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

Journées des Statistiques 2022

Tom Rohmer, Anne Ricard & Ingrid David

13-17 juin 2022

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

Plan

Plan

multitrait genetic animal model

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

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

multitrait genetic animal model

▷ When 2 phenotypes are simultaneously observed, multivariate mixed models are widely used in animal genetics to deal with genetic and environmental effects:

$$\begin{cases} \mathsf{y}_1 = X_1\beta_1 + Z\mathsf{a}_1 + \varepsilon_1 \\ \mathsf{y}_2 = X_2\beta_2 + Z\mathsf{a}_2 + \varepsilon_2 \end{cases},$$

where

- y_j are the phenotype vectors of size n
- β_j the related unknown parameter vector; X_j the design matrices for the fixed effect
- a_j the random vector of breeding values (BV) to predict; Z the design matrix for the random effect (genetic part);

multitrait genetic animal model

Particularly, the BVs are

$$a_{i,j} = 0.5(a_{i_S,j} + a_{i_D,j}) + M_{i,j},$$

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.

- The distribution of the breeding vector (a_1, a_2) is assumed to be Gaussian

$$\begin{pmatrix} \mathrm{a}_1 \\ \mathrm{a}_2 \end{pmatrix} \sim \mathcal{N}(0, G \otimes A),$$

with G a 2×2 unknown genetic covariance matrix to be estimate, and A the genetic relationship matrix, of size $N \times N$, $N \geq n$ related to the pedigree;
the distribution of the Mendelian sampling term are $(M_{i,1}, M_{i,2}) \sim \mathcal{N}(0, G/2)$.

multitrait genetic animal model

- Finally, the residual $(\varepsilon_1, \varepsilon_2)$ follow a standard bivariate Gaussian distribution with covariance matrix $E \otimes I_n$ with E the 2×2 residual covariance matrix to be estimate.

The REML to estimate G and E consists to maximized the restricted log-likelihood

$$\lambda_R(\sigma^2_{a_1}, \sigma^2_{a_2}, \sigma_{a_{12}}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma_{e_{12}}) = -\frac{1}{2}\log[|V|] - \frac{1}{2}\log\left[\left|X^T V^{-1} X\right|\right] - \frac{1}{2} y^T P y$$

where V covariance matrix of $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ and

$P = V^{-1} - V^{-1}X\left(X^T V^{-1} X\right)^{-1} X^T V^{-1}$ with $X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}$.

multitrait genetic animal model

- Finally, the residual $(\varepsilon_1, \varepsilon_2)$ follow a standard bivariate Gaussian distribution with covariance matrix $E \otimes I_n$ with E the 2×2 residual covariance matrix to be estimate.

The REML to estimate G and E consists to maximized the restricted log-likelihood

$$\lambda_R(\sigma^2_{a_1}, \sigma^2_{a_2}, \sigma_{a_{12}}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma_{e_{12}}) = -\frac{1}{2}\log[|V|] - \frac{1}{2}\log\left[\left|X^T V^{-1} X\right|\right] - \frac{1}{2} y^T P y$$

where V covariance matrix of $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ and

$P = V^{-1} - V^{-1}X\left(X^T V^{-1} X\right)^{-1} X^T V^{-1}$ with $X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix}$.

▷ What about if the distribution of the random vectors $(\varepsilon_1, \varepsilon_2)$ is not Gaussian?

multitrait genetic animal model

Then a BLUE estimator for the fixed effect $\beta = (\beta_1, \beta_2)$ and a BLUP prediction for the BVs $\mathrm{a} = (\mathrm{a}_1, \mathrm{a}_2)$ are obtained by solving the Henderson's equations:

$$\begin{bmatrix} \mathrm{X}'\mathrm{X} & \mathrm{X}'\mathrm{Z}' \\ \mathrm{Z}'\mathrm{X} & \mathrm{Z}'\mathrm{Z} + \hat{\mathrm{G}}^{-1} \otimes \mathrm{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathrm{a}} \end{bmatrix} = \begin{bmatrix} \mathrm{X}'\mathrm{y} \\ \mathrm{Z}'\mathrm{y} \end{bmatrix}.$$

Selection of the animals for breeding is done in such a way to maximizing a linear combination of the EBVs for the two traits.

▷ In this topic, we studied the robustness of the REML estimations which assumes the normality for the multivariate traits, face to a non-Gaussian dependence structure (copula) for the residuals.

Copulas

Theorem of [Sklar(1959)]

Let $\boldsymbol{X} = (X_1, X_2)$ be a 2-dimensional random vector with c.d.f. $\boldsymbol{F}$ and let F_1, F_2 be the marginal c.d.f. of $\boldsymbol{X}$ assuming continuous. Then it exists a unique function $C : [0,1]^2 \to [0,1]$ such that:

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

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

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.

Some mathematics, 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:

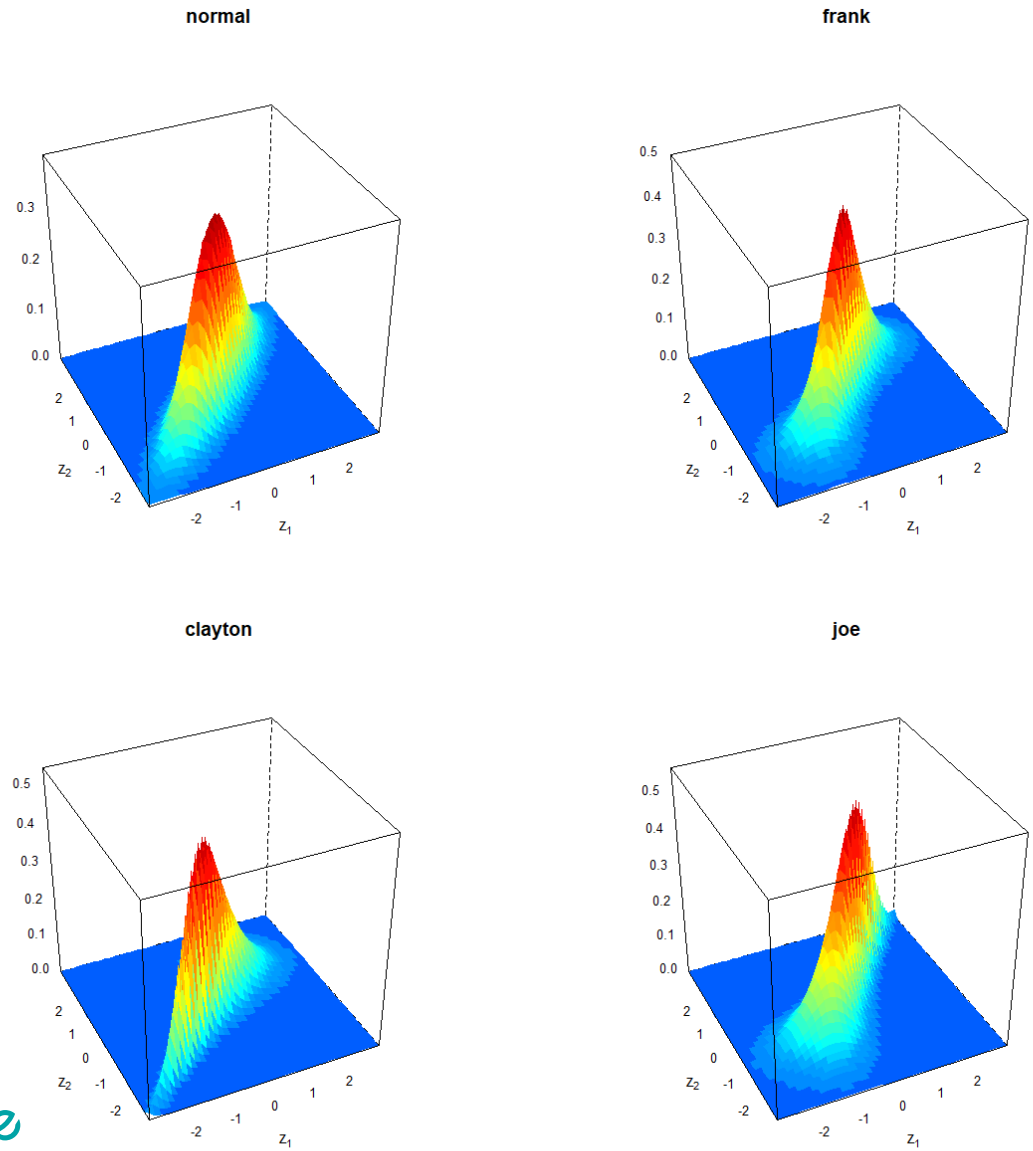
$$C_\theta^F(u,v) = \frac{1}{\theta}\log\left(1 + \frac{(\exp(-u\theta)-1)(\exp(-v\theta)-1)}{\exp(-\theta-1)}\right), \quad \theta \in \mathbb{R}^\star,$$

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

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

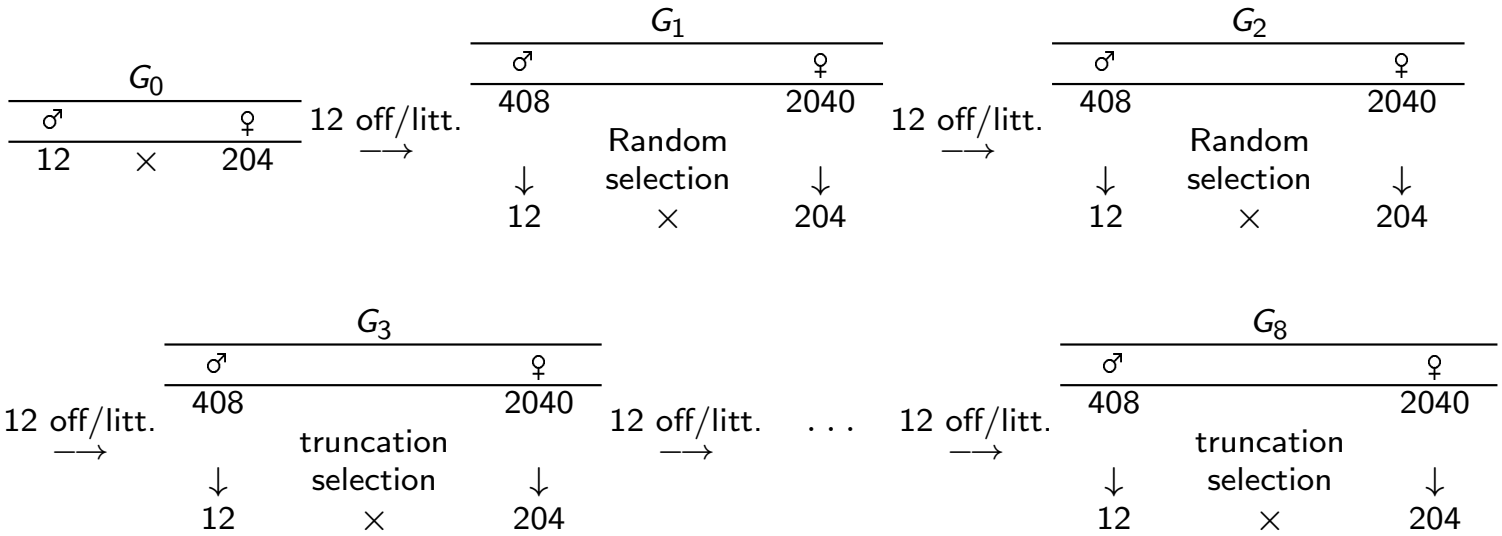
In the simulations, we considered residual from bivariate distribution with Gaussian margins and Gaussian, Frank, Clayton and Joe copula with kendall's correlation of $\tau = 0.7$.

Surface plots of bivariate distributions with Gaussian margins and several copula



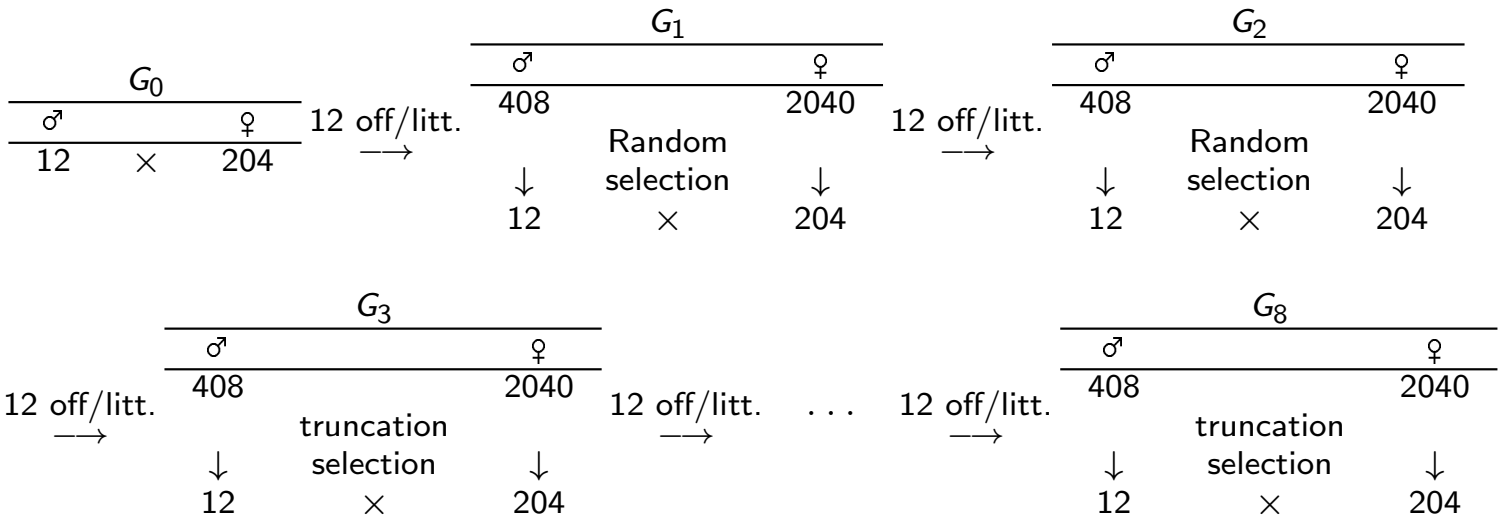
Plan

Population undergoing selection



- unrelated animals in G_0
- each female produced 12 offspring: 2 males and 10 females
- All progeny were candidates for selection

Population

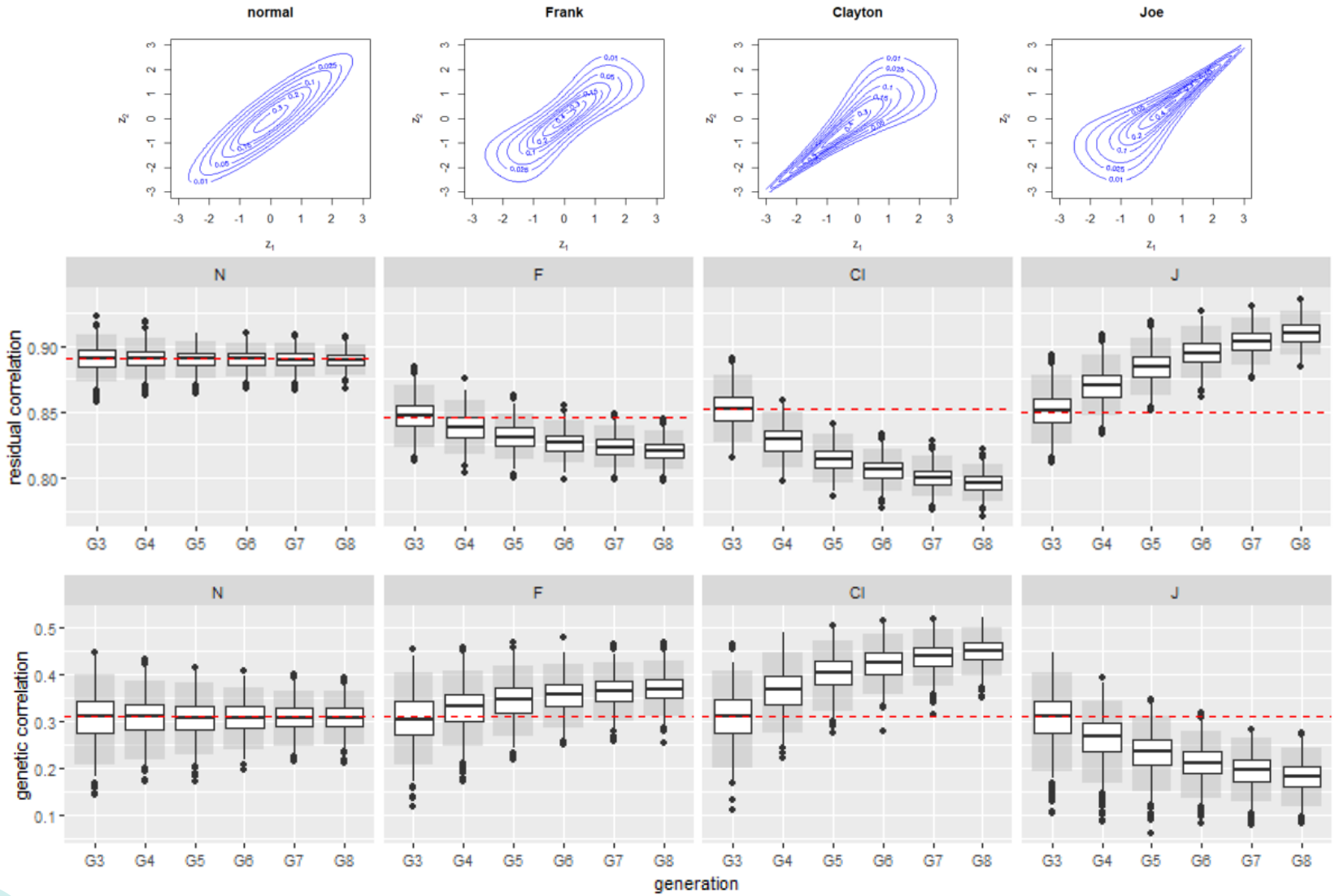


Selection:

- Selection was made among the offspring of one male
- $G_1 - G_3$ random selection
- $G_4 - G_8$ selection from a combination of their EBV (BLUP prediction)
- selection rate: 2.9% for the males and 10% for the females

Plan

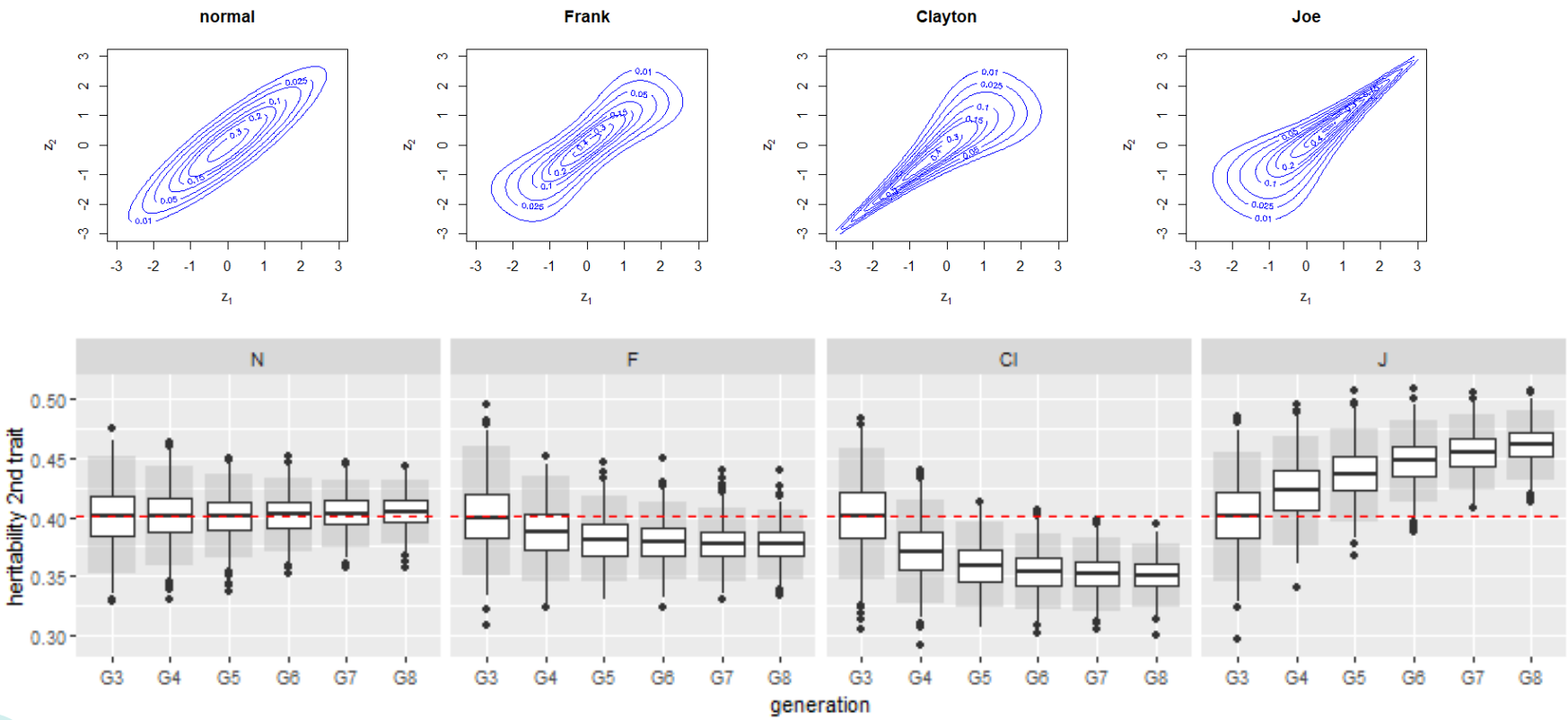
1000 runs, Estim. correlations, $h_1^2 = h_2^2 = 0.40, \rho_a = 0.31$



1000 runs, Esti. heritability

$h_1^2 = 0.15, h_2^2 = 0.40, \rho_a = 0.59$

$$h_j^2 = \frac{\sigma_{a_j}^2}{\sigma_{a_j}^2 + \sigma_{e_j}^2}, \quad j = 1, 2$$



Conclusion

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 for the trait with medium heritability
     - Significant impact on residual correlations;
     - medium biases (but non-significant) on genetic correlations;
     - no impact on low heritability