr2$V2),]
tr=cbind(tr1,tr2$V3)
write.table(tr,"test_2tr.phen",row.names=F,col.names=F,quote=F,sep="\t")

# Bivariate GREML
gcta64 --grm test --pheno test_2tr.phen --reml-bivar 1 2 --out test_2tr --thread-num 50
```

Questions and discussion

Lecture #5

MLM based association analysis

Jian Yang
Institute for Molecular Bioscience
The University of Queensland

To recap

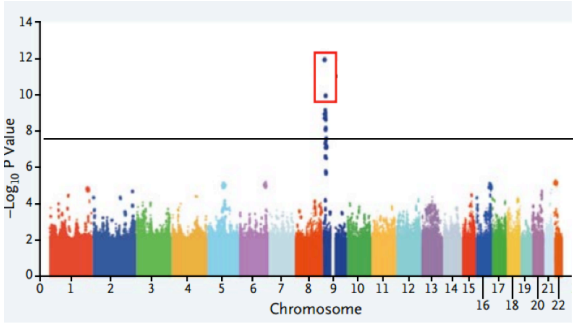
- Linear regression

  $y = b_0 + x_1 b_1 + e$

  y = trait value; x_1 = SNP genotype (0, 1 or 2)

  b_1-hat = $\mathbf{X}_1^T\mathbf{y} / (\mathbf{X}_1^T\mathbf{X}_1) = cov(x_1,y) / var(x_1)$

  SE(b_1-hat) = $\sigma^2_e / [n\ var(x_1)]$



Manolio 2008 NEJM

Inflated test-statistics due to population structure

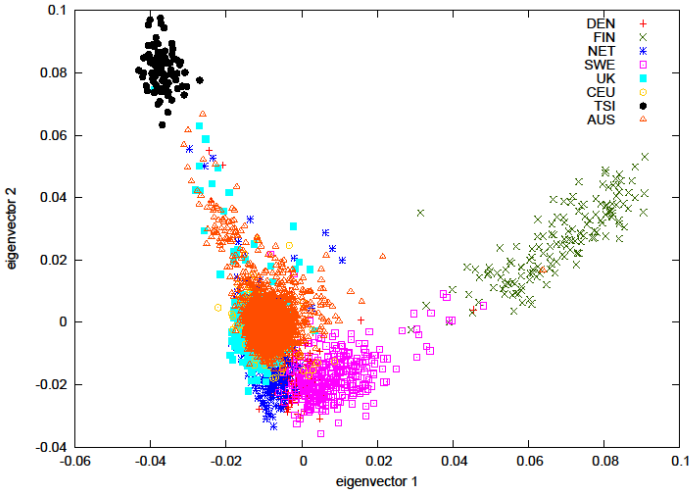
- Assumption underlying Linear regression: e_i and e_j are independent and identically distributed.

  $y_i = b_0 + x_i b_1 + e_i$

- Two issues:
  - Population stratification
  - Cryptic relatedness

- Solutions:
  - Fitting PCs (Price et al. 2007 Nat Genet)
  - Genomic control (Devlin & Roeder 1999 Biometrics)

Population stratification inferred from SNP data

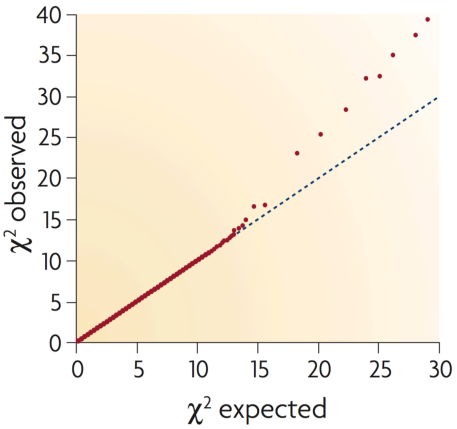


Yang et al. 2010 Nat Genet

Problem: PCs are unable to capture relatedness

Genomic control

- Chi-squared statistics / λ
  λ = mean chi-squared
  λ = median(chi-squared) / 0.455



McCarthy et al. 2008 Nat Rev Genet

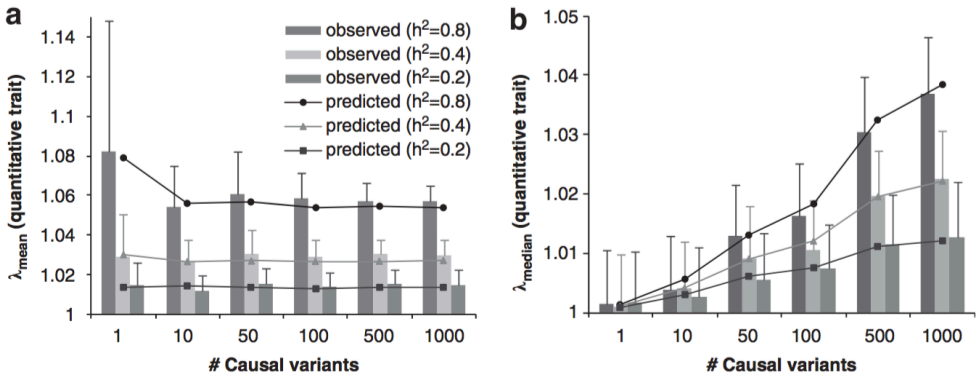
- Assumption: only a few true signals. If there is no inflation, the expected value

- Genomic inflation is expected under polygenic model
  E(chi-squared) = 1 + NCP = 1 + n * h^2 / m

Yang et al. 2011 EJHG

Genomic inflation factors under polygenic model

- λ_{median} ~ h^2 and number of causal variants



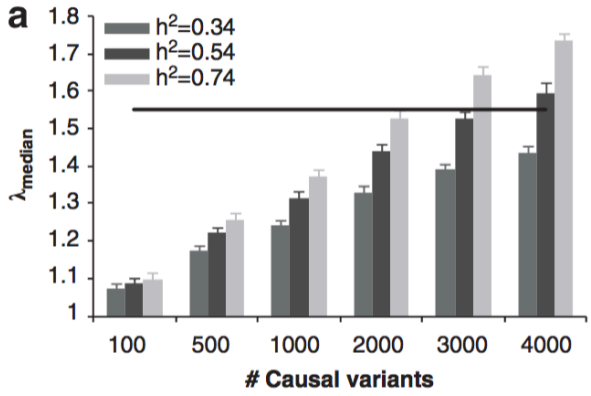
Yang et al. 2011 EJHG

Estimating heritability and #causal from genomic inflation factor

- h^2: $\lambda_{\text{mean}}^{\text{QT}} \approx 1 + \frac{N h^2 \overline{r^2 s}}{n}$

  (LD score regression is a more elegant solution)

- #causal



Yang et al. 2011 EJHG

MLM based association analysis

- $\mathbf{y} = \mathbf{x}b + \mathbf{Zu} + \mathbf{e}$ or $\mathbf{y} = \mathbf{x}b + \mathbf{g} + \mathbf{e}$

  $\mathbf{V} = \text{var}(\mathbf{y}) = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$

- Testing for fixed effects given sample structure

  b-hat = $(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{V}^{-1}\mathbf{y}$

  var(b-hat) = $\sigma^2_e(\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x})^{-1}$ Kang et al. 2010 Nat Genet

- Issue: a SNP is fitted twice (MLMi: MLM association including the target SNP in GRM).

MLMe: MLM association excluding the target SNP from the GRM

- Expected chi-squared values

$$\lambda_{\text{mean}}(\text{LR}) = 1 + Nh_g^2 / M$$

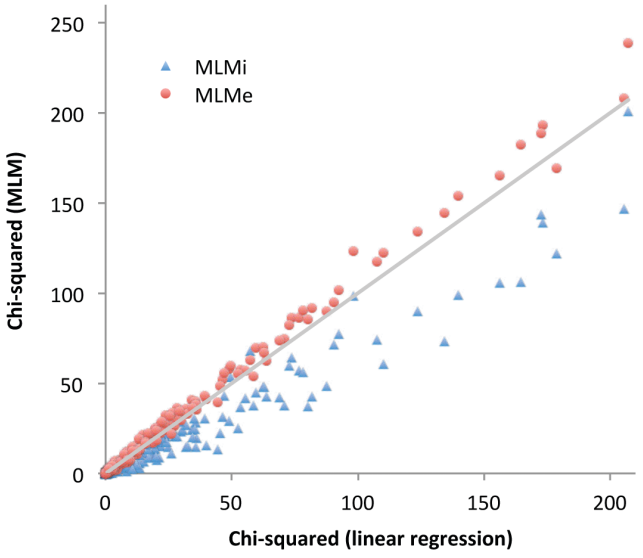
$$\lambda_{\text{mean}}(\text{MLMi}) = 1$$ Deflation: E(chi-squared) < 1 for null SNPs

$$\lambda_{\text{mean}}(\text{MLMe}) = 1 + \frac{Nh_g^2 / M}{1 - r^2 h_g^2}$$

	Linear regression	MLMi	MLMe	MLMe / MLMi
Causal markers (M_q)	$1 + Nh_g^2 / M_q$	$\frac{Nh_g^2 / M_q + 1 - r^2 h_g^2}{Nh_g^2 / M + 1 - r^2 h_g^2}$	$1 + \frac{Nh_g^2 / M_q}{1 - r^2 h_g^2}$	$1 + \frac{Nh_g^2 / M}{1 - r^2 h_g^2}$
Null markers (M–M_q)	1	$\frac{1 - r^2 h_g^2}{Nh_g^2 / M + 1 - r^2 h_g^2}$	1	$1 + \frac{Nh_g^2 / M}{1 - r^2 h_g^2}$
All markers (M)	$1 + Nh_g^2 / M$	1	$1 + \frac{Nh_g^2 / M}{1 - r^2 h_g^2}$	$1 + \frac{Nh_g^2 / M}{1 - r^2 h_g^2}$

Yang et al. 2014 Nat Genet

Power comparison at causal variants



Selection of SNPs to compute the GRM

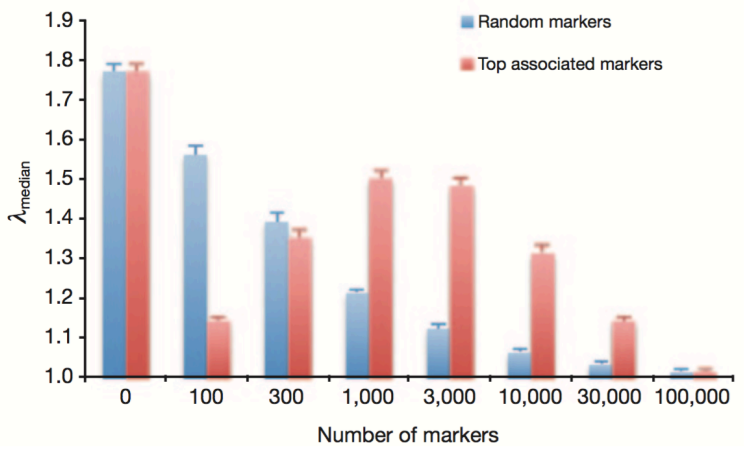


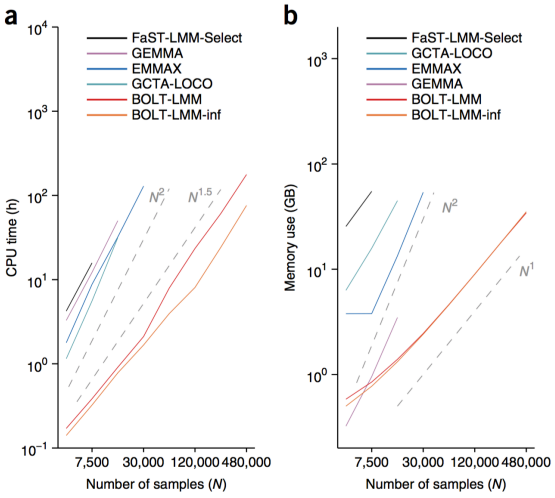
Figure 2 Effectiveness of mixed linear models using random or top associated markers in correcting for stratification. We report average λ_{median} (± s.e.m.) in 100 simulations with population stratification based on N = 10,000 samples, M = 100,000 markers, 2 discrete subpopulations with fixation index (F_{ST}) = 0.005 and a mean trait difference of 0.25 s.d. between subpopulations. Calibration of small P values is reported in **Supplementary Table 4**.

Yang et al. 2014 Nat Genet

Computational challenge

# samples (N)	# markers (M)	GCTA-MLMi	GCTA-LOCO
5,000	50,000	0.3hr / 2.0GB	1.0hr / 4.1GB
5,000	100,000	0.6hr / 3.0GB	1.4hr / 5.2GB
10,000	50,000	1.3hr / 5.9GB	7.2hr / 14.3GB
10,000	100,000	2.5hr / 7.9GB	8.2hr / 16.3GB

Yang et al. 2014 Nat Genet



Loh et al. 2015 Nat Genet

BOLT-LMM

- Computationally efficient when the number of SNPs is not large.

- It uses a cross validation approach prediction approach to specify models (infinitesimal model vs. mixture normal model).

- Leave-one-chromosome-out analysis as the default.

Computer practical

- Simulating phenotypes based on a real GWAS data set in GCTA

- Linear regression analysis in PLINK

- GCTA-MLMA analysis (MLMi)

- GCTA-MLMA-LOCO

- BOLT-LMM

Questions and discussion

Lecture 6:
BLUP and Genomic Prediction
1: BLUP

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
9 – 10 Feb 2017

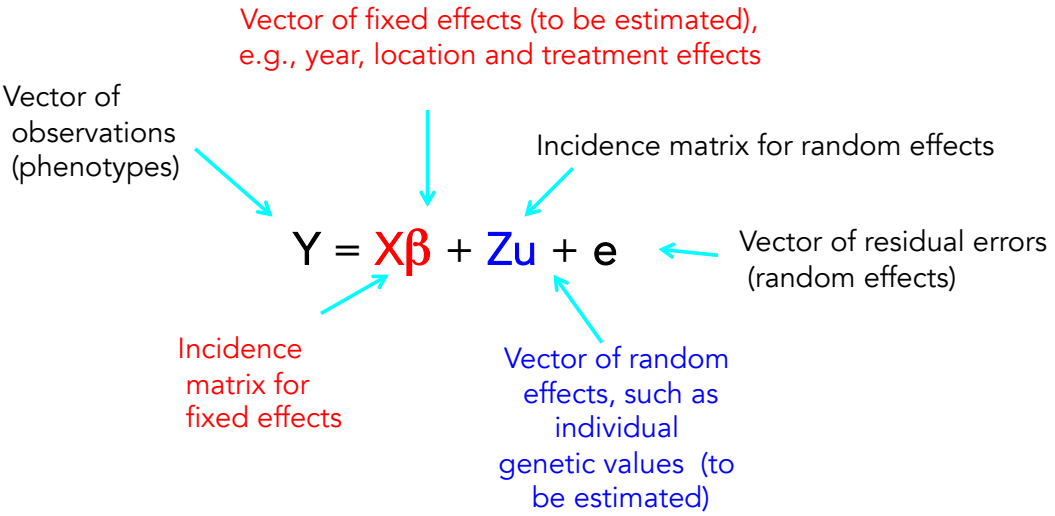
Estimation of Var(A) and Breeding Values in General Pedigrees

The classic designs (ANOVA, P-O regression) for variance components are simple, involving only a single type of relative comparison. Further, they assume balanced designs, with the number of offspring the same in each family.

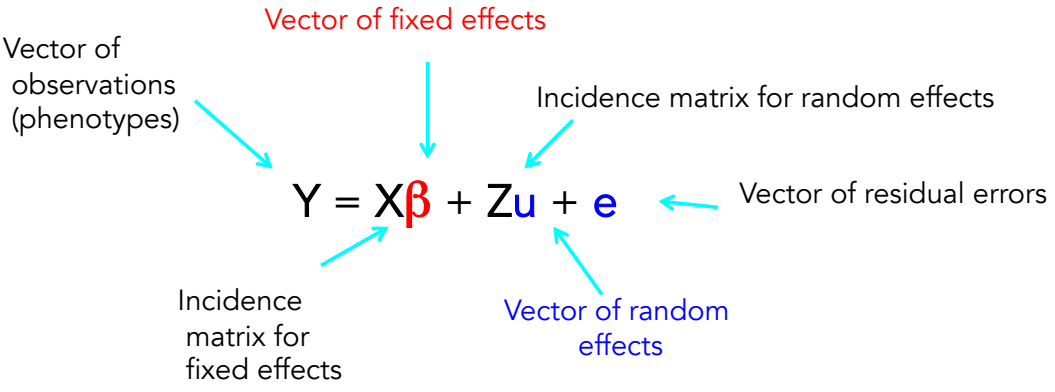
In the real world, we often have a pedigree of relatives, with a very unbalanced design. Fortunately, the general mixed model (so called because it includes both fixed and random effects), offers an ideal platform for both estimating genetic variances as well a predicting the breeding values of individuals.

Almost all animal breeding is based on such models, with REML (restricted max likelihood) used to estimated variances and BLUP (best linear unbiased predictors) used to predict BV

The general mixed model



The general mixed model



Observe **y, X, Z**.

Estimate fixed effects **β**

Estimate random effects **u, e**

Example

Suppose we wish to estimate the breeding values of three sires (fathers), each of which is mated to a random female (dam), producing two offspring, some reared in environment one, others in environment two. The data are

Observation	Value	Sire	environment
Y_{111}	9	1	1
Y_{121}	12	1	2
Y_{211}	11	2	1
Y_{212}	6	2	1
Y_{311}	7	3	1
Y_{321}	14	3	2

Here the basic model is

$$Y_{ijk} = \beta_j + u_i + e_{ijk}$$

Effect of environment j

Breeding value of sire i

The mixed model vectors and matrices become

$$\mathbf{y} = \begin{pmatrix} y_{1,1,1} \\ y_{1,2,1} \\ y_{2,1,1} \\ y_{2,1,2} \\ y_{3,1,1} \\ y_{3,2,1} \end{pmatrix} = \begin{pmatrix} 9 \\ 12 \\ 11 \\ 6 \\ 7 \\ 14 \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

Means & Variances for $y = X\beta + Zu + e$

Means: $E(u) = E(e) = 0$, $E(y) = X\beta$

Variances:

Let R be the covariance matrix for the residuals. We typically assume $R = \sigma^2_e * I$

Let G be the covariance matrix for the breeding values (the vector **u**)

The covariance matrix for y becomes

$V = ZGZ^T + R$

Estimating fixed Effects & Predicting Random Effects

For a mixed model, we observe **y**, X, and Z

β, u, R, and G are generally unknown

Two complementary estimation issues

(i) Estimation of **β** and **u**

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$ Estimation of fixed effects

BLUE = Best Linear Unbiased Estimator

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$ Prediction of random effects

BLUP = Best Linear Unbiased Predictor

Recall $V = ZGZ^T + R$

Henderson's Mixed Model Equations

$y = X\beta + Zu + e$, $u \sim (0,G)$, $e \sim (0, R)$, $\text{cov}(u,e) = 0$,

If X is n x p and Z is n x q

p x p p x q

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

q x p q x q

The whole matrix is (p+q) x (p+q)

Easier to numerically work with than BLUP/BLUE equations

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

$V = ZGZ^T + R$

Inversion of an n x n matrix

Standard Errors

A relatively straightforward extension of Henderson's mixed-model equations provides estimates of the standard errors of the fixed and random effects. Let the inverse of the leftmost matrix in Equation 26.5 be

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \quad (26.6)$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{12}$, and $\mathbf{C}_{22}$ are, respectively, $p \times p$, $p \times q$, and $q \times q$ submatrices. Using this notation, Henderson (1975) showed that the sampling covariance matrix for the BLUE of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\sigma}(\hat{\boldsymbol{\beta}}) = \mathbf{C}_{11} \quad (26.7a)$$

that the sampling covariance matrix of the prediction errors $(\hat{\mathbf{u}} - \mathbf{u})$ is given by

$$\boldsymbol{\sigma}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{22} \quad (26.7b)$$

and that the sampling covariance of estimated effects and prediction errors is given by

$$\boldsymbol{\sigma}(\hat{\boldsymbol{\beta}}, \hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{12} \quad (26.7c)$$

(We consider $\hat{\mathbf{u}} - \mathbf{u}$ rather than $\hat{\mathbf{u}}$ as the latter includes variance from both the prediction error and the random effects $\mathbf{u}$ themselves.)

Let's redo our example on slide 6 using Henderson's Equation

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} = \frac{1}{6}\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} = \left(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\right)^T = \frac{1}{6}\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$\mathbf{G}^{-1}+\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} = \frac{5}{6}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 33 \\ 26 \end{pmatrix}, \quad \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 21 \\ 17 \\ 21 \end{pmatrix}$$

The MM equations become

$$\begin{pmatrix} 4 & 0 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 & 1 \\ 1 & 1 & 5 & 0 & 0 \\ 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \begin{pmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{pmatrix}$$

Taking the inverse gives

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \frac{1}{18}\begin{pmatrix} 148 \\ 235 \\ -1 \\ 2 \\ -1 \end{pmatrix}$$

As found above

The Animal Model, $y_i = \mu + a_i + e_i$

Here, the individual is the unit of analysis, with y_i the phenotypic value of the individual and a_i its BV

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \mu, \quad \mathbf{u} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} \quad \mathbf{G} = \sigma_A^2\,\mathbf{A},$$

Where the additive genetic relationship matrix A is given by $A_{ij} = 2\theta_{ij}$, namely twice the coefficient of coancestry

Assume $R = \sigma^2_e * I$, so that $R^{-1} = 1/(\sigma^2_e) * I$.
Likewise, $G = \sigma^2_A * A$, so that $G^{-1} = 1/(\sigma^2_A) * A^{-1}$.

The "animal" model estimates the breeding value for each individual, even for a plant or tree! Same approach also works to estimate line (genotypic) values for inbreds.

Returning to the animal model

Henderson's mixed model equations

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z}+\lambda\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

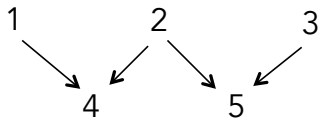
here $\lambda = \sigma^2_e / \sigma^2_A = (1-h^2)/h^2$

This reduces to

$$\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I}+\lambda\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \sum^n y_i \\ \mathbf{y} \end{pmatrix}$$

Example

Suppose our pedigree is



$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 1 & 1/4 \\ 0 & 1/2 & 1/2 & 1/4 & 1 \end{pmatrix}$$

Suppose $\lambda = 1$ (corresponds to $h^2 = 0.5$). In this case,

$$\mathbf{I}+\lambda\mathbf{A}^{-1} = \begin{pmatrix} 5/2 & 1/2 & 0 & -1 & 0 \\ 1/2 & 3 & 1/2 & -1 & -1 \\ 0 & 1/2 & 5/2 & 0 & -1 \\ -1 & -1 & 0 & 3 & 0 \\ 0 & -1 & -1 & 0 & 3 \end{pmatrix}$$

Suppose the vector of observations is

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{pmatrix}$$

Here n = 5, Σ y = 41, and Henderson's equation becomes

$$\begin{pmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 5/2 & 1/2 & 0 & -1 & 0 \\ 1 & 1/2 & 3 & 1/2 & -1 & -1 \\ 1 & 0 & 1/2 & 5/2 & 0 & -1 \\ 1 & -1 & -1 & 0 & 3 & 0 \\ 1 & 0 & -1 & -1 & 0 & 3 \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ \widehat{a}_5 \end{pmatrix} = \begin{pmatrix} 41 \\ 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{pmatrix}$$

Solving gives

$$\widehat{\mu} = \frac{440}{53} \simeq 8.302, \qquad \begin{pmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} -662/689 \\ 4/53 \\ 610/689 \\ -732/689 \\ 381/689 \end{pmatrix} \simeq \begin{pmatrix} -0.961 \\ 0.076 \\ 0.885 \\ -1.062 \\ 0.553 \end{pmatrix}$$

More on the animal model

- Under the animal model
  - $\mathbf{y} = \mathbf{X\beta} + \mathbf{Za} + \mathbf{e}$
  - $\mathbf{a} \sim (0, \sigma_A^2\mathbf{A})$, $\mathbf{e} \sim (0, \sigma_e^2\mathbf{I})$
  - BLUP(**a**) = $\sigma_A^2\mathbf{AZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})$
  - Where $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R} = \sigma_A^2\mathbf{ZAZ}^T + \sigma_e^2\mathbf{I}$
- Consider the simplest case of a single observation on one individual, where the only fixed effect is the mean m, which is assumed known
  - Here $\mathbf{Z} = \mathbf{A} = \mathbf{I} = (1)$,
  - $\mathbf{V} = \sigma_A^2 + \sigma_e^2$
  - $\sigma_A^2\mathbf{AZ}^T\mathbf{V}^{-1} = \sigma_A^2/(\sigma_A^2 + \sigma_e^2) = h^2$
  - BLUP(a) = $h^2(y-\mu)$

- More generally, with single observations on n unrelated individuals,
  - $A = Z = I_{n \times n}$
  - $V = \sigma_A^2 ZAZ^T + \sigma_e^2 I = (\sigma_A^2 + \sigma_e^2) I$
  - $\sigma_A^2 AZ^T V^{-1} = h^2 I$
  - $BLUP(a) = \sigma_A^2 A Z^T V^{-1} (y - X\beta) = h^2 (y - \mu)$
- Hence, the predicted breeding value of individual i is just $BLUP(a_i) = h^2(y_i - \mu)$
- When at least some individuals are related and/or inbred (so that $A \neq I$) and/or missing or multiple records (so that $Z \neq I$), then the estimates of the BV differ from this simple form, but BLUP fully accounts for this

BLUP is a shrinkage estimator

- For a single observation on one individual, $BLUP(a) = h^2(y - \mu)$
  - The difference between the observed value (y) and the mean (μ) is shrunk by the factor h^2 --- shrinks the estimate back towards the mean (zero in the case of BVs)
- More generally, $BLUP(a) = G Z^T V^{-1} (y - X\beta)$
  - First adjusts observations (y) for fixed effects ($X\beta$) and then regresses this difference back towards zero (the mean BV), as $Cov * Var^{-1}$ is a generalized regression coefficient

The Relationship Matrix A

- Typically given from a pedigree, but increasingly being estimated from marker data
- The diagonal elements indicate the amount of inbreeding
  - $A_{ii} = 1 + F_i$, where F_i is inbreeding coefficent for individual i.
  - For a fully-inbred, $A_{ii} = 2$

Marker-based relationship matrices

- There are two reasons for using a marker-estimated relationship matrix
  - Pedigree either unknown or poorly known
  - With very dense markers, provides a better estimate than a known pedigree. Why?
    - Consider two (non-inbred) full-sibs. The expectation under a pedigree is that they share exactly half their genes.
    - However, there is a sampling variance about this expected value, so that some pair of sibs may share more than 50%, while another may share less. Using markers to detect such pairs improves the estimated values
    - This is called G-BLUP (in animal breeding) and is a form of genomic selection

Marker-based relationship matrix

Simplest case is to consider a very large number (L) of SNPs, and treat alike in state as IBD, and then compute the probability f_{xy} that x and y share a randomly-drawn allele for each SNP marker. Twice the average over all markers is the entry for x and y in the relationship matrix (as $A_{xy} = 2f_{xy}$)

SNP genotype for x

SNP genotype for y

	00	01	11
00	1	0.5	0
01	0.5	0.5	0.5
11	0	0.5	1

Values for f_{xy} given the SNP genotypes

Estimation of R and G

A second estimation issue concerns the covariance matrix for residuals R and for breeding values G

As we have seen, both matrices have the form σ^2*B, where the variance σ^2 is unknown, but B is known

For example, for residuals, $R = \sigma^2_e$*I

For breeding values, $G = \sigma^2_A$*A, where A is given from the pedigree

REML Variance Component Estimation

REML = Restricted Maximum Likelihood.

Standard ML variance estimation assumes fixed factors are known without error. Results in downward bias in variance estimates

REML maximizes that portion of the likelihood that does not depend on fixed effects

Basic idea: Use a transformation to remove fixed effects, then perform ML on this transformed vector

Simple variance estimate under ML vs. REML

$$\text{ML} = \frac{1}{n}\sum_{i+1}^{n}(x-\overline{x})^2, \quad \text{REML} = \frac{1}{n-1}\sum_{i+1}^{n}(x-\overline{x})^2$$

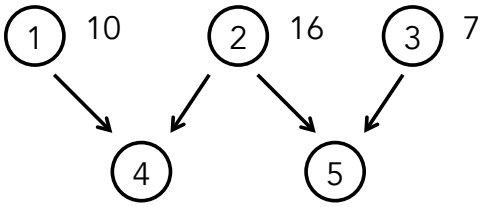
REML adjusts for the estimated fixed effect, in this case, the mean

With balanced design, ANOVA variance estimates are equivalent to REML variance estimates

Example

Suppose individuals 1 - 3 are measured, 4 & 5 are not.

Assume only a single fixed effect, the mean μ.



$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 1 & 1/4 \\ 0 & 1/2 & 1/2 & 1/4 & 1 \end{pmatrix}$$

Model becomes

$$\begin{pmatrix} 10 \\ 16 \\ 8 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix}$$

Here

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Letting Var(A) = 100, Var(e) = 100, $V = ZGZ^T + R = 200 * I$

Solving gives

$$\hat{\mathbf{a}} = \begin{pmatrix} -0.50 \\ 2.50 \\ -2.00 \\ 1.00 \\ 0.25 \end{pmatrix}$$

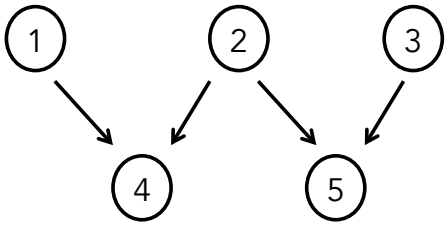
Average base pop EBVs = 0

EBVs for individuals (4,5) with no phenotypic records

Key: Information from relatives provides estimates for BV of unmeasured relatives.

G-BLUP

Suppose we have maker data.
How does this change EBVs?



Suppose marker data gives A as

2 slightly inbred

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0.5 & 0 \\ 0 & 1.2 & 0 & 0.5 & 0.5 \\ 0 & 0 & 1 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 1 & 0.2 \\ 0 & 0.5 & 0.5 & 0.2 & 1 \end{pmatrix}$$

4 & 5 slightly less related than 1/2 sibs

G-BLUP

$$\hat{\mathbf{a}} = \begin{pmatrix} -0.69 \\ 2.62 \\ -1.59 \\ 0.80 \\ 0.30 \end{pmatrix}$$

Pedigree-BLUP

$$\hat{\mathbf{a}} = \begin{pmatrix} -0.50 \\ 2.50 \\ -2.00 \\ 1.00 \\ 0.25 \end{pmatrix}$$

Lecture #6B

Genomic risk prediction

Jian Yang
Institute for Molecular Bioscience
The University of Queensland

Conceptual difference between estimation and prediction

- Estimation: only a few parameters (e.g. σ^2_g and σ^2_e) are required to be estimated. Estimate !~ sample; SE ~ sample size.

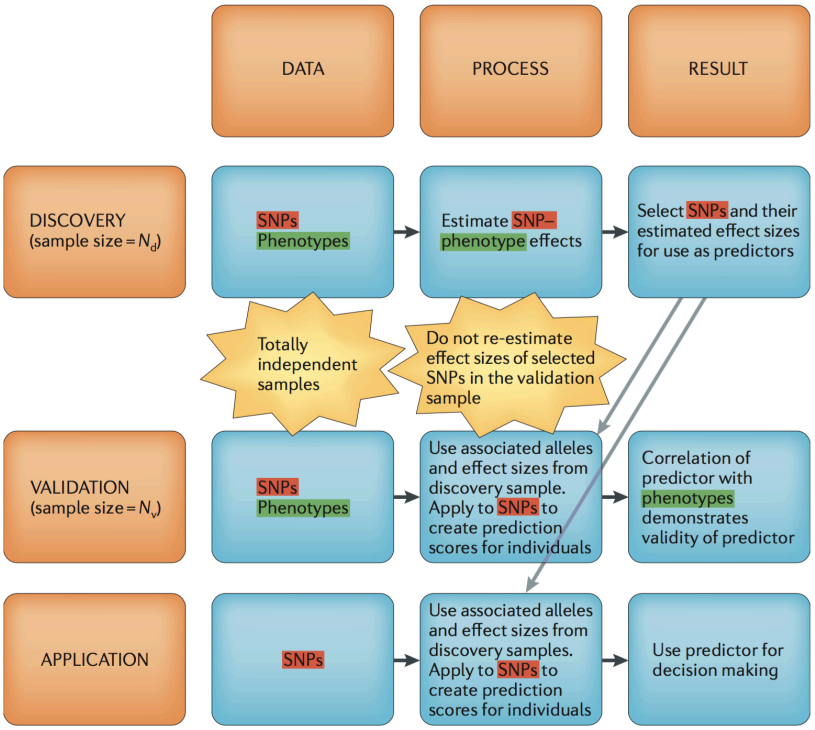
  $\mathbf{y} = \mathbf{g} + \mathbf{e}$

  $\text{var}(\mathbf{y}) = \mathbf{A}\sigma^2_g + \mathbf{I}\sigma^2_e$

- Prediction: all the SNP effects need to be estimated with little errors. Prediction accuracy ~ discovery sample size; SE ~ validation sample size.

  $\mathbf{y} = \mathbf{Zu} + \mathbf{e}$

Genetic risk prediction



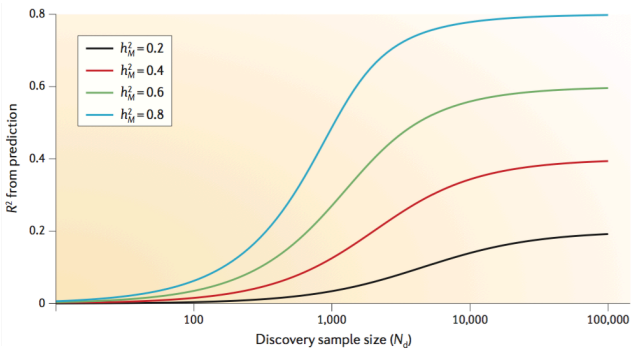
Heritability is the upper limit

- The accuracy of GBLUP prediction depends on SNP-based heritability and M / N_d ratio

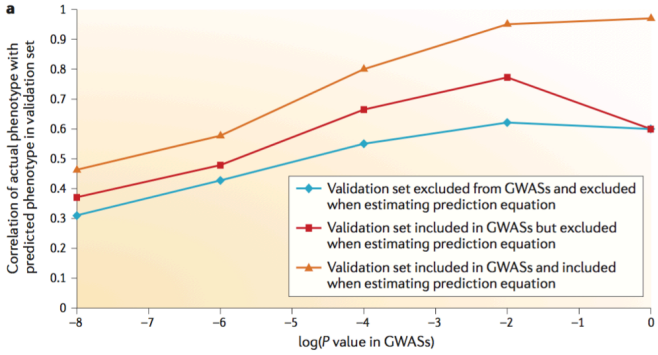
$$R^2 = \frac{h^2_M}{1 + \frac{M}{N_d h^2_M}(1 - R^2).}$$

Wray et al. 2013 Nat Rev Genet

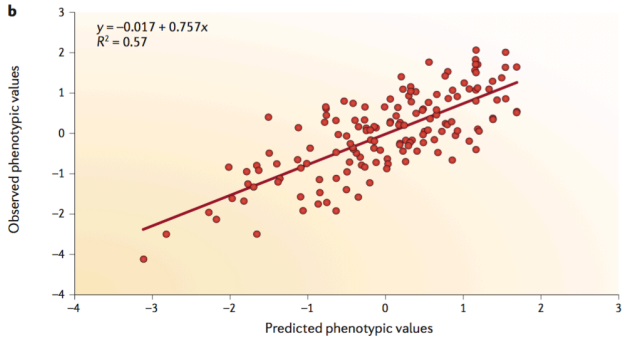
where M is the effective number of independent SNPs.



A common pitfall – sample overlap



Two scenarios where the prediction accuracy is inflated due to sample overlap



In-sample prediction (top 10 SNPs)

Wray et al. 2013 Nat Rev Genet

Genomic prediction vs. mid-parental prediction

Predicting human height by Victorian and genomic methods

Yurii S Aulchenko*[,1,2,7], Maksim V Struchalin[1,3,7], Nadezhda M Belonogova[2,4], Tatiana I Axenovich[2], Michael N Weedon[5], Albert Hofman[1], Andre G Uitterlinden[6], Manfred Kayser[3], Ben A Oostra[1], Cornelia M van Duijn[1], A Cecile JW Janssens[1] and Pavel M Borodin[2,4]

[1]*Department of Epidemiology and Biostatistics and Clinical Genetics, Erasmus MC, Rotterdam, The Netherlands;* [2]*Laboratory of Recombination and Segregation Analysis, Institute of Cytology and Genetics SD RAS, Novosibirsk, Russia;* [3]*Department of Forensic Molecular Biology, Erasmus MC, Rotterdam, The Netherlands;* [4]*Department of Cytology and Genetics, Novosibirsk State University, Novosibirsk, Russia;* [5]*Department of Genetics of Complex Traits and Diabetes Genetics, Peninsula College of Medicine and Dentistry, Exeter, UK;* [6]*Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands*

In the Victorian era, Sir Francis Galton showed that 'when dealing with the transmission of stature from parents to children, the average height of the two parents, … is all we need care to know about them' (1886). One hundred and twenty-two years after Galton's work was published, 54 loci showing strong statistical evidence for association to human height were described, providing us with potential genomic means of human height prediction. In a population-based study of 5748 people, we find that a 54-loci genomic profile explained 4–6% of the sex- and age-adjusted height variance, and had limited ability to discriminate tall/short people, as characterized by the area under the receiver-operating characteristic curve (AUC). In a family-based study of 550 people, with both parents having height measurements, we find that the Galtonian mid-parental prediction method explained 40% of the sex- and age-adjusted height variance, and showed high discriminative accuracy. We have also explored how much variance a

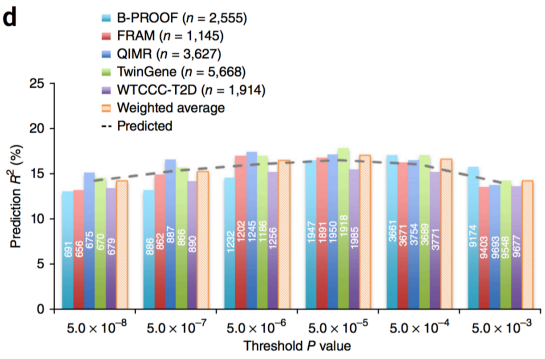
1) Common environmental effects?
2) Late onset diseases?
3) Parental data are missing?

Aulchenko et al. 2009 EJHG

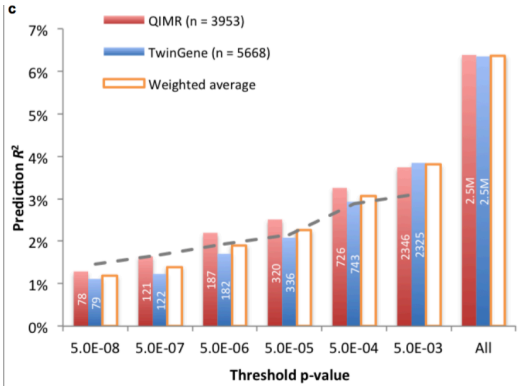
Current GWAS

Accuracy of prediction based on GWAS result is still very limited

Height n = ~250,000



BMI n > 300,000



Wood et al. 2014 Nat Genet
Locke et al. 2015 Nature

Possible reasons

- The polygenic architecture: too many variants of small effects (the average variance explained by the top associated height SNPs is ~0.02%).
- Sample heterogeneity: the effective sample size is likely smaller than the reported.
- Modelling: one SNP is fitted at a time.

Summary-data based BLUP (sBLUP)

- GCTA-COJO: re-estimation of SNP effects using summary-level data from GWAS/meta-analysis and LD correlation between SNPs from a reference sample (Yang et al. 2012 Nat Genet)

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \text{ and } \text{var}(\hat{\mathbf{b}}) = \sigma^2_{\text{J}}(\mathbf{X}'\mathbf{X})^{-1}$$ Multiple regression

$$\hat{\boldsymbol{\beta}} = \mathbf{D}^{-1}\mathbf{X}'\mathbf{y} \text{ and } \text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2_{\text{M}}\mathbf{D}^{-1}$$ Simple regression in matrix form

$$\tilde{\mathbf{b}} = \mathbf{B}^{-1}\mathbf{D}\hat{\boldsymbol{\beta}} \text{ and } \text{var}(\tilde{\mathbf{b}}) = \sigma^2_{\text{J}}\mathbf{B}^{-1}$$

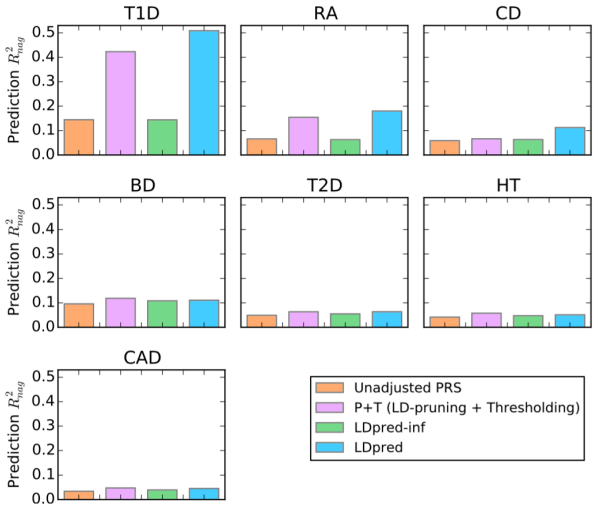
- Similar idea can be applied to performed a summary-data based BLUP analysis (ridge regression) – GCTA-COJO-sBLUP

$$\widehat{\mathbf{b}}_{\text{R}} = (\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{D}\widehat{\mathbf{b}} = (\mathbf{R} + \mathbf{I}\lambda/n)^{-1}\widehat{\mathbf{b}}$$

Peters et al. 2015 Nat Commu
Robinson et al. 2017 Nat Hum Behav

Summary-data based mixture model

LDpred: a mixture of random effect models



Vilhjálmsson et al. 2015 AJHG

Questions and discussion

Lecture 07: Models with Multiple Random Effects: Repeated Measures, Maternal and Associative effects

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
9 – 10 Feb 2017

Often there are several vectors of random effects

- Repeatability models
  - Multiple measures
- Common family effects
  - Cleaning up residual covariance structure
- Maternal effects models
  - Maternal effect has a genetic (i.e., breeding value) component

Multiple random effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{u} + \mathbf{e}$$

y is a n x 1 vector of observations

$\boldsymbol{\beta}$ is a q x 1 vector of fixed effects

a is a p x 1 vector of random effects

u is a m x 1 vector of random effects

X is n x q, **Z** is n x p, **W** is n x m

y, **X**, **Z**, **W** observed. $\boldsymbol{\beta}$, **a**, **u**, **e** to be estimated

Covariance structure

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{W}\mathbf{u} + \mathbf{e}$$

Defining the covariance structure key in any mixed-model

Suppose $\mathbf{e} \sim (0, \sigma_e^2 \mathbf{I})$, $\mathbf{u} \sim (0, \sigma_u^2 \mathbf{I})$, $\mathbf{a} \sim (0, \sigma_A^2 \mathbf{A})$,
as with breeding values

These covariances matrices are still not sufficient, as we have yet to give describe the relationship between **e**, **a**, and **u**. If they are independent:

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$$

$y = X\beta + Za + Wu + e$ $\quad \begin{pmatrix} \mathbf{a} \\ \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$

Covariance matrix for the vector of observations **y**

$$\mathbf{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{Z}\mathbf{A}\mathbf{Z}^T\sigma_A^2 + \mathbf{W}\mathbf{W}^T\sigma_u^2 + \mathbf{I}\,\sigma_e^2$$

Note that if we ignored the second vector u of random effects, and assumed **y = Xβ + Za + e***, then **e* = Wu + e**, with Var(**e***) = σ_e^2 I + σ_u^2 **WW**^{T}

Consequence of ignoring random effects is that these are incorporated into the residuals, potentially compromising its covariance structure

Mixed-model Equations

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{W} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda_A\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{Z} & \mathbf{W}^T\mathbf{W} + \lambda_u\mathbf{I} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{pmatrix}$$

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_u^2}$$

The repeatability model

- Often, multiple measurements (aka "records") are collected on the same individual
- Such a record for individual k has three components
  - Breeding value a_k
  - Common (permanent) environmental value p_k
  - Residual value for ith observation e_{ki}
- Resulting observation is thus
  - $z_{ki} = \mu + a_k + p_k + e_{ki}$
- The repeatability of a trait is $r = (\sigma_A^2+\sigma_p^2)/\sigma_z^2$
- Resulting variance of the residuals is $\sigma_e^2 = (1-r)\,\sigma_z^2$

Resulting mixed model

$$y = X\beta + Za + Zp + e$$

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_p^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$$

Notice that we could also write this model as
$y = X\beta + Z(a + p) + e = y = X\beta + Zv + e,\ v = a+p$

In class question: Why can we obtain separate estimates of **a** and **p**?

The careful reader might notice that the two vectors of random effects, the breeding values $\mathbf{a}$ and permanent environment effects $\mathbf{p}$, enter the model as $\mathbf{Za}$ and $\mathbf{Zp}$, respectively. Why then do we simply not combine these, e.g., $\mathbf{Zu}$ where $\mathbf{u} = \mathbf{a} + \mathbf{p}$? The reason we cannot do this (and indeed the reason we can estimate $\mathbf{a}$ and $\mathbf{p}$ separately!) is that $\mathbf{a}$ and $\mathbf{p}$ have *different covariance structures*, $\sigma^2_A \mathbf{A}$ versus $\sigma^2_p \mathbf{I}$. Thus, we assume that permanent environment effects are uncorrelated across individuals and are homoscedastic. On the other hand, breeding values generate covariances in relatives. Again, the critical importance of the covariance matrix to a mixed model analysis is apparent.

The incident matrix Z

Suppose we have a total of 7 observations/records, with 3 measures from individual 1, 2 from individual 2, and 2 from individual 3. Then:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

Why? Matrix multiplication. Consider y_{21}.

$$y_{21} = \mu + A_2 + p_2 + e_{21}$$

Consequences of ignoring p

- Suppose we ignored the permanent environment effects and assumed the model $\mathbf{y = X\beta + Za + e^*}$
  - Then $\mathbf{e^* = Zp + e}$,
  - Var($\mathbf{e^*}$) = $\sigma_e^2 I + \sigma_p^2 ZZ^T$
- Assuming that Var($\mathbf{e^*}$) = $\sigma_e^2 I$ gives an incorrect model
- We could either
  - use $\mathbf{y = X\beta + Za + e^*}$ with the correct error structure (covariance) for $e^* = \sigma_e^2 I + \sigma_p^2 ZZ^T$
  - Or use $\mathbf{y = X\beta + Za + Zp + e}$, where $e = \sigma_e^2 I$

The repeatability model was used by Estany et al. (1989) to examined the selection response for litter size in rabbits. Their model assumed two groups of fixed effects, d_t the year-season (environmental) effect which had 22 levels in this experiment and the reproductive state l_i of the doe (l has three levels: l_1 for primiparious does, l_2 for lactating does, and l_2 for non-primiparious and non-lactating does). Since only two of these l_x factors are estimable, l_1 was assigned a value zero. Their model had three random effects, a_k and p_k for the additive genetic and permanent environmental effect of the kth doe, and the residual e, giving the overall model as

$$y_{tk\ell i} = \mu + l_i + d_t + a_k + p_k + e_{tk\ell i}$$

where $y_{tk\ell i}$ denotes the litter size for the ℓth litter of doe k in reproductive state i in season-year t.

In matrix form, the mixed-model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Za} + \mathbf{Zp} + \mathbf{e}$$

where $\mathbf{a}$ and $\mathbf{p}$ are $n \times 1$ vectors corresponding to the n does, $\mathbf{Var}(\mathbf{a}) = \sigma_A^2 \mathbf{A}$, $\mathbf{Var}(\mathbf{p}) = \sigma_p^2 \mathbf{I}$, and $\mathbf{Var}(\mathbf{e}) = \sigma_e^2 \mathbf{I}$. $\mathbf{X}$ and $\mathbf{Z}$ are incident matrices, and the vector of fixed effects is

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ l_1 \\ l_2 \\ d_1 \\ \vdots \\ d_{22} \end{pmatrix}$$

Resulting mixed-model equations

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z}+\lambda_A\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\mathbf{Z}+\lambda_u\mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} = \frac{1-r}{h^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_p^2} = \frac{1-r}{r-h^2}$$

Common family effects

- Sibs in the same family also share a common environment
  - Cov(full sibs) = $\sigma_A^2/2 + \sigma_D^2/4 + \sigma_{ce}^2$
- Hence, if the model assumes $y_i = \mu + a_i + c_i + e_i$, with $\mathbf{a} \sim 0, \sigma_A^2\mathbf{A}$, $c \sim 0, \sigma_{cf}^2 I$. If there are records for different sibs from the same family, Var($\mathbf{e}$) is no longer $\sigma_e^2 I$
- $\mathbf{y} = X\beta + Za + Wc + e$
- Again, if common family effect ignored (we assume $\mathbf{y} = X\beta + Za + e^*$) the error structure is $e^* = \sigma_e^2 I + \sigma_{cf}^2 WW^T$
  - Where $\sigma_{cf}^2 = \sigma_D^2/4 + \sigma_{ce}^2$
  - The common family effect may contain both environment and non-additive genetic components

Example: Measure 7 individuals, first five are from family one, last two from family 2

$$y = X\beta + Za + Wc + e$$

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}, \quad \mathbf{Z} = \mathbf{I}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Z = I as every individual has a single record.
If there are missing and/or repeated records,
Z does not have this simple structure

$$y = X\beta + Za + Wc + e$$

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}, \quad \mathbf{Z} = \mathbf{I}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Again, matrix multiplication gives us the form of the Z and W matrices. Consider y_6:

$$y_6 = \mu + A_6 + c_2 + e_6$$

Maternal effects with genetic components

- The phenotype of an offspring can be influenced by its mother beyond her genetic contribution
- For example, two offspring with identical genotypes will still show potentially significant differences in size if they receive different amounts of milk from their mothers
- Such maternal effects can be quite important
- While we have just discussed models with common family effects, these are potentially rather different that maternal effects models
  - Common family environmental effects are assumed not to be inherited across generations.

- Consider milk yield. The heritability for this trait is around 30% and the milk yield of the mother has a significant impact on the weight of her offspring
- Offspring with high breeding values for milk will tend to have daughters with above -average milk yield, and hence above -average maternal effects
- The value of an offspring can be considered to consist of two components
  - A direct effect (intrinsic breeding value)
  - A maternal contribution

Phenotypic value = direct value + maternal value

$$P_z = P_d + P_m$$

Observable

Latent (unseen) values

Both of the latent values can be further decomposed into breeding plus residual (environmental + non- additive genetic) values

$$P_d = \mu + A_d + E_d, \qquad P_m = \mu + A_m + E_m,$$

The direct breeding value A_d appears in the phenotype of its carrier

The maternal breeding value A_m DOES NOT appear in the phenotype of its carrier, but rather in the phenotype of her offspring

Direct vs. maternal breeding values

- The direct and maternal contributions are best thought of as two separate, but potentially correlated, traits.
  - Hence, we need to consider $\sigma(A_d,A_m)$ in addition to $\sigma^2(A_d)$ and $\sigma^2(A_m)$. This changes the form of the mixed-model equations
- The direct BV (A_d) is expressed in the individual carrying it
- The maternal BV (A_m) is only expressed in the offspring trait value (and only mom's A_m appears)

Covariance structure

$$\begin{pmatrix} \mathbf{a}_d \\ \\ \mathbf{a}_m \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2(A_d)\,\mathbf{A} & \sigma(A_d, A_m)\,\mathbf{A} \\ \\ \sigma(A_d, A_m)\,\mathbf{A} & \sigma^2(A_m)\,\mathbf{A} \end{pmatrix}$$

This is often written using the Kronecker (or direct) product:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & {}_{mn}\mathbf{B} \end{pmatrix}$$

Giving

$$\begin{pmatrix} \mathbf{a}_d \\ \\ \mathbf{a}_m \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \\ \mathbf{0} \end{pmatrix}, \mathbf{G} \otimes \mathbf{A} \qquad \mathbf{G} = \begin{pmatrix} \sigma^2(A_d) & \sigma(A_d, A_m) \\ \\ \sigma(A_d, A_m) & \sigma^2(A_m) \end{pmatrix}$$

The mixed-model becomes

Direct effects
breeding values

$$y = X\beta + Z_d a_d + Z_m a_m + e$$

Maternal effects
breeding values

The error structure needs a little care, as the direct E_d and maternal E_m residual values can be correlated*. Initially, we will assume Var(**e**) ~ σ_e^2**I**

*See Bijma 2006 J. Anim. Sci. 84:800-806 for treatment of correlated environmental residuals under this model

The resulting mixed-model equations become

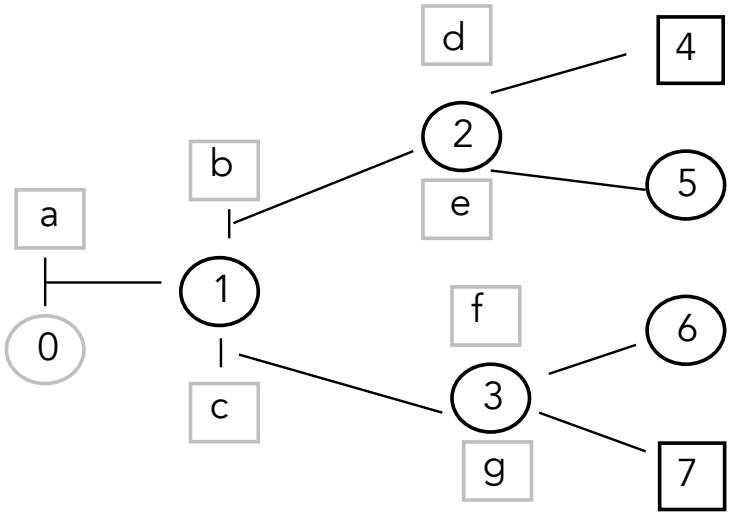
$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_d & \mathbf{X}^T\mathbf{Z}_s \\ \mathbf{Z}_d\mathbf{X}^T & \mathbf{Z}_d^T\mathbf{Z}_d + \lambda_1\mathbf{A}^{-1} & \mathbf{Z}_d^T\mathbf{Z}_m + \lambda_2\mathbf{A}^{-1} \\ \mathbf{Z}_m\mathbf{X}^T & \mathbf{Z}_m^T\mathbf{Z}_d + \lambda_2\mathbf{A}^{-1} & \mathbf{Z}_m^T\mathbf{Z}_m + \lambda_3\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_d^T\mathbf{y} \\ \mathbf{Z}_m^T\mathbf{y} \end{pmatrix}$$

where the weights λ_i are related to elements in the inverse of $\mathbf{G}$, viz.,

$$\begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_2 & \lambda_3 \end{pmatrix} = \sigma_e^2\,\mathbf{G}^{-1} = \sigma_e^2 \begin{pmatrix} \sigma^2(A_d) & \sigma(A_d, A_m) \\ \sigma(A_d, A_m) & \sigma^2(A_m) \end{pmatrix}^{-1}$$

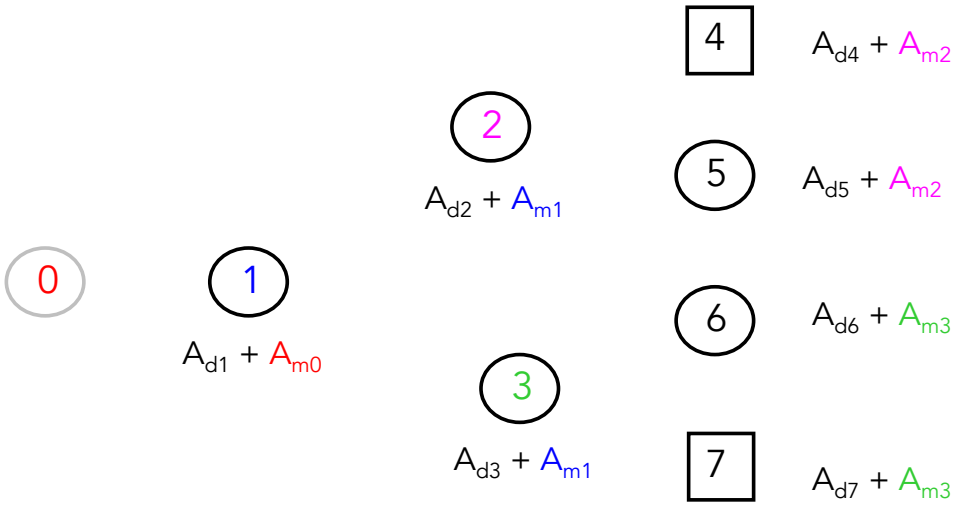
Filling out the maternal effects incident matrix Z_m

A little bookkeeping care is needed when filling out Z_m, because the A_m associated with a record (measured individual) is that of their mother.



1-7 have records

All sires unrelated



The observed values are y_1 through y_7.
What we can estimate are A_{d1} through A_{d7}, A_{m0} through A_{m3}

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}, \qquad \mathbf{a}_d = \begin{pmatrix} A_{d,1} \\ A_{d,2} \\ A_{d,3} \\ A_{d,4} \\ A_{d,5} \\ A_{d,6} \\ A_{d,7} \end{pmatrix}, \qquad \mathbf{Z}_d = \mathbf{I}, \qquad \mathbf{a}_m = \begin{pmatrix} A_{m.o} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \end{pmatrix}$$

Note that we estimate A_{m0} even though we don't have a record (observation) on her.

Since $\mathbf{Z_m a_m}$ must be a 7 x 1 matrix, $\mathbf{Z_m}$ is 7 x 4 (as $\mathbf{a_m}$ is 4 x 1)

Record 1 is associated with A_{m0}

Records 2 and 3 are associated with A_{m1}

Records 4 and 5 are associated with A_{m2}

Records 6 and 7 are associated with A_{m3}

Record 1 is associated with A_{m0}

Records 2 and 3 are associated with A_{m1}

Records 4 and 5 are associated with A_{m2}

Records 6 and 7 are associated with A_{m3}

$$\mathbf{Z}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \text{as} \quad \mathbf{Z}_m\mathbf{a}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \end{pmatrix} = \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,1} \\ A_{m,2} \\ A_{m,2} \\ A_{m,3} \\ A_{m,3} \end{pmatrix}$$

What about A_{m4} through A_{m7}?

Although we have records that only directly relate A_{m0} to A_{m3}, through the use of A we can (in theory) also estimate the maternal breeding values for individuals 4 through 7. Note this includes the maternal BVs for the two males (5 & 7), as they can pass this onto their daughters.

$$\mathbf{Z}_m^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{a}_m^* = \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \\ A_{m,4} \\ A_{m,5} \\ A_{m,6} \\ A_{m,7} \end{pmatrix}$$

Note that

$$\mathbf{Z}_m^* \mathbf{a}_m^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \\ A_{m,4} \\ A_{m,5} \\ A_{m,6} \\ A_{m,7} \end{pmatrix} = \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,1} \\ A_{m,2} \\ A_{m,2} \\ A_{m,3} \\ A_{m,3} \end{pmatrix}$$

All this raises the question about what can, and cannot, be estimated from the data (**y**) and the design (Z_m, Z_d)?

First issue: Is the structure of the design such that we can estimate all of the variance components. This is the issue of identifiability

Estimability vs. Identifiability

Details: Identifiability of Variance Components

Due to potential confounding of effects, any particular design might not allow for all variables of interest to be uniquely estimated. For the vector $\boldsymbol{\beta}$ of fixed effects, this is the concept of **estimability** (LW Chapter 26). For $\mathbf{z} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, the vector of fixed effects is estimable (all have unique values) if $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ exists. Otherwise, some of the fixed effects are confounded and cannot be separated by the design ($\mathbf{X}$) being used. With (co)variance components (often called **dispersal parameters**), a similar concept, **identifiability**, also exists. If variance components are not identifiable in the design, then BLUPs for their associated vectors of random effects do not exist.

Conditions for identifiability of REML estimates of (co)variance components are given by Rothenberg (1971), Jiang (1996), and Cantet and Cappa (2008). Before presenting these, we first review a few details about REML. Recall (LW Chapter 27) that REML estimates are those that maximize that part of the likelihood function that is independent of the fixed effects (this is often stated as being the **translation invariant** part). Let $\mathbf{V}$ be the covariance matrix of $\mathbf{z}$, which is a function of its variance components. As detailed in LW Chapter 27, Harville (1977) shows that (if it exists) the transformation provided by the matrix

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} \tag{1a}$$

plays a critical role in REML estimates. That this matrix can remove fixed effects can be seen by noting that

$$\mathbf{P}\mathbf{z} = \mathbf{V}^{-1}\left(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{1b}$$

yields a vector that is the data vector adjusted by the (estimated) fixed effects. Now consider covariance structures of the form

$$\mathbf{V} = \sum_{i=1}^{n} \mathbf{V}_i \theta_i \tag{2a}$$

where $\mathbf{V}_i$ is a matrix of known constants and the θ_i are unknown variances and covariances to be estimated.

The equations to maximize the likelihood over the restricted space (the REML estimates) are given by LW Equations 27.18 and 27.19, and are solved iteratively. These equations involve the **trace** (sum of the diagonal elements) of matrix products involving $\mathbf{P}$ and the $\mathbf{V}_i$. Recall (LW Appendix 4) that for a vector $\boldsymbol{\Theta}$ of n unknowns, the Fisher information matrix $\mathbf{F}$ (the matrix of second partial derivatives of the likelihood with respect to the parameters) can be used to provide large-sample standard errors. The resulting $n \times n$ information matrix for REML estimates of the unknown θ_i in Equation 2a is

$$F_{ij} = \text{trace}\,(\mathbf{P}\mathbf{V}_i\mathbf{P}\mathbf{V}_j) \tag{2b}$$

Much in the same fashion that the existence of $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ informs us that all fixed effects are estimable in a given design, all variance components θ_i are identifiable if all of the eigenvalues of $\mathbf{F}$ are positive, that is, that $\mathbf{F}$ is positive-definite (Rothenberg 1971, Jiang 1996). For the maternal effects mixed model, Equation 2a becomes

$$\mathbf{V} = \mathbf{V}_1\,\sigma^2(A_d) + \mathbf{V}_2\,\sigma(A_d, A_s) + \mathbf{V}_3\,\sigma^2(A_s) + \mathbf{V}_4\,\sigma_e^2 \tag{3a}$$

where

$$\mathbf{V}_1 = \mathbf{Z}_d\mathbf{A}\mathbf{Z}_d^T, \quad \mathbf{V}_2 = \left(\mathbf{Z}_d\mathbf{A}\mathbf{Z}_m^T + \mathbf{Z}_m\mathbf{A}\mathbf{Z}_d^T\right), \quad \mathbf{V}_3 = \mathbf{Z}_m\mathbf{A}\mathbf{Z}_s^T, \quad \mathbf{V}_4 = \mathbf{I} \tag{3b}$$

Substituting Equations 1a and 3b into Equation 2b fills out the $\mathbf{F}$ matrix (which is only 4×4 in this case given the four unknown variance components). For any particular design, the eigenvalues of this matrix can be computed to determine if the variance components are all identifiable.

Second issue, connectivity

Even if the design is such that we can estimate all the genetic variances, whether we can estimate all of the β, a_d, and a_m in the model depends on whether a unique inverse exists for the MME

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_d & \mathbf{X}^T\mathbf{Z}_s \\ \mathbf{Z}_d\mathbf{X}^T & \mathbf{Z}_d^T\mathbf{Z}_d + \lambda_1\mathbf{A}^{-1} & \mathbf{Z}_d^T\mathbf{Z}_m + \lambda_2\mathbf{A}^{-1} \\ \mathbf{Z}_m\mathbf{X}^T & \mathbf{Z}_m^T\mathbf{Z}_d + \lambda_2\mathbf{A}^{-1} & \mathbf{Z}_m^T\mathbf{Z}_m + \lambda_3\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_d^T\mathbf{y} \\ \mathbf{Z}_m^T\mathbf{y} \end{pmatrix}$$

Unique estimates of all the β require $(X^TV^{-1}X)^{-1}$ exists

If $(X^TV^{-1}X)^{-1}$ does not exist, a generalized inverse is used which can uniquely estimate k linear combinations of the β where k is the rank of $X^TV^{-1}X$

Likewise, if the MME equation does not have an inverse (and this is not due to constraints on β), then a generalized inverse can be used to estimate unique estimates of certain linear combinations of the a_d and a_m.

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_d & \mathbf{X}^T\mathbf{Z}_s \\ \mathbf{Z}_d\mathbf{X}^T & \mathbf{Z}_d^T\mathbf{Z}_d + \lambda_1\mathbf{A}^{-1} & \mathbf{Z}_d^T\mathbf{Z}_m + \lambda_2\mathbf{A}^{-1} \\ \mathbf{Z}_m\mathbf{X}^T & \mathbf{Z}_m^T\mathbf{Z}_d + \lambda_2\mathbf{A}^{-1} & \mathbf{Z}_m^T\mathbf{Z}_m + \lambda_3\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_d^T\mathbf{y} \\ \mathbf{Z}_m^T\mathbf{y} \end{pmatrix}$$

A key role in ensuring that unique estimates of a_d and a_m exist is played by the relationship matrix A. If individuals with records and individuals without records are sufficiently well connected (non-zero entries in A for their pair-wise relatedness), then we usually can estimate values of un-observed individuals (although their precision is another issue)

Associative effects models

- A very powerful recent development in quantitative genetics (although the idea dates back to Griffin's work in the 1960s) is the notion of direct vs. associative (or social, or indirect genetic) effects
- This idea unifies kin and group selection, offers models for the evolution of social (group-level) traits, and shows why selection can often fail
- The basic idea is that the phenotype of a target individual is a function of some intrinsic direct value and also the phenotypes of those individuals with which it interacts.

Direct & Associative effects

- Consider egg production from chickens raised in cages. Production is a function of both a chicken's own genetics and the environment (her other cage-mates)
  - Direct effects = intrinsic egg production
  - Associative effects = competitive ability
- Suppose our focal individual (i) interacts with n-1 others in a group

$$z_i = P_{d,i} + \sum_{j \neq i}^{n} P_{j,s}$$

Direct and associative effects can be antagonistic

- Consider a plant with a trait that allows it to more efficiently garner resources
- This gives it a high direct effect but a negative associative effect --- it reduces the trait values in those individuals with which it interacts
- Thus, the best performing single plants can have very low average plot performance

Breeding values for direct (A_d) and associative (A_s) effects

- Can express the phenotype of i in terms of its direct breeding value ($A_{d,i}$) and the associative breeding values ($A_{s,j}$) of its group mates

$$z_i = \mu + (A_{d_i} + E_{d_i}) + \sum_{j \neq i} (A_{s_j} + E_{s_j})$$

$$z_i = \mu + A_{d_i} + \sum_{j \neq i} A_{s_j} + e_i, \qquad e_i = E_{d_i} + \sum_{j \neq i} E_{s_j}$$

Total response

The trait mean equals the mean of the direct effects plus the means of the associative effects,

$$\mu_z = \mu_{A_d} + (n-1)\mu_{A_s}$$

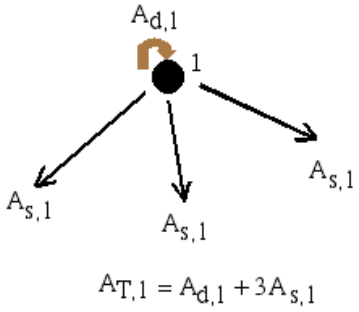
Total response is the sum of the response R_d in the direct breeding values plus the sum of the responses R_s in the associative effects breeding values,

$$R_z = R_d + (n-1)R_s$$

Total breeding value

The key to predicting response is the total breeding value of an individual, where

$$A_{T,i} = A_{d,i} + (n-1)A_{s,i}$$



Note that part ($A_{s,i}$) of the total breeding value of i never appears in its phenotype. Must either use informative from relatives or the group to estimate it.

h^2 and τ^2

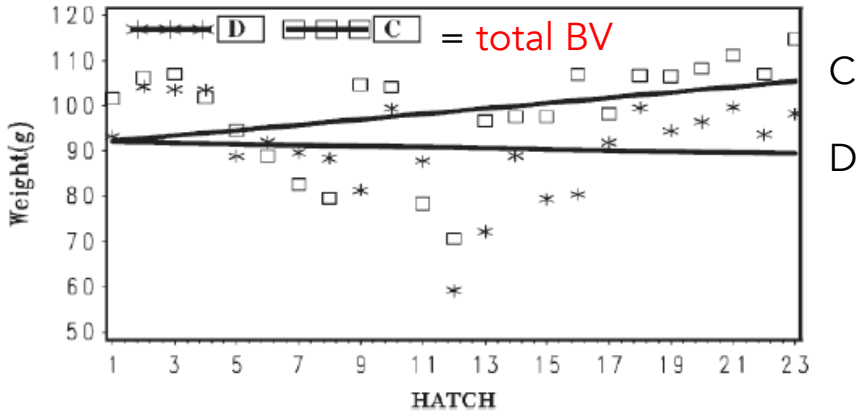
- τ^2, the analog for h^2, is the ratio of the total breeding value to the individual phenotypic variance
  - $\tau^2 = Var(A_T)/Var(z)$
- Note that, unlike h^2, τ^2 can exceed one,
- Why? A potentially large fraction of A_T never appears in z, and hence Var(z)
  - $Var(A_T) = Var(A_d) + (n-1)Var(A_s)$
  - $\tau^2 = Var(A_d)/Var(z) + (n-1)Var(A_s)/Var(z)$
  - $= h^2 + (n-1)Var(A_s)/Var(z)$

BLUP estimation

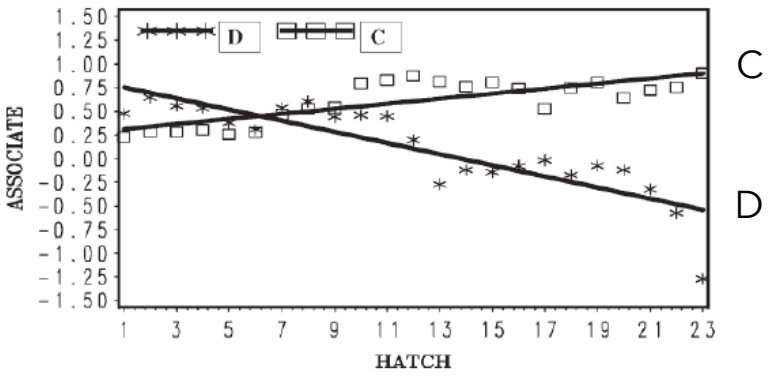
- While the total breeding value cannot be estimated directly from an individual's phenotype, using an appropriate mixed model, we can obtain
  - BLUPs of Direct breeding values (A_d)
  - BLUPs of Associative (or social) BVs (A_s)
  - REML estimates of $\sigma^2(A_d)$, $\sigma^2(A_s)$, and the direct-associate effects covariance $\sigma(A_d, A_s)$

This works: Muir's result

- Bill Muir (Purdue University) selection on six-week weight in Japanese quail over 23 generations using two different schemes
  - BLUP selection on estimated direct BV (D)
    - Denoted by D-BLUP
  - BLUP selection on estimated total BV
    - Denoted by C-BLUP



Weighted increased under selection using total BV (C), decreased under selection using direct BV (D).



C

D

Under BLUP selection on direct BV (D), significant decline in the mean social value, which over-rode the positive response in the direct value

Under BLUP selection of total BV (C), *both* increase

The mixed model

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_d\,\mathbf{a}_d + \mathbf{Z}_s\,\mathbf{a}_s + \mathbf{e}$$

Example: Individuals 1-4 and 5-8 are half sibs from unrelated families

$$\mathbf{A} = \begin{pmatrix} 1 & 0.25 & 0.25 & 0.25 & 0 & 0 & 0 & 0 \\ 0.25 & 1 & 0.25 & 0.25 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 1 & 0.25 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0.25 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 & 0.25 & 1 & 0.25 & 0.25 \\ 0 & 0 & 0 & 0 & 0.25 & 0.25 & 1 & 0.25 \\ 0 & 0 & 0 & 0 & 0.25 & 0.25 & 0.25 & 1 \end{pmatrix}$$

Filling out Z_s

- Suppose group one contains individuals 1, 2, 5, 6. The resulting values for these individuals become
  - $z_1 = m + A_{d1} + A_{s2} + A_{s5} + A_{s6} + e$
  - $z_2 = m + A_{d2} + A_{s1} + A_{s5} + A_{s6} + e$
  - $z_5 = m + A_{d5} + A_{s1} + A_{s2} + A_{s6} + e$
  - $z_6 = m + A_{d6} + A_{s1} + A_{s2} + A_{s5} + e$
  - The result Z_d and Z_s incident matrices become

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_d\,\mathbf{a}_d + \mathbf{Z}_s\,\mathbf{a}_s + \mathbf{e}$$

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{a}_d = \begin{pmatrix} A_{d,1} \\ A_{d,2} \\ A_{d,3} \\ A_{d,4} \\ A_{d,5} \\ A_{d,6} \\ A_{d,7} \\ A_{d,8} \end{pmatrix}, \quad \mathbf{Z}_d = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} = \mathbf{I}_8$$

Group one contains individuals 1,2,5,6; while group two contains 3,4,7,8.

$$\mathbf{Z}_s = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{a}_s = \begin{pmatrix} A_{s,1} \\ A_{s,2} \\ A_{s,3} \\ A_{s,4} \\ A_{s,5} \\ A_{s,6} \\ A_{s,7} \\ A_{s,8} \end{pmatrix}$$

Lots of hidden variation to exploit

- Bergsma et al. (2008) examined four traits in 14,000 pigs grown in pens of 6-12 animals.
- Heritability for these traits was estimated in a model without social effects,

	Growth	Back fat	Muscle	Intake
$\sigma^2(A)$	2,583	2.83	7.94	41,275
h^2	0.37	0.36	0.25	0.41

Next, a model was fit allowing for heritable social effects, $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_d\mathbf{a}_d + \mathbf{Z}_s\mathbf{a}_s + \mathbf{Z}_c\mathbf{c} + \mathbf{e}$, which gave estimates of

	Growth	Back fat	Muscle	Intake
$\sigma^2(A_d)$	1,522	2.75	6.68	16,950
h^2_d	0.21	0.35	0.21	0.17
$\sigma^2(A_s)$	51	0.01	0.03	596
$\sigma^2(A_T)$	5,208	3.19	10.35	68,687
τ^2	0.71	0.41	0.32	0.70

Here $h^2_d = \sigma^2(A_d)/\sigma^2(z)$, while $\tau^2 = \sigma^2(A_T)/\sigma^2(z)$. h^2_d measures the response potential under phenotypic selection, while $\tau^2 \geq h^2_d$ measures the total genetic potential for improvement under specialized selection designs.

	Growth	Back fat	Muscle	Intake
$\sigma^2(A)$	2,583	2.83	7.94	41,275
h^2	0.37	0.36	0.25	0.41

Hence, for growth and food intake, lots of additional genetic variation for trait response lies “hidden” in associative effects.

Lecture 8:

Infinite-dimensional/Function-valued Traits: Covariance Functions and Random Regressions

Bruce Walsh lecture notes
Introduction to Quantitative Genetics
SISG, Brisbane
9 – 10 Feb 2017

Longitudinal traits

- Many classic quantitative traits are longitudinal -- measured at multiple time points --- milk yield, body size, etc.
- We have already examined the repeated-measures design wherein an identical trait (assumed to be unchanging) is measured multiple times.
- For most longitudinal traits, we expect the trait to change over time, such as a growth curve.
- These are function-valued traits, also called infinite-dimensional traits.
- One critical feature of such traits is that their additive variances change with t, and trait values from different time points have different correlations.

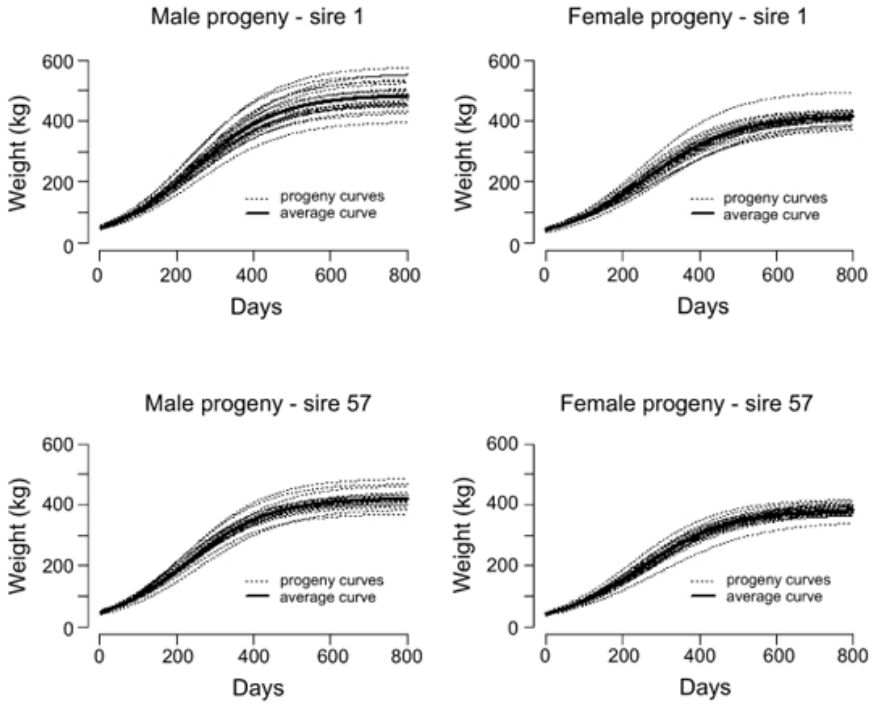


Figure 3 - Mixed logistic growth curves (---) fitted for all progeny of sire 1 (24 males and 32 females) and all progeny of sire 57 (20 males and 59 females) and associated average growth curves (—).

Sci Agric. 66: 85-89

Norms of reaction

- The other type of function-valued trait is one indexed by some continuous environmental variable (as opposed to time), such as adult body weight as a function of temperature or grain yield as a function of total rainfall.
- The measurement of such traits generally requires replication of individuals over environments (versus the sequential evaluation of a single individual with longitudinal traits). As with G x E, this can be done
  - Using clones/pure lines
  - Using family members
- Such curves are common in ecology & evolution and are called norms of reaction, and are measures of G x E
  - Norms of reaction measure phenotypic plasticity --- variation that can be expressed from a fixed genotype, which is often an important adaptation in changing environments.

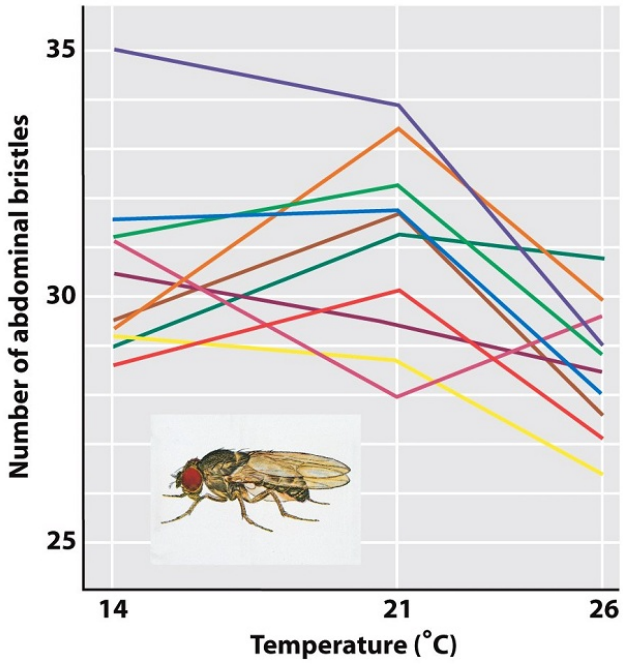


Figure 18-6
Introduction to Genetic Analysis, Ninth Edition

How to model such traits?

- One obvious approach is to treat the trait measured at discrete time points as a series of correlated traits.
  - Makes sense to do this for something like parity (litter number), as individuals are all measured at the same event, i.e., parity one, parity two, etc.
  - However, with a trait like a growth or some performance curve, we often expect to have different time measurements for different individuals.
    - We could either lump these into groups (reducing precision) or treat each different time/tuning variable value as a different trait (much missing data).
  - Better solution: estimate the trait covariance function, where $C(t_1,t_2) = Cov[z(t_1),z(t_2)]$ or $Cov[A(t_1),A(t_2)]$

Covariance function approach

- Kirkpatrick popularized the use of covariance functions (largely in evolutionary biology) in the mid-late 1980's.
- He noted that traits measured with respect to some continuous indexing variable (such as time or temperature) have effectively infinite dimensions, as one could (in theory) always consider finer and finer time scales.
  - Thus, rather than treat them as a (potentially) every-expanding set of discrete correlated traits, better to simply consider the covariance $C(t_1,t_2)$ between any two time points within the range of the sampled data. Note that $C(t_1,t_1)$ is the trait variance at time t_1.
  - $C(t_1,t_2)$ is the covariance function, the logical extension of the covariance matrix C(i,j) used for correlated traits, using continuous, rather than integer, indexes.

Covariance functions (cont)

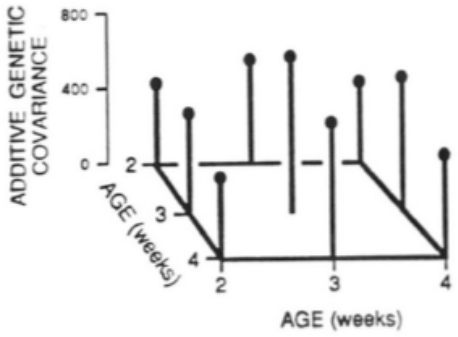
- As with any quantitative trait, the covariance between the values at two time points can be decomposed into an additive-genetic (breeding value) covariance function and a residual (or environmental) covariance function,
  - $C_z(t_1,t_2) = C_A(t_1,t_2) + C_E(t_1,t_2)$
- The issue in the estimation of the additive covariance function is how one proceeds from an additive-covariance matrix estimate **G** from discrete time points to a continuous function covering all possible values with the span of time sampled to estimate **G**.
  - Basic (initial) idea: Use curve-fitting based on low-degree polynomials to use **G** to fit a covariance function
  - This is typically done by using Legendre polynomials as the basis function.

Riska et al. (1984) data on breeding values for log(body weight)

The basic idea was illustrated by Kirkpatrick with a data set on mouse body weight measured at ages 2, 3, and 4 weeks. Riska et al. estimated the G matrix as

$$\hat{G} = \begin{array}{c} \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 436 & 522 & 424 \\ 522 & 808 & 665 \\ 424 & 665 & 558 \end{bmatrix} \end{array}$$

Plotting these values on a lattice at these discrete time points gives



Ideally, would like some sort of smooth curve for this data.

Towards the covariance function

- Suppose we assume the breeding value at time t (for $2 \leq t \leq 4$ weeks) is in the form of a quadratic, so that individual's i breeding value is given by
  - $A_i(t) = a_{io} + a_{i1}\, t + a_{i2}\, t^2$.
  - Here the a_{ij} (for $0 \leq j \leq 2$) are regression coefficients unique to individual i, and are unchanging over time.
- A different individual (j) also has a quadratic regression, but with different coefficients
  - $A_j(t) = a_{jo} + a_{j1}\, t + a_{j2}\, t^2$.
  - the a_{ij} are referred to as random regression coefficients, as they are random (drawn from some distribution) OVER individuals, but constant over time WITHIN an individual.

Towards the covariance function (cont)

We can think of these random regression coefficients as being drawn from a distribution:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \sim \mathbf{0}, \mathbf{C_G}, \quad \text{where} \quad \mathbf{C_G} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Ideally, we would like to use our estimate of **G** to make inferences on the elements in $\mathbf{C_G}$.

We can write the additive value in time t for individual i as $\mathbf{a_i}^T * \mathbf{t}$, where = $\mathbf{a_i}^T = (a_{i0}, a_{i1}, a_{i2})$ and $\mathbf{t}^T = (1, t, t^2)$

Towards the covariance function

The regression $A_i(t) = a_{io} + a_{i1}t + a_{i2}t^2 = \mathbf{a_i}^T\mathbf{t}$ yields the covariance function, as the value of the vector t for different times are constants, giving

$$\begin{aligned} \text{Cov}[A_i(t_1), A_i(t_2)] &= \text{Cov}[\mathbf{a_i}^T\mathbf{t}_1, \mathbf{a_i}^T\mathbf{t}_2] \\ &= \mathbf{t}_1^T \text{Cov}(\mathbf{a_i}, \mathbf{a_i}) \mathbf{t}_2 \\ &= \mathbf{t}_1^T \mathbf{C_G} \mathbf{t}_2 \end{aligned}$$

This is a bilinear form (the generalization of a quadratic form).

$$\mathrm{Cov}[A(t_1), A(t_2)] = \mathbf{t}_1^T \mathbf{C_G}\, \mathbf{t}_2$$

$$= \begin{pmatrix} 1 & t_1 & t_1^2 \end{pmatrix} \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ t_2 \\ t_2^2 \end{pmatrix}$$

Expanding gives

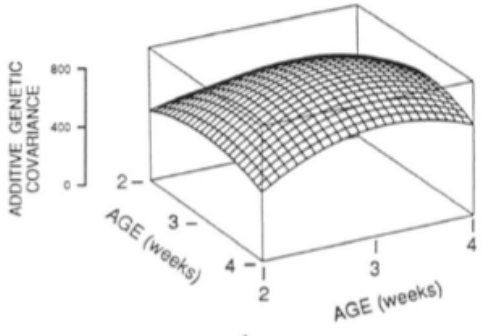
$$\mathrm{Cov}[A(t_1), A(t_2)] = \sigma_0^2 + \sigma_{01}(t_1 + t_2) + \sigma_{02}(t_1^2 + t_2^2) + \sigma_1^2 t_1\, t_2 + \sigma_{12}(t_1^2\, t_2 + t_1\, t_2^2) + \sigma_2^2\, t_1^2\, t_2^2$$

More generally, fitting an m-th degree polynomial for A gives the product of two m-degree polynomials for the covariance function

$$A_i(t) = \sum_{j=0}^{m} a_{ij}\, t^j$$

$$\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \sum_{j=0}^{m} \sum_{k=0}^{m} a_{jk}\, t_1^j\, t_2^k$$

Kirkpatrick estimated to covariance function for the Riska data by assuming an individual's breeding value over time can be modeled by 2nd degree polynomial. The resulting covariance function gives the following surface:



Estimated additive-genetic covariance function

Details

- Before building on these basic ideas to estimate the covariance function, some background on Legendre polynominals is required, as these are used as the basis functions (building blocks) for curve-fitting instead of the set $(1, t, t^2, \ldots t^k)$
  - Specifically, we could approximate a function f(t) by the k-th degree polynomial $f(t) = \Sigma^k a_i t^i$.
  - Instead, we approximate it by a weighted sum of the functions $\phi_0(t), \phi_1(t), \ldots, \phi_k(t)$, where $\phi_j(t)$ is a polynomial of degree j (the Legendre polynomial of order j, for $0 \leq j \leq k$), using $f(t) = \Sigma^k b_i \phi_i(t)$.

Legendre Polynomials

For curve-fitting, orthogonal polynomials are often used, where $\phi_k(t)$ denotes a k-th degree polynomial. The set of these building blocks $\phi_o(t), \phi_1(t), \ldots \phi_k(t)$ .. are defined to be orthogonal in the sense that the integral of $\phi_i(t)\,\phi_j(t) = 0$ when i and j are not equal. We also assume they are scaled to have unit length, with the integral $\phi_i^2(t) = 1$.

For $-1 \leq t \leq 1$, the first five scaled Legendre polynomials are given by

$$\phi_0(t) = 0.7071$$
$$\phi_1(t) = 1.2247\ t$$
$$\phi_2(t) = -0.7906 + 2.3717\ t^2$$
$$\phi_3(t) = -2.8062\ t + 4.6771\ t^3$$
$$\phi_4(t) = 0.7955 - 7.9550\ t^2 + 9.2808\ t^4$$
$$\phi_5(t) = 4.2973\ t - 20.5205\ t^3 + 18.4685\ t^5$$

For example, the curve $y = a + b\,t$ can be written as
$y = a/(0.7071)\,\phi_0(t) + b/(1.2247)\,\phi_1(t)$ for $-1 \leq t \leq 1$.
More generally, any k-th degree polynomial can be written as
$\Sigma^k a_i \phi_i(t)$

$\phi_0(t) = 0.7071$
$\phi_1(t) = 1.2247\ t$
$\phi_2(t) = -0.7906 + 2.3717\ t^2$
$\phi_3(t) = -2.8062\ t + 4.6771\ t^3$
$\phi_4(t) = 0.7955 - 7.9550\ t^2 + 9.2808\ t^4$
$\phi_5(t) = 4.2973\ t - 20.5205\ t^3 + 18.4685\ t^5$

In matrix form, $\boldsymbol{\phi} = \mathbf{M}\mathbf{t}$, where $\boldsymbol{\phi} = \begin{pmatrix} \phi_0(t) \\ \phi_1(t) \\ \phi_2(t) \\ \phi_3(t) \\ \phi_4(t) \\ \phi_5(t) \end{pmatrix}$, $\mathbf{t} = \begin{pmatrix} 1 \\ t \\ t^2 \\ t^3 \\ t^4 \\ t^5 \end{pmatrix}$

j-th row of **M** are the coefficients for the jth Legendre polynomial

Row 4 = coefficients for ϕ_4.

$$\mathbf{M} = \begin{pmatrix} 0.7071 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2247 & 0 & 0 & 0 & 0 \\ -0.7906 & 0 & 2.3717 & 0 & 0 & 0 \\ 0 & -2.8062 & 0 & 4.5777 & 0 & 0 \\ 0.7944 & 0 & -7.9950 & 0 & 9.2808 & 0 \\ 0 & 4.2973 & 0 & -20.5205 & 0 & 18.4685 \end{pmatrix}$$

$1 \quad t \quad t^2 \quad t^3 \quad t^4 \quad t^5$

How do we write the following 5th order polynomial in terms of Legendre polynomials?

$$y = 4 - 6x + 14x^2 + 26x^3 + 50x^4 - 110x^5$$

Note that $y = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a} = \begin{pmatrix} 4 \\ -6 \\ 14 \\ 26 \\ 50 \\ -110 \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{pmatrix}$

$$\begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \\ \phi_5(x) \end{pmatrix} = \mathbf{M} \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{pmatrix} \quad \text{implies} \quad \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \\ \phi_5(x) \end{pmatrix}$$

Giving $\mathbf{x} = \mathbf{M}^{-1}\boldsymbol{\phi}$. Since $y = \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{M}^{-1}\boldsymbol{\phi}$, weights on Legendre polynomials are given by $\mathbf{a}^T\mathbf{M}^{-1}$

Weights are given by $\mathbf{a}^T\mathbf{M}^{-1}$

R returns

```
> M
        [,1]    [,2]    [,3]     [,4]   [,5]    [,6]
[1,]  0.7071  0.0000  0.0000   0.0000 0.0000  0.0000
[2,]  0.0000  1.2247  0.0000   0.0000 0.0000  0.0000
[3,] -0.7906  0.0000  2.3717   0.0000 0.0000  0.0000
[4,]  0.0000 -2.8062  0.0000   4.5777 0.0000  0.0000
[5,]  0.7944  0.0000 -7.9950   0.0000 9.2808  0.0000
[6,]  0.0000  4.2973  0.0000 -20.5205 0.0000 18.4685
> t(a)%*%solve(M)
         [,1]     [,2]     [,3]      [,4]     [,5]      [,6]
[1,] 26.51006 -32.1633 24.06409 -21.01970 5.387467 -5.956087
```

Giving y = 26.51006*$\phi_0(x)$ -32.1633 *$\phi_1(x)$ + 24.06409 *$\phi_2(x)$ -21.01970 *$\phi_3(x)$ + 5.387467 *$\phi_4(x)$ -5.956087 *$\phi_5(x)$

More generally, any k-degree polynomial $\mathbf{y} = \mathbf{a}^T\mathbf{x}_k$ can be expressed as a weighted series of the first k+1 Legendre polynomials ϕ_0, .., ϕ_k, where the weights are $\mathbf{a}^T\mathbf{M}^{-1}$. $\mathbf{M}$ is (k+1) x (k+1), with the jth row being the coefficients on x for the j-th order Legendre polynomial.

The Covariance function in terms of Legendre polynomials

- Express the trait breeding value for individual i at time t_j by an m-th order polynomial,
  - $A_i(t_j) = \Sigma_k^m a_{ik} \phi_k(t_j)$, where $\mathbf{a}_i \sim 0, \mathbf{C}_G$
  - Define the vectors
    - $\boldsymbol{\phi}_m(t) = (\phi_0(t), \phi_1(t), \ldots, \phi_m(t))^T$, which we often write as just $\boldsymbol{\phi}_m$ or $\boldsymbol{\phi}$ for brevity
    - $\mathbf{a}_i = (a_{i0}, a_{i1}, \ldots, a_{im})^T$.
- Hence $A_i(t_j) = \boldsymbol{\phi}_m(t)^T\mathbf{a}_i = \mathbf{a}_i^T\boldsymbol{\phi}_m(t)$.
  - $\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \mathrm{Cov}[\mathbf{a}_i^T\boldsymbol{\phi}_m(t_1), \mathbf{a}_i^T\boldsymbol{\phi}_m(t_2)]$
  - $\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \boldsymbol{\phi}_m(t_1)^T \mathbf{C}_G \boldsymbol{\phi}_m(t_2)$

Covariance function (cont)

- $Cov[A_i(t_1), A_i(t_2)] = \phi_m(t_1)^T C_G \phi_m(t_2)$
- Recall for $\mathbf{t}_m = (1, t, t^2, \ldots, t^m)^T$ that
  - $\phi_m(t) = M\mathbf{t}_m$, where M is the (m+1) x (m+1) matrix of coefficients for the first (m+1) Legendre polynomials
- Substituting in $\phi(t) = \mathbf{M}t$ yields
  - $Cov[A_i(t_1), A_i(t_2)] = t_1^T M^T C_G M t_2$, or
  - $Cov[A_i(t_1), A_i(t_2)] = t_1^T H t_2$, with $H = M^T C_G M$
    - This allows us to express the covariance function in terms t_1 and t_2 directly

From G to C_G

- The key component to the covariance function is the covariance matrix C_G for the additive genetic random regression coefficients. How do we obtain this?
- We start with what Kirkpatrick called the "full estimate"
  - Given an estimated G matrix of the trait measured at m time points, we can describe trait breeding value as an m-1 degree polynomial
  - This is done as a weighted combination of the first m Legendre polynomials, $\phi_0, \phi_1, \ldots \phi_{m-1}$.
  - $G_{ij} = Cov[A(t_i), A(t_j)] = \phi_m(t_i) C_G \phi_m(t_j)^T$

The full estimate does an element-by-element matching of **G** to functions of $\phi_m(t_i)$ (which are known constants) and C_G.

$$\mathbf{G} = \begin{pmatrix} G_{11} & \cdots & G_{1m} \\ \vdots & \ddots & \vdots \\ G_{m1} & \cdots & G_{mm} \end{pmatrix}, \quad \text{where} \quad G_{ij} = \boldsymbol{\phi}^T(t_i)\mathbf{G_C}\boldsymbol{\phi}(t_j)$$

$$= \begin{pmatrix} \boldsymbol{\phi}^T(t_1)\mathbf{G_C}\boldsymbol{\phi}(t_1) & \cdots & \boldsymbol{\phi}^T(t_1)\mathbf{G_C}\boldsymbol{\phi}(t_m) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\phi}^T(t_m)\mathbf{G_C}\boldsymbol{\phi}(t_1) & \cdots & \boldsymbol{\phi}^T(t_m)\mathbf{G_C}\boldsymbol{\phi}(t_m) \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\phi}^T(t_1) \\ \vdots \\ \boldsymbol{\phi}^T(t_m) \end{pmatrix} \mathbf{G_C} \begin{pmatrix} \boldsymbol{\phi}(t_1) \\ \vdots \\ \boldsymbol{\phi}(t_m) \end{pmatrix} = \boldsymbol{\Phi}^T\mathbf{G_C}\boldsymbol{\Phi}$$

$$\mathbf{G} = \boldsymbol{\Phi}^T\mathbf{G_C}\boldsymbol{\Phi} \quad \text{implies} \quad \mathbf{G_C} = \left(\boldsymbol{\Phi}^T\right)^{-1}\mathbf{G}\boldsymbol{\Phi}^{-1}$$

where

$$\boldsymbol{\Phi}^T = \begin{pmatrix} \boldsymbol{\phi}^T(t_1) \\ \boldsymbol{\phi}^T(t_2) \\ \vdots \\ \boldsymbol{\phi}^T(t_m) \end{pmatrix} = \begin{pmatrix} \phi_0(t_1) & \phi_1(t_1) & \cdots & \phi_{m-1}(t_1) \\ \phi_0(t_2) & \phi_1(t_2) & \cdots & \phi_{m-1}(t_2) \\ \vdots & \vdots & \cdots & \vdots \\ \phi_0(t_m) & \phi_1(t_m) & \cdots & \phi_{m-1}(t_m) \end{pmatrix}$$

Note that $\boldsymbol{\Phi}$ is a matrix of constants --- the Legendre polynomials evaluated at the sample time points. Note that time points are scaled to be within (-1, 1), so ordering time on the original scale as $T_1 < \ldots < T_m$, scaled values are given by $t_i = 2(T_i - T_1)/(T_m - T_1) - 1$

Example: Riska's data

$$\mathbf{G} = \begin{pmatrix} 436.0 & 522.3 & 424.2 \\ 522.3 & 808.0 & 664.7 \\ 424.2 & 664.7 & 558.0 \end{pmatrix}$$

$$\boldsymbol{\Phi}^T = \begin{pmatrix} \phi_0(-1) & \phi_1(-1) & \phi_2(-1) \\ \phi_0(0) & \phi_1(0) & \phi_2(0) \\ \phi_0(1) & \phi_1(1) & \phi_2(1) \end{pmatrix}$$ ⟵ 2 weeks, t = -1 ⟵ 3 weeks, t = 0 ⟵ 4 weeks, t = 1

$$= \begin{pmatrix} 0.7071 & -1.2247 & 1.5811 \\ 0.7071 & 0 & -0.7906 \\ 0.7071 & 1.2247 & 1.5811 \end{pmatrix}$$

$$\mathbf{G_C} = \left(\boldsymbol{\Phi}^T\right)^{-1} \mathbf{G}\boldsymbol{\Phi}^{-1} = \begin{pmatrix} 1348.1 & 66.6 & -111.7 \\ 66.6 & 24.2 & -14.0 \\ -111.7 & -14.0 & 14.5 \end{pmatrix}$$

```
> G<-matrix(c(436.0,522.3,424.2,522.3,808.0,664.7,424.2,664.7,558.0),nrow=3)
> G
      [,1]  [,2]  [,3]
[1,] 436.0 522.3 424.2
[2,] 522.3 808.0 664.7
[3,] 424.2 664.7 558.0
> Phi<-matrix(c(0.7071,0.7071,0.7071,-1.2247,0,1.2247,1.5811,-0.7906,1.5811),nrow=3)
> Phi
       [,1]    [,2]    [,3]
[1,] 0.7071 -1.2247  1.5811
[2,] 0.7071  0.0000 -0.7906
[3,] 0.7071  1.2247  1.5811
>  solve(Phi)%*% G %*% solve(t(Phi))
           [,1]      [,2]       [,3]
[1,] 1348.14866  66.55166 -111.68492
[2,]   66.55166  24.26844  -14.01216
[3,] -111.68492 -14.01216   14.50677
```

The resulting covariance function becomes

$$\mathrm{Cov}(t_1, t_2) = \boldsymbol{\phi}^T(t_1)\mathbf{G_C}\boldsymbol{\phi}(t_2)$$

$$= \begin{pmatrix} \phi_0(t_1) & \phi_1(t_1) & \phi_2(t_1) \end{pmatrix} \begin{pmatrix} 1348.1 & 66.6 & -111.7 \\ 66.6 & 24.2 & -14.0 \\ -111.7 & -14.0 & 14.5 \end{pmatrix} \begin{pmatrix} \phi_0(t_1) \\ \phi_1(t_1) \\ \phi_2(t_1) \end{pmatrix}$$

This bilinear form expresses the covariance function in terms of the Legendre polynomials. Usually we would like to express this as a polynomial in t_1 & t_2:

One could do this by first substituting in the polynomial form for $\phi_i(t)$, expanding and collecting terms. However, much easier to do this in matrix form. Recall the coefficient matrix **M** from earlier in the notes, where $\boldsymbol{\phi} = \mathbf{M}\mathbf{t}$. Writing the covariance function as $\boldsymbol{\phi}_1^T \mathbf{G_C} \boldsymbol{\phi}_2 = (\mathbf{M}\mathbf{t}_1)^T \mathbf{G_C}(\mathbf{M}\mathbf{t}_2) = \mathbf{t}_1^T \mathbf{M}^T\mathbf{G_C}\mathbf{M}\,\mathbf{t}_2 = \mathbf{t}_1^T \mathbf{H}\,\mathbf{t}_2$, where $\mathbf{H} = \mathbf{M}^T\mathbf{C_G}\mathbf{M}$.

The covariance function becomes $\mathbf{t}_1^T \mathbf{H}\,\mathbf{t}_2$, with $\mathbf{H} = \mathbf{M}^T\mathbf{C_G}\mathbf{M}$

Since the first three Legendre polynomials are used, **M** is 3 x 3

$$\mathbf{M} = \begin{pmatrix} 0.7071 & 0 & 0 \\ 0 & 1.2247 & 0 \\ -0.7906 & 0 & 2.3717 \end{pmatrix}$$

$\mathbf{H} = \mathbf{M}^T\mathbf{C_G}\mathbf{M}$ gives

$$\mathbf{H} = \begin{pmatrix} 808.0 & 71.2 & -214.5 \\ 71.2 & 36.4 & -40.7 \\ -214.5 & -40.7 & 81.6 \end{pmatrix}$$

Expanding this out gives

$\mathrm{Cov}(A_1,A_2) = 808 + 71.2(t_1 + t_2) + 36.4\, t_1\, t_2$
$- 40.7(t_1^2\, t_2 + t_1 t_2^2) - 215.0(t_1^2 + t_2^2)$
$+ 81.6 t_1^2 t_2^2$

More generally, the coefficient on $t_1^{i-1}\, t_2^{j-1}$ in the covariance expansion is given by H_{ij}. -- the (i,j)-th element of **H**.

The Eigenstructure of C_G

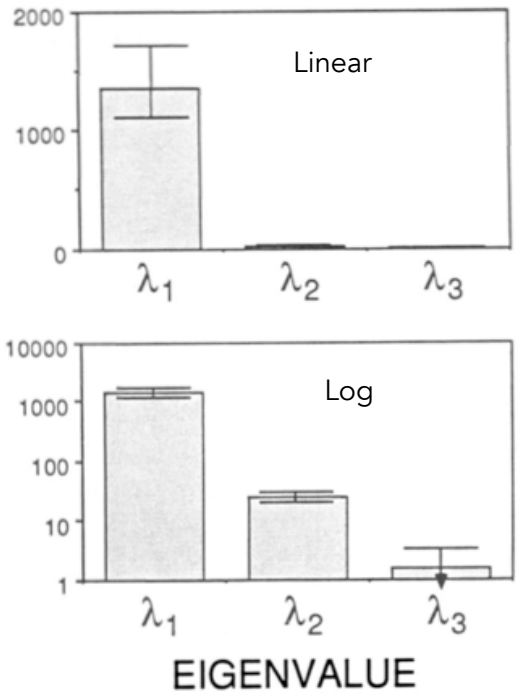
- The variance-covariance matrix C_G of the random regression coefficients is extremely information on the nature of variation for the function-valued trait.
- The function-valued analogue of the eigenvector is the eigenfunction, which also has an associated eigenvalue. Akin to the eigvenvector associated with the largest eigenvalue accounting for the largest single direction of variation, the eigenfunction associated with the largest eigenvalue is the functional curve associated with the most variation.
- The eigenvalues of C_G are the same as those for the covariance function, while the associated eigenvectors of C_G give the weights on the orthogonal polynomials that recover the eigenfunctions of the covariance function.

Back to Riska's data

$$\mathbf{G_C} = \left(\boldsymbol{\Phi}^T\right)^{-1} \mathbf{G}\boldsymbol{\Phi}^{-1} = \begin{pmatrix} 1348.1 & 66.6 & -111.7 \\ 66.6 & 24.2 & -14.0 \\ -111.7 & -14.0 & 14.5 \end{pmatrix}$$

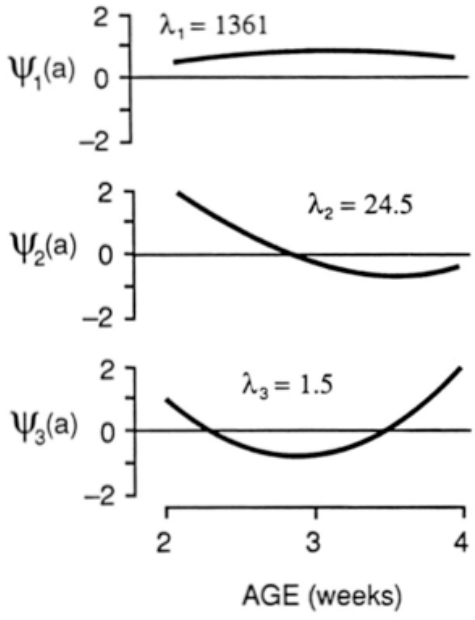
```
> eigen(CG)
$values
[1] 1360.844364   24.544765    1.534744

$vectors
           [,1]        [,2]        [,3]
[1,] -0.99526560  0.07934234 -0.05613532
[2,] -0.05042796 -0.91529538 -0.39961406
[3,]  0.08308671  0.39489133 -0.91496308
```

First eigenvector



CG has a dominant eigenvalue --- most of the variation in the breeding value for growth is along one curve

Associated eigenfunctions for C_G for the Riska dataset



Eigenfunctions of C_G

- If $\mathbf{e}_i$ denotes the eigenvector associated with the ith eigenvalue λ_i of C_G, then for the covariance function
  - λ_i is the ith eigenvalue
  - associated eigenfunction is $\underline{\phi}_m(t)^T \mathbf{e}_i$
  - $= e_{i1}\phi_0(t) + e_{i2}\phi_1(t) + \cdots + e_{im}\phi_{m-1}(t)$
  - Since $\phi = \mathbf{M}t$, we have $(\mathbf{M}t)^T \mathbf{e}_i = t^T (M^T \mathbf{e}_i)$, giving the weights on $(1, t, t^2, .. , t^{m-1})$ as $M^T \mathbf{e}_i$
  - For Riska's data, the leading eigenfunction is
  - $\psi_1(t) = 0.7693 - 0.0617\ t - 0.1971\ t^2$

Eigenfunctions: $\psi_i(t) = t^T (M^T e_i)$

$$\mathbf{M} = \begin{pmatrix} 0.7071 & 0 & 0 \\ 0 & 1.2247 & 0 \\ -0.7906 & 0 & 2.3717 \end{pmatrix}$$

$$\mathbf{e}_1 = \begin{pmatrix} 0.995 \\ 0.050 \\ -0.083 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} -0.079 \\ 0.915 \\ -0.395 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0.056 \\ 0.400 \\ 0.915 \end{pmatrix}$$

$$\mathbf{M}^T\mathbf{e}_1 = \begin{pmatrix} 0.769 \\ 0.062 \\ -0.197 \end{pmatrix}, \quad \mathbf{M}^T\mathbf{e}_2 = \begin{pmatrix} 0.256 \\ 1.121 \\ -0.937 \end{pmatrix}, \quad \mathbf{M}^T\mathbf{e}_3 = \begin{pmatrix} -0.684 \\ 0.490 \\ 2.170 \end{pmatrix}$$

$\psi_2(t) = 0.256 + 1.121*t - 0.937*t^2$

$\psi_3(t) = -0.684 + 0.490*t + 2.170*t^2$

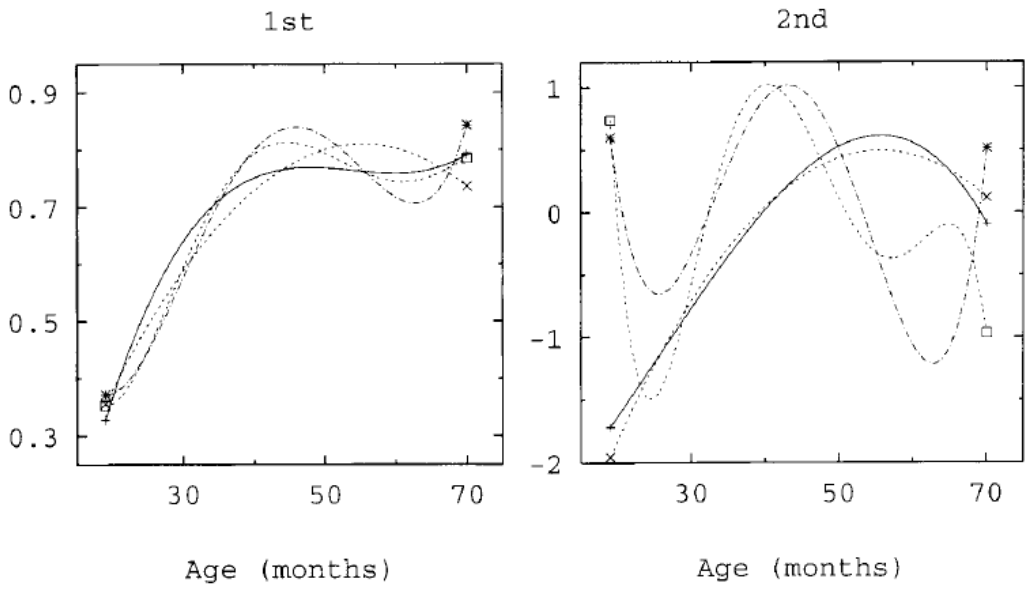


Figure 3. Estimated first and second eigenfunction of the genetic covariance function, for orders of polynomial fit of 3 (×), 4 (+), 5 (∗) and 6 (□), respectively (rank 3 estimates of the coefficient matrices).

Meyer's data on Cattle Weight

Over-fitting G_C?

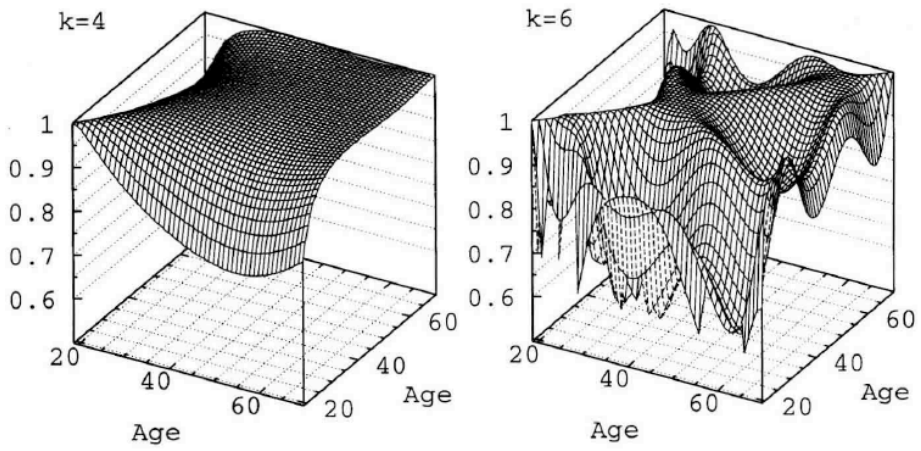


Figure 5. Estimates of genetic correlations for orders of polynomial fit (k) of 4 and 6.

Meyer's data showing how increasing the degree of polynomial used results in over-fitting. In her words: "surfaces become 'wiggly' "

Reduced estimation of $\mathbf{C_G}$

- While the full estimate (rank $\mathbf{C_G}$ = rank of observed $\mathbf{G}$) is (relatively) straightforward, this likely results in an overfit of the data, as the covariance function is forced to exactly fit the observed values for all t_1, t_2, some of which are sampling noise
  - Results in a less smooth covariance function than one based on using a reduced dimension.
  - Kirkpatrick originally suggested a least-squares approach, while Meyer & Hill suggested a REML-based approach
  - Key breakthrough, first noticed by Goddard, and fully developed by Meyer, is the connection between covariance functions and random regressions.
  - This should not be surprising given that we started with random regressions to motivate covariance functions.
  - The key is that standard BLUP approaches (for multivariate traits) can be used for random regressions.

Mixed-Models (BLUPs) for Longitudinal traits

- Simplest setting is the repeatability model, the trait breeding and residual (permanent environmental) values are assumed constant over time. The jth observation on i is
  - $y_{ij} = u + a_i + pe_i + e_{ij}$
  - $\mathbf{a} \sim \mathbf{0}$, Var(A)$\mathbf{A}$
- At the other extreme is the multiple-trait approach, where each sampled time point is considered as a separate, but correlated, trait. Here y_{ij} is the jth "trait" (sampled time point) for individual i.
  - $y_{ij} = u + a_{ij} + e_{ij}$
  - $\mathbf{a} \sim \mathbf{0}$, $\mathbf{G} \otimes \mathbf{A}$
- In the middle are random-regressions, where for the jth observation (time t_j) on individual i is
  - $y_{ij} = u + \Sigma_k^n a_{ik}\phi_k(t_j) + \Sigma_k^m pe_{ik}\phi_k(t_j) + e_{ij}$
  - $\mathbf{a_i} \sim \mathbf{0}$, $\mathbf{C_G}$ and $\mathbf{p_i} \sim \mathbf{0}$, $\mathbf{C_E}$

The repeatability model

- The repeatability model assumes that the trait is unchanging between observations, but multiple observations (records) are taken over time to smooth out sampling noise (e)
- Such a record for individual k has three components
  - Breeding value a_k
  - Common (permanent) environmental value p_k
  - Residual value for ith observation e_{ki}
- Resulting observation is thus
  - $z_{ki} = \mu + a_k + p_k + e_{ki}$
- The repeatability of a trait is $r = (\sigma_A^2+\sigma_p^2)/\sigma_z^2$
- Resulting variance of the residuals is $\sigma_e^2 = (1-r)\,\sigma_z^2$

Mixed-model $y = X\beta + Za + Zp + e$

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_p^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$$

Mixed-model equations

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda_A\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\mathbf{Z} + \lambda_u\mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}$$

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} = \frac{1-r}{h^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_p^2} = \frac{1-r}{r-h^2}$$

The multiple-trait model

- With a clearly discrete number of stages (say k), a longitudinal trait could be modeled as k correlated traits, so that individual i has values y_{i1}, y_{i2}, .., y_{ik}.
- In this case, there is no need for permanent environmental effects, as these now appear in correlations among the residuals, the within-individual environmental correlations (which are estimated by REML).
- This can be put into standard Mixed Model equations by simply "stacking" the vectors for each trait to create one vector for each random effect.

For trait j ($1 \leq j \leq k$), the mixed model becomes

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{a}_i + \mathbf{e}_j$$

$$\begin{pmatrix}\mathbf{a}_j \\ \mathbf{e}_j\end{pmatrix} \sim \begin{pmatrix}\mathbf{0} \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\sigma^2_{A_j}\,\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{e_j}\,\mathbf{I}\end{pmatrix}$$

We can write this as $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{a} + \mathbf{e}$, where

$$\begin{pmatrix}\mathbf{y}_1 \\ \vdots \\ \mathbf{y}_k\end{pmatrix} = \begin{pmatrix}\mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_k\end{pmatrix}\begin{pmatrix}\boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k\end{pmatrix} + \begin{pmatrix}\mathbf{Z}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_k\end{pmatrix}\begin{pmatrix}\mathbf{a}_1 \\ \vdots \\ \mathbf{a}_k\end{pmatrix} + \begin{pmatrix}\mathbf{e}_1 \\ \vdots \\ \mathbf{e}_k\end{pmatrix}$$

Again, the BLUP for the vector of all EBVs is given by

$$\hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

With $\mathbf{V}$ the covariance structure for this model

Covariance structure for EBVS

The resulting covariance structure for the stacked vector of breeding values is

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_k\end{pmatrix} = \begin{pmatrix}\sigma^2(A_1)\mathbf{A} & \cdots & \sigma(A_1,A_k)\mathbf{A}\\ \vdots & \ddots & \vdots\\ \sigma(A_k,A_1)\mathbf{A} & \cdots & \sigma^2(A_k)\mathbf{A}\end{pmatrix} = \mathbf{G}\otimes\mathbf{A}$$

where $\otimes$ denotes the Kronecker (or direct) product (LW Chapter 26) and

$$\mathbf{G} = \begin{pmatrix}\sigma^2(A_1) & \cdots & \sigma(A_1,A_k)\\ \vdots & \ddots & \vdots\\ \sigma(A_k,A_1) & \cdots & \sigma^2(A_k)\end{pmatrix}$$

is the matrix of genetic covariances of interest.

The genetic variance-covariance matrix G accounts for the genetic covariances among traits. G has k variances and k(k-1)/2 covariances, which must be estimated (REML) from the data.

Covariance structure for residuals

Similarly, the covariance structure for the stacked vectors of residuals is

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_k\end{pmatrix} = \mathbf{E}\otimes\mathbf{I}, \quad \text{where} \quad \mathbf{E} = \begin{pmatrix}\sigma^2(e_1) & \cdots & \sigma(e_1,e_k)\\ \vdots & \ddots & \vdots\\ \sigma(e_k,e_1) & \cdots & \sigma^2(e_k)\end{pmatrix}$$

Finally, we need to specify any covariances between $\mathbf{a}$ and $\mathbf{e}$. By construction $\sigma(a_z,e_z) = \sigma(a_w,e_w) = 0$, while the standard assumption is $\sigma(A_z,e_w) = \sigma(A_w,e_z) = 0$, giving the covariance structure as

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_k\\ \mathbf{e}_1\\ \vdots\\ \\ \mathbf{e}_k\end{pmatrix} = \begin{pmatrix}\mathbf{G}\otimes\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{E}\otimes\mathbf{I}\end{pmatrix}$$

Here the matrix E accounts for within-individual correlations in the environmental (or residual) values.

Random regressions

- Random regression models are basically a hybrid between repeated records models and multiple-trait models.
  - The basic structure of the model is that the trait at time t is the sum of potentially time-dependent fixed effects $\mu(t)$, a time-dependent breeding value $a(t)$, a time-dependent permanent environmental effect $p(t)$, and a residual error e. These last three are random effects
  - $y(t) = \mu(t) + a(t) + p(t) + e$
  - $a(t)$ and $p(t)$ are both approximated by random regressions, of order n and m, respectively (usually n = m)
  - $a_i(t_j) = \Sigma_k^n a_{ik}\phi_k(t_j)$ and $p_i(t_j) = \Sigma_k^m b_{ik}\phi_k(t_j)$
  - The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ for individual i are handled in a multiple-trait framework, with covariance matrices $\mathbf{C}_G$ and $\mathbf{C}_E$ for the within-individual vectors of additive and permanent environmental effects.

To build up the random regression model, consider the q_i observations from different times for individual i

$$\mathbf{y}_i = \begin{pmatrix} y(t_{i1}) \\ \vdots \\ y(t_{iq_i}) \end{pmatrix} = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{i1}\mathbf{a}_i + \mathbf{Z}_{i2}\mathbf{p}_i + \mathbf{e}_i$$

$$\mathbf{a}_i = \begin{pmatrix} a_{i0} \\ \vdots \\ a_{im} \end{pmatrix}, \quad \mathbf{p}_i = \begin{pmatrix} p_{i0} \\ \vdots \\ p_{im} \end{pmatrix}, \quad \mathbf{e}_i = \begin{pmatrix} e_{i0} \\ \vdots \\ e_{im} \end{pmatrix}$$

Here are fitting m-degree polynomials ($m < q_i$) for both the breeding value and permanent environmental value regressions. We also assume that any fixed-effects are not time dependent. Both of these assumptions are easily relaxed.

Model & covariance structure for vector $\mathbf{y}_i$ of observations from individual i

$$\mathbf{y}_i = \begin{pmatrix} y(t_{i1}) \\ \vdots \\ y(t_{iq_i}) \end{pmatrix} = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{i1}\mathbf{a}_i + \mathbf{Z}_{i2}\mathbf{p}_i + \mathbf{e}_i$$

$$\mathbf{a}_i = \begin{pmatrix} a_{i0} \\ \vdots \\ a_{im} \end{pmatrix}, \quad \mathbf{p}_i = \begin{pmatrix} p_{i0} \\ \vdots \\ p_{im} \end{pmatrix}, \quad \mathbf{e}_i = \begin{pmatrix} e_{i0} \\ \vdots \\ e_{im} \end{pmatrix}$$

Covariance structure

$$\begin{pmatrix} \mathbf{a}_i \\ \mathbf{p}_i \\ \mathbf{e}_i \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{C_G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_E} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2\mathbf{I} \end{pmatrix}$$

The design matrix for the regression coefficients on the breeding values is very information

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{i1}\mathbf{a}_i + \mathbf{Z}_{i2}\mathbf{p}_i + \mathbf{e}_i$$

$$\mathbf{Z}_{i1} = \begin{pmatrix} \phi_0(t_{i1}) & \cdots & \phi_m(t_{i1}) \\ \phi_0(t_{i2}) & \cdots & \phi_m(t_{i2}) \\ \vdots & \ddots & \vdots \\ \phi_0(t_{iq_i}) & \cdots & \phi_m(t_{iq_i}) \end{pmatrix}$$

$\mathbf{Z_{i1}}$ is a q_i x (m+1) matrix of fixed constants that depend on the values of order zero through m Legendre polynomials, where the jth row represents these evaluated at time t_{ij}.
A KEY FEATURE is that this set of times could be different for each individual, yet the mixed model does all the bookkeeping to fully account for this.

As with the multiple trait model, stacking the individual vectors allows us to put this model in standard form. Note that while the vectors stacked for the multiple trait model represented the vectors for each trait separately, here the stacked vectors are the observations for each individual.

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e}$$

$\mathbf{Z}_1$, $\mathbf{Z}_2$ Block diagonal $\quad \mathbf{Z}_1 = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{12} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{Z}_{1n} \end{pmatrix}$

Full Model & covariance structure

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e}$$

Covariance structure

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{A} \otimes \mathbf{C_G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \otimes \mathbf{C_E} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \mathbf{I} \end{pmatrix}$$

More generally, we can replace $\sigma_e^2 \mathbf{I}$ by $\mathbf{R}$.

Mixed-model equations (slightly more generalized covariance structure)

$$\mathbf{H}\begin{pmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{a}}\\ \widehat{\mathbf{p}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{y}\end{pmatrix}\quad\begin{pmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{pmatrix}\sim\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix},\begin{pmatrix}\mathbf{A}\otimes\mathbf{C_G} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\otimes\mathbf{C_E} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{pmatrix}$$

where

$$\mathbf{H}=\begin{pmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_2\\ \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{Z}_1+\mathbf{A}^{-1}\otimes\mathbf{C_G^{-1}} & \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{Z}_2\\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_2+\mathbf{I}\otimes\mathbf{C_E^{-1}}\end{pmatrix}$$

Model-fitting issues

- A central issue is what degree m of polynomials to use.
- Standard likelihood tests can be used (compare m = k with m = k + 1).
- Meyer suggests that tests should be comparing k with k + 2, as often going from odd to even does not improve fit, but going from even to even (k+2) does, and vice-versa.

Response to selection

- Standard BLUP selection can be used, based on some criteria for an optimal functional value (curve) in the offspring.
- The expected response in the offspring is simply obtained by substituting the average of the parental breeding values into the polynomial regression for the breeding value to generate an expected offspring curve.