



copula model to improve genetic parameter estimation for dependent traits.

StatMathAppli, 2025, Fréjus

Tom Rohmer, Inrae Toulouse,
joint work with Victoria Bruning & Estelle Kuhn

Sept 2025

➤ multitrait G+E animal model

- ▷ Every phenotypic observation on an animal is determined by environmental and genetic factors and may be defined by the following model:

Phenotypic observation

$$= \text{envir. effects} + \text{genetic effects} + \text{resid. effects}$$

➤ bivariate mixed model (G+E)

With bivariate phenotypes, the most classical inference model in genetic is

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

$\mathbf{y}_j$ the phenotype vectors (neither identically distributed nor independent)

$\boldsymbol{\beta}_j$ parameter vectors to estimate, $\mathbf{X}_j$ design matrices related to fixed effects,

$\mathbf{a}_j$ unobservable (latent) vectors, $\mathbf{Z}_j$ incidence matrices related to genetic effects

$\boldsymbol{\varepsilon}_j$ residual vectors with components assumed i.i.d.



➤ bivariate mixed model (G+E)

With bivariate phenotypes, the most classical inference model in genetic is

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

$\mathbf{y}_j$ the phenotype vectors (neither identically distributed nor independent)

$\boldsymbol{\beta}_j$ parameter vectors to estimate, $\mathbf{X}_j$ design matrices related to fixed effects,

$\mathbf{a}_j$ unobservable (latent) vectors, $\mathbf{Z}_j$ incidence matrices related to genetic effects

$\boldsymbol{\varepsilon}_j$ residual vectors with components assumed i.i.d.

Particularly, the genetic effects (referred to as breeding values) are

$$\mathbf{a}_{i,j} = 0.5(\mathbf{a}_{i_S,j} + \mathbf{a}_{i_D,j}) + \mathbf{M}_{i,j},$$

where $\mathbf{a}_{i_S,j}$ and $\mathbf{a}_{i_D,j}$ are the BVs of the sire and dam $\mathbf{M}_{i,j}$ are the Mendelian sampling terms with distribution

$$(\mathbf{M}_{i,1}, \mathbf{M}_{i,2}) \sim \mathcal{N}(0, \mathbf{G}/2) \quad \text{if no inbreeding}$$

➤ From the classical mixed model to Copula mixed model

The variance of the random terms (latent) is

$$\text{var}(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{G} \otimes \mathbf{A} = \begin{pmatrix} \sigma_{a_1}^2 \mathbf{A} & \sigma_{a_{12}} \mathbf{A} \\ \sigma_{a_{12}} \mathbf{A} & \sigma_{a_2}^2 \mathbf{A} \end{pmatrix}, \quad \mathbf{A} \text{ the kinship matrix,}$$

➤ From the classical mixed model to Copula mixed model

The variance of the random terms (latent) is

$$\text{var}(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{G} \otimes \mathbf{A} = \begin{pmatrix} \sigma_{a_1}^2 \mathbf{A} & \sigma_{a_{12}} \mathbf{A} \\ \sigma_{a_{12}} \mathbf{A} & \sigma_{a_2}^2 \mathbf{A} \end{pmatrix}, \quad \mathbf{A} \text{ the kinship matrix,}$$

and for $i = 1, \dots, n$

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2), \quad (\varepsilon_{i,1}, \varepsilon_{i,2}) \text{ is assumed jointly Gaussian}$$

Then, REML procedures are used to estimate parameters.

➤ From the classical mixed model to Copula mixed model

The variance of the random terms (latent) is

$$\text{var}(\mathbf{a}_1, \mathbf{a}_2) = \mathbf{G} \otimes \mathbf{A} = \begin{pmatrix} \sigma_{a_1}^2 \mathbf{A} & \sigma_{a_{12}} \mathbf{A} \\ \sigma_{a_{12}} \mathbf{A} & \sigma_{a_2}^2 \mathbf{A} \end{pmatrix}, \quad \mathbf{A} \text{ the kinship matrix,}$$

and for $i = 1, \dots, n$

$$\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2), \quad (\varepsilon_{i,1}, \varepsilon_{i,2}) \text{ is assumed jointly Gaussian}$$

Then, REML procedures are used to estimate parameters.

➡ The multivariate-Gaussian assumption is often unrealistic due to the dependence structure of the observations, we propose the generalization:

$$\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_j^2), \quad (\varepsilon_{i,1}, \varepsilon_{i,2}) \text{ has copula } C_\theta$$

Copulas

Definition: A **copula** $C : [0, 1]^d \rightarrow [0, 1]$ is the multivariate cumulative distribution function (c.d.f.) of a random vector whose marginal distributions are uniforms on $[0, 1]$.

> Copulas

Definition: A **copula** $C : [0, 1]^d \rightarrow [0, 1]$ is the multivariate cumulative distribution function (c.d.f.) of a random vector whose marginal distributions are uniforms on $[0, 1]$.

Theorem of [Sklar(1959)]

Let $\mathbf{X} = (X_1, \dots, X_d)$ be a d -dimensional random vector with c.d.f. $\mathbf{F}$ and let $F_1, \dots, F_d$ be the marginal c.d.f. of $\mathbf{X}$ assumed continuous. Then it exists a unique copula C such that:

$$\mathbf{F}(\mathbf{x}) = C\{F_1(x_1), \dots, F_d(x_d)\}, \quad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d.$$



➤ Copulas

Definition: A **copula** $C : [0, 1]^d \rightarrow [0, 1]$ is the multivariate cumulative distribution function (c.d.f.) of a random vector whose marginal distributions are uniforms on $[0, 1]$.

Theorem of [Sklar(1959)]

Let $\mathbf{X} = (X_1, \dots, X_d)$ be a d -dimensional random vector with c.d.f. $\mathbf{F}$ and let $F_1, \dots, F_d$ be the marginal c.d.f. of $\mathbf{X}$ assumed continuous. Then it exists a unique copula C such that:

$$\mathbf{F}(\mathbf{x}) = C\{F_1(x_1), \dots, F_d(x_d)\}, \quad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

- The copula C characterizes the dependence structure of vector $\mathbf{X}$.

➤ Copulas

Definition: A **copula** $C : [0, 1]^d \rightarrow [0, 1]$ is the multivariate cumulative distribution function (c.d.f.) of a random vector whose marginal distributions are uniforms on $[0, 1]$.

Theorem of [Sklar(1959)]

Let $\mathbf{X} = (X_1, \dots, X_d)$ be a d -dimensional random vector with c.d.f. $\mathbf{F}$ and let $F_1, \dots, F_d$ be the marginal c.d.f. of $\mathbf{X}$ assumed continuous. Then it exists a unique copula C such that:

$$\mathbf{F}(\mathbf{x}) = C\{F_1(x_1), \dots, F_d(x_d)\}, \quad \mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d.$$

- The copula C characterizes the dependence structure of vector $\mathbf{X}$.

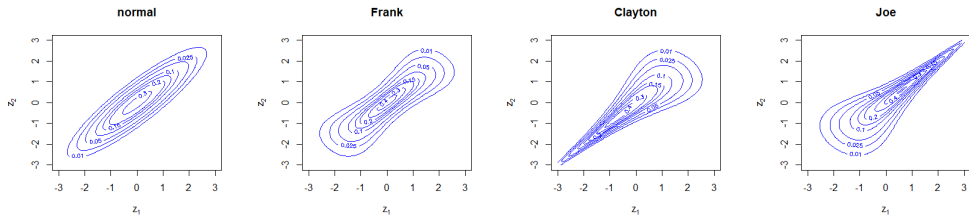


A. Sklar.

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 of bivariate distributions with Gaussian margins and several copula



The lack of consideration for an appropriate dependence structure (e.g., wrongly assuming a Gaussian distribution) may lead to poor estimation of variance parameters.



Rohmer, T., Ricard, A., David, I.

Copula miss-specification in REML multivariate genetic animal model estimation,
Genetics Selection Evolution, May 2022

➤ On real data from growing pig I

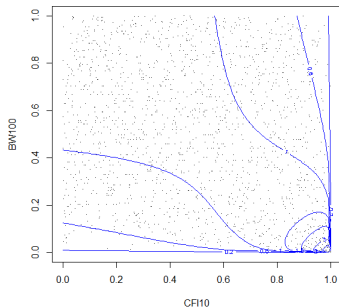


Figure: Contours plot of the fitted copula from pseudo-observations for CFI10 and BW100 (rotated 90 degrees Clayton copula): (left) uniform scale, (middle) gaussian scale, (right) Large-white pigs!

➤ Maximum likelihood estimation

Because the BVs $\mathbf{a}_j$ are unobservable, the log-density is

$$\log f_{\mathbf{Y}}(\mathbf{y}) = \log \int_{\mathbb{R}^{2N}} f_{\mathbf{Y}|\mathbf{a}}(\mathbf{y}|\mathbf{a}) f_{\mathbf{a}}(\mathbf{a}) d\mathbf{a}.$$

The MLE is solution with respect to $\xi = (\beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \theta, \sigma_{a_1}^2, \sigma_{a_2}^2, \sigma_{a_{12}})$ of the system equations

$$\frac{\partial}{\partial \xi} \log f_{\mathbf{Y}}(\mathbf{y}; \xi) = 0.$$

Remember that

$$\begin{aligned} \mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2) &\sim \mathcal{N}_{2N}(\mathbf{0}, \mathbf{G} \otimes \mathbf{A}); \\ Y_{ij}|Z_i \mathbf{a}_j &\sim \mathcal{N}(Z_i \mathbf{a}_j + \mathbf{x}_{ij} \beta_j, \sigma_j^2) \\ (Y_{i,1}, Y_{i,2})|(Z_i \mathbf{a}_1, Z_i \mathbf{a}_2) &\text{ has copula } C_\theta; \end{aligned}$$

➤ Stochastic gradient descent algorithm

Remember we have to solve

$$\frac{\partial}{\partial \xi} \log f_{\mathbf{Y}}(\mathbf{y}; \xi) = 0.$$

First note that the Fisher's identity states

$$\frac{\partial}{\partial \xi} \log f_{\mathbf{Y}}(\mathbf{y}; \xi) = \mathbb{E}_{\mathbf{a}|\mathbf{y}} \left(\frac{\partial}{\partial \xi} \log f_{(\mathbf{Y}, \mathbf{a})}(\mathbf{y}, \mathbf{a}; \xi) \right).$$

➤ Stochastic gradient descent algorithm

Remember we have to solve

$$\frac{\partial}{\partial \xi} \log f_{\mathbf{Y}}(\mathbf{y}; \xi) = 0.$$

First note that the Fisher's identity states

$$\frac{\partial}{\partial \xi} \log f_{\mathbf{Y}}(\mathbf{y}; \xi) = \mathbb{E}_{\mathbf{a}|\mathbf{y}} \left(\frac{\partial}{\partial \xi} \log f_{(\mathbf{Y}, \mathbf{a})}(\mathbf{y}, \mathbf{a}; \xi) \right).$$

An SGD algorithm is

for $m \in 1, \dots, M$ do:

- ▶ Simulate $\mathbf{a}^{(m)}$ from the conditional distribution of $\mathbf{a}|\mathbf{y}$
- ▶ Update the parameter:

$$\xi^{(m)} = \xi^{(m-1)} + \gamma_m \frac{\partial}{\partial \xi} \log f_{(\mathbf{Y}, \mathbf{a})}(\mathbf{y}, \mathbf{a}^{(m)}; \xi^{(m-1)})$$

for a well-chosen learning rate γ_m .

> Simplifications

The update step can be rewritten as

$$\mathbf{G}^{(m)} = \mathbf{G}^{(m-1)} + \gamma_{1,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{a}}(\mathbf{a}^{(m)}),$$

$$(\boldsymbol{\beta}, \sigma_j^2)^{(m)} = (\boldsymbol{\beta}, \sigma_j^2)^{(m-1)} + \gamma_{2,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{Y}|\mathbf{a}}(\mathbf{y} | \mathbf{z}^T \mathbf{a}^{(m)})$$

and

$$\theta^{(m)} = \theta^{(m-1)} + \gamma_{3,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{Y}|\mathbf{a}}(\mathbf{y} | \mathbf{z}^T \mathbf{a}^{(m)}).$$

> Simplifications

The update of the covariance matrix for the genetic effects is

$$\mathbf{G}^{(m)} = \mathbf{G}^{(m-1)} + \gamma_{1,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{a}}(\mathbf{a}^{(m)}),$$

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

$$\begin{aligned} & \frac{\partial}{\partial \xi} \log f(\mathbf{a}^{(m)}) \\ &= \frac{1}{2} \left(\text{trace} \left(\left(\mathbf{G}^{(m-1)} \otimes A \right) \times \left(\nabla \mathbf{G}^{-1(m-1)} \right) \otimes A^{-1} \right) - \left(\mathbf{a}^{(m)} \right)^T \left(\left(\nabla \mathbf{G}^{-1(m-1)} \right) \otimes A^{-1} \right) \mathbf{a}^{(m)} \right) \\ &= \frac{1}{2} \left(N \times \text{trace} \left(\mathbf{G}^{(m-1)} \times \left(\nabla \mathbf{G}^{-1(m-1)} \right) \right) - \left(\mathbf{a}^{(m)} \right)^T \left(\left(\nabla \mathbf{G}^{-1(m-1)} \right) \otimes A^{-1} \right) \mathbf{a}^{(m)} \right) \end{aligned}$$

> Simplifications

It can be rewritten as can be rewritten as

$$\mathbf{G}^{(m)} = \mathbf{G}^{(m-1)} + \gamma_{1,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{a}}(\mathbf{a}^{(m)}),$$

↪ based on trace of sparse matrix

$$(\boldsymbol{\beta}, \sigma_j^2)^{(m)} = (\boldsymbol{\beta}, \sigma_j^2)^{(m-1)} + \gamma_{2,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{Y}|\mathbf{a}}(\mathbf{y}|z^T \mathbf{a}^{(m)})$$

and

$$\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} + \gamma_{3,m} \frac{\partial}{\partial \xi} \log f_{\mathbf{Y}|\mathbf{a}}(\mathbf{y}|z^T \mathbf{a}^{(m)}).$$

↪ ↪ more complex analytic formulations (derivatives of copula density) but no real challenge to efficiently compute it.



➤ Simulation of the conditional distribution of BVs given observations

- ▶ for Gaussian copula, $\mathbf{a}|\mathbf{Y}$ has explicit Gaussian distribution with covariance $((\mathbf{ZG} \otimes \mathbf{AZ}^T)^{-1} + (\mathbf{R} \otimes \mathbf{I}_n)^{-1})^{-1}$.
- ↪ That can be sampled using a Cholesky decomposition



➤ Simulation of the conditional distribution of BVs given observations

- ▶ for Gaussian copula, $\mathbf{a}|\mathbf{Y}$ has explicit Gaussian distribution with covariance $((\mathbf{ZG} \otimes \mathbf{AZ}^T)^{-1} + (\mathbf{R} \otimes \mathbf{I}_n)^{-1})^{-1}$.

➡ That can be sampled using a Cholesky decomposition

- ▶ for non-Gaussian copula, $\mathbf{a}|\mathbf{Y}$ does not have an explicit distribution ($\propto f(\mathbf{Y}|\mathbf{a})f(\mathbf{a})$), and there is not easy resampling scheme.

➡ Hybrid MCMC-Metropolis-Gibbs block sampling



Rohmer, T., Bruning, V. and Kuhn, Estelle

G+E copula model to improve the estimation of the genetic parameters in bivariate mixed model,
submitted, 2025

➤ On simulations

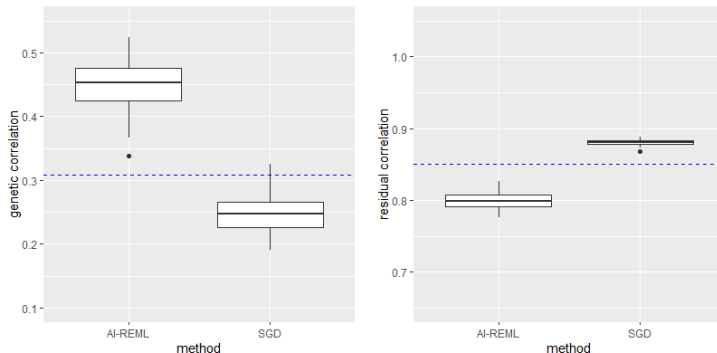
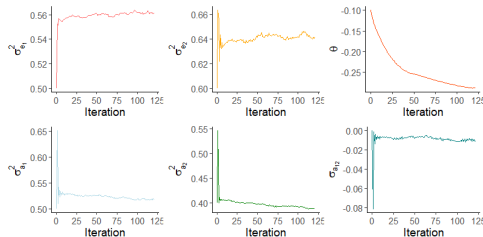


Figure: Boxplot of the estimated genetic correlation (left), residual correlation (right), using $B = 50$ runs. The Clayton model using MLE (SGD) and the Gaussian model using REML are compared. For the two models, data are sampled through Clayton copula for the residual part from 9900 animals under selection using a truncation selection process. Blue dotted line is the true value of the parameter

> On real data II

varcomp	$\sigma_{a_1}^2$	$\sigma_{a_2}^2$	$\sigma_{a_{12}}$	$\sigma_{e_1}^2$	$\sigma_{e_2}^2$	θ	h_{CFI10}^2	h_{BW100}^2	ρ_e	iterations
AI-REML	0.56	0.37	-0.02	0.38	0.54	-0.18	0.59	0.41	-0.39	7
rC-SGD	0.52	0.39	-0.01	0.56	0.64	-0.28	0.48	0.38	-0.20	121

Table: Estimation of the variance components using Gaussian inference model with AI-REML procedure and using rotated 90 degree Clayton inference model with SGD procedure, using 3 generations, $n = 1749$, $N = 4653$.



Next steps

1. From bivariate to multivariate analysis.
↳ Preconditionning the GD step by Fisher's information.



Next steps

1. From bivariate to multivariate analysis.
↳ Preconditionning the GD step by Fisher's information.
2. R package
↳ for users



Next steps

1. From bivariate to multivariate analysis.

↪ Preconditionning the GD step by Fisher's information.

2. R package

↪ for users

3. Robustness to a model misspecification? (e.g. copula, or marginal distribution)

