Medical Genomics Practical #2: Multi-omic analyses in R

International Agency for Research on Cancer
Lyon, France

Nicolas Alcala, PhD
Scientist, Genetics section

4th of November 2020

International Agency for Research on Cancer

World Health Organization

alcalan@iarc.fr

Plan

Practical (3.5 hrs)

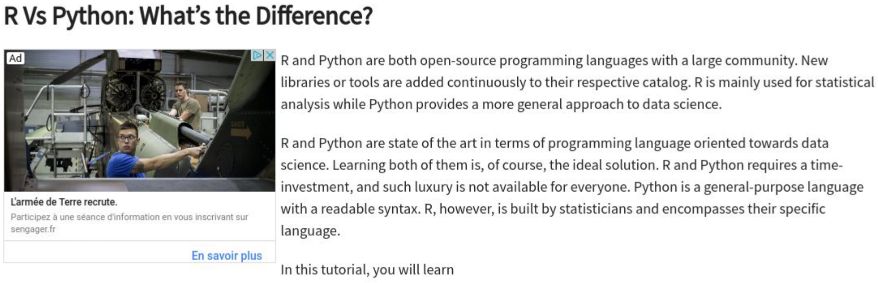
- **Concepts**: MOFA2 implementation
- **Practical**

Projects (0.5 hrs)

- **Choice**
- **Quickstart**

Practical | *MOFA implementation*

Why choose between python and R?



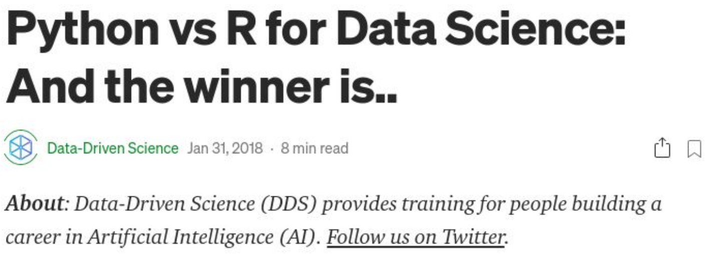
R Vs Python: What's the Difference?

R and Python are both open-source programming languages with a large community. New libraries or tools are added continuously to their respective catalog. R is mainly used for statistical analysis while Python provides a more general approach to data science.

R and Python are state of the art in terms of programming language oriented towards data science. Learning both of them is, of course, the ideal solution. R and Python requires a time-investment, and such luxury is not available for everyone. Python is a general-purpose language with a readable syntax. R, however, is built by statisticians and encompasses their specific language.

In this tutorial, you will learn

- R
- Python
- Popularity index
- Job Opportunity
- Analysis done by R and Python
- Percentage of people switching
- Difference between R and Python
- R or Python Usage



Python vs R for Data Science: And the winner is..

Data-Driven Science Jan 31, 2018 · 8 min read

***About**: Data-Driven Science (DDS) provides training for people building a career in Artificial Intelligence (AI). Follow us on Twitter.*

R vs Python for Data Analysis — An Objective Comparison

Practical | *MOFA implementation*

Why choose between python and R?

- The **core of the method is in python** (mofapy, mofapy2) and uses the powerful python machine learning packages (e.g., scikit-learn)
- The **downstream analyses** and **graphical functions are in R** and leverage the contributions of the enormous R community of computational biologists and statisticians

Issues

- Need to correctly **interface python and R**
- **Many dependencies** in both languages, making the code **difficult to set up and fragile**

Practical | *MOFA implementation*

Why choose between python and R?

- The **core of the method is in python** (mofapy, mofapy2) and uses the powerful python machine learning packages (e.g., scikit-learn)
- The **downstream analyses** and **graphical functions are in R** and leverage the contributions of the enormous R community of computational biologists and statisticians

Solutions

1. (earlier MOFA and MOFA+ releases) **R package reticulate:** allows to specify a python install or conda env, run python functions, transfer R and Pandas data frames, or R matrices and NumPy arrays
2. (newer MOFA+ releases) **R bioconductor package basilisk**: allows R to directly create and handle conda environments with specific python dependencies, allowing smooth usage of incompatible python installs within a same R session

Projects | *MOFA implementation*

Different flavors of computational biology for medical genomics

- I have new data that I want to process through existing workflows => **Project 1**
- I have additional processed data that I want to integrate in my analyses => **Projects 2 and 5**
- I have scripts for a software that I want to implement in a reproducible workflow => **Project 3**
- I have heard of new analyses techniques that I want to try on my data => **Project 4**

Reminder from Medical Genomics #2: Transcriptomics, multi-omics and beyond

International Agency for Research on Cancer
Lyon, France

International Agency for Research on Cancer

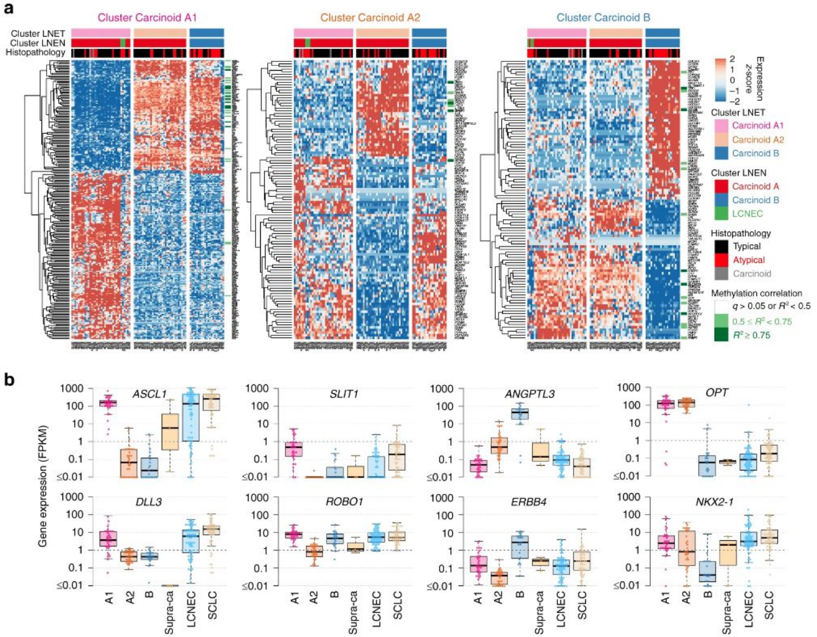
World Health Organization

Part I. Transcriptomics | *Analysis*

Supervised analyses (i): differential expression analysis

Goal: explain biological differences between different conditions

- Fitting model, correcting for confounding variables like batch, or accounting for clinical variables such as sex, age, environmental exposure (e.g., edgeR, DESeq2, limma)
- Analyzing list of genes obtained to understand differences (e.g., gene-set enrichment) or identify therapeutic targets
- *Example: differential expression between molecular subtypes of lung cancer*



Differential expression analysis of lung neuroendocrine tumors. a. Heatmaps of DE genes. **b**. DE genes with clinical relevance. *Source: Alcala, Leblay, Gabriel, et al. Nature Communications 2019.*

International Agency for Research on Cancer

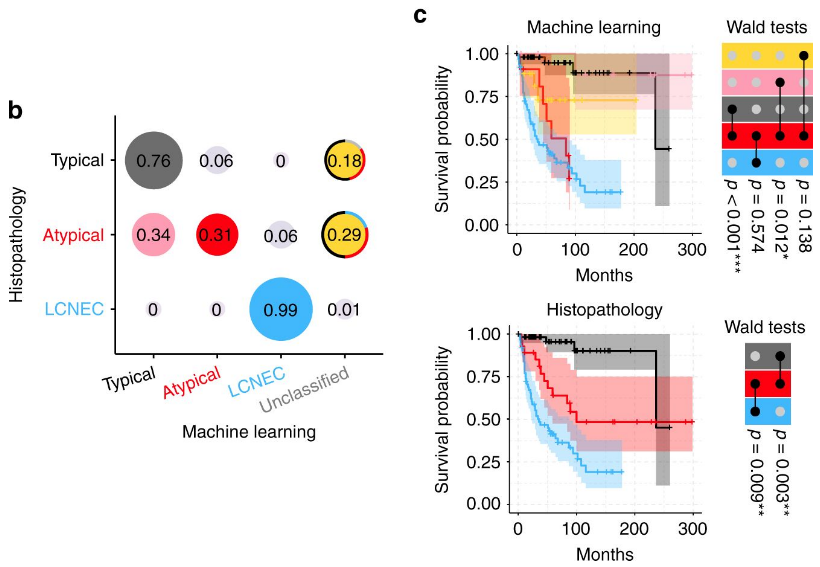
World Health Organization

Part I. Transcriptomics | *Analysis*

Supervised analyses (ii): machine learning

Goal: predict biological or clinical features using molecular data

- Normalization of expression (e.g., Variance Stabilization)
- Training model (e.g., random forest, support vector machine, neural network)
- Testing model
- *Example: predict tumor histopathological types based on molecular data.*



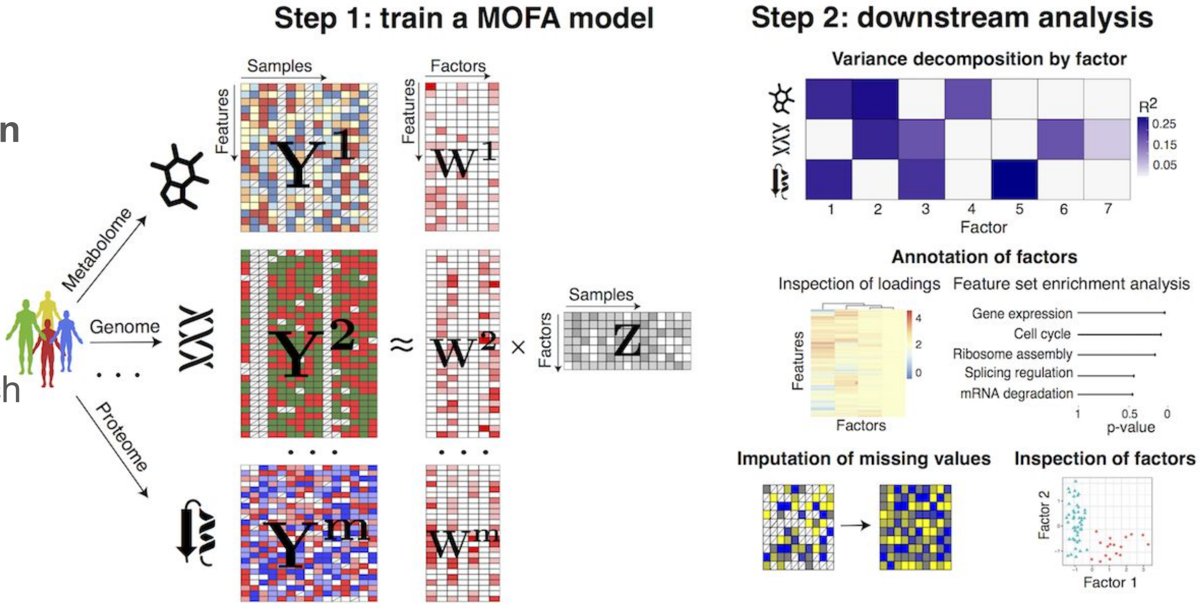
A random forest classifier stratifies atypical carcinoids into good- and bad-prognosis. b. Confusion matrix of the classifier**. c**. Kaplan-Meier survival curves. Model trained on 186 transcriptomes. *Source: Alcala, Leblay, Gabriel, et al. Nature Communications 2019.*

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Identify **latent factors (unknown continuous variables)** representing biological variation **shared between modalities (e.g., genome, transcriptome)**
- Identify in which ‘omic’ layer each factor is active
- Downstream analysis to **understand what each factor represents**

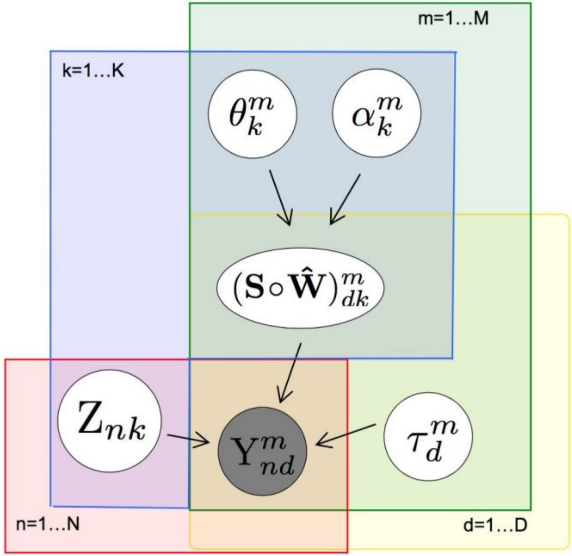


Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- **Generalization of Principal Component Analysis to multiple modalities** *M*
- model $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$,
- where $\mathbf{Y}^m$ is the matrix of observations for each sample *n* (rows) and each feature *d* (columns) for modality *m* (e.g., genomic alterations, expression)
- $\mathbf{Z}$ is the latent factors matrix (*N* by *K*) shared by all modalities *m*
- $\mathbf{W}^m$ is the weights (loadings) matrix (*K* by *M*) of *m*
- $\boldsymbol{\varepsilon}^m$ is the residual noise (column vector of size *N*)



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

Part II. Multi-omics | *Analysis*

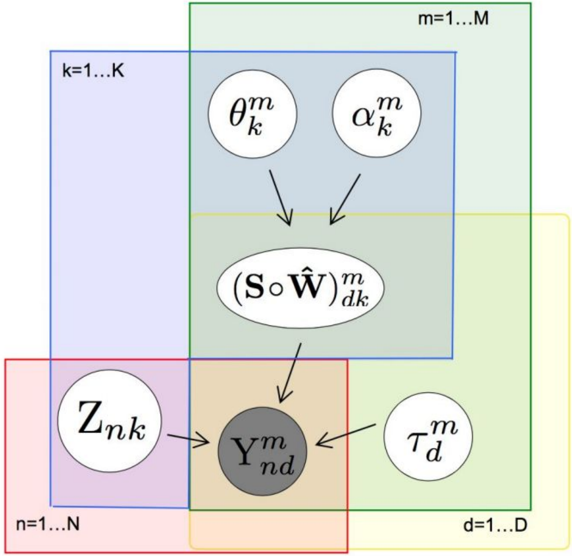
Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Model $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$

Bayesian inference of elements of **Z** and $\mathbf{W}^m$

- *Sparse* (Automatic Relevance Determination X "spike-and-slab") priors on weights $w^m_{d,k}=s^m_{d,k}\hat{w}^m_{d,k}$, with $s^m_{d,k}$ following a Bernoulli prior and $\hat{w}^m_{d,k}$ following a Normal prior with precision α^m_k, **so if the density of *s* is close to 0 the factor is not active in modality *m* (e.g., the factor does not explain any variation in expression data)**



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

International Agency for Research on Cancer

World Health Organization

Part II. Multi-omics | *Analysis*

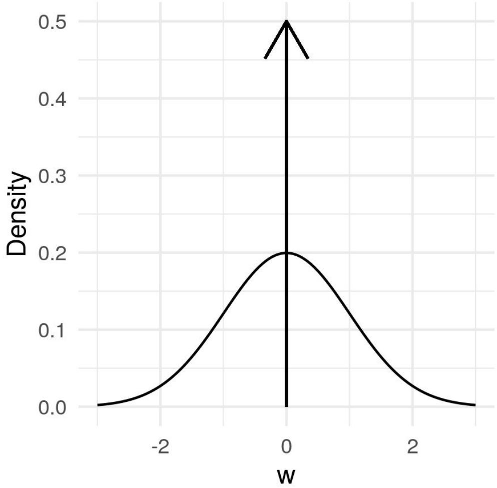
Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Model $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$

Bayesian inference of elements of **Z** and $\mathbf{W}^m$

- *Sparse* (Automatic Relevance Determination X "spike-and-slab") priors on weights $w^m_{d,k}=s^m_{d,k}\hat{w}^m_{d,k}$, with priors $s^m_{d,k}$ ~ Bernoulli(Θ^m_k) and $\hat{w}^m_{d,k}$~ Normal(0,1/α^m_k), **so in modality *m*, if Θ^m_k is close to 0, factor *k* is sparse (most features have 0 weights), and if α^m_k is large factor k not active (e.g., the factor does not explain any variation in expression data)**



Spike and slab prior. The arrow represents a Dirac point mass at 0.

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

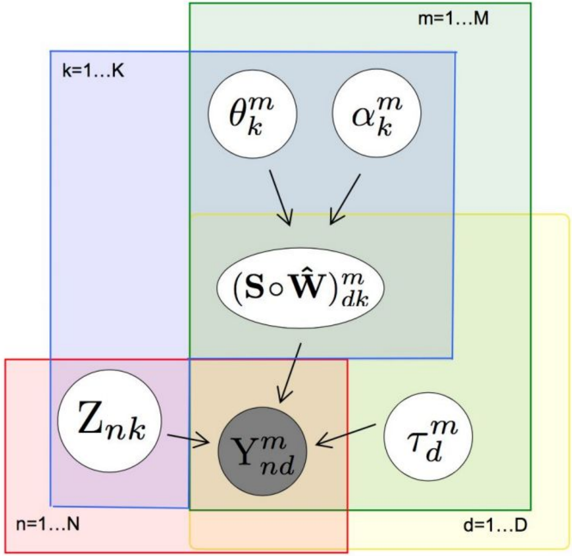
Multi-Omics Factor Analysis (MOFA)

- Model **Y^m = ZW^m + ε^m**

Bayesian inference of elements of **Z** and **W^m**

- *Gaussian* (for continuous data, e.g. normalized expression data and methylation *M* values), *Bernoulli* (for binary data, e.g. genomic alterations), or *Poisson* (for count data, e.g. as expression in read counts)

  prior distributions on noise ε^m_n



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

International Agency for Research on Cancer

World Health Organization

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

Variational Bayes (or VI) implementation:

- *Rationale*: when fitting complex Bayesian models, the posterior distribution of the parameters is often intractable; we **need an approximation**
- *Method (VI)*: a **lower bound on the model likelihood (the Evidence Lower Bound--ELBO) is optimized** (E-M algorithm), using a simpler factorized form for the posterior
- *Note*: less computer-intensive alternative to the popular Monte Carlo Markov Chains (MCMC)

Part II. Multi-omics | *Analysis*

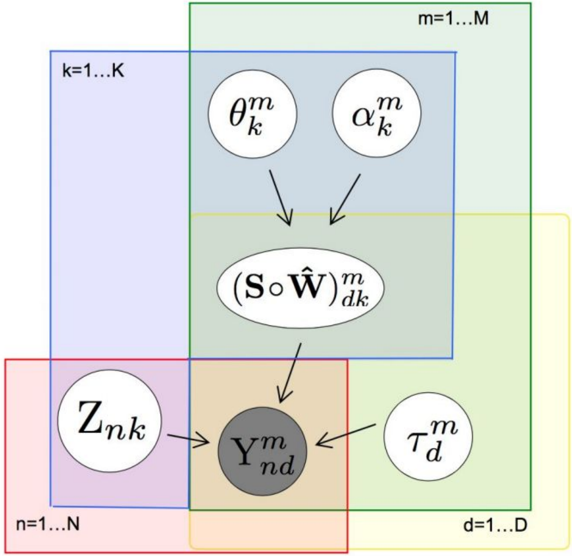
Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Model $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$

Important points:

- Because **Z** is estimated from all ‘omic’ layers *m* and features *d*, the **model handles missing data naturally**
- The sparsity assumptions perform **automatic feature and factor selection**
- **Technical artifacts**, usually restricted to a single modality *k*, are separated from variation with **evidence from multiple modalities**
- **Correlations between modalities** are found (e.g., expression QTLs)



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

Part II. Multi-omics | *Analysis*

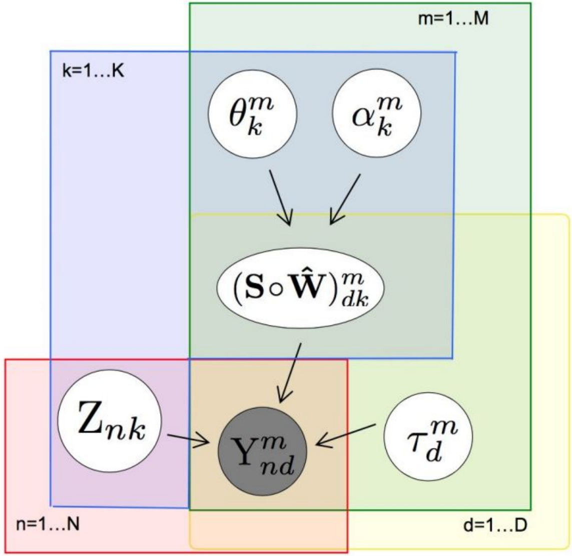
Tools for integration: unsupervised analyses

Multi-Omics Factor Analysis (MOFA)

- Model $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$

Important points:

- the likelihood formulation implicitly gives more weight to modalities with many features, so **beware of imbalance between input data matrices** (e.g., a mutation matrix of 20 features will not influence much **Z** if an expression matrix with 10,000 features is also provided)



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*