Proposition de sujet de thèse

Titre/*Title* Méthodes à copules pour l'inférence de réseaux de régulation multi-omiques/ *Copula-based network inference for multi-omics data*

Résumé/*Summary* Pour mieux comprendre les interactions entre les différents objets qui constituent un réseau biologique (gènes, protéines, etc), les biologistes réalisent des mesures sur des variables de différents types : catégorielles, ordinales, continues. La "découverte" des inter-dépendances dans ces données hétérogènes (on parle de données "multi-omiques" en biologie) est un défi majeur à la fois en biologie et en statistique. L'objectif de cette thèse est de construire un modèle statistique pour inférer ces inter-dépendances en modélisant l'hétérogénéité des données à l'aide de copules. Ces dernières sont des fonctions qui permettent de lier entre elles les variables de différents types. La méthode d'inférence devra être étudiée théoriquement et numériquement, avant de l'appliquer sur un jeu de données multi-omiques produit à l'INRAE. *To better understand the relationships between different objects that comprise a biological network (genes, proteins, etc), biologists observe variables of various types: categorical, ordinal, continuous. The "discovery" of the inter-dependencies in those heterogeneous data (also known as "multi-omics" data in biology) is a genuine challenge both in biology and statistics. The goal of this PhD thesis is to build a statistical model to infer those inter-dependencies by modeling the heterogeneity in the data with copulas, which are functions that can couple variables of varying types. The estimation method will be examined both theoretically and numerically, and will be applied to a multi-omics dataset produced by INRAE.*

Environnement et déroulement de la thèse Le futur doctorant sera rattaché à l'Ecole Doctorale de Mathématique Hadamard et Université Paris-Saclay, et membre des unités MAIAGE (Mathématiques et Informatique Appliquées du Génome à l'Environnement) et GABI (Génétique Animale et Biologie Intégrative) de INRAE, Allée de Vilvert, 78350 Jouy-en-Josas, France. Il sera encadré par Gildas Mazo (MAIAGE) et Florence Jaffrézic (GABI) et pourra bénéficier de collaborations avec Andrea Rau (unités GABI and BioEcoAgro, INRAE), Dimitris Karlis (Athens University of Economics and Business) et plus largement avec les partenaires du projet DINAMIC, financé par l'INRAE. Il est prévu que la thèse, qui dure trois ans, démarre entre le 1er septembre et le 1er décembre 2022.

Candidat recherché Le candidat est issu d'un Master 2 en statistique ou biostatistique. Il a de bonnes connaissances en statistique mathématique et inférentielle et de bonnes aptitudes en programmation.

Contact Pour envoyer votre candidature (CV, lettre de motivation, liste des cours suivis et toutes les notes de master) ou demander des informations supplémentaires, contactez Gildas Mazo (`gildas.mazo@inrae.fr`) ou Florence Jaffrézic (`florence.jaffrezic@inrae.fr`).

1 Context, positioning and objectives

1.1 Context

Next generation sequencing (NGS) technologies have given rise to a tsunami of biological data at unprecedented resolution, accuracy and scale. The extraordinary complexity of these data brings about immense challenges for mathematicians, statisticians and computational biologists. For instance, high-throughput transcriptome sequencing (RNA-seq) data typically comprise tens or hundreds of millions of reads, from which transcript expression levels, co-expression patterns, and co-regulatory dynamics must be inferred. These difficult tasks are often part of a grander goal: discovering regulation networks from multi-omics data (that is, representing different levels of molecular variability: genomics, transcriptomics, proteomics, metabolomics, epigenomics, etc.) However, despite the rapid development of statistical tools to handle NGS data, the field is still young and more research is needed to unlock the full potential of NGS data to improve our knowledge of complex biological systems[1].

[1] K. A. Do, Z. S. Qin and M. Vannucci (Eds.). (2013). *Advances in statistical bioinformatics: models and integrative inference for high-throughput data.* Cambridge University Press.

	goal	
type of data	*classical inference*	*network inference*
continuous	classical statistics	glasso, continuous transformations
discrete	P1	P2
mixed	P3	P4

Table 1: Schematic state of current network research.

One key statistical challenge in this area is the problem of network inference. One aims to identify the inter-dependencies among the random variables in a biological system to generate biological insight into regulation pathways. The standard approach for continuous data is the *glasso* (graphical least absolute shrinkage and selection operator[2]), which is an extension of Lauritzen's estimation method in Gaussian Markov graphical models[3]. For Gaussian distributions, the conditional independence relationships characterising the underlying graph are defined with respect to (w.r.t.) the precision matrix (the inverse of the variance-covariance matrix). Since the likelihood function is in fact a function of the precision matrix alone (up to marginal parameters), maximizing the likelihood function amounts to finding the true precision matrix, and thus, the true underlying graph. The *glasso* incorporates a penalty term to the log-likelihood so as to favor sparse networks, a necessary assumption in higher dimensions.

Like many statistical methods, the *glasso* relies on the assumption of Gaussian distributions. If the data are not Gaussian, as is typically the case of complex, multi-omics data, then the correspondence between the Markov graph and the precision matrix breaks down, and the *glasso* may not work. To address this issue, Liu et al[4] assumed that the data could be assumed to be Gaussian after applying a suitable continuous transformation, allowing *glasso* to be used for non-Gaussian but continuous data. This method is theoretically valid if the marginal cumulative distribution functions (c.d.f.'s) of the data are continuous. If however, they have discontinuities, as is the case with discrete data, then such a transformation cannot permit a return to the Gaussian space. Many NGS data are discrete (e.g., read counts of transcripts or genes, or categorical genotype data) and hence have discontinuities in their c.d.f.'s. How, then, can multivariate models be constructed for inferring dependency networks among discrete variables? A second problem arises when integrating multi-omic data from different biological levels (transcriptomics, proteomics, etc.), with potentially different types of measures (continuous, count, binary): how can one "couple" these heterogeneous, multi-omic datasets?

1.2 Positioning

A schematic state of research in inference for continuous, discrete and mixed-data is given in Table 1. Here classical inference refers to the case where the correlation matrix characterising the model is known up to the value of its parameters, while network inference refers to the case where the structure of the correlation matrix is itself unknown. Methods have been proposed to address both challenges with discrete (P1, P2) and mixed data (P3, P4), but, contrary to the *glasso*, no generic method has emerged. For instance, latent variable models can deal with discrete variables, but every conditional distribution must be specified w.r.t. each one, and it is unclear what they represent. Moreover, a model built for one type of data may not be transferable to another, meaning that a new model must be designed for each type. Concerning Lauritzen's mixed graphical models, they assume that continuous data are Gaussian given the discrete data, but the joint distribution of the discrete data must be specified, which is precisely one of the issue one wants to solve here.

The goal is to reduce the problems P1, P2, P3 and P4 of Table 1 to a single one, thanks to copula theory. A copula is a function C that can "couple" marginal c.d.f.'s to define a multivariate c.d.f. with prescribed marginal distributions. More precisely, if $F_1, \ldots, F_d$ are arbitrary c.d.f.'s and C is an arbitrary copula, then

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \tag{1}$$

[2]J. Friedman, T. Hastie, T. and R. Tibshirani (2008). *Sparse inverse covariance estimation with the graphical lasso.* Biostatistics.

[3]S. L. Lauritzen (1996). *Graphical Models.* Oxford Science Publications.

[4]H. Liu, J. Lafferty and L. Wasserman (2009). *The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs.* Journal of Machine Learning Research.

is a valid multivariate c.d.f., called a copula-based model. For instance, one may couple negative binomial, Bernoulli and Gaussian c.d.f.'s together with a copula to yield a multivariate model, the dependency structure of which would depend on the chosen copula. It will be sufficient here to consider the Gaussian copula (not to be confounded with the Gaussian distribution!), parameterized by a correlation matrix.

Doing the inference in copula models for discrete or mixed data is challenging. One has to derive the likelihood function of (1), which for discrete data is a sum with 2^d terms (intractable in high dimensions) and for mixed data is quite nonstandard (one would have to manage integrals w.r.t. a mix of counting and Lebesgue measures).

1.3 Objectives

The main objective is to develop theoretically valid and scalable computational methods to perform network inference in copula-based models of the form (1) for data of any type, including mixed-type data. The driving motivation behind this methodological development is the task of elucidating regulation networks from complex multi-omics data. In particular, these statistical and computational developments will be guided by the use of several real-world multi-omics data from funded projects at INRAE, including epigenomic, transcriptomic, and phenotypic data from a study focused on bull fertility[5] and genotypic, metabolomic, and proteomic data from a study focused on maize tolerance to temperature stress[6]. In both of these projects, a key question of scientific interest that remains thus far unaddressed is the discovery and estimation of inter-dependencies among the ensemble of multi-omic variables.

2 Methods and expected results

The first task consists of developing new statistical methods to infer relationships between variables of arbitrary type in copula-based models of the form (1). To do this, instead of specifying every conditional distribution between the discrete and continuous variables, one can work directly with the Radon-Nikodym density w.r.t. the product measure $\mu^{d_1} \otimes \lambda^{d_2}$, where μ and λ denote the counting measure and the Lebesgue measure, respectively, and d_1 and d_2 denote the number of discrete and continuous variables, respectively. To estimate the elements of the copula correlation matrix, one can use a tractable surrogate of the likelihood function, called the pairwise likelihood, that needs only bivariate densities. (The pairwise likelihood is known to yield $\sqrt{n}$-consistent estimators under mild conditions, where n is the number of observations of the statistical experiment.)

To infer the relationships between the variables, one can explore two strategies. In the first, recent algorithms can be extended to discover homogeneous blocks in the estimated correlation matrix, allowing to group the variables that tend to be active together. Whether a joint estimation of the parameters and the blocks can be carried out in a single step may also be investigated by adding a penalty term to the pairwise log-likelihood. In the second, the discrete c.d.f.'s can be replaced by continuous approximations in the estimated copula arguments to generate synthetic continuous data with the same estimated correlations as the original data, allowing for the use of *glasso* after renormalization. A mathematical analysis will be undertaken to find the conditions under which the proposed methods yield consistent inference.

The second task focuses on scaling up the proposed statistical algorithms to address real-world, high-dimensional, multi-omics datasets. The challenge consists of reducing the computational complexity without deteriorating the precision of the estimator. One can leverage the method proposed in (Mazo et al, 2021) that maximizes a randomized version of the pairwise log-likelihood, allowing to control the tradeoff between the precision of the parameter estimates and the computational complexity. The standard error of the estimator is of order $\sqrt{\alpha/n}$, where here α denotes the amount by which the computing time is divided. Thus, the computational gain is necessarily offset by a loss of precision of the estimator. However, using the divide-and-conquer philosophy, it can be conjectured that the factor α can be removed (leading to a standard error of order $\sqrt{1/n}$) if the pairwise log-likelihood is split into a large number of small independent random functions maximized simultaneously on many

[5] INRAE BREED Research Unit; ANR SeQuaMol project

[6] INRAE BioEcoAgro Research Unit; ANR-PIA AMAIZING

computing processors. This conjecture will be checked both in theory and on extensive simulations across a computer cluster.

3 References related to the project

• **Mazo, G.**, **Karlis, D.**, **Rau, A.** (2021) A randomized pairwise likelihood method for complex statistical inferences.`https://hal.archives-ouvertes.fr/hal-03126620`.
• Hulot A., Laloë D., **Jaffrézic F.** (2021). A unified framework for the integration of multiple hierarchical clusterings or networks from multi-source data. BMC Bioinformatics, 22(1):392.