Copula Integration for Genetic Parameter Estimation in Bivariate Linear Mixed Models

Victoria Bruning, Estelle Kuhn, Tom Rohmer

Copula Integration for Genetic Parameter Estimation in Bivariate Linear Mixed Models

Victoria Bruning*, Estelle Kuhn**, Tom Rohmer*

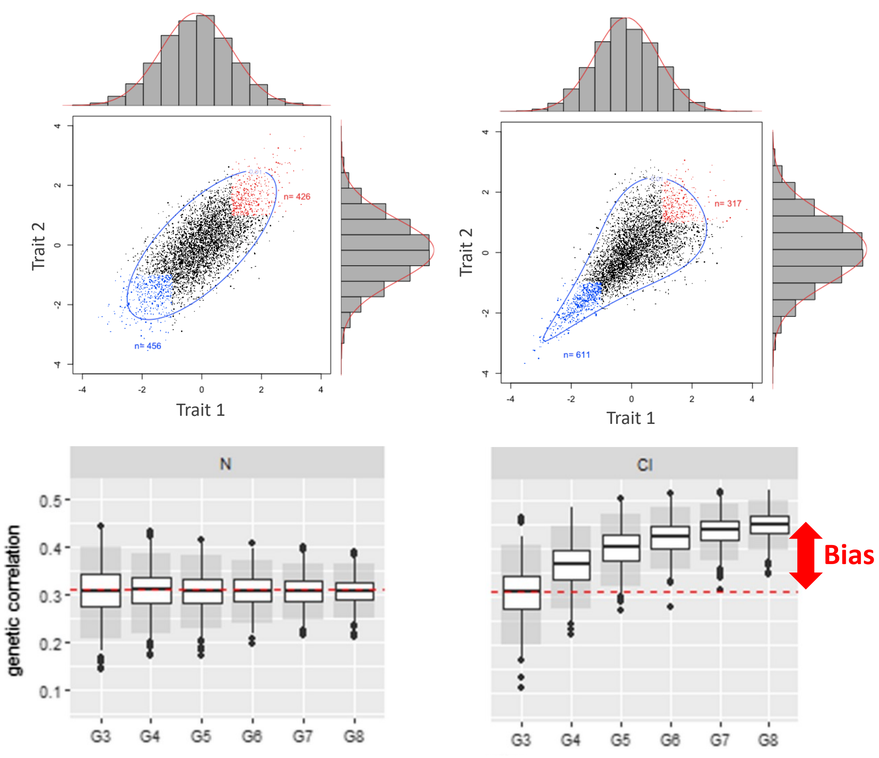
** GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan, France*
*** INRAE, MaIAGE, Domaine de Vilvert 78352 Jouy-en-Josas Cedex FRANCE*

In the bivariate genetic animal model, the estimation methods assume that the traits are gaussian, but in practice, the joint distribution of the two traits can be non-Gaussian. This induces a bias in the genetic parameters estimation when the reproducers are non-randomly selected. The aim of this project was to integrate copula functions, characterising the joint distribution, in the estimation of genetic parameters, and so, reduce this bias.

The problem

Induction of a bias in the REML estimation of the genetic parameters when the residual joint distribution is non-gaussian and the reproducers are non-randomly selected from generation G3.

Gaussian joint distribution **Clayton-type distribution**



Rohmer at al., 2022

The goal

Estimation **based on the joint Gaussian** distribution of the data

Multi-normality assumption

Estimation **adapted to any** distribution of the data

Copula assumption

The genetic animal model

Each trait is determined by an environmental and genetic factor. The breeding value a is the average additive effects of genes an individual receives from both parents. Thus, it is the primary component that can be selected and is the main focus.

Bivariate Mixed Model

$$\begin{cases} \mathbf{y}_1 = X_1\beta_1 + Z\mathbf{a}_1 + \mathbf{e}_1 \\ \mathbf{y}_2 = X_2\beta_2 + Z\mathbf{a}_2 + \mathbf{e}_2 \end{cases}$$

Traits — Environment component — Genetic component — Residual term

$$\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2) \sim \mathcal{N}(0, \Sigma) \quad \text{with} \quad \Sigma = \mathbf{G} \otimes \mathbf{A}$$

Kinship matrix

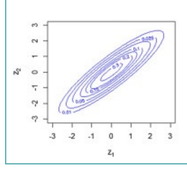
$$\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2) \sim \mathcal{N}(0, \mathbf{R}) \quad \text{with} \quad \mathbf{R} = \mathbf{E} \otimes \mathbf{I}_n$$

Goal : $\mathbf{e}$ is defined by a copula function and Gaussian marginals

$$\mathbf{G} = \begin{pmatrix} \sigma^2_{a_1} & \sigma_{a_{12}} \\ \sigma_{a_{12}} & \sigma^2_{a_2} \end{pmatrix} \quad \mathbf{E} = \begin{pmatrix} \sigma^2_{e_1} & \sigma_{e_{12}} \\ \sigma_{e_{12}} & \sigma^2_{e_2} \end{pmatrix} \Bigg\} \hat{\boldsymbol{\theta}}$$

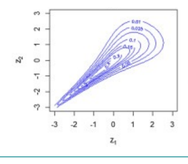
The tool : Copula functions

Copulas are functions characterize the joint distribution of the data



Gaussian copula

$$C_\rho(u, v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$$



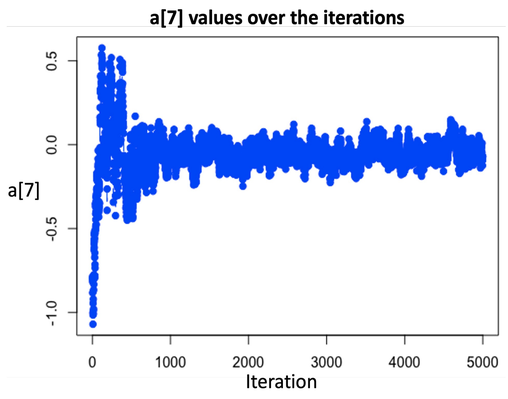
Clayton copula

$$C_\rho(u, v) = \left[\max\left((u^{-\rho} + v^{-\rho} - 1), 0\right)\right]^{-1/\rho}$$

The estimation algorithm : Stochastic Gradient with Copulas

The stochastic gradient is an optimization method that updates the parameters using the gradient of the likelihood computed from a simulated conditional distribution

1- **Simulate** $\mathbf{a}^{(k)} \sim f(\mathbf{a} \mid \mathbf{y})$, the conditional distribution by Gibbs sampling



2- **Maximize** $f(\mathbf{y}; \theta)$, the likelihood function

$$\log f(\mathbf{y}; \theta) = \log \int f(\mathbf{y} \mid \mathbf{a}; \theta) f(\mathbf{a}; \theta)\, d\mathbf{a}$$

With Gaussian assumption

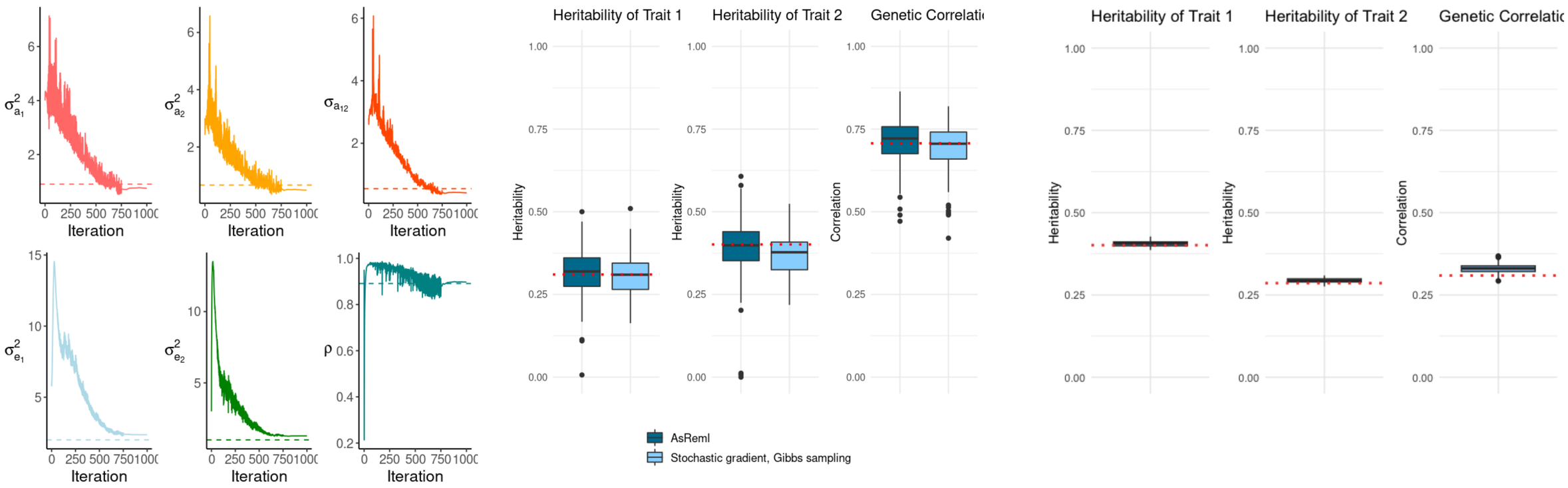
$$f(\mathbf{y} \mid \mathbf{a}) = \frac{1}{\sqrt{(2\pi)^{2n} \cdot \det(\boldsymbol{R})}} \exp\left(-\frac{1}{2}(\mathbf{y} - \mathbf{a})^\top \boldsymbol{R}^{-1} (\mathbf{y} - \mathbf{a})\right)$$

Integrating a Copula function

$$f(\mathbf{y} \mid \mathbf{a}) = \sum_i c\left(\Phi_1\left(\mathbf{y}_{i1} \mid \mathbf{a}_{i1}\right), \Phi_2\left(\mathbf{y}_{i2} \mid \mathbf{a}_{i2}\right)\right) \times \prod_{j=1}^{2} f_j\left(\mathbf{y}_{ij} \mid \mathbf{a}_{ij}\right)$$

$$\theta^{(k+1)} = \theta^{(k)} + \text{learning rate} \times \nabla_\theta \log f(\mathbf{y}, \mathbf{a}^{(k)}; \theta^{(k)})$$

Results & Discussion



Estimation integrating a Gaussian copula

The simulated data has a Gaussian dependence structure with 720 animals, and the estimation included a Gaussian copula. A- Convergence graph for a single run, B- Boxplot of the estimates obtained for 100 runs by ASReml software and by the developed method. The dotted lines are the true values.

Estimation integrating a Clayton copula

The simulated data has a Clayton-type dependence structure with 7560 animals, and the estimation included a Clayton copula. Boxplot of the estimates obtained for 100 runs by the developed method.

Conclusion

The linear mixed model integrating copulas enables the estimation of genetic parameters when the joint distribution is either Gaussian or defined through another dependence structure, here a Clayton-type copula. The convergence of the stochastic gradient algorithm is observed toward a critical point, and expected toward the MLE which may be closed to the true values for large sample size.

Perspectives

The learning rate will be enriched by a Fisher scoring matrix to standardize the gradient descent direction. This algorithm will be applied on real data. In future work, this could be extended to more than two traits and on data with non-Gaussian marginal distributions.