Impact of a non-Gaussian dependence structure on REML estimation of the bivariate genetic animal model

T. Rohmer[1*], A. Ricard[2,3] and I. David[1]

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosane, France; [2] Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas, France; [3] Institut Français du Cheval et de l'Equitation, Pôle développement, Innovation et Recherche, Exmes, France; [*]tom.rohmer@inrare.fr

Abstract

In multiple-trait animal models, variance components are frequently estimated using Restricted Maximum Likelihood method (REML). Such an approach assumes the multivariate normality for the phenotype even if, in practice, this hypothesis is not always realistic. We assessed, using simulation, the impact of a non-Gaussian distribution for the residual term of the mixed model, on the REML estimations. The non-Gaussian distributions were simulated using a copula-based approach. Large populations over 8 generations were simulated using random selection for the 3 first generations and using a truncation selection for the following. Results obtained highlighted the robustness of REML when random selection is performed. On the contrary with a truncation selection process, we observed significant differences with the true parameters, particularly with asymmetric bivariate distributions on the residual part.

Introduction

In a genetic animal context, mixed models are widely used to dissociate the genetic and environmental variance part on the studied phenotypes and perform selection. Extension to the multiple trait model permits to consider the correlations between traits (Meyer (1991), Mrode (2014)).

The principal assumption of this model is the multi normality for the observed phenotype vectors. Since Henderson (1975), the variance-covariance parameters of the model are frequently estimated by REstricted Maximum Likelihood (REML) method, that, under the Gaussian assumption, provides unbiased estimators. However, even if the marginal distributions are Gaussian, the distribution of the multivariate phenotype can be non-Gaussian due to a non-Gaussian dependence structure between the components of the residual vectors of the mixed model.

More precisely let $\boldsymbol{X} = (X_1, X_2)$ be a random vector with cumulative distribution function (c.d.f.) F and marginal c.d.f.s F_1 and F_2 assumed to be continuous. According to Sklar's theorem (Sklar, 1959), it exists a unique function $C: [0,1]^d \to [0,1]$ such that

$$F(\boldsymbol{x}) = C\{F_1(x_1), F_2(x_2)\}, \qquad \boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2.$$

The function C is called the copula of the random vector $\boldsymbol{X}$ and characterizes the dependence structure of the vector $\boldsymbol{X}$. Typical examples of copulas are normal copula (C^N), Clayton copula (C^C), Frank copula (C^F) or Joe copula (C^J), see Nelson (2007). The contour plot of the bivariate distributions with C^C, C^F and C^J copula and Gaussian margins are given in Figure 1. Clayton and Joe distribution are asymmetric (radial sense), and have high dependence in the tail (resp. lower and upper) of the distribution. Frank distribution (with Gaussian margins) is symmetric with slightly more variation in the tails than the Gaussian distribution.

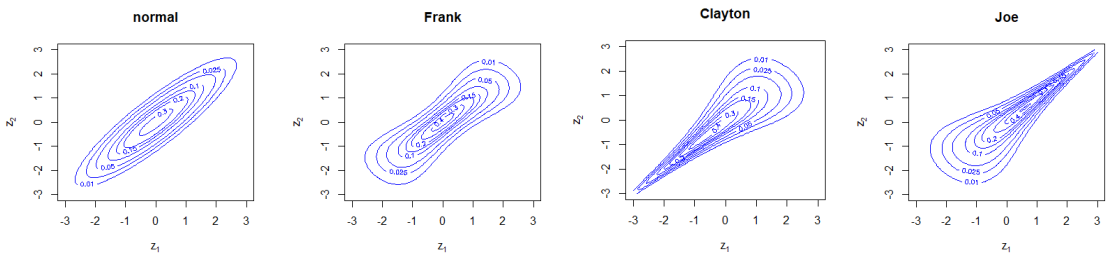

Figure 1. Contour plot of bivariate distributions with Gaussian margins and for copula, the normal copula, Frank's copula, Clayton's copula and Joe's Copula with Kendall's correlation $\tau = 0.7$

The aim of this paper is to assess the effect of a Gaussian misspecification for the residual part on the REML estimations of the heritabilities and correlations (genetic and residual), when the copula is C^C, C^F or C^J but the margins remain Gaussian.

Materials & Methods

We simulated simplified pig breeding schemes, following Gonzales et al. (2020). The total number of animal animals over 8 generations, including founders was 19 800. In the first 3 generations, choice of reproducers was carried out at random. In the following 5, reproducers were chosen by truncation selection (intra-family selection) from a combination of their 2 breeding values (BVs) estimated by BLUP, with equal weight for each trait, that is to say, from the 3rd generation to the 8th, REML estimation of the variance components and BLUP of the BVs were performed using ASReml software (Gilmour et al., 2015). All the progeny had 2 observed phenotypes and were candidates for selection. Selection rate of 2.9% for the males and 10% for the females were considered (204 breeder females and 10 breeder males by generation including founders). Full-sibs and half-sibs were never mated to limit the inbreeding.

For the founders, the BVs $a_{i,1}, a_{i,2}$ for the ith animal were simulated according to a bivariate Gaussian distribution with fixed covariance matrix G. For the other generations, the BVs were simulated through a Mendelian sampling term: $a_{i,j} = 0.5\left(a_{i_f,j} + a_{i_m,j}\right) + M_{ij}, j = 1,2$ where i_f and i_m were respectively the indices of the sire and dam of the ith animal, and (M_{i1}, M_{i2}) followed a bivariate Gaussian distribution with covariance matrix $G/2$. The residual vectors $(\varepsilon_{i,1}, \varepsilon_{i,2})$ were sampled from a bivariate distribution with standard Gaussian margins and C^N, C^C, C^F or C^J copula whose Kendall's correlation was 0.7 (corresponding Pearson's correlations respectively were 0.891, 0.846, 0.852 and 0.850).

We carried out 1 000 Monte Carlo simulations and estimated the heritabilities for the two traits and the genetic and residual correlations. The true genetic variances were 0.18 or 0.67 leading to low and medium heritabilities of 0.153 or 0.401. The true genetic correlations ρ_a were 0.309 or 0.588.

Results

At the end of the 3rd generation (where reproducers were carried out at random), the absolute bias for the heritabilities ranged between 6.77×10^{-6} and 1.43×10^{-3} with SEs between 0.020 and 0.030. The absolute bias for the correlations ranged between 2.4×10^{-7} and 0.010 with SEs between 0.004 and 0.093. No differences were significant (for the t-test at level 5%).

The average estimation biases and SEs for the heritabilities evaluated at the end of the 8th generation are shown in Table 1.

Table 1. Table of bias and SE of estimated heritabilities

True parameters				Estimated heritabilities							
				Trait 1				Trait 2			
h_1^2	h_2^2	ρ_a		C^N	C^F	C^C	C^J	C^N	C^F	C^C	C^J
0.153	0.153	0.309	bias	0.002	-0.001	-0.004	0.018	0.002	0.000	-0.004	0.019
			SE	0.012	0.012	0.011	0.017	0.012	0.012	0.011	0.016
0.153	0.401	0.309	bias	0.001	0.005	0.015	-0.015	0.004	-0.023	-0.052*	0.057*
			SE	0.014	0.015	0.015	0.013	0.013	0.015	0.014	0.015
0.401	0.401	0.309	bias	0.004	-0.000	-0.004	0.027	0.004	0.000	-0.004	0.027
			SE	0.016	0.016	0.016	0.017	0.015	0.016	0.016	0.018
0.153	0.153	0.588	bias	0.002	0.002	0.002	0.010	0.002	0.003	0.003	0.009
			SE	0.012	0.012	0.011	0.015	0.012	0.012	0.012	0.014
0.153	0.401	0.588	bias	0.001	0.004	0.012	-0.006	0.004	-0.024	-0.051*	0.059*
			SE	0.014	0.015	0.015	0.012	0.014	0.015	0.014	0.015
0.401	0.401	0.588	bias	0.004	0.006	0.008	0.020	0.003	0.005	0.009	0.020
			SE	0.016	0.016	0.016	0.017	0.016	0.016	0.016	0.018

Biases and SEs were obtained from 1 000 simulations. Residual copulas were normal(C^N), Frank(C^F), Clayton(C^C) and Joe(C^J). True genetic correlations were $\rho_a \in \{0.309, 0.588\}$.
'*': significant difference between estimated and true heritability for the t-test at level $\alpha = 0.05$

When the residual terms followed a bivariate distribution with C^C or C^J copula, we observed significant difference with the true heritability on the trait with medium heritability as soon as the heritability for the two traits are different. In this case, REML over-estimated the heritability for C^J copula and under-estimated the heritability for C^C copula. For C^F copula, we did not observe any significant difference with the theoretical values. The average estimation biases and SEs for the correlations evaluated at the end of the 8th generation are shown in Table 2.

Table 2. Table of bias and SE of estimated correlations

True parameters				Estimated parameters							
				genetic correlations				residual correlations			
h_1^2	h_2^2	ρ_a		C^N	C^F	C^C	C^J	C^N	C^F	C^C	C^J
0.153	0.153	0.309	bias	-0.005	0.048	0.130*	-0.161*	-0.001	-0.006	-0.016*	0.023*
			SE	0.051	0.051	0.046	0.058	0.003	0.004	0.004	0.007
0.153	0.401	0.309	bias	-0.003	0.009	0.059	0.031	-0.001	-0.012*	-0.031*	0.013*
			SE	0.038	0.038	0.038	0.036	0.004	0.005	0.006	0.006
0.401	0.401	0.309	bias	-0.003	0.060*	0.140*	-0.128*	-0.001	-0.025*	-0.056*	0.060*
			SE	0.029	0.030	0.027	0.032	0.006	0.008	0.007	0.009
0.153	0.153	0.588	bias	-0.003	0.034	0.089*	-0.108*	-0.000	-0.004	-0.010*	0.013*
			SE	0.035	0.034	0.031	0.044	0.003	0.003	0.003	0.005
0.153	0.401	0.588	bias	-0.001	0.003	0.030	0.039	-0.000	-0.009*	-0.023*	0.011*
			SE	0.027	0.028	0.026	0.026	0.004	0.005	0.005	0.005
0.401	0.401	0.588	bias	-0.001	0.049*	0.106*	-0.128*	-0.001	-0.018*	-0.040*	0.052*
			SE	0.021	0.020	0.018	0.027	0.005	0.006	0.006	0.008

Biases and SEs were obtained from 1 000 simulations. Residual copulas were Gaussian(C^N), Frank(C^F), Clayton(C^C) and Joe(C^J). True genetic correlations were $\rho_a \in \{0.309, 0.588\}$. True residual correlations were 0.891, 0.846, 0.852 and 0.850 respectively for C^N, C^F, C^C and C^J.
'*': significant difference between estimated and true correlations for the t-test at level $\alpha = 0.05$

Concerning the residual correlations, as soon as at least one of the heritability was medium (0.401), all the non-normal copulas led to significant differences with the true residual correlation. In theses case, C^F and C^C copula led to an under-estimation of the residual correlations and C^J led to an over-estimation. Concerning the genetic correlation, as soon as the theoretical heritabilities were equal, we observed significant under-estimation with C^J copula and over-estimation with C^C copula. Using C^F copula, the difference with the theoretical values appeared significant (over-estimation) for medium heritability only.

Discussion

To qualify the robustness of the REML facing deviation from normality in the bivariate case, we look for symmetric (C^F), and asymmetric bivariate distributions (C^C and C^J) for the residual term.

With random selection, the robustness of the REML was remarkable despite the non-Gaussian distribution on the residuals; neither the asymmetry nor the heaviness of the distributions affected the REML estimations.

At the end of the 8th generation (truncation selection from the 3rd generation), we observed systematic bias for the variance-covariance parameters using asymmetric distribution: for the estimated heritability when the true heritabilities for the 2 traits were different and for the estimated correlation, when the true heritabilities were the same and medium. Even for the symmetric non-Gaussian distribution (C^F), we observed significant differences with the true parameter for the estimated correlations. Hence, with truncation selection based on EBVs of 2 observed phenotypes, the asymmetry of the residual part but also the heaviness of the upper tail affects the REML estimations. Clearly, the estimation biases are caused in particular, by the selection process carried out in the upper right tail of the distribution. In the case of same heritability for the 2 traits, high (resp. low) dependences in the upper right tail of the residual distribution (using C^J resp. C^C) will lead to an overestimation (resp. underestimation) of the residual correlations and consequently an underestimation (resp. overestimation) of the genetic correlations. In the case of different heritability for the 2 traits, presumably the selection alters the variance of the BV for the trait with medium heritability which causes under or over estimation of the heritability using C^J and C^C.

References

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S., & Thompson, R. (2015). ASReml user guide release 4.1 structural specification. *Hemel hempstead: VSN international ltd*. Available at: https://www.hpc.iastate.edu/sites/default/files/uploads/ASREML/UserGuideStructural.pdf

González-Diéguez, D., Tusell, L., Bouquet, A., Legarra, A., & Vitezica, Z. G. (2020). *G3*, *10*(8) 2829-2841. https://doi.org/10.1534/g3.120.401376

Henderson, C. R. (1975). *Biometrics*, 423-447. https://doi.org/10.2307/2529430

Meyer, K. (1991. *GSE*, *23*(1), 67-83. https://doi.org/10.1186/1297-9686-23-1-67

Mrode, R. A. (2014). Linear models for the prediction of animal breeding values. *Cabi*

Nelsen RB. (2007). An introduction to copulas. *Springer Science & Business Media*.

Sklar, M. (1959). Fonctions de répartition à n-dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, *8*, 229-231