A semi-parametric Gaussian copula model for heterogeneous network inference: an application to multi-omics data

Ekaterina Tomilina, Gildas Mazo, Florence Jaffrézic

A semi-parametric Gaussian copula model for heterogeneous network inference: an application to multi-omics data

Ekaterina Tomilina[*a,b], Gildas Mazo[†a], and Florence Jaffrézic[‡b]

[a]Université Paris-Saclay, INRAE,
MaIAGE, Jouy-en-Josas, France
[b]Université Paris-Saclay, INRAE,
AgroParisTech, GABI, Jouy-en-Josas, France

Abstract

Large-scale heterogeneous data integration for network inference is a key methodological challenge, especially in the context of multi-omics data analysis. We propose here a novel procedure based on Gaussian copula methods which allows the joint analysis of data of various types (continuous and discrete). The proposed estimation procedure is semi-parametric, and does not require any explicit assumption concerning the distribution of the marginals. This offers great flexibility for the analysis of biological data that may not follow perfectly any pre-specified parametric distribution. We present a detailed proof of the pairwise likelihood calculation in the context of mixed type data. We show the equivalence between the presence of a block-wise diagonal structure in the copula correlation matrix and block-wise mutual independence in the observed data. We characterize the lower and upper extreme values of the copula parameter in terms of the observed data when a Bernoulli distribution is involved. In an extensive simulation study, we showed that the proposed estimation procedure, based on a pairwise-likelihood approach, was able to accurately estimate the copula correlation matrix, even for a large number of variables (several hundreds) and a small number of replicates (several dozens). The proposed method was also applied to a real ICGC dataset on breast cancer, and is implemented in a freely available R package `heterocop`.

1 Introduction

The recent development of high-throughput technologies provides access to a large amount of omics data of various types (transcriptomics, proteomics,

[*] ekaterina.tomilina@inrae.fr
[†] gildas.mazo@inrae.fr
[‡] florence.jaffrezic@inrae.fr

metabolomics, metagenomics, epigenetics). Despite the access to this rich and varied data, our knowledge on the functioning of the genome and its link with phenotypic characteristics is still incomplete. One tool to better understand biological regulatory mechanisms is network inference. A challenge of systems biology is to have a comprehensive view of regulatory phenomena. Genes encoded by DNA (genotype) are transcribed into mRNAs (transcriptome) which are translated into proteins (proteome). It is known that methylation (methylome) and metabolites (metabolome) also play an important role in these regulatory processes. One would therefore expect to find strong links between these different biological entities. In order to have a better understanding of the biological system it is therefore necessary to infer multi-omics networks. We will focus in this work on correlation network inference.

A major statistical challenge underlying correlation network inference is the heterogeneous nature of the data. Indeed, RNA-seq data for instance are count data, whereas protein abundances are continuous and mutation encoding is often binary. Existing methods such as WGCNA (Langfelder and Horvath, 2008) rely on Pearson's correlation coefficient, and are therefore limited to linear relationships between variables. Another possibility would be to base correlation network inference on Spearman's rho. However, this coefficient is not well-adapted when at least one variable is discrete (Nešlehová, 2007; Mesfioui et al., 2022). Moreover, the lack of an underlying model may be an impediment to further statistical investigation, for instance the simulation of new data or the inclusion of covariables in the experimental design. The goal of this paper is therefore to propose a model for heterogeneous correlation networks, based on the copula theory.

The Gaussian copula model relies on the assumption that the observed variables are transformations of a hidden Gaussian vector, and enables to link their joint cumulative distribution function (CDF) to a Gaussian CDF while preserving their marginal distributions. The Gaussian copula model corresponds to the Nonparanormal distribution in the continuous case (Liu et al., 2009), but can be extended to discrete and mixed variables. With this approach, for instance, it is possible to build a joint CDF for a Poisson, a Negative Binomial and a Gamma random variable, which enables to deal with biological data of various nature.

As biological data do not perfectly follow a pre-defined distribution, a semi-parametric approach has been proposed in (Fan et al., 2017; Dey and Zipunnikov, 2022). Indeed, Spearman's rho and Kendall's tau are estimated first on the observed data, and bridge functions that link these correlation coefficients to the copula correlation coefficients are presented.

We introduce a more direct, likelihood-based approach. We provide an explicit expression of the pseudo-likelihood in the mixed case of continuous and discrete variables, and give a detailed theoretical proof of its calculation. As multi-omics data are often high-dimensional, a pairwise likelihood estimator is built. In order to avoid assumptions on the distribution of the marginals, we estimate the CDFs empirically. We show the equivalence between the presence of a block-wise diagonal structure in the copula correlation matrix and block-wise mutual independence in the observed data. We characterize the lower and upper extreme values of the copula parameter in terms of the observed data when a Bernoulli distribution is involved. This provides an interpretation of the copula correlation coefficients in terms of association relationships between the observed variables. The performance of the proposed method is illustrated in an extensive simulation study. An application to a real ICGC (Zhang et al., 2019) dataset containing tumoral samples of women affected by breast cancer is carried out.

The rest of the paper is as follows. Section 2 presents the model and its dependence properties. The estimation method is given in Section 3. Section 4 presents the simulation studies and Section 5 the real data analysis. A discussion section closes the paper.

2 The model

Let $(X_1, \ldots, X_d)$ be a random vector with cumulative distribution function (CDF) given by

$$\begin{aligned} F(x_1, ..., x_d) &= C_\Sigma(F_1(x_1), \ldots, F_d(x_d)) \\ &\equiv \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_d(x_d))) \end{aligned} \tag{1}$$

where $F_1, \ldots, F_d$ denote the marginal CDFs of the variables $X_1, \ldots, X_d$, C_Σ denotes the Gaussian copula parameterized by the correlation matrix Σ, Φ_Σ the centered Gaussian multivariate CDF of correlation matrix Σ, and Φ^{-1} the inverse of the standard Normal CDF Φ. It can be checked that the right-hand side of (1) indeed is a well-defined CDF with marginals $F_1, \ldots, F_d$ (Sklar, 1973; Nelsen, 2007). One can note that model (1) corresponds to a latent Gaussian variable structure where, if $(Z_1, ..., Z_d) \sim \mathcal{N}(0, \Sigma)$ is a centered Gaussian vector with correlation matrix Σ, then each X_j can be expressed as $X_j = F_j^{\leftarrow}(\Phi(Z_j))$. Note that $F_j^{\leftarrow}$ denotes the generalized inverse function of F_j, that is, $F_j^{\leftarrow}(u) = \inf\{x : F_j(x) \geq u\}$. With model (1) we do not assume that the observed variables $X_1, \ldots, X_d$ are Gaussian. Only the latent variables $Z_1, \ldots, Z_d$ are. In model (1) the marginal distributions $F_1, \ldots, F_d$ of the observed variables $X_1, \ldots, X_d$ are arbitrary. In particular, there can be a mix of continuous and discrete variables. Model (1) also provides us with an explicit expression of the joint CDF of the variables as a function of their marginal CDFs and thus enables us to see how the distribution of each variable impacts the joint distribution. Note that when all the variables are continuous, model (1) corresponds to the Nonparanormal distribution defined in Liu et al. (2009).

2.1 Joint density

An expression of the multivariate density can be derived from model (1). Below, we say that a random variable is continuous if its CDF is increasing, and discrete if its CDF has a countable support.

Proposition 1 *Without loss of generality, suppose that the first p variables are continuous and that the remaining $d-p$ are discrete. Then the multivariate density of (1) can be written as*

$$f(x_1,\dots,x_d) =$$
$$\left(\prod_{j=1}^{p} f_j(x_j)\right) \times \left(\sum_{j_{p+1}=0}^{1} \cdots \sum_{j_d=0}^{1} (-1)^{j_{p+1}+\cdots+j_d} \times C_\Sigma^p(F_1(x_1),\dots,F_p(x_p),u_{p+1,j_{p+1}},\dots,u_{d,j_d})\right), \tag{2}$$

where f_j denotes the density of X_j, $u_{j,0} = F_j(x_j)$ and $u_{j,1} = F_j(x_j-)$, x_j- denotes the previous point from x_j in the ordered support of F_j, and C_Σ^p denotes the derivative of the copula with respect to the p continuous marginal CDFs, that is $C_\Sigma^p(u_1,\dots,u_d) = \partial^p C_\Sigma(u_1,\dots,u_d)/\partial u_1 \cdots \partial u_p$. If x_j is the least point (if there is one), we set by convention that $F_j(x_j-) = 0$. Also by convention we set that if $p = d$ then the second factor in the right-hand side of (2) is replaced by $c_\Sigma(F_1(x_1),\dots,F_p(x_p))$, where $c_\Sigma(u_1,\dots,u_p) = C_\Sigma^p(u_1,\dots,u_p)$ is the density of C_Σ. If $p = 0$, the first factor in (2) is replaced by 1 and $C_\Sigma^p(u_1,\dots,u_d) = C_\Sigma(u_1,\dots,u_d)$.

The formula (2) appears in Song (2007) without proof. A proof of Proposition 1 is given in Section C.1 of the Supplementary material.

2.2 Dependence properties

Having an expression of the multivariate density in equation (2) enables us to study the (in)dependence relationships between $X_1,\dots,X_d$.

2.2.1 Multivariate dependence properties

Proposition 2 *Let $G_1,\dots,G_k$ be a partition of $D = \{1,\dots d\}$, and denote $X_G = (X_j : j \in G)$ for $G \subset D$. Then, $X_{G_1},\dots,X_{G_k}$ are mutually independent if and only if Σ is a block matrix of the form*

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & ... & 0 \\ 0 & \Sigma_2 & ... & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \Sigma_k \end{pmatrix}$$

where each Σ_i is a block of size $|G_i| \times |G_i|$.

A proof of Proposition 2 can be found in Section C.2 of the Supplementary material. We see that the correlation matrix of the copula encodes mutual independencies between groups of variables. Note that the standard Pearson's correlation matrix of the observed variables does not satisfy this property.

2.2.2 Bivariate dependence properties

Let X_1 and X_2 be a pair of variables distributed according to the Gaussian copula model (1) with

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this case the copula is simply denoted by C_ρ. Using (1), it is easy to see that the density c_ρ of C_ρ, that is, $c_\rho(u,v) = \partial^2 C_\rho(u,v)/\partial u \partial v$, $0 < u, v < 1$, is given by

$$c_\rho(u,v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{2\rho\Phi^{-1}(u)\Phi^{-1}(v) - \rho^2(\Phi^{-1}(u)^2 + \Phi^{-1}(v)^2)}{2(1-\rho^2)} \right).$$

By definition, the parameter ρ measures the correlation between the latent Gaussian variables, but how can it be interpreted for the observed variables? By taking $d = 2$ in Proposition 2 we see that $\rho = 0$ if and only if X_1 and X_2 are independent. We shall see that the lower and upper extreme values of ρ can also be characterized in terms of the observed variables when the discrete variables follow a Bernoulli distribution. Remember that X_1 and X_2 are said to be *comonotonic* if one of them is almost surely an increasing function of the other, and *countermonotonic* if they are almost surely a decreasing function of each other (Nelsen, 2007).

Proposition 3 *Suppose that one of the three cases below holds:*

(i) X_1 *and* X_2 *are continuous;*

(ii) $X_1 \sim \mathcal{B}(p_1)$, $0 < p_1 < 1$, *and* X_2 *continuous;*

(iii) $X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$, $0 < p_1 \leq p_2 < 1$ *and* $p_1 + p_2 \geq 1$.

Then

$$\rho = 1 \text{ iff } \begin{cases} (X_1, X_2) \text{ is comonotonic} & \text{case (i);} \\ (X_1, \mathbf{1}_{\{X_2 > F_2^{-1}(1-p_1)\}}) \text{ is comonotonic} & \text{case (ii) ;} \\ X_1 \leq X_2 & \text{case (iii).} \end{cases}$$

and

$$\rho = -1 \text{ iff } \begin{cases} (X_1, X_2) \text{ is countermonotonic} & \text{case (i);} \\ (X_1, \mathbf{1}_{\{X_2 > F_2^{-1}(p_1)\}}) \text{ is countermonotonic} & \text{case (ii) ;} \\ X_1 + X_2 > 0 & \text{case (iii).} \end{cases}$$

A proof of Proposition 3 is given in Section C.3 of the Supplementary material. In case (ii) for $\rho = 1$, the variable X_2 exceeds a certain threshold only if $X_1 = 1$. A similar pattern holds for $\rho = -1$. In case (iii), $\rho = 1$ indicates that X_1 is dominated by X_2, and $\rho = -1$ indicates that at least one of the variables has to be non-null. For example, if X_1 and X_2 encode the presence of two mutations, then $\rho = 1$ indicates that presence of the first mutation implies presence of the second. A visual representation of Proposition 3 is depicted in Figures S1 ($\rho = 1$) and S2 ($\rho = -1$) of the Supplementary material.

3 Inference of Σ

Let $X^i = (X_1^i, \dots, X_d^i)$, $i = 1, \dots, n$, be n i.i.d. observations in $\mathbb{R}^d$ drawn from the distribution defined in model (1). As it is often the case, we suppose that for all j in $\{1, \dots, d\}$, we have no information regarding the marginal distributions F_j which are replaced by the empirical distributions

$$\hat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_j^i \leq x)$$

where $\mathbb{1}$ denotes the indicator function. Hence, our inference is performed in a semi-parametric framework. In a high dimensional setting, computing the full multivariate density has a high computational cost. We therefore propose to estimate the copula correlation matrix Σ by extending the pairwise maximum likelihood estimator (Mazo et al., 2024) to mixed and non-parametric marginals. In other words, we compute

$$\hat{\Sigma} = \arg\max_{\Sigma} \frac{1}{n} \sum_{i=1}^{n} \sum_{j<j'} \log \hat{f}_{jj'}(X_j^i, X_{j'}^i, \rho_{jj'}). \tag{3}$$

In the expression above, $\rho_{jj'}$ denotes the element of Σ at the jth row and j'th column and $\hat{f}_{jj'}$ denotes an estimate of the density of the bivariate marginal CDF corresponding to $(X_j, X_{j'})$ with respect to $\lambda \otimes \lambda$ if both variables are continuous, $\mu \otimes \mu$ if both variables are discrete, and $\lambda \otimes \mu$ measure if X_j is continuous and $X_{j'}$ is discrete, with λ the Lebesgue measure and μ the counting measure. Above we said that $\hat{f}_{jj'}$ is an *estimate* of $f_{jj'}$, the density of $(X_j, X_{j'})$. Indeed, as we shall see below, the density $f_{jj'}$ depends on the marginal CDFs F_j and $F_{j'}$. But since we substitute the empirical CDFs $\hat{F}_j$ and $\hat{F}_{j'}$ for F_j and $F_{j'}$, the resulting function $\hat{f}_{jj'}(\cdot, \cdot; \rho_{jj'})$ is only an estimate of the true density $f_{jj'}(\cdot, \cdot; \rho_{jj'})$.

The formulas of the densities $f_{jj'}$ in the three cases (X_j and $X_{j'}$ continuous, X_j continuous and $X_{j'}$ discrete, X_j and $X_{j'}$ discrete) are given next. Rewrite

$$C_{\rho_{jj'}}(u, v) = C_\Sigma(1, \dots, 1, u, 1, \dots, 1, v, 1, \dots, 1)$$

(u and v at the jth and j'th positions, respectively) so that the bivariate CDF of $(X_j, X_{j'})$ is given by $C_{\rho_{jj'}}(F_j(x_j), F_{j'}(x_{j'}))$. Let $c_{\rho_{jj'}}(u, v)$ denote the density of $C_{\rho_{jj'}}(u, v)$, that is, $c_{\rho_{jj'}}(u, v) = \partial^2 C_{\rho_{jj'}}(u, v)/\partial u \partial v$, $0 < u, v < 1$, $-1 < \rho_{jj'} < 1$. Let f_j be the marginal density of variable X_j. If X_j and $X_{j'}$ are continuous, then $f_{jj'}$ can be expressed as

$$f_{jj'}(x_j, x_{j'}) = f_j(x_j) f_{j'}(x_{j'}) \times c_{\rho_{jj'}}(F_j(x_j), F_{j'}(x_{j'})).$$

If both variables are discrete, then the density takes the following form:

$$\begin{aligned} f_{jj'}(x_j, x_{j'}) = \mathbb{P}(X_j = x_j, X_{j'} = x_{j'}) = C_{\rho_{jj'}}(F_j(x_j), F_{j'}(x_{j'})) \\ + C_{\rho_{jj'}}(F_j(x_j-), F_{j'}(x_{j'}-)) \\ - C_{\rho_{jj'}}(F_j(x_j-), F_{j'}(x_{j'})) \\ - C_{\rho_{jj'}}(F_j(x_j), F_{j'}(x_{j'}-)). \end{aligned}$$

Finally, if X_j is continuous and $X_{j'}$ is discrete, then we get the following form:

$$f_{jj'}(x_j, x_{j'}) = f_j(x_j) \int_{F_{j'}(x_{j'}-)}^{F_{j'}(x_{j'})} c_{\rho_{jj'}}(F_j(x_j), v)\mathrm{d}v.$$

The estimated density $\hat{f}_{jj'}$ is obtained by substituting $\hat{F}_j$ and $\hat{F}_{j'}$ for F_j and $F_{j'}$, respectively, in the formulas above.

4 Simulations

The goal of this simulation study was to illustrate several properties of the proposed copula model and estimation procedure. We first considered the bivariate case. Then, we extended our estimation to a high-dimensional setting. The simulations from this section were run with our `heterocop` R package available on CRAN.

4.1 Simulation study in the bivariate case

We simulated four variables with a joint cumulative distribution function corresponding to a Gaussian copula as in model (1) and with marginals detailed below:

- a Poisson distribution $\mathcal{P}(1)$ of mean and variance 1
- a Negative Binomial distribution, denoted $NB(1, 0.5)$, where 1 is the number of successful trials and 0.5 is the probability of success
- a centered normal distribution with variance 1 $\mathcal{N}(0, 1)$
- a Bernoulli distribution $\mathcal{B}(0.5)$ of mean 0.5

The four variables make 6 pairs, studied separately. Let ρ denote the copula parameter of the pair considered and $\hat{\rho}$ its estimate obtained from (3). The Mean Squared Error (MSE) of $\hat{\rho}$ is defined as: $\text{MSE}(\hat{\rho}) = \mathbb{E}[(\hat{\rho} - \rho)^2]$. The MSE can be decomposed into the sum of the variance and the squared bias of $\hat{\rho}$ as follows:

$$\text{MSE}(\hat{\rho}) = \mathbb{E}[(\hat{\rho} - \rho)^2] = \underbrace{\mathbb{E}[\hat{\rho}^2] - \mathbb{E}[\hat{\rho}]^2}_{Var(\hat{\rho})} + \underbrace{(\mathbb{E}[\hat{\rho} - \rho])^2}_{Bias(\hat{\rho})^2}$$

For each of the 6 pairs, the MSE, variance and squared bias of our estimator were empirically estimated by running $N = 500$ simulations for different sample sizes $n = 20, 50, 100, 500, 1000$ and copula coefficients $\rho = 0.3, 0.6, 0.8$. The results are depicted in Figure 1.

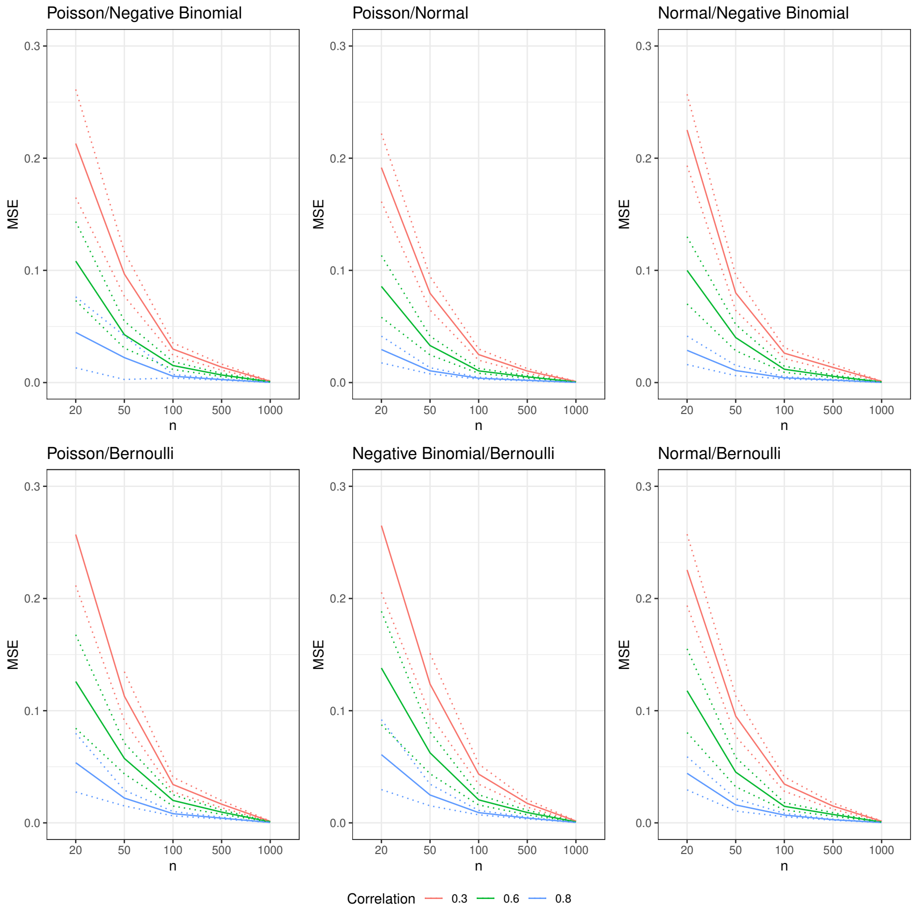


Figure 1: Averaged MSE with 95% confidence intervals of $\hat{\rho}$ for the various values of ρ and sample sizes, and each of the 6 pairs of the 4 distributions.

Figures S3 and S4 of the Supplementary material represent the evolution of the variance and of the squared bias depending on the sample size. One can see that the variance of the estimators decreases to zero as the sample size increases. It is also interesting to note that it is higher for lower values of the correlation coefficient ($\rho = 0.3$) than for the higher ones ($\rho = 0.8$). The variances do not seem to be impacted by the types of the variables, and a similar pattern is observed for all the distributions considered here. It can be noticed that the squared biases are all very close to zero. Although slightly higher in the discrete/discrete case for $n = 20$, they remain extremely low and do not significantly differ from zero as soon as the sample size exceeds 50.

We compared our semi-parametric approach with a fully parametric one in which the parametric families of the marginal distributions are known. We considered the case of the Normal and Negative Binomial distributions, with parameters specified above. The copula correlation coefficient was now estimated in a fully parametric way, i.e. the parameters of the marginals were estimated by maximum likelihood in a first step (each marginal separately) and the copula parameter was estimated in a second step from the likelihood with the estimated parametric marginals plugged in. Figure S5 in the Supplementary material presents the variances and squared biases of both the parametric and semi-parametric estimates, for $N = 500$ simulations. The variances were found to be slightly higher for our semi-parametric estimator for low sample sizes of 20 and 50, but quite similar otherwise. Both the parametric and semiparametric estimators have a negligible bias, compared to the variance.

In real data analyses, the parametric families of the marginal distributions is rarely known. We therefore assessed the robustness of our semi-parametric method against miss-specification of the marginal distributions. We simulated the data as previously but estimated the marginal parameters assuming a Poisson distribution instead of a Negative Binomial one, a common situation in genomics. Figure S6 of the Supplementary material shows that the estimates of ρ obtained by the fully parametric approach are biased while our semi-parametric method remains robust. Our proposed approach will therefore be useful for practical applications when the parametric distribution of the data cannot be specified.

Finally, we assessed the ability of the copula correlation coefficient to capture complex dependence relationships. Let $X_1 \sim \mathcal{N}(0, 3)$ and $X_2 = \mathbb{1}_{\{X_1 \geq t\}}$, where $t \in \mathbb{R}$ is some fixed threshold. It is shown in Section D of the supplementary material that the random vector (X_1, X_2) belongs to model (1) with copula correlation $\rho = 1$. By comparison, the numerical values of Pearson's ρ^P and Spearman's ρ^S for the threshold values $t = 0, 2, 4, 6$ are 0.79, 0.62, 0.27, 0.06, and 0.87, 0.57, 0.18, 0.03, respectively. See Section D of the supplementary material for the calculations. The higher the threshold, the less the ability of Pearson and Spearman coefficients to capture the dependence relationship. For illustration, we generated $N = 500$ samples of size $n = 1000$ of the pair (X_1, X_2) for $t = 0, 2, 4, 6$. An histogram of the realizations of X_1 is presented in Figure S7 in the Supplementary material. The proportions of ones for the realizations of X_2 averaged over the 500 samples is given in Table S1 of the Supplementary material. Pearson's ρ^P, Spearman's ρ^S and the copula correlation coefficient of model (1) were estimated from each sample. The distribution of the estimates is depicted in Figure 2, where we see that the numerical calculations are confirmed. The proposed copula correlation estimation seems therefore more robust when binary variables have to be analyzed, especially in the case of rare events as observed in mutation data for example, as presented in the next section.

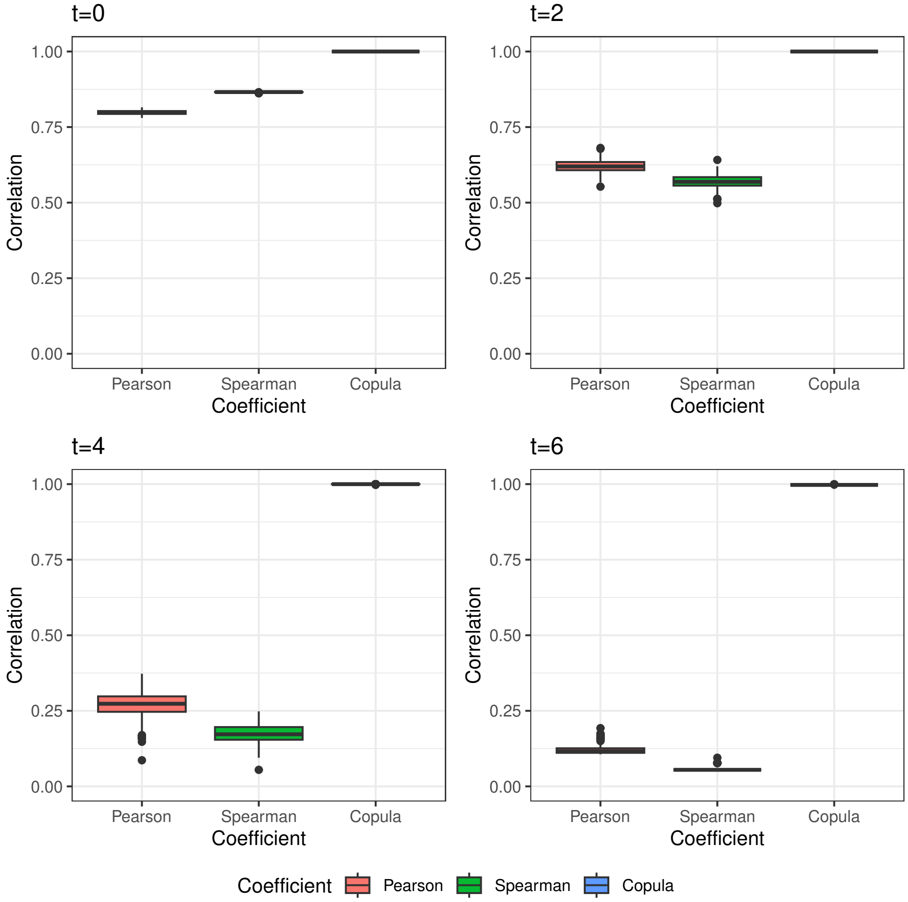


Figure 2: Comparison of the estimation of different correlation coefficients (Pearson, Spearman, Copula) for N=500 replications between $X \sim \mathcal{N}(0, 3)$ and $Y = \mathbb{1}_{\{X \geq t\}}$ for different thresholds t.

4.2 Simulation study in the high-dimensional case

4.2.1 Simulation protocol

Five different sample sizes were considered $n = 20, 50, 100, 500, 1000$, for $d = 30$ and $d = 300$ variables. In each case, one third of the variables were distributed according to a $\mathcal{N}(0, 1)$, one third were $NB(1, \frac{1}{2})$ and the last ones were $\mathcal{B}(\frac{1}{2})$. Two structures were considered for the copula correlation matrix. The first is a block-diagonal structure as specified below:

$$\Sigma_{blocks} = \begin{pmatrix} \underbrace{0.8}_{7\times 7} & 0 & 0 & 0 & 0 \\ 0 & \underbrace{0.6}_{10\times 10} & 0 & 0 & 0 \\ 0 & 0 & \underbrace{0.5}_{2\times 2} & 0 & 0 \\ 0 & 0 & 0 & \underbrace{0.7}_{6\times 6} & 0 \\ 0 & 0 & 0 & 0 & \underbrace{0.3}_{5\times 5} \end{pmatrix}$$

In this matrix, the order of the variables was randomly defined to have blocks of correlated variables of different types. For $d = 300$ variables, the size of each block was multiplied by 10. The second structure is a sparse structure. A matrix Σ is generated through a modified Cholesky decomposition as described in Algorithm 1.

Algorithm 1 Simulation of a positive definite sparse matrix Σ

```
Require: $\gamma \in [0,1],\ m > 0$
  Define a $m \times m$ matrix of zeroes $\Sigma$
  Simulate $\frac{m(m-1)}{2}$ uniform $\mathcal{U}(0.3,1)$ coefficients
  Randomly set a proportion $\gamma$ of coefficients to 0
  Fill the upper triangular part of $\Sigma$ with the coefficients
  $\Sigma \leftarrow \Sigma^T\Sigma$
  $\Sigma_{ij} \leftarrow \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}}\sqrt{\Sigma_{jj}}}$
  return $\Sigma$
```

By varying γ in Algorithm 1, we can generate matrices with different proportions of zeroes. We let γ_F denote the obtained proportion of zeroes of the final matrix Σ. We call γ_F the sparsity coefficient. Regarding the sparsity of the correlation matrix, we have considered a final proportion γ_F of null coefficients of around 20%, 50% and 80%, by empirically setting the γ parameter at 0.61, 0.79 and 0.91. The simulated matrices were denoted $\Sigma_{0.2}, \Sigma_{0.5}$, and $\Sigma_{0.8}$. For each correlation matrix, simulations were run $N = 500$ times.

4.2.2 Numerical results

Results are first presented for $d = 30$ variables. The estimation accuracy was evaluated using the normalized Root Mean Squared Error (RMSE) and the normalized Mean Absolute Error (MAE) calculated as follows for a $d \times d$ matrix Σ:

$$\mathrm{RMSE}(\hat{\Sigma}) = \frac{1}{N}\sum_{k=1}^{N}\sqrt{\frac{1}{d(d-1)}\sum_{1\leq i\neq j\leq d}(\hat{\Sigma}_{ij}^{k} - \Sigma_{ij})^2}$$

$$\mathrm{MAE}(\hat{\Sigma}) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{d(d-1)}\sum_{1\leq i\neq j\leq d}|\hat{\Sigma}_{ij}^{k} - \Sigma_{ij}|$$

where $\hat{\Sigma}^k$ corresponds to the kth estimation of Σ. Note that the same Σ was kept for the $N = 500$ simulations. As shown in Table S2a from the Supplementary material, the normalized RMSE decreases as the sample size increases. It does not exceed 20% for samples larger than $n = 50$ and remains below 5% for sample sizes greater than $n = 500$. It seems to be robust to the specification of the correlation structure and the amount of sparsity in the matrix. Similarly to the normalized RMSE, the normalized MAE values given in Table S2b from the Supplementary material decrease when the sample size increases, and remain below 5% for sample sizes larger than $n = 500$. The normalized MAE also seems robust to the structure of the correlation matrix and its sparsity.

The same metrics were also evaluated in a higher-dimensional setting, for $d = 300$ variables. In order to reduce computational time, we chose to study only a matrix of sparsity close to 0.8 and four different sample sizes $n = 20, 50, 100, 500$. Table S3 from the Supplementary material shows the obtained normalized RMSE and MAE also averaged over $N = 500$ repetitions. The proposed estimation procedure was found to be robust to an increase of the number of variables. Normalized RMSE and MAE values were indeed close to the values previously obtained with 30 variables, even for a small sample size. This result is promising for applying the proposed method to the analysis of real-life examples.

In the perspective of applying the proposed procedure to construct biological networks, we evaluate its ability to discriminate between small and large values of the copula correlation coefficient. Given a fixed threshold $t \in [0, 1]$, a copula correlation coefficient estimate $\hat{\rho}$ is classified as belonging to the first group if $\hat{\rho} < t$, and as belonging to the second otherwise. By an abuse of language, we call the estimates classified into the first group the predicted zeroes, and those classified into the second group the predicted non-zeroes. Threshold t was here arbitrarily set to 0.3.

The sensitivity to the identification of the non-zeroes, also known as true positive rate, and its specificity in the detection, also known as the true negative rate, were measured. Let TP and FN denote the detected non-zeroes and detected zeroes, respectively, among the real non-zeroes. Similarly, let TN and FP denote the detected zeroes and detected non-zeroes among the real zeroes. The true positive rate (TPR) is equal to the proportion of detected non-zeroes among the real non-zeroes, that is, TPR=TP/(TP+FN). The true negative rate (TNR) is equal to the proportion of detected zeroes among the true zeroes, that is, TNR=TN/(TN+FP). The false negative rate (FNR) is defined as the proportion of detected non-zeroes among the real zeroes, that is, FNR=1-TNR. The false positive rate (FPR) is the proportion of detected zeroes among the real non-zeroes, that is, FPR = 1-TPR. A contingency table is available in Table S4 of the Supplementary material for visual aid.

The Receiver Operating Characteristic (ROC) is a measure of global performance of a given classification rule, or classifier. It is a plot of the TPR against the FPR for each value of t. For instance, when $t = 0$, all the estimated coefficients are classified as non-zeroes and hence TPR=1, FNR=1. When $t = 1$, all the estimated coefficients are classified as zeroes and TPR=0, FNR=0. The AUC, Area Under Curve criterion enables us to quantify the performance of the classifier by evaluating the area under the ROC curve. The closer it is to 1, the better the performance.

The ROC curves are presented in Figure 3 for the four correlation structures considered for $d = 30$ variables and each sample size, after averaging over $N = 500$ simulations. Figure 3 shows the results for $d = 300$ variables for a matrix of 0.8 sparsity and Table S5 from the Supplementary material sums up the AUC values in each case. As expected, the AUC values increase with the sample size. They are already good for a low sample size of 20, close to 0.8 even for $d = 300$ variables, and increase to around 0.9 for a sample size of 50, and close to 1 even for a sample size of 100. It can also be noticed that the accuracy is improved for a sparser correlation structure, which is often the case of interest in the context of biological network inference.

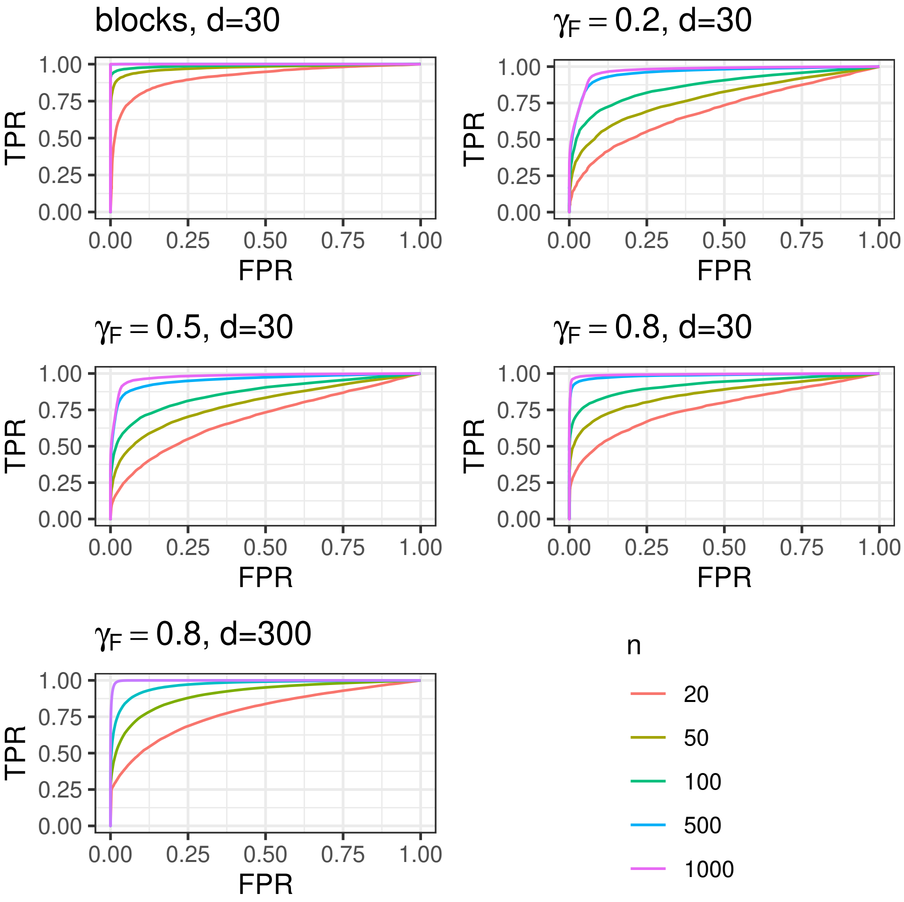


Figure 3: Average ROC curves for $N = 500$ simulations for the classification of the estimates of the copula correlation coefficients for different sample sizes. Four different sparse matrices were considered for $d = 30$ variables (block-wise and sparse matrices with sparsity $\gamma_F = 0.2, 0.5, 0.8$). A matrix with sparsity $\gamma_F = 0.8$ was considered for $d = 300$ variables.

5 Application on real data

We applied the proposed methodology to a data set from the International Cancer Genome Consortium (ICGC, see Zhang et al. (2019)) regarding Breast Cancer in the United States with 990 donors. On each individual, several samples were collected on both healthy and tumoral tissue. Our variables of interest here are RNA-seq counts, protein abundance, and mutations. We kept for further analysis only the samples collected on tumoral tissue, and averaged the normalized protein expression and the RNA-seq counts per individual. The binary encoding was kept for the presence of the mutations for each individual. The initially selected variables prior to pre-processing contained:

- RNA-seq counts for 20 501 genes observed on 939 individuals
- normalized protein abundance for 115 genes observed on 260 individuals
- presence of 107 249 mutations observed on 918 individuals

5.1 Data pre-processing

First, the RNA-seq counts were normalized via the `DESeq2` R package (Love et al., 2014), which enables to study gene differential expression, and rounded to the next integer. The number of donors was reduced to 250 after intersecting the available data for all types of variables. For network inference, in order to reduce the dimension while allowing a biological interpretation of the results, we restricted the analysis to the 108 genes found in common between the RNA-seq and protein data. Concerning the mutation data, we kept those present in at least two donors, reducing their number to 62. The genes associated to each mutation were then identified via the `ensembldb` R package (Rainer et al., 2019). As there were only 4 common genes involving the mutations, RNA-seq and protein data, we decided to keep all 62 mutations for network inference. Our final dataset therefore contained 250 individuals and 278 variables: 108 discrete RNA-seq counts, 108 continuous protein data and 62 binary mutations. Note that for the mutations, the proportion of ones has gone from 0.001 to 0.013 after data pre-processing.

Finally, the copula correlation coefficients of model (1) were estimated through (3) from the final dataset. For comparison, we also estimated the Spearman's ρ^S coefficients.

5.2 Results

5.2.1 Comparison of the copula correlation coefficient with Spearman's ρ^S

Figure S8 from the Supplementary material shows an histogram of the estimates of the coefficients of Spearman's ρ^S and the proposed copula. We can see that the copula correlation coefficient seems to span the entire range of possible values from -1 to 1, while Spearman's ρ^S seems to take smaller absolute values. To understand the difference between Spearman's ρ^S and the copula correlation coefficient, we compare the estimates by type. Remember that there are three variable types: discrete RNA-seq counts (D), continuous protein abundance (C) and binary mutations (B), and hence 6 possible combinations of types for each pair: DD, DC, DB, CC, CB, BB. RNA-seq data, although discrete, have a large number of distinct values, which makes them nearly continuous. Hence we grouped the DD, DC and CC coefficient estimates, leaving three combinations CC (which also contains DD and DC), CB and BB. A scatterplot is displayed for each of these combinations in Figure 4. Panel A of Figure 4 confirms that the differences between Spearman's ρ^S and the copula does not come from the combinations of types DD, DC and CC. We see that the differences are explained by the combinations involving the binary variables. The narrow range of Spearman's ρ^S is explained by the fact that this coefficient applied to two Bernoulli variables with parameters p_1 and p_2 is bounded by $3p_1(1-p_2)$ in absolute value (Mesfioui et al., 2022).

5.2.2 Dependence relationships between the binary variables

Let X_j denote the presence of the jth mutation in some individual ($X_j = 1$ when the mutation is present and 0 otherwise) and let $p_j = \mathbb{P}(X_j = 1)$ denote the Bernoulli parameter of X_j ($j = 1, \ldots, 62$). In the data all p_j are less than 0.15. Thus the conditions of case (iii) of Proposition 3 are satisfied by every pair of binary variables $(1 - X_{j'}, 1 - X_j)$. Indeed $1 - p_j + 1 - p_{j'} \geq 2 - 0.3 = 1.7 > 1$. Thus when the copula correlation coefficient is close to minus one, case (iii) of Proposition 3 predicts that $1 - X_j + 1 - X_{j'} > 0$ and hence $X_j + X_{j'} \leq 1$, that is, no two mutations can co-occur. Case (iii) of Proposition 3 also predicts that when the copula correlation coefficient is close to one then $1 - p_j < 1 - p_{j'}$ implies $1 - X_j \leq 1 - X_{j'}$ and hence $X_j \geq X_{j'}$, that is, the rarest mutation cannot occur without the more common one. A look at the data confirms these predictions, see Table S6 in the Supplementary material for an illustration.

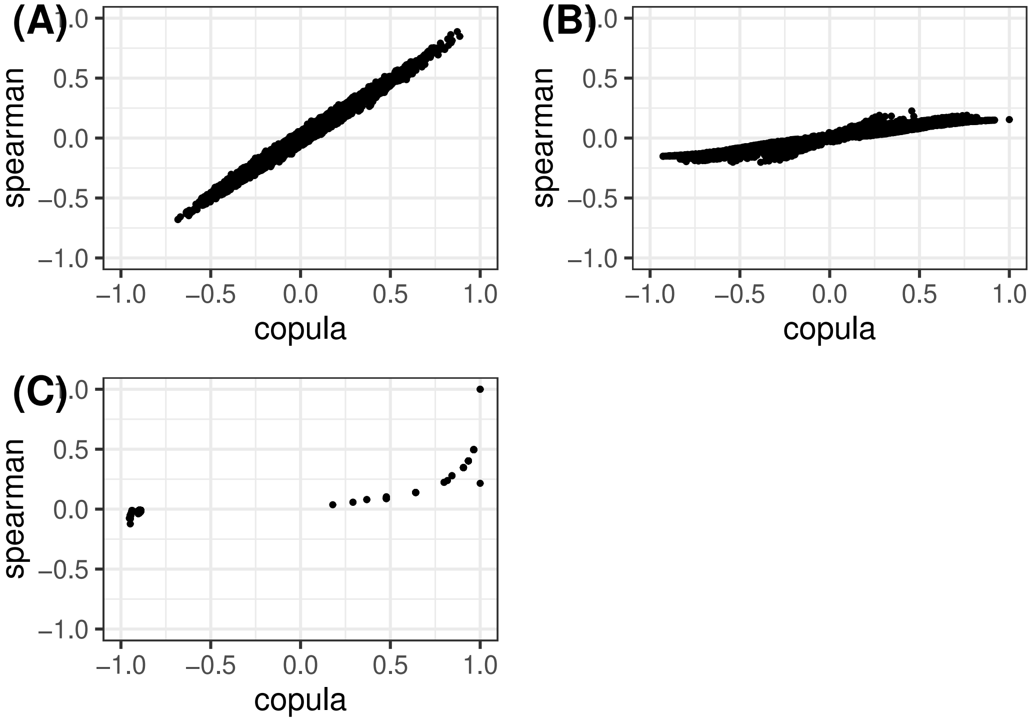


Figure 4: Copula correlation coefficient versus Spearman correlation coefficient for each combination of variable types: continuous/continuous (A), binary/continuous (B) and binary/binary (C). The RNA-seq data have been grouped with the continuous data

5.2.3 Network inference

From a set of estimates of copula correlation coefficients we can build a network by linking highly dependent variables. More precisely, the network is a graph in which the nodes represent the variables and the edges the copula correlation coefficient estimates. One draws an edge between two variables if the absolute value of their copula parameter is greater than some chosen threshold.

One can do the same with the estimates of the Spearman's ρ^S coefficients, and comparison of the inferred networks by the two methods was investigated. The number of detected edges as a function of the threshold, separately for each combination of data types (CC, CD, CB, DD, DB, BB) is depicted in Figure 5. As illustrated in Figure 5, for CC, CD and DD the proposed copula approach and Spearman coefficient behave similarly and identify a similar number of links. When the binary mutation data are involved, however, the copula model detects more links than the Spearman approach, which agrees with the previous remark that Spearman's ρ^S between two binary variables in general cannot reach the endpoints of the interval $[-1, 1]$. When looking at the overlap between the links identified by both methods, it can be noticed that all pairs of binary variables with a non-null Spearman correlation coefficient are included in the subset of pairs detected by the copula approach, for all threshold values. Hence, the copula not only detects the interactions already detected by Spearman, but enables the discovery of new interactions.

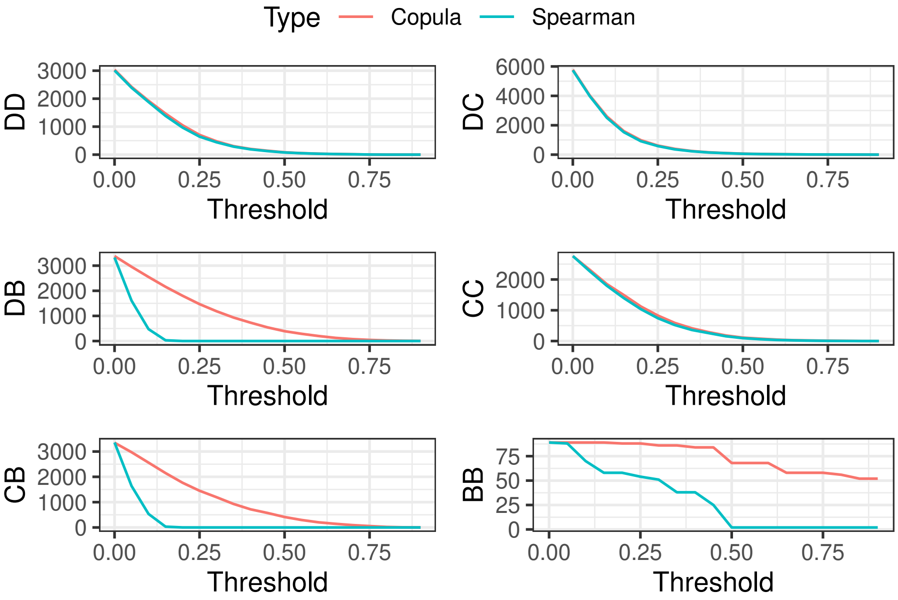
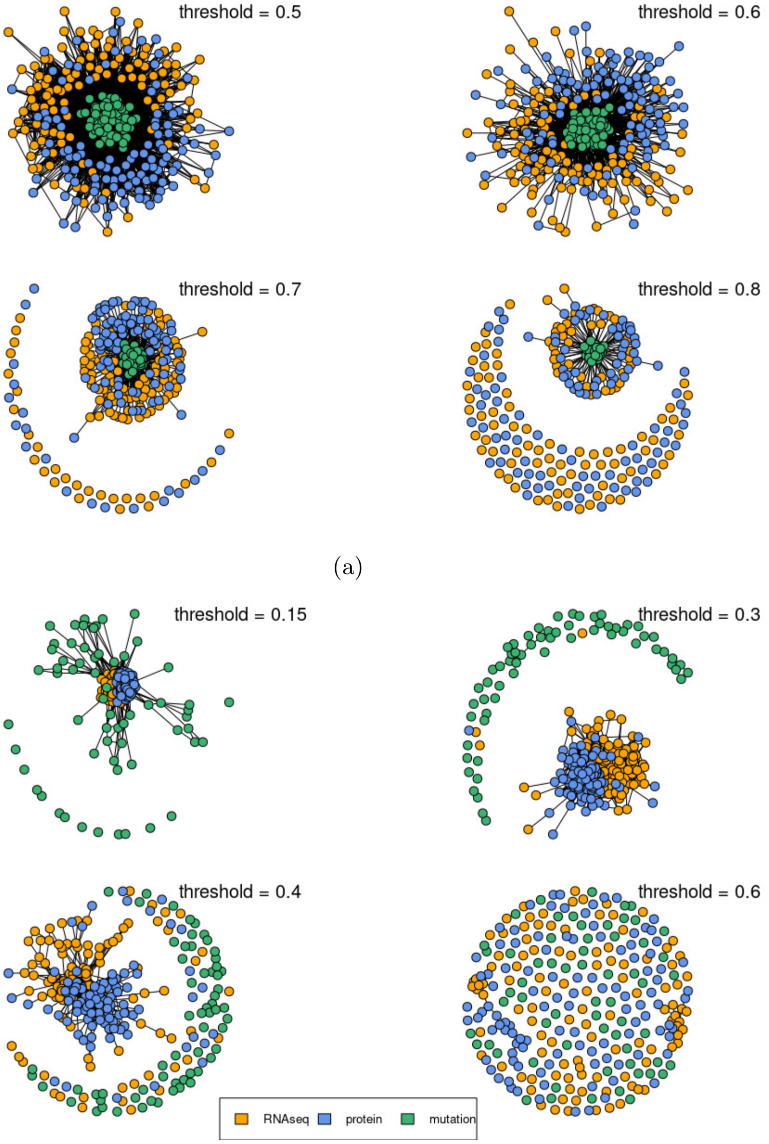


Figure 5: Number of detected links by the copula correlation coefficient versus the Spearman correlation coefficient for threshold values ranging from 0 to 0.9, by combination of types where D stands for discrete (RNA-seq count data), C for continuous (protein data), B for binary (mutation data).

Figure 6 presents the inferred networks obtained from model (1) and the Spearman method, for different threshold values. Regarding the copula model, it can be noted that for a high threshold value of 0.8, edges are identified mainly between mutations. For a lower value of 0.7, edges between pairs of variables with mixed type are detected. In order to have a similar number of links in the network inferred by the Spearman method, the chosen threshold values had to be lower. Indeed, even at the threshold value of 0.6, the Spearman network was found to be very sparse, with only a few edges between proteins and RNA-seq data. Considering a threshold value of 0.4 leads to a larger number of interactions. As expected, very few links were identified with the Spearman approach for the binary mutation data.



(a)

(b)

Figure 6: Copula (a) and Spearman (b) correlation networks for different threshold values. The nodes (variables) are colored by biological type (RNA-seq, protein, mutation). An edge is drawn between two nodes if the absolute value of the corresponding estimated correlation coefficient is above the threshold.

In order to go one step further in the biological interpretation of the inferred networks, we considered the nodes with the highest number of links for the graph returned by the copula model, for a threshold of 0.7 in Figure 6. Figure S9 from the Supplementary material shows the distribution of the degrees of all the nodes in the network. Four of them have a degree (number of associated edges) greater than 10. These four variables correspond to mutations MU4777833, MU5153080, MU17289, and MU5551967. The associated genes as identified by `ensembldb` are shown in Table S7 in the Supplementary Material. We performed a literature search for each of these genes, and they were all found to be involved in cancer development. Indeed, gene ARHGEF11 has been identified as playing a key role in the migration and growth of invasive breast cancer cells (Itoh et al., 2017). Gene SLC7A9 belongs to the SLC7 family which is known for its role in cancer cell metabolism (Yan et al., 2022). Similarly, gene CDKN1B affects protein p27 which is linked to the production of breast cancer cells (Cusan et al., 2018) and finally, PQBP1 is usually overexpressed in breast cancer patients (Liu et al., 2024). The identified hubs of the copula network therefore seem to highlight interesting mutations, that were not identified with the Spearman approach.

6 Discussion

The joint analysis of heterogeneous data is a key methodological topic, especially in the context of multi-omic analyses. We proposed here an innovative approach based on copula methods that allows to infer biological networks from various types of data (continuous and discrete). The idea is to assume an underlying Gaussian latent structure and estimate its corresponding correlation matrix. The proposed method is semi-parametric, with no explicit assumption concerning the distribution of the marginals, which makes it very flexible for biological data analysis. The estimation procedure is based here on a computationally efficient pairwise likelihood approach, and is implemented in a freely available R package called `heterocop`.

We theoretically derived properties of the copula correlation coefficients to make the link with the dependence relationships in the observed data. In particular, we showed that a block-wise structure in the copula correlation matrix is equivalent to block-wise mutual independence in the observed data. We characterized the lower and upper extreme values of the copula parameter in terms of the observed data when a Bernoulli distribution is involved, thus providing an interpretation of the copula parameters.

In an extensive simulation study, we showed that under various experimental designs the Gaussian copula correlation matrix was estimated with a good accuracy with only dozens of observations even for a large number of variables (several hundreds). We also showed that it provided more accurate results than classical correlation coefficients such as Pearson or Spearman, especially for the analysis of binary data. This result was also observed in the real data analysis regarding a breast cancer study including binary mutation data.

Regarding the block-wise mutual independence property, it would be interesting in a further work to propose a sound statistical procedure to identify independent blocks in the data. Theoretical consistency and asymptotic normality of the estimator could also be studied in a future work. This would open the gate to statistical testing and model selection.

Our focus was here on the correlation matrix estimation. In order to obtain the direct links in the networks, the next step would be to propose an estimation procedure for the precision matrix, using the computational efficiency of the pairwise likelihood approach, with a Lasso penalty to obtain a sparse network inference.

References

Cusan, M., Mungo, G., De Marco Zompit, M., Segatto, I., Belletti, B., and Baldassarre, G. (2018). Landscape of CDKN1B mutations in luminal breast cancer and other hormone-driven human tumors. *Frontiers in Endocrinology* **9,** 393.

Dey, D. and Zipunnikov, V. (2022). Semiparametric gaussian copula regression modeling for mixed data types (sgcrm). arXiv:2205.06868 [stat].

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79,** 405–421.

Itoh, M., Radisky, D., Hashiguchi, M., and Sugimoto, H. (2017). The exon 38-containing ARHGEF11 splice isoform is differentially expressed and is required for migration and growth in invasive breast cancer cells. *Oncotarget* **8,** 92157–92170.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research* **10,** 2295–2328.

Liu, X., Zhang, J., Wang, Z., Yan, M., Xu, M., Li, G., Shender, V., Wei, J., Li, J., Shao, C., Zhang, S., Kong, B., Sun, K., and Liu, Z. (2024). Splicing factor PQBP1 curtails BAX expression to promote ovarian cancer progression. *Advanced Science* **11,** e2306229.

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550.

Mazo, G., Karlis, D., and Rau, A. (2024). A randomized pairwise likelihood method for complex statistical inferences. *Journal of the American Statistical Association* **119,** 2317–2327.

Mesfioui, M., Trufin, J., and Zuyderhoff, P. (2022). Bound on Spearman's rho when at least one random variable is discrete. *European Actuarial Journal* **12,** 321–348.

Nelsen, R. (2007). *An Introduction to Copulas.* Springer Series in Statistics. Springer New York.

Nešlehová, J. (2007). On rank correlation measures for non-continuous random variables. *Journal of Multivariate Analysis* **98,** 544–567.

Rainer, J., Gatto, L., and Weichenberger, C. (2019). ensembldb: an R package to create and use ensembl-based annotation resources. *Bioinformatics* **35,** 3151–3153.

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika* **09,** 449–460.

Song, P. (2007). *Correlated data analysis: modeling, analytics, and applications.* Springer.

Yan, L., He, J., Liao, X., Liang, T., Zhu, J., Wei, W., He, Y., Zhou, X., and Peng, T. (2022). A comprehensive analysis of the diagnostic and prognostic value associated with the SLC7A family members in breast cancer. *Gland surgery* **11,** 389–411.

Zhang, J., Bajari, R., and D., A. (2019). The international cancer genome consortium data portal. *Nature Biotechnoogy.* **37,** 367–369.