第二天回顾

- 微效等位基因向后代的传递
- 方差、协方差
- 多元正态分布
- 广义最小二乘的原理
- BLUE 是固定效应的广义最小二乘估计
- BLUP 和混合线性方程组

第三讲： A 及其逆矩阵、模型举例，基因组选择

于希江

挪威生命科学大学畜牧与水产系

二〇一九 • 十月
青岛

目录

两个半同胞个体

$$\mathrm{Var}(\mathrm{y}) = \mathrm{ZGZ}' + \mathrm{R}$$

$$\mathrm{Z} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathrm{I}$$

$$\mathrm{R} = \mathrm{I}\sigma_e^2$$

$$\mathrm{G} = \begin{bmatrix} 1 & .25 \\ .25 & 1 \end{bmatrix} \sigma_a^2$$

$$= \mathrm{A}\sigma_a^2$$

IBD 和 IBS

IBD Identical by descent. 指两个基因是同一个基因的复制品。

IBS Identical by state. 指两个基因（序列）相同。

- 通常用一个个体基因组中 IBD 的比例来表示个体的近交程度。
- 通常用两个个体之间的 IBD 比例来表示这两个个体之间的亲缘关系。

加性关系矩阵

- A 可以用表格法递推
- A^{-1} 也可以直接构建。
  - Henderson 1975
  - Meuwissen and Luo 1992
- A 的元素相当于 Wright 遗传相关的分子
  - 因此又称分子血缘相关系数矩阵。

加性关系矩阵

- A 可以用表格法递推
- A^{-1} 也可以直接构建。
  - Henderson 1975
  - Meuwissen and Luo 1992
- A 的元素相当于 Wright 遗传相关的分子
  - 因此又称分子血缘相关系数矩阵。

Table: 一个排序后的系谱示例

个体	父	母
A	–	–
B	–	–
C	–	–
D	A	B
E	A	C
F	E	D

表格法

	−,−	−,−	−,−	A,B	A,C	E,D
	A	B	C	D	E	F
A	1	0	0			
B	0	1	0			
C	0	0	1			
D						
E						
F						

- 父母未知的个体
  - 令其对角线为元素 1
  - 互相之间的关系为 0

表格法续一

	–,–	–,–	–,–	A,B	A,C	E,D
	A	B	C	D	E	F
A	1	0	0	.5	.5	.5
B	0	1	0	.5	0	.25
C	0	0	1	0	.5	.25
D	.5	.5	0			
E	.5	0	.5			
F	.5	.25	.25			

$$a_{\mathrm{AD}} = .5(a_{\mathrm{AA}} + a_{\mathrm{AB}}) = .5(1+0) = .5$$

$$a_{\mathrm{AE}} = .5(a_{\mathrm{AA}} + a_{\mathrm{AC}}) = .5(1+0) = .5$$

$$a_{\mathrm{AF}} = .5(a_{\mathrm{AE}} + a_{\mathrm{AD}}) = .5(.5+.5) = .5$$

一般的，若 j 的父母为 p 和 q，则：

$$a_{ij} = .5(a_{ip} + a_{iq})$$

表格法续二

	-,-	-,-	-,-	A,B	A,C	E,D
	A	B	C	D	E	F
A	1	0	0	.5	.5	.5
B	0	1	0	.5	0	.25
C	0	0	1	0	.5	.25
D	.5	.5	0	1	.25	.625
E	.5	0	.5	.25	1	.625
F	.5	.25	.25	.625	.625	1.125

父母已知的个体，依定义，其近交系数 F_i 为其父母的加性遗传关系的一半。如：

$$F_{\mathrm{D}} = .5a_{\mathrm{AB}} = 0$$

个体 D 与其本身的加性遗传关系为：

$$a_{\mathrm{DD}} = 1 + F_{\mathrm{D}} = 1 + 0$$

一般的，若个体 i 的父母为 p 和 q:

$$a_{ii} = 1 + .5a_{pq}$$

一般规则

- 当 A 维数小时，可以直接计算 A^{-1}
  - 如果是根据（SNP）基因型计算的 G，则只能硬算 G^{-1}
  - 或者用特殊的数值算法
- 根据系谱，有简单的方法直接计算 A^{-1}，而不用计算A
  - Henderson 1976；Quaas，1976
- 见 day-3-a-mat.ipynb

目录

公牛模型

$$y = xb + zu + e$$

- 有来自 2 个奶牛场的 2 头公牛的 5 头女儿的记录如下：

个体号	场	公牛	表型值
1	1	1	11
2	1	1	15
3	2	1	10
4	1	2	19
5	2	2	25

公牛模型

$$y = xb + zu + e$$

- 有来自 2 个奶牛场的 2 头公牛的 5 头女儿的记录如下：

个体号	场	公牛	表型值
1	1	1	11
2	1	1	15
3	2	1	10
4	1	2	19
5	2	2	25

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

线性混合模型观测值举例

- 公牛模型

$$
\begin{array}{ccccccccc}
\text{y} & = & \text{y} & = & \text{Xb} & + & \text{Zu} & + & \text{e} \\
\begin{bmatrix} y_{111} \\ y_{112} \\ y_{213} \\ y_{121} \\ y_{222} \end{bmatrix} & = & \begin{bmatrix} 11 \\ 15 \\ 10 \\ 19 \\ 25 \end{bmatrix} & = & \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} & + & \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} & + & \begin{bmatrix} e_{111} \\ e_{112} \\ e_{213} \\ e_{121} \\ e_{222} \end{bmatrix}
\end{array}
$$

混合线性模型方程组

令 $h^2 = .25$

$$\begin{aligned}\lambda &= \frac{\sigma_e^2}{\sigma_u^2} \\ &= \frac{(1-.25h^2)\sigma_P^2}{.25h^2\sigma_P^2} \\ &= \frac{1-.25h^2}{.25h^2} \\ &= \frac{4-h^2}{h^2} \\ &= \frac{4-.25}{.25} = 15\end{aligned}$$

若两头公牛无亲缘关系，则 $\mathrm{A} = \mathrm{I} = \mathrm{A}^{-1}$：

$$\begin{bmatrix} 3 & 0 & 2 & 1 \\ 0 & 2 & 1 & 1 \\ 2 & 1 & 3+\lambda & 0 \\ 1 & 1 & 0 & 2+\lambda \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} 45 \\ 35 \\ 36 \\ 44 \end{bmatrix}$$

MME

$$\begin{bmatrix} 3 & 0 & 2 & 1 \\ 0 & 2 & 1 & 1 \\ 2 & 1 & 18 & 0 \\ 1 & 1 & 0 & 17 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} 45 \\ 35 \\ 36 \\ 44 \end{bmatrix}$$

MME

$$\begin{bmatrix} 3 & 0 & 2 & 1 \\ 0 & 2 & 1 & 1 \\ 2 & 1 & 18 & 0 \\ 1 & 1 & 0 & 17 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} 45 \\ 35 \\ 36 \\ 44 \end{bmatrix}$$

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} 15.22 \\ 17.50 \\ -0.66 \\ 0.66 \end{bmatrix}$$

动物模型

模型

$$y_i = \mu + a_i + e_i$$

假设

1. 群体足够大
2. 后代无选择
3. 假定只有加性作用
4. 除了模型中的效应，无其它作用

$$\mathrm{E}(a_i) = 0, \qquad \mathrm{E}(e_i) = 0$$

$$\mathrm{Var}(a_i) = \sigma_a^2$$

$$\mathrm{Var}(e_i) = \sigma_e^2$$

$$\mathrm{Var}(\mathrm{a}) = \mathrm{A}\sigma_a^2$$

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

$$\lambda = \frac{\sigma_e^2}{\sigma_a^2}$$

示例

ID	Pa	Ma	PT	ID	Pa	Ma	PT
1	0	0	-	9	5	6	36.0
2	0	0	-	10	7	8	66.4
3	0	0	-	11	5	8	28.9
4	0	0	-	12	7	6	73.0
5	1	2	38.5	13	1	6	44.2
6	3	4	48.9	14	3	8	53.4
7	3	2	64.3	15	5	4	33.6
8	1	4	50.5	16	7	8	49.5

$$\sigma_a^2 = 36$$

$$\sigma_e^2 = 64$$

$$h^2 = .36$$

$$\lambda = \frac{16}{9}$$

每个表型观测值的模型

$$
\begin{bmatrix} 1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1 \end{bmatrix}\mu +
\begin{bmatrix}
0&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1
\end{bmatrix}
\begin{bmatrix} u_1\\u_2\\u_3\\u_4\\u_5\\u_6\\u_7\\u_8\\u_9\\u_{10}\\u_{11}\\u_{12}\\u_{13}\\u_{14}\\u_{15}\\u_{16} \end{bmatrix}
+
\begin{bmatrix} e_5\\e_6\\e_7\\e_8\\e_9\\e_{10}\\e_{11}\\e_{12}\\e_{13}\\e_{14}\\e_{15}\\e_{16} \end{bmatrix}
=
\begin{bmatrix} 38.5\\48.9\\64.3\\50.5\\36.0\\66.4\\28.9\\73.0\\44.2\\53.4\\33.6\\49.5 \end{bmatrix}
$$

$$
A = \frac{1}{16}\begin{bmatrix}
16 &&&&&&&&&&&&&&& \\
0 & 16 &&&&& s & y & m & m & e & t & r & y && \\
0 & 0 & 16 &&&&&&&&&&&&& \\
0 & 0 & 0 & 16 &&&&&&&&&&&& \\
8 & 8 & 0 & 0 & 16 &&&&&&&&&&& \\
0 & 0 & 8 & 8 & 0 & 16 &&&&&&&&&& \\
0 & 8 & 8 & 0 & 4 & 4 & 16 &&&&&&&&& \\
8 & 0 & 0 & 8 & 4 & 4 & 0 & 16 &&&&&&&& \\
4 & 4 & 4 & 4 & 8 & 8 & 4 & 4 & 16 &&&&&&& \\
4 & 4 & 4 & 4 & 4 & 4 & 8 & 8 & 4 & 16 &&&&&& \\
8 & 4 & 0 & 4 & 10 & 2 & 2 & 10 & 6 & 6 & 18 &&&&& \\
0 & 4 & 8 & 4 & 2 & 10 & 10 & 2 & 6 & 6 & 2 & 18 &&&& \\
8 & 0 & 4 & 4 & 4 & 8 & 2 & 6 & 6 & 4 & 5 & 5 & 16 &&& \\
4 & 0 & 8 & 4 & 2 & 6 & 4 & 8 & 4 & 6 & 5 & 5 & 5 & 16 && \\
4 & 4 & 0 & 8 & 8 & 4 & 2 & 6 & 6 & 4 & 7 & 3 & 4 & 3 & 16 & \\
4 & 4 & 4 & 4 & 4 & 4 & 8 & 8 & 4 & 8 & 6 & 6 & 4 & 6 & 4 & 16
\end{bmatrix}
$$

X, Z, y 已知，$G = A_{16}\sigma_a^2$，$R = I_{12}\sigma_e^2$

混合线性模型方程组

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

目录

若干概念

EBV 传统的估计育种值，如来自 BLUP 的。

DGV Direct genmoic BV.

GEBV Genomically enhanced/ enabled/estimated BV

SNP Single nucleotype polymorphism

Chips 芯片

若干概念

EBV 传统的估计育种值，如来自 BLUP 的。

DGV Direct genmoic BV.

GEBV Genomically enhanced/ enabled/estimated BV

SNP Single nucleotype polymorphism

Chips 芯片

一个著名结果

Table: Meuwissen et al., 2001

	$r_{\text{TBV,EBV}}$	$b_{\text{TBV,EBV}}$
LS	0.32	0.29
genomic BLUP	0.73	0.90
Bayes A	0.80	0.83
Bayes B	0.85	0.95

基因型信息

基因组选择之前

- 用于淘汰或固定某些基因，例如遗传缺陷、疾病
- 有一些标记被用于标记辅助选择，以改进育种值的估计

基因型信息

基因组选择之前

- 用于淘汰或固定某些基因，例如遗传缺陷、疾病
- 有一些标记被用于标记辅助选择，以改进育种值的估计

基因组选择

- 用到成千上万的基因座位
- 对单个基因的作用通常不再感兴趣
- 目的要完全通过这些标记基因（与 QTL 的 LD）来估计育种值

一般模型

$$\mathrm{y} = 1_{\mathrm{n}}\mu + \mathrm{Z}\mathrm{g} + \mathrm{e}$$

- 其中

y	$n \times 1$ 观察值向量
1_n	元素为 1 的向量
μ	未知的一般平均数
Z	$n \times m$ 的标记基因型矩阵
g	$m \times 1$ （随机的）SNP 标记效应
e	随机残差向量

基因组选择

- 需要的数据
  - 基因型
  - 表现型
  - 系谱【可选】
- 关于 Z
  - 筛除低质量的标记
  - 有很多基因型缺失的座位（如：$> 20\%$）
  - 设置 MAF，如 1%
  - 缺失基因型的填充（imputation）
  - 根据系谱检查基因型的正确性。
  - 去除有争议的基因型

目录

SNP-BLUP

$$\mathrm{y} = 1_{\mathrm{n}}\mu + \mathrm{Zg} + \mathrm{e}$$

- 其中

  y $n \times 1$ 观察值向量

  1_n 元素为 1 的向量

  μ 未知的一般平均数

  Z $n \times m$ 的标记基因型矩阵

  g $m \times 1$ （随机的）SNP 标记效应

  e 随机残差向量

- 假定：

  $\mathrm{e} \sim \mathrm{MVN}(0, \mathrm{R})$ 其中 $\mathrm{R} = \mathrm{R}_0\sigma_e^2$，且残差方差 σ_e^2 已知。

  $\mathrm{g} \sim \mathrm{MVN}(0, \mathrm{I}\sigma_g^2)$ 有已知的 σ_g^2。

SNP-BLUP 混合模型方程组

估计标记基因型效应的 MME

$$\begin{bmatrix} 1'R^{-1}1 & 1'R^{-1}Z \\ Z'R^{-1}1 & Z'R^{-1}Z + I^{-1}/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

当 $R = I\sigma_e^2$， $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$

$$\begin{bmatrix} 1'1 & 1'Z \\ Z'1 & Z'Z + I\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

SNP-BLUP 混合模型方程组

估计标记基因型效应的 MME

$$\begin{bmatrix} 1'R^{-1}1 & 1'R^{-1}Z \\ Z'R^{-1}1 & Z'R^{-1}Z + I^{-1}/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}$$

当 $R = I\sigma_e^2$，$\lambda = \frac{\sigma_e^2}{\sigma_g^2}$

$$\begin{bmatrix} 1'1 & 1'Z \\ Z'1 & Z'Z + I\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

- 基因组育种值
  - 训练组（training set）

$$\hat{u}_t = 1\hat{\mu} + Z_t\hat{g}$$

  - 候选组（candidate set）

$$\hat{u}_c = 1\hat{\mu} + Z_c\hat{g}$$

目录

G-BLUP

G-BLUP

- 直接估计育种值，而不估计标记效应
- 使用基因组（关系）矩阵
- 得到与 SNP-BLUP 相同的结果

G-BLUP

由 $\mathrm{a} = \mathrm{Zg}$

G-BLUP

- 直接估计育种值，而不估计标记效应
- 使用基因组（关系）矩阵
- 得到与 SNP-BLUP 相同的结果

$$\mathrm{y} = 1\mu + \mathrm{a} + \mathrm{e}$$

$$\Rightarrow \mathrm{Var(a)} = \mathrm{Var(Zg)} = \mathrm{ZVar(g)Z'}$$

$$\text{if } \mathrm{Var(g)} = \mathrm{I}\sigma_g^2$$

$$\Rightarrow \mathrm{a} \sim \mathrm{MVN}(0, \mathrm{G}\sigma_g^2)$$

$$\text{where } \mathrm{G} = \mathrm{ZZ'}$$

所有个体的 G-BLUP

与动物模型相似，候选个体有基因组信息，但是没有观测结果。
我们假定：

$$\mathrm{y} = 1\mu + \mathrm{Wa} + \mathrm{e}$$

W 是 $n \times q$的指示矩阵。
$q = n_t + n_c$。其中 n_t 是有观测值的训练组。n_c 是有基因型无表型观测值的候选群体。
$\mathrm{a} = \mathrm{Zg}$，$\mathrm{Z} = \begin{bmatrix} \mathrm{Z}_t \\ \mathrm{Z}_c \end{bmatrix}$，$\mathrm{a} = \begin{bmatrix} \mathrm{a}_t \\ \mathrm{a}_c \end{bmatrix}$。

所有个体的 G-BLUP

与动物模型相似，候选个体有基因组信息，但是没有观测结果。
我们假定：

$$\mathrm{y} = 1\mu + \mathrm{Wa} + \mathrm{e}$$

W 是 $n \times q$的指示矩阵。
$q = n_t + n_c$。其中 n_t 是有观测值的训练组。n_c 是有基因型无表型观测值的候选群体。
$\mathrm{a} = \mathrm{Zg}$，$\mathrm{Z} = \begin{bmatrix} \mathrm{Z}_t \\ \mathrm{Z}_c \end{bmatrix}$，$\mathrm{a} = \begin{bmatrix} \mathrm{a}_t \\ \mathrm{a}_c \end{bmatrix}$。 $\mathrm{a} \sim \mathrm{MVN}(0, \mathrm{G}\sigma_g^2)$，$\mathrm{G} = \mathrm{ZZ}'$

MME，仅训练组

MME

$$\begin{bmatrix} 1_n' R^{-1} 1_n & 1_n' R^{-1} \\ R^{-1} 1_n & R^{-1} + G_t^{-1}/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_t \end{bmatrix} = \begin{bmatrix} 1_n' R^{-1} y \\ R^{-1} y \end{bmatrix}$$

当 $R = I\sigma_e^2$

$$\begin{bmatrix} 1_n' 1_n & 1_n' \\ 1_n & I_n + \lambda G_t^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_t \end{bmatrix} = \begin{bmatrix} 1_n' y \\ y \end{bmatrix}$$

$\lambda = \frac{\sigma_e^2}{\sigma_g^2}$

MME，所有个体

MME

$$\begin{bmatrix} 1_n' R^{-1} 1_n & 1_n' R^{-1} W \\ W' R^{-1} 1_n & W' R^{-1} W + G_t^{-1}/\sigma_g^2 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_t \end{bmatrix} = \begin{bmatrix} 1_n' R^{-1} y \\ W' R^{-1} y \end{bmatrix}$$

当 $R = I\sigma_e^2$

$$\begin{bmatrix} 1_n' 1_n & 1_n' W \\ W' 1_n & W' W + \lambda G_t^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_t \end{bmatrix} = \begin{bmatrix} 1_n' y \\ W' y \end{bmatrix}$$

由于 $a' = [a_t' \; a_c'] \Rightarrow W = [I \; 0]$

MME，所有个体

$$\begin{bmatrix} 1'_n 1_n & 1'_n & 0 \\ 1_n & I_n + \lambda\{G^{-1}\}_{tt} & \lambda\{G^{-1}\}_{tc} \\ 0 & \lambda\{G^{-1}\}_{ct} & \lambda\{G^{-1}\}_{cc} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} 1'_n y \\ y \\ 0 \end{bmatrix}$$

其中

$$G^{-1} = \begin{bmatrix} \{G^{-1}\}_{tt} & \{G^{-1}\}_{tc} \\ \{G^{-1}\}_{ct} & \{G^{-1}\}_{cc} \end{bmatrix}$$