Estimation des paramètres génétiques

Jean Pierre Bidanel

Cours Supérieur d 'Amélioration Génétique des animaux Domestiques

2ème session

Grignon, 20-24 mars 2006

Estimation des paramètres génétiques

J.P. BIDANEL
Institut National de la Recherche Agronomique
Département de Génétique Animale
Station de génétique quantitative et appliquée
78352 Jouy-en-Josas Cedex - France

tél: 01-34-65-22-84
E-mail: jean-pierre.bidanel@jouy.inra.fr
Web : http://inra-sgqa.jouy.inra.fr

Paramètres génétiques – composantes de variance

Composantes de la variance	Notations	Paramètres génétiques
Variance phénotypique	$V_P=\sigma_P^2=\Sigma_J V_J$ J=A, R,D,I,M,P,C	
Variance génétique additive Variance résiduelle Variance de dominance Variance d'épistasie	$V_a=\sigma_a^2$ $V_r=\sigma_r^2$ $V_d=\sigma_d^2$ $V_i=\sigma_i^2$	Héritabilité $h^2 = V_a / V_P$
Variance maternelle Variance d'environnement permanent Variance de milieu commun	$V_m=\sigma_m^2$ $V_{pe}=\sigma_{pe}^2$ $V_c=\sigma_c^2$	$m^2 = V_m / V_P$ $P^2 = V_{pe} / V_P$ $c^2 = V_c / V_P$

Variances et covariances

- Variance : mesure l'étendue des différences entre individus
- Covariance : mesure l'importance des différences communes (entre individus ou entre caractères)

	Ressemblance familiale										
	Nulle				Moyenne				Totale		
Père	1	2	3		1	2	3		1	2	3
D1	1	1	1		2	2	1		1	2	3
D2	2	2	2		3	1	3		1	2	3
D3	3	3	3		1	2	3		1	2	3
V(EF)	Nulle				Moyenne				Forte		
V(IF)	Forte				Moyenne				Nulle		

Paramètres génétiques – composantes de covariance

Composantes de la covariance entre caractères	Notations	Paramètres
Covariance phénotypique	$cov_P = \sum_J cov_P$ J=A, R,D,I,M,P,C	Corrélation phénotypique $\rho_{P12} = \frac{cov_{P12}}{\sigma_{P1} \cdot \sigma_{P2}}$
Covariance génétique additive Covariance résiduelle Covariance de dominance Covariance d'épistasie	cov_a cov_r cov_d Cov_i	Corrélation génétique $\rho_{a12} = \frac{cov_{a12}}{\sigma_{a1} \cdot \sigma_{a2}}$
Covariance maternelle Covariance d'environnement permanent Covariance de milieu commun	cov_m cov_{pe} cov_c	ρ_m ρ_{pe} P_c

Estimation des paramètres génétiques

Pourquoi ?

- Meilleure connaissance du déterminisme génétique des caractères
- Nécessaires pour :
  - ✓ prédire la valeur génétique des reproducteurs

Indice de sélection:

$$A = \frac{cov(P,A)}{\sigma_P^2}(P - \mu_P) + e = h^2(P - \mu_P) + e$$

BLUP

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

$R = I \otimes R_0$

Matrice de variances-covariances résiduelles

$G = A \otimes G_0$

Matrice de variances-covariances génétiques

Estimation des paramètres génétiques

Exemple : évaluation génétique nationale porcine

genet

2032,9	14,85	0,17	2045,4	-9,938	-26,5	5,61	-125	9,95
14,85	1,538	0,14	-6,825	0,215	-2,65	0,33	0,24	1,05
0,169	0,142	0,02	-2,817	0,01	-0,33	0,04	0,1	0,09
2045,4	-6,825	-2,82	2588,7	-14,07	16,47	4,44	-174	-2,9
-9,938	0,215	0,01	-14,07	0,781	-0,06	-0,11	1,15	-0
-26,52	-2,647	-0,33	16,465	-0,055	7,272	-0,68	0,07	-2,3
5,612	0,331	0,04	4,439	-0,105	-0,68	1,36	0,02	0,15
-125	0,24	0,1	-174,4	1,153	0,067	0,02	18,4	-0,4
9,95	1,05	0,09	-2,863	-0,027	-2,3	0,15	-0,4	1

resid

3987,1	11,73	0	0	0	0	0	0	0
11,733	0,529	0	0	0	0	0	0	0
0	0	0,03	-6,422	0,023	-0,05	0,01	0	0
0	0	-6,42	4804,4	2,303	-25,4	-5,5	0	0
0	0	0,02	2,303	1,06	0,169	-0,02	0	0
0	0	-0,05	-25,44	0,169	2,532	0,07	0	0
0	0	0,01	-5,503	-0,022	0,074	4,69	0	0
0	0	0	0	0	0	0	64,5	1,94
0	0	0	0	0	0	0	1,94	1,22

portee

649,31	2,732	0	0	0	0	0	0	0
2,732	0,054	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	1290,2	0	0	0	0	0
0	0	0	0	0,135	0	0	0	0
0	0	0	0	0	1,685	0	0	0
0	0	0	0	0	0	0,38	0	0
0	0	0	0	0	0	0	12,1	0,36
0	0	0	0	0	0	0	0,36	0,17

Estimation des composantes de la variance

Pourquoi ?

- Nécessaires pour :
  - ✓ Prédire la réponse à la sélection et optimiser les programmes de sélection

$$\Delta G = \frac{i\rho\,\sigma_A}{t}$$

Ou plus généralement :

$$\Delta G = \frac{i_{mm}\rho_{mm} + i_{mf}\rho_{mf} + i_{fm}\rho_{fm} + i_{ff}\rho_{ff}}{t_{mm} + t_{mf} + t_{fm} + t_{ff}}\,\sigma_A$$

Attention: prédit correctement la réponse à court terme, mais pas à long terme

Estimation des paramètres génétiques

Exemple : évaluation génétique nationale porcine

genet

2032,9	14,85	0,17	2045,4	-9,938	-26,5	5,61	-125	9,95
14,85	1,538	0,14	-6,825	0,215	-2,65	0,33	0,24	1,05
0,169	0,142	0,02	-2,817	0,01	-0,33	0,04	0,1	0,09
2045,4	-6,825	-2,82	2588,7	-14,07	16,47	4,44	-174	-2,9
-9,938	0,215	0,01	-14,07	0,781	-0,06	-0,11	1,15	-0
-26,52	-2,647	-0,33	16,465	-0,055	7,272	-0,68	0,07	-2,3
5,612	0,331	0,04	4,439	-0,105	-0,68	1,36	0,02	0,15
-125	0,24	0,1	-174,4	1,153	0,067	0,02	18,4	-0,4
9,95	1,05	0,09	-2,863	-0,027	-2,3	0,15	-0,4	1

resid

3987,1	11,73	0	0	0	0	0	0	0
11,733	0,529	0	0	0	0	0	0	0
0	0	0,03	-6,422	0,023	-0,05	0,01	0	0
0	0	-6,42	4804,4	2,303	-25,4	-5,5	0	0
0	0	0,02	2,303	1,06	0,169	-0,02	0	0
0	0	-0,05	-25,44	0,169	2,532	0,07	0	0
0	0	0,01	-5,503	-0,022	0,074	4,69	0	0
0	0	0	0	0	0	0	64,5	1,94
0	0	0	0	0	0	0	1,94	1,22

portee

649,31	2,732	0	0	0	0	0	0	0
2,732	0,054	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	1290,2	0	0	0	0	0
0	0	0	0	0,135	0	0	0	0
0	0	0	0	0	1,685	0	0	0
0	0	0	0	0	0	0,38	0	0
0	0	0	0	0	0	0	12,1	0,36
0	0	0	0	0	0	0	0,36	0,17

Estimation des paramètres génétiques

Les paramètres génétiques variant notamment selon le caractère, la population et le milieu où elle est élevée

Quand doit-on (ré)estimer les composantes de (co)variances ?

- Nouveaux caractères
- Nouvelles populations
- Changements des (co)variances au cours du temps:

  - ✓Changement de milieu
  - ✓ Évolutions génétiques
    - Sélection
    - Définition/mesure des caractères

Estimation des paramètres génétiques

Les méthodes classiques

- régression
- analyse de la variance

Principe

- **Modèle statistique**
  - ✓ estimation de la (co)variation entre les performances d'individus apparentés, le plus souvent :
    - Parents/descendants
    - Demi-frères ou pleins-frères

- **Modèle génétique**
  - ✓ interprétation génétique de la (co)variation observée
    On égale l'estimation à son espérance sous l'hypothèse d'un déterminisme polygénique infinitésimal

Estimation des paramètres génétiques

Modèle génétique

Pour une population panmictique non consanguine et en équilibre de liaison, la covariance entre les valeurs génétiques g_i et g_j des individus i et j peut s'écrire (Malécot, 1948; Cockerham, 1954):

$$\mathrm{cov}(g_i, g_j) = \sum_{1 \le h+k \le n} (2\varphi_{ij})^h (d_{ij})^k \sigma^2_{hk}$$

où σ_{hk}^2 = variance génétique d'effets de gènes seuls à h loci et de couples de gènes à k loci

En l'absence d'épistasie:

$$\mathrm{cov}(g_i, g_j) = 2\varphi_{ij}\sigma^2_a + (\varphi_{R,P_j}\varphi_{M_i,M_j} + \varphi_{R,M_j}\varphi_{M_i,P_j})\sigma^2_d$$

Estimation des composantes de la variance

Exemples de relations entre apparentés

	Coefficient de				
Relation de parenté	σ^2_A	σ^2_D	σ^2_{AA}	σ^2_{AD}	σ^2_{DD}
Parent/enfant	0,5	0	0,25	0	0
Grand-parent / petit enfant	0,25	0	0,0625	0	0
Demi – germains	0,25	0	0,0625	0	0
Oncle – neveu	0,25	0	0,0625	0	0
Cousins germains	0,125	0	0,0156	0	0
Jumeaux monozygotes	1	1	1	1	1
Pleins frères	0,5	0,25	0,25	0,125	0,0625
Doubles cousins germains	0,25	0,0625	0,0625	0,0156	0,0039

Estimation des paramètres génétiques

Modèle génétique

Exemples :

Variance entre familles de demi-frères =

covariance entre demi-frères = 1/4 V_a

(ils ont en espérance 25% de gènes ident. Par asc.)

Variance intra-famille de demi-frères =

variance résiduelle = V_p - 1/4 V_a = 3/4 V_a + V_e + V_d

Variance entre familles de pleins-frères =

covariance entre pleins-frères = 1/2 V_a + V_c + 1/4 V_d

(ils ont en espérance 50% de gènes ident. par asc.)

Variance intra-famille de pleins-frères =

variance résiduelle = 1/2 V_a + V_e + 3/4 V_d

Estimation des paramètres génétiques

Régression parents-descendants (1)

Père	Mère	Moyenne	Descendant
20	26	23	17
15	17	16	11
10	20	15	16
18	6	12	8
16	12	14	9
14	14	14	10
10	8	9	10
11	20	15,5	15
16	25	20,5	14
17	11	14	12
8	11	9,5	14
18	13	15,5	15

3 possibilités

- Régression père – descendant
- Régression mère – descendant
- Régression parent moyen - descendant

Estimation des paramètres génétiques

Régression parent-descendants (2)

Régression père (ou mère) - descendant

Sous certaines hypothèses (absence d'effet de milieu important), les données peuvent être décrites par le modèle statistique:

$$y_i = \mu + b(x_i - x_m) + e_i \qquad \hat{b} = \frac{cov(x,y)}{\sigma_x^2}$$

Dans notre petit exemple :

$$\hat{b}_P = -0{,}06 \qquad \hat{b}_M = 0{,}34$$

Estimation des paramètres génétiques

Régression parent-descendants (3)

Sous les hypothèses (modèle génétique) :
d'un déterminisme polygénique infinitésimal
d'une population panmictique, non consanguine et sans d.l.

$$cov(x,y) = cov(P,D) = 1/2\ \sigma_a^2$$

Et: $$E(\hat{b}) = = \frac{0,5\ \sigma_a^2}{\sigma_y^2} = \frac{h^2}{2}$$

Donc : $$\hat{h}^2 = 2\,\hat{b}$$

Dans notre petit exemple :

$$\hat{h}_P^2 = -0,12 \qquad \hat{h}_M^2 = 0,68$$

Estimation des composantes de la variance

Régression parent-descendants (4)

Régression parent moyen - descendant

Sous les hypothèses d'un même déterminisme génétique, d'une égalité des variances dans les 2 sexes et absence d'effet de milieu important), les données peuvent être décrites par le modèle statistique:

$$y_i = \mu + b\, x_i + e_i$$

avec: $x_i = 0{,}5\,(y_{père} + y_{mère})$

On a:

$$\text{var}(x_i) = 0{,}25\,[\text{var}(y_{père}) + \text{var}(y_{mère}) + 2\text{cov}(y_{père}, y_{Mère})]$$

Estimation des composantes de la variance

Régression parent-descendants (5)

Si les pères et les mères sont accouplés au hasard et non apparentés

$$\text{cov}(y_{\text{père}}, y_{\text{Mère}}) = 0$$

Si les pères et les mères ne sont pas sélectionnés

$$\text{var}(y_{\text{père}}) = \text{var}(y_{\text{père}}) = \sigma_y^2$$

et :

$$\text{var}(x_i) = 0{,}5\,\sigma_y^2$$

D'où : $$\hat{b} = \frac{\text{cov}(x,y)}{\sigma_x^2}$$

Estimation des paramètres génétiques

Régression parent-descendants (6)

Par ailleurs, sous les hypothèses (modèle génétique) :

d'un déterminisme polygénique infinitésimal

d'une population panmictique, non consanguine et sans d.l.

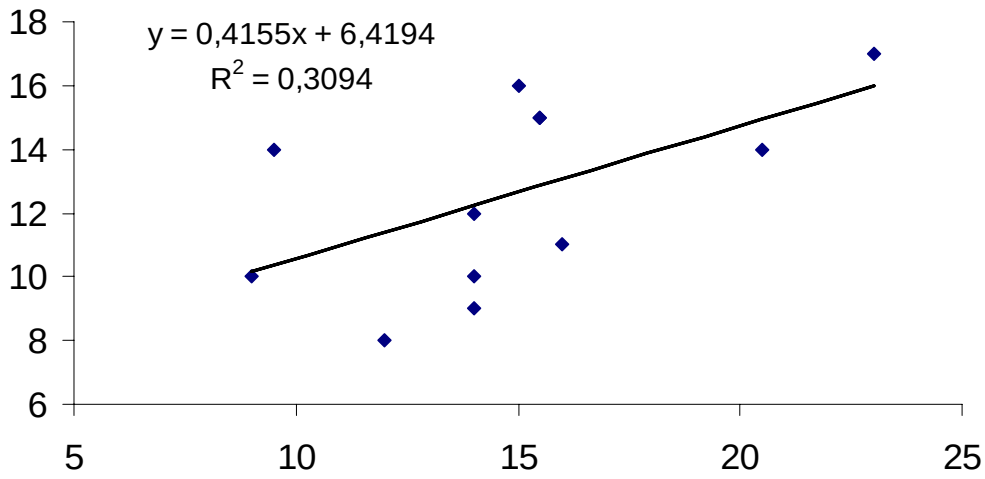
$$\mathrm{cov}(x,y) = \mathrm{cov}(\bar{P},D) = 1/2\ \sigma_a^2$$

Et l'on aura donc :

$$E(\hat{b}) = = \frac{0{,}5\ \sigma_a^2}{0{,}5\ \sigma_y^2} = h^2$$

Soit, pour l'exemple numérique

$$\hat{h}^2 = 0{,}42$$



Estimation des paramètres génétiques

Analyse de variance (1)

Principe :

Détecter l'importance des différentes sources de variation

Cette importance est déterminée par la contribution de chaque effet à la variation

Elle est obtenue à partir des sommes de carrés et des DDL

Exemple :

père	P1	P1	P1	P2	P2	P2	P3	P3	P3
$y_{descendants}$	8	12	10	7	13	12	15	11	14

Modèle : $Y_i = \mu + p_i + e_{ij}$

avec : $E(p) = 0$ $E(e) = 0$ $v(p)=I\sigma_p^2$ $v(e)=I\sigma_e^2$ $cov(p,e)=0$

Estimations : $\mu = 11$ $p_1 = -2$ $p_2 = 0$ $p_3 = 2$

Estimation des paramètres génétiques

Analyse de variance (2)

Calcul des sommes de carrés :

Moyenne	$N\mu^2$	9×11^2	1089	SCM
Père	Σp_i^2	$(-2)^2+(-2)^2+(-2)^2+2^2+2^2+2^2$	24	SCP
Résiduelle	Σe_i^2	$(-1)^2+1^2+(-4)^2+3^2+1^2+2^2+(-2)^2$	36	SCE
Total	Σy_i^2	$8^2+9^2+10^2+7^2+14^2+12^2+15^2+11^2+13^2$	1149	SCT

Tableau d'analyse de variance :

	SC	DDL	CM	E(CM)	
Moyenne	1089	1	1089		
Père	24	2	12	$\sigma_e^2 + 3\,\sigma_p^2$	Var. entre pères
Résiduelle	36	6	6	σ_e^2	Var. intra-père
Total	1149	9			

Estimation des paramètres génétiques

Analyse de variance (3)

$$\hat{\sigma}_e^2 = CM_R = 6 \qquad \hat{\sigma}_P^2 = (CM_P - CM_R)/k = (16-6)/3 = 2$$

Sous les hypothèses suivantes :

- déterminisme polygénique infinitésimal
- population panmictique, non consanguine et sans d.l.
- les parents ne sont pas apparentés
- les parents ne sont pas sélectionnés

et donc :

$$\sigma_p^2 = \text{cov}(DF) = 1/4\sigma_a^2$$

$$\hat{h}^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} = \frac{4\sigma_p^2}{4\sigma_p^2 + \sigma_e^2}$$

$$\hat{h}^2 = \frac{4*2}{(4*2)+6} = 0{,}57$$

Estimation des paramètres génétiques

Précision :
erreur standard des estimations de l'héritabilité

Nombre de données	Héritabilité vraie	
	0,1	0,3
100	0,18	0,30
500	0,08	0,14
1000	0,06	0,10
5000	0,03	0,04

Estimation des paramètres génétiques

Précision :
erreur standard des estimations de l'héritabilité

Asymptotiquement :

$$\mathrm{Var}(CM_x) \cong \frac{2CM_x^2}{ddl_x + 2}$$

$$\mathrm{Var}(\hat{\sigma}_P^2) \cong \mathrm{Var}\left(\frac{CM_P - CM_R}{k}\right)$$

Comme les 2 CM sont distribués indépendamment :

$$\mathrm{Var}(\hat{\sigma}_P^2) \cong \frac{2}{k^2}\left(\frac{CM_P}{N+1} + \frac{CM_R}{T-N+2}\right)$$

Estimation des composantes de la variance

Cas de l'analyse de variance
Modèle père-mère –cas équilibré

$$Y_{ijk} = \mu + s_i + m_{ij} + e_{ijk}$$

Source de variation	ddl	Carré moyen (CM)	E(CM)
Entre pères	s-1	MS_S	$\sigma_R^2 + k\sigma_M^2 + d\sigma_S^2$
Entre mères (intra-père)	s(d-1)	MS_M	$\sigma_R^2 + k\sigma_M^2$
Intra-mère	sd(k-1)	MS_R	σ_R^2

Estimation des composantes de la variance

Analyse de variance (Henderson I)
Modèle père-mère –cas déséquilibré

Source de variation	Sommes de carrés	Espérance			
		μ	σ_s^2	σ_d^2	σ_e^2
Moyenne	SCM= $y^2_{...}/N$	N	$q_1=\Sigma_i\, n^2_{i.}/N$	$q_2= \Sigma_{ij}\, n^2_{ij}/N$	1
Entre pères	SCS= $\Sigma\Sigma_i\, y^2_{i..}/n_{i.}$	N	N	$q_3= \Sigma_i\, (\Sigma_j\, n^2_{ij}/\, n_{i.})$	n_s
Entre mères	SCD= $\Sigma_{ij}\, y^2_{ij.}/n_{ij}$	N	N	N	n_d
Total	SCT=$\Sigma_{ijk}\, y^2_{ijk}$	N	N	N	N

$\hat{\sigma}_e^2$ = (SCT – SCD)/(N- n_d)

$\hat{\sigma}_d^2$ = (SCT – SCS-(N- n_s) $\hat{\sigma}_e^2$)/(N- q_3)

$\hat{\sigma}_s^2$ = (SCS – SCM-(n_s-1) $\hat{\sigma}_e^2$ -(q_3- q_2) $\hat{\sigma}_d^2$)/(N- q_1)

Estimation des composantes de la variance

Analyse de variance (Henderson I) - Exemple

Type 1 Analysis of Variance

Source	DF	Sum of Squares	Mean Square
pere	9	107.326190	11.925132
mere(pere)	10	65.583333	6.558333
Error	15	83.833333	5.588889
Corrected Total	34	256.742857	.

Source	Expected Mean Square
pere	Var(Error) + 2.0524 Var(mere(pere)) + 3.4349 Var(pere)
mere(pere)	Var(Error) + 1.45 Var(mere(pere))
Error	Var(Error)

Variance Component	Estimate
Var(pere)	1.44517
Var(mere(pere))	0.66858
Var(Error)	5.58889

Estimation des composantes de la variance

Composantes observées	Interprétation génétique des composantes observées	
Pères	σ_S^2=cov(DF)	$1/2\sigma_a^2$
Mères	σ_M^2=cov(PF)-cov(HS)	$1/4\sigma_a^2$+ $1/4\sigma_d^2$ + σ_C^2
Descendants	σ_R^2=Var(P)-cov(FS)	$1/2\sigma_a^2$+ $3/4\sigma_d^2$ + σ_e^2
Total	σ_T^2= σ_S^2+ σ_M^2 + σ_R^2 = Var(P)	σ_a^2+σ_d^2 +σ_c^2+σ_e^2
Pères + mères	σ_S^2+ σ_M^2 = cov(PF)	$1/2\sigma_a^2$+ $1/4\sigma_d^2$ + σ_C^2

Estimation des composantes de la variance

Cas général : la méthode d' Henderson III

- Le nombre d'observations par famille est déséquilibré
- Les données sont influencées par des effets de milieu importants
- Principe :
  - Modèle statistique : modèle linéaire mixte

$$Y_{ijk} = \mu + b_i + s_j + e_{ijk}$$

  - Formalisme matriciel

$$Y = Xb + Zu + e$$

  - Les sommes de carrés deviennent des formes quadratiques

$$SCT = \sum_{ijk} y_{ijk}^2 = Y'Y$$

Estimation des composantes de la variance

Cas général : la méthode d' Henderson III

- Présente dans de nombreux logiciels

Ex: SAS Proc varcomp (method = type I)

Source	DF	Sum of Squares	Mean Square
elev	2	3.642857	1.821429
pere	9	104.772817	11.641424
mere(pere)	9	68.993850	7.665983
Error	14	79.333333	5.666667
Corrected Total	34	256.742857	

elev	Var(Error) + 1.9473 Var(mere(pere)) + 1.2495 Var(pere) + Q(elev)
pere	Var(Error) + 1.9525 Var(mere(pere)) + 3.1573 Var(pere)
mere(pere)	Var(Error) + 1.2783 Var(mere(pere))
Error	Var(Error)

Variance Component	Estimate
Var(pere)	0.92516
Var(mere(pere))	1.56406
Var(Error)	5.66667

Estimation des composantes de la variance

Méthode d' Henderson III : inconvénients

- Basée sur l'espérance de formes quadratiques
  - ne peut s'appliquer à un modèle animal
  - sensible aux effets de la sélection, notamment des pères
    - ex : corrélation entre ½ germains avec sélection des pères

Var_A(entre pères) = $(1 - kh^2)\sigma^2_A$

Var_P(descendants) = $(1 - 1/4kh^4)\ \sigma^2_A$

d'où :

$$\frac{Var_A}{Var_P} = \frac{(1-kh^2)}{(1-1/4kh^4)}h^2 \neq h^2$$

Sous estimation car :

$kh^2 > 1/4kh^4$

Résultats d'une simulation
(h2=0,25 ; 25% mâles sélectionnés)

Valeur vraie = 10	Pas de sélection	Sélection
Henderson III	10,02	6,51
MIVQUE-MA	9,96	10,02

Estimation des composantes de la variance

Les méthodes du maximum de vraisemblance

Chaque observation a une densité de probabilité, caractérisée par:

- sa distribution
- sa moyenne
- sa variance

- Exemple :distribution normale de moyenne μ et de variance σ^2

$$f(y)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{1(y-\mu)^2}{2\ \sigma^2}}$$

f donne la probabilité de l'observation sachant μ et σ^2

Vraisemblance : on « renverse » le raisonnement pour obtenir la « probabilité » (vraisemblance) des paramètres sachant y

Estimation des composantes de la variance

Les méthodes du maximum de vraisemblance

Cas de données décrites par un modèle linéaire mixte

$$y = Xb + Za + e$$

La vraisemblance s'écrit :

$$p(y|b,\sigma^2) = (2\pi)^{-N/2} |V|^{-1/2} e^{[-(y-Xb)'V^{-1}(y-Xb)/2]}$$

La vraisemblance étant une fonction positive, on peut travailler sur son logarithme:

$$L(b,V|X,y) = -\frac{1}{2} N \log(2\pi) - \frac{1}{2} \log(|V|) - \frac{1}{2} (y-Xb)'V^{-1}(y-Xb)$$

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)

On travaille sur des combinaisons linéaires des données corrigées pour les effets fixés

Sauf dans le cas de modèles très simples, il n'y a pas d'expression analytique des solutions :

On recherche le maximum de la fonction de vraisemblance
De façon itérative à partir de valeurs de départ

Il existe de nombreux algorithmes pour la recherche du maximum de la fonction de vraisemblance

- utilisant les dérivées premières/secondes
  (Newton-Raphson, EM, scores de Fisher, AIREML, …)
- sans utilisation des dérivées
  (DF-REML)

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)
Recherche du maximum de la fonction

Exemple simple:

$$Y = Xb + Za + e$$

1 – on résout les EMM en utilisant une valeur a priori pour les composantes de variance

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

2 – on estime les composantes à partir des solutions des EMM

$$\sigma_a^2 = \left[\hat{a}'A^{-1}\hat{a} + tr(A^{-1}C)\sigma_e^2\right] / q$$

$$\sigma_e^2 = \left[y'y - b'X'y - \hat{a}Z'y\right] / (N - r(X))$$

3 – on utilise un nouveau $\lambda = \sigma_e^2 / \sigma_a^2$ pour une nouvelle itération

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)
Intérêt / méthodes d'Henderson

- Basé sur les équations du modèle mixte
  - Permet d'utiliser un modèle animal
    - Prend en compte sous certaines conditions les effets de la sélection
    - combine automatiquement et de façon optimale les informations issues des différentes relations de parenté du pedigree
    - permet une grande flexibilité de modèles
      - Modèles multicaractères
      - Modèles avec effet maternel

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)
Nombreux logiciels disponibles

Sous SAS

Proc Mixed ou Proc Varcomp (mais pas de matrice de parenté)

Logiciels spécialisés

VCE (Groeneveld et al, 1998; Kovac et al, 2002)
ASREML (Gilmour et al, 2002)
BGF90 (Misztal et al, 2002)
DMU (Jansen et Madsen, 1994)
DFREML (Meyer, 1988)
MATVECH (Kachman et Fernando, 2002)
WOMBAT (Meyer, 2006)

Estimation des composantes de la variance

REML – modèle père-mère – Exemple SAS Varcomp

Variance Components Estimation Procedure
REML Iterations

Iteration	Objective	Var(pere)	Var(mere(pere))	Var(Error)
0	63.9163615044	0.4010991400	3.0949260591	4.8690200882
1	63.8237808004	0.5081473542	2.3111096739	5.2535803502
2	63.8229819350	0.5243462441	2.3648866120	5.2117619517
3	63.8229499359	0.5260194651	2.3519734480	5.2187507789
4	63.8229499353	0.5260068351	2.3520355904	5.2187195224

Variance Component	Estimate	Henderson III
Var(pere)	0.52601	0.92516
Var(mere(pere))	2.35204	1.56406
Var(Error)	5.21872	5.66667
h2	0.26	0.45

Estimation des composantes de la variance

REML – modèle animal – Exemple VCE

```
.....VCE 4.2.5.....
Mon Sep  2 11:56:45 2002                                  page     1

 ***********************************************************************
 *            G E N E R A L  I N F O R M A T I O N            *
 ***********************************************************************
  VCE was started on :  at Mon Sep  2 11:56:45 by : ugenjpb
  Comments:

  _________
 .  ***********************************************
 .  *                  VCE4                       *
 .  *              version 4.2.5                  *
 .  *           04-déc-1998 @ 08:41:2             *
 .  *            AIX 2 000039128900               *
 .  *                written by                   *
 .  *             Eildert Groeneveld              *
 .  *               eg@tzv.fal.de                 *
 .  ***********************************************
 .
  Files involved:

  _________________
  current directory      : /ugen/ugenjpb/BOURGES
  parameter file         : tsmb10
  data input file        : /prod19/jpb/BOURGES/data3.b10   n =  1958
  pedigree input file    : /prod19/jpb/BOURGES/gen.b10     n =  1355
  log list file          : L_tsmb10
  binary parameter log   : S_tsmb10
  solutions file         : sol_tsmb10
```

Version de VCE (il en existe une plus récente)

Information sur les fichiers (nom, emplacement,...

Fichier paramètre

Fichier de données

Fichier de données

Fichiers sorties

Estimation des composantes de la variance

REML – modèle animal – Exemple VCE

```
*****************************************************************
*                  D A T A   I N F O R M A T I O N              *
*****************************************************************
```

General statistics:

traits	# rec.	min.	max.	avg.	std.
car1	1933	132.00000	331.00000	242.33109	24.55644
car2	1958	0.62571	2.20000	1.33052	0.21144
car3	1958	0.80000	23.55000	12.64998	3.67101
car4	1945	0.00000	0.82731	0.21423	0.07653
aget_reg	1958	309.00000	795.00000	506.18182	87.22604
consm_reg	1958	0.00000	0.30599	0.04756	0.04422
cons_reg	1958	0.00000	0.34402	0.05680	0.04469

Variables analysées

Covariables

Pattern of Traits:

The following frequencies were counted:

car1	car2	car3	car4	COUNT
x	x	x	x	1920
-	x	x	x	25
x	x	x	-	13

Structure des données : l'essentiel des animaux (1920) ont été mesurés pour les 4 caractères

Estimation des composantes de la variance

****
* P E D I G R E E I N F O R M A T I O N *
****

Inbreeding	# of animals
0	742
0 < 5	319
5 < 10	198
10 < 15	82
15 < 20	9
25 < 30	5

Average inbreeding : 6. (of inbred animals)
Maximum inbreeding : 27.

Information sur
la consanguinité,
qui est prise en compte
dans le calcul de
la matrice de parenté

Estimation des composantes de la variance

```
************************************************************************
*   C O V A R I A N C E   M A T R I X   I N F O R M A T I O N      *
************************************************************************
```

The following (co)variance matrices will be estimated (starting values):

Resid./pro	car1	car2	car3	car4
	0.15E+03	0.10E-02	0.10E-02	0.10E-02
		0.11E-01	0.10E-02	0.10E-02
			3.4	0.10E-02
				0.15E-02

portee	car1	car2	car3	car4
	0.15E+03	0.10E-02	0.10E-02	0.10E-02
		0.11E-01	0.10E-02	0.10E-02
			3.4	0.10E-02
				0.15E-02

envp	car1	car2	car3	car4
	0.15E+03	0.10E-02	0.10E-02	0.10E-02
		0.11E-01	0.10E-02	0.10E-02
			3.4	0.10E-02
				0.15E-02

animal	car1	car2	car3	car4
	0.15E+03	0.10E-02	0.10E-02	0.10E-02
		0.11E-01	0.10E-02	0.10E-02
			3.4	0.10E-02
				0.15E-02

Thus, optimization is in 40 dimensions.

Valeurs de départ des composantes de (co)variance
Option par défaut
Variances : valeurs de départ égales pour toutes les composantes
Covariances: proches de zéro (milieu de l'espace des paramètres)

Estimation des composantes de la variance

```
*************************************************************************
*                    M O D E L   I N F O R M A T I O N                  *
*************************************************************************
 Factor       T  nested   # skp   car1   car2  car3  car4
 ___________________________________________________________________
 aget_reg     C npo          2      x     x     x     x
 consm_reg    C              1      x     x     x     x
 cons_reg     C              1      x     x     x     x
 common       R            603      x     x     x     x
 perm         R           1131      x     x     x     x
 an           A           1355      x     x     x     x
 bande        F             98      x     x     x     x
 npom         F              2      x     x     x     x

*************************************************************************
*   C O E F F I C I E N T   M A T R I X   I N F O R M A T I O N     *
*************************************************************************
 Setting up of equations:
 ___________________________
 # of equations                :    12772
 rank of system                :    12767
 equations set to zero         :    bande/49
 # of nonzero coefficients (HS) :    477552
 fill of coefficient matrix    :     0.293%
 # of NZE in factor            :     1066282
 Total storage required        :     3794166
 Total storage defined (total) :     5000000
 CPU-time for solving    (per rnd):    6.89
 CPU-time for inverting (per rnd):     9.77
 MFLOPs during factorization   :     83.82
```

Effets du modèle
C: covariable
R: aléatoire autre que A
A : aléatoire proportionnel à A
F: fixé

Tous les caractères sont décrits par le même modèle

Informations techniques sur l'analyse

Estimation des composantes de la variance

```
**********************************************************************
*          E S T I M A T E S   I N F O R M A T I O N          *
**********************************************************************
 Mon Sep  2 12:38:31 2002                CPU time used: 00:41:14

 AG Log likelihood :   9794.2976 status : 1 at iteration:    163/  163
 The following covariance matrices were estimated:
 ----------------------------------------------------------

 Resid./pro   car1       car2       car3       car4
              176.283    0.169      13.678     0.020
                         0.024      -0.012     -0.002
                                    8.666      0.008
                                               0.005
 portee       car1       car2       car3       car4
              13.897     -0.036     0.587      -0.040
                         0.000      -0.008     0.000
                                    0.443      -0.002
                                               0.000
 envp         car1       car2       car3       car4
              57.451     -0.189     -2.213     0.064
                         0.003      -0.008     0.000
                                    1.844      0.018
                                               0.000
 animal       car1       car2       car3       car4
              92.742     0.595      6.223      0.074
                         0.013      0.093      0.002
                                    1.709      0.019
                                               0.001
```

Les résultats : variances et covariances

Estimation des composantes de la variance

Les résultats en % variance phénotypique et corrélations

Les écart types d'échantillonnage (approchés)

these are the corresponding ratios:

.....VCE 4.2.5.....

Resid./pro	car1	car2	car3	car4
	0.518	0.082	0.350	0.022
		0.604	-0.026	-0.219
			0.684	0.039
				0.802

portee	car1	car2	car3	car4
	0.041	-0.718	0.237	-0.996
		0.005	-0.846	0.777
			0.035	-0.322
				0.020

envp	car1	car2	car3	car4
	0.169	-0.467	-0.215	0.415
		0.072	-0.116	0.075
			0.146	0.647
				0.071

animal	car1	car2	car3	car4
	0.272	0.548	0.494	0.311
		0.320	0.631	0.653
			0.135	0.594
				0.107

h^2 r_G

Standard errors of ratios:

Resid./pro	car1	car2	car3	car4
	0.019	0.021	0.019	0.021
		0.020	0.021	0.020
			0.021	0.021
				0.021

portee	car1	car2	car3	car4
	0.017	0.451	0.270	0.046
		0.005	0.312	0.416
			0.015	0.396
				0.010

envp	car1	car2	car3	car4
	0.023	0.168	0.079	0.096
		0.017	0.132	0.191
			0.023	0.128
				0.021

animal	car1	car2	car3	car4
	0.031	0.070	0.081	0.095
		0.024	0.081	0.086
			0.021	0.087
				0.019

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)

ATTENTION

- AU MODELE D'ANALYSE
  - Les qualités des estimateurs supposent que le modèle de description des données est correct
  - Oublier un facteur de variation important -> biais
    - effets de milieu commun de la portée de naissance, ou de milieu permanent (données répétées) oubliés -> surestimation de h2
  - A l'inverse, des modèles trop complexes et surparamétrés sont moins robustes
  - Modèle génétique
    - Gène majeur -> biais potentiel

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)

ATTENTION

- AUX DONNEES ANALYSEES
  - Nombre de données (totales, par groupe de contemporains,…)
  - Confusion d'effets
  - Traitements préférentiels non corrigés
  - interactions G x E non corrigées
  - Qualité des généalogies
    - effet maternel : qualité des généalogies maternelles
    - effet de milieu permanent : données répétées
    - effet de milieu commun portée : plusieurs descendants par portée
    - Corrélation r(d,m) : données sur les mères

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)
Importance du modèle de description des données

Exemple : analyse du poids au sevrage chez (Suffolk)
9 700 données; 15 000 génalogies

Modéle	Avec effets maternels (EM)		Sans EM
	ρ(d,m) estimé	ρ(d,m)=0	
VarP	23,45	23,26	23,94
h^2 directe	0,25 (0,04)	0,19 (0,03)	0,44 (0,03)
h^2 maternelle	0,28 (0,04)	0,18 (0,02)	-
ρ(d,m)	-0,44 (0,10)	-	-

Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)

Comparaison de modèles

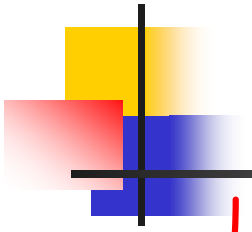
Modèles « emboités » : test du rapport de vraisemblance (TRV)

$$RV = \quad -2\ln\frac{MV(\text{mod}\,\text{èle réduit})}{MV(\text{mod}\,\text{èle complet})}$$

RV suit une distribution du chi-2 à n ddl

ddl = différence de nombre de paramètres entre le modèle réduit et le modèle complet

ATTENTION : certains logiciels (VCE) ne donnent pas la valeur de la vraisemblance, mais uniquement d'une fraction de celle-ci les valeurs sont donce inutilisables pour un TRV



Estimation des composantes de la variance

Le maximum de vraisemblance restreint (REML)

Comparaison de modèles

- Modèles non emboités
  - Critères de décision (Akaike, Schwartz, BIC,...)
- Autres éléments à considérer:
  - ➢ considérations théoriques
  - ➢ informations a priori sur les facteurs de variation
  - ➢ information fournie par les données
  - ➢ Etude de la robustesse des modèles
  - ➢ Analyse des résidus
  - ➢ Techniques de validation croisées

Estimation des composantes de la variance

Autres méthodes

- MIVQUE = « MInimum Variance QUadratic Estimation »

  Equivalent à la première itération d'un REML

  - Méthode pas très performante si l'on part de valeurs très différentes des valeurs vraies
  - SAS Varcomp (method = Mivque0) : part de valeurs nulle des composantes (autres que résiduelle)
- MINQUE = « MInimum Norm QUadratic Estimation »
- Méthode « R » = méthode empirique
- Méthodes bayésiennes
  - avec utilisation des méthodes de type « Monte Carlo »
  - Très puissant et flexible, mais très lourd sur le plan calculatoire

Estimation des composantes de la variance

Aspects pratiques
1 – population de base sélectionnée

- Population de base ?
  - ➢ En général définie par rapport aux données disponibles
- Performances manquantes, mais pas les généalogies
  - ➢ L'inclusion de généalogies réduit les biais
- Performances et généalogies manquantes
  - ➢ Pas de solution globalement satisfaisante
  - ➢ Dépend de l'information manquante
  - ➢ Quelques propositions
    - inclusions de groupes de parents inconnus
      - convergence ralentie et/ou plus difficile
    - considérer les individus de base comme fixés

Estimation des composantes de la variance

Aspects pratiques

2 – Stratégies de traitement des fichiers de grande taille

- Situation idéale
  - Utilisation des mêmes données que pour l'évaluation génétique (ensemble des données)
- Des astuces calculatoires permettent dans certains cas de se ramener à des analyses univariates
  - Nécessite une programmation spécifique (non disponible dans les logiciels « généraux » existants)
- Echantillonnage
  - Doit être bien réfléchi pour éviter une sélection liée au choix des données
  - Scission du fichier global en sous-fichiers
  - Calcul sur des sous-ensembles de caractères