



Inference in copula models

séminaire de de statistiques, Institut de Mathématiques de Marseille

Tom Rohmer (Inrae)

12 mai 2025



Plan

Short introduction to copulas

Non parametric estimation

Parametric estimation

Copulas in quantitative genetics



Plan

Short introduction to copulas

Non parametric estimation

Parametric estimation

Copulas in quantitative genetics



Multivariate observations

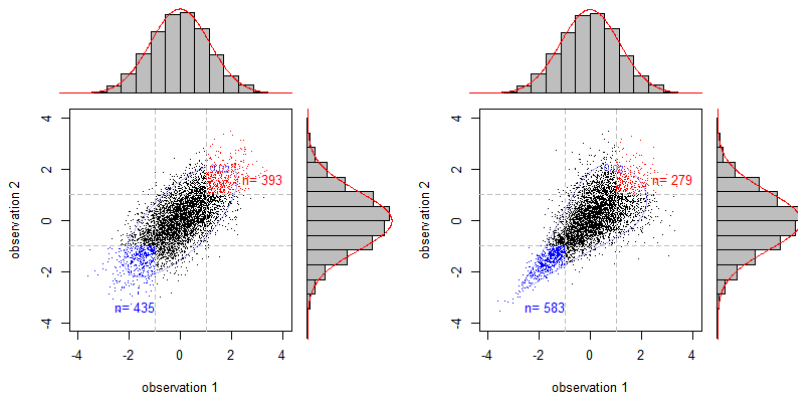


Figure: Simulation of $n = 5000$ bivariate observations whose the univariate distributions are Gaussian and with Spearman's correlation $\rho = 0.7$. (Left) Multivariate Gaussian distribution. (Right) Clayton-type distribution

➤ Tail dependence measure

Lower tail dependence:

$$\lambda_\ell = \lim_{u \rightarrow 0} \lambda_\ell(u), \quad \lambda_\ell(u) = \mathbb{P}(Y_2 \leq F_2^{-1}(u) \mid Y_1 \leq F_1^{-1}(u))$$

Upper tail dependence:

$$\lambda_u = \lim_{u \rightarrow 1} \lambda_u(u), \quad \lambda_u(u) = \mathbb{P}(Y_2 > F_2^{-1}(u) \mid Y_1 > F_1^{-1}(u))$$

Gaussian case: $\lambda_\ell = \lambda_u = 0$ regardless of ρ

Clayton case: $\lambda_\ell = 0.61$, $\lambda_u = 0$ for $\rho = 0.7$

In the previous example, these 'empirical coefficients' are

Gaussian case: $\hat{\lambda}_\ell(0.01) = 0.28$, $\hat{\lambda}_u(0.99) = 0.24$

Clayton case: $\hat{\lambda}_\ell(0.01) = 0.70$, $\hat{\lambda}_u(0.99) = 0.04$



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 $\mathbf{X}$. Then it exists a copula 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 $\mathbf{X}$. Then it exists a copula 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 $\mathbf{X}$.
- ▶ The copula C can be expressed as follows:

$$C(\mathbf{u}) = \mathbf{F}\{F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)\}, \quad \mathbf{u} = (u_1, \dots, u_d) \in [0, 1]^d.$$

➤ 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 $\mathbf{X}$. Then it exists a copula 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 $\mathbf{X}$.
- ▶ The copula C can be expressed as follows:

$$C(\mathbf{u}) = \mathbf{F}\{F_1^{-1}(u_1), \dots, F_d^{-1}(u_d)\}, \quad \mathbf{u} = (u_1, \dots, u_d) \in [0, 1]^d.$$



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.

INRAE

Inference in copula models

12 mai 2025 / Tom Rohmer, Inrae Toulouse, France

➤ Some classical (bivariate) copulas

Normal copula:

$$C_{\rho}^N(u, v) = \Phi_{\rho}(\Phi^{-1}(u), \Phi^{-1}(v)), \quad (u, v) \in [0, 1]^2,$$

where Φ and Φ_{ρ} stand for the c.d.f. of the standard univariate Gaussian distribution and the bivariate Gaussian distribution with correlation ρ .

Frank, Clayton, Joe copulas:

$$C_{\theta}^F(u, v) = \frac{1}{\theta} \log \left(1 + \frac{(\exp(-u\theta) - 1)(\exp(-v\theta) - 1)}{\exp(-\theta) - 1} \right), \quad \theta \in \mathbb{R}^*,$$

$$C_{\theta}^{Cl}(u, v) = \max \left(\left(u^{-\theta} + v^{-\theta} - 1 \right)^{-1/\theta}, 0 \right), \quad \theta \in [-1, 0) \cup (0, +\infty),$$

$$C_{\theta}^J(u, v) = 1 - \left[(1-u)^{\theta} + (1-v)^{\theta} - (1-u)^{\theta}(1-v)^{\theta} \right]^{1/\theta} \quad \theta \geq 1.$$

https://testmyshinyapply.shinyapps.io/Shiny_copula



dependence's measure and copula

Kendall's τ (measure of concordance) and Spearman's ρ_S (rank based correlation measure) are often employed to measure a non-linear relation between to variables The Kendall's (but also Spearman's) correlation is related with copula by

$$\tau = 4 \int_{[0,1]^2} C_{\theta}(u, v) dC_{\theta}(u, v) - 1,$$

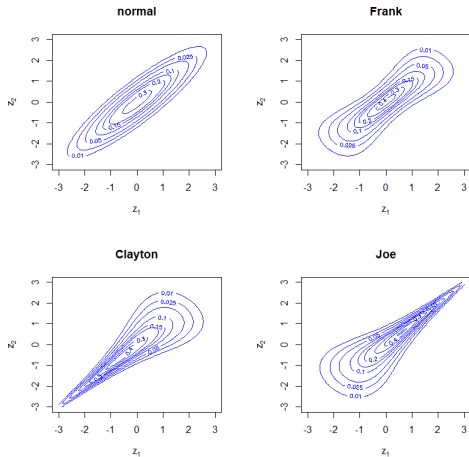
$$\rho_S = 12 \int_{[0,1]^2} C_{\theta}(u, v) dudv - 3,$$

Concerning the tail dependence indexes:

$$\lambda_{\ell} = \lim_{u \rightarrow 0^+} \frac{C(u, u)}{u}, \quad \lambda_u = \lim_{u \rightarrow 1^-} \frac{1 - 2u + C(u, u)}{1 - u}$$



Contour plots of bivariate distributions with Gaussian margins and several copula





To summarize

1. Copula links marginales distributions to the joint distribution
2. it captures and characterizes dependence structure between variables
3. it makes possible to modelize jointly multiple dependent phenomena, (continuous data, discrete data or combination)
4. it is not restricted to the bivariate case and are especially useful in the case of tail dependence.



R packages

-  M. Hofert, I. Kojadinovic, M. Maechler and J. Yan
copula: Multivariate Dependence with Copulas
R package version 1.1-4, 2024
-  Nagler T., Schepsmeier U., Stoeber J.
VineCopula: Inference of Vine Copulas
R package version 2.5.1, 2024
-  Rohmer, T.
npcopTest: Non Parametric Test for Detecting Changes in the Copula
R package version: 1.03, 2018
Analyse de la dépendance avec R, une brève introduction aux copules

Plan

Short introduction to copulas

Non parametric estimation

Parametric estimation

Copulas in quantitative genetics



➤ Copulas and test for break detections

In many contexts, we want to test

$\mathcal{H}_0$:

$\exists F$ such that $\mathbf{X}_1, \dots, \mathbf{X}_n$ have c.d.f. F .

Sklar's theorem allows to rewrite $\mathcal{H}_0$ as $\mathcal{H}_{0,m} \cap \mathcal{H}_{0,c}$ where

$\mathcal{H}_{0,m} \cap \mathcal{H}_{0,c}$:

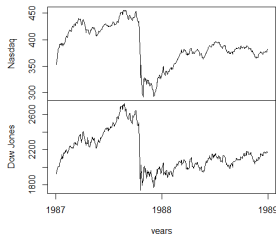
$\mathcal{H}_{0,c}$: $\exists C$, such that $\mathbf{X}_1, \dots, \mathbf{X}_n$ have copula C

$\mathcal{H}_{0,m}$: $\exists F_1, \dots, F_d$ such that $\mathbf{X}_1, \dots, \mathbf{X}_n$ have m.c.d.f. $F_1, \dots, F_d$.

Hence $\mathcal{H}_0$ can be rejected because $\neg \mathcal{H}_{0,c}$ and/or $\neg \mathcal{H}_{0,m}$

> examples

Example 1: Nasdaq, Dow Jones and the "black Monday" (1987-10-19),  library QRM



Example 2: Identification of changes in drinking behaviour and/or physical activity of pig farming that could indicate a technical or health-related event



Hamelin, G., Rohmer, T., Gaillard, C.

Amélioration d'un outil de détection d'événements techniques ou sanitaires à l'échelle de plusieurs bandes de truies
Poster Journées de la recherche Porcine, 2023 (JRP23)

INRAE

Inference in copula models

12 mai 2025 / Tom Rohmer, Inrae Toulouse, France

➤ Non-parametric estimation of the copula

In the bivariate case, for an i.i.d. sample, the identification of the copula of $(X_{i,1}, X_{i,2})$ can be done using graphical plot of the pseudo-observations $(\hat{U}_{i,1}, \hat{U}_{i,2})$ where

$$\hat{U}_{i,j} = F_{n,j}(X_{i,j})$$

and $F_{n,j}$ is the empirical cumulative distribution of $X_{1,j}, \dots, X_{n,j}$.

➤ Non-parametric estimation of the copula

In the bivariate case, for an i.i.d. sample, the identification of the copula of $(X_{i,1}, X_{i,2})$ can be done using graphical plot of the pseudo-observations $(\hat{U}_{i,1}, \hat{U}_{i,2})$ where

$$\hat{U}_{i,j} = F_{n,j}(X_{i,j})$$

and $F_{n,j}$ is the empirical cumulative distribution of $X_{1,j}, \dots, X_{n,j}$.

Note that:

1. copulas are invariant under increasing transformations of the margins!
2. A consistent non-parametric estimation of the copula is given by

$$C_n(u_1, \dots, u_d) = \frac{1}{n} \sum_{i=1}^n \prod_{j=1}^d 1(\hat{U}_{i,j} \leq u_j) \quad (u_1, \dots, u_d) \in [0, 1]^d$$

The selection can also be done using AIC criterion, see for example the R function `BiCopCompare` of the package `VineCopula`



➤ Example on pig data

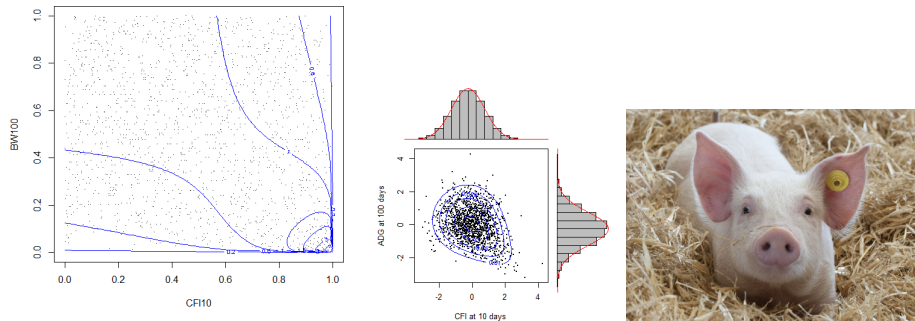


Figure: Contours plot of the fitted copula from pseudo-observations for CFI10 and BW100 (rotated 90 degrees Clayton copula): (left) uniform scale, (middle) gaussian scale, (right) Large-white pigs!

Plan

Short introduction to copulas

Non parametric estimation

Parametric estimation

Copulas in quantitative genetics



➤ Log-likelihood

let consider for $i = 1, \dots, n$, $\underline{y}_i = (y_{i,1}, \dots, y_{i,s})$ be independent random vectors with copula C_θ and marginal distributions $F_{i,j}$ assumed continuous, which depend on parameters β_j . The density of $\underline{y}_i$ is

$$f_i(\underline{y}_i | \beta, \theta) = c_\theta(F_{i,1}(y_{i,1} | \beta_1), \dots, F_{i,s}(y_{i,s} | \beta_s)) \prod_{j=1}^s f_{i,j}(y_{i,j} | \beta_j),$$

where c_θ is the density of the copula given by:

$$c_\theta(u_1, \dots, u_s) = \frac{\partial^s C_\theta(u_1, \dots, u_s)}{\partial u_1 \dots \partial u_s}.$$

The marginal parameters $\beta_1, \dots, \beta_s$ and the copula parameter θ are unknown and are to be estimate.



➤ Log-likelihood

The log-density can be written as

$$\log f_i(\underline{\mathbf{y}}_i | \boldsymbol{\beta}, \theta) = \log c_\theta(F_{i,1}(y_{i,1} | \boldsymbol{\beta}_1), \dots, F_{i,s}(y_{i,s} | \boldsymbol{\beta}_s)) + \sum_{j=1}^s \log f_{i,j}(y_{i,j} | \boldsymbol{\beta}_j).$$

Let $\log \mathcal{L}_j$ be the j -marginal contributions over i , i.e.,

$$\log \mathcal{L}_j(\boldsymbol{\beta}_j | \mathbf{y}) = \sum_{i=1}^n \log f_{i,j}(\boldsymbol{\beta}_j | y_{i,j}). \quad (1)$$

Hence the full log-likelihood $\log \mathcal{L}$ of $\mathbf{y} = (\underline{\mathbf{y}}_1, \dots, \underline{\mathbf{y}}_n)^T$ can be expressed as

$$\log \mathcal{L}(\boldsymbol{\beta}, \theta | \mathbf{y}) = \underbrace{\sum_{i=1}^n \log c_\theta(F_{i,1}(y_{i,1} | \boldsymbol{\beta}_1), \dots, F_{i,s}(y_{i,s} | \boldsymbol{\beta}_s))}_{(a)} + \underbrace{\sum_{j=1}^s \log \mathcal{L}_j(\boldsymbol{\beta}_j | \mathbf{y})}_{(b)}. \quad (2)$$

> Estimation

1. MLE consists in finding the solution of

$$\left(\frac{\partial \log \mathcal{L}}{\partial \beta_1}, \dots, \frac{\partial \log \mathcal{L}}{\partial \beta_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\beta, \theta | \mathbf{y}) = 0.$$

Found the MLE using gradient descent can be time consuming for a high number of variables or very huge n .

> Estimation

1. MLE consists in finding the solution of

$$\left(\frac{\partial \log \mathcal{L}}{\partial \beta_1}, \dots, \frac{\partial \log \mathcal{L}}{\partial \beta_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\beta, \theta | \mathbf{y}) = 0.$$

Found the MLE using gradient descent can be time consuming for a high number of variables or very huge n .

2. Inference for margins estimators (IFM) [Xu96] consists in finding the solution of

$$\left(\frac{\partial \log \mathcal{L}_1}{\partial \beta_1}, \dots, \frac{\partial \log \mathcal{L}_s}{\partial \beta_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\beta, \theta | \mathbf{y}) = 0.$$

That is to say, optimize (b) then (a).

➤ Estimation

1. MLE consists in finding the solution of

$$\left(\frac{\partial \log \mathcal{L}}{\partial \beta_1}, \dots, \frac{\partial \log \mathcal{L}}{\partial \beta_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\beta, \theta | \mathbf{y}) = 0.$$

Found the MLE using gradient descent can be time consuming for a high number of variables or very huge n .

2. Inference for margins estimators (IFM) [Xu96] consists in finding the solution of

$$\left(\frac{\partial \log \mathcal{L}_1}{\partial \beta_1}, \dots, \frac{\partial \log \mathcal{L}_s}{\partial \beta_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\beta, \theta | \mathbf{y}) = 0.$$

That is to say, optimize (b) then (a).



Xu, J. J.

Statistical modelling and inference for multivariate and longitudinal discrete response data.
PhD thesis, University of British Columbia, 1996.

➤ with no identically distributed samples

When the marginal distributions depends on categorical explanatory variable with a high number of modalities, IFM can be again very time consuming due to the optimization problem (b).

➤ with no identically distributed samples

When the marginal distributions depends on categorical explanatory variable with a high number of modalities, IFM can be again very time consuming due to the optimization problem (b).

For marginal distributions inside the exponential family, we proposed in [Brouste et al. 2023, Brouste et al. 2025] an alternative IFM procedure that works with

1. a consistent (and explicit) guess estimator for the marginal parameters [Brouste et al. 2022]

➤ with no identically distributed samples

When the marginal distributions depends on categorical explanatory variable with a high number of modalities, IFM can be again very time consuming due to the optimization problem (b).

For marginal distributions inside the exponential family, we proposed in [Brouste et al. 2023, Brouste et al. 2025] an alternative IFM procedure that works with

1. a consistent (and explicit) guess estimator for the marginal parameters [Brouste et al. 2022]
2. a one-step (Fisher's scoring) procedure to obtain an efficient estimation $\hat{\beta}^{OS}$ of β and finally a consistent estimation $\hat{\theta}$ of θ such that

$$\frac{\partial \log \mathcal{L}}{\partial \theta}(\hat{\beta}^{OS}, \hat{\theta} | \mathbf{y}) = 0.$$





with no identically distributed samples



Brouste, A., Dutang, C. and Rohmer, T.

One-step closed-form estimator for generalized linear model with categorical explanatory variables
Statistics, Simulation and computation, 2022.



Brouste, A., Dutang, C., Hovsepyan, L. and Rohmer, T.

One-step closed-form estimator for generalized linear model with categorical explanatory variables
Statistics and Computing, 2023.



Brouste, A., Dutang, C., Hovsepyan, L. and Rohmer, T.

Fast inference in copula models with categorical explanatory variables using the one-step procedure
submitted, 2025.



Hovsepyan, L.

Fast and efficient estimation methods for generalized linear models with applications in insurance
ph.D. thesis 2025



Alexandre Brouste, Christophe Dutang & Tom Rohmer

glmtools: Tools to fit generalized linear models with explicit expressions

Dépôt logiciel d'invention, 2022, disponible à la demande

Next version: multivariate glm using copulas.

➤ Discrete and mixed case in bivariate model

- ▷ When all the 2 margins are discrete, the p.m.f. can be rewritten as:

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2) &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2) - \mathbb{P}(X_1 \leq x_1^-, X_2 \leq x_2) \\ &\quad - \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2^-) + \mathbb{P}(X_1 \leq x_1^-, X_2 \leq x_2^-) \\ &= C_\theta(F_1(x_1), F_2(x_2)) - C_\theta(F_1(x_1^-), F_2(x_2)) \\ &\quad - C_\theta(F_1(x_1), F_2(x_2^-)) + C_\theta(F_1(x_1^-), F_2(x_2^-))\end{aligned}$$



➤ Discrete and mixed case in bivariate model

- ▷ In the mixed case, when X_1 is continuous and X_2 discrete the density is

$$f(x_1, x_2) = f(x_1)(c_{\theta, u_1}(F_1(x_1), F_2(x_2)) - c_{\theta, u_1}(F_1(x_1), F_2(x_2^-)))$$

with

$$c_{\theta, u_1}(u_1, u_2) = \frac{\partial}{\partial u_1} C_{\theta}(u_1, u_2)$$

➤ Discrete and mixed case in bivariate model

▷ In the mixed case, when X_1 is continuous and X_2 discrete the density is

$$f(x_1, x_2) = f(x_1)(c_{\theta, u_1}(F_1(x_1), F_2(x_2)) - c_{\theta, u_1}(F_1(x_1), F_2(x_2^-)))$$

with

$$c_{\theta, u_1}(u_1, u_2) = \frac{\partial}{\partial u_1} C_{\theta}(u_1, u_2)$$

Example:



Tomilina, E. and Mazo, G. and Jaffrézic, F.

A semi-parametric Gaussian copula model for heterogeneous network inference: an application to multi-omics data
<https://hal.inrae.fr/hal-04847648>, 2024.

Plan

Short introduction to copulas

Non parametric estimation

Parametric estimation

Copulas in quantitative genetics



➤ 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

= envir. effects + genetic effects + resid. effects

➤ Copula mixed model (G+E)

The inference model is

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

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

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

$\mathbf{Z}$ $n \times N$ incidence matrix; $\mathbf{a}_j$ unobservable BV vectors of size $N \geq n$.

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



➤ Copula mixed model (G+E)

The inference model is

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

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

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

$\mathbf{Z}$ $n \times N$ incidence matrix; $\mathbf{a}_j$ unobservable BV vectors of size $N \geq n$.

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

Particularly, the BVs 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

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

➤ Copula mixed model (G+E)

Because founders are assumed unrelated, the variance of the random terms (latent) is

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

➤ Copula mixed model (G+E)

Because founders are assumed unrelated, the variance of the random terms (latent) is

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

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

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

The lack of consideration for an adequate dependence structure (e.g., assuming a Gaussian structure) 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

Reformulation

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

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

$$\log f(\mathbf{y}) = \log \int_{2N} f(\mathbf{y} | \mathbf{a}) f(\mathbf{a}) d\mathbf{a}.$$

The MLE is solution with respect to $\xi = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \sigma_1^2, \sigma_2^2, \theta, \sigma_{a_1}^2, \sigma_{a_2}^2, \sigma_{a_{12}}^2)$ of the system equations

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

➤ Stochastic gradient descent algorithm

Remember we have to solve

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

First note that the Fisher's identity states

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

➤ Stochastic gradient descent algorithm

Remember we have to solve

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

First note that the Fisher's identity states

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

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:

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

for a well-chosen learning rate γ_m .

➤ Simplifications

It can be rewritten as can be rewritten as

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

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} :

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

> Simplifications

It can be rewritten as can be rewritten as

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

↪ based on trace of sparse matrix

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

and

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

↪ ↪ 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



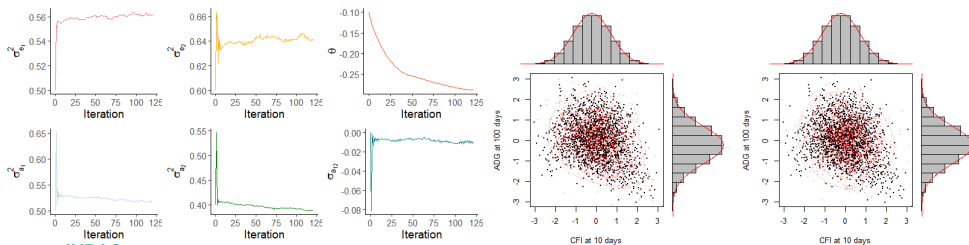
Rohmer, T., Bruning, V. and Kuhn, Estelle

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

➤ On real data

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, $n = 1749$, $N = 4653$.



INRAE

Inference in copula models

12 mai 2025 / Tom Rohmer, Inrae Toulouse, France