Inference in Copula multi-trait animal model to Improve the genetic selection

Séminaire Maths-bio-santé

from a joint work with

Tom Rohmer, Anne Ricard & Ingrid David

and

Victoria Bruning, Estelle Kuhn

07 juin 2024

Plan

Introduction

Copula miss-specification in the inference model
from coll. with Anne Ricard & Ingrid David

Inference in copula genetic model
from coll. with V. Bruning & E. Kuhn

Plan

Introduction

Copula miss-specification in the inference model
from coll. with Anne Ricard & Ingrid David

Inference in copula genetic model
from coll. with V. Bruning & E. Kuhn

Multivariate observations

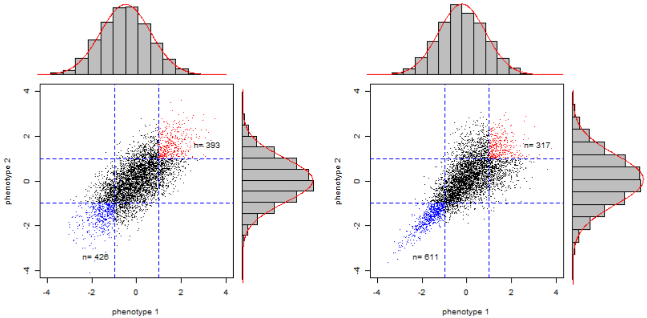


Figure: Simulation of $n = 5000$ bivariate observations whose the univariate distributions are Gaussian and with Spearman's correlation $\rho = 0.7$. (Left) Multivariate Gaussian distribution. (Right) Clayton-type distribution

Copulas

Theorem of Sklar 1959, bivariate case

Let $\boldsymbol{Y} = (Y_1, Y_2)$ be a 2-dimensional random vector with c.d.f. $\boldsymbol{F}(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2)$ and let $F_1(y_1) = P(Y_1 \leq y_1)$, $F_2(y_2) = P(Y_2 \leq y_2)$ be the marginal c.d.f. of $\boldsymbol{Y}$ assuming continuous. Then it exists a unique function $C : [0,1]^2 \to [0,1]$ such that:

$$\boldsymbol{F}(\boldsymbol{y}) = C\{F_1(y_1), F_2(y_2)\}, \qquad \boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2.$$

- The copula C characterizes the dependence structure of vector $\boldsymbol{Y}$.

Copulas

Theorem of Sklar 1959, bivariate case

Let $\boldsymbol{Y} = (Y_1, Y_2)$ be a 2-dimensional random vector with c.d.f. $\boldsymbol{F}(y_1, y_2) = P(Y_1 \leq y_1, Y_2 \leq y_2)$ and let $F_1(y_1) = P(Y_1 \leq y_1)$, $F_2(y_2) = P(Y_2 \leq y_2)$ be the marginal c.d.f. of $\boldsymbol{Y}$ assuming continuous. Then it exists a unique function $C : [0,1]^2 \to [0,1]$ such that:

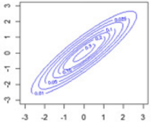
$$\boldsymbol{F}(\boldsymbol{y}) = C\{F_1(y_1), F_2(y_2)\}, \qquad \boldsymbol{y} = (y_1, y_2) \in \mathbb{R}^2.$$

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

$$\begin{aligned} C(u_1, u_2) &= \boldsymbol{F}\{F_1^{-1}(u_1), F_2^{-1}(u_2)\} \\ &= P(F_1(Y_1) \leq u_1,\ F_2(Y_2) \leq u_2) \end{aligned}$$

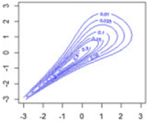
Some copulas

normal



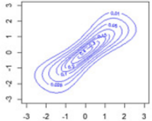
$$C_\rho(u,v) = \Phi_\rho\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$$

Clayton



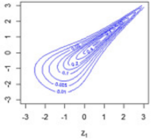
$$C_\rho(u,v) = \left[\max\left((u^{-\rho} + v^{-\rho} - 1), 0\right)\right]^{-1/\rho}$$

Frank



$$C_\rho(u,v) = -\frac{1}{\rho}\ln\left[1 + \frac{(\exp(-\rho u) - 1)(\exp(-\rho v) - 1)}{\exp(-\rho) - 1}\right]$$

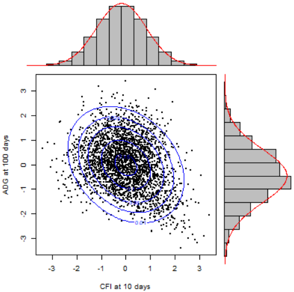
Joe

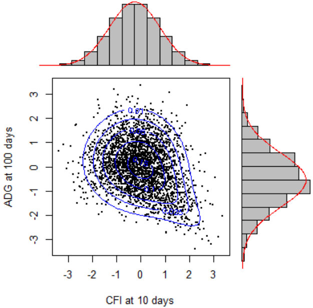


$$C_\rho(u,v) = 1 - \left[(1 - (1-u)^\rho)^{1/\rho} + (1 - (1-v)^\rho)^{1/\rho} - 1\right]^\rho$$

Example, Large-White dataset, pseudo-observations

- Even if the marginals are Gaussian, the bivariate distribution may be non-Gaussian.





- The bivariate normality is questionable
- What about the REML estimations of the bivariate animal model, which assume the bivariate normality?

Classical multivariate inference model

- Multivariate Gaussian model
- Multinomial model / multivariate threshold model
- Multivariate Poisson log-normal

Classical multivariate inference model

- Multivariate Gaussian model
- Multinomial model / multivariate threshold model
- Multivariate Poisson log-normal

- Copula makes possible to define an <u>infinit</u>y of multivariate inference model.
  - Nevertheless the estimation procedures can be difficult (optimization problems) or time-consuming.

Classical multivariate inference model

- Multivariate Gaussian model
- Multinomial model / multivariate threshold model
- Multivariate Poisson log-normal

- Copula makes possible to define an infinity of multivariate inference model.
  - Nevertheless the estimation procedures can be difficult (optimization problems) or time-consuming.

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan & Tom Rohmer
Fast inference in copula models with categorical explanatory variables using one-step procedures
Article in progress, 2023

Victoria Bruning, Estelle Kuhn, Tom Rohmer
Inference in copula genetic models
Article in progress, 2024

Plan

Introduction

Copula miss-specification in the inference model
from coll. with Anne Ricard & Ingrid David

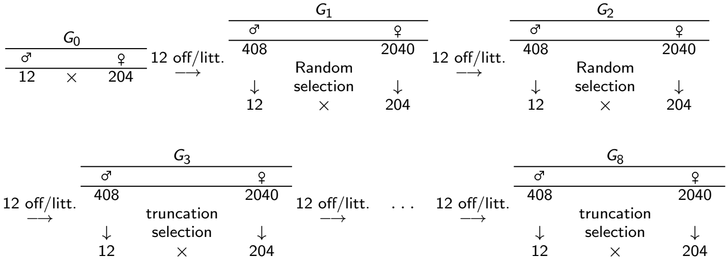
Inference in copula genetic model
from coll. with V. Bruning & E. Kuhn

Articles

Rohmer, T., Ricard, A, David, I
Copula miss-specification in REML multivariate genetic animal model estimation,
Genetics Selection Evolution, May 2022

Rohmer, T.
An R Markdown for phenotypes simulation, multitrait and Random Regression models with Asreml
http://genoweb.toulouse.inra.fr/~trohmer/dyna_phen_RR_3.html

Simulation of populations undergoing selection



G_0

♂ ♀

12 × 204

12 off/litt. ⟶

G_1

♂ ♀

408 2040

Random selection

↓ ↓

12 × 204

12 off/litt. ⟶

G_2

♂ ♀

408 2040

Random selection

↓ ↓

12 × 204

12 off/litt. ⟶

G_3

♂ ♀

408 2040

truncation selection

↓ ↓

12 × 204

12 off/litt. ⟶ . . . 12 off/litt. ⟶

G_8

♂ ♀

408 2040

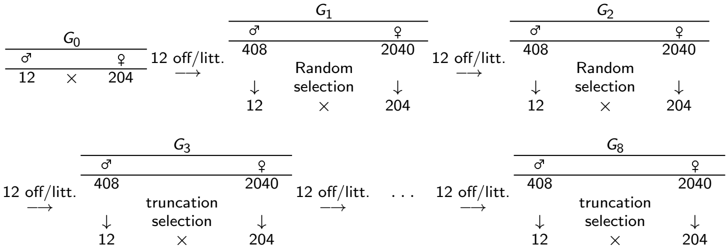
truncation selection

↓ ↓

12 × 204

- unrelated animals in G_0
- each female produced 12 offspring: 2 males and 10 females

Population



G_0

♂ ♀

12 × 204

12 off/litt. ⟶

G_1

♂ ♀

408 2040

Random selection

↓ ↓

12 × 204

12 off/litt. ⟶

G_2

♂ ♀

408 2040

Random selection

↓ ↓

12 × 204

12 off/litt. ⟶

G_3

♂ ♀

408 2040

truncation selection

↓ ↓

12 × 204

12 off/litt. ⟶ . . . 12 off/litt. ⟶

G_8

♂ ♀

408 2040

truncation selection

↓ ↓

12 × 204

Selection:

- $G_1 - G_3$ the reproducers were chosen at random
- $G_4 - G_8$ selection from a combination of their EBV
- Full/half siblings were not mated
- selection rate: 2.9% for the males and 10% for the females

Phenotypes simulation → Shiny simulation link

The phenotype vectors $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n,j})$, $j = 1, 2$ were obtained following the bivariate animal model:

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

$\boldsymbol{X}_j$ the design matrices for the fixed effects and β_j associated parameter.

Simulated distribution:

- $a_{i,j} = 0.5(a_{i_S,j} + a_{i_D,j}) + M_{i,j}$,
  where $a_{i_S,j}$ and $a_{i_D,j}$ are the BVs of the sire and dam and $M_{i,j}$ are the Mendelian sampling terms, with
  $$(M_{i,1}, M_{i,2}) \sim \mathcal{N}(0, G/2).$$
  The distribution of $(\boldsymbol{a}_1, \boldsymbol{a}_2)$ is assumed to be $\mathcal{N}(0, G \otimes A)$ with A the relationship genetic matrix.

Phenotypes simulation → Shiny simulation link

The phenotype vectors $\boldsymbol{y}_j = (y_{1,j}, \dots, y_{n,j})$, $j = 1, 2$ were obtained following the bivariate animal model:

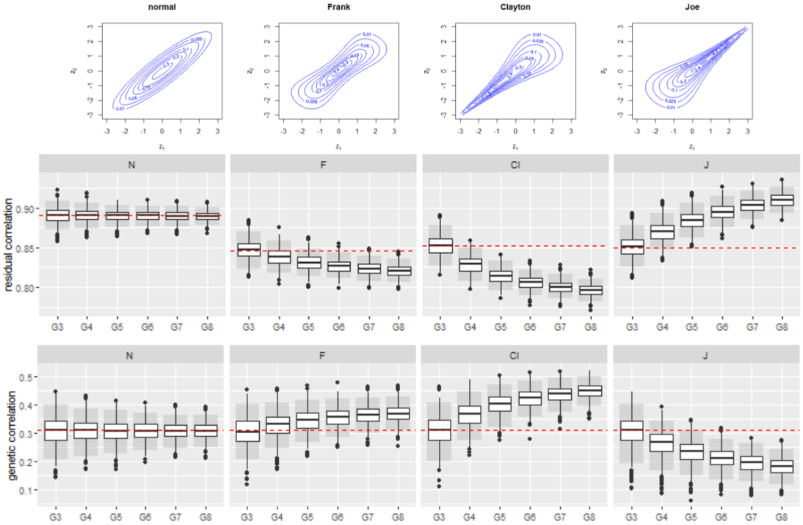
$$\begin{cases} \boldsymbol{y}_1 = \boldsymbol{X}_1 \beta_1 + \boldsymbol{a}_1 + \varepsilon_1 \\ \boldsymbol{y}_2 = \boldsymbol{X}_2 \beta_2 + \boldsymbol{a}_2 + \varepsilon_2. \end{cases}$$

$\boldsymbol{X}_j$ the design matrices for the fixed effects and β_j associated parameter.

Simulated distribution:

- $a_{i,j} = 0.5(a_{i_S,j} + a_{i_D,j}) + M_{i,j}$,
  where $a_{i_S,j}$ and $a_{i_D,j}$ are the BVs of the sire and dam and $M_{i,j}$ are the Mendelian sampling terms, with
  $$(M_{i,1}, M_{i,2}) \sim \mathcal{N}(0, G/2).$$
  The distribution of $(\boldsymbol{a}_1, \boldsymbol{a}_2)$ is assumed to be $\mathcal{N}(0, G \otimes A)$ with A the relationship genetic matrix.
- $(\varepsilon_{i,1}, \varepsilon_{i,2})$, $i = 1, \dots, n$, have standard Gaussian margins and copula C.

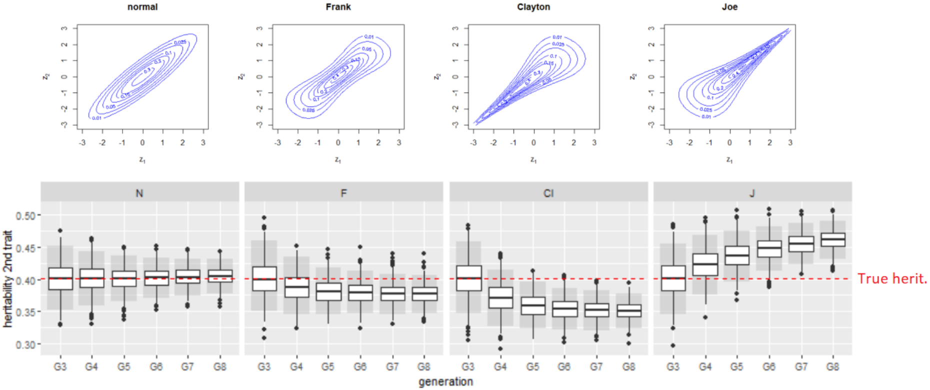
1000 runs, Estim. correlations, $h_1^2 = h_2^2 = 0.40, \tau_e = 0.7$



True res. cor.

True gen. cor.

1000 runs, Estim. heritability $h_1^2 = 0.15, h_2^2 = 0.40$



Results

1. With Random selection: no impact of the copula
2. With truncation selection;
    - For balanced heritabilities:
        - Significant impact on correlations;
        - ▷ very low biases for heritability
    - For unbalanced heritabilities:
        - Significant impact on the estim. heritabilities for the trait with moderate heritability
        - Significant impact on residual correlations;
        - ▷ moderate biases (but non-significant) on genetic correlations;
        - ▷ no impact on the estim. heritabilities for the trait with low heritability

Plan

Introduction

Copula miss-specification in the inference model
from coll. with Anne Ricard & Ingrid David

Inference in copula genetic model
from coll. with V. Bruning & E. Kuhn

Inference in copula genetic model

Let define the genetic covariance matrix

$$G = \begin{pmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2} \\ \sigma_{a_1 a_2} & \sigma_{a_2}^2 \end{pmatrix}$$

All of the animal of the pedigree are phenotyped that is to say $N = n$. Consider for $j = 1, 2$, $i = 1, \dots, n$

$$\begin{aligned} (a_{1,1}, \dots, a_{N,1}, a_{1,2}, \dots, a_{N,2}) &\sim \mathcal{N}_{2N}\left(0, G \otimes A\right) && (1) \\ Y_{ij} | a_{i,j} &\sim \mathcal{N}(a_{i,j}, \sigma_{e_j}^2) && (2) \\ (Y_{i,1}, Y_{i,2}) | (a_{i,1}, a_{i,2}) & \text{ has copula } C_\theta && (3) \end{aligned}$$

Where A the kinship matrix. The complete log-likelihood of y is

$$\log \mathcal{L}(y) = \log \int \mathcal{L}(y|a)\mathcal{L}(a)da$$

Inference in copula genetic model

We have

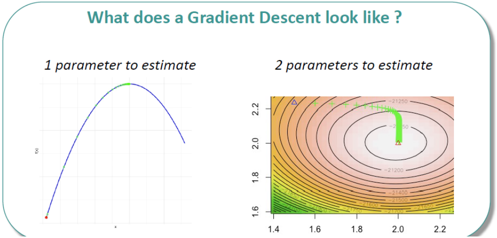
$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha},\theta;\boldsymbol{y}|\boldsymbol{a}) &= \prod_{i=1}^{n}\mathcal{L}(\boldsymbol{\alpha},\theta;(y_{i,1},y_{i,2})|(a_{i,1},a_{i,2}))\\ &= \prod_{i=1}^{n} c_{\theta}\left(\Phi_1(0,\sigma^2_{\mathrm{e}_1};y_{i1}|a_{i1}),\Phi_2(0,\sigma^2_{\mathrm{e}_2};y_{i2}|a_{i2})\right)\times\prod_{i=1}^{n}\prod_{j=1}^{2}\mathcal{L}_j(\sigma^2_{\mathrm{e}_j};y_{ij}|a_{i,j}).\\ \mathcal{L}(G;\boldsymbol{a}) &= \frac{1}{(2\pi)^n(det(G\otimes A))^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{a}^T G^{-1}\otimes A^{-1}\boldsymbol{a}\right),\end{aligned}$$

c_θ is the copula density given by

$$c_\theta\left(u_1,u_2\right)=\frac{\partial^2 C_\theta\left(u_1,u_2\right)}{\partial u_1\partial u_2}.$$

and Φ_1, Φ_2 marginal c.d.f.s (here Gaussian)

Gradient Descent



What does a Gradient Descent look like ?

1 parameter to estimate

2 parameters to estimate

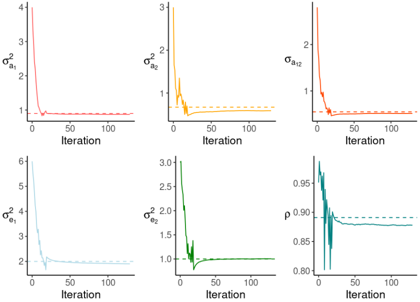
- For Gaussian copula, i.e. in multivariate Gaussian setting, $\boldsymbol{Y}$ is Gaussian.
    - (Guilmour et al., 2003) proposed a Fisher-scoring type algorithm (AI-REML) to estimate the variance components.

- For non-Gaussian copula, the log-likelihood of $\boldsymbol{Y}$ does not have analytic expression
    - We propose a stochastic gradient method to estimate the variance components.

Stochastic gradient algorithm

- Initialization $\eta^{(0)} = (\sigma^2_{a_1}, \sigma^2_{a_2}, \sigma_{a_{12}}, \sigma^2_{e_1}, \sigma^2_{e_1}, \theta)^{(0)}$
- Define a learning rate γ_0
- STEP $m \geq 1$.
  - simulate $\boldsymbol{a}^{(m-1)}$ from the conditional distribution $\boldsymbol{a}|\boldsymbol{Y}$
  - Update the parameter

$$\boldsymbol{\eta}^{(m)} = \boldsymbol{\eta}^{(m-1)} + \gamma_{m-1}\nabla_\eta \log \mathcal{L}(\boldsymbol{\eta}^{(m-1)}; \boldsymbol{y}, \boldsymbol{a}^{(m-1)})$$



Some difficulties

1. A is a very huge matrix, working with A can be numerically complex.
   ▷ but A^{-1} is very sparse! With some simplifications, we can work only with A^{-1}.

Some difficulties

1. A is a very huge matrix, working with A can be numerically complex.
   ▷ but A^{-1} is very sparse! With some simplifications, we can work only with A^{-1}.

   For example

$$\nabla_{(\sigma^2_{a_1},\sigma^2_{a_2},\sigma_{12})} \log \mathcal{L}(G^{(m-1)}; \boldsymbol{a}^{(m-1)})$$

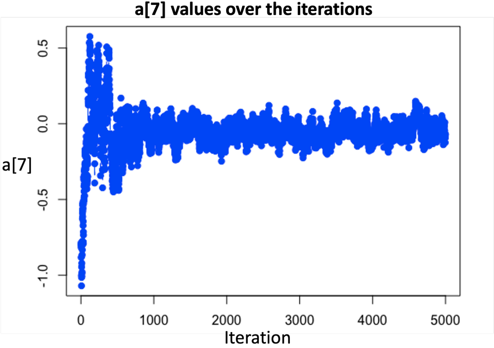
$$= \frac{1}{2}\Big(trace((G^{(m-1)} \otimes A) \times (\nabla G^{-1(m-1)}) \otimes A^{-1}) - (\boldsymbol{a}^{(m-1)})^T((\nabla G^{-1(m-1)}) \otimes A^{-1})\boldsymbol{a}$$

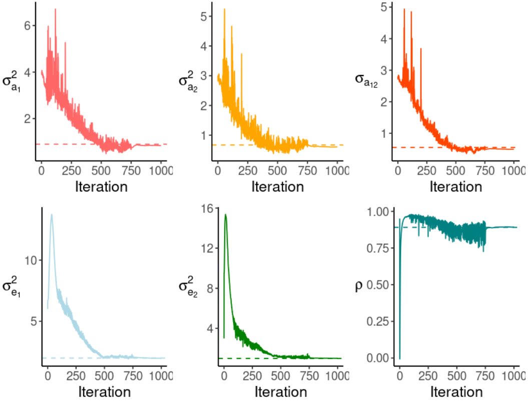
$$= \frac{1}{2}\Big(N \times trace(G^{(m-1)} \times (\nabla G^{-1(m-1)})) - (\boldsymbol{a}^{(m-1)})^T((\nabla G^{-1(m-1)}) \otimes A^{-1})\boldsymbol{a}^{(m-1)}\Big)$$

Some difficulties

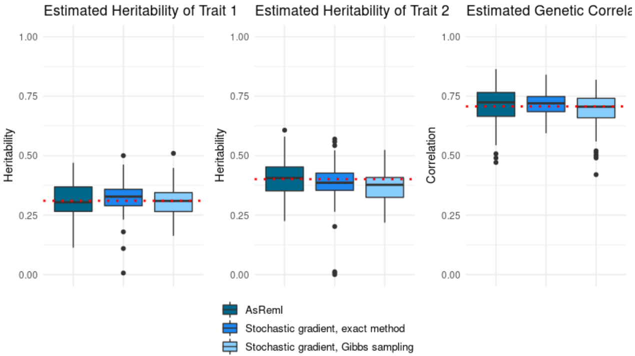
2. For non Gaussian copula, we do not have access to simulations from the conditional distribution $\boldsymbol{a}|\boldsymbol{Y}$.

   ▷ MCMC (block)-Gibbs sampling procedure!

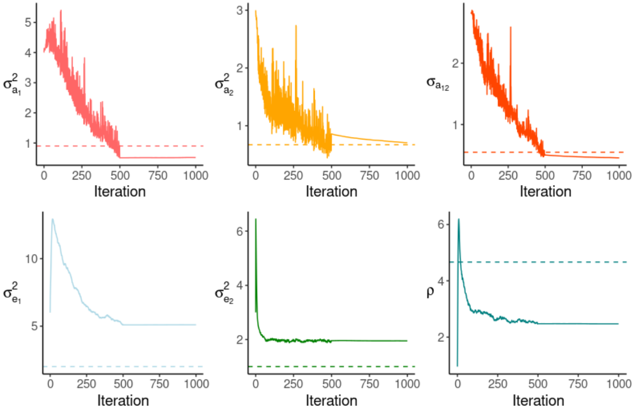

a[7] values over the iterations



Results, 100 runs, Gaussian copula, $n = 720$ animals



Calibration using Clayton copula not finished yet..


$\sigma^2_{a_1}$
Iteration
$\sigma^2_{a_2}$
Iteration
$\sigma_{a_{12}}$
Iteration
$\sigma^2_{e_1}$
Iteration
$\sigma^2_{e_2}$
Iteration
ρ
Iteration