Selective Breeding

-A bird's eye view

吴晓平

2019-12-15

Contents

- **Basic concepts of Quantitative Genetics**

  Quantitative traits, selection, genotype value, breeding value, variance components, genetic progress

- **Traditional selective breeding**

  Selection index， BLUP

- **Genomic Selection**

  GBLUP， Bayesian model， ssGBLUP， feature model

- **Genomic Selection using DMU**

Basic concepts of Quantitative Genetics

Quantitative traits

What is the "best" animal?

- **Quantitative Trait**

  Continuous, affected by environment, numerous genes

  Milk yield, weight, egg weight…

- **Genotype provides genetic background for phenotype**

  P = G + E

Breeding by selection and mating

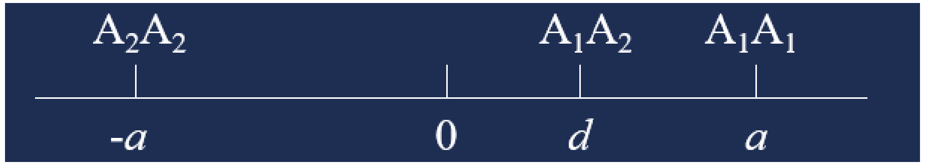
How do we breed animals so that there descendants will be , if not “best”, at least better than today’s animal?

= how can we genetically improve animal **population**?

- **Selection** (long term genetic change) -> change allele frequency -> improve population performance

- Mating (inbreeding)

Genotype value



Genotype	Frequency	Genotypic value
A_1A_1	p^2	a
A_1A_2	2pq	d
A_2A_2	q^2	-a

Genotype value: Deviation from the average of two homozygotes

$$G_{A_1A_1} = a - [a(p-q) + 2pqd] \\ = 2q(a - pd)$$

$$G_{A_1A_2} = d - [a(p-q) + 2pqd] \\ = a(q - p) + d(1 - 2pq)$$

$$G_{A_2A_2} = -a - [a(p-q) + 2pqd] \\ = -2p(a - qd)$$

Population mean

$$\mu_G = p^2a + 2pqd - q^2a = a(p-q) + 2pqd$$

Substitution effect

- The average effect of an allele, for A_1

M|(1 allele=A_1)

$$\alpha_1 = pa + qd - [a(p-q) + 2pqd]$$
$$= q[a + d(q-p)]$$

- The average effect of an allele , for A_2

$$\alpha_2 = pd - qa - [a(p-q) + 2pqd]$$
$$= -p[a + d(q-p)]$$

- Average effect of a gene substitution (α)

$$\alpha = a + d(q-p)$$

$$\alpha_1 = q\alpha \qquad \alpha_2 = -p\alpha$$

- When d=0, α=a
- We will mainly consider additive genetic effect and ignore non-additive genetic effect

Breeding value （BV）

- **Breeding value** (additive genetic value) of a animal = 2 ×(the expected phenotypic value of offspring of the animal when it mated randomly, expressed as deviation from the population mean)

$$A_{A_1A_1} = 2\alpha_1 = 2q\alpha$$

$$A_{A_1A_2} = \alpha_1 + \alpha_2 = (q-p)\alpha$$

$$A_{A_2A_2} = 2\alpha_2 = -2p\alpha$$

- **Mean breeding value of a population** MBV = $p^2 2q\alpha - q^2 2p\alpha + 2pq(q-p)\alpha = 0$

- **Variance of breeding values** V(BV) = $p^2\mathrm{Var}(2q\alpha) + 2pq\mathrm{Var}((q-p)\alpha) + q^2\mathrm{Var}(-2p\alpha)$
$= 2pq\alpha^2$

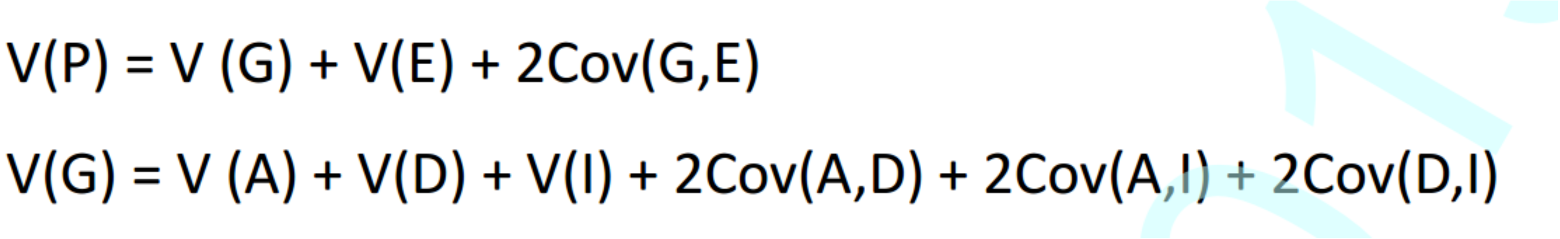
- **Considering a trait determined by a number of genes** $BV = \sum BV_i$ $V(BV) = \sum 2p_iq_i\alpha_i^2$

Breeding value （BV）

- Value of breeding
- Performance of the next generation, not animals itself

Variance components

- $P = G + E$
- E: fixed, random
- $G = A + D + I$



$$V(P) = V(G) + V(E) + 2\mathrm{Cov}(G,E)$$

$$V(G) = V(A) + V(D) + V(I) + 2\mathrm{Cov}(A,D) + 2\mathrm{Cov}(A,I) + 2\mathrm{Cov}(D,I)$$

- $h^2 = V_A/V_p$

Genetic progress

$$\Delta G = \frac{i r_{ia} \sigma_a}{L}$$

ΔG *:* genetic progress

i :selection intensity (留种率)

r_{ia}: accuracy (cor(EBV,TBV), h^2, data, model)

σ_a:variation（选择的前提，保持遗传变异)

L: generation interval

$$\text{Accuracy} = cor(\text{EBV,a}) = \frac{\text{cov(EBV,EBV} + \varepsilon)}{\sigma_{EBV}\sigma_a} = \frac{\sigma_{EBV}}{\sigma_a}$$

ε: Prediction error

For unbiased EBV

Cov(EBV, ε) =0

Reliability = Accuracy^2

Traditional selective Breeding

Phenotypic Selection

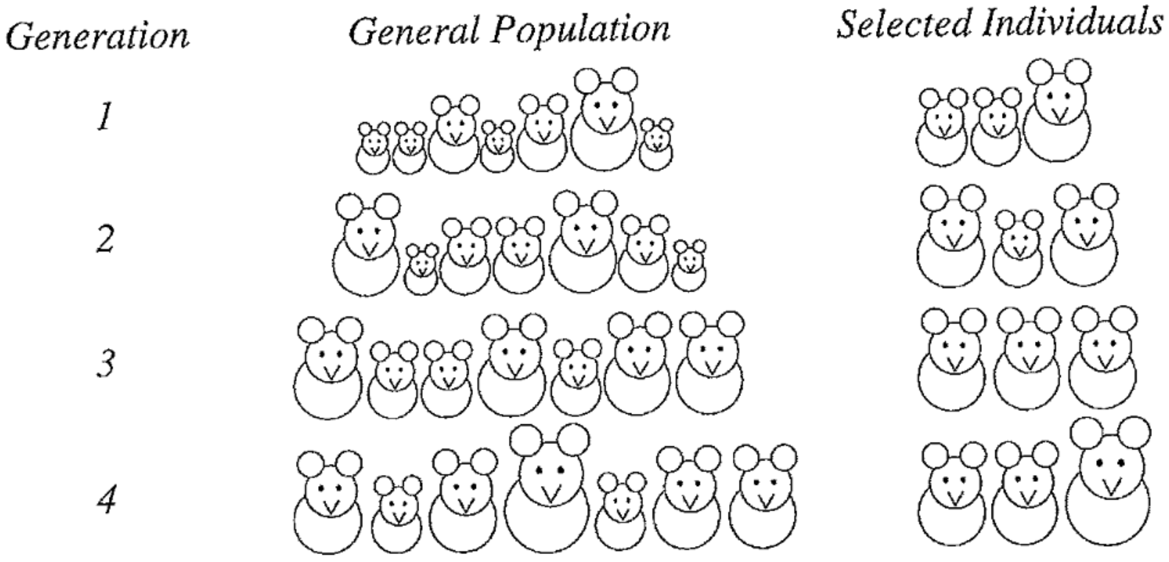


Figure 1.1. Illustration of phenotypic selection for increased body size in mice

High h^2

Selection by EBV

- Low/medium h^2
- Estimate EBV for animals without directive records
- Selection using difference resource of record
  - Animal’s Own Performance
  - Progeny Records
  - Pedigree
  - Another trait
  - Selection Index

BV prediction – Selection index

$$EBV = \hat{A} = b_{AP}(P^* - \overline{P})$$

Phenotypic information

$$b_{AP} = \frac{Cov(A, P^*)}{\sigma^2_{p^*}} = \frac{r_A n h^2}{1+(n-1)r_P}$$

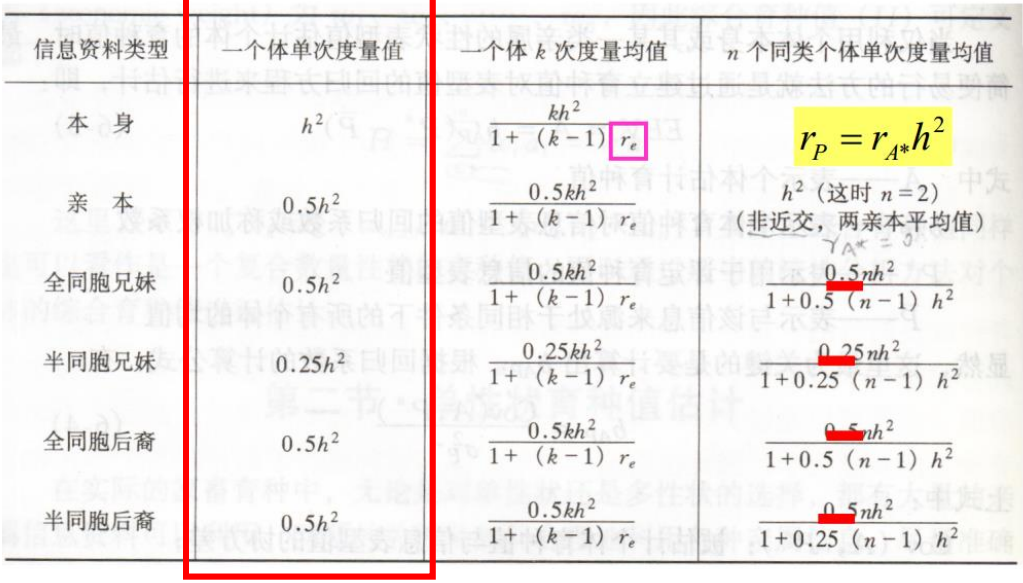
b_{AP} is the regression of true breeding value on phenotypic performance

r_A relationship coefficient

n repeat records number

r_p repeatability

h^2 heritability



信息资料类型	一个体单次度量值	一个体 k 次度量均值	n 个同类个体单次度量均值
本　身	h^2	$\frac{kh^2}{1+(k-1)\boxed{r_e}}$	$r_P = r_{A*}h^2$
亲　本	$0.5h^2$	$\frac{0.5kh^2}{1+(k-1)r_e}$	h^2（这时 $n=2$） （非近交，两亲本平均值） $r_{A*}=0$
全同胞兄妹	$0.5h^2$	$\frac{0.5kh^2}{1+(k-1)r_e}$	$\frac{0.5nh^2}{1+0.5(n-1)h^2}$
半同胞兄妹	$0.25h^2$	$\frac{0.25kh^2}{1+(k-1)r_e}$	$\frac{0.25nh^2}{1+0.25(n-1)h^2}$
全同胞后裔	$0.5h^2$	$\frac{0.5kh^2}{1+(k-1)r_e}$	$\frac{0.5nh^2}{1+0.5(n-1)h^2}$
半同胞后裔	$0.5h^2$	$\frac{0.5kh^2}{1+(k-1)r_e}$	$\frac{0.5nh^2}{1+0.25(n-1)h^2}$

$$r_{A\hat{A}} = r_{AP} = b_{AP}\frac{\sigma_{p^*}}{\sigma_A} = r_A\sqrt{\frac{nh^2}{1+(n-1)r_P}}$$

BV prediction – Selection index

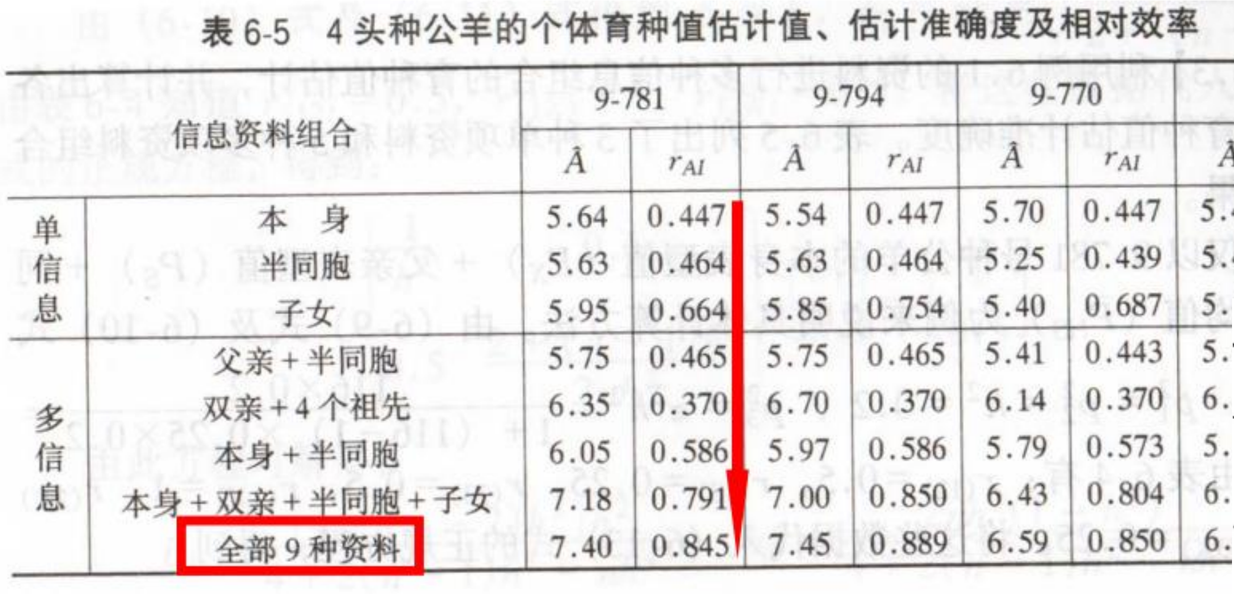


表 6-5　4 头种公羊的个体育种值估计值、估计准确度及相对效率

信息资料组合		9-781		9-794		9-770		9-752	
		$\hat{A}$	r_{AI}	$\hat{A}$	r_{AI}	$\hat{A}$	r_{AI}	$\hat{A}$	r_{AI}
单信息	本　身	5.64	0.447	5.54	0.447	5.70	0.447	5.48	0.447
	半同胞	5.63	0.464	5.63	0.464	5.25	0.439	5.49	0.447
	子女	5.95	0.664	5.85	0.754	5.40	0.687	5.	
多信息	父亲＋半同胞	5.75	0.465	5.75	0.465	5.41	0.443	5.	
	双亲＋4 个祖先	6.35	0.370	6.70	0.370	6.14	0.370	6.	
	本身＋半同胞	6.05	0.586	5.97	0.586	5.79	0.573	5.	
	本身＋双亲＋半同胞＋子女	7.18	0.791	7.00	0.850	6.43	0.804	6.	
	全部 9 种资料	7.40	0.845	7.45	0.889	6.59	0.850	6.	

$$r_{A\hat{A}} = r_{AP} = b_{AP}\frac{\sigma_{p*}}{\sigma_{A}} = r_{A}\sqrt{\frac{nh^{2}}{1+(n-1)r_{P}}}$$

Key factors affecting accuracy of EBV : h^2, data, model

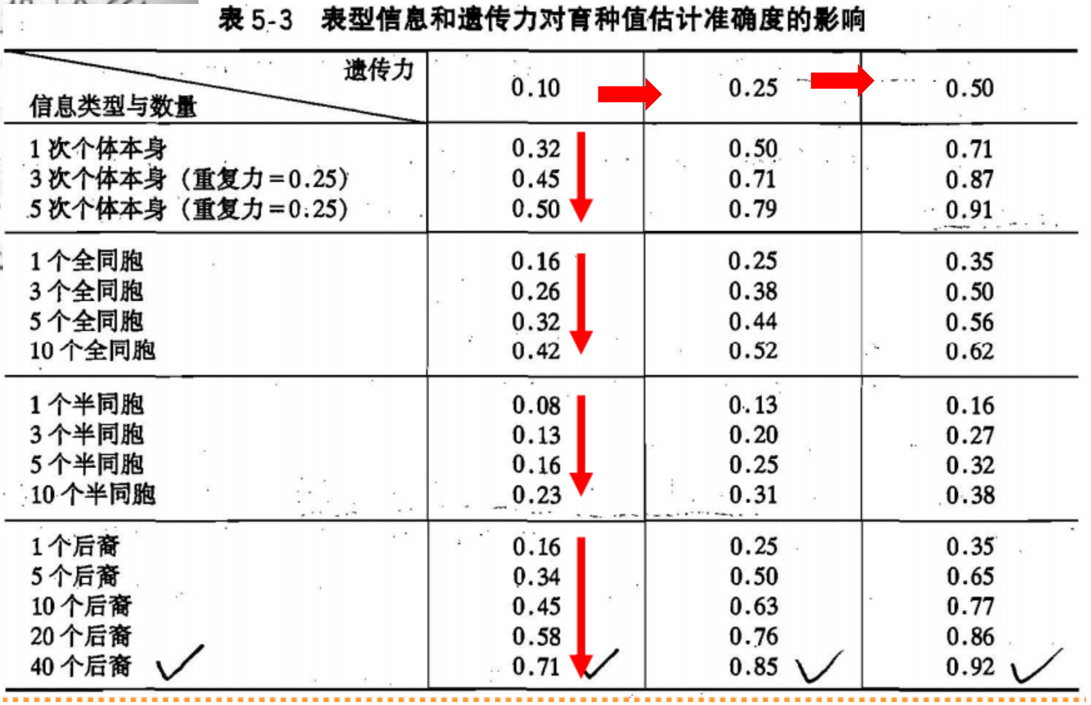


表 5-3　表型信息和遗传力对育种值估计准确度的影响

遗传力 信息类型与数量	0.10	0.25	0.50
1 次个体本身	0.32	0.50	0.71
3 次个体本身（重复力＝0.25）	0.45	0.71	0.87
5 次个体本身（重复力＝0.25）	0.50	0.79	0.91
1 个全同胞	0.16	0.25	0.35
3 个全同胞	0.26	0.38	0.50
5 个全同胞	0.32	0.44	0.56
10 个全同胞	0.42	0.52	0.62
1 个半同胞	0.08	0.13	0.16
3 个半同胞	0.13	0.20	0.27
5 个半同胞	0.16	0.25	0.32
10 个半同胞	0.23	0.31	0.38
1 个后裔	0.16	0.25	0.35
5 个后裔	0.34	0.50	0.65
10 个后裔	0.45	0.63	0.77
20 个后裔	0.58	0.76	0.86
40 个后裔 ✓	0.71 ✓	0.85 ✓	0.92 ✓

BV prediction - Selection index

Use of information from animal and all relatives

$$\hat{A} = b_1(P_1 - \overline{P}_1) + b_2(P_2 - \overline{P}_2) + \ldots\ldots + b_n(P_n - \overline{P}_n)$$

$$\mathbf{Vb} = \mathbf{c} \rightarrow \underline{\hat{\mathbf{b}}} = \mathbf{V}^{-1}\mathbf{c} \quad (\mathbf{c} = \mathbf{r}\sigma_A^2)$$

V: phenotype variance-covariance matrix, c: covariance vector between EBV and phenotype, r: relationship

Multiple trait

$$I_T = \hat{A}_T = \sum b_i(p_i - \overline{p}_i) = b'(p - \overline{p})$$

$$Vb = DAw \rightarrow \hat{b} = V^{-1}DAw$$

V: phenotype variance-covariance matrix, A: covariance vector between object trait and information trait, D: relationship, w: weight

BV prediction –**Selection index**

- Using of information from animal and all relatives
- Records may have to be pre-adjusted for fixed or environmental factors (non-genetic factors)
- Assume known genetic parameter
- Estimated individual and information animals come from same population
- Solutions to the index equations require the inverse of the covariance matrix for observations and this may not be computationally feasible for large data sets

BV prediction - BLUP

- Henderson (1949) developed, best linear unbiased prediction (BLUP)
- Fixed effects and breeding values can be simultaneously estimated
- Best – means it maximizes the correlation between true (a) and predicted breeding value or minimizes prediction error variance (PEV).
- Linear – predictors are linear functions of observations.
- Unbiased – estimation of realized values for a random variable, such as animal breeding values, and of estimable functions of fixed effects are unbiased.
- Prediction – involves prediction of true breeding value.

BV prediction - BLUP

Mixed linear model **y = Xb + Za + e**

Genetic relationship matrix

$$a \;\square\; \mathrm{N}(0,\mathbf{A}\sigma_a^2) \qquad \mathbf{e} \;\square\; \mathrm{N}(0,\mathbf{I}\sigma_e^2)$$

Mixed model equation $\mathbf{G} = \mathbf{A}\sigma_a^2$ $\quad \mathbf{R} = \mathbf{I}\sigma_e^2$ $\quad \alpha = \sigma_e^2/\sigma_a^2$ or $(1-h^2)/h^2$

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

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

- Using information of all relatives through **A**
- All animals in the pedigree get EBV

Henderson (1950)

BLUP – BLUP

$$a = EBV + \varepsilon$$

$$\operatorname{cov}(EBV, \varepsilon) = 0$$

$$\sigma_a^2 = \sigma_{EBV}^2 + \sigma_\varepsilon^2$$

$$r_{EBV}^2 = cor^2(EBV, a)$$

$$= \frac{\sigma_a^2 - \sigma_\varepsilon^2}{\sigma_a^2}$$

$$= 1 - \frac{\sigma_\varepsilon^2}{\sigma_a^2}$$

$$\sigma_\varepsilon^2 = PEV$$

Inverse of coefficient of MME

$$\begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$$

Henderson (1975) showed that the PEV is:

$$\text{PEV} = \operatorname{var}(a - \hat{a}) = C^{22}\sigma_e^2$$

BV prediction - BLUP

- Single-trait model
- Multiple-trait model (correlation, h^2, records number). Large benefit for the traits with low h^2 (e.g., fertility) and small number or records. Fx: FCR
- Model including direct & maternal additive genetic effects (calving ease, birth weight, early growth)
- Random regression model (G by E, longitudinal data)
- Bayesian inference with Gibbs sampling
- Generalized linear mixed model (Logistic model, Probit model (binary trait))

BV prediction - BLUP

Compare with index selection, BLUP

- Adjusted environment effect
- Using all relative information
- EBV for different population (genetic correlation exists among population)
- High accuray

But

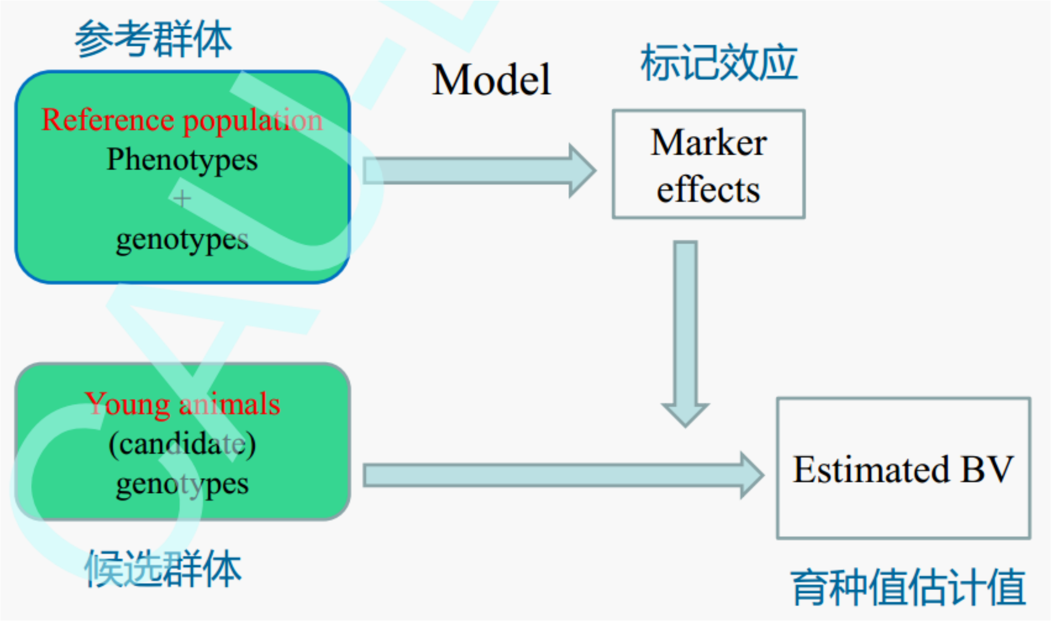
- High accuracy of selection often companies with long generation interval (e.g., progeny test in cattle)
- Low accuracy for the individual without own or offspring phenotype

Genomic Selection

Genomic Selection

Proposed by Meuwissen et.al 2001

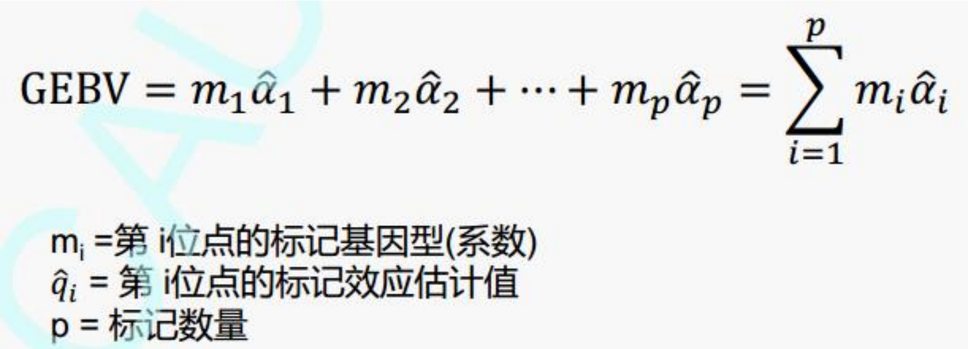
- Selection early
- Shorten generation interval (cow)
- Accuracy (pig)



Idea:

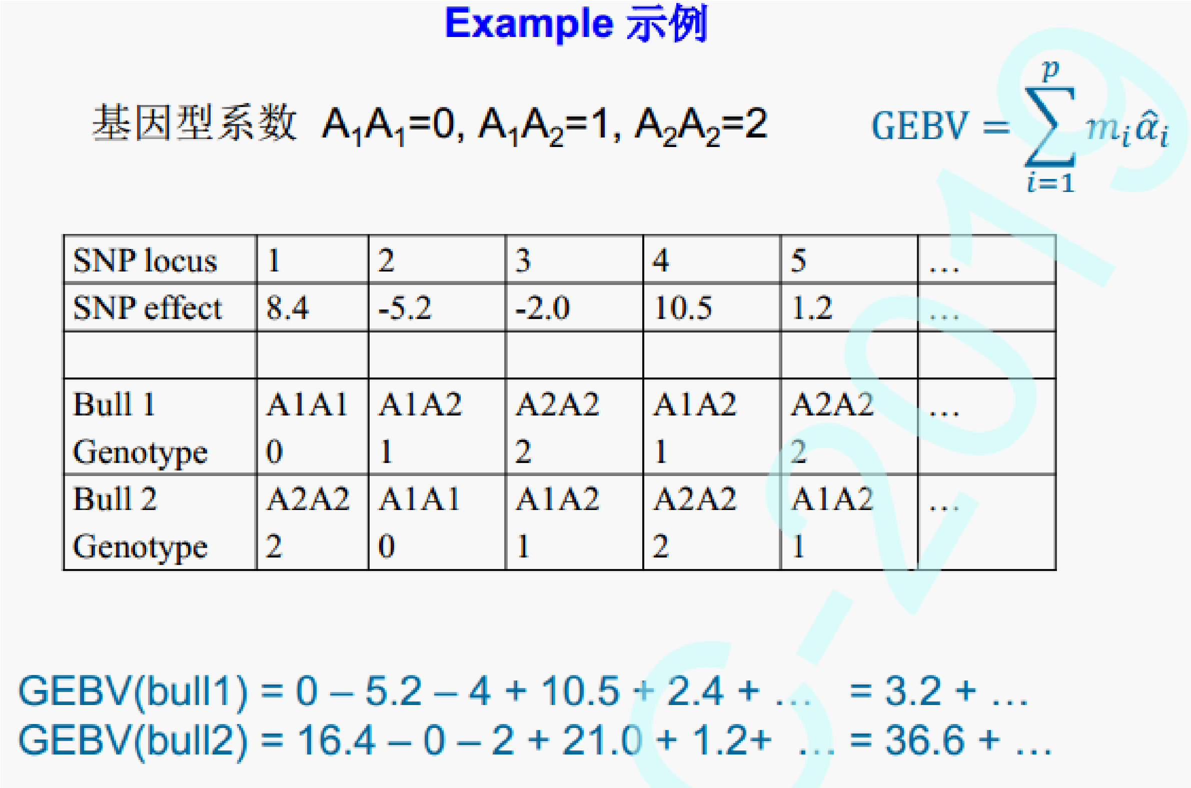
- Marker cover whole genome
- LD between marker and gene
- Use population information
- Account for Mendelian sampling term even without own or offspring’s phenotypes

GEBV



$$\mathrm{GEBV} = m_1\hat{\alpha}_1 + m_2\hat{\alpha}_2 + \cdots + m_p\hat{\alpha}_p = \sum_{i=1}^{p} m_i\hat{\alpha}_i$$

m_i =第 i位点的标记基因型(系数)
$\hat{q}_i$ = 第 i位点的标记效应估计值
p = 标记数量



Example 示例

基因型系数 A_1A_1=0, A_1A_2=1, A_2A_2=2

$$\mathrm{GEBV} = \sum_{i=1}^{p} m_i\hat{\alpha}_i$$

SNP locus	1	2	3	4	5	…
SNP effect	8.4	-5.2	-2.0	10.5	1.2	…
Bull 1 Genotype	A1A1 0	A1A2 1	A2A2 2	A1A2 1	A2A2 2	…
Bull 2 Genotype	A2A2 2	A1A1 0	A1A2 1	A2A2 2	A1A2 1	…

GEBV(bull1) = 0 – 5.2 – 4 + 10.5 + 2.4 + … = 3.2 + …
GEBV(bull2) = 16.4 – 0 – 2 + 21.0 + 1.2+ … = 36.6 + …

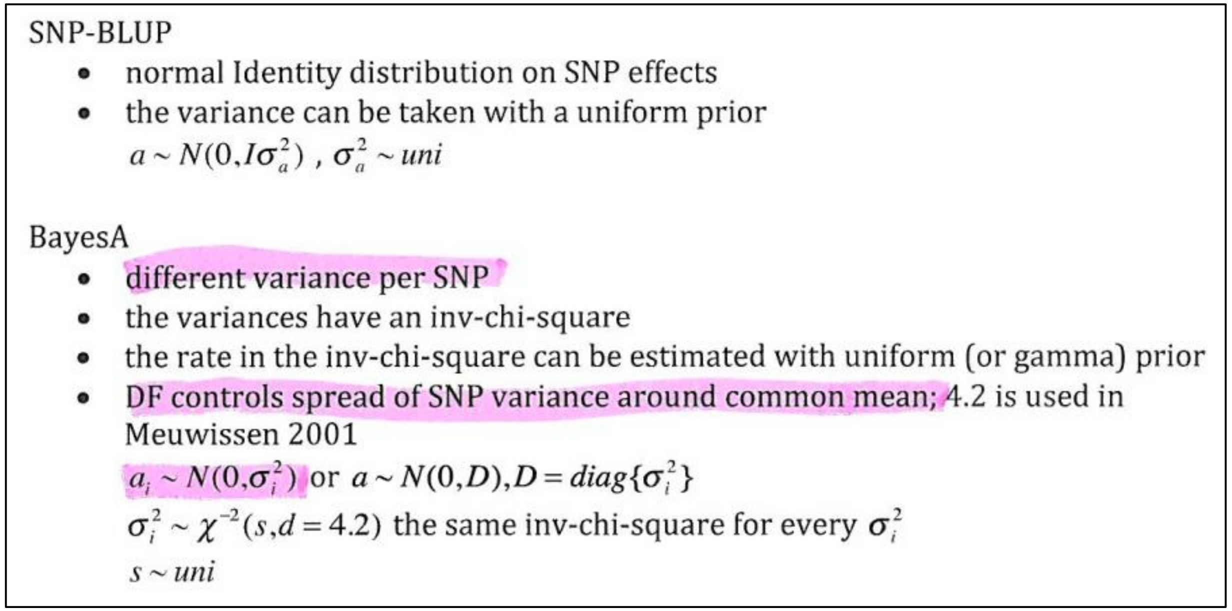
Estimate SNP effect - Models

- SNP-BLUP, GBLUP
- Bayes A,B,C,R, mixture, Lasso
- Genomic feature model, weighted GBLUP

These models all take SNP effects as random effects

These models differ in the assumption on distribution of SNP effects

Estimate SNP effect - Models



SNP-BLUP

- normal Identity distribution on SNP effects
- the variance can be taken with a uniform prior

  $a \sim N(0, I\sigma_a^2)$, $\sigma_a^2 \sim uni$

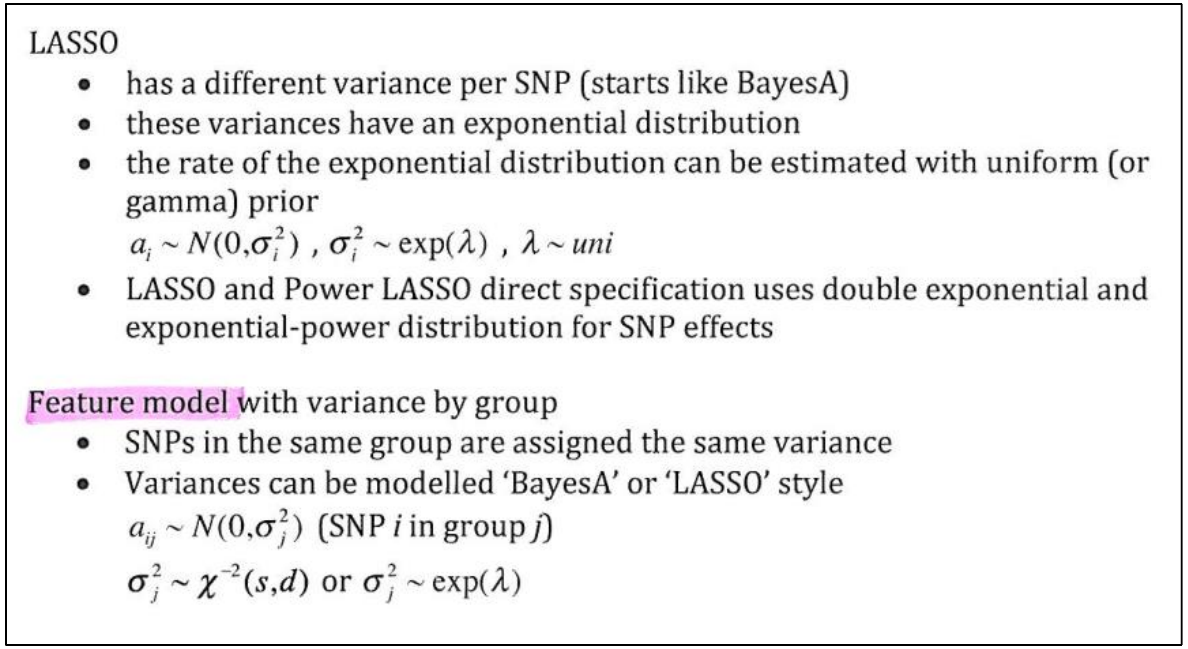
BayesA

- different variance per SNP
- the variances have an inv-chi-square
- the rate in the inv-chi-square can be estimated with uniform (or gamma) prior
- DF controls spread of SNP variance around common mean; 4.2 is used in Meuwissen 2001

  $a_i \sim N(0, \sigma_i^2)$ or $a \sim N(0, D), D = diag\{\sigma_i^2\}$

  $\sigma_i^2 \sim \chi^{-2}(s, d = 4.2)$ the same inv-chi-square for every σ_i^2

  $s \sim uni$



LASSO

- has a different variance per SNP (starts like BayesA)
- these variances have an exponential distribution
- the rate of the exponential distribution can be estimated with uniform (or gamma) prior

  $a_i \sim N(0, \sigma_i^2)$, $\sigma_i^2 \sim \exp(\lambda)$, $\lambda \sim uni$

- LASSO and Power LASSO direct specification uses double exponential and exponential-power distribution for SNP effects

Feature model with variance by group

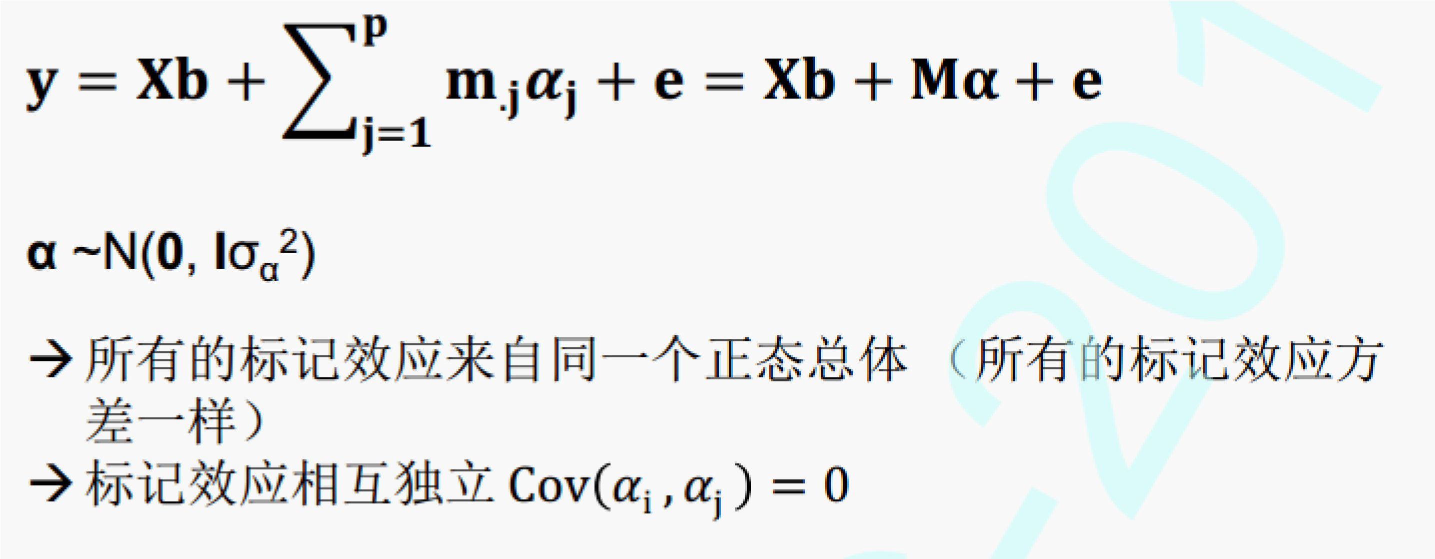
- SNPs in the same group are assigned the same variance
- Variances can be modelled 'BayesA' or 'LASSO' style

  $a_{ij} \sim N(0, \sigma_j^2)$ (SNP *i* in group *j*)

  $\sigma_j^2 \sim \chi^{-2}(s, d)$ or $\sigma_j^2 \sim \exp(\lambda)$

From Luc et al. 2014

SNP-BLUP



$$\mathbf{y} = \mathbf{Xb} + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{m}_{.\mathbf{j}}\boldsymbol{\alpha}_{\mathbf{j}} + \mathbf{e} = \mathbf{Xb} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{e}$$

$\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}\sigma_{\alpha}^{2})$

→ 所有的标记效应来自同一个正态总体（所有的标记效应方差一样）

→ 标记效应相互独立 $\mathrm{Cov}(\alpha_i, \alpha_j) = 0$

Good: Simple and Fast

Bad: assumption of a common normal distribution might not be appropriate

GBLUP

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

$$\mathbf{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$$

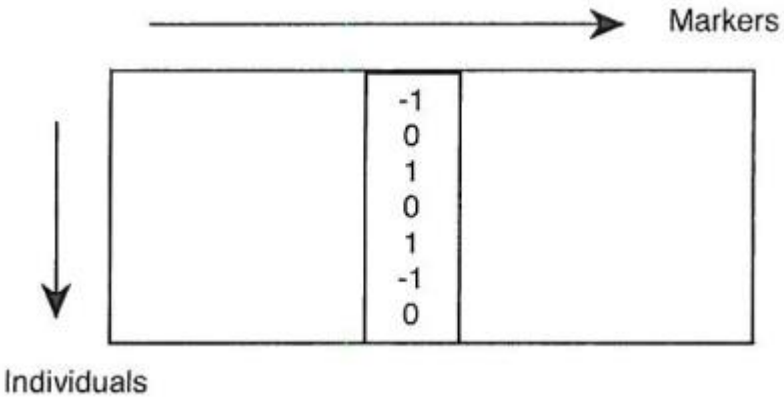
Assume variance for all SNP are the same ($\sigma_{\alpha j}^2$= σ_α^2)
total genomic variance σ_g^2 is

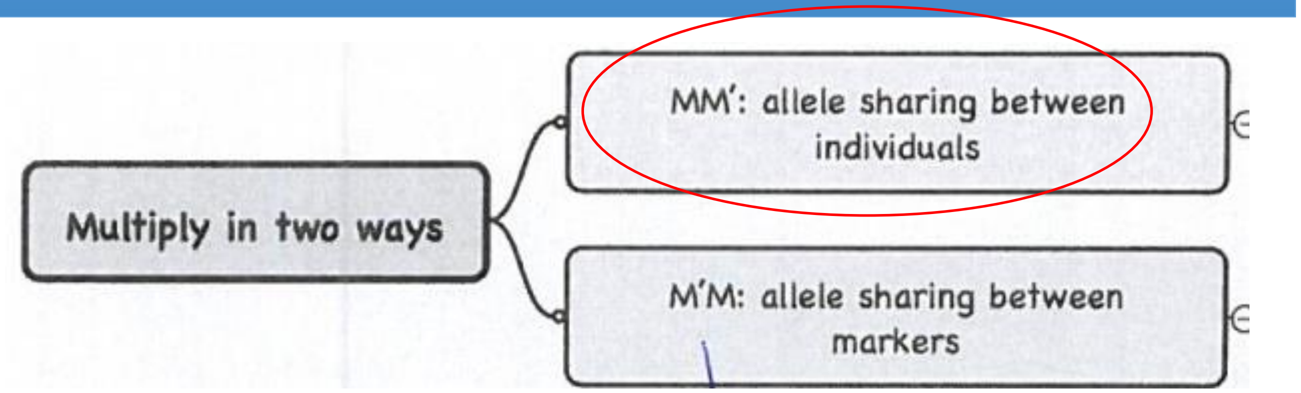
$$\sigma_g^2 = \sum_{j=1}^{p} 2p_j q_j \sigma_\alpha^2$$

- SNP-BLUP： variance contribution per SNP
- GBLUP： total variance of all SNPs

SNP-BLUP = GBLUP, because every SNP counts equal in G matrix

G matrix




Multiply in two ways

MM': allele sharing between individuals

M'M: allele sharing between markers

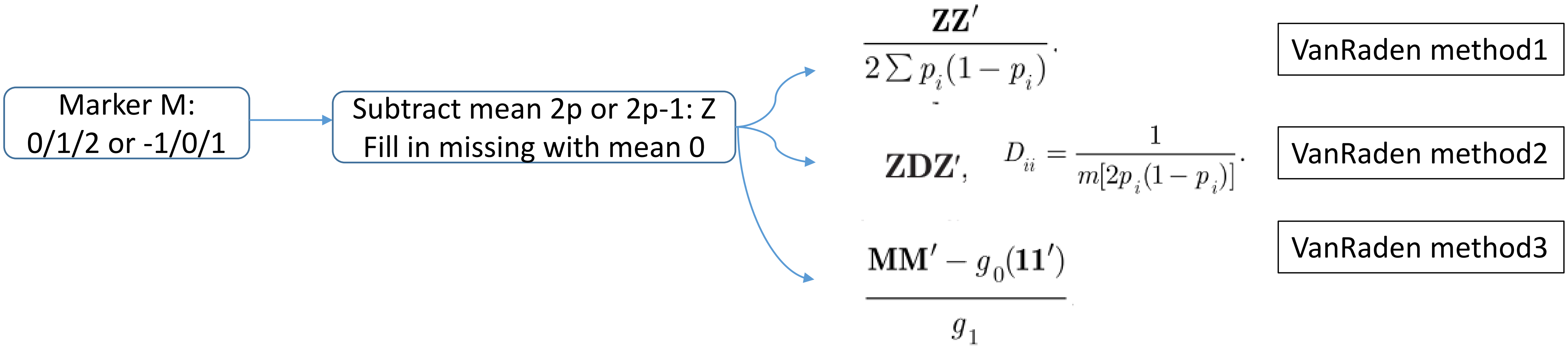
	11	12	22
code	-1	0	1
count	n1	n2	n3
Centered value	-1-(2p-1) = -2p	0-(2p-1) = 1-2p	1-(2p-1) = 2-2p
Frequency	$(1-p)^2$	2p(1-p)	p^2

Mean=$-1*(1-p)^2+0*2p(1-p)+1*p^2 = 2p-1$

	11	12	22
code	0	1	2
count	n1	n2	N3
Centered value	-2p	1-2p	2-2p
Frequency	$(1-p)^2$	2p(1-p)	p^2

Mean=2p

G matrix



Marker M: 0/1/2 or -1/0/1

Subtract mean 2p or 2p-1: Z
Fill in missing with mean 0

$$\frac{\mathbf{ZZ}'}{2\sum p_i(1-p_i)}.$$

VanRaden method1

$$\mathbf{ZDZ}', \quad D_{ii} = \frac{1}{m[2p_i(1-p_i)]}.$$

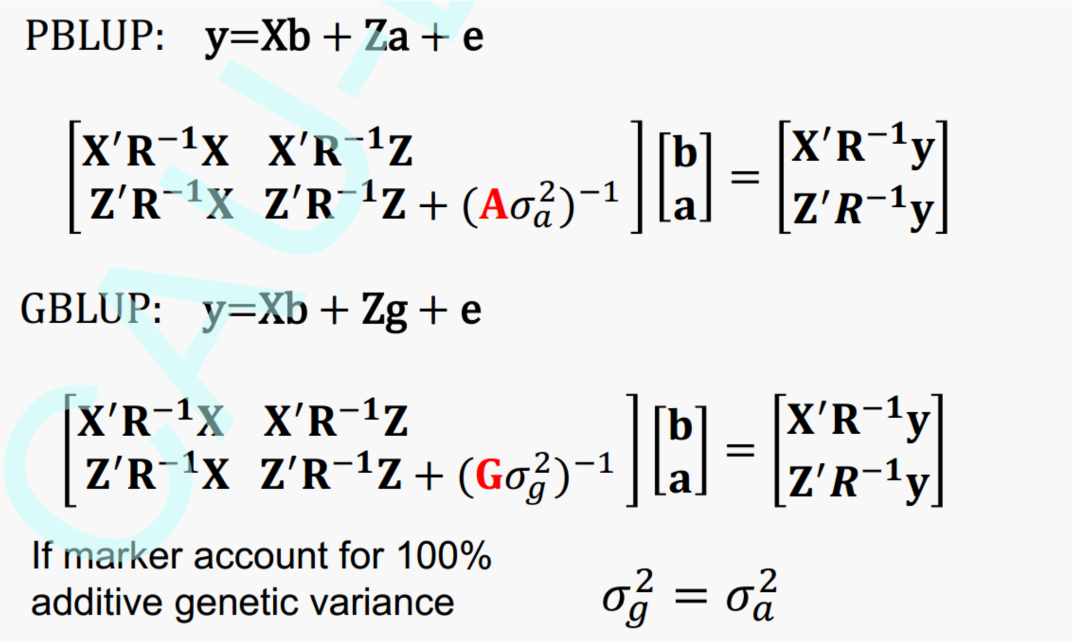
VanRaden method2

$$\frac{\mathbf{MM}' - g_0(\mathbf{11}')}{g_1}$$

VanRaden method3

- Add a small value to diagonal of G matrix in order to make G-matrix being positive definitive. Usually 0 - 0.02.
- Add a value to all elements of G matrix. It is said that adding a very small values can improve the relationship matrix but it has not been confirmed.
- $\mathbf{G}_\omega = \omega\mathbf{A} + (1-\omega)\mathbf{G}$

PBLUP vs GBLUP



PBLUP: $\mathbf{y}=\mathbf{Xb}+\mathbf{Za}+\mathbf{e}$

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+(\mathbf{A}\sigma_a^2)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'R^{-1}\mathbf{y} \end{bmatrix}$$

GBLUP: $\mathbf{y}=\mathbf{Xb}+\mathbf{Zg}+\mathbf{e}$

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+(\mathbf{G}\sigma_g^2)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'R^{-1}\mathbf{y} \end{bmatrix}$$

If marker account for 100% additive genetic variance $\sigma_g^2 = \sigma_a^2$

- A matrix is an expected relationship matrix
- G matrix is a realized relationship matrix
- G matrix capture the relationship due to Mendelian sampling error

GBLUP with polygenic effect

- Marker may not account all genetic variation
- Model including polygenic effect to account the remaining genetic effect
- GBLUP$_{AG}$ $\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \mathbf{Z}\mathbf{g} + \mathbf{e}$
- *GBLUP$_{AG}$ $\mathbf{y}=\mathbf{1}\mu+\mathbf{Z}\mathbf{g}_\omega+\mathbf{e}.$

Residual Genetic variance $\omega\,\sigma^2_{g_\omega}$

$$\mathbf{g}_\omega = \mathbf{u} + \mathbf{g}, \ \mathrm{Var}(\mathbf{g}_\omega) = \mathbf{A}\sigma_u^2 + \mathbf{G}\sigma_g^2$$

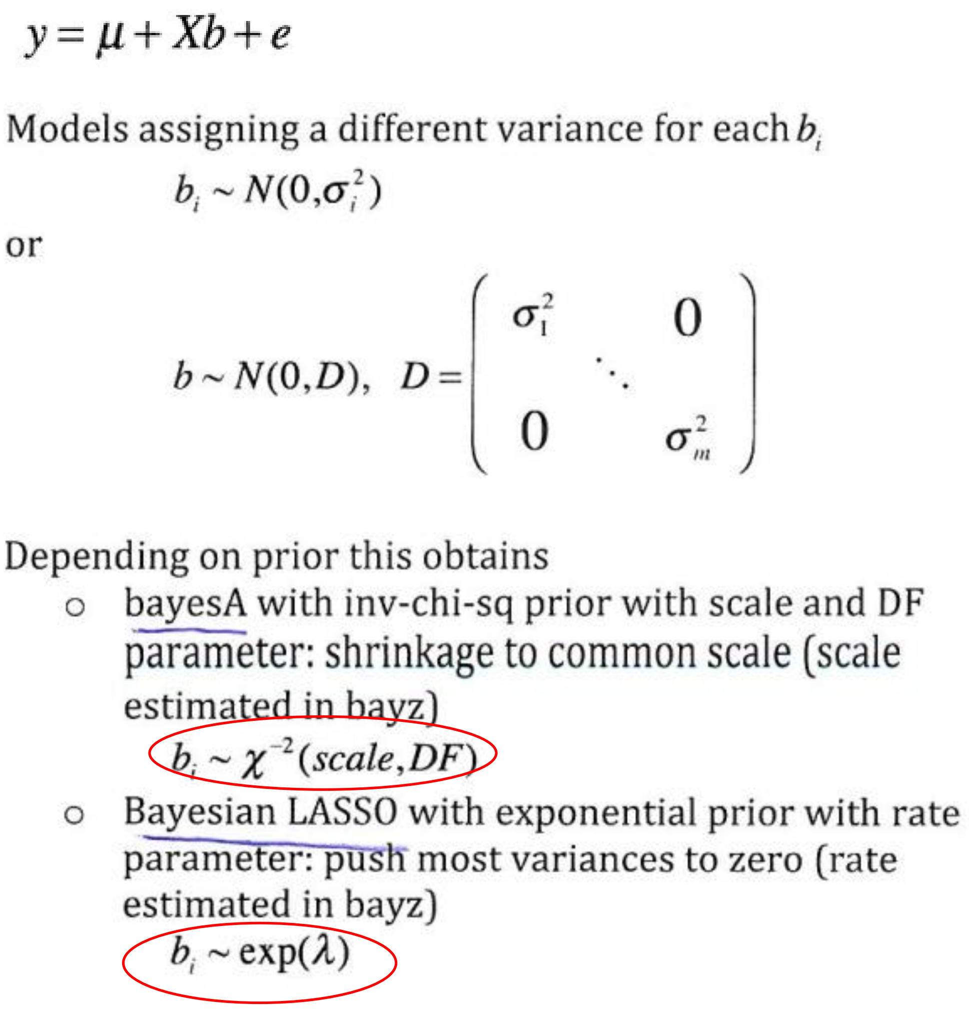
Genetic variance by marker $(1-\omega)\sigma^2_{g_\omega}$

ω is the ratio of residual polygenic to total additive genetic variance

GBLUP

- Algorithm same as pedigree based model
- G inverse was not assured
- When n > number of marker, G is not positive

Bayesian



$$y = \mu + Xb + e$$

Models assigning a different variance for each b_i

$$b_i \sim N(0, \sigma_i^2)$$

or

$$b \sim N(0, D), \quad D = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_m^2 \end{pmatrix}$$

Depending on prior this obtains

- bayesA with inv-chi-sq prior with scale and DF parameter: shrinkage to common scale (scale estimated in bayz)
  $$b_i \sim \chi^{-2}(scale, DF)$$
- Bayesian LASSO with exponential prior with rate parameter: push most variances to zero (rate estimated in bayz)
  $$b_i \sim \exp(\lambda)$$

- Assume effects of different SNP having different variances
- Machine learning to determine if a SNP should be included in model during the procedure of analysis
- Efficiently differentiate SNP with large effect or null effect/small effect

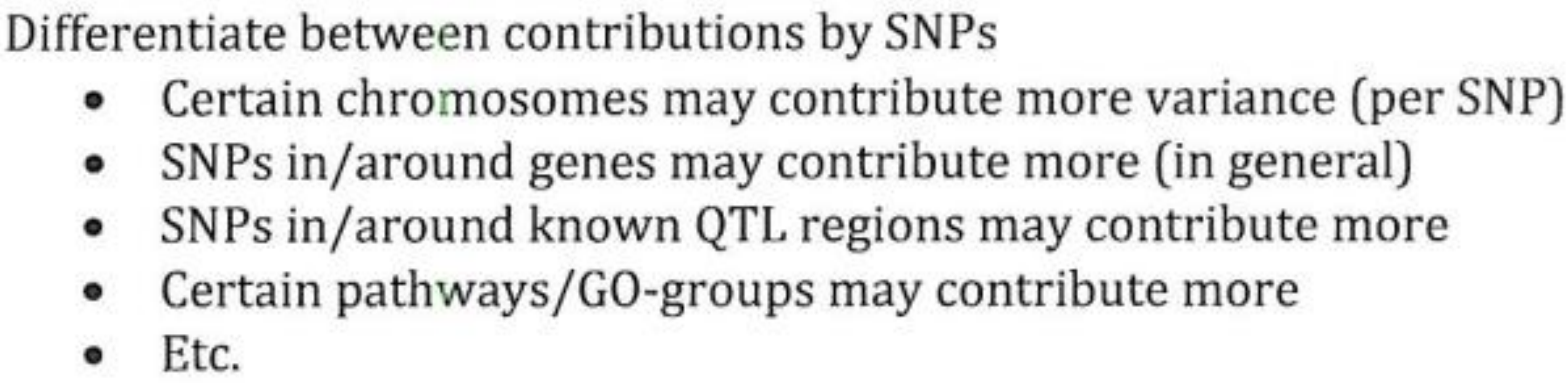
Good: The assumption of SNP effect is more consistent with distribution of QTL effect

Bad: Heavy computing time

Compare GBLUP with Bayesian model

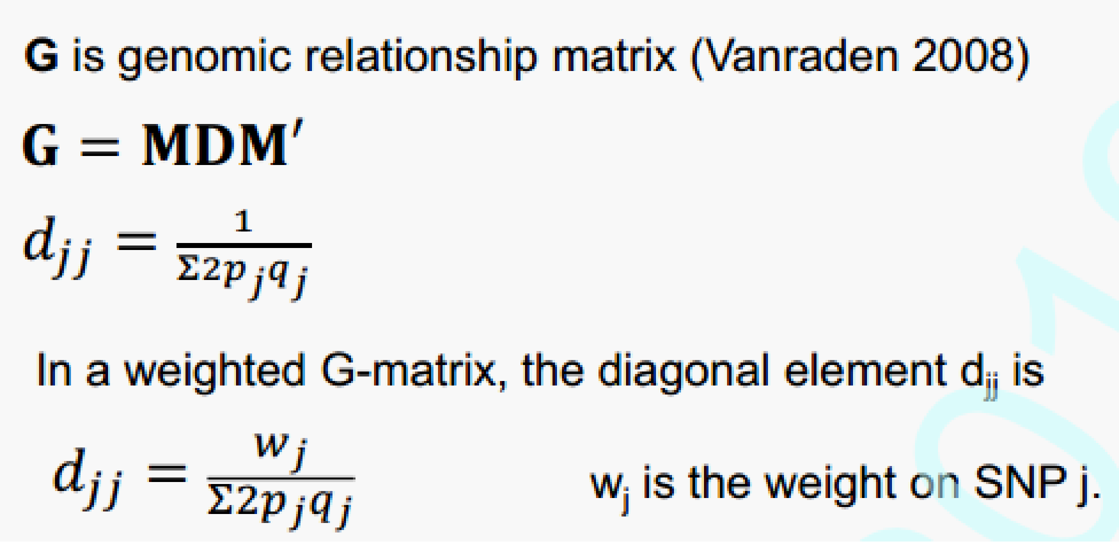
- Based on simulation data, Bayesian variable selection models are much better than BLUP model
- Based on real data, Bayesian variable selection models slightly better than or similar to linear mixed models
- The advantage of Bayesian models over BLUP model depend on number of QTL with large effect on a trait.

Genomic feature model



Differentiate between contributions by SNPs

- Certain chromosomes may contribute more variance (per SNP)
- SNPs in/around genes may contribute more (in general)
- SNPs in/around known QTL regions may contribute more
- Certain pathways/GO-groups may contribute more
- Etc.



G is genomic relationship matrix (Vanraden 2008)

$$\mathbf{G} = \mathbf{MDM'}$$

$$d_{jj} = \frac{1}{\Sigma 2p_j q_j}$$

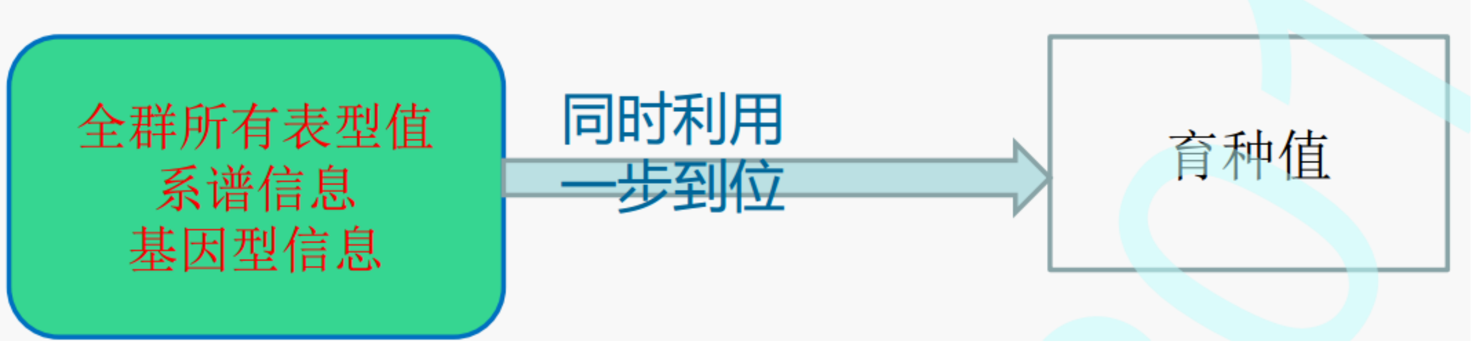
In a weighted G-matrix, the diagonal element d_{jj} is

$$d_{jj} = \frac{w_j}{\Sigma 2p_j q_j}$$

w_j is the weight on SNP j.

ssGBLUP

- Legarra et al., 2009
- Limited number of animal genotyped
- predition model: use information both from genotyped and ungenotyped animal
- Combined genotype-pedigree relationship matrix (H matrix)



全群所有表型值
系谱信息
基因型信息

同时利用
一步到位

育种值

ssGBLUP

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + (\mathbf{H}\sigma_a^2)^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'}R^{-1}\mathbf{y} \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{G}_\omega & \mathbf{G}_\omega \mathbf{A}_{11}^{-1} \mathbf{A}_{12} \\ \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{G}_\omega & \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{G}_\omega \mathbf{A}_{11}^{-1} \mathbf{A}_{12} + \mathbf{A}_{22} - \mathbf{A}_{21} \mathbf{A}_{11}^{-1} \mathbf{A}_{12} \end{bmatrix}$$

Proposed by Legarra 2009

A11: submatrix of A for **genotyped animals**

A22: submatrix of A for **non-genotyped animals**

A12 or A21: sub-matrix of A for relationships between genotyped and non-genotyped animals

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{G}_\omega^{-1} - \mathbf{A}_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \mathbf{A}^{-1}$$

ssGBLUP

$$\mathbf{G}_{\omega} = (1-\omega)\mathbf{G} + \omega\mathbf{A}_{11}$$

- Marker may not explain all genetic variance

ω=0.1~0.3 for most of trait

Vitezica ZG et al. 2011

- To ensure **G** positive definite

ω=0.05

Tsuruta et al. 2011

- Scales of **A** and **G** matrix may differ

$$G_a = \beta G + \alpha,$$

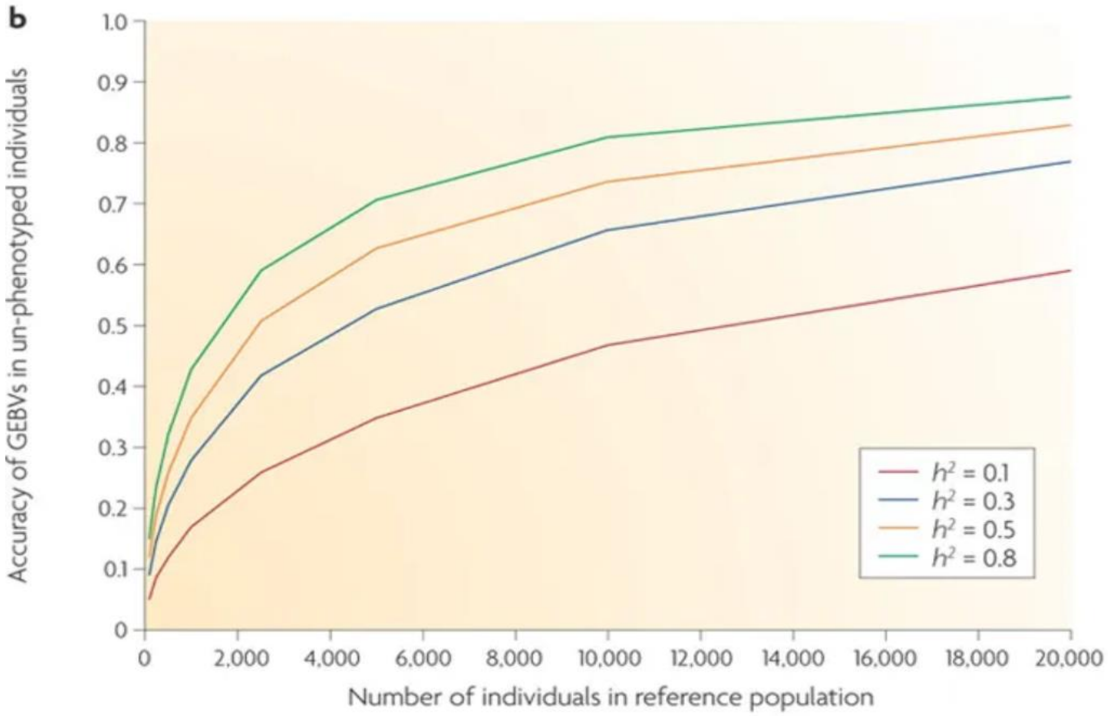
where β and α solved the system of equations

$$\text{Avg}(\text{diag}(G))\beta + \alpha = \text{Avg}(\text{diag}(A_{11})),$$
$$\text{Avg}(G)\beta + \alpha = \text{Avg}(A_{11}).$$

Christensen, GSE 2012

Factors affect GP accuracy

1. **Model**
2. **Number of individual in reference population**
3. **Accurate phenotype records**



Goddard and Hayes, 2009

Factors affect GP accuracy

4. Genetic relationship within reference population

Cover genetic variation of the whole population

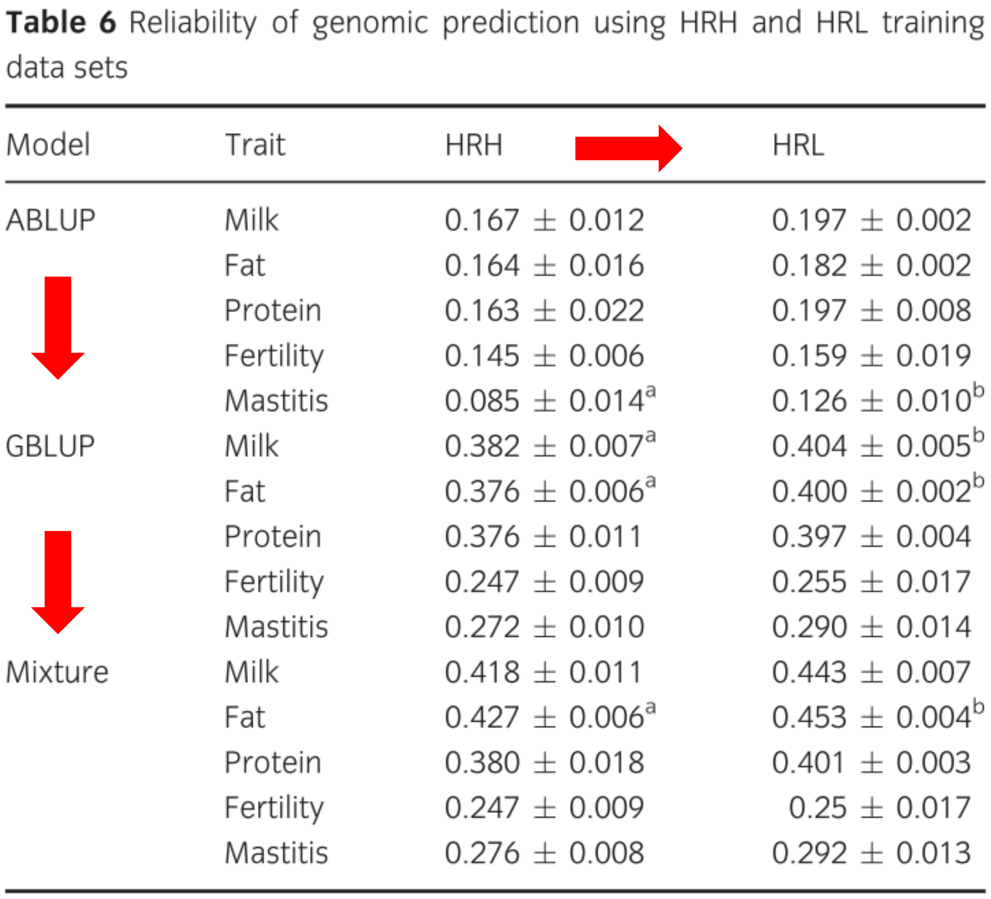


Table 6 Reliability of genomic prediction using HRH and HRL training data sets

Model	Trait	HRH	HRL
ABLUP	Milk	0.167 ± 0.012	0.197 ± 0.002
	Fat	0.164 ± 0.016	0.182 ± 0.002
	Protein	0.163 ± 0.022	0.197 ± 0.008
	Fertility	0.145 ± 0.006	0.159 ± 0.019
	Mastitis	0.085 ± 0.014[a]	0.126 ± 0.010[b]
GBLUP	Milk	0.382 ± 0.007[a]	0.404 ± 0.005[b]
	Fat	0.376 ± 0.006[a]	0.400 ± 0.002[b]
	Protein	0.376 ± 0.011	0.397 ± 0.004
	Fertility	0.247 ± 0.009	0.255 ± 0.017
	Mastitis	0.272 ± 0.010	0.290 ± 0.014
Mixture	Milk	0.418 ± 0.011	0.443 ± 0.007
	Fat	0.427 ± 0.006[a]	0.453 ± 0.004[b]
	Protein	0.380 ± 0.018	0.401 ± 0.003
	Fertility	0.247 ± 0.009	0.25 ± 0.017
	Mastitis	0.276 ± 0.008	0.292 ± 0.013

Wu et al. 2015

Factors affect GP accuracy

5. Genetic relationship between reference and validate population : Closer relationship, more consistent LD phase

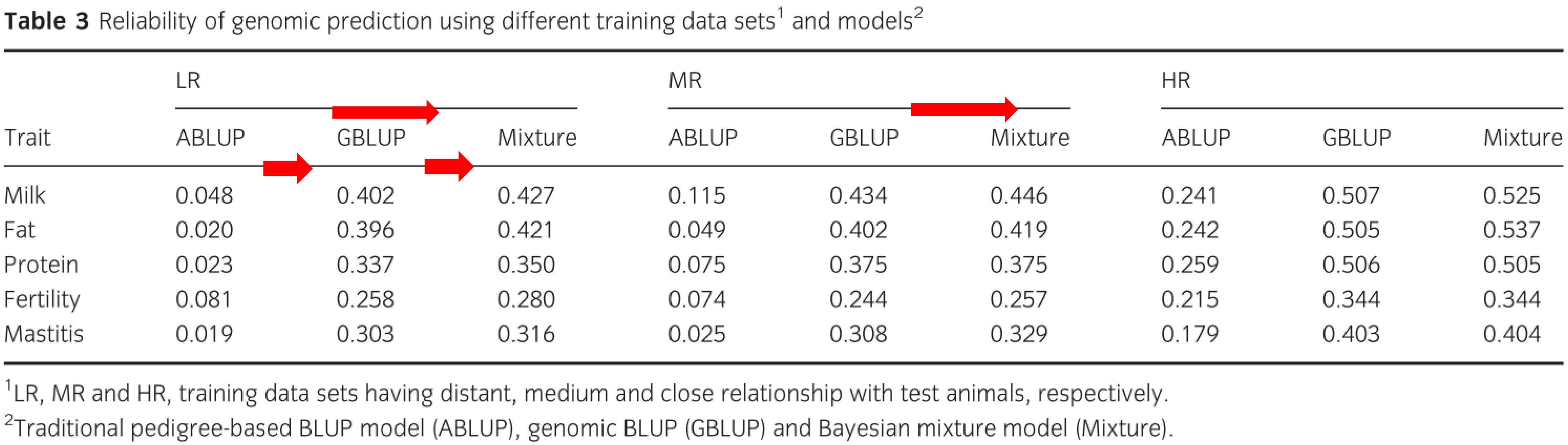


Table 3 Reliability of genomic prediction using different training data sets[1] and models[2]

	LR			MR			HR		
Trait	ABLUP	GBLUP	Mixture	ABLUP	GBLUP	Mixture	ABLUP	GBLUP	Mixture
Milk	0.048	0.402	0.427	0.115	0.434	0.446	0.241	0.507	0.525
Fat	0.020	0.396	0.421	0.049	0.402	0.419	0.242	0.505	0.537
Protein	0.023	0.337	0.350	0.075	0.375	0.375	0.259	0.506	0.505
Fertility	0.081	0.258	0.280	0.074	0.244	0.257	0.215	0.344	0.344
Mastitis	0.019	0.303	0.316	0.025	0.308	0.329	0.179	0.403	0.404

[1] LR, MR and HR, training data sets having distant, medium and close relationship with test animals, respectively.

[2] Traditional pedigree-based BLUP model (ABLUP), genomic BLUP (GBLUP) and Bayesian mixture model (Mixture).

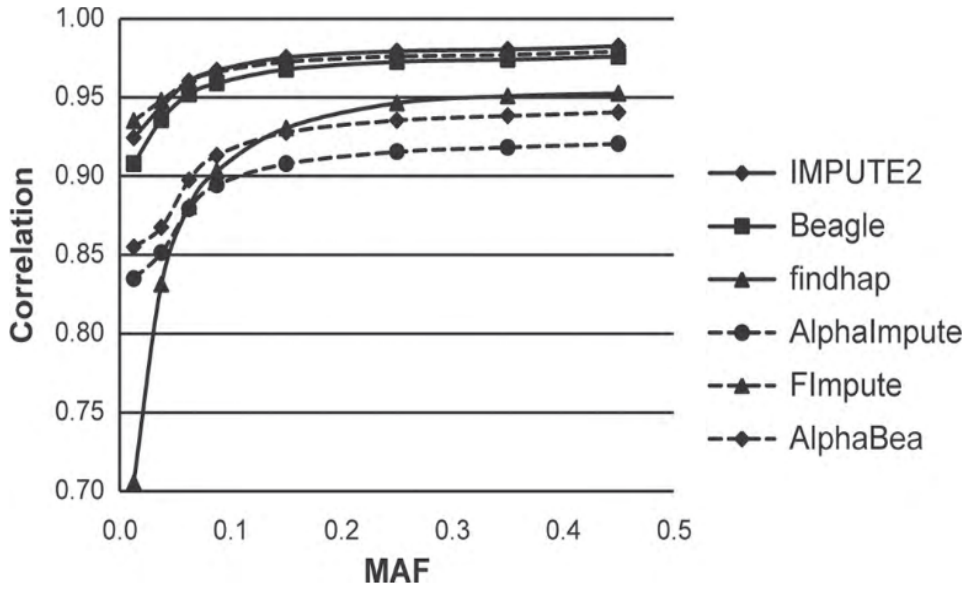
Reference population should be updated gradually

Wu et al. 2015

Factors affect GP accuracy

6. Genotype density, imputation accuracy

- **SNP chip**
  - Low-density, Moderate-density , high-density, whole genome sequence
  - Call rate > 80%, maf > 0.01, gencall score > 60%
- **Imputation**
  - using population LD information
  - Accuracy: Genotype/allele corrected rate , correlation, >80%
  - Minimac, Eagle: Good for imputation od sequencing data



(Ma et al., J. Dairy Sci. 96)

Factors affect GP accuracy

- Imputation from LD to MD, then to whole genome sequence level

- Factors affect imputation accuracy：
  - Marker density of the chips
  - Number of common SNP in different chips.
  - Number of animals genotyped with target density (higher density)
  - The relationship between the animals with lower density and the animals with high density
  - Effective population size
  - Minor allele frequency (MAF)

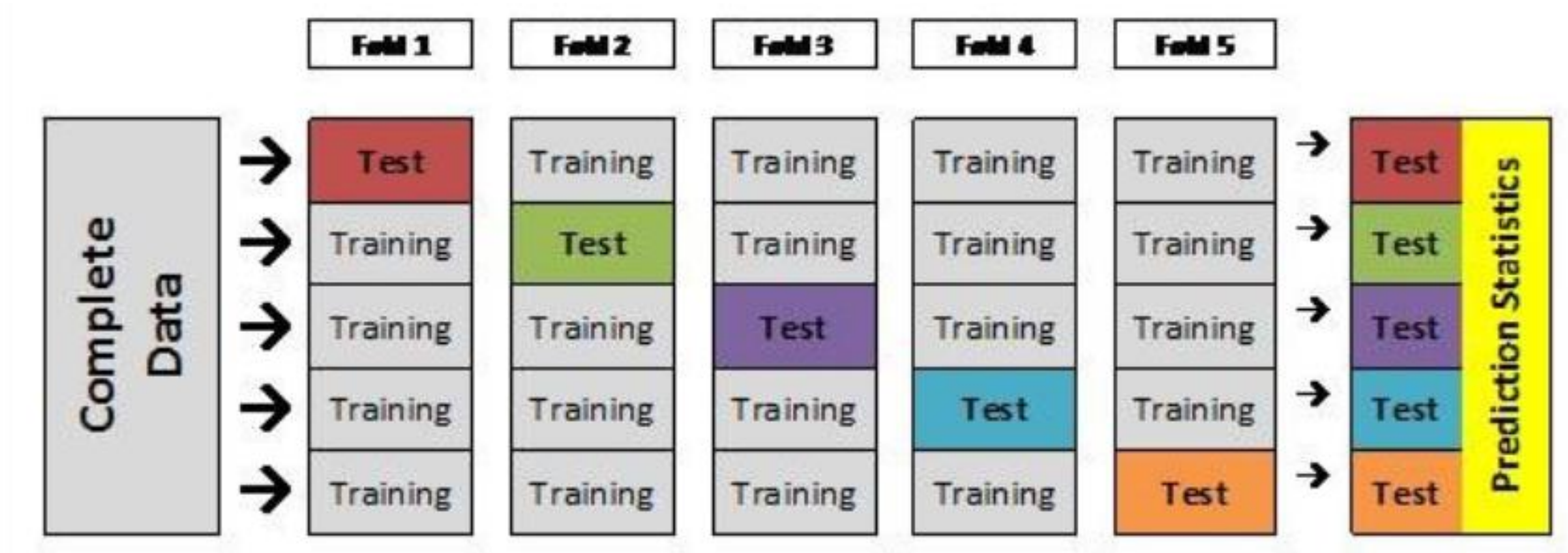
How to check GP accuracy

1. Divided the whole dataset into training data and test data
2. Using training data to predict BV of animals in test data
3. Calculate correlation between GEBV with observation or pseudo observation for animals in test data
4. Calculate intercept and regression coefficient of observations on GEBV

- **Popular strategies of validation on GEBV**
    - K-fold cross validation
    - Leave-young animals-out validation

How to check GP accuracy

- **Good**: Keep both training data and test data large
- **Bad**: not following the real life situation where candidates are young animals



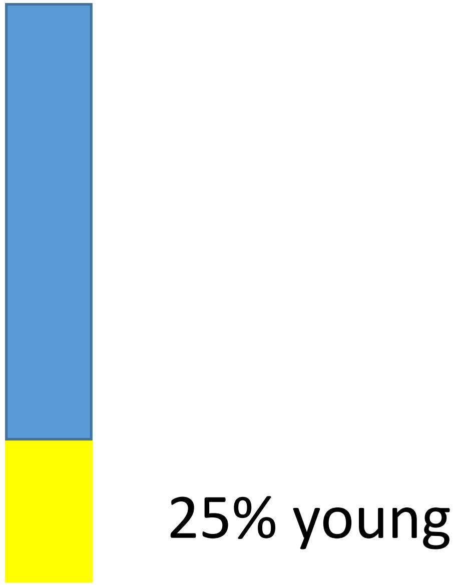
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Test	Training	Training	Training	Training
Training	Test	Training	Training	Training
Training	Training	Test	Training	Training
Training	Training	Training	Test	Training
Training	Training	Training	Training	Test

Schematic representation of 5-fold cross-validation
Golden Helix.

How to check GP accuracy

Leave-young animals-out validation

- **Good:** Consistent with real life scenario
- **Bad**: less power of validation. If the whole data is not large, test data could be too small to get reliable test



GP reliability

- **Reliability:**

$$\mathrm{R}_{\mathrm{GEBV}}^{2} = accuracy^{2} = \mathrm{cor}^{2}(GEBV, a) = \frac{\mathrm{cor}^{2}(GEBV, y)}{r_{y}^{2}}$$

y=a+e, if y is a single record of phenotype, $\mathrm{r}_{y}^{2} = h^{2} = \frac{V_a}{V_y}$

Because $\frac{\mathrm{cor}^{2}(GEBV,y)}{r_{y}^{2}} = \frac{\mathrm{cov}^{2}(GEBV,a+e)}{V_{GEBV}V_{y}\mathrm{r}_{y}^{2}}=\frac{\mathrm{cov}^{2}(GEBV,a)}{V_{GEBV}V_{a}}=\mathrm{cor}^{2}(GEBV, a)$

$\mathrm{r}_{y}^{2}$is the reliability of phenotype

y， DRP， DYD， EBV…

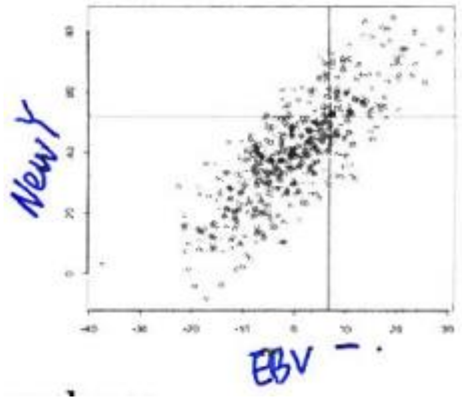
GP - Unbiasedness

y = b_0 + b_1GEBV + residual

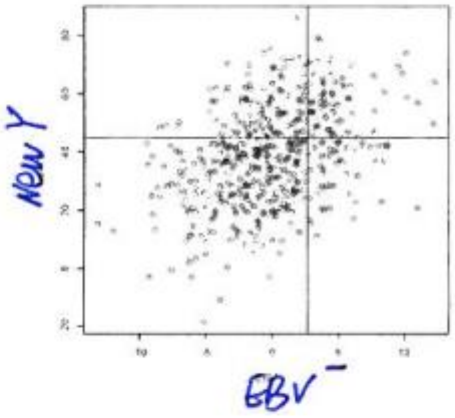
$$b_1 = \frac{cov(GEBV, y)}{V_{GEBV}}$$

if unbiased, b_0 close to 0, and b_1 close to 1.

b_1<1, GEBV inflation, V_{GEBV} big
b_1>1, GEBV deflation, V_{GEBV} small



unbiased



biased

DMU

DMU Introduction

- A package for **Mu**ltivariate analysis by restricted maximum likelihood based on a **D**erivative-free approach

- Author: Per Madsen & Just Jensen, QGG, Aarhus University

- Estimate variance
- Estimate parater
- EBV

Introduction

Modules：

- **dmu1**: Prepare program
- **dmuai** : AI-REML estimation of (co)-variance components
- **dmu4** : BLUE and BLUP in core
- **dmu5**: BLUE and BLUP iteration on data
- **rjmc**: Bayesian analysis of linear and binary traits

Input files:

- Phenotype file
- Pedigree file
- G matrix
- Parameter file: .DIR

Input file

Phenotype file

- Only number, no character
- Integer first, real later
- Missing value for interger: 0
- Missing value for real: -999/…

G matrix or G inverse matrix

- 3 columns
- First column and second column is ID
- The third column is relationship

Input file

Pedigree file

- 4 column

  Col1: individual ID

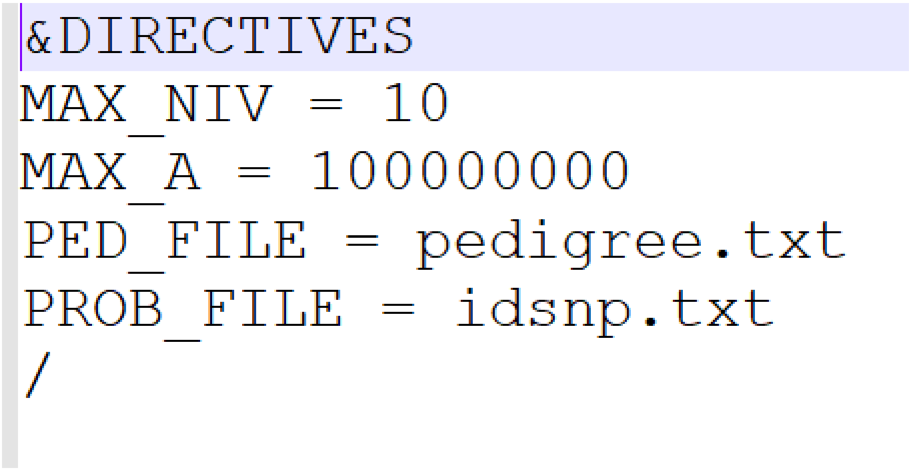
  Col2: sire ID

  col3: Dam/MGS ID

  Col4: birthdate

- Missing value: 0

Package dmutrace



```
&DIRECTIVES
MAX_NIV = 10
MAX_A = 100000000
PED_FILE = pedigree.txt
PROB_FILE = idsnp.txt
/
```

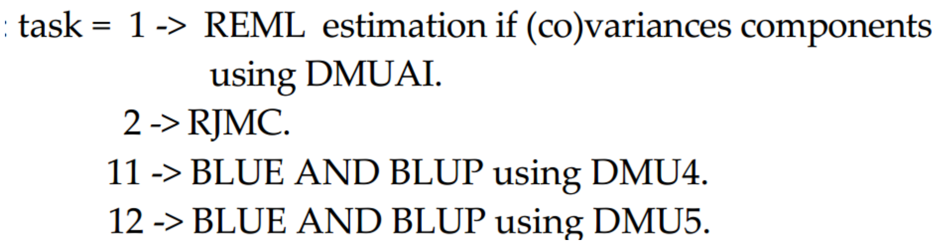
Rdmutrace XXX

.DIR

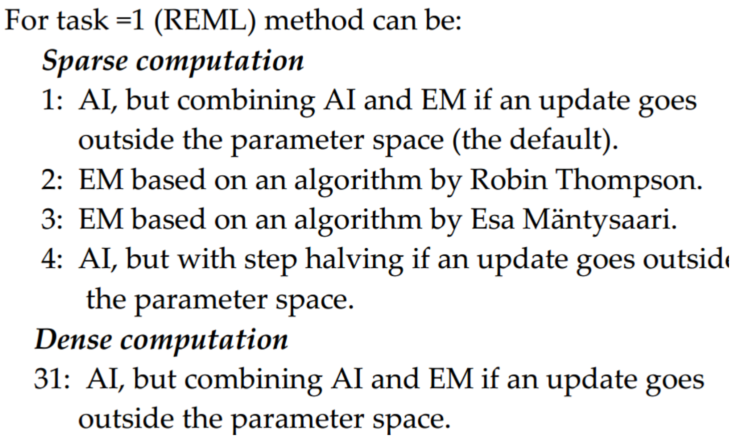
```
$COMMENT
GBLUP model for IMF
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2           3
age weight IMF
$MODEL
1
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
$DMUAI
```

1 $COMMENT	注释
2 $ANALYSE	分析方法
3 $DATA	记录数据文件
4 $VARIABLE	记录数据的变量名
5 $MODEL	模型
6 $VAR_STR	遗传方差结构定义
7 $PRIOR	方差协方差的初始值
8 $DMUAI	不同DMU模块的选项

```
$COMMENT
IMF = sex + batch + b*age + b*weight + e
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2         3
age weight IMF
$MODEL
1
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
$DMUAI
```



task = 1 -> REML estimation if (co)variances components using DMUAI.
2 -> RJMC.
11 -> BLUE AND BLUP using DMU4.
12 -> BLUE AND BLUP using DMU5.



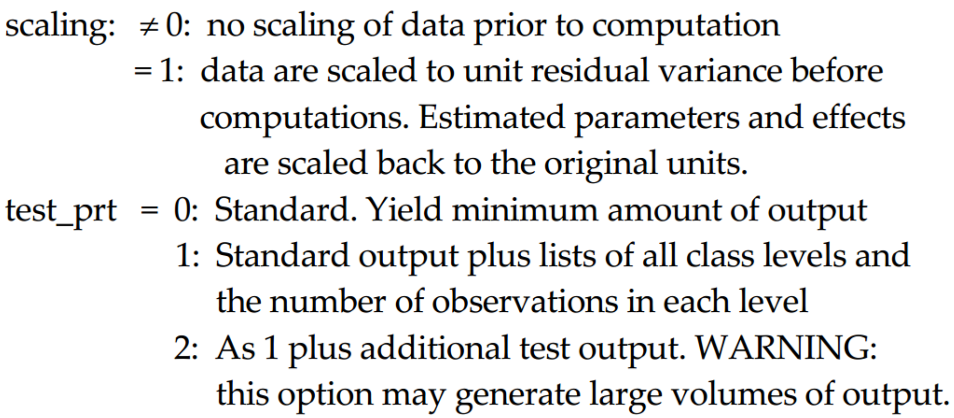
For task =1 (REML) method can be:

Sparse computation

1: AI, but combining AI and EM if an update goes outside the parameter space (the default).
2: EM based on an algorithm by Robin Thompson.
3: EM based on an algorithm by Esa Mäntysaari.
4: AI, but with step halving if an update goes outsid the parameter space.

Dense computation

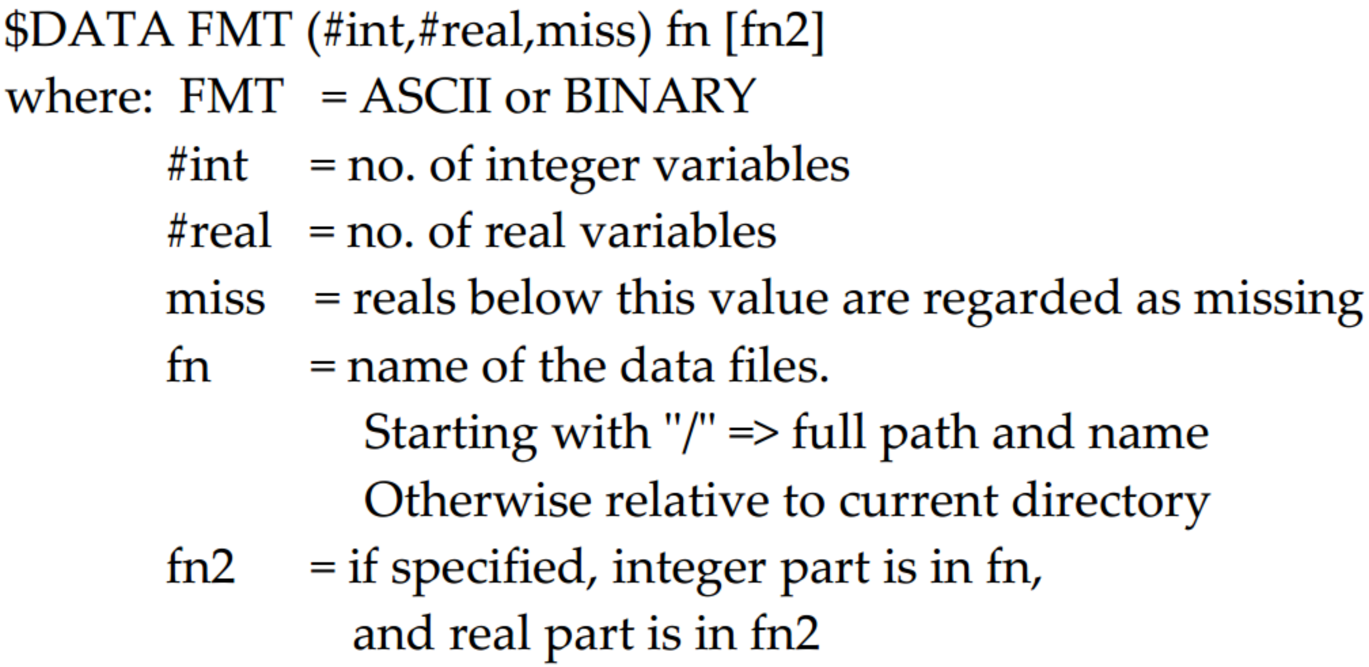
31: AI, but combining AI and EM if an update goes outside the parameter space.



scaling: ≠0: no scaling of data prior to computation
=1: data are scaled to unit residual variance before computations. Estimated parameters and effects are scaled back to the original units.

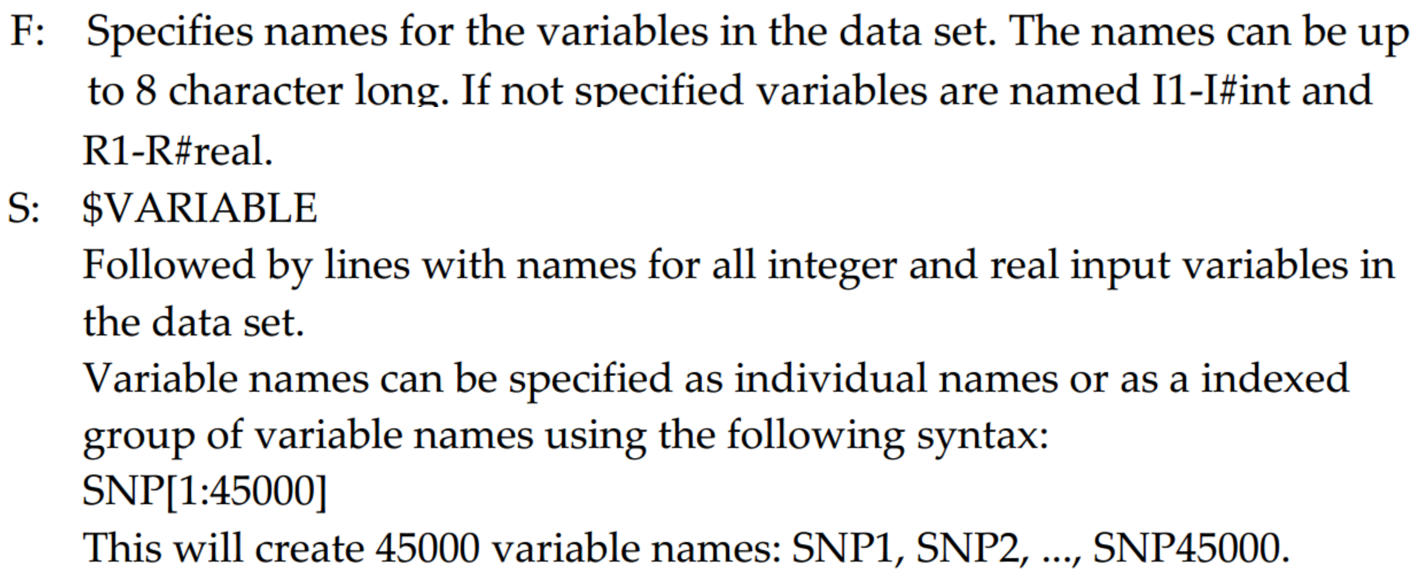
test_prt = 0: Standard. Yield minimum amount of output
1: Standard output plus lists of all class levels and the number of observations in each level
2: As 1 plus additional test output. WARNING: this option may generate large volumes of output.

```
$COMMENT
IMF = sex + batch + b*age + b*weight + e
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2          3
age weight IMF
$MODEL
1
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
$DMUAI
```



```
$DATA FMT (#int,#real,miss) fn [fn2]
where:  FMT    = ASCII or BINARY
        #int   = no. of integer variables
        #real  = no. of real variables
        miss   = reals below this value are regarded as missing
        fn     = name of the data files.
                 Starting with "/" => full path and name
                 Otherwise relative to current directory
        fn2    = if specified, integer part is in fn,
                 and real part is in fn2
```

```
$COMMENT
IMF = sex + batch + b*age + b*weight + e
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2          3
age weight IMF
$MODEL
1
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
$DMUAI
```



F: Specifies names for the variables in the data set. The names can be up to 8 character long. If not specified variables are named I1-I#int and R1-R#real.

S: $VARIABLE

Followed by lines with names for all integer and real input variables in the data set.

Variable names can be specified as individual names or as a indexed group of variable names using the following syntax:

SNP[1:45000]

This will create 45000 variable names: SNP1, SNP2, ..., SNP45000.

```
$COMMENT
IMF = sex + batch + b*age + b*weight + e
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2          3
age weight IMF
$MODEL
1 1 0 0 0
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
$DMUAI
```

- Traits
- Absorbs
- Model
- Random
- Regression
- Residual covariance

```
$MODEL
1 1 0 0 0
0
3 0 3 2 3 1
1 1
2 1 2
0
```

- # traits # Gaussian # right Censored # categorical # binary
- Only for DMU5

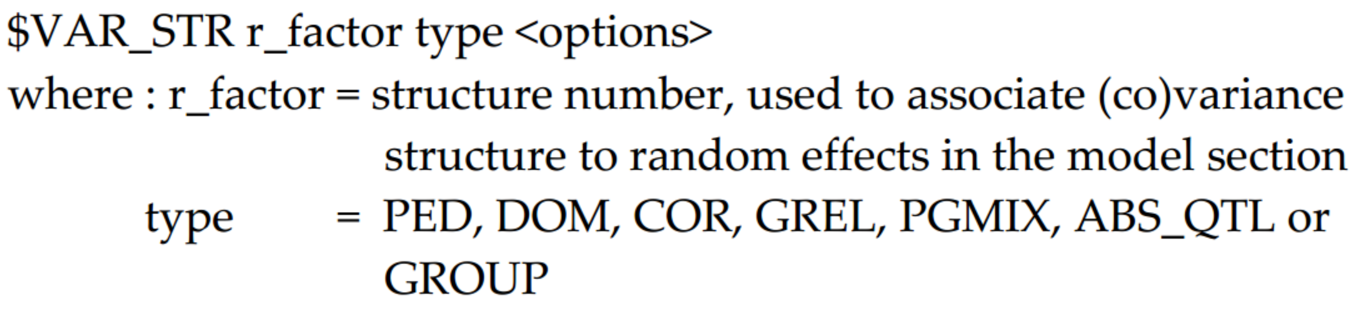
```
$MODEL
1 1 0 0 0
0
3 0 3 2 3 1
1 1
2 1 2
0
```

- **Model**
  - 1st value is real input number for the trait
  - 2nd value is real input number for a weight variable. If no weight variable is used specify zero (0)
  - 3rd value is the number of class variables (fixed and random) in the model for this trait
  - On the rest of the line integer input numbers for each class variable in the model is specified (fixed effect before random)
- **Random effect**
  - The first value is the number of random effects in the model for this trait, followed by a numbering of the random factors

```
$MODEL
1 1 0 0 0
0
3 0 3 2 3 1
1 1
2 1 2
0
```

- **Regression**
  - The 1st value is the number of regressions for this trait. If no covariables are desired for this trait, a zero must be specified
  - On the rest of the line the real input numbers for the covariables must be specified
- **number of non-existing residual covariances**

```
$COMMENT
IMF = sex + batch + b*age + b*weight + e
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2          3
age weight IMF
$MODEL
1
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
$DMUAI
```



$VAR_STR r_factor type <options>

where : r_factor = structure number, used to associate (co)variance structure to random effects in the model section

type = PED, DOM, COR, GREL, PGMIX, ABS_QTL or GROUP

- For GBLUP: $VAR_STR 1 GREL ASCII ginv-full.dat
- For PBLUP: $VAR_STR 1 PED 1 ASCII pedigree.dat
- For ssGBLUP: $VAR_STR 1 PGMIX 1 ASCII pedigree.dat idsnp.dat gmat-sub.dat 0.2 G-ADJUST
- For feature model:
  $VAR_STR 1 COR ASCII ginv-1.dat
  $VAR_STR 2 COR ASCII ginv-2.dat

```
$COMMENT
IMF = sex + batch + b*age + b*weight + e
$ANALYSE 1 31 0 0
$DATA  ASCII (5,3,-999) ref.txt
$VARIABLE
#1   2      3 4  5
id sex batch  birth  sladate
# 1   2          3
age weight IMF
$MODEL
1
0
3 0 3 2 3 1
1 1
2 1 2
0
$VAR_STR 1 GREL ASCII Gsparseinv.txt
$PRIOR
1  1  1   0.10261700
2  1  1   0.58650634
$DMUAI
```

(co)variance matrix value

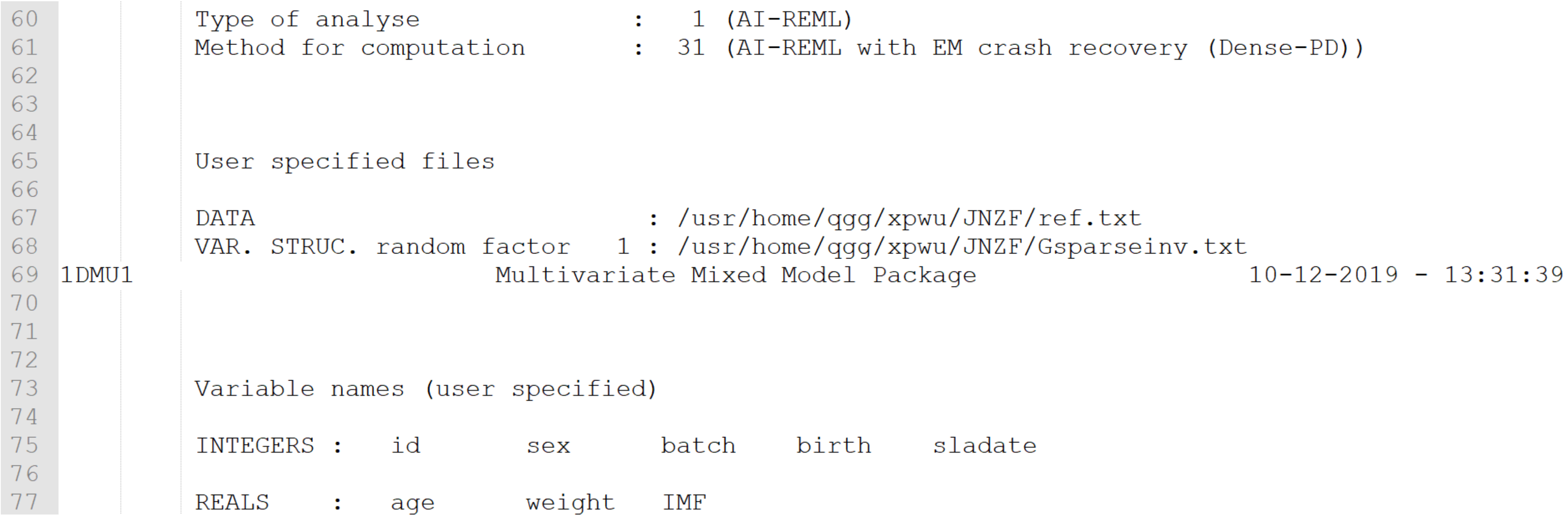
Run DMU

- **Linux**
  rdmuai *[rdmu4, rdmu5, rrjmc]* xxx.DIR

- **Windows**
  run_dmuai *[run_dmu4, run_dmu5, run_rjmc]* xxx.DIR

Output file

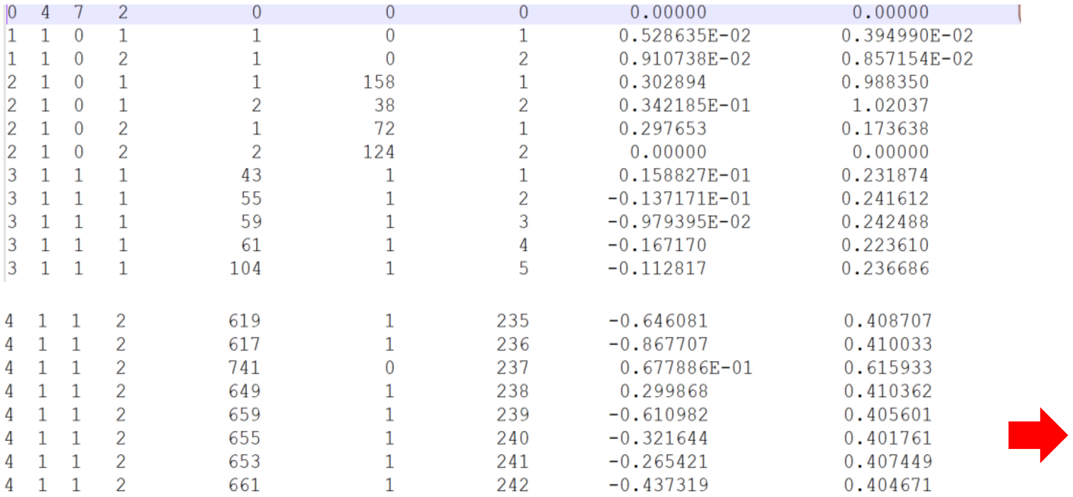
- Log file: .lst



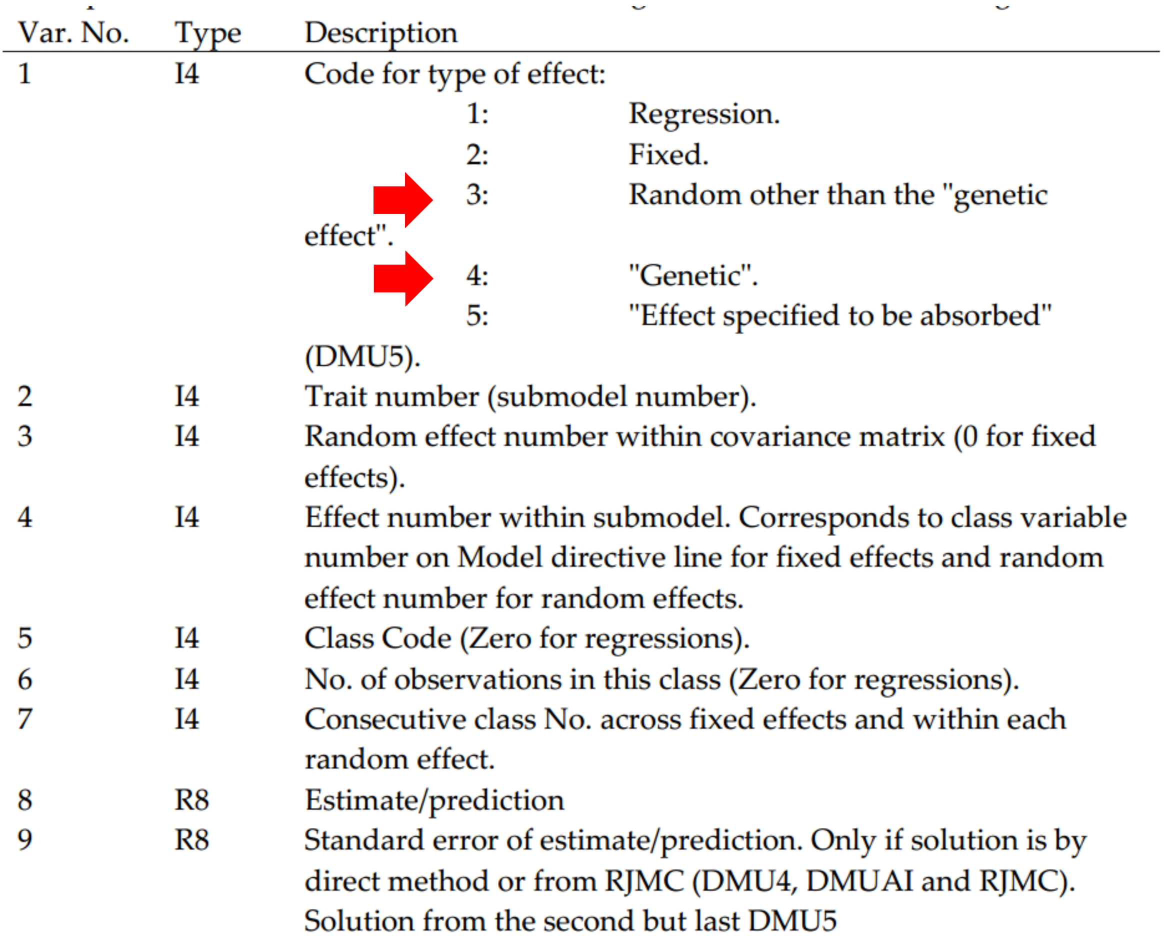
```
         Type of analyse               :    1 (AI-REML)
         Method for computation        :   31 (AI-REML with EM crash recovery (Dense-PD))



         User specified files

         DATA                           : /usr/home/qgg/xpwu/JNZF/ref.txt
         VAR. STRUC. random factor   1 : /usr/home/qgg/xpwu/JNZF/Gsparseinv.txt
1DMU1                        Multivariate Mixed Model Package                  10-12-2019 - 13:31:39



         Variable names (user specified)

         INTEGERS :    id       sex      batch    birth    sladate

         REALS    :    age      weight   IMF
```

Output file

- Solution for effects: .SOL



```
0   4   7   2          0          0          0    0.00000          0.00000
1   1   0   1          1          0          1   0.528635E-02     0.394990E-02
1   1   0   2          1          0          2   0.910738E-02     0.857154E-02
2   1   0   1          1        158          1   0.302894         0.988350
2   1   0   1          2         38          2   0.342185E-01      1.02037
2   1   0   2          1         72          1   0.297653         0.173638
2   1   0   2          2        124          2    0.00000          0.00000
3   1   1   1         43          1          1   0.158827E-01     0.231874
3   1   1   1         55          1          2  -0.137171E-01     0.241612
3   1   1   1         59          1          3  -0.979395E-02     0.242488
3   1   1   1         61          1          4  -0.167170         0.223610
3   1   1   1        104          1          5  -0.112817         0.236686

4   1   1   2        619          1        235  -0.646081         0.408707
4   1   1   2        617          1        236  -0.867707         0.410033
4   1   1   2        741          0        237   0.677886E-01     0.615933
4   1   1   2        649          1        238   0.299868         0.410362
4   1   1   2        659          1        239  -0.610982         0.405601
4   1   1   2        655          1        240  -0.321644         0.401761
4   1   1   2        653          1        241  -0.265421         0.407449
4   1   1   2        661          1        242  -0.437319         0.404671
```



Var. No.	Type	Description
1	I4	Code for type of effect: 1: Regression. 2: Fixed. 3: Random other than the "genetic effect". 4: "Genetic". 5: "Effect specified to be absorbed" (DMU5).
2	I4	Trait number (submodel number).
3	I4	Random effect number within covariance matrix (0 for fixed effects).
4	I4	Effect number within submodel. Corresponds to class variable number on Model directive line for fixed effects and random effect number for random effects.
5	I4	Class Code (Zero for regressions).
6	I4	No. of observations in this class (Zero for regressions).
7	I4	Consecutive class No. across fixed effects and within each random effect.
8	R8	Estimate/prediction
9	R8	Standard error of estimate/prediction. Only if solution is by direct method or from RJMC (DMU4, DMUAI and RJMC). Solution from the second but last DMU5

Reference

- Guosheng Su, 2019. Application of modern quantitative genetics to animal breeding.
- Luc Janss. 2014. Quantitative Genomics.
- 家畜育种学。张沅

Thank you for your attention!

Any question or suggestion?

Model reliability

$$\mathrm{R}^2_{\mathrm{GEBV}} = accuracy^2 = \mathrm{cor}^2(GEBV, a) = \frac{\mathrm{cov}^2(GEBV, a)}{V_{GEBV}V_{\mathrm{a}}} = \frac{V_{GEBV}}{V_{\mathrm{a}}}$$

assume GEBV is unbiased, $a = GEBV + \varepsilon,\ cov(GEBV, a) = 0, \varepsilon$ is prediction error.

$\mathrm{R}^2_{\mathrm{GEBV}} = \frac{V_{GEBV}}{V_{\mathrm{a}}} = 1 - \frac{\mathrm{PEV}}{V_{\mathrm{a}}}$, PEV is prediction error variance , V_{a}=$V_{GEBV} + \mathrm{PEV}$

- Does not consider phenotypic values, but data structure and estimated variances
- overestimate reliability of GEBV (SNP!=QTL, LD pattern difference between reference and test data,…)

模型预测能力 Accuracy： Cor(GEBV,a)
模型拟合能力 Unbiasedness: b = Cov(GEBV,a)/V(GEBV)