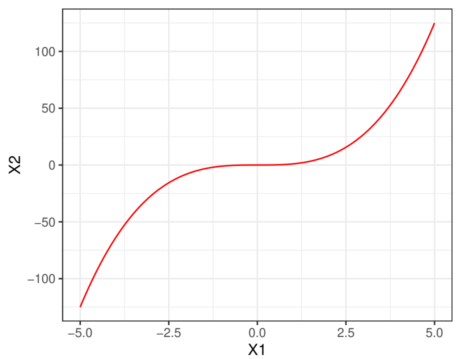
Supplementary material for A semi-parametric Gaussian copula model for heterogeneous network inference : an application to multi-omics data by Tomilina, Mazo and Jaffrézic

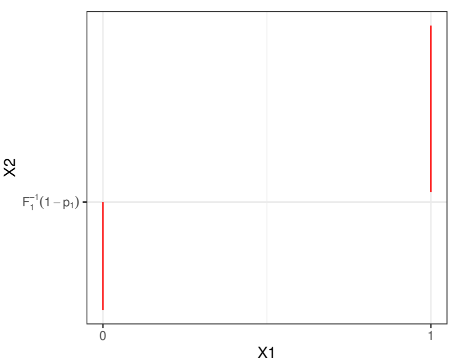
December 19, 2024

Contents

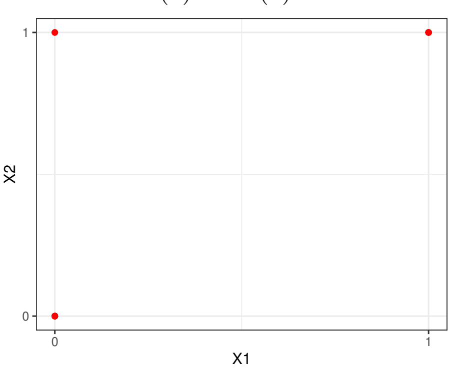
A Supplementary Figures



(a) Case (i)

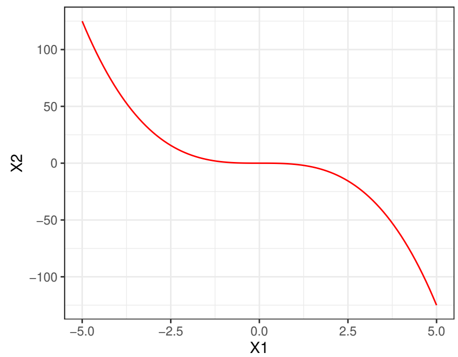


(b) Case (ii)

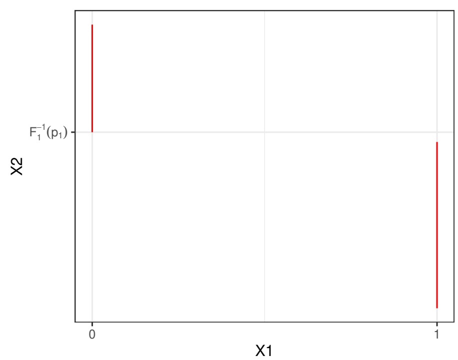


(c) Case (iii)

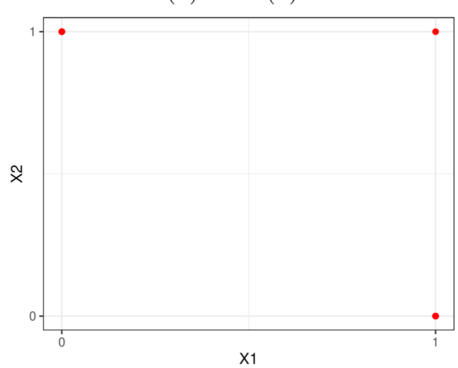
Figure S1: Illustration of comonotonicity for cases (i), (ii) and (iii) of Proposition 3



(a) Case (i)



(b) Case (ii)



(c) Case (iii)

Figure S2: Illustration of countermonotonicity for cases (i), (ii) and (iii) of Proposition 3

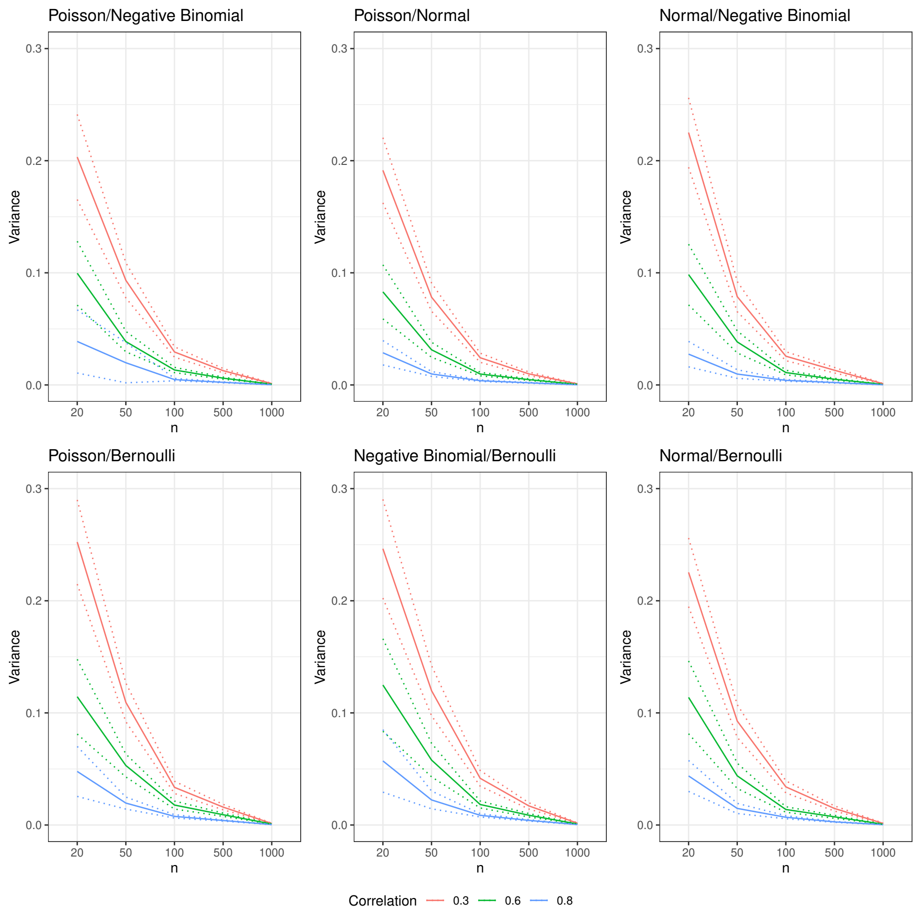


Figure S3: Averaged variances and 95% confidence intervals (for N=500 replications) of the Gaussian copula correlation coefficient estimators defined in (3) for $\rho = 0.3, 0.6, 0.8$ between $P(1)$, $NB(1, \frac{1}{2})$, $\mathcal{N}(0,1)$ and $\mathcal{B}(\frac{1}{2})$ depending on the sample size.

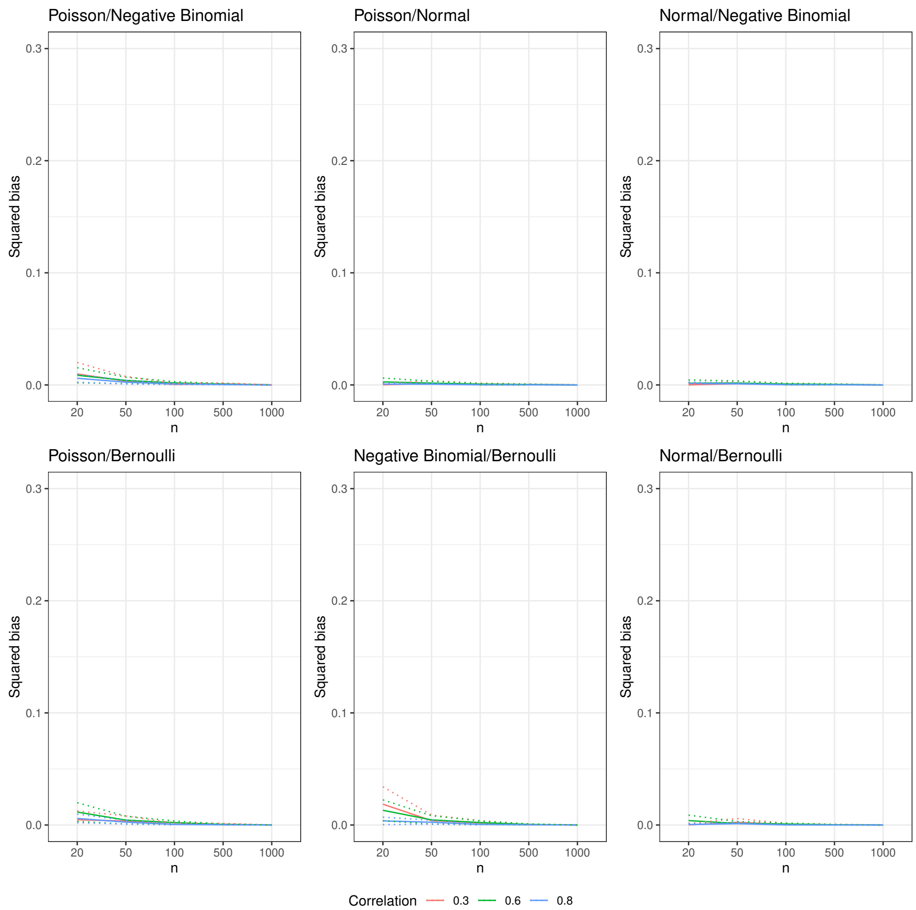


Figure S4: Averaged squared biases and 95% confidence intervals (for N=500 replications) of the Gaussian copula correlation coefficient estimators defined in (3) for $\rho = 0.3, 0.6, 0.8$ between $P(1)$, $NB(1, \frac{1}{2})$, $\mathcal{N}(0, 1)$ and $\mathcal{B}(\frac{1}{2})$ depending on the sample size.

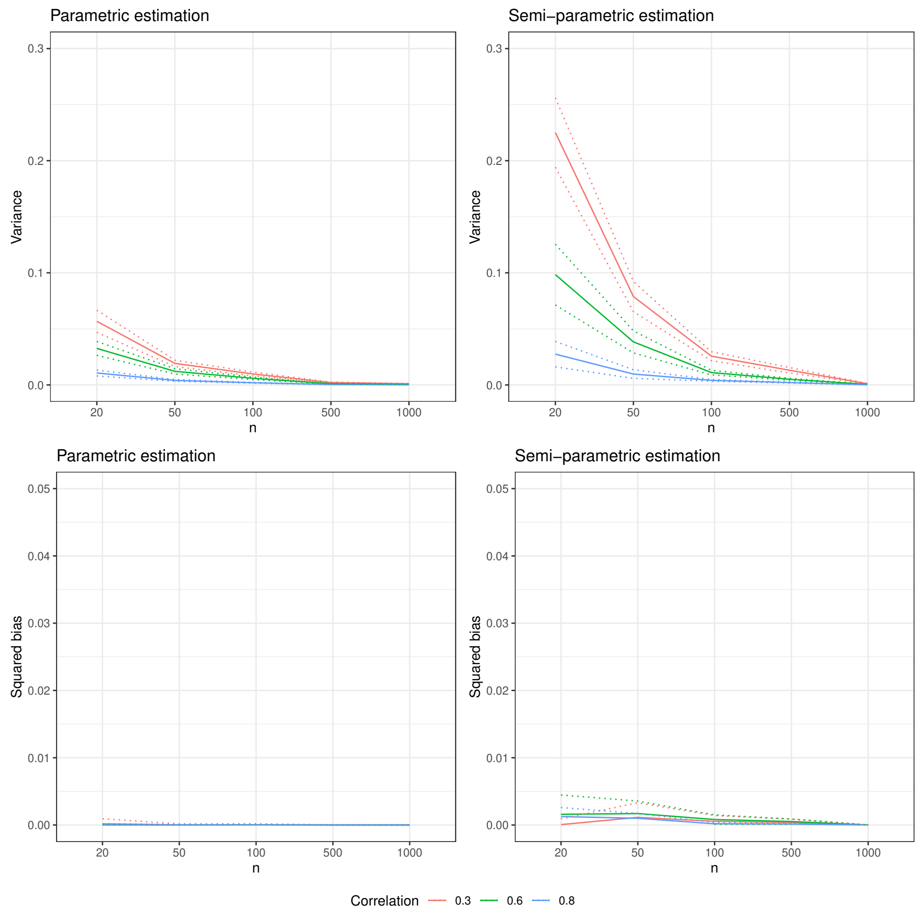


Figure S5: Averaged variances and squared biases with 95% confidence intervals of $\hat{\rho}$ obtained with the semi-parametric method and the parametric method when the marginals are correctly specified for different values of ρ and sample sizes.

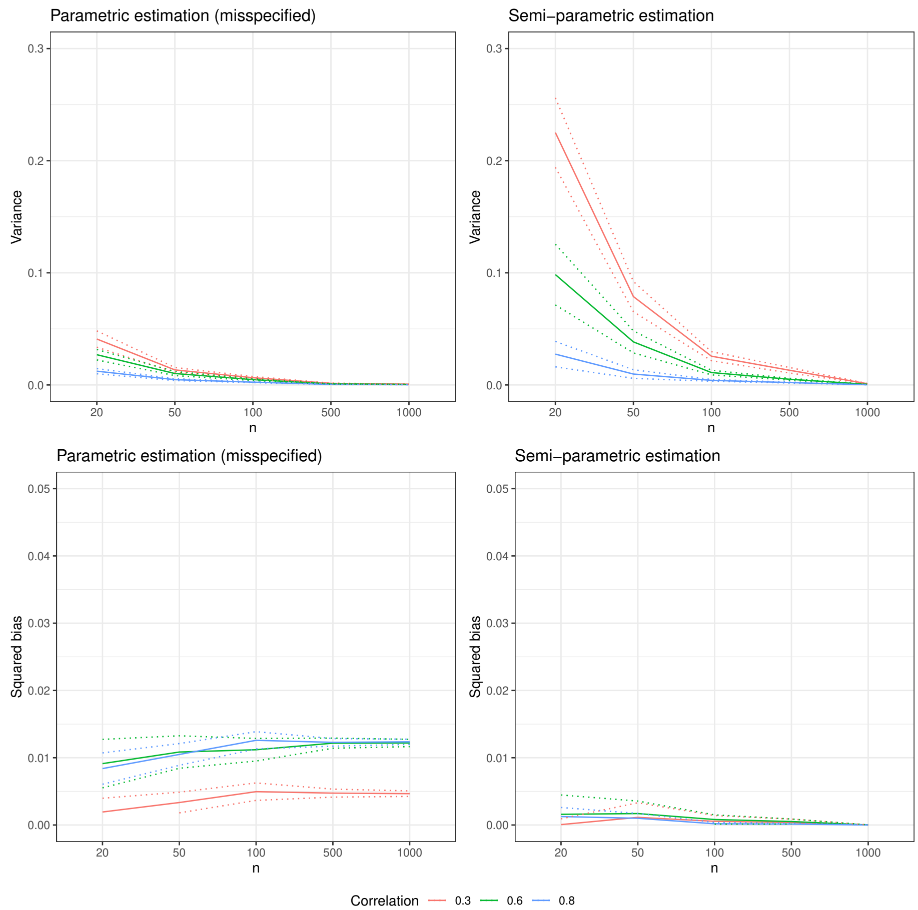


Figure S6: Averaged variances and squared biases with 95% confidence intervals of $\hat{\rho}$ obtained with the semi-parametric method and parametric method when the marginals are misspecified for different values of ρ and sample sizes.

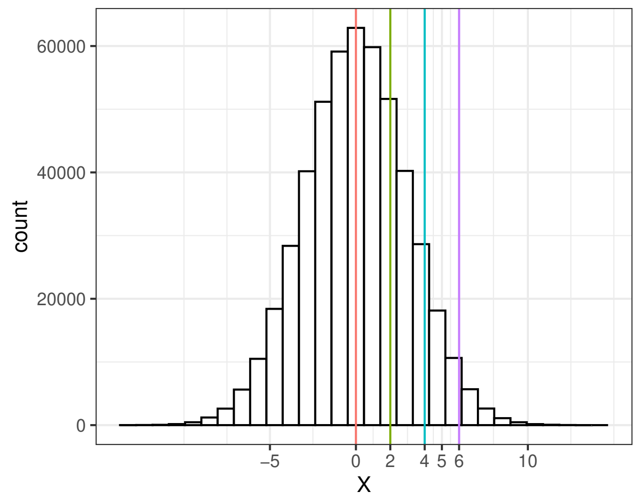


Figure S7: Distribution of variable $X \sim \mathcal{N}(0, 3)$ for 500 replications of sample size n=1000, with threshold lines at $t = 0, 2, 4, 6$ in order to visualize the associated binary variable $Y = \mathbb{1}_{\{X \geq t\}}$.

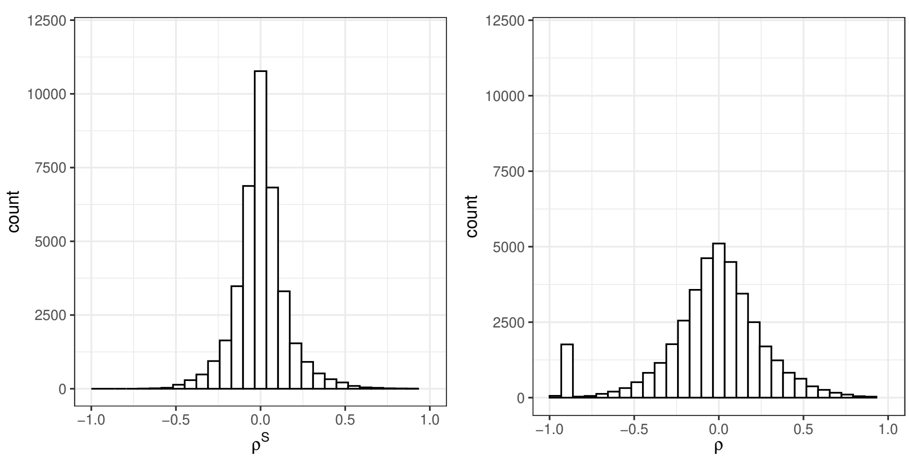


Figure S8: Histograms of the estimated correlation coefficients for Spearman's ρ^S (left) and the copula (right)

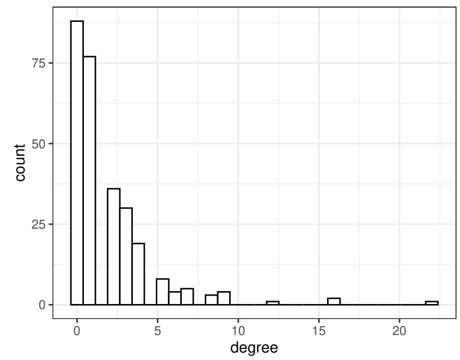


Figure S9: Histogram of the degrees of the nodes from the copula correlation network for a threshold value of 0.7, represented in Figure 6.

B Supplementary Tables

Threshold	0	2	4	6
Bernoulli parameter	0.5	0.25	0.09	0.02

Table S1: Bernoulli parameter estimated for $Y = \mathbb{1}_{\{X \geq t\}}$, for $t = 0, 2, 4, 6$, averaged over $N = 500$ simulations. These values correspond to the standardized areas under the histogram on the right of threshold values t in Figure S7.

sample size	20	50	100	500	1000
Σ_{blocks}	0.329 (0.02)	0.188 (0.02)	0.127 (0.01)	0.054 (0.004)	0.038 (0.003)
$\Sigma_{0.2}$	0.315 (0.02)	0.187 (0.01)	0.127 (0.01)	0.054 (0.004)	0.038 (0.003)
$\Sigma_{0.5}$	0.332 (0.01)	0.195 (0.01)	0.132 (0.01)	0.056 (0.004)	0.039 (0.002)
$\Sigma_{0.8}$	0.349 (0.02)	0.2 (0.01)	0.133 (0.01)	0.057 (0.006)	0.04 (0.002)

(a) Root Mean Squared Error (RMSE)

sample size	20	50	100	500	1000
Σ_{blocks}	0.265 (0.02)	0.148 (0.01)	0.102 (0.01)	0.043 (0.005)	0.03 (0.003)
$\Sigma_{0.2}$	0.262 (0.02)	0.15 (0.01)	0.101 (0.01)	0.043 (0.003)	0.03 (0.007)
$\Sigma_{0.5}$	0.279 (0.02)	0.155 (0.01)	0.105 (0.007)	0.04 (0.003)	0.031 (0.002)
$\Sigma_{0.8}$	0.283 (0.01)	0.16 (0.008)	0.107 (0.005)	0.045 (0.002)	0.032 (0.002)

(b) Mean Average Error (MAE)

Table S2: Average normalized Root Mean Squared Error (a) and normalized Mean Absolute Error (b) values for $N = 500$ replications for the copula correlation pairwise estimator for $d = 30$ variables, for a block-wise matrix and for three different matrices of respective sparsity $\gamma_F = 0.2, 0.5, 0.8$, for different sample sizes. The simulation standard deviations are specified in parentheses.

sample size	20	50	100	500
RMSE	0.354 (0.004)	0.202 (0.002)	0.136 (0.001)	0.058 (0.0005)
MAE	0.294 (0.003)	0.162 (0.001)	0.108 (0.0007)	0.046 (0.0004)

Table S3: Average normalized Root Mean Squared Error and normalized Mean Absolute Error values for $N = 500$ replications for the copula correlation pairwise estimator for $d = 300$ variables, for a matrix of sparsity $\gamma_F = 0.8$, for different sample sizes. The simulation standard deviations are specified in parentheses.

	Predicted zero	Predicted non-zero
Real zero	True negatives (TN)	False positives (FP)
Real non-zero	False negatives (FN)	True positives (TP)

Table S4: Contingency matrix

p	sample size	20	50	100	500	1000
30	Σ_{blocks}	0.91 (0.04)	0.97 (0.01)	0.99 (0.007)	0.999 (0.0004)	0.999 (0.0005)
30	$\Sigma_{0.2}$	0.72 (0.03)	0.82 (0.02)	0.89 (0.02)	0.96 (0.01)	0.98 (0.004)
30	$\Sigma_{0.5}$	0.76 (0.04)	0.88 (0.03)	0.94 (0.01)	0.99 (0.003)	0.994 (0.002)
30	$\Sigma_{0.8}$	0.84 (0.03)	0.94 (0.02)	0.98(0.01)	0.999 (0.001)	0.999(0.0004)
300	$\Sigma_{0.8}$	0.79 (0.01)	0.90 (0.007)	0.97 (0.004)	0.998 (0.0002)	NE

Table S5: Average AUC values for $N = 500$ simulations for the copula pairwise correlation coefficients for different sample sizes. Four different sparse matrices were evaluated for $d = 30$ variables (block-wise matrix, final sparsity $\gamma_F = 0.2, 0.5, 0.8$). A matrix of sparsity $\gamma_F = 0.8$ was considered for $d = 300$ variables. The standard deviations are specified in parentheses, and the NE acronym stands for Not Evaluated.

	0	1
0	437156	5723
1	7621	0

(a)

	0	1
0	248	0
1	0	2

(b)

	0	1
0	213	0
1	35	2

(c)

Table S6: Contingency tables for the mutation variables corresponding to the points for which the copula correlation coefficient was close to -1 (S6a) (1802 points), for which both the copula and Spearman had correlation coefficients of 1 (S6b) (one point), and for which the copula had a correlation coefficient close to 1 and Spearman close to 0 (S6c) (one point).

Variable	Gene
MU4777833	ARHGEF11
MU5153080	SLC7A9
MU17289	CDKN1B
MU5551967	PQBP1

Table S7: Mutation variables with a degree greater than 10 in the copula correlation network for a threshold value of 0.7, and their corresponding genes.

C Proofs of Propositions

C.1 Proof of Proposition 1

We are going to show that $f(x_1, ..., x_d)$ from equation (2) corresponds to the density of the joint cumulative distribution $F(x_1, \ldots, x_d)$ from model (1) with respect to the $\lambda^{\otimes p} \otimes \mu^{\otimes(d-p)}$ measure, where $\lambda^{\otimes p} = \underbrace{\lambda \times \cdots \times \lambda}_{p \text{ times}}$ with λ the Lebesgue measure and $\mu^{\otimes(d-p)} = \underbrace{\mu \times \cdots \times \mu}_{d-p \text{ times}}$ denotes the counting measure. For easier notations, we denote $S^{a,b}$, where $a < b$ and $a, b \in \{1, \ldots, d\}$, the set $]-\infty; x_a] \times ... \times]-\infty; x_b]$. Let us find f that satisfies:

$$\begin{aligned} C_\Sigma(F_1(x_1), ..., F_d(x_d)) &= \int_{S^{1,d}} f(y_1, ..., y_d) \mathrm{d}(\lambda^{\otimes p} \otimes \mu^{\otimes(d-p)})(y_1, ..., y_d) \\ &= \int_{S^{1,p}} \left(\int_{S^{p+1,d}} f(y_1, ..., y_d) \mathrm{d}\mu^{\otimes(d-p)}(y_{p+1}, ..., y_d) \right) \mathrm{d}\lambda^{\otimes p}(y_1, ..., y_p). \end{aligned}$$

The second equality can be directly obtained by Fubini-Tonelli's theorem. By differentiating both sides with respect to the p continuous variables, we get:

$$\begin{aligned} \prod_{k=1}^{p} f_k(x_k) C_\Sigma^p(F_1(x_1), ..., F_d(x_d)) &= \int_{S^{p+1,d}} f(x_1, ..., x_p, y_{p+1} ..., y_d) \mathrm{d}\mu^{\otimes(d-p)}(y_{p+1}, ..., y_d) \\ &= \sum_{\substack{y_{p+1} \leq x_{p+1} \\ y_{p+1} \in S_{p+1}}} \cdots \sum_{\substack{y_d \leq x_d \\ y_d \in S_d}} f(x_1, ..., x_p, y_{p+1}, ... y_d) \end{aligned}$$

where f_j denotes the density of X_j, C_Σ^p denotes the differential of C_Σ with respect to the p continuous variables, and S_j denotes the support of F_j. The Möbius inversion formula (Rota, 1964) provides us with the following expression

$$\begin{aligned} f(x_1, ..., x_d) = & \sum_{\substack{y_{p+1} \leq x_{p+1} \\ y_{p+1} \in S_{p+1}}} \cdots \sum_{\substack{y_d \leq x_d \\ y_d \in S_d}} \prod_{k=1}^{p} f_k(x_k) C_\Sigma^p(F_1(x_1), ..., F_p(x_p), F_{p+1}(y_{p+1}), ..., F_d(y_d)) \\ & \times m_{p+1}(y_{p+1}, x_{p+1}) \cdots m_d(y_d, x_d) \end{aligned}$$

where

$$m_j(y_j, x_j) = \begin{cases} 1 & \text{if } x_j = y_j \\ -1 & \text{if } y_j \text{ precedes } x_j \text{ in } S_j \\ 0 & \text{otherwise.} \end{cases}$$

In our case, one can notice that y_j can only take the values x_j and x_j-, where x_j- denotes the point that precedes x_j in S_j. It corresponds to the expression of the multivariate density below where

$$u_{k,j_k} = \begin{cases} F_k(x_k) & \text{if } j_k = 0 \\ F_k(x_k-) & \text{if } j_k = 1 \end{cases}$$

$$f(x_1, \dots, x_d) = \prod_{k=1}^{p} f_k(x_k) \sum_{j_{p+1}=0}^{1} \dots \sum_{j_d=0}^{1} (-1)^{j_{p+1}+\dots+j_d} \times C_\Sigma^p(F_1(x_1), ..., F_p(x_p), u_{p+1,j_{p+1}}, ..., u_{d,j_d})$$

The density is unique up to all sets of measure zero with respect to our measure $\lambda^{\otimes p} \otimes \mu^{\otimes d-p}$.

C.2 Proof of Proposition 2

Let us show first that if the correlation matrix of the copula is of the form

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \Sigma_k \end{pmatrix}$$

then the multivariate density can be factorized as

$$f(x_1, \dots, x_d) = \prod_{i=1}^{k} g_i(x_{G_i}) \tag{1}$$

for some functions g_i, $i = 1, \dots, k$. We know that for all $(u_1, \dots, u_d) \in (0,1)^d$,

$$\frac{\partial^p C_\Sigma(u_1, \dots, u_d)}{\partial u_1 \dots \partial u_p} = \int_0^{u_{p+1}} \cdots \int_0^{u_d} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{v}^T(\Sigma^{-1} - I)\mathbf{v}\right) \mathrm{d}q_{p+1} \dots \mathrm{d}q_d \tag{2}$$

where $\mathbf{v} = (v_1, \dots, v_d)$ such that

$$v_i = \begin{cases} \Phi^{-1}(u_i) \text{ if } i \in \{1, \dots, p\} \\ \Phi^{-1}(q_i) \text{ if } i \in \{p+1, \dots, d\} \end{cases}$$

and I denotes the identity matrix. Let us split the vector $\mathbf{v}$ as $\mathbf{v}^T = (\mathbf{v}_1^T, \dots, \mathbf{v}_k^T)$ where each $\mathbf{v}_l$ is of size $|G_l| \times 1$. Hence, the right-hand side of (2) can be written as

$$\int_0^{u_{p+1}} \cdots \int_0^{u_d} \left(\prod_{l=1}^{k} |\Sigma_l|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{v}_l^T(\Sigma_l^{-1} - I|_{G_l})\mathbf{v}_l\right)\right) \mathrm{d}q_{p+1} \dots \mathrm{d}q_d \tag{3}$$

where $I|_{G_l}$ denotes the identity matrix of size $|G_l|$. By noticing that the lth factor only depends on $\mathbf{v}_l$, and that each q_j can only belong to exactly one $\mathbf{v}_l$, Fubini-Tonelli's theorem enables us to factor (2). Let us denote by $\mathcal{D}(G_l)$ the set of indexes corresponding to the discrete variables in G_l, that is $\mathcal{D}(G_l) = \{p+1,\dots,d\} \cap G_l$. Denote $d_l = |\mathcal{D}(G_l)|$. We get that (3) can be written as

$$\prod_{l=1}^{k} \int_0^{u_{j_1(l)}} \cdots \int_0^{u_{j_{d_l}(l)}} |\Sigma_l|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{v}_l^T(\Sigma_l^{-1} - I|_{G_l})\mathbf{v}_l\right) \mathrm{d}q_{j_{d_l}(l)} \cdots \mathrm{d}q_{j_1(l)},$$

where above $j_1(l),\dots,j_{d_l}(l)$ is an enumeration of the elements of $\mathcal{D}(G_l)$. We use the convention that an integral over the empty set $\mathcal{D}(G_l) = \emptyset$ is replaced by its integrand. Let us define

$$P_l(\mathbf{u}_{\mathcal{C}(G_l)}, \mathbf{u}_{\mathcal{D}(G_l)}) = \int_0^{u_{j_1(l)}} \cdots \int_0^{u_{j_{d_l}(l)}} |\Sigma_l|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{v}_l^T(\Sigma_l^{-1} - I|_{G_l}))\mathbf{v}_l\right) \mathrm{d}q_{j_{d_l}(l)} \cdots \mathrm{d}q_{j_1(l)},$$

where $\mathcal{C}(G_l)$ is the set of indexes corresponding to the continuous variables in group G_l, $\mathbf{u}_{\mathcal{D}(G_l)} = \{u_j : j \in \mathcal{D}(G_l)\}$ and $\mathbf{u}_{\mathcal{C}(G_l)} = \{u_j : j \in \mathcal{C}(G_l)\}$. Thus, remembering the notation $C_\Sigma^p(u_1,\dots,u_d) \equiv \partial^p C/\partial u_1 \cdots \partial u_p$, we have got

$$C_\Sigma^p(u_1,\dots,u_d) = \prod_{l=1}^{k} P_l(\mathbf{u}_{\mathcal{C}(G_l)}, \mathbf{u}_{\mathcal{D}(G_l)}) \tag{4}$$

for all $(u_1,\dots,u_d) \in \mathbb{R}^d$. Choose and fix $(x_1,\dots,x_d) \in \mathbb{R}^d$. Remember that we want to prove (1) for some functions g_i. From (2) in the main text, we have

$$f(x_1,\dots,x_d) =$$
$$\left(\prod_{k=1}^{p} f_k(x_k)\right)\left(\sum_{\alpha_{p+1}=0}^{1} \cdots \sum_{\alpha_d=0}^{1} (-1)^{\alpha_{p+1}+\cdots+\alpha_d} C_\Sigma^p(\zeta(1,0),\dots,\zeta(p,0),\zeta(p+1,\alpha_{p+1}),\dots,\zeta(d,\alpha_d))\right),$$

where each ζ is seen as a function on $\{p+1,\dots,d\} \times \{0,1\}$ such that $\zeta(i,0) = F_i(x_i)$ and $\zeta(i,1) = F_i(x_i-)$. By (4), we have

$$C_\Sigma^p(\zeta(1,0),\dots,\zeta(p,0),\zeta(p+1,\alpha_{p+1}),\dots,\zeta(d,\alpha_d))$$
$$= \prod_{l=1}^{k} P_l(\zeta(i_1(l),0),\dots,\zeta(i_{c_l}(l),0),\zeta(j_1(l),\alpha_{j_1(l)}),\dots,\zeta(j_{d_l}(l),\alpha_{j_{d_l}(l)})) =: \prod_{l=1}^{k} P_l(\boldsymbol{\alpha}_{\mathcal{D}(G_l)}),$$

where above $i_1(l), \dots, i_{c_l}(l)$ is an enumeration of the elements of $\mathcal{C}(G_l)$ with $c_l = |\mathcal{C}(G_l)|$. It follows

$$\begin{aligned}
f(x_1, \dots, x_d) &= \prod_{i=1}^{p} f_i(x_i) \sum_{\alpha_{p+1}=0}^{1} \cdots \sum_{\alpha_d=0}^{1} (-1)^{\alpha_{p+1}+\cdots+\alpha_d} \prod_{l=1}^{k} P_l(\boldsymbol{\alpha}_{\mathcal{D}(G_l)}) \\
&= \left(\prod_{l=1}^{k} \prod_{i \in \mathcal{C}(G_l)} f_i(x_i) \right) \sum_{\alpha_{p+1}=0}^{1} (-1)^{\alpha_{p+1}} \cdots \sum_{\alpha_d=0}^{1} (-1)^{\alpha_d} \prod_{l=1}^{k} P_l(\boldsymbol{\alpha}_{\mathcal{D}(G_l)}) \\
&= \prod_{l=1}^{k} \left(\prod_{i \in \mathcal{C}(G_l)} f_i(x_i) \sum_{\substack{\alpha_m=0 \\ m \in \mathcal{D}(G_l)}}^{1} (-1)^{\sum_{\alpha_m \in \mathcal{D}(G_l)} \alpha_m} P_l(\boldsymbol{\alpha}_{\mathcal{D}(G_l)}) \right).
\end{aligned}$$

This proves that the multivariate density f factorizes.

Conversely, this factorization implies a k-block-diagonal structure for Σ by unicity of the density with respect to the product measure $\underbrace{\lambda \times \cdots \times \lambda}_{p \text{ times}} \times \underbrace{\mu \times \cdots \times \mu}_{d-p \text{ times}}$.

The proof of Proposition 2 is complete.

C.3 Proof of Proposition 3

Case (i) is well-known and hence not shown here. Let us show cases (ii) and (iii). Remember that for all $0 \le s \le 1$ and $t \in \mathbb{R}$, it holds $F_j(F_j^{\leftarrow}(s)) \ge s$, $F_j^{\leftarrow}(s) \le t$ if and only if $s \le F_j(t)$, and $F_j(t) < s$ if and only if $t < F_j^{\leftarrow}(s)$; see e.g., Resnick (1987). Observe also that if $X_j \sim \mathcal{B}(p_j)$ then

$$F_j(X_j) = \begin{cases} 1 - p_j & \text{if } X_j = 0 \\ 1 & \text{if } X_j = 1, \end{cases} \qquad F_j^{\leftarrow}(u) = \begin{cases} 0 & \text{if } 0 < u \le 1 - p_j \\ 1 & \text{if } 1 - p_j < u \le 1, \end{cases}$$

for all $0 < u < 1$. Remember that $X_j = F_j^{\leftarrow}(\Phi(Z_j))$, $j = 1, 2$, where Z_1 and Z_2 are standard normal random variables with correlation ρ. Remember also that $Z_1 = Z_2$ almost surely (that is, with probability one) if and only if $\rho = 1$, and $Z_1 = -Z_2$ almost surely if and only if $\rho = -1$. Let $X_1 \sim \mathcal{B}(p_1)$.

Case (ii). For the first claim of the proposition, note that $(X_1, \mathbf{1}_{\{X_2 > F_2^{-1}(1-p_1)\}})$ is comonotonic if and only if $\mathbb{P}(X_1 = 0, X_2 > F_2^{-1}(1 - p_1)) + \mathbb{P}(X_1 = 1, X_2 \le F_2^{-1}(1 - p_1)) = 0$. This implies

$$\begin{aligned}
0 &= \mathbb{P}(X_1 = 0, X_2 > F_2^{-1}(1 - p_1)) \\
&= \mathbb{P}(F_1^{\leftarrow}(\Phi(Z_1)) = 0, \Phi(Z_2) > 1 - p_1) \\
&= \mathbb{P}(Z_1 \le \Phi^{-1}(1 - p_1), Z_2 > \Phi^{-1}(1 - p_1)),
\end{aligned}$$

which is false unless $\rho = 1$. Conversely, if $\rho = 1$ then $Z_1 = Z_2$ almost surely and

$$\begin{aligned}
&\mathbb{P}(X_1 = 0, X_2 \leq F_2(1-p_1)) + \mathbb{P}(X_1 = 1, X_2 > F_2^{-1}(1-p_1)) \\
&= \mathbb{P}(F_1^{\leftarrow}(\Phi(Z_1)) = 0, \Phi(Z_1) \leq 1-p_1) + \mathbb{P}(F_1^{\leftarrow}(\Phi(Z_1)) = 1, \Phi(Z_1) > 1-p_1) \\
&= \mathbb{P}(\Phi(Z_1) \leq 1-p_1) + \mathbb{P}(\Phi(Z_1) > 1-p_1) \\
&= 1.
\end{aligned}$$

For the second claim of the proposition, note that $(X_1, \mathbf{1}_{\{X_2 > F_2^{-1}(p_1)\}})$ is countermonotonic if and only if $\mathbb{P}(X_1 = 0, X_2 \leq F_2^{-1}(p_1)) + \mathbb{P}(X_1 = 1, X_2 > F_2^{-1}(p_1)) = 0$. This implies $\mathbb{P}(\Phi(Z_1) \leq 1-p_1, \Phi(Z_2) \leq p_1) = 0$, which is false unless $\rho = -1$. Conversely, if $\rho = -1$ then

$$\begin{aligned}
&\mathbb{P}(X_1 = 0, X_2 > F_2^{-1}(p_1)) + \mathbb{P}(X_1 = 1, X_2 \leq F_2^{-1}(p_1)) \\
&= \mathbb{P}(\Phi(Z_1) \leq 1-p_1, \Phi(-Z_1) > p_1) + \mathbb{P}(\Phi(Z_1) > 1-p_1, \Phi(-Z_1) \leq p_1) \\
&= \mathbb{P}(\Phi(Z_1) \leq 1-p_1) + \mathbb{P}(\Phi(Z_1) > 1-p_1) \\
&= 1.
\end{aligned}$$

This proves case (ii).

Case (iii). For the first claim of the proposition, note that $X_1 \leq X_2$ almost surely if and only if $\mathbb{P}(X_1 = 1, X_2 = 0) = 0$. But

$$\begin{aligned}
\mathbb{P}(X_1 = 1, X_2 = 0) &= \mathbb{P}(F_1^{\leftarrow}(\Phi(Z_1)) = 1, F_2^{\leftarrow}(\Phi(Z_2)) = 0) \\
&= \mathbb{P}(1-p_1 < \Phi(Z_1), \Phi(Z_2) \leq 1-p_2),
\end{aligned}$$

which is null if and only if $\rho = 1$ (because $p_1 \leq p_2$). For the second claim of the proposition, note that $X_1 + X_2 > 0$ almost surely if and only if

$$0 = \mathbb{P}(X_1 = 0, X_2 = 0) = \mathbb{P}(\Phi(Z_1) \leq 1-p_1, \Phi(Z_2) \leq 1-p_2).$$

Since $\Phi^{-1}(1-p_1) \leq -\Phi^{-1}(1-p_2)$, the latter probability is null if and only if $\rho = -1$.

D Correlation coefficient computation

Let $X \sim \mathcal{N}(0,3)$ and $Y = \mathbb{1}_{\{X \geq t\}}$ for a given threshold t. Let us show first that (X, Y) belongs to model (1). Let F and G denote the CDFs of X and Y, respectively. We know that $F(x) = \Phi(x/\sqrt{3})$ and

$$G(y) = \begin{cases} 0 & \text{if } y < 0 \\ \Phi(t/\sqrt{3}) & \text{if } 0 \leq y < 1 \\ 1 & \text{otherwise.} \end{cases}$$

It is easy to see that

$$G^{\leftarrow}(u) = \begin{cases} 0 & \text{if } 0 < u \leq \Phi(t/\sqrt{3}) \\ 1 & \text{otherwise.} \end{cases}$$

for all $u \in (0,1]$. It suffices to exhibit a standard Gaussian random vector (Z_1, Z_2) with correlation ρ such that $X = F^{\leftarrow}(\Phi(Z_1))$ and $Y = G^{\leftarrow}(\Phi(Z_2))$. But this is easily checked for $\rho = 1$.

It is known that the Pearson correlation coefficient between X and Y is given by:

$$\rho^P(X,Y) = \frac{\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]}{\sigma(X)\sigma(Y)}$$

where $\mathbb{E}(X)$ (resp. $\mathbb{E}(Y)$) denotes the expectation of X (resp. Y) and $\sigma(X)$ (resp. $\sigma(Y)$) denotes the standard deviation of X (resp. Y). We know that $\mathbb{E}(X) = 0$ and $\sigma(X) = \sqrt{3}$. Moreover, we have

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(\mathbb{1}_{\{X \geq t\}}) \\ &= \mathbb{P}(X \geq t) \\ &= 1 - \Phi(\tfrac{t}{\sqrt{3}})\end{aligned}$$

and $\sigma(Y) = \sqrt{\Phi(\frac{t}{\sqrt{3}})(1-\Phi(\frac{t}{\sqrt{3}}))}$ because we can recognize that $Y \sim \mathcal{B}(1 - \Phi(\frac{t}{\sqrt{3}}))$. Hence, we get:

$$\begin{aligned}\rho^P(X,Y) &= \frac{\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]}{\sigma(X)\sigma(Y)} \\ &= \frac{\mathbb{E}[X(Y-1+\Phi(\frac{t}{\sqrt{3}}))]}{\sqrt{3\Phi(\frac{t}{\sqrt{3}})(1-\Phi(\frac{t}{\sqrt{3}}))}} \\ &= \frac{\mathbb{E}(XY) - \mathbb{E}(X) + \Phi(\frac{t}{\sqrt{3}})\mathbb{E}(X)}{\sqrt{3\Phi(\frac{t}{\sqrt{3}})(1-\Phi(\frac{t}{\sqrt{3}}))}} \\ &= \frac{\mathbb{E}(XY)}{\sqrt{3\Phi(\frac{t}{\sqrt{3}})(1-\Phi(\frac{t}{\sqrt{3}}))}}.\end{aligned}$$

Let us compute $\mathbb{E}(XY)$.

$$\begin{aligned}\mathbb{E}(XY) &= \mathbb{E}(X\mathbb{1}_{\{X \geq t\}}) \\ &= \int_{-\infty}^{\infty} x\mathbb{1}_{\{x \geq t\}} \frac{1}{\sqrt{3 \times 2\pi}} \exp(\frac{-x^2}{2 \times 3})dx \\ &= \int_t^{\infty} x \frac{1}{\sqrt{6\pi}} \exp(\frac{-x^2}{6})dx \\ &= [\frac{-3}{\sqrt{6\pi}} \exp(\frac{-x^2}{6})]_t^{+\infty} \\ &= \frac{3}{\sqrt{6\pi}} \exp(\frac{-t^2}{6}).\end{aligned}$$

So we end up with

$$\rho^P(X,Y) = \frac{\exp(\frac{-t^2}{6})}{\sqrt{2\pi\Phi(\frac{t}{\sqrt{3}})(1-\Phi(\frac{t}{\sqrt{3}}))}}.$$

Similarly, Spearman's rho is given by

$$\rho^s(X,Y) = \frac{\mathbb{E}[(F(X)-\mathbb{E}(F(X)))(G(Y)-\mathbb{E}(G(Y)))]}{\sigma(F(X))\sigma(G(Y))}.$$

We have

$$\begin{aligned}\mathbb{E}(G(Y)) &= \Phi(\frac{t}{\sqrt{3}})\mathbb{P}(Y=0)+\mathbb{P}(Y=1)\\ &= \Phi(\frac{t}{\sqrt{3}})^2+1-\Phi(\frac{t}{\sqrt{3}})\end{aligned}$$

and by the same reasoning, $\mathbb{E}(G(Y)^2) = \Phi(\frac{t}{\sqrt{3}})^3+1-\Phi(\frac{t}{\sqrt{3}})$, which leads to

$$\begin{aligned}\sigma(G(Y)) &= \sqrt{\mathrm{Var}(G(Y))}\\ &= \sqrt{\mathbb{E}(G(Y)^2)-\mathbb{E}(G(Y))^2}\\ &= \sqrt{\Phi(\frac{t}{\sqrt{3}})^3+1-\Phi(\frac{t}{\sqrt{3}})-[\Phi(\frac{t}{\sqrt{3}})^2+1-\Phi(\frac{t}{\sqrt{3}})]^2}\\ &= \sqrt{-\Phi(\frac{t}{\sqrt{3}})^4+3\Phi(\frac{t}{\sqrt{3}})^3-3\Phi(\frac{t}{\sqrt{3}})^2+\Phi(\frac{t}{\sqrt{3}})}\end{aligned}$$

Moreover, we know that $F(X)\sim\mathcal{U}[0,1]$ so $\mathbb{E}(F(X)) = \dfrac{1}{2}$ and $\sigma(F(X)) = \sqrt{\dfrac{1}{12}}$.

So, we get:

$$\begin{aligned}\rho^S(X,Y) &= \frac{\mathbb{E}[(F(X)-\mathbb{E}(F(X)))(G(Y)-\mathbb{E}(G(Y)))]}{\sigma(F(X))\sigma(G(Y))}\\ &= \frac{\mathbb{E}[(F(X)G(Y)]-\mathbb{E}(F)\mathbb{E}(G)}{\sqrt{\frac{1}{12}(-\Phi(\frac{t}{\sqrt{3}})^4+3\Phi(\frac{t}{\sqrt{3}})^3-3\Phi(\frac{t}{\sqrt{3}})^2+\Phi(\frac{t}{\sqrt{3}}))}}\\ &= \frac{\mathbb{E}[(F(X)G(Y)]-\frac{1}{2}(\Phi(\frac{t}{\sqrt{3}})^2+1-\Phi(\frac{t}{\sqrt{3}}))}{\sqrt{\frac{1}{12}(-\Phi(\frac{t}{\sqrt{3}})^4+3\Phi(\frac{t}{\sqrt{3}})^3-3\Phi(\frac{t}{\sqrt{3}})^2+\Phi(\frac{t}{\sqrt{3}}))}}\end{aligned}$$

Let us compute $\mathbb{E}(F(X)G(Y))$.

$$\begin{aligned}
\mathbb{E}(F(X)G(Y)) &= \mathbb{E}(F(X)G(\mathbb{1}_{\{X\geq t\}})) \\
&= \int_{-\infty}^{\infty} \Phi(\frac{x}{\sqrt{3}})G(\mathbb{1}_{\{x\geq t\}})\frac{1}{\sqrt{3\times 2\pi}}\exp(\frac{-x^2}{2\times 3})dx \\
&= \int_{-\infty}^{t} \Phi(\frac{x}{\sqrt{3}})\Phi(\frac{t}{\sqrt{3}})\frac{1}{\sqrt{6\pi}}\exp(\frac{-x^2}{6})dx + \int_{t}^{\infty} \Phi(\frac{x}{\sqrt{3}})\frac{1}{\sqrt{6\pi}}\exp(\frac{-x^2}{6})dx \\
&= [\frac{\Phi^2(\frac{x}{\sqrt{3}})\Phi(\frac{t}{\sqrt{3}})}{2}]_{-\infty}^{t} + [\frac{\Phi^2(\frac{x}{\sqrt{3}})}{2}]_{t}^{+\infty} \\
&= \frac{1+\Phi^3(\frac{t}{\sqrt{3}})-\Phi^2(\frac{t}{\sqrt{3}})}{2}
\end{aligned}$$

So we end up with

$$\rho^S(X,Y) = \frac{\Phi^3(\frac{t}{\sqrt{3}})-2\Phi^2(\frac{t}{\sqrt{3}})+\Phi(\frac{t}{\sqrt{3}})}{\sqrt{\frac{1}{3}(-\Phi(\frac{t}{\sqrt{3}})^4+3\Phi(\frac{t}{\sqrt{3}})^3-3\Phi(\frac{t}{\sqrt{3}})^2)}}.$$

References

Resnick, S. I. (1987). *Extreme Values, Regular Variation, and Point Processes.* Springer.

Rota, G.-C. (1964). On the foundations of combinatorial theory i. *Z. Wahrseheinlichkeitstheorie* **2,** 340–368.