Bayesian Copula Directional Dependence for causal inference on gene expression data

Vasiliki Vamvaka, Dr. Clara Grazian
School of Mathematics & Statistics, University of New South Wales, Sydney
v.vamvaka@student.unsw.edu.au, c.grazian@unsw.edu.au

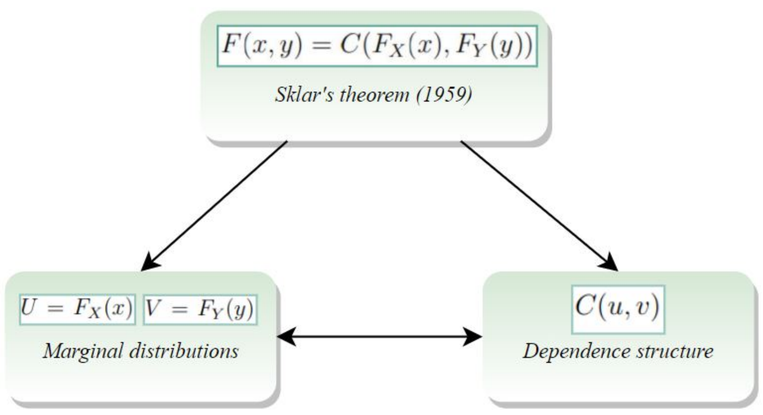
1. Introduction

The process of gene expression is observed in each cell of every living organism to determine the cell's functionality and survival. Regulatory networks are the essential building blocks that control both the expression of proteins and the creation of different types of cells. Copula models are probabilistic tools able to construct multivariate probability distribution functions by combining multiple uniformly distributed marginals. Bayesian methods analyzing copula models usually apply specific copula families to the data, increasing estimation biases and model miss-specifications. Herein, the Bayesian Copula Directional Dependence (Bayesian CDD), a novel method for assessing the direction of influence between genes based on their dependence structure, which avoids the definition of their copula function.

2. Why Copulas?

- Separate marginal distributions from dependence structure
- Unveil the true nature of dependence between variables
- Able to model asymmetric dependency between extreme events
- Flexible for multivariate modelling
- If the variables are continuous then their copula is unique

A copula $C : [0,1]^2 \to [0,1]$ is a cumulative distribution function with uniform marginals, which constructs the joint distribution of (X, Y) [1].



3. Copula Directional Dependence

A pair (X, Y) is directionally dependent in the joint behaviour if and only if the conditional copulas $C_v(u)$ and $C_u(v)$ are different [2].
The copula regressions in each direction are:

$$\mu_{U|V(v)} = E(U|V = v) = 1 - \int_0^1 C_v(u)du$$

$$\mu_{V|U(u)} = E(V|U = u) = 1 - \int_0^1 C_u(v)dv$$

By introducing $Beta(\mu, \kappa)$ regression [3, 4] the conditional means become:

$$\mu_{U|V(v)} = \frac{exp(\beta_0 + \beta_1 v)}{1 + exp(\beta_0 + \beta_1 v)}$$

$$\mu_{V|U(u)} = \frac{exp(\beta_0 + \beta_1 u)}{1 + exp(\beta_0 + \beta_1 u)}$$

while the copula is modelled as an error term, $\epsilon \sim MVN(0, \Omega)$, where Ω corresponds to the form of the dependence.
The strength of the directional dependence is measured by:

$$\rho^2_{U\to V} = \frac{Var(\mu_{V|U})}{Var(V)}$$

$$\rho^2_{V\to U} = \frac{Var(\mu_{U|V})}{Var(U)}$$

with higher values indicating stronger direction.

4. Bayesian Copula Directional Dependence method

The Bayesian CDD method extends the frequentist CDD [3] to the Bayesian framework by introducing uncertainty and more interpretable results.

1. Transform genes (X, Y) to $(U, V) \in [0,1]^2$
2. Initialize $(\beta_0, \beta_1, \kappa)$
3. Set $\mu_U = \frac{exp(\beta_0+\beta_1 v)}{1+exp(\beta_0+\beta_1 v)} \quad \mu_V = \frac{exp(\beta_0+\beta_1 u)}{1+exp(\beta_0+\beta_1 u)}$
4. Likelihoods $(U|V = v) \sim Beta(\mu_U, \kappa) \quad (V|U = u) \sim Beta(\mu_V, \kappa)$
5. Priors $(\beta_0, \beta_1) \sim \mathcal{N}(0, 10)$ and $\kappa \sim \Gamma(1, 1)$
6. Posterior $\propto$ priors $\times$ likelihood
7. **Output:** posterior samples for the coefficients β_0 and β_1
8. Use **step 7** samples to calculate

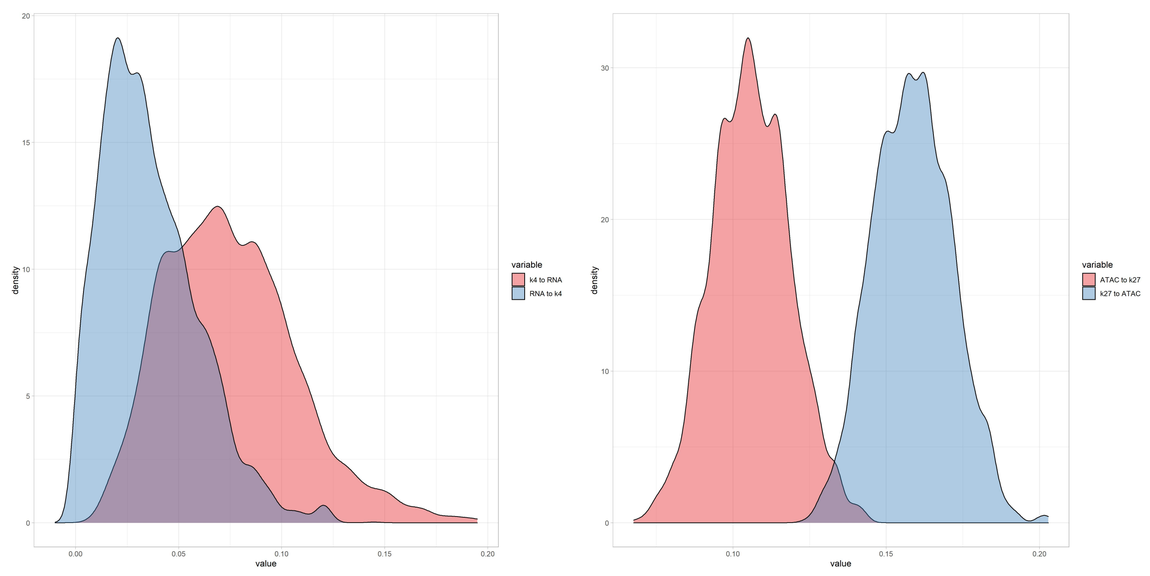
$$\rho^2_{V\to U} = 12Var(\mu_U) \quad \rho^2_{U\to V} = 12Var(\mu_V)$$

9. If $\rho^2_{U\to V} > \rho^2_{V\to U}$ then the direction is $U \to V$

5. Results

The plots illustrate the posterior directional dependence densities from two directions; $\rho^2_{U\to V}$ and $\rho^2_{V\to U}$, for genes of the *Drosophila* embryos bulk epigenome data. The density with the higher values indicates the stronger direction.

- **Left plot:** Histone mark enrichment K4 (red) $\to$ RNA (blue); mark enrichment k4 is always present where genes are expressed
- **Right plot:** Histone mark enrichment K27 (blue)$\to$ ATAC-seq (red); k27 is enriched at regions that bind proteins that open up ATAC-seq; the gene that regulates chromatin accessibility



6. Conclusion

The novel Bayesian CDD method, can be used for construction of gene interactions. The advantages of the method are:

- Uncertainty, robustness and no parametric assumptions on the estimates of the dependence structure
- Able to measure strength of the dependence
- Can be used in higher dimensions for inference of full regulatory gene networks

7. References

[1] M Sklar. Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universit de Paris*, 8:229–231, 1959.

[2] Engin A Sungur. A note on directional dependence in regression setting. *Communications in Statistics—Theory and Methods*, 34(9-10):1957–1965, 2005.

[3] Jong-Min Kim and Sun-Young Hwang. Directional dependence via gaussian copula beta regression model with asymmetric garch marginals. *Communications in Statistics-Simulation and Computation*, 46(10):7639–7653, 2017.

[4] Namgil Lee and Jong-Min Kim. Copula directional dependence for inference and statistical analysis of whole-brain connectivity from fmri data. *Brain and behavior*, 9(1):e01191, 2019.