Impact d'une structure de dépendance résiduelle non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié

Tom Rohmer, Anne Ricard & Ingrid David

Genotoul biostat-bioinfo day 2021
12 october 2021

Summary

1. Introduction
   - Genetic context
   - Copulas

2. Population sampling

3. Results

Summary

1 Introduction
- Genetic context
- Copulas

2 Population sampling

3 Results

Mixed model

In a genetic animal context, multivariate mixed animal models are widely used, to deal with genetic and environmental effect.
That is for the bivariate case, for $\boldsymbol{y}_j = (y_{1j}, \dots, y_{nj})$ vector of phenotypes,

$$\begin{cases} \boldsymbol{y}_1 = \boldsymbol{X}_1 \beta_1 + Z_1 \boldsymbol{a}_1 + \varepsilon_1 \\ \boldsymbol{y}_2 = \boldsymbol{X}_2 \beta_2 + Z_2 \boldsymbol{a}_2 + \varepsilon_2 \end{cases},$$

where the genetic effects (BV)

$$\begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{pmatrix} \sim \mathcal{N}\left(0, G \otimes A\right) \quad \text{with } G = \begin{pmatrix} \sigma^2_{\boldsymbol{a_1}} & \sigma_{\boldsymbol{a_{12}}} \\ \sigma_{\boldsymbol{a_{12}}} & \sigma^2_{\boldsymbol{a_2}} \end{pmatrix}$$

and A the $n \times n$ additive genetic relationship matrix associated to the pedigree.
Finally the residuals (envirnonment effect)

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \mathcal{N}\left(0, E \otimes I_n\right) \quad \text{with } E = \begin{pmatrix} \sigma^2_{\boldsymbol{e_1}} & \sigma_{\boldsymbol{e_{12}}} \\ \sigma_{\boldsymbol{e_{12}}} & \sigma^2_{\boldsymbol{e_2}} \end{pmatrix}.$$

and $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ parameters for fixed effect (for example pen/batch or other.)

Estimation in the mixed model

Genetic coefficients to be estimate

- Heritabilities for the two traits that are

$$h_j^2 = \frac{\sigma^2_{\mathbf{a}_j}}{\sigma^2_{\mathbf{a}_j} + \sigma^2_{\mathbf{e}_j}} \quad j = 1, 2;$$

- Genetic and residual correlations

$$\rho_a = \frac{\sigma_{\mathbf{a_{12}}}}{(\sigma_{\mathbf{a_1}}\sigma_{\mathbf{a_2}})^{1/2}}, \quad \rho_e = \frac{\sigma_{\mathbf{e_{12}}}}{(\sigma_{\mathbf{e_1}}\sigma_{\mathbf{e_2}})^{1/2}}\,.$$

Estimation in the mixed model

Genetic coefficients to be estimate

- Heritabilities for the two traits that are

$$h_j^2 = \frac{\sigma^2_{\boldsymbol{a}_j}}{\sigma^2_{\boldsymbol{a}_j} + \sigma^2_{\boldsymbol{e}_j}} \quad j = 1, 2;$$

- Genetic and residual correlations

$$\rho_a = \frac{\sigma_{\boldsymbol{a_{12}}}}{(\sigma_{\boldsymbol{a_1}}\sigma_{\boldsymbol{a_2}})^{1/2}}, \quad \rho_e = \frac{\sigma_{\boldsymbol{e_{12}}}}{(\sigma_{\boldsymbol{e_1}}\sigma_{\boldsymbol{e_2}})^{1/2}}.$$

REML for estimate these coeeficients consists to maximized the restricted log-likelihood

$$\lambda_R(\sigma^2_{a_1}, \sigma^2_{a_2}, \sigma_{a_{12}}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma_{e_{12}}) = -\frac{1}{2}\log[|V|] - \frac{1}{2}\log\left[\left|X^T V^{-1} X\right|\right] - \frac{1}{2} y^T P y$$

where V covariance matrix of $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ and
$P = V^{-1} - V^{-1}X\left(X^T V^{-1} X\right)^{-1} X^T V^{-1}$ with $X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}$

The Sklar's theorem

Theorem of [Sklar(1959)]

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a d-dimensional random vector with c.d.f. $\boldsymbol{F}$ and let $F_1, \ldots F_d$ be the marginal c.d.f. of $\boldsymbol{X}$ assuming continuous. Then it exists a unique function $C : [0,1]^d \to [0,1]$ such that :

$$\boldsymbol{F}(\boldsymbol{x}) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

- The copula C characterizes the dependence structure of vector $\boldsymbol{X}$.
- The copula C can be expressed as follows :

$$C(\boldsymbol{u}) = \boldsymbol{F}\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\}, \qquad \boldsymbol{u} = (u_1, \ldots, u_d) \in [0,1]^d.$$

A. Sklar.
Fonctions de répartition à n dimensions et leurs marges.
Publications de l'Institut de Statistique de l'Université de Paris, 8 : 229–231, 1959.

Contour plots

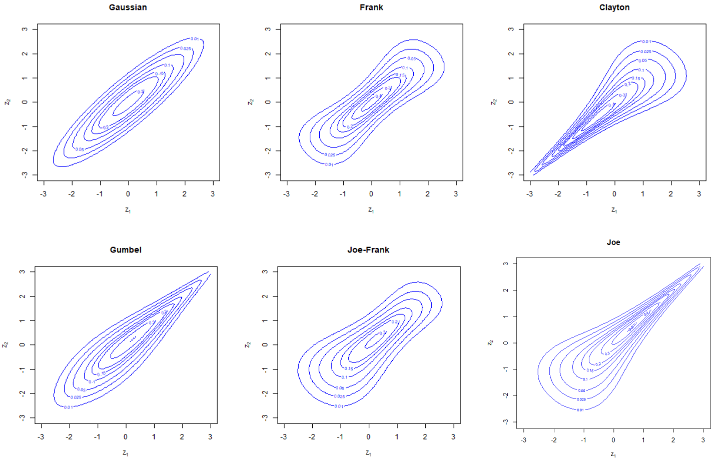


Figure – Contour plot of bivariate distributions with Gaussian margins and Gaussian copula (N), Frank copula (F), Clayton copula (C), Gumbel-Hougaard copula (GH), Joe-Frank copula ($\lambda = 6$) copula (BB8) and Joe copula (J) with Kendall's tau $\tau = 0.7$

differences with bivariate Gaussian ditribution

$\tau = 0.4$	Skewness	kurtosis	lower tail	upper tail
N	0.02	7.98	-	-
F	0.03	8.66	-	-
C	0.69	8.61	0.59	-
GH	0.23	8.51	-	0.48
BB8	0.07	8.69	-	-
J	0.78	8.73	-	0.63

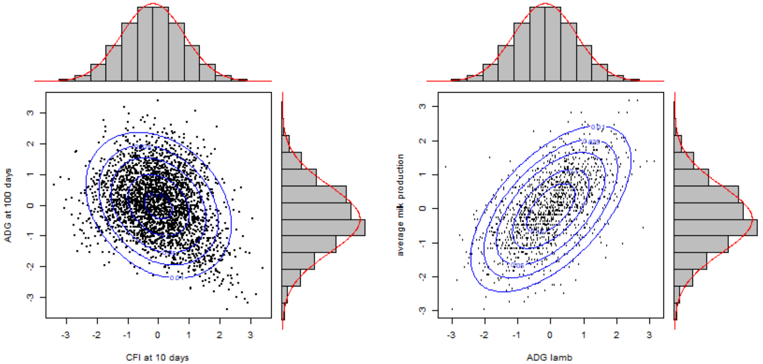
$\tau = 0.7$	Skewness	kurtosis	lower tail	upper tail
N	0.02	7.99	-	-
F	0.07	11.01	-	-
C	2.31	11.58	0.86	-
GH	0.56	9.43	-	0.77
BB8	1.83	11.59	-	-
J	2.35	11.67	-	0.86

Table – Estimated Mardia skewness and Kurtosis based on 1000 Monte Carlo simulations

What about non gaussian model ?

Model robust to margins misspecifications.

▷ what about more general bivariate distribution misspecification ?



(a) Large White pigs

(b) Lambs

Figure – Plot of Gaussian quantiles of the ranked observations over n for the two illustrations. Contour plots of the Gaussian copula with respective Pearson's correlation $\rho = -0.27, 0.57$. Skewness were 0.4, 0.03 and kurtosis were 8.15 and 8.67.

What about non gaussian model ?

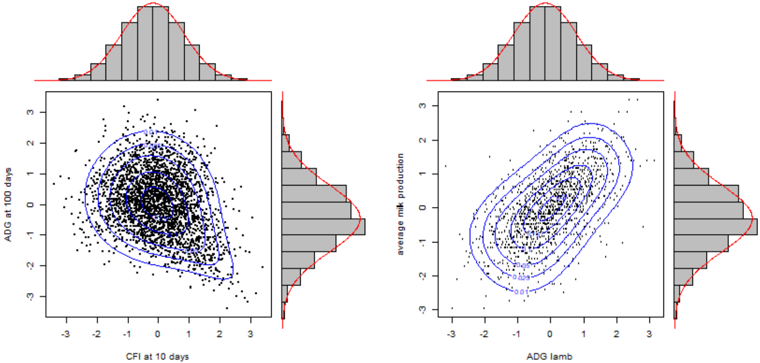
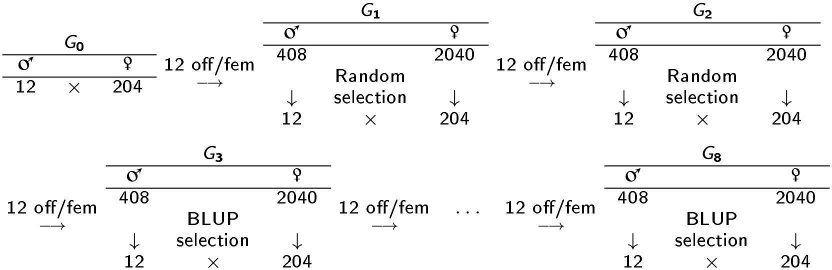


Figure – Plot of Gaussian quantiles of the ranked observations over *n* for the two illustrations. (a)Contour plots of the Rotated BB8 270deg copula with parameters $\delta = -1.4$, $\lambda = -1$ (b) contour plots of the BB8 copula with parameter $(\delta, \lambda) = (6, 0.56)$

Summary

1. Introduction
   - Genetic context
   - Copulas
2. Population sampling
3. Results

Mating schema



G_0
♂ ♀
12 × 204
12 off/fem →

G_1
♂ ♀
408 2040
Random selection
↓ ↓
12 × 204
12 off/fem →

G_2
♂ ♀
408 2040
Random selection
↓ ↓
12 × 204

12 off/fem →

G_3
♂ ♀
408 2040
BLUP selection
↓ ↓
12 × 204
12 off/fem →
. . .
12 off/fem →

G_8
♂ ♀
408 2040
BLUP selection
↓ ↓
12 × 204

Figure – Mating schema for the construction of the data used in the simulations

- 19800 phenotyped animals
- intra-sire selection
- no full-sibs nor half-sibs were matted.

Mandelian sampling

Genetic part :

G_0 $(a_{i,1}, a_{i,2}) \sim_{i.i.d.} \mathcal{N}_2(0, G)$

G_k for $k = 1, \ldots, 8$

$$(a_{i,j})_{j=1,2} = (0.5(a_{S_i,j} + a_{D_i,j}))_{j=1,2} + \mathcal{N}_2(0, G/2) \quad j = 1, 2$$

Residual part :

- For $i = 1, \ldots, n$, $(\varepsilon_{i,1}, \varepsilon_{i,2}) \sim C_\theta(\Phi(x), \Phi(y))$, Φ c.d.f. of the standard Gaussian distribution

Phenotype :

$$\begin{cases} \boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + Z_1\boldsymbol{a}_1 + \boldsymbol{\varepsilon}_1 \\ \boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + Z_2\boldsymbol{a}_2 + \boldsymbol{\varepsilon}_2 \end{cases} .$$

Summary

1. Introduction
   - Genetic context
   - Copulas
2. Population sampling
3. Results

Impact of misspecification on heritability

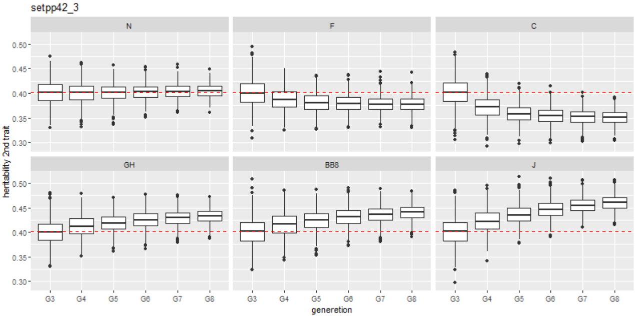


Figure – Boxplot of estimated heritabilities from generation G3 to generation G8 for the set corresponding to $\tau_a = 0.4$ and $\tau_e = 0.7$, $h_1^2 = 0.153$ and $h_2^2 = 0.401$. 1,000 simulations were performed.

Impact of misspecification on genetic correlation

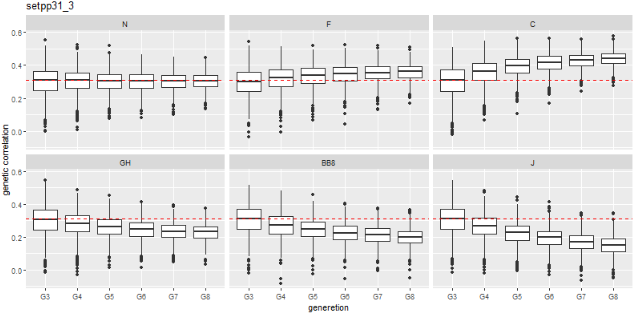


Figure – Boxplot of estimated heritabilities from generation G3 to generation G8 for the set corresponding to $\tau_a = 0.2$ and $\tau_e = 0.7$, $h_1^2 = 0.153$ and $h_2^2 = 0.153$. 1,000 simulations were performed.

Impact of misspecification on residual correlation

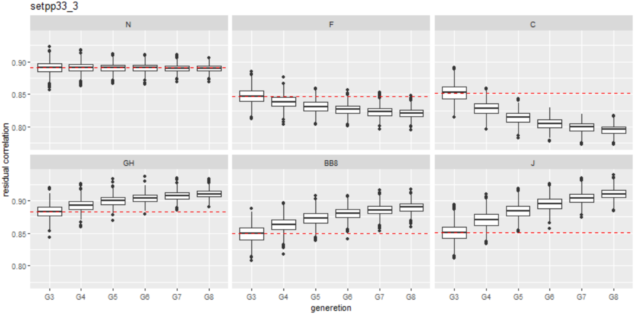


Figure – Boxplot of estimated heritabilities from generation G3 to generation G8 for the set corresponding to $\tau_a = 0.2$ and $\tau_e = 0.7$, $h_1^2 = 0.401$ and $h_2^2 = 0.401$. 1,000 simulations were performed.

Biases for the case $\tau_a = 0.2$ and $\tau_e = 0.7$ based on the whole of the population

(h_1^2/h_2^2)	est	N	F	C	GH	BB8	J
(0.15/0.15)	h_1^2	0.002	-0.001	-0.004	0.009	0.012	0.018
	h_2^2	0.002	0.000	-0.004	0.010	0.013	0.019
	$corr_a$	-0.005	0.048	0.130*	-0.082	-0.111*	-0.161*
	$corr_e$	-0.001	-0.006	-0.016*	0.010*	0.015*	0.023*
(0.15/0.40)	h_1^2	0.001	0.005	0.015	-0.008	-0.011	-0.015
	h_2^2	0.004	-0.023	-0.052*	0.030*	0.039*	0.057*
	$corr_a$	-0.003	0.009	0.059	0.017	0.014	0.031
	$corr_e$	-0.001	-0.012*	-0.031*	0.004	0.008	0.013*
(0.40/0.40)	h_1^2	0.004	-0.000	-0.004	0.014	0.020	0.027
	h_2^2	0.004	0.000	-0.004	0.014	0.020	0.027
	$corr_a$	-0.003	0.060*	0.140*	-0.064*	-0.089*	-0.128*
	$corr_e$	-0.001	-0.025*	-0.056*	0.027*	0.041*	0.060*

Conclusion

1. Without selection :
   ▷ : low biases on the estimated values.

2. With a selection process with low residual correlation
   ▷ : low biases on the estimated heritabilities but some significant differences for genetic and residual correlations (C, GH, J copulas).

3. With a selection process, and for high residual correlation :
   - ▷ significant biases on the estimated heritabilities ; absolute relative differences with the theoretical values up to 15% (J copula) and 13% (C copula)
   - ▷ significant biases on the estimated genetic correlations ; absolute relative differences with the theoretical values up to 27% (J copula) and 24% (C copula)
   - ▷ significant biases on the estimated residual correlations ; absolute relative differences with the theoretical values up to 7% (J copula) and 6% (C copula)