Multi-omics regulation network inference via the Gaussian copula

Ekaterina TOMILINA

supervised by Gildas MAZO and Florence JAFFRÉZIC

November 25th, 2025

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

INRAE

Table of Contents

Table of Contents

From a biological problem...

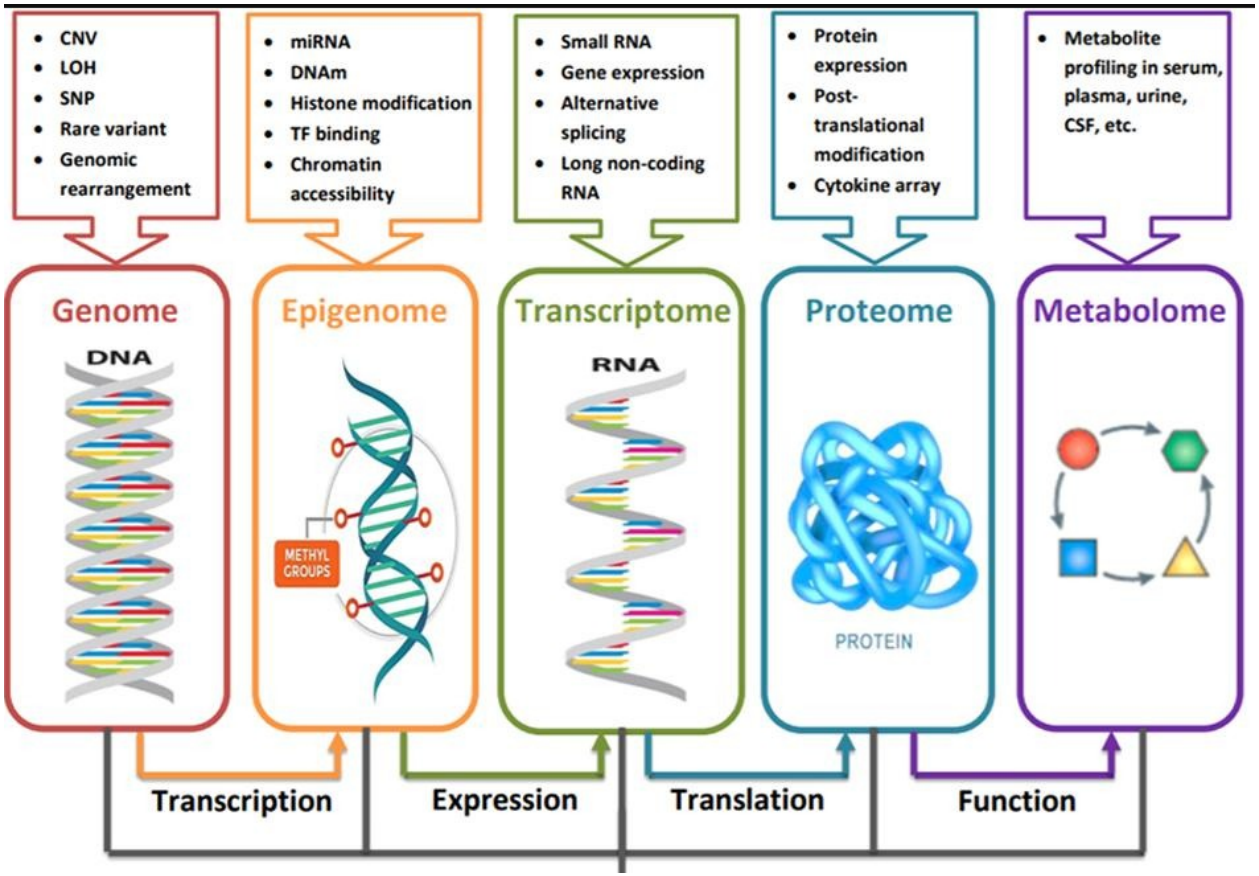


Figure: Main omics levels

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

...to a statistical problem

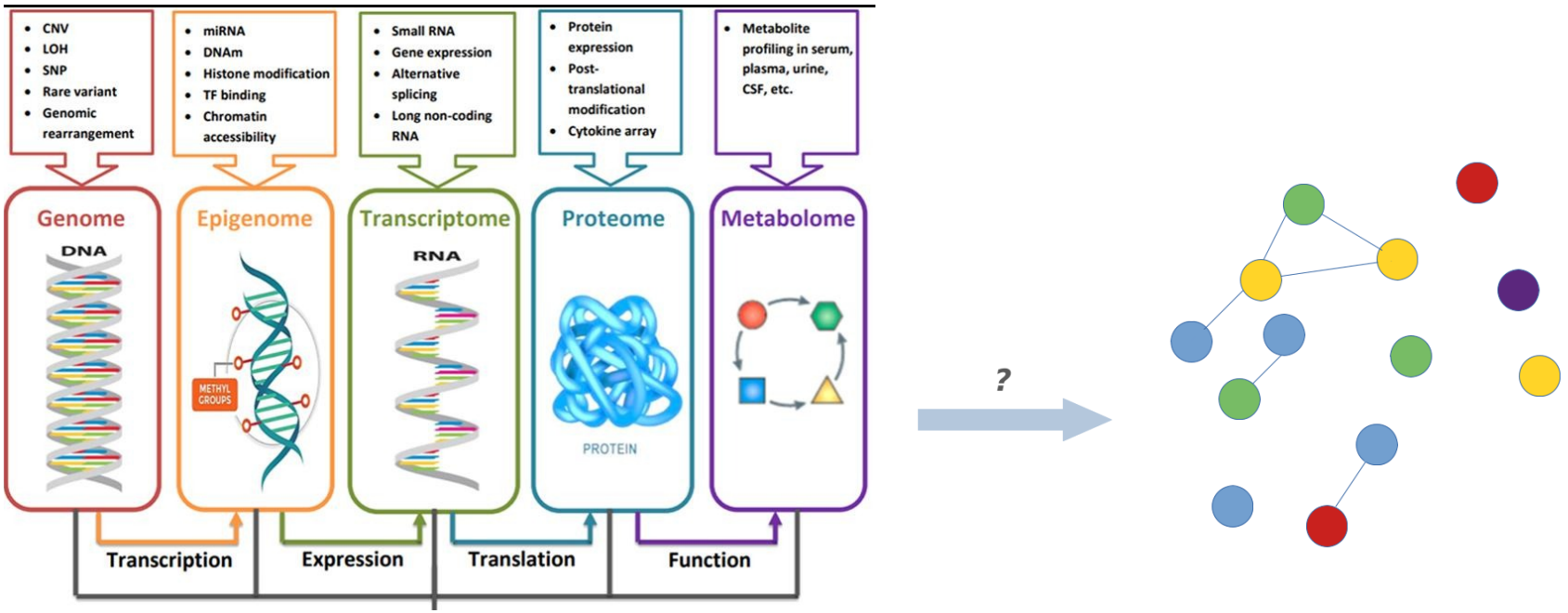


Figure: How to understand the underlying regulation network?

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

...to a statistical problem

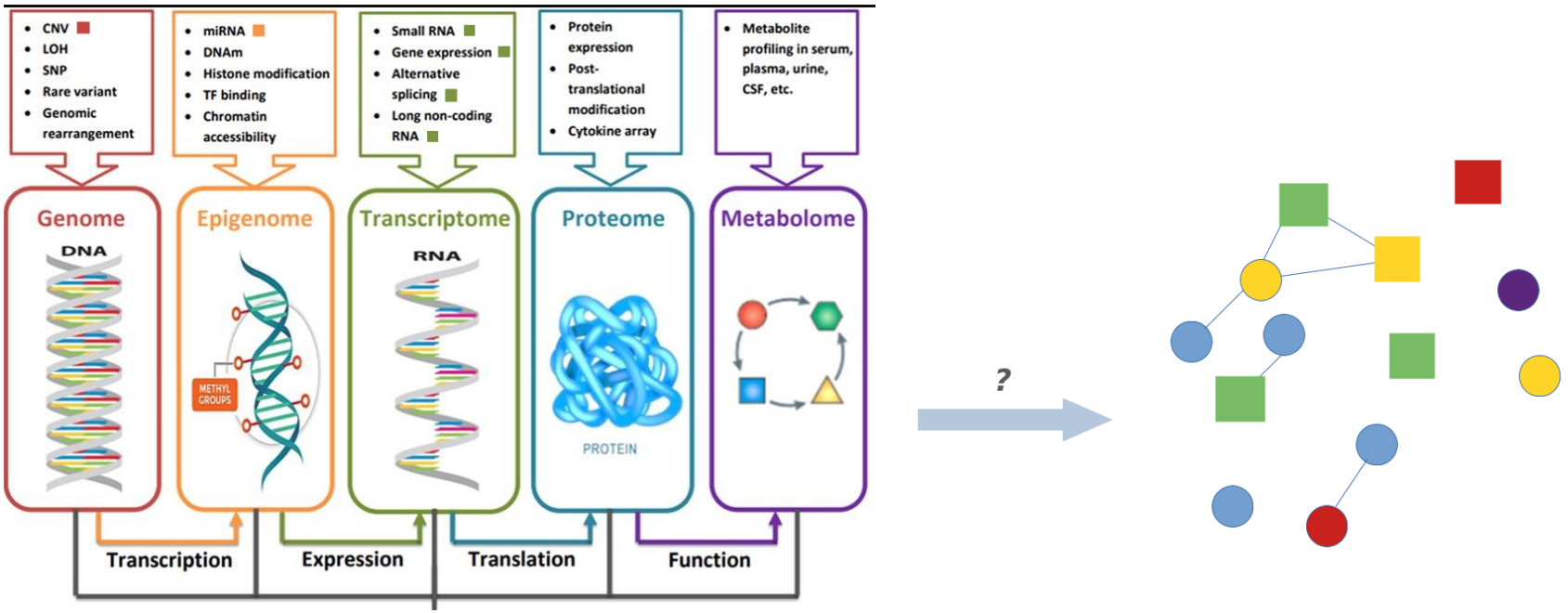


Figure: How to understand the underlying regulation network?

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

...to a statistical problem

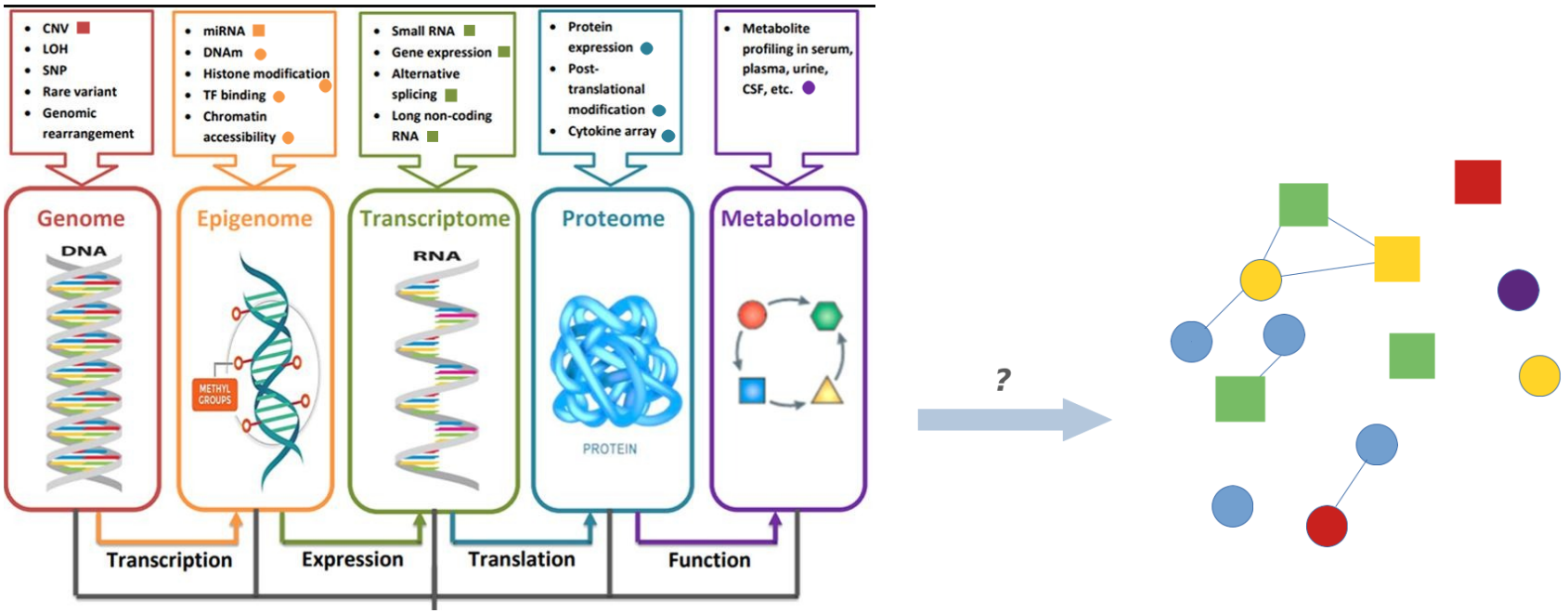


Figure: How to understand the underlying regulation network?

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

...to a statistical problem

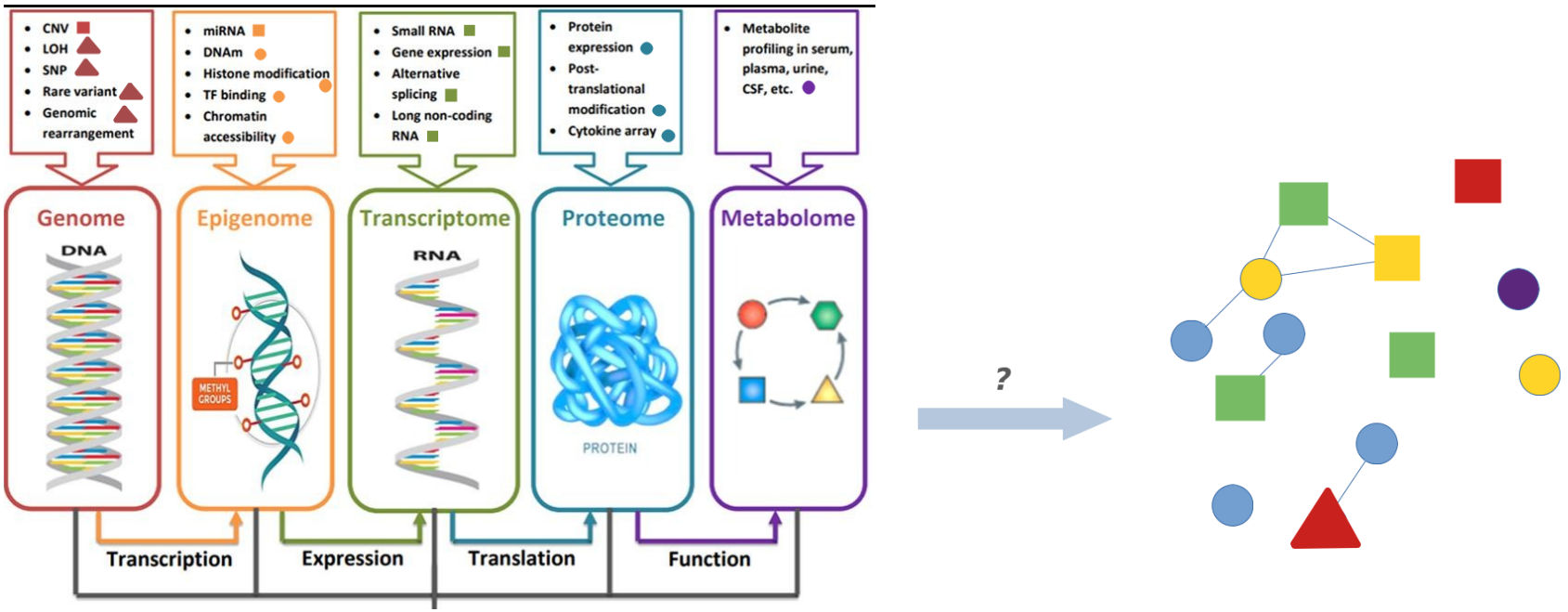


Figure: How to understand the underlying regulation network?

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

→ Mixed variables

...to a statistical problem

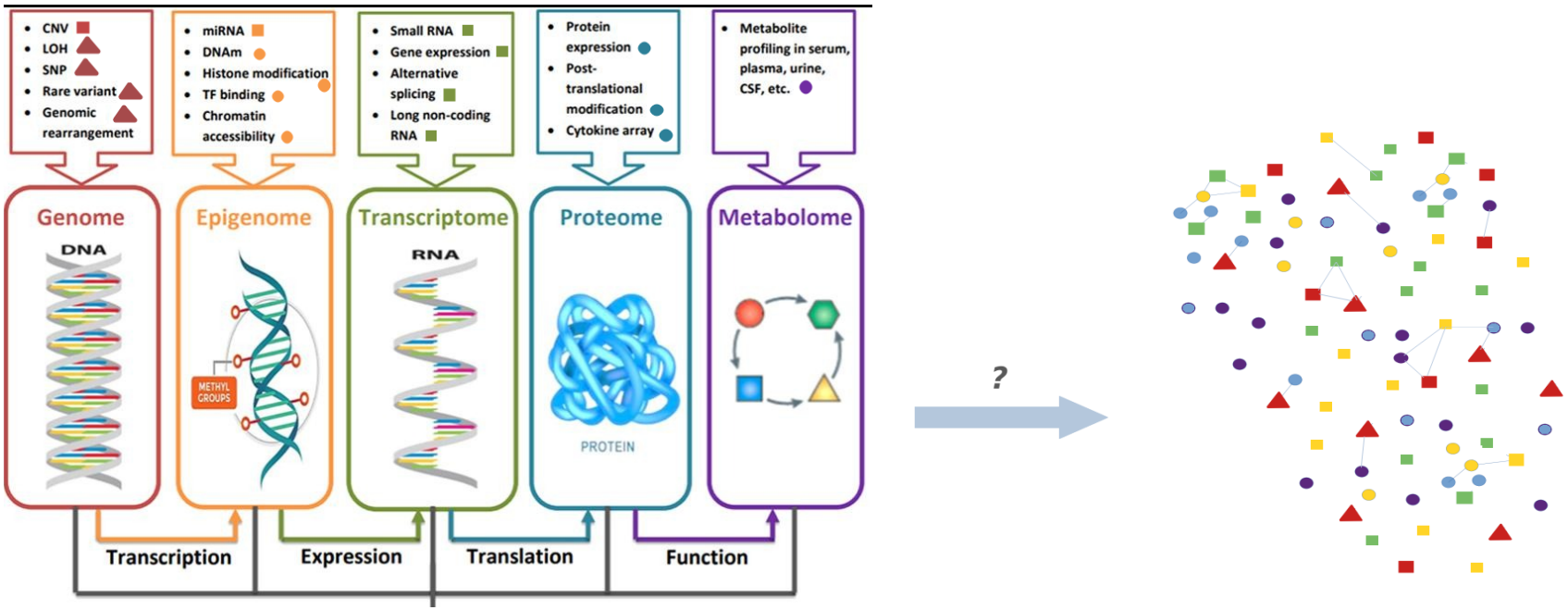


Figure: How to understand the underlying regulation network?

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

$\rightarrow$ Mixed variables, high dimension $d > n$

...to a statistical problem

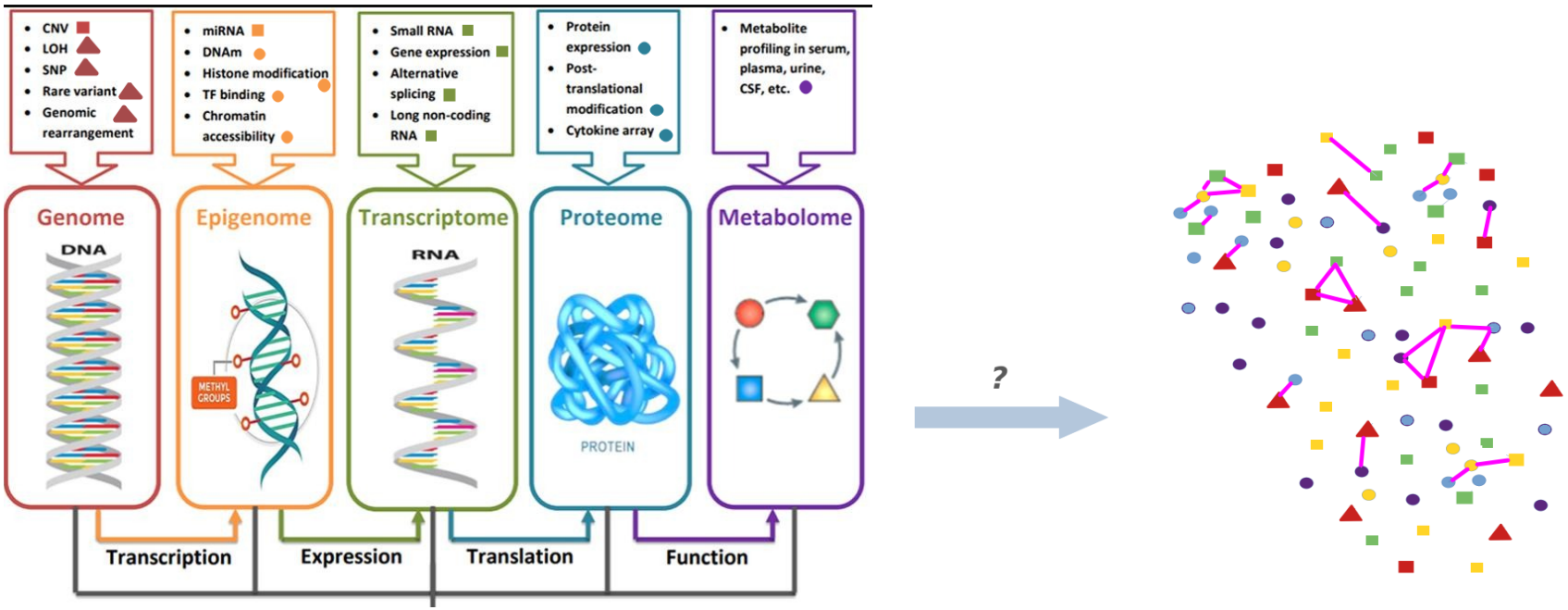


Figure: How to understand the underlying regulation network?

Source: Momeni et al., A survey on single and multi omics data mining methods in cancer data classification, Journal of Biomedical Informatics, 2020

$\rightarrow$ Mixed variables, high dimension $d > n$: what kind of links?

Table of Contents

Table of Contents

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**
  - Pearson's ρ^P:
    - ★ Eisen et al. (1998), discrete RNAseq
    - ★ Love et al. (2014), transformed RNAseq

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**
  - Pearson's ρ^P:
    - ★ Eisen et al. (1998), discrete RNAseq
    - ★ Love et al. (2014), transformed RNAseq
  - Spearman's ρ^S: Langfelder & Horvath (2008), discrete RNAseq

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**
  - Pearson's ρ^P:
    - ★ Eisen et al. (1998), discrete RNAseq
    - ★ Love et al. (2014), transformed RNAseq
  - Spearman's ρ^S: Langfelder & Horvath (2008), discrete RNAseq

  $\rightarrow$ do not handle non-monotonic relationships, nor discrete variables with few categories

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**
  - Pearson's ρ^P:
    - ⋆ Eisen et al. (1998), discrete RNAseq
    - ⋆ Love et al. (2014), transformed RNAseq
  - Spearman's ρ^S: Langfelder & Horvath (2008), discrete RNAseq

  $\rightarrow$ do not handle non-monotonic relationships, nor discrete variables with few categories
- **Mutual information**

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**
  - Pearson's ρ^P:
    - ★ Eisen et al. (1998), discrete RNAseq
    - ★ Love et al. (2014), transformed RNAseq
  - Spearman's ρ^S: Langfelder & Horvath (2008), discrete RNAseq

  $\rightarrow$ do not handle non-monotonic relationships, nor discrete variables with few categories
- **Mutual information**
  Kullback-Leibler divergence between the joint density and the product of the marginals.
  - Butte & Kohane (2000), "binned" RNAseq
  - Margolin (2006), continuous RNAseq
  - Gao et al. (2017), mixed data

Relevance networks

Association relationships between pairs of variables.

- **Correlation coefficients**
  - Pearson's ρ^P:
    - ★ Eisen et al. (1998), discrete RNAseq
    - ★ Love et al. (2014), transformed RNAseq
  - Spearman's ρ^S: Langfelder & Horvath (2008), discrete RNAseq

  $\rightarrow$ do not handle non-monotonic relationships, nor discrete variables with few categories
- **Mutual information**
  Kullback-Leibler divergence between the joint density and the product of the marginals.
  - Butte & Kohane (2000), "binned" RNAseq
  - Margolin (2006), continuous RNAseq
  - Gao et al. (2017), mixed data

  $\rightarrow$ only positive values, tedious estimation

Thesis contributions

Multi-omics network inference

- **Novel approach for simultaneous multi-omics integration.**
- Application to a real-life INRAE data set.
- Availability on CRAN.

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- Mixed density and independence properties.
- Novel estimation approach of the correlation matrix.

Definition

Assume that $F(X_1, \ldots, X_d) = C_\Sigma(F_1(x_1), ..., F_d(x_d))$.

$$\begin{aligned}(X_1, \ldots, X_d) &\sim C_\Sigma(F_1(x_1), ..., F_d(x_d)) \\ &\equiv \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_d(x_d)))\end{aligned}$$

where Φ_Σ represents the CDF of $\mathcal{N}(0, \Sigma)$, Φ the CDF of $\mathcal{N}(0, 1)$, and $F_1, \ldots, F_d$ the marginal CDFs of $X_1, \ldots, X_d$.

Definition

Assume that $F(X_1, \ldots, X_d) = C_\Sigma(F_1(x_1), ..., F_d(x_d))$.

$$\begin{aligned}(X_1, \ldots, X_d) &\sim C_\Sigma(F_1(x_1), ..., F_d(x_d)) \\ &\equiv \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_d(x_d)))\end{aligned}$$

where Φ_Σ represents the CDF of $\mathcal{N}(0, \Sigma)$, Φ the CDF of $\mathcal{N}(0, 1)$, and $F_1, \ldots, F_d$ the marginal CDFs of $X_1, \ldots, X_d$.

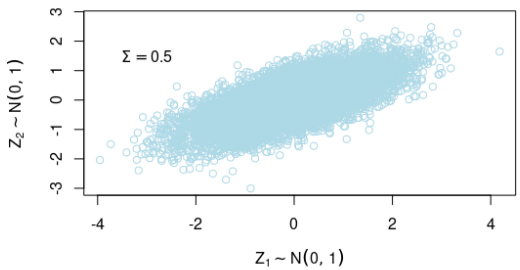
This model is equivalent to d Gaussian $\mathcal{N}(0, \Sigma)$ vectors $Z_1, \ldots, Z_d$ of $\mathbb{R}^n$, such that X_j is defined as $F_j^{\leftarrow}(\Phi(Z_j))$, where $F_j^{\leftarrow}(u) = \inf\{x : F_j(x) \geq u\}$.

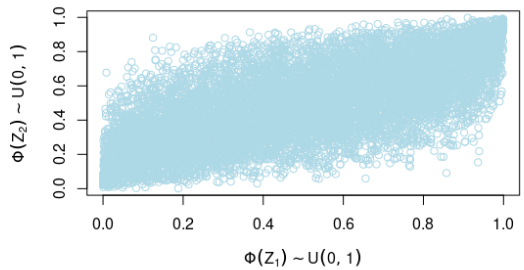
Definition

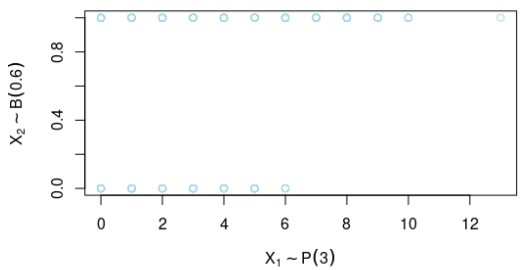
Assume that $F(X_1, \ldots, X_d) = C_\Sigma(F_1(x_1), ..., F_d(x_d))$.

$$\begin{aligned}(X_1, \ldots, X_d) &\sim C_\Sigma(F_1(x_1), ..., F_d(x_d)) \\ &\equiv \Phi_\Sigma(\Phi^{-1}(F_1(x_1)), ..., \Phi^{-1}(F_d(x_d)))\end{aligned}$$

where Φ_Σ represents the CDF of $\mathcal{N}(0, \Sigma)$ and Φ the CDF of $\mathcal{N}(0, 1)$.







Thesis contributions

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- Application to a real-life INRAE data set.
- Availability on CRAN.

Gaussian copula model

- **Interpretation of extreme parameter values for mixed variables.**
- Mixed density and independence properties.
- Novel estimation approach of the correlation matrix.

Comonotonicity

Recall that X_j and X_k are said to be *comonotonic* if one of them is almost surely an increasing function of the other.

Proposition (Tomilina, Mazo, Jaffrézic, 2024)

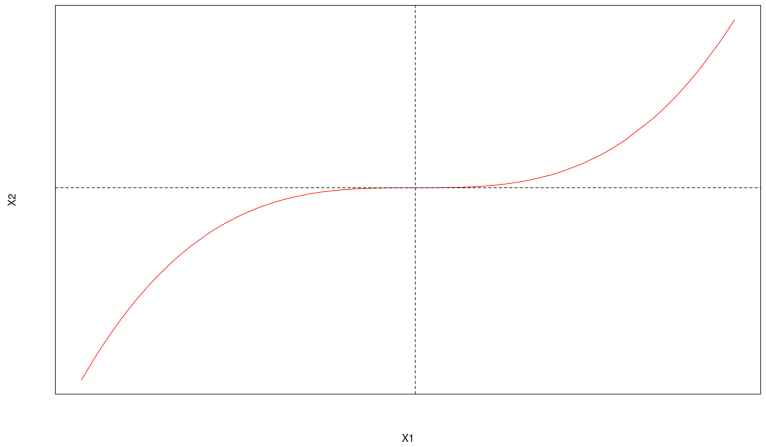
Suppose that one of the three cases below holds:

1. X_j and X_k are continuous;
2. $X_j \sim \mathrm{Ber}(p_j)$ and X_k is continuous;
3. $X_j \sim \mathrm{Ber}(p_j)$, $X_k \sim \mathrm{Ber}(p_k)$, $p_j \leq p_k$ and $p_j + p_k > 1$.

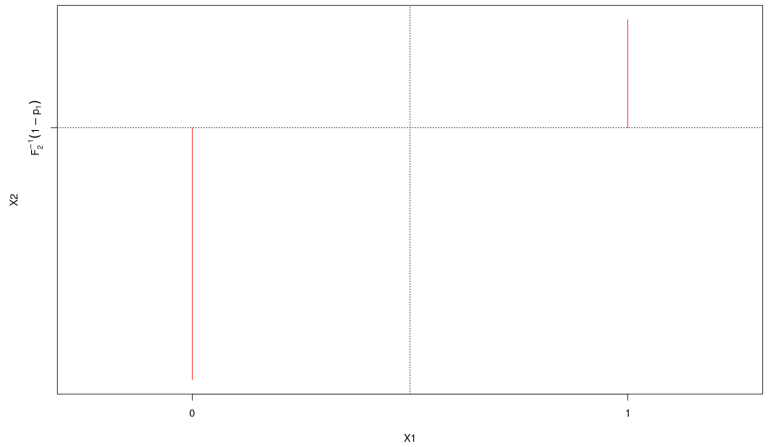
Then

$$\Sigma_{jk} = 1 \text{ iff } \begin{cases} X_j \text{ and } X_k \text{ are comonotonic} & \text{case (1)}; \\ (X_j, \mathbf{1}_{\{X_k > F_k^{-1}(1-p_j)\}}) \text{ is comonotonic} & \text{case (2)}; \\ X_j \leq X_k & \text{case (3)}. \end{cases}$$

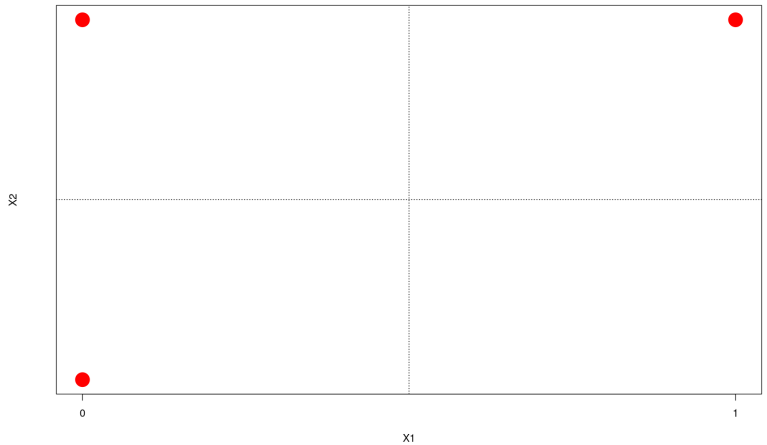
Comonotonicity



Case 1



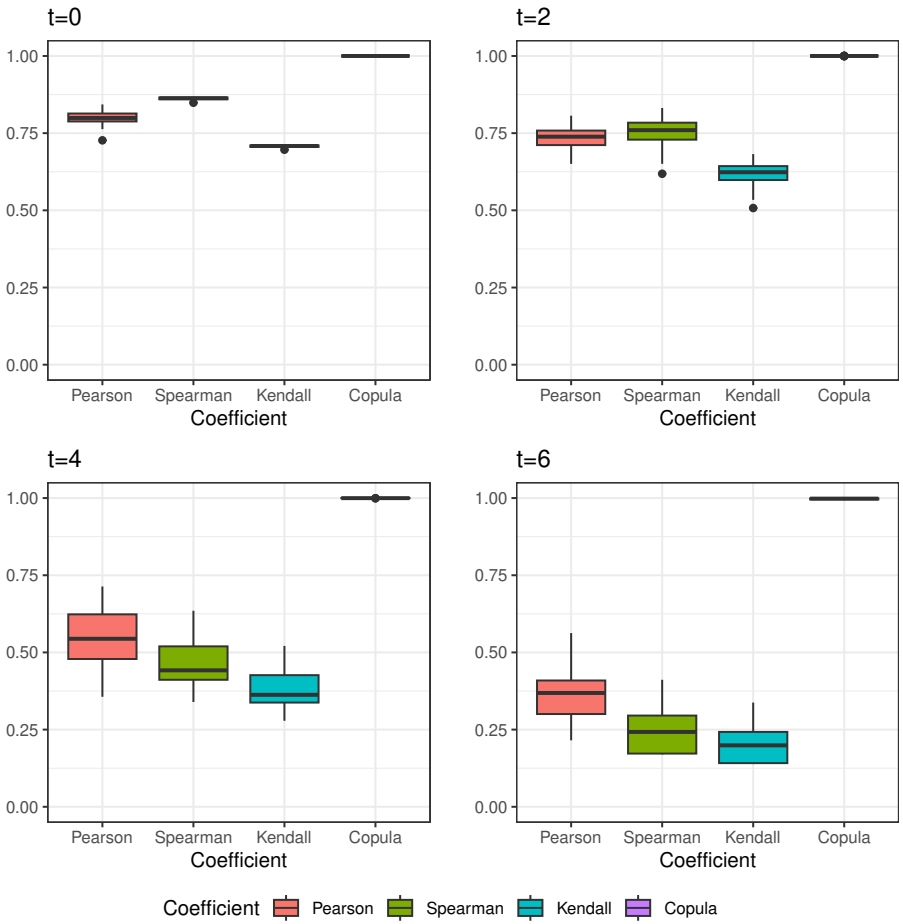
Case 2



Case 3

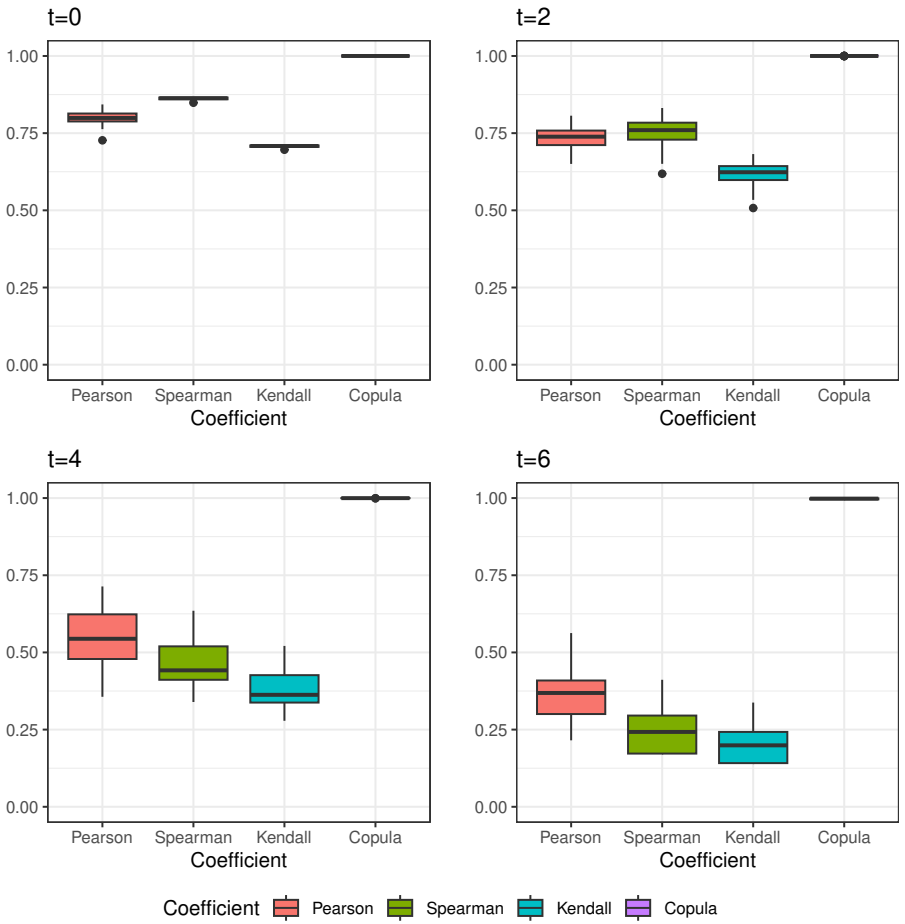
Comparison with other correlation coefficients

Example (case 2): $N = 500$ replications of $n = 10000$ observations, $X \sim \mathcal{N}(0, 3)$, $Y = \mathbb{1}_{\{X \geq t\}}$.



Comparison with other correlation coefficients

Example (case 2): $N = 500$ replications of $n = 10000$ observations, $X \sim \mathcal{N}(0, 3)$, $Y = \mathbb{1}_{\{X \geq t\}}$. Mesfioui et al. (2022): $|\rho^S| \leq \sqrt{3p(1-p)}$.



Countermonotonicity

X_j and X_k are said to be *countermonotonic* if they are almost surely a decreasing function of each other.

Proposition (Tomilina, Mazo, Jaffrézic, 2024)

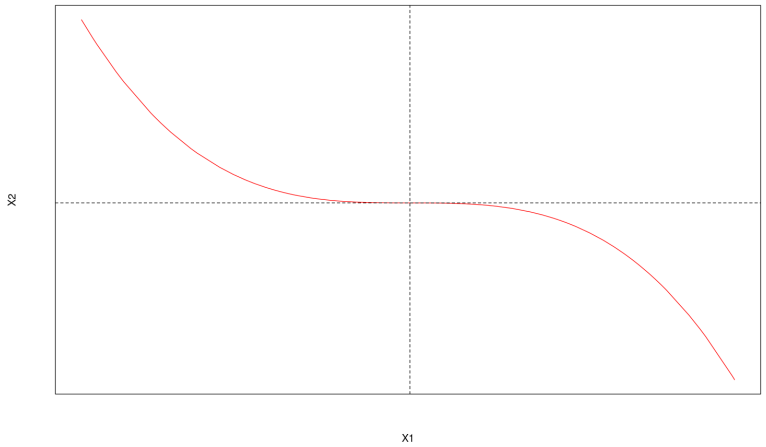
Suppose that one of the three cases below holds:

1. X_j and X_k are continuous;
2. $X_j \sim \mathrm{Ber}(p_j)$ and X_k is continuous;
3. $X_j \sim \mathrm{Ber}(p_j)$, $X_k \sim \mathrm{Ber}(p_k)$, $p_j \leq p_k$ and $p_j + p_k > 1$.

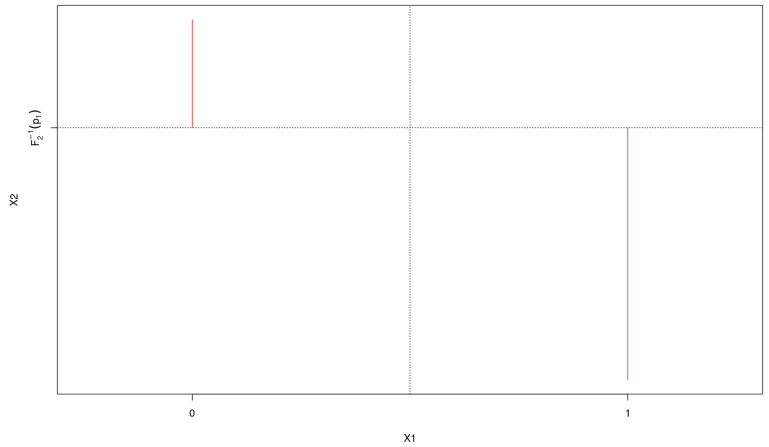
Then

$$\Sigma_{jk} = -1 \text{ iff } \begin{cases} X_j \text{ and } X_k \text{ are countermonotonic} & \text{case (1)}; \\ (X_j, \mathbf{1}_{\{X_k > F_k^{-1}(p_j)\}}) \text{ is countermonotonic} & \text{case (2)}; \\ X_j + X_k \geq 1 & \text{case (3)}. \end{cases}$$

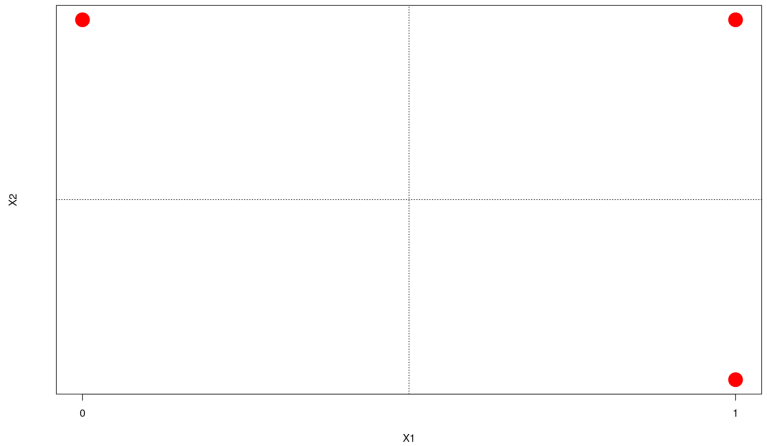
Countermonotonicity



Case 1



Case 2



Case 3

Thesis contributions

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- Application to a real-life INRAE data set.
- Availability on CRAN.

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- **Mixed density and independence properties.**
- Novel estimation approach of the correlation matrix.

Mixed density

Suppose that $X_1, \ldots, X_p$ are continuous, $X_{p+1}, \ldots, X_d$ are discrete, and $\Sigma_{jk} \in (-1, 1)$ for any j, k. The corresponding density is given by[1] :

$$f(x_{i1}, ..., x_{id}, \Sigma) = \prod_{k=1}^{p} f_k(x_{ik}) \sum_{j_{p+1}=1}^{2} \cdots \sum_{j_d=1}^{2} (-1)^{j_{p+1}+...+j_d} \times C_{\Sigma}^{p}(F_1(x_{i1}), ..., F_p(x_{ip}), u_{p+1,j_{p+1}}, ..., u_{d,j_d})$$

where $u_{j,1} = F_j(x_{ij})$, $u_{j,2} = F_j(x_{ij}-)$ and C_{Σ}^{p} represents the derivative of C_{Σ} depending on the p continuous variables.

[1] Song, P., An Introduction to Copulas, 2007

Independence encoding

Proposition (Tomilina, Mazo, Jaffrézic, 2024)

For any $X_1, \ldots, X_d$, let $(G_1, \ldots, G_k)$ be a partition of $\{1, \ldots d\}$, and $\forall i = 1, \ldots, k$, $X_{G_i} = (X_j : j \in G_i)$. Then, $X_{G_1}, \ldots, X_{G_k}$ are mutually independent IFF Σ is a block-wise matrix as below:

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

where each Σ_i is a $s_i \times s_i$ block such that $\sum_{i=1}^{k} s_i = p$ and where $s_i = \#G_i$.

Independence encoding

Proposition (Tomilina, Mazo, Jaffrézic, 2024)

For any $X_1, \ldots, X_d$, let $(G_1, \ldots, G_k)$ be a partition of $\{1, \ldots d\}$, and $\forall i = 1, \ldots, k$, $X_{G_i} = (X_j : j \in G_i)$. Then, $X_{G_1}, \ldots, X_{G_k}$ are mutually independent IFF Σ is a block-wise matrix as below:

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

where each Σ_i is a $s_i \times s_i$ block such that $\sum_{i=1}^{k} s_i = p$ and where $s_i = \#G_i$.

$\rightarrow$ **Contribution:** The independence property also holds in the presence of discrete variables.

How to estimate Σ?

- MCMC methods (Dobra and Mohammadi, 2018)

How to estimate Σ?

- MCMC methods (Dobra and Mohammadi, 2018)
- Methods based on bridge functions, with a statistic $\hat{r}$ computed on the observed data such that $\hat{r} = B(\hat{\Sigma})$.

How to estimate Σ?

- MCMC methods (Dobra and Mohammadi, 2018)
- Methods based on bridge functions, with a statistic $\hat{r}$ computed on the observed data such that $\hat{r} = B(\hat{\Sigma})$.
  - Kruskal (1958): continuous data

$$\hat{r} = \hat{\tau}, \tau = B^{cc}(\Sigma) = \frac{2}{\pi} \arcsin(\Sigma) \quad \checkmark$$

How to estimate Σ?

- MCMC methods (Dobra and Mohammadi, 2018)
- Methods based on bridge functions, with a statistic $\hat{r}$ computed on the observed data such that $\hat{r} = B(\hat{\Sigma})$.
  - Kruskal (1958): continuous data

$$\hat{r} = \hat{\tau}, \tau = B^{cc}(\Sigma) = \frac{2}{\pi} \arcsin(\Sigma)$$ ✔

  - Fan et al. (2017): binary data

$$r = B^{bb}(\Sigma) = 2\{\Phi(\Delta_j, \Delta_k, \Sigma) - \Phi(\Delta_j)\Phi(\Delta_k)\}, \text{where } \Delta_j = \Phi^{-1}(F_j(C_j))$$ ✔

How to estimate Σ?

- MCMC methods (Dobra and Mohammadi, 2018)
- Methods based on bridge functions, with a statistic $\hat{r}$ computed on the observed data such that $\hat{r} = B(\hat{\Sigma})$.
  - Kruskal (1958): continuous data

$$\hat{r} = \hat{\tau}, \tau = B^{cc}(\Sigma) = \frac{2}{\pi} \arcsin(\Sigma)$$ ✔

  - Fan et al. (2017): binary data

$$r = B^{bb}(\Sigma) = 2\{\Phi(\Delta_j, \Delta_k, \Sigma) - \Phi(\Delta_j)\Phi(\Delta_k)\}, \text{where } \Delta_j = \Phi^{-1}(F_j(C_j))$$ ✔

  - Zhang et al. (2021): multinomial data of L and M categories

$$r = B^{dd}(\Sigma) = 2\left\{\sum_{l=1}^{L}\sum_{m=1}^{M}\Phi(\Delta_{jl}, \Delta_{km}, \Sigma) - \sum_{l=1}^{L}\Phi(\Delta_{jl})\sum_{m=1}^{M}\Phi(\Delta_{km})\right\}$$

How to estimate Σ?

- MCMC methods (Dobra and Mohammadi, 2018)
- Methods based on bridge functions, with a statistic $\hat{r}$ computed on the observed data such that $\hat{r} = B(\hat{\Sigma})$.
  - Kruskal (1958): continuous data

$$\hat{r} = \hat{\tau}, \tau = B^{cc}(\Sigma) = \frac{2}{\pi} \arcsin(\Sigma)$$ ✔

  - Fan et al. (2017): binary data

$$r = B^{bb}(\Sigma) = 2\{\Phi(\Delta_j, \Delta_k, \Sigma) - \Phi(\Delta_j)\Phi(\Delta_k)\}, \text{where } \Delta_j = \Phi^{-1}(F_j(C_j))$$ ✔

  - Zhang et al. (2021): multinomial data of L and M categories

$$r = B^{dd}(\Sigma) = 2\left\{\sum_{l=1}^{L}\sum_{m=1}^{M}\Phi(\Delta_{jl}, \Delta_{km}, \Sigma) - \sum_{l=1}^{L}\Phi(\Delta_{jl})\sum_{m=1}^{M}\Phi(\Delta_{km})\right\}$$

→ under debate (Dey and Zippunikov, 2022), does not have a solution in $(-1, 1)$ for high numbers of categories.

How to estimate Σ?

- MLE approach:

$$L_n(F_1, \ldots, F_d, \Sigma) = \frac{1}{n} \sum_{i=1} \log f(x_{i1}, \ldots, x_{id}, \Sigma)$$

[2] Mazo, Karlis and Rau, JASA, 2024

How to estimate Σ?

- MLE approach:

$$L_n(F_1, \ldots, F_d, \Sigma) = \frac{1}{n} \sum_{i=1} \log f(x_{i1}, \ldots, x_{id}, \Sigma)$$

$\rightarrow$ Issue: high computational cost, evaluation of n $d-p-$dimensional integrals: $O(n(d-p)^3)$ (Genz and Bretz, 2002).

[2] Mazo, Karlis and Rau, JASA, 2024

How to estimate Σ?

- MLE approach:

$$L_n(F_1, \ldots, F_d, \Sigma) = \frac{1}{n} \sum_{i=1} \log f(x_{i1}, \ldots, x_{id}, \Sigma)$$

$\rightarrow$ Issue: high computational cost, evaluation of n $d-p-$dimensional integrals: $O(n(d-p)^3)$ (Genz and Bretz, 2002).

$\rightarrow$ Solution: consider the **pairwise MLE** [2]: $O(n(d-p)^2)$.

[2] Mazo, Karlis and Rau, JASA, 2024

Thesis contributions

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- Application to a real-life INRAE data set.
- Availability on CRAN.

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- Mixed density and independence properties.
- **Novel estimation approach of the correlation matrix.**

Pairwise MLE approach

$$L_n(F_1, \ldots, F_d, \Sigma) = \frac{1}{n} \sum_{i=1}^{} \sum_{j<j'} \log f^{(jj')}(x_{ij}, x_{ij'}, \Sigma_{jj'})$$

with the pairwise copula density:

$$f^{(j,j')}(x_{ij}, x_{ij'}, \Sigma_{jj'}) =$$

$$\begin{cases} c_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'})) f_j(x_j) f_{j'}(x_{j'}) & \text{when both variables are continuous} \\ f_j(x_j) \int_{F_{j'}(x_{j'}-)}^{F_{j'}(x_{j'})} c_{\Sigma_{jj'}}(F_j(x_j), v) d\lambda(v) & \text{when the pair is mixed} \\ C_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'})) - C_{\Sigma_{jj'}}(F_j(x_j-), F_{j'}(x_{j'})) & \\ -C_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'}-)) + C_{\Sigma_{jj'}}(F_j(x_j-), F_{j'}(x_{j'}-)) & \text{otherwise} \end{cases}$$

Pairwise MLE approach

$\frac{d(d-1)}{2}$ separate problems:

$$\hat{\Sigma}_{jj'} = \underset{\Sigma_{jj'} \in (-1,1)}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1} \log f^{(jj')}(x_{ij}, x_{ij'}, \Sigma_{jj'}).$$

$$f^{(j,j')}(x_{ij}, x_{ij'}, \Sigma_{jj'}) =$$

$$\begin{cases} c_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'})) f_j(x_j) f_{j'}(x_{j'}) & \text{when both variables are continuous} \\ f_j(x_j) \int_{F_{j'}(x_{j'}-)}^{F_{j'}(x_{j'})} c_{\Sigma_{jj'}}(F_j(x_j), v) d\lambda(v) & \text{when the pair is mixed} \\ C_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'})) - C_{\Sigma_{jj'}}(F_j(x_j-), F_{j'}(x_{j'})) & \\ -C_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'}-)) + C_{\Sigma_{jj'}}(F_j(x_j-), F_{j'}(x_{j'}-)) & \text{otherwise} \end{cases}$$

Pairwise MLE approach

$\frac{d(d-1)}{2}$ separate problems:

$$\hat{\Sigma}_{jj'} = \underset{\Sigma_{jj'} \in (-1,1)}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1} \log f^{(jj')}(x_{ij}, x_{ij'}, \Sigma_{jj'}).$$

$$f^{(jj')}(x_{ij}, x_{ij}, \Sigma_{(jj')}) =$$

$$\begin{cases} c_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'}))f_j(x_j)f_{j'}(x_{j'}) \text{ when both variables are continuous} \\ f_j(x_j) \int_{F_{j'}(x_{j'}-)}^{F_{j'}(x_{j'})} c_{\Sigma_{jj'}}(F_j(x_j), v)d\lambda(v) \text{ when the pair is mixed} \\ C_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'})) - C_{\Sigma_{jj'}}(F_j(x_j-), F_{j'}(x_{j'})) \\ -C_{\Sigma_{jj'}}(F_j(x_j), F_{j'}(x_{j'}-)) + C_{\Sigma_{jj'}}(F_j(x_j-), F_{j'}(x_{j'}-)) \text{ otherwise} \end{cases}$$

Marginal estimation

How to estimate the marginals $F_1, \ldots, F_d$?

1. If their parametric form is known, estimate the parameters for each marginal
2. Otherwise, estimate $\hat{F}_1, \ldots, \hat{F}_d$ where $\hat{F}_j = \frac{1}{n+1} \sum_{i=1} \mathbb{1}_{\{X_j^i \leq x\}}$

Pairwise semi-parametric MLE approach

$\frac{d(d-1)}{2}$ separate problems:

$$\hat{\Sigma}_{jj'} = \underset{\Sigma_{jj'} \in (-1,1)}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1} \log f^{(jj')}(x_{ij}, x_{ij'}, \Sigma_{jj'}).$$

$$f^{(jj')}(x_{ij}, x_{ij'}, \Sigma_{jj'}) =$$

$$\begin{cases} c_{\Sigma_{jj'}}(\hat{F}_j(x_j), \hat{F}_{j'}(x_{j'})) f_j(x_j) f_{j'}(x_{j'}) & \text{when both variables are continuous} \\ f_j(x_j) \int_{\hat{F}_{j'}(x_{j'}-)}^{\hat{F}_{j'}(x_{j'})} c_{\Sigma_{jj'}}(\hat{F}_j(x_j), v) d\lambda(v) & \text{when the pair is mixed} \\ C_{\Sigma_{jj'}}(\hat{F}_j(x_j), \hat{F}_{j'}(x_{j'})) - C_{\Sigma_{jj'}}(\hat{F}_j(x_j-), \hat{F}_{j'}(x_{j'})) & \\ -C_{\Sigma_{jj'}}(\hat{F}_j(x_j), \hat{F}_{j'}(x_{j'}-)) + C_{\Sigma_{jj'}}(\hat{F}_j(x_j-), \hat{F}_{j'}(x_{j'}-)) & \text{otw.} \end{cases}$$

where $\hat{F}_j = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}_{\{x_{ij} \leq x\}}$.

Simulation study

Simulations:

1. **Pairwise** Simulation of four variables:
   - $X_1 \sim \mathcal{N}(0, 1)$
   - $X_2 \sim \mathcal{B}(0.5)$
   - $X_3 \sim \mathcal{P}(1)$
   - $X_4 \sim \mathcal{NB}(1, 0.5)$

   that are linked by a Gaussian copula of correlation coefficient ρ, i.e.

$$F(X_1, X_2, X_3, X_4) = C_\Sigma(F_1(X_1), F_2(X_2), F_3(X_3), F_4(X_4)),$$

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

Simulation study

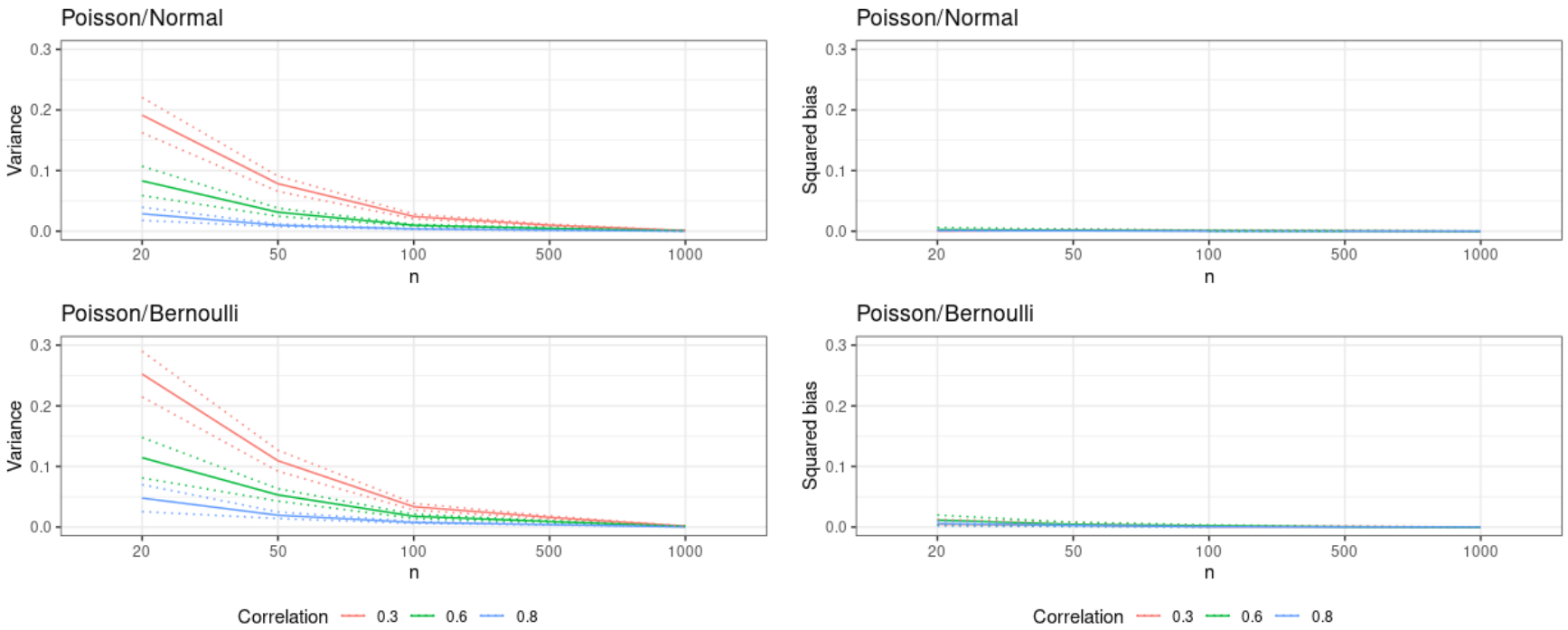


Figure: Variance and squared bias for pairs of variables for N=500 estimations

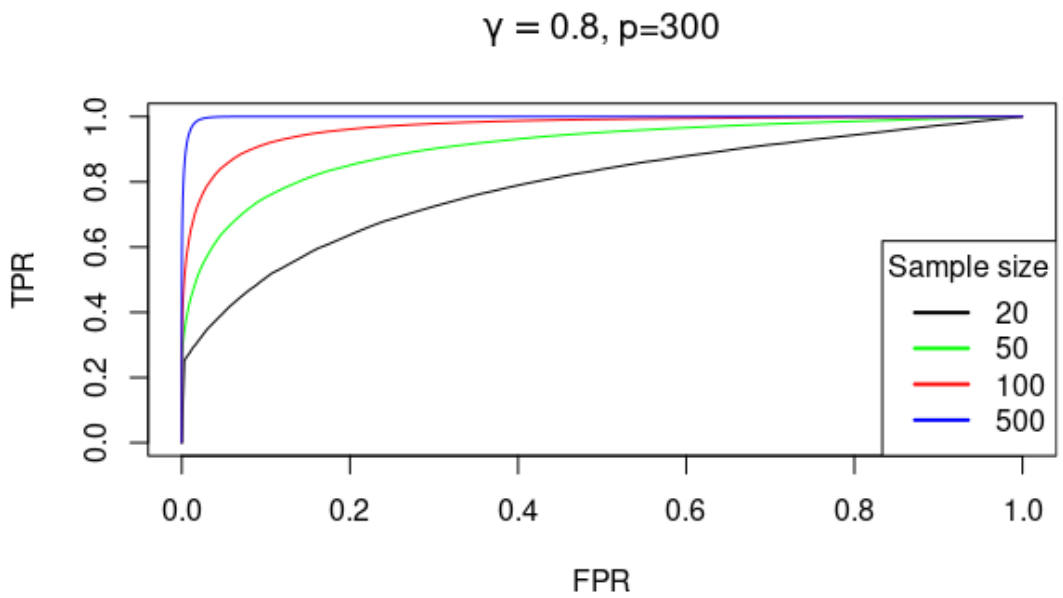
Simulation study

Simulations:

2 **High dimensional** Simulation of $d = 300$ variables:
- $X_1, \ldots, X_{100} \sim \mathcal{N}(0, 1)$
- $X_{101}, \ldots, X_{200} \sim \mathcal{NB}(1, 0.5)$
- $X_{201}, \ldots, X_{300} \sim \mathcal{B}(0.5)$

linked by a Gaussian copula of correlation matrix of sparsity 0.8.

Simulation study



sample size	AUC
20	0.79 (0.01)
50	0.90 (0.007)
100	0.97 (0.004)
500	0.998 (0.0002)

Figure: ROC curve and AUC for 300 variables and different sample sizes

FPR: proportion of coefficients detected as significant among the non-significant ones
TPR: proportion of coefficients correctly detected as significant among the significant ones

Real data

→ International Cancer Genome Consortium dataset[3]

Different:

- cancer types
- countries
- datasets

→ BRCA-US (after pre-processing):

- 250 individuals
- 108 RNAseq (counts)
- 108 proteins (continuous expression)
- 62 mutations (binary)

[3] Zhang et al., The International Cancer Genome Consortium Data Portal., 2019

Coefficient detection

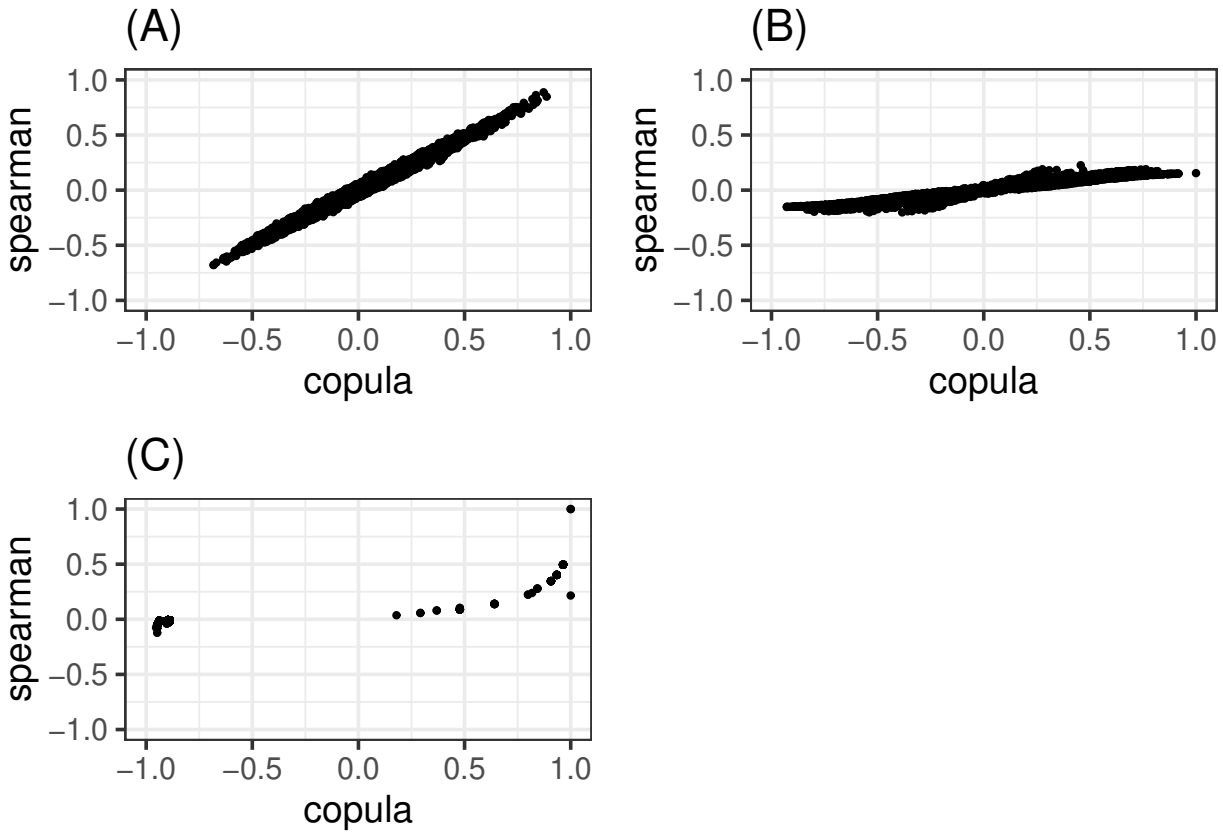


Figure: Comparison of Spearman's ρ^S and the copula correlation coefficient for (A) continuous-continuous pairs, (B) mixed pairs, (C) binary-binary pairs. The RNAseq counts have been classified as continuous because of their high number of values.

Network visualization

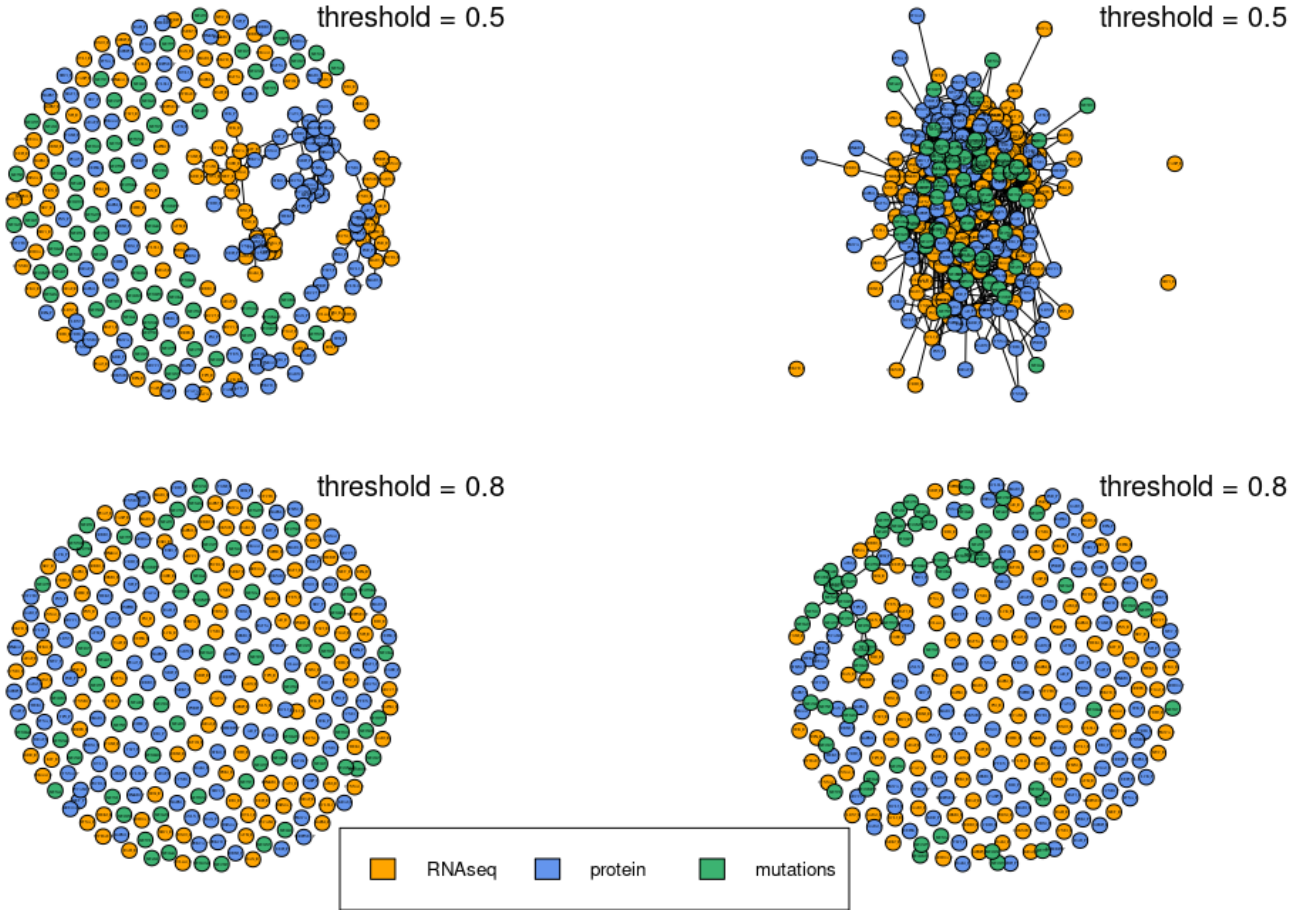


Figure: Correlation networks for Spearman's ρ^S (left) and the copula (right)

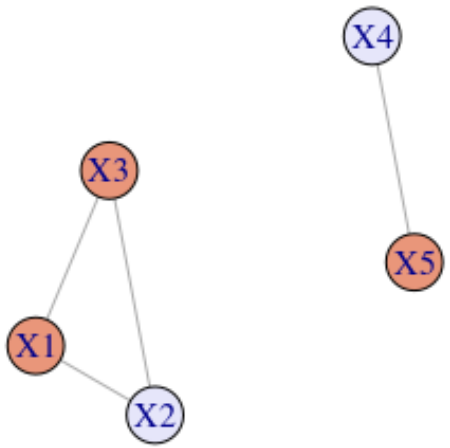
Table of Contents

Table of Contents

Influence of the remaining variables.

X_1, X_3, X_5: RNAseq (counts), X_2 and X_4: transcription factors (binary).

Pairwise relationships

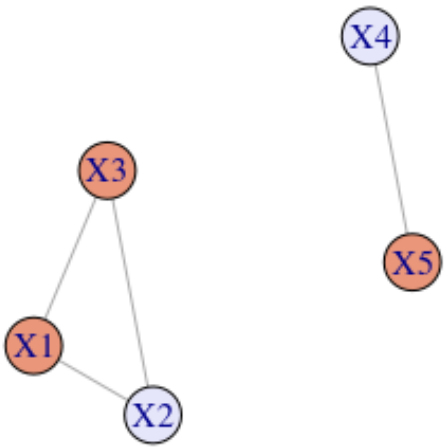


no edge: independence between X_j and $X_k \iff \Sigma_{jk} = 0$

Influence of the remaining variables.

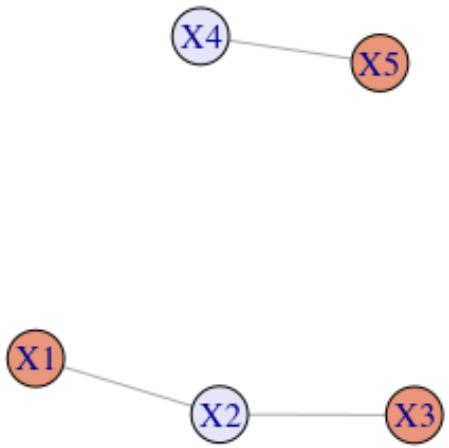
X_1, X_3, X_5: RNAseq (counts), X_2 and X_4: transcription factors (binary).

Pairwise relationships



no edge: independence between X_j and $X_k \iff \Sigma_{jk} = 0$

Global relationships



no edge: independence between X_j and X_k conditionally to the rest of the variables

Graphical models

1. GGM (Krämer et al. (2009)): Gaussian data
2. NPN (Liu et al. (2009)): Gaussian copula for continuous variables

$\rightarrow$ conditional independence relationships encoded in $\Omega = \Sigma^{-1}$

Graphical models

1. GGM (Krämer et al. (2009)): Gaussian data
2. NPN (Liu et al. (2009)): Gaussian copula for continuous variables

$\rightarrow$ conditional independence relationships encoded in $\Omega = \Sigma^{-1}$

What about mixed variables with a Gaussian copula structure?

Graphical models

1. GGM (Krämer et al. (2009)): Gaussian data
2. NPN (Liu et al. (2009)): Gaussian copula for continuous variables

$\rightarrow$ conditional independence relationships encoded in $\Omega = \Sigma^{-1}$

What about mixed variables with a Gaussian copula structure?The **latent conditional dependencies** are encoded in $\Omega = \Sigma^{-1}$.

Penalized inversion

Graphical lasso [4]:

$$\hat{\Omega}_\lambda = \underset{\Omega}{\text{argmin}} \left(\text{tr}(\hat{\Sigma}\Omega) - \log\det(\Omega) + \lambda||\Omega||_1 \right)$$

λ : penalization parameter

[4] Friedman, Hastie, Tibshirani, Biostatistics, 2008
[5] Wang, Kim et Li, The Annals of Statistics, 2013

Penalized inversion

Graphical lasso [4]:

$$\hat{\Omega}_\lambda = \underset{\Omega}{\text{argmin}} \left(\text{tr}(\hat{\Sigma}\Omega) - \log\det(\Omega) + \lambda||\Omega||_1 \right)$$

λ : penalization parameter

Optimal λ choice by the HBIC criterion: [5]

$\lambda_{\text{opt}} = \text{argmin} \left(\text{tr}(\hat{\Sigma}\hat{\Omega}_\lambda) - \log\det(\hat{\Omega}_\lambda) + \log(\log(n))\frac{\log(d)}{n} s_\lambda \right)$ where s_λ is the number of non-null edges in $\hat{\Omega}_\lambda$

[4] Friedman, Hastie, Tibshirani, Biostatistics, 2008
[5] Wang, Kim et Li, The Annals of Statistics, 2013

Application to our estimator

Recall

$$\hat{\Sigma}_{jj'} = \underset{\Sigma_{jj'} \in (-1,1)}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1} \log f^{(jj')}(x_{ij}, x_{ij'}, \Sigma_{jj'}).$$

Graphical lasso:

$$\hat{\Omega}_\lambda = \underset{\Omega}{\operatorname{argmin}} \left(\operatorname{tr}(\hat{\Sigma}\Omega) - \log\det(\Omega) + \lambda ||\Omega||_1 \right)$$

λ : penalization parameter

Optimal λ choice by the HBIC criterion: [5]

$\lambda_{\text{opt}} = \operatorname{argmin} \left(\operatorname{tr}(\hat{\Sigma}\hat{\Omega}_\lambda) - \log\det(\hat{\Omega}_\lambda) + \log(\log(n)) \frac{\log(d)}{n} s_\lambda \right)$ where s_λ is the number of non-null edges in $\hat{\Omega}_\lambda$

[5] Wang, Kim et Li, The Annals of Statistics, 2013

Simulation study

Simulations:

3 **Latent conditional dependence structure** Simulation of $d = 30$ variables:

- $X_1, \ldots, X_{10} \sim \mathcal{N}(0, 1000)$
- $X_{11}, \ldots, X_{20} \sim \mathcal{NB}(1000, 0.3)$
- $X_{21}, \ldots, X_{30} \sim \mathcal{B}(0.5)$

linked by a latent partial correlation matrix P of sparsity 0.8.

Simulation study

Simulations:

3 **Latent conditional dependence structure** Simulation of $d = 30$ variables:

- $X_1, \ldots, X_{10} \sim \mathcal{N}(0, 1000)$
- $X_{11}, \ldots, X_{20} \sim \mathcal{NB}(1000, 0.3)$
- $X_{21}, \ldots, X_{30} \sim \mathcal{B}(0.5)$

linked by a latent partial correlation matrix P of sparsity 0.8.

$$F(X_1, \ldots, X_{30}) = C_\Sigma(F_1(X_1), \ldots, F_{30}(X_{30})),$$

where $\Omega = \Sigma^{-1}$ is of sparsity 0.8 and so $P_{jk} = \dfrac{-\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}}$ also is of sparsity 0.8.

Simulation study

PPMLE vs. Bridge function (Zhang et al., 2021): ROC curve for $\lambda \in \log\{1.01, \ldots, 3\}$.

Sample size	20	50	200	500
PPMLE	0.67	0.71	0.77	0.78
Bridge	0.63	0.66	0.72	0.76

Table: Average AUC values over all pairs for N=100 replications.

Simulation study

PPMLE vs. Bridge function (Zhang et al., 2021): ROC curve for $\lambda \in \log\{1.01, \ldots, 3\}$.

Sample size	20	50	200	500
PPMLE	0.67	0.71	0.77	0.78
Bridge	0.63	0.66	0.72	0.76

Table: Average AUC values over all pairs for N=100 replications.

	AUC on discrete pairs		AUC on continuous-discrete pairs	
Sample size	PPMLE	Bridge	PPMLE	Bridge
20	0.65	0.60	0.84	0.73
50	0.70	0.60	0.91	0.76
200	0.77	0.64	0.93	0.83
500	0.80	0.69	0.93	0.92

Table: Average AUC values over all pairs involving discrete variables for N=100 replications.

ICGC data network

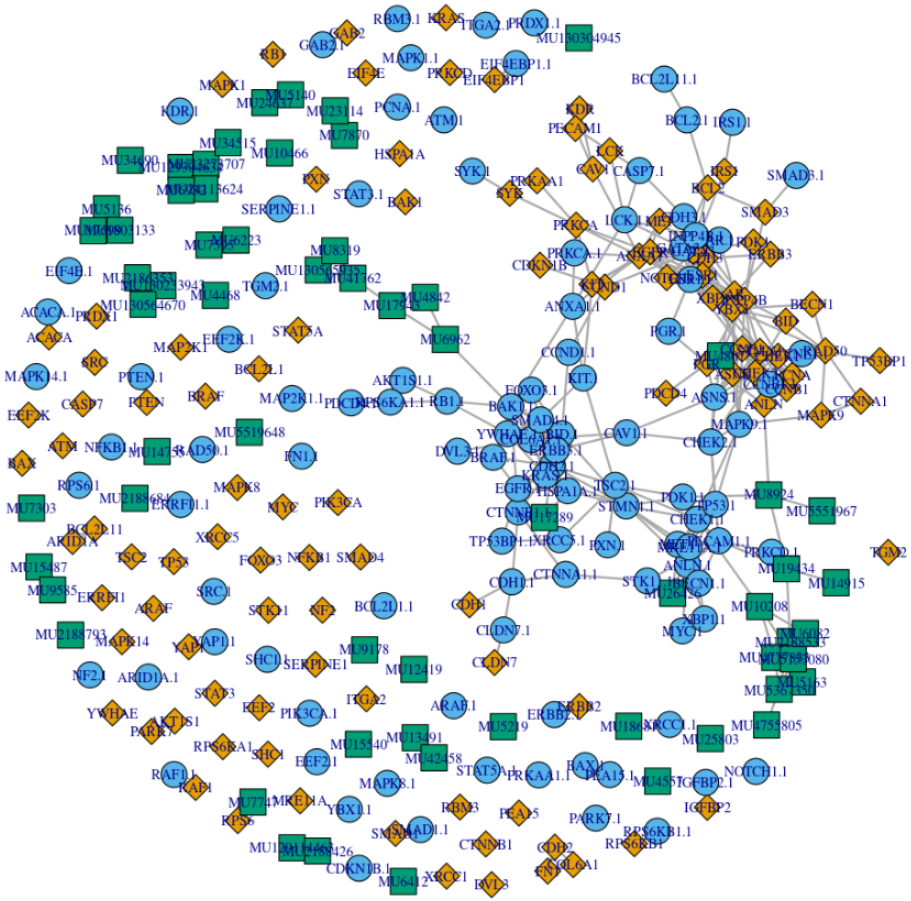


Figure: Latent partial correlation graph

Table of Contents

Table of Contents

Thesis contributions

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- **Application to a real-life INRAE data set.**
- Availability on CRAN.

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- Mixed density and independence properties.
- Novel estimation approach of the correlation matrix.

Real data

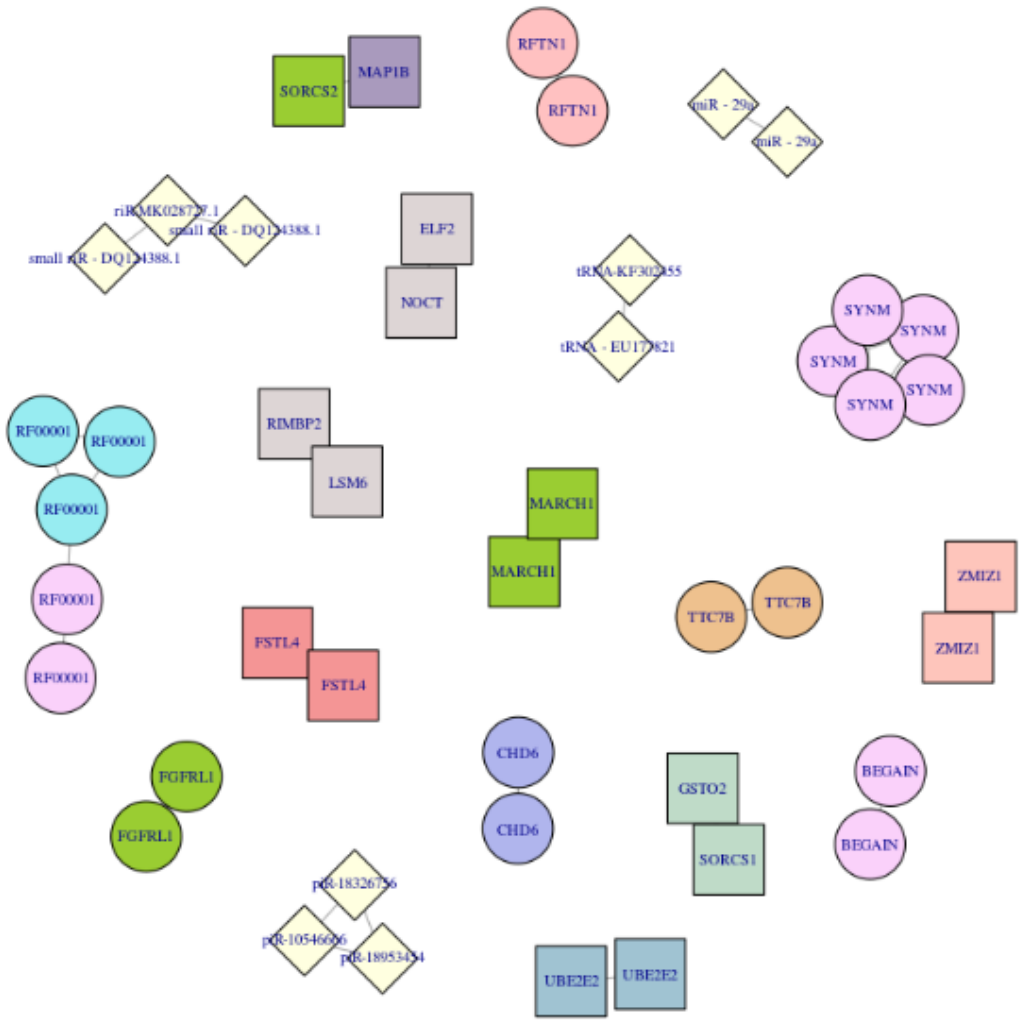
INRAE data set:[6]

1. NR56 (=pregnancy rate after two cycles) measured per bull, $n = 519$
2. two groups: fertile ($n_F = 51$) and subfertile ($n_S = 47$)
3. $d = 485$ variables: 183 CpGs, 159 SNPs, 143 sncRNAs

$\rightarrow$ are there any differences in the regulation networks?

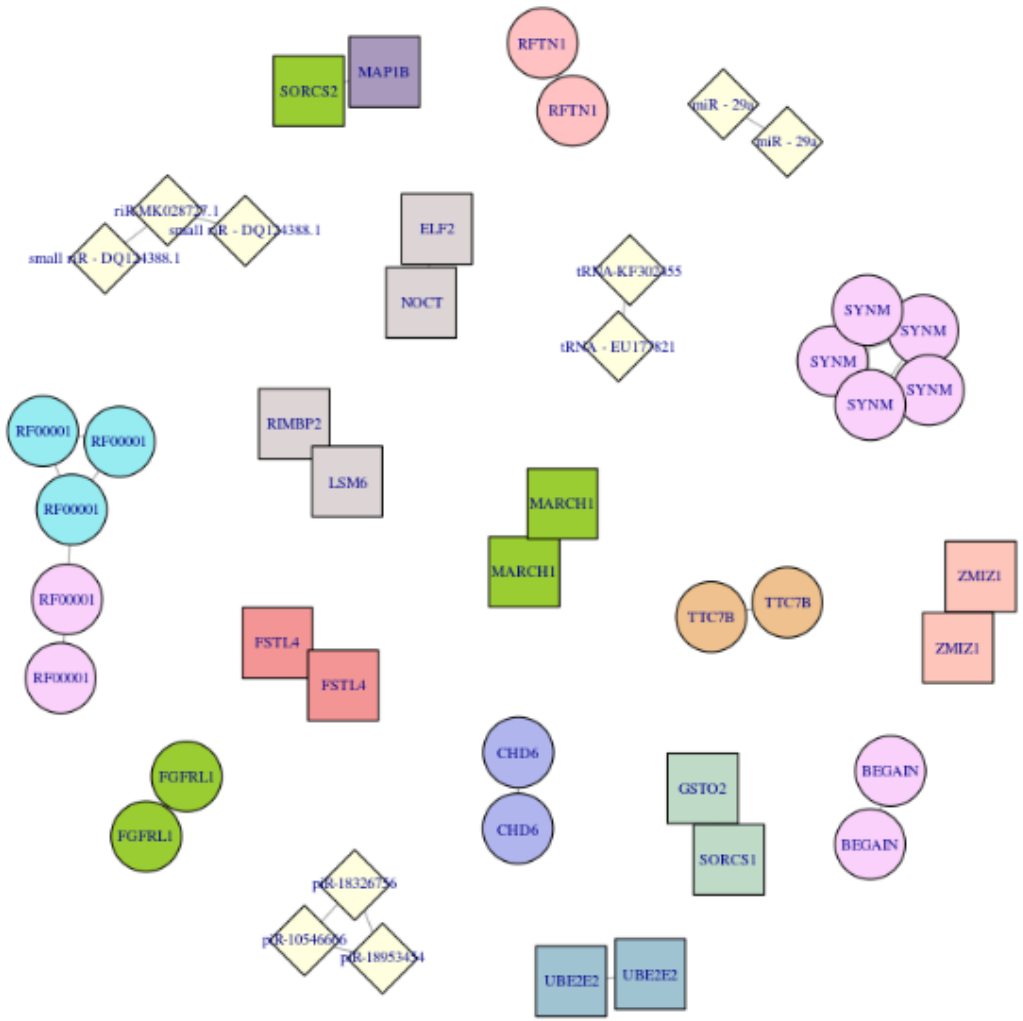
[6] Costes et al. 2022

Fertile bulls



circle: CpG, square: SNP, diamond: sncRNA, color: chromosome

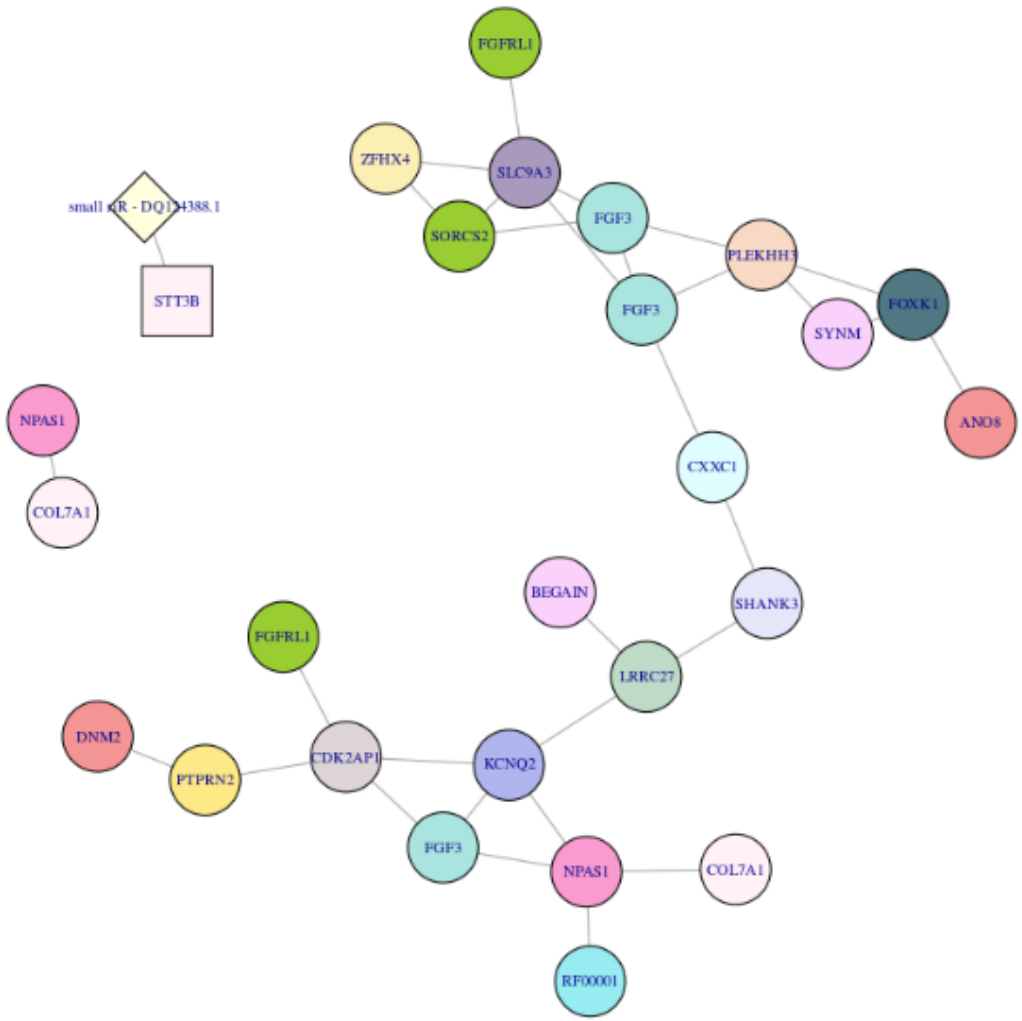
Fertile bulls



circle: CpG, square: SNP, diamond: sncRNA, color: chromosome

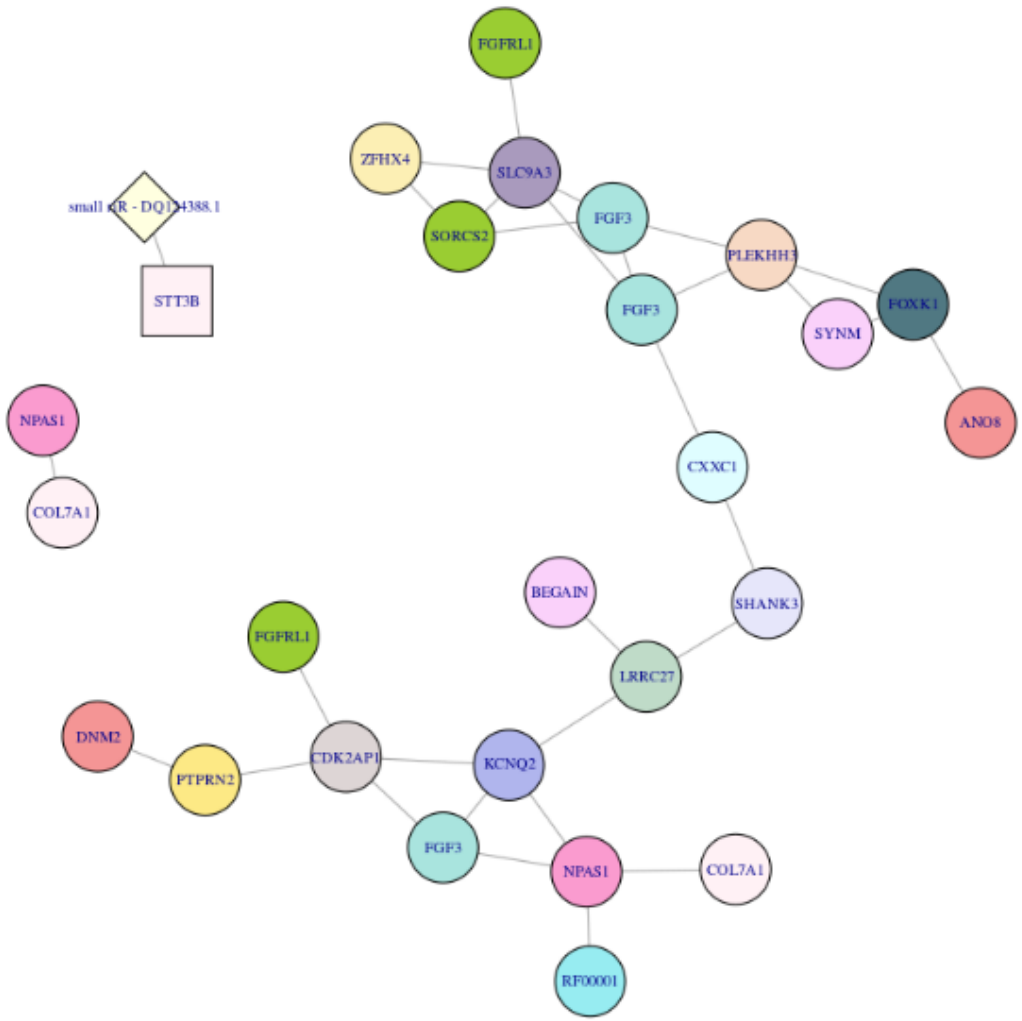
MAP1B& SORCS2: brain development

Subfertile bulls



circle: CpG, square: SNP, diamond: sncRNA, color: chromosome

Subfertile bulls



circle: CpG, square: SNP, diamond: sncRNA, color: chromosome COL7A1: skin disorder, NPAS1: central nervous system dvpt

Subfertile bulls

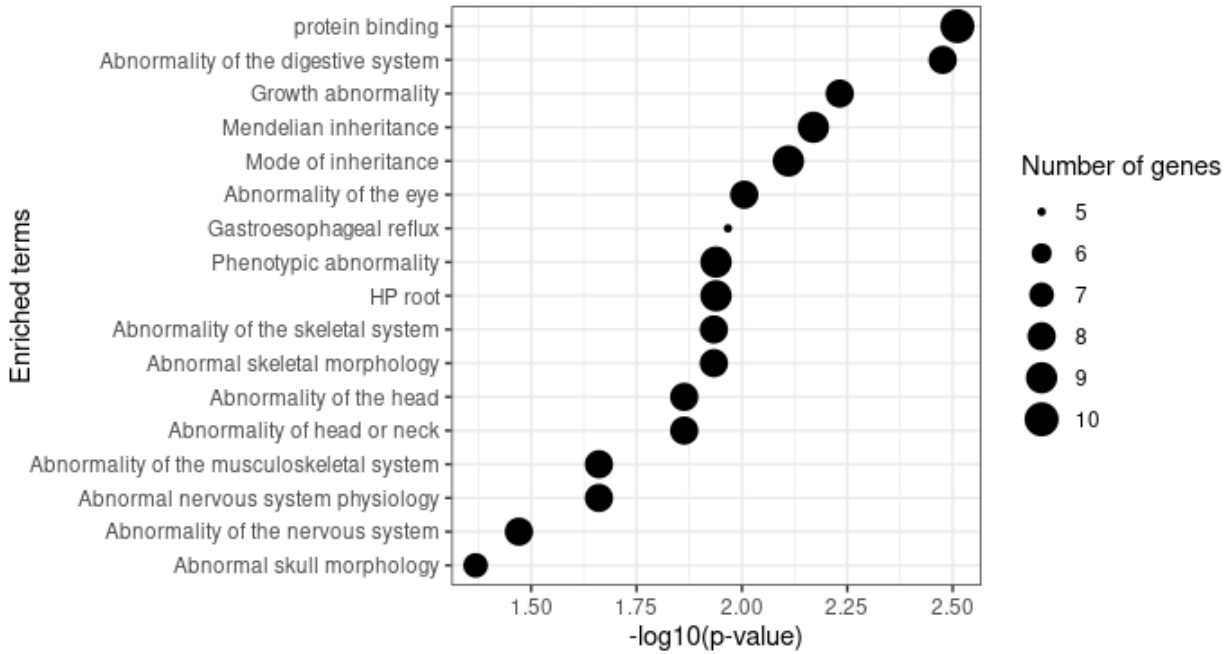


Figure: Functional enrichment dot plot obtained via the `ensembldb` package

Table of Contents

Table of Contents

Thesis contributions

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- Application to a real-life INRAE data set.
- Availability on CRAN.

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- Mixed density and independence properties.
- Novel estimation approach of the correlation matrix.

Availability

- R package `heterocop` available on CRAN
- RShiny version `https://shiny-heterocop.sk8.inrae.fr/`

Availability

- R package `heterocop` available on CRAN
- RShiny version `https://shiny-heterocop.sk8.inrae.fr/`

Pre-prints:

- Estimation and properties of Σ: Gaussian copula correlation network analysis with application to multi-omics data, `https://hal.inrae.fr/hal-04847648`. Under revision.
- Estimation of Ω and comparison with the bridge functions method: Multi-omics network inference with a Gaussian copula model, `https://hal.inrae.fr/hal-05173829v1`. Under revision.

Work in progress

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- Application to a real-life INRAE data set.
- Availability on CRAN.
- **Biological prior integration.**

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- Mixed density and independence properties.
- Novel estimation approach of the correlation matrix.
- **Theoretical properties of the estimator.**

Work in progress

1. Integration of a biological prior on Ω via the coefficient-wise penalized gLasso approach

$$\hat{\Omega}_\lambda = \underset{\Omega}{\operatorname{argmin}} \left(\operatorname{tr}(\hat{\Sigma}\Omega) - \log\det(\Omega) + ||\Lambda\Omega||_1 \right),$$

where Λ is a matrix of the size of Σ and Ω - is it better to favor close biological entities? How can we define "closeness"?

Work in progress

2 Proof of consistency/asymptotical normality

Work in progress

2. Proof of consistency/asymptotical normality

Sketch of proof:

1. Check that in all cases, for a pair (X_j, X_k), the empirical mean of the score function has a root on $(-1, 1)$, and uniformly converges to a function which has a unique root on $(-1, 1)$.

Work in progress

2. Proof of consistency/asymptotical normality

Sketch of proof:

1. Check that in all cases, for a pair (X_j, X_k), the empirical mean of the score function has a root on $(-1, 1)$, and uniformly converges to a function which has a unique root on $(-1, 1)$.
2. Check conditions from Bickel et al. (1993) for consistency and asymptotical normality of $GM-$estimates.

Work in progress

2. Proof of consistency/asymptotical normality

Sketch of proof:

1. Check that in all cases, for a pair (X_j, X_k), the empirical mean of the score function has a root on $(-1, 1)$, and uniformly converges to a function which has a unique root on $(-1, 1)$.
2. Check conditions from Bickel et al. (1993) for consistency and asymptotical normality of $GM-$estimates.

Tools:

- Boundedness of the first and second-order derivatives of the score function.
- Term decomposition inspired by Ryumgaart (1974), based on an independent copy of the observations.

Remaining questions

Multi-omics network inference

- Novel approach for simultaneous multi-omics integration.
- Application to a real-life INRAE data set.
- Availability on CRAN.
- **Biological prior integration.**
- **How to handle non-monotonic relationships?**
- **How to interpret the values of Ω?**

Gaussian copula model

- Interpretation of extreme parameter values for mixed variables.
- Mixed density and independence properties.
- Novel estimation approach of the correlation matrix.
- **Theoretical properties of the estimator.**
- **How to estimate the variance of $\hat{\Sigma}$ and perform thresholding?**
- **Is the HBIC a suitable criterion?**

Additional slides

Expressions of CC and MI

$$\rho^P(X,Y) = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{(\mathbb{E}[X^2] - \mathbb{E}[X]^2)(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2)}}$$

$$\rho^S(X,Y) = \rho^P(r(X), r(Y))$$

$$MI(X,Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy$$

Expression of C_Σ^p

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

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

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

and I denotes the identity matrix.

Parametric misspecification

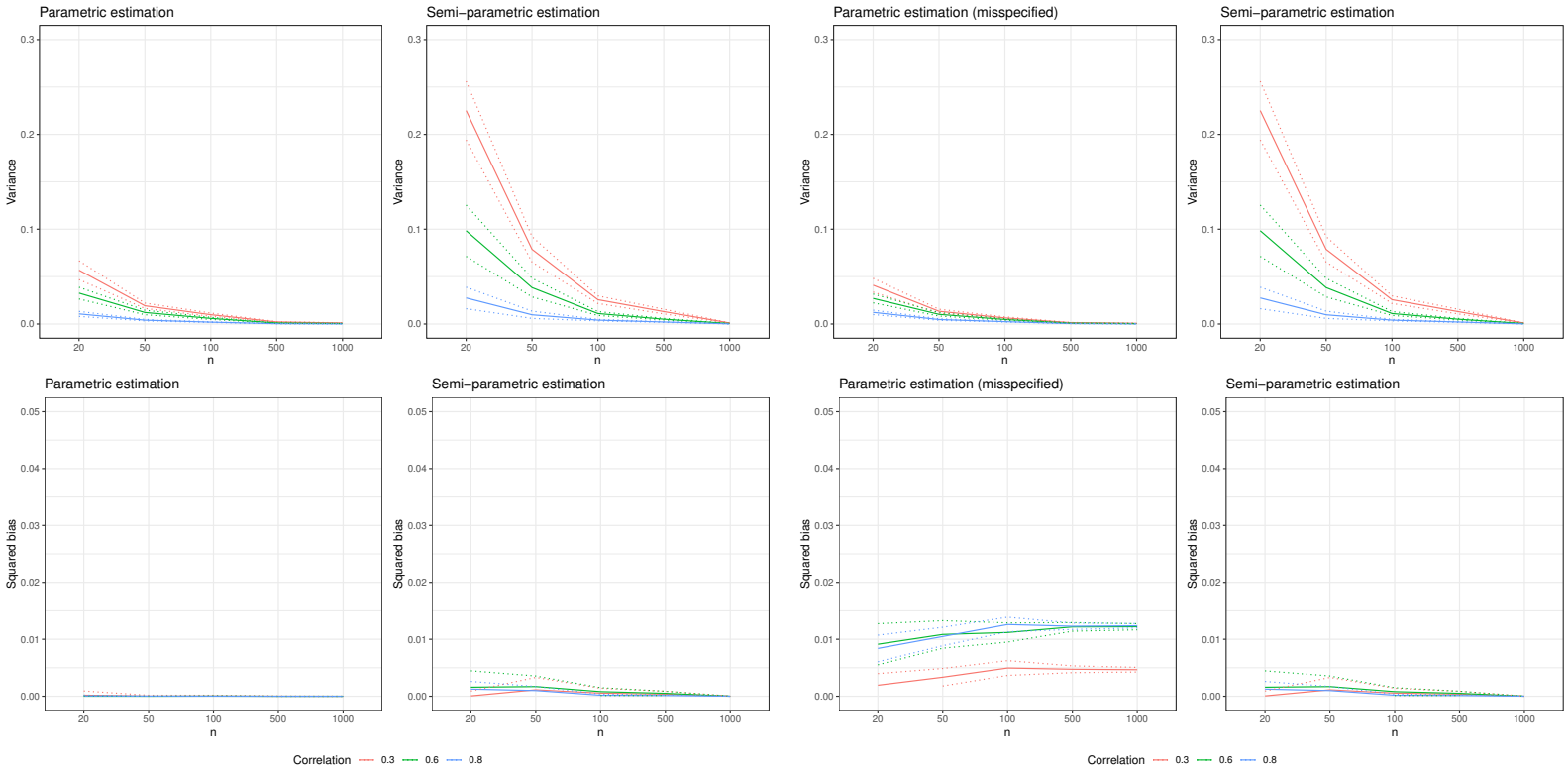


Figure: Variance and bias of $\rho = 0.3, 0.6, 0.8$ parametrically estimated between a $\mathcal{N}(0, 1)$ and a $NB(1, 0.5)$ distribution, where the NB has been correctly specified (left) or misspecified for a poisson (right).

Computational time

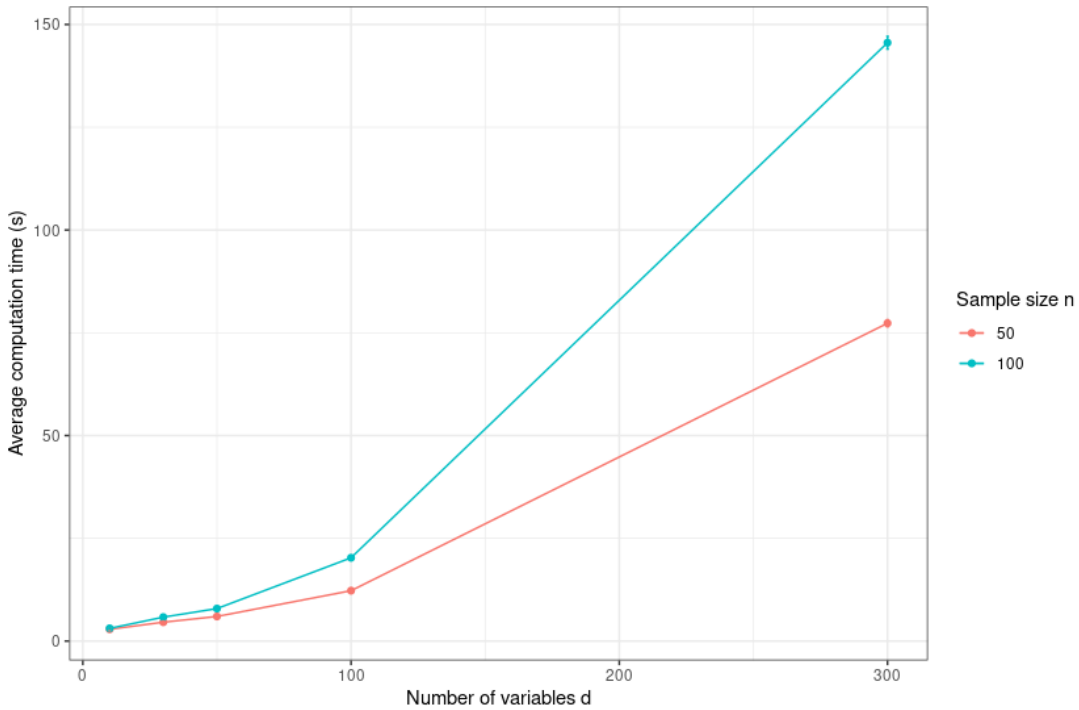
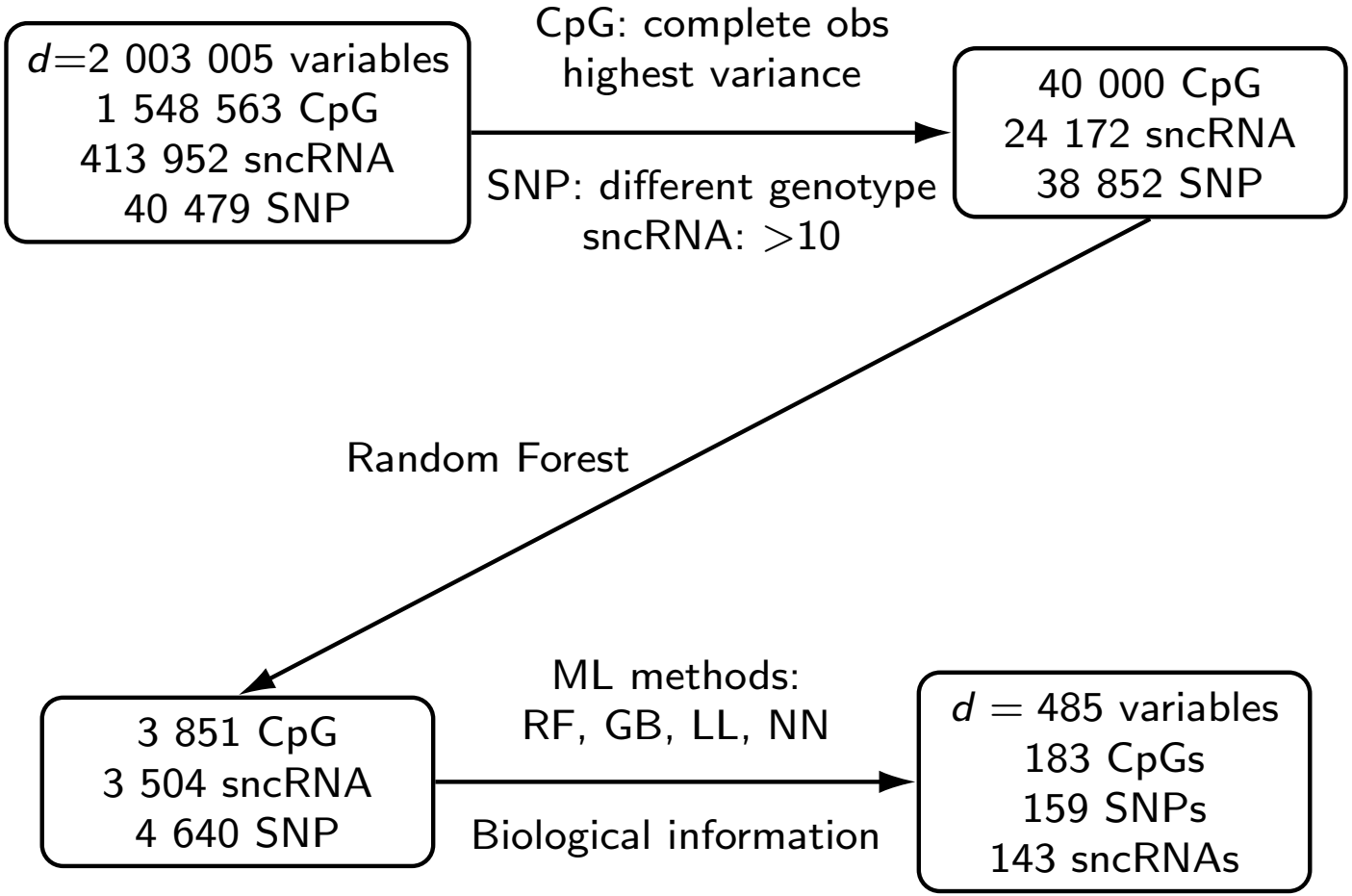


Figure: Computational time for the estimation of Σ for the PPMLE.

INRAE bull fertility data set

Pre-processing [7]



[7] Costes et al. 2022