Copula miss-specification in REML multivariate genetic animal model estimation

DINAMIC

joint work with
Tom Rohmer, Anne Ricard & Ingrid David

7 nov 2022

Plan

Introduction

Simulation

Results

Plan

Introduction

Simulation

Results

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.html

Hamelin, G., Rohmer, T., Gaillard, C.
Amélioration d'un outil de détection d'évènements techniques ou sanitaires à l'échelle de plusieurs bandes de truies
Poster Journées de la recherche Porcine, 2023 (JRP23)

- Echelle multivariée, une perturbation peut affecter un/des phénotypes et/ou l'interaction entre les phénotypes (copules)

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:

Phenotypic observation

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

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:

$$\begin{cases} y_1 = X_1\beta_1 + Za_1 + \varepsilon_1 \\ y_2 = X_2\beta_2 + Za_2 + \varepsilon_2 \end{cases},$$

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

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \sim \mathcal{N}(0, G \otimes A),$$

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

$$\boldsymbol{F}(\boldsymbol{x}) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

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

Copulas

Theorem of [Sklar(1959)]

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

$$\boldsymbol{F}(\boldsymbol{x}) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

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

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

Copulas

Theorem of [Sklar(1959)]

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

$$\boldsymbol{F}(\boldsymbol{x}) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad \boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

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

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

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.

Some mathematics, copulas

https://testmyshinyapply.shinyapps.io/Shiny_copula

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:

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

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

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

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

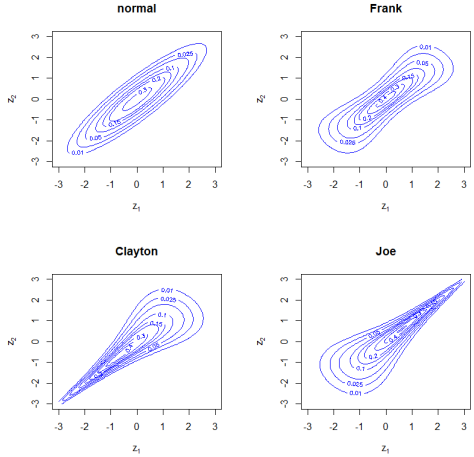
$$\tau = 4 \int_{[0,1]^2} C_\theta(u, v) dC_\theta(u, v) - 1,$$

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

and for Gaussian distribution, we have the relation with the Pearson's correlation ρ_P

$$\rho_P = \sin(\frac{\pi}{2}\tau).$$

Contour plots of bivariate distributions with Gaussian margins and several copula



Example, Large-White dataset

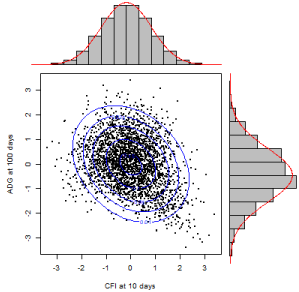
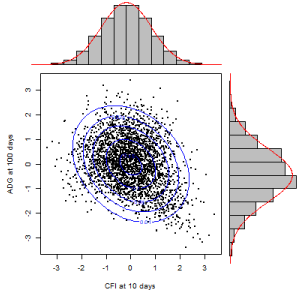


Figure: Plot of gaussian quantiles of the ranks for 2 phenotypes and contour plot of a bivariate Gaussian distribution

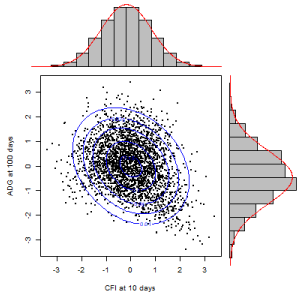
- the hypothesis of bivariate normality for the bivariates phenotypes seems to be unrealistic.
- an asymmetric bivariate distribution for the phenotypes.

Example, Large-White dataset



- Mardia's normality test based on the asymmetry (skewness) of the distribution: p-value was 10^{-6} leading to a rejection because of the asymmetry of the distribution

Example, Large-White dataset



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

Example, Large-White dataset

- Even if the marginals are Gaussian, the bivariate distribution may be non-Gaussian.
- In fact, the *copula* of the random vectors is not the Gaussian Copula

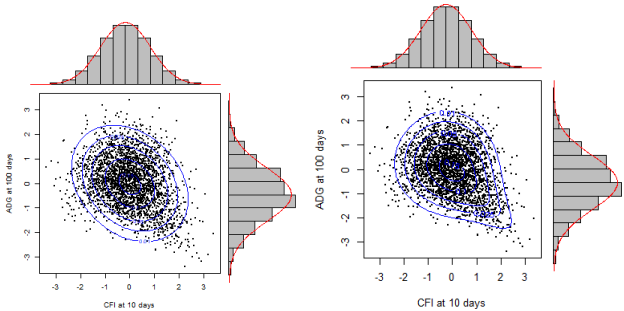


Figure: (right) contour plot of a Joe-Frank copula

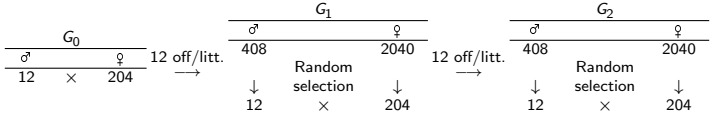
Plan

Introduction

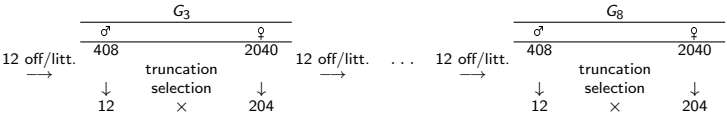
Simulation

Results

Simulation of populations undergoing selection

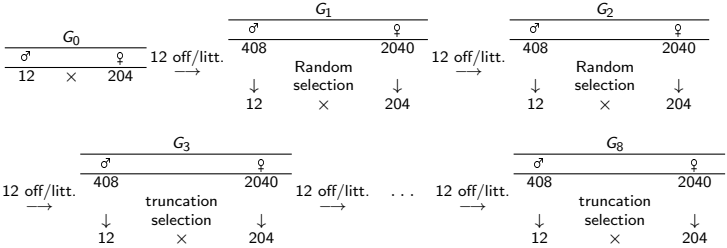


https://testmyshinyapply.shinyapps.io/Shiny_bivariate_phen/



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

Population



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

Simulated distribution:

- $(\boldsymbol{a}_1, \boldsymbol{a}_2) \sim \mathcal{N}(0, G \otimes A)$.

Phenotypes simulation

Simulated distribution:

- $(\boldsymbol{a}_1, \boldsymbol{a}_2) \sim \mathcal{N}(0, G \otimes A)$.
- $(\varepsilon_{i,1}, \varepsilon_{i,2})$, $i = 1, \ldots, n$, have standard Gaussian margins and copula C.

Copula C were Gaussian, Frank, Clayton and Joe with Kendall's correlation to 0.4 or 0.7 The phenotypes vector $\boldsymbol{y}_j = (y_{1,j}, \ldots, y_{n,j})$, $j = 1, 2$ were obtained following the bivariate animal model:

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

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

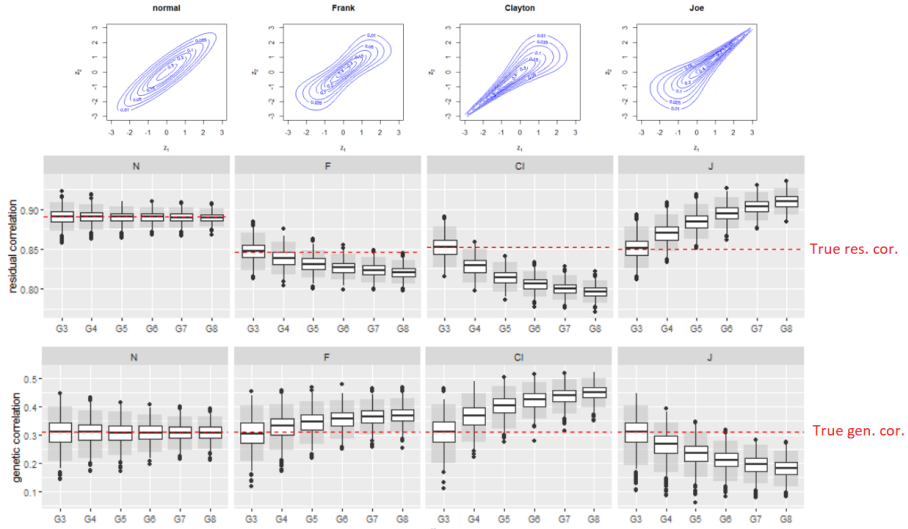
Plan

Introduction

Simulation

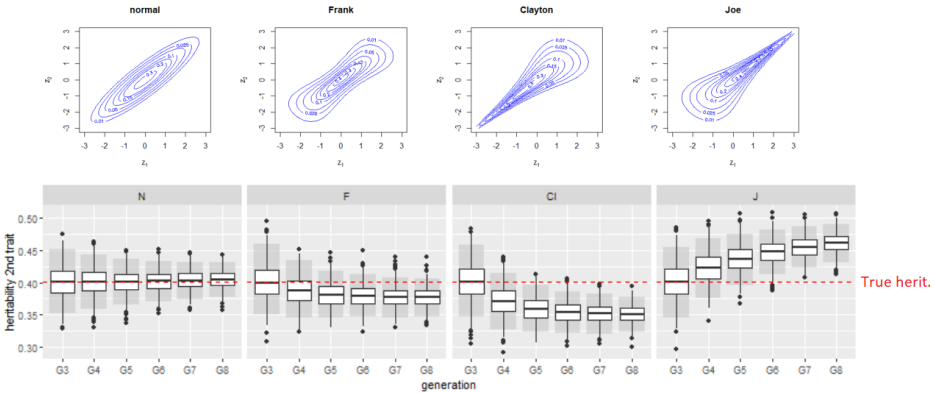
Results

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



True parameters					Estimated parameters							
					genetic correlations				residual correlations			
h_1^2	h_2^2	τ_a	τ_e		N	F	CI	J	N	F	CI	J
0.153	0.153	0.4	0.4	bias	-0.001	0.015	0.052	-0.060	-0.000	-0.002	-0.006	0.007
				SE	0.043	0.042	0.038	0.049	0.006	0.007	0.007	0.007
0.153	0.401	0.4	0.4	bias	0.001	0.013	0.039	-0.033	-0.000	-0.004	-0.013	0.015
				SE	0.032	0.032	0.029	0.035	0.008	0.008	0.008	0.008
0.401	0.401	0.4	0.4	bias	-0.000	0.021	0.068*	-0.101*	-0.000	-0.009	-0.029*	0.038*
				SE	0.025	0.024	0.022	0.031	0.010	0.011	0.011	0.012
0.153	0.153	0.4	0.7	bias	-0.003	0.034	0.089*	-0.108*	-0.000	-0.004	-0.010*	0.013*
				SE	0.035	0.034	0.031	0.044	0.003	0.003	0.003	0.005
0.153	0.401	0.4	0.7	bias	-0.001	0.003	0.030	0.039	-0.000	-0.009*	-0.023*	0.011*
				SE	0.027	0.028	0.026	0.026	0.004	0.005	0.005	0.005
0.401	0.401	0.4	0.7	bias	-0.001	0.049*	0.106*	-0.128*	-0.001	-0.018*	-0.040*	0.052*
				SE	0.021	0.020	0.018	0.027	0.005	0.006	0.006	0.008

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



True parameters					Estimated heritabilities							
					Trait 1				Trait 2			
h_1^2	h_2^2	τ_a	τ_e		N	F	Cl	J	N	F	Cl	J
0.153	0.153	0.4	0.4	bias	0.003	0.004	0.008	-0.002	0.002	0.005	0.009	-0.002
				SE	0.012	0.013	0.012	0.013	0.012	0.013	0.013	0.012
0.153	0.401	0.4	0.4	bias	0.001	0.005	0.013	-0.010	0.004	-0.001	-0.009	0.029
				SE	0.013	0.014	0.014	0.012	0.016	0.017	0.015	0.017
0.401	0.401	0.4	0.4	bias	0.003	0.011	0.020	-0.003	0.003	0.012	0.021	-0.003
				SE	0.016	0.017	0.016	0.018	0.017	0.017	0.017	0.018
0.153	0.153	0.4	0.7	bias	0.002	0.002	0.002	0.010	0.002	0.003	0.003	0.009
				SE	0.012	0.012	0.011	0.015	0.012	0.012	0.012	0.014
0.153	0.401	0.4	0.7	bias	0.001	0.004	0.012	-0.006	0.004	-0.024	-0.051*	0.059*
				SE	0.014	0.015	0.015	0.012	0.014	0.015	0.014	0.015
0.401	0.401	0.4	0.7	bias	0.004	0.006	0.008	0.020	0.003	0.005	0.009	0.020
				SE	0.016	0.016	0.016	0.017	0.016	0.016	0.016	0.018

Rotated 270° copula

True parameters					Estimated heritability							
					Trait 1				Trait 2			
h_1^2	h_2^2	τ_a	τ_e		N	F	Cl	J	N	F	Cl	J
0.153	0.153	0.4	-0.4	bias	0.002	-0.004	-0.012	0.007	0.002	-0.004	0.007	-0.013
				SE	0.010	0.010	0.009	0.012	0.010	0.010	0.011	0.009
0.153	0.401	0.4	-0.4	bias	0.002	-0.003	-0.010	0.005	0.004	-0.010	0.028	-0.034*
				SE	0.011	0.011	0.009	0.012	0.013	0.014	0.015	0.013
0.401	0.401	0.4	-0.4	bias	0.004	-0.007	-0.028*	0.027	0.004	-0.007	0.026	-0.030*
				SE	0.014	0.015	0.014	0.016	0.013	0.014	0.015	0.013
0.153	0.153	0.4	-0.7	bias	0.002	-0.009	-0.024*	0.010	0.002	-0.009	0.011	-0.025*
				SE	0.009	0.009	0.008	0.012	0.008	0.009	0.011	0.008
0.153	0.401	0.4	-0.7	bias	0.002	-0.004	-0.017*	0.016	0.005	-0.013	0.032*	-0.047*
				SE	0.010	0.010	0.009	0.013	0.010	0.012	0.014	0.011
0.401	0.401	0.4	-0.7	bias	0.005	-0.006	-0.038*	0.038*	0.004	-0.006	0.037*	-0.039*
				SE	0.013	0.013	0.012	0.015	0.012	0.013	0.014	0.011

Result

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

Perspective

- parametric model $Y = X\beta + ZU + \varepsilon$, where $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})$ have c.d.f. $F_e(x, y) = C_\theta(\Phi_1(x), \Phi_2(y))$.
  - For $j = 1, 2$, Φ_j are the c.d.f. of the centred normal distribution with variance $\sigma^2_{e,j}$
  - C_θ is a parametric copula function (known) with parameter θ (to estimate)
  - joint estimation of G, $\sigma^2_{e,1}, \sigma^2_{e,2}$ and θ