Simulation of data-driven multi-omic benchmark data for cellular deconvolution methods evaluation

Hugo Barbot, David Causeur, Yuna Blum, Magali Richard

Simulation of data-driven multi-omic benchmark data for cellular deconvolution methods evaluation

Hugo Barbot[1], David Causeur[1], Yuna Blum[2], Magali Richard[3]

[1] IRMAR - UMR CNRS 6625, [2] IGDR - UMR CNRS 6290, , [3] TIMC - UMR CNRS 5525

Cell Deconvolution

Cellular heterogeneity in a bulk:

→ refers to the variety of cell types within the bulk,

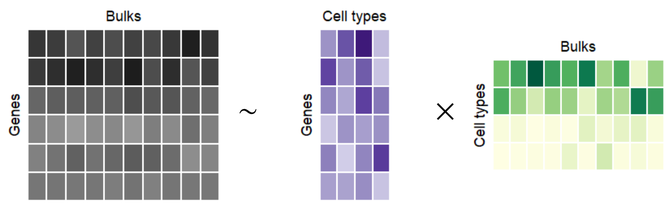
→ reflects progression of **disease state**,

→ is a **complex mixture** signal,

→ is **difficult to assess** from bulk molecular profiles.

⇒ Cell deconvolution **infers** relative abundance of cell types using **one or more -omic data** [1].

Y bulk matrix $M \times N$ ~ X reference matrix $M \times K$ × β proportion matrix $K \times N$



More robust methods + testing limits ⟹ Simulations → recreating the complex variability; → from many omics data types

Dependance with high dimensionnality

For now, we want **control on 4 hypotheses** of our deconvolution model based on Ordinary Least Squares optimisation:

$$\begin{cases} \forall i \in [\![1; N]\!] \quad Y_i = X\beta_i + \varepsilon_i, \\ \mathcal{L}(\varepsilon_i) = \mathcal{N}\left(0, \sigma^2 I_M\right). \end{cases}$$

Normality, Centrality, Homoscedasticity, Independence

u.c. for each β_i $\begin{cases} \sum_{k=1}^{K} \beta_{ik} = 1, \\ 0 \leq \beta_{ik} \leq 1. \end{cases}$

Leads to a variety of algorithmic solutions.

Natural way to deal with dependant data is by using a multivariate normal law. However, inferring a conditional correlation matrix with ~ 20000 or ~ 800000 features (gene/probe) is time consuming and quite inoperable for simulation with this approach.

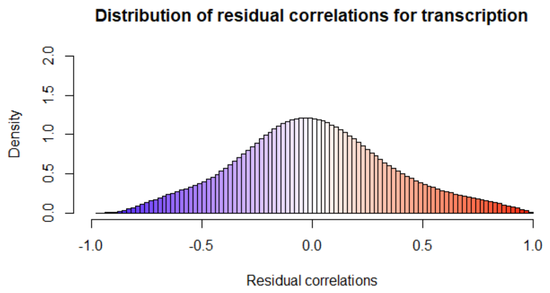
Benchmark dataset

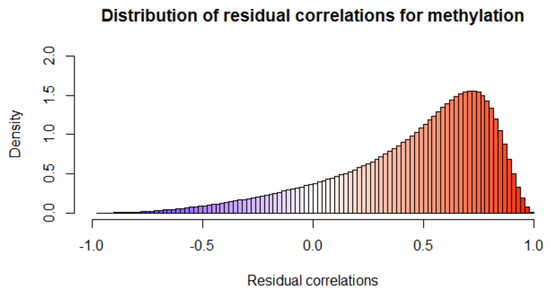
A benchmark **dataset generated *in vitro*** is accessible (from COMETH project [2]) with:

- 21104 gene expressions,
- ~ 800000 CpG probes methylation,
- $N = 30$ **independent** bulk,
- $K = 9$ cell types commonly found in **PDAC**.

The true proportions of each cell type in each bulk are controlled and therefore can be assumed to be known.

Moreover, both omics have significative conditional two by two correlations between features:



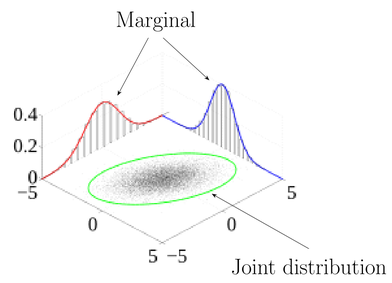


We propose **two different simulation methodologies** for dependent data following specific marginals distribution functions:

Copulas

Thanks to Sklar theorem [3], Copulas:

- **defines how the joint behavior** of multiple random variables **is structured**, regardless of their individual distributions,
- allow us to **characterise various complex forms of dependence**, such as non-linear or tail dependence between multiple variables.



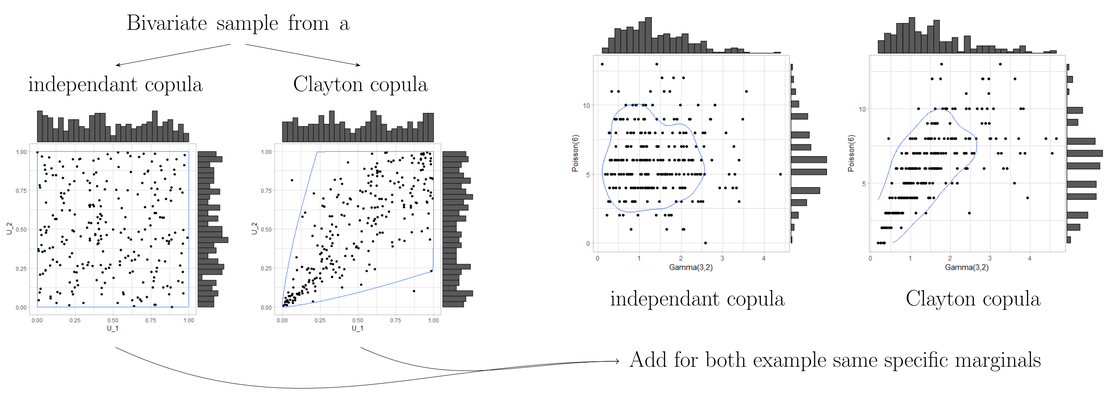
Generation procedure

→ Copulas generate dependant **uniform** random vectors $(U_1, \ldots, U_M)$

This allows us to see those uniform values as **quantile**. Then we only have to **define or infer the marginal law of each feature to simulate any data types**, thanks to the inversion of the Cumulative Distribution Function:

$$\left(X_1, \ldots, X_M\right) = \left(F_1^{-1}(U_1), \ldots, F_M^{-1}(U_M)\right),$$

where here X_j represents values for the **specific** feature j (gene/probe) and F_j is the marginal distribution function defined or inferred for the **specific** feature j (gene/probe) with its **specific** parameter which can change for each feature.



Factor model

Based on a **low-rank factor approximation** [4] of R the square conditional correlation matrix between features:

$$R = \Psi + \underbrace{BB'}_{\text{Shared variance}}, \qquad \Psi \text{ Specific variance}$$

- $\Psi \in \mathcal{M}_{M,M}(\mathbb{R})$ diagonal,
- $B \in \mathcal{M}_{M,q}(\mathbb{R})$,

where $1 \leqslant q < N$ is the number of factors chosen. Matrix B can be seen as a matrix of loadings for each feature on each factor.

Generation procedure

$$\varepsilon_{generated} = \mathcal{N}(0, \Psi) + B \times \mathcal{N}\left(0, I_q\right), \ \varepsilon_{generated} \in \mathcal{M}_{M,N}(\mathbb{R}).$$

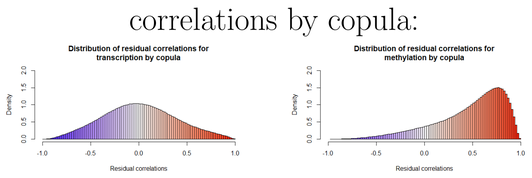
Since Ψ and B result from a decomposition of a correlation matrix, each line of $\varepsilon_{generated}$ is centered and scaled residuals.

$$\varepsilon_{generated} \leftarrow \left(\varepsilon_{generated} \times \hat{\sigma}_{feature}\right) + \hat{\mu}_{feature}$$

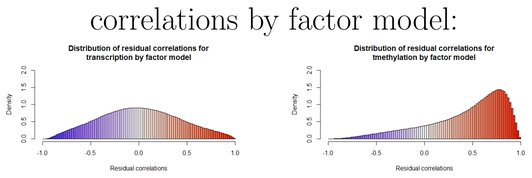
The procedure mimics the behaviour of the data provided.

Results

Distributions of simulated conditionnal correlations by copula:



Distributions of simulated conditionnal correlations by factor model:



Score of NNLS, a basic deconvolution algorithm with late omics integration, on:

		rmse	mae
Benchmark dataset	nnlsmultimodal	0.0856219	0.0544927
No dependant simulated dataset	nnlsmultimodal	0.0209714	0.0148217
Copula simulated dataset	nnlsmultimodal	0.0796089	0.0509681
Factor model simulated dataset	nnlsmultimodal	0.0801569	0.0519525

Perspectives

Both methodologies:

→ are computationally **fast**,

→ **reproduce** and make explicit hypothesis on **different levels of complexity** (dependencies, intrinsic nature of data, ...),

→ need at least one *in vivo* or *in vitro* dataset with **known and controlled ground truth**, however more datasets are needed to avoid overfitted simulation procedure.

Ongoing works:

Here, Copulas and factor model methodology capture dependance structure empirically. We focus now on defining controlled parameters for each approach to simulate different scenarios.

[1] Clémentine Decamps, Alexis Arnaud, Florent Petitprez, et al. DECONbench: a benchmarking platform dedicated to deconvolution methods for tumor heterogeneity quantification. *BMC Bioinformatics*, 22(1):473, October 2021.

[2] Yuna Blum, Jérôme Cros, Sergio Escalera et al. COMETH – COmputational METHods in Health.

[3] M Sklar. Fonctions de répartition à n dimensions et leurs marges. *Annales de l'ISUP*, 8(3):229-231, 1959.

[4] Chloé Friguet, Maela Kloareg, and David Causeur. A factor model approach to multiple testing under dependence. *Journal of the American Statistical Association*, 104(488):1406-1415, 2009.

L'INSTITUT agro Rennes Angers

65 rue de Saint Brieuc,
35042 Rennes,
hugo.barbot@institut-agro.fr