Impact d'une copule non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié pour des populations sous sélection

Tom Rohmer, Anne Ricard, Ingrid David

Impact d'une copule non-Gaussienne dans l'estimation REML du modèle génétique animal bivarié pour des populations sous sélection

Tom Rohmer [1], Anne Ricard [2,3] & Ingrid David [1]

[1] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan.*
[2] *Université Paris-Saclay, INRAE, AgroParisTech, GABI, Jouy-en-Josas*
[3] *Institut Français du Cheval et de l'Equitation, Pôle Développement, Innovation et Recherche, Exmes*
{*tom.rohmer, anne.ricard, ingrid.david*}*@inrae.fr*

Résumé. Dans les modèles multi-caractères utilisés en génétique animal, les composantes de variances sont très souvent estimées par des méthodes de maximum de vraisemblance restreinte (REML) sous l'hypothèse de normalité jointe des caractères, bien qu'en pratique cette hypothèse n'est pas toujours réaliste. Nous avons simulé des populations mimant un schéma de sélection classique rencontré dans les élevages porcins et mesuré l'impact d'une distribution multivariée non Gaussienne sur les résidus (de part une copule non-Gaussienne) sur les estimations réalisées en pratique. Les résultats ont montré que lorsque les reproducteurs sont sélectionnés au hasard, nous n'observons aucun impact significatif sur les estimés, malgré l'hypothèse Gaussienne sous-jacente. Néanmoins, lorsque les reproducteurs sont sélectionnés de façon à améliorer les deux caractères d'intérêt par un processus de troncation basé sur les prédictions BLUP des valeurs génétiques, on observe des différences significatives avec les paramètres théoriques, en particulier avec des distributions bivariées asymétriques sur la partie résiduelle.

Mots-clés. Modèle génétique animal, multi-caractères, copule.

Abstract. In the classical multiple-trait animal models used in a genetic context, variance components are frequently estimated using Restricted Maximum Likelihood method (REML) under the Gaussian assumption for the bivariate traits. Nevertheless, in practice, this hypothesis is not always realistic. We simulated populations mimicking a selection schema encountered in pig farming and assessed the impact of a non-Gaussian multivariate distribution on the residual part generated by a non-Gaussian copula. When the breeders were selected at random, we did not observe any significant impact on the estimated parameters despite the underlying Gaussian assumption. Nevertheless, when the breeders were selected in order to improve the two traits, with a truncation selection process based on BLUP estimation of the breeding values, we observed significant differences with the true parameters, particularly with asymmetric bivariate distributions on the residual part.

Keywords. genetic animal model, multi-trait, copula

1 Introduction

En génétique animal, les modèles mixtes sont fréquemment utilisés afin de dissocier la part de variabilité due à la génétique et la part de variabilité due à l'environnement sur les phénotypes étudiés. Nous citons ici à titre d'exemple Mrode (2014). Ces derniers sont étendus au cas bivarié pour tenir compte des covariances entre caractères lorsque deux phénotypes $y_{i,1}, y_{i,2}$ sont observés pour l'animal $i = 1, \ldots, n$ voir par ex. Meyer (1991). Dans le modèle mixte, la matrice de variance-covariance entre les vecteurs $\boldsymbol{a}_j = (a_{1,j}, \ldots, a_{N,j})$, $N \geq n$, supposés Gaussiens, constitués des valeurs génétiques des animaux du pédigrée pour le caractère $j = 1, 2$, est la suivante :

$$\text{var} \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix} = G \otimes A, \quad \text{avec } G = \begin{pmatrix} \sigma^2_{\boldsymbol{a}_1} & \sigma_{\boldsymbol{a}_{12}} \\ \sigma_{\boldsymbol{a}_{12}} & \sigma^2_{\boldsymbol{a}_2} \end{pmatrix},$$

avec A la matrice des coefficients de parenté associée au pédigrée des animaux et '$\otimes$' désigne le produit de Kronecker.

Les composantes de variances sont généralement estimées par des méthodes de maximum de vraisemblance restreinte (REML) sous l'hypothèse de normalité jointe des phénotypes bien qu'en pratique cette hypothèse n'est pas toujours réaliste, en particulier de part des résidus non nécessairement Gaussiens. Et même si chacune des marges a une distribution Gaussienne, cela ne garantit pas la normalité des vecteurs bivariés.

Plus précisément, considérons un vecteur aléatoire $\boldsymbol{\mathcal{X}} = (\mathcal{X}_1, \mathcal{X}_2)$ de fonction de répartition (f.d.r.) $\boldsymbol{F}$, et de fonctions de répartitions marginales F_1, F_2 supposées continues. Le théorème de Sklar (1959) permet d'affirmer qu'il existe une unique fonction C : $[0, 1]^d \to [0, 1]$ telle que :

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

Cette fonction C dite copule caractérise la structure de dépendance du vecteur aléatoire $\boldsymbol{\mathcal{X}}$. En particulier une structure de dépendance non-Gaussienne entraîne une distribution bivariée non-gaussienne. Les contour plots des distributions considérées dans cette étude pour la partie résiduelle du modèle sont représentés dans la Figure 1.

Le but de ce papier est de mesurer l'impact d'une copule non-Gaussienne dans la partie résiduelle du modèle mixte, sur les estimations REML des composantes de variance (supposant le modèle Gaussien), lorsque les marges sont Gaussiennes.

2 Simulations

Les populations d'animaux apparentés ont été générées en suivant Gonzales et al. (2020), mimant un schéma de sélection pouvant être rencontré dans un élevage porcin. Les fondateurs (G_0) consistaient en 204 femelles et 12 mâles non apparentés. Dans les générations

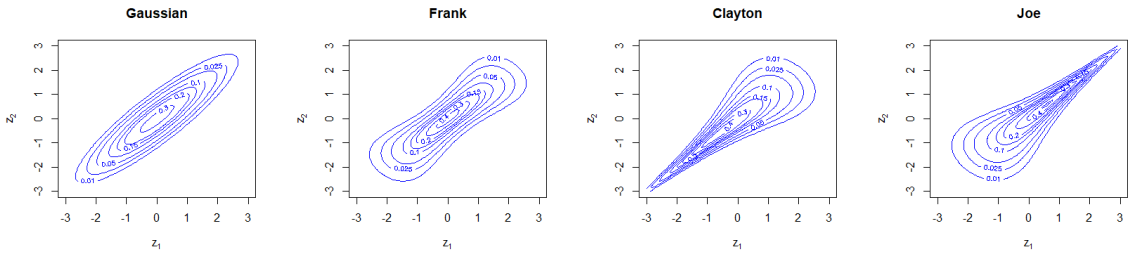


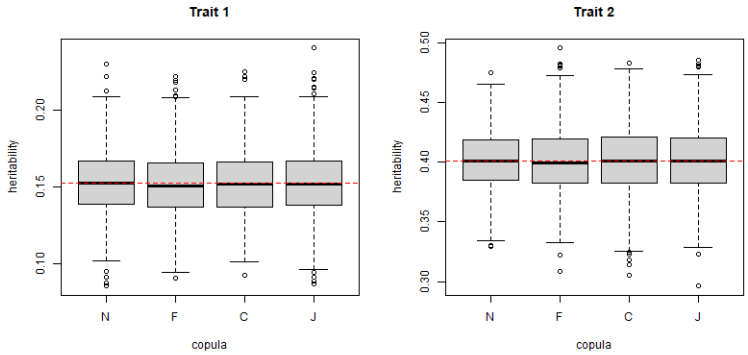
FIGURE 1 – Contours plot des distributions bivariées avec marge Gaussienne standard et copule normale, Frank, Clayton et Joe avec une corrélation de Kendall $\tau = 0.7$

suivantes (G_1–G_8), chaque mâle reproducteur a été accouplé à 17 femelles pour produire 10 femelles et 2 mâles phénotypés par femelle, soit $n = 19800$ individus ($N = n$) à l'issu de 8 générations. Dans les 3 premières générations, les mâles reproducteurs ont été choisi au hasard. Dans les 5 générations suivantes, pour chaque animal, les valeurs génétiques pour les deux caractères ont été prédites par BLUP (voir Henderson, 1975). Au sein de chaque famille de même père, les individus dont la somme de leurs BLUP étaient les plus fortes ont été sélectionné (sélection par troncation). Dans chaque génération, un ratio de sélection pour la reproduction de 2.9% de mâles et 10% des femelles a été considéré. Les accouplements ont été réalisé au hasard de telle sorte que les pleins frères et demi-frères ne soient jamais accouplés afin de réduire la consanguinité.

Les valeurs génétiques des fondateurs ont été simulées indépendamment selon une Gaussienne bivariée de matrice de covariance G. Pour les générations G_1–G_8, les valeurs génétiques $(a_{i,1}, a_{i,2})$, $i = 217, \ldots, n$ ont été simulé selon une loi normale bivariée respectant un schéma Mendélien : pour $j = 1, 2$, $a_{i,j} = 0.5(a_{i_p,j} + a_{i_m,j}) + M_{ij}$, où i_p et i_m sont les indices du père et de la mère de l'animal i et les vecteurs (M_{i1}, M_{i2}) sont simulés selon une Gaussienne multivariée de matrice de variance-covariance $G/2$.

Les vecteurs des résidus du modèle mixte ont été échantillonnés indépendamment selon une distribution bivariée avec des marges Gaussiennes de variance σ_j^2 fixée à 1 et de copule normal(F), Frank(F), Clayton(C) et Joe(J) avec corrélation de Kendall fixée à 0.7, voir Nelson 2007. La corrélation génétique théorique a été fixée à 0.59 (Kendall à 0.4). Les variances génétiques théoriques étaient de 0.18 ou 0.67 amenant des héritabilité théoriques $h_j^2 = \sigma_{a_j}^2/(\sigma_{a_j}^2 + \sigma_{e_j}^2)$, $j = 1, 2$ de 0.153 ou 0.401 pour les caractères simulés. De la génération G_3 à la génération G_8, nous avons estimé l'héritabilité des caractères ainsi que les corrélations génétiques et résiduelles à partir des estimations REML des composantes de variance sous l'hypothèse de normalité jointe des phénotypes.

a. 3ème génération (sélection aléatoire)



b. 8ème génération (sélection par troncation)

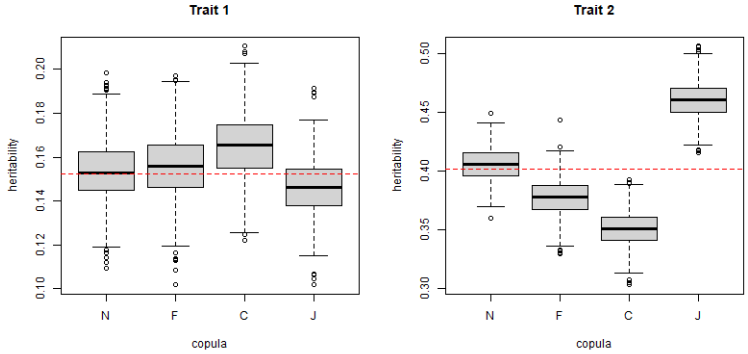


FIGURE 2 – Boxplot des héritabilités estimées des deux caractères (a.) sans sélection (b.) avec selection, pour les copules Normale (N), Frank (F), Clayton (C) et Joe(J) avec une dépendance résiduelle forte (corrélation de Kendall à 0.7) et une dépendance génétique modérée (corrélation de Kendall à 0.4). Les héritabilités théoriques (lignes rouges) des deux caractères sont respectivement 0.153 et 0.401.

3 Résultats

Les résultats ont montré qu'avec une sélection aléatoire des reproducteurs, quel que soit la copule sur les résidus, les biais pour les héritabilités estimées (Figure 2.a) ainsi que les biais pour les corrélations estimées (Table 1.a) étaient extrêmement faibles, en valeur absolue entre 6.77×10^{-6} et 1.43×10^{-3} pour les héritabilités (SE entre 0.020 et 0.030) et entre 2.4×10^{-7} et 0.010 pour les corrélations (SE entre 0.004 et 0.093). Aucunes différences entre les valeurs théoriques et estimées n'apparaissent significatives pour un t-test au niveau $\alpha = 5\%$.

À contrario, dans le cas d'une population soumise à une stratégie de sélection non aléatoire (Figure 2.b et Table 1.b), nous avons observé des différences significatives entre les paramètres estimés et ceux simulés lorsque les résidus n'étaient pas gaussiens et les

corrélations résiduelles suffisamment importantes. Notamment les différences avec les valeurs théoriques pour les héritabilités apparaissaient significatives pour le caractère ayant l'héritabilité la plus élevé, lorsque les héritabilités théoriques étaient différentes pour les deux caractères et pour les copules asymétriques C et J. Le biais estimé était de -0.051 pour la copule C (SE de 0.014) marquant une sous-estimation et de 0.059 pour la copule J (SE de 0.015) marquant une surestimation. Le biais estimé pour la copule F apparaît plus modéré, -0.024 (SE de 0.015).

Les différences avec les valeurs théoriques pour les corrélations étaient les plus fortes lorsque les héritabilités des deux caractères étaient identiques, et modérées ($h_j^2 = 0.401$, $j = 1, 2$). Nous remarquons dans ce cas une sous-estimation de la corrélation génétique pour J (biais estimé de -0.128, SE à 0.027) et une surestimation de la corrélation génétique pour C et G (bais resp. de 0.106 et 0.049, SE resp. de 0.018 et 0.020).

Ces résultats indiquent que le processus de sélection effectué dans les queues droites des distributions, améliorant simultanément les deux caractères, va entraîner des surestimations des corrélations résiduelles, lorsque la distribution jointe, portée par les résidus, a des corrélations dans la queue droite de la distribution plus fortes que le cas Gaussien (c'est le cas par exemple de la copule J) et des sous-estimation dans le cas contraire (par exemple C et F). De là, une sur(resp. sous)-estimation de la corrélation résiduelle va entraîner une sous(resp. sur)-estimation de la partie génétique. Vraisemblablement, dans le cas de différentes héritabilités pour les deux caractères, le processus de sélection mis en œuvre en présence d'une copule non-Gaussienne (en particulier asymétrique et à queue lourde) va altérer les valeurs génétiques du caractère avec la plus grande héritabilité et ainsi causer des sur/sous estimations des héritabilités.

Bibliographie

González-Diéguez D., Tusell L., Bouquet A., Legarra A., and Vitezica Z. G. (2020), Purebred and crossbred genomic evaluation and mate allocation strategies to exploit dominance in pig crossbreeding schemes, *G3 : Genes, Genomes, Genetics*, vol. 10, no. 8, pp. 2829–2841.

Henderson C. R. (1975), Best linear unbiased estimation and prediction under a selection model, *Biometrics*, pp. 423–447.

Meyer K. (1991) Estimating variances and covariances for multivariate animal models by restricted maximum likelihood, *Genetics Selection Evolution*, vol. 23, no. 1, pp. 1–17.

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

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

Sklar M. (1959) Fonctions de répartition an dimensions et leurs marges, *Publ. inst. statist. univ. Paris*.

TABLE 1 – Biais et standard error (SE) des corrélations génétiques et résiduelles estimées à la génération G_3 et G_8.

a. 3ème génération (sélection aléatoire)

	hérit. théorique			Corrélations estimées							
				Corrélation génétique				Corrélation résiduelle			
	h_1^2	h_2^2		N	F	C	J	N	F	C	J
	0.153	0.153	biais	-0.004	-0.007	-0.004	-0.003	-0.000	0.001	0.000	0.001
			SE	0.062	0.063	0.067	0.063	0.004	0.006	0.006	0.006
G_3	0.153	0.401	biais	-0.001	-0.003	-0.001	-0.001	0.000	0.001	0.000	0.001
			SE	0.049	0.051	0.054	0.051	0.007	0.008	0.009	0.009
	0.401	0.401	biais	-0.001	-0.003	-0.001	-0.002	0.000	0.001	0.000	0.001
			SE	0.036	0.037	0.040	0.039	0.008	0.010	0.011	0.011

b. 8ème génération (sélection par troncation)

	hérit. théorique			Corrélations estimés							
	h_1^2	h_2^2		N	F	C	J	N	F	C	J
	0.153	0.153	biais	-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
G_8	0.153	0.401	biais	-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	biais	-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

Les biais et les SE ont été obtenus à partir de 1000 simulations. Les copules des résidus du modèle mixte étaient normale(N), Frank(F), Clayton(Cl) and Joe(J). La corrélation génétique théorique était de 0.588, correspondant à un tau de Kendall de 0.4. Les corrélations résiduelles pour N, F, C and J étaient respectivement de 0.891, 0.846, 0.852 et 0.850 correspondant à un tau de Kendall de 0.7. '*' : Différence significative entre corrélation estimée et théorique pour le t-test au level $\alpha = 0.05$.