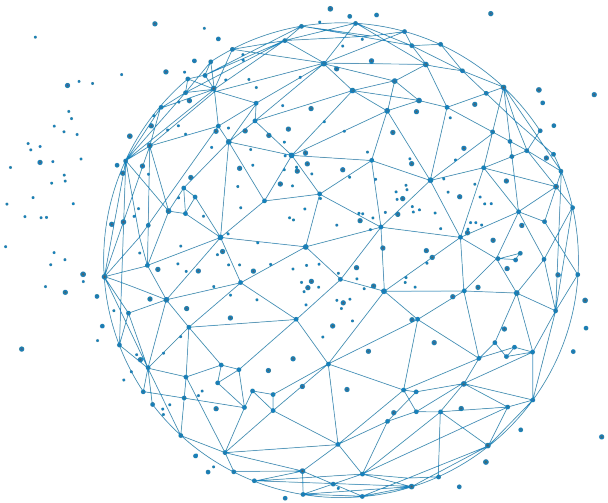
Second edition 2024 in Fréjus

ÉCOLE THÉMATIQUE ETBII BIOINFORMATIQUE INTÉGRATIVE IFB



MOFA: Multi-omic Factor Analysis

Jimmy Vandel (PLBS - Bilille)
Arnaud Gloaguen (CNRGH-CEA)
Vincent Guillemot (Institut Pasteur)

DOI version final

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir FRANCE
cnrs INRAE Inserm cea INVESTIR L'AVENIR

Relative publications

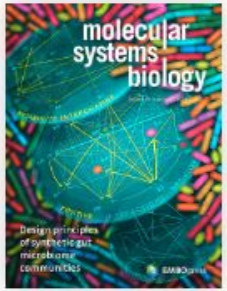
Method | 20 June 2018 | OPEN ACCESS TRANSPARENT PROCESS

Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, Oliver Stegle

Author Information

Molecular Systems Biology (2018) 14: e8124 | https://doi.org/10.15252/msb.20178124



About the cover

Volume 14
Issue 6
1 June 2018

IN THIS ISSUE

Method | Open access | Published: 11 May 2020

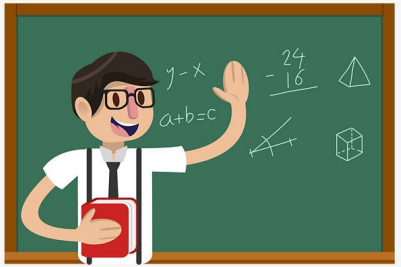
MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data

Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni & Oliver Stegle

Genome Biology **21**, Article number: 111 (2020) | Cite this article

44k Accesses | **237** Citations | **118** Altmetric | Metrics

Matrix multiplication, back to school



$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 10 & 11 \\ 20 & 21 \\ 30 & 31 \end{bmatrix}$$

$$= \begin{bmatrix} 1x10 + 2x20 + 3x30 & 1x11 + 2x21 + 3x31 \\ 4x10 + 5x20 + 6x30 & 4x11 + 5x21 + 6x31 \end{bmatrix}$$

$$= \begin{bmatrix} 10+40+90 & 11+42+93 \\ 40+100+180 & 44+105+186 \end{bmatrix} = \begin{bmatrix} 140 & 146 \\ 320 & 335 \end{bmatrix}$$

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} + a_{14}b_{41}$$

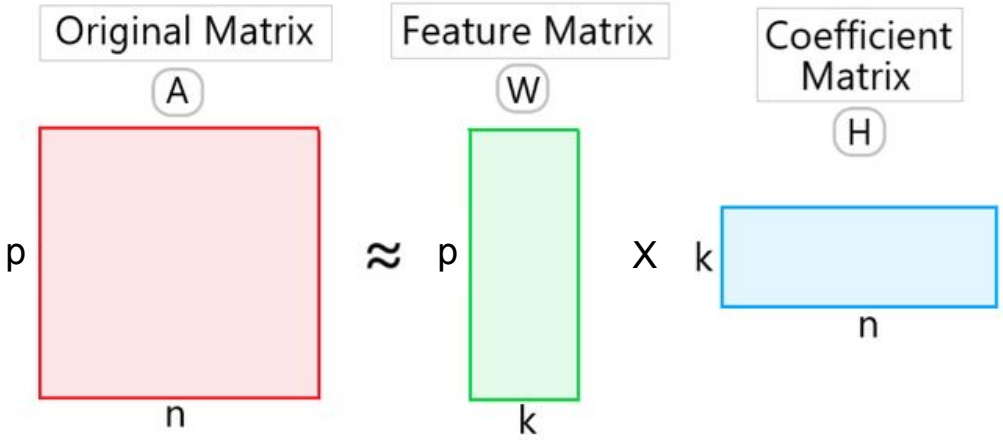
$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

Matrix dimensions : 2 x **4** **4** x 3 2 x 3

$$c_{22} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} + a_{24}b_{42}$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \\ b_{41} & b_{42} & b_{43} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

Matrix factorization



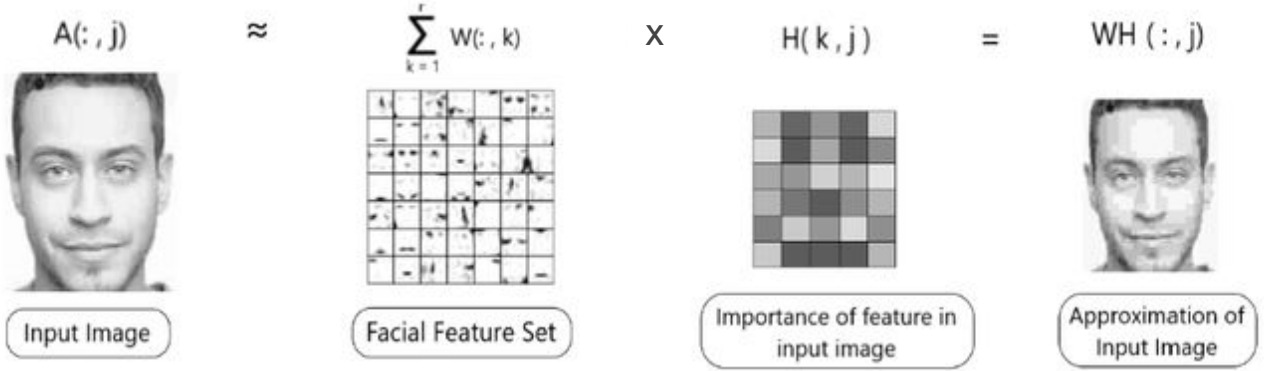
→ approximate the large data matrix A using the product of 2 smaller matrices W and H

$$A = W \times H + \varepsilon$$

Matrix factorization applications

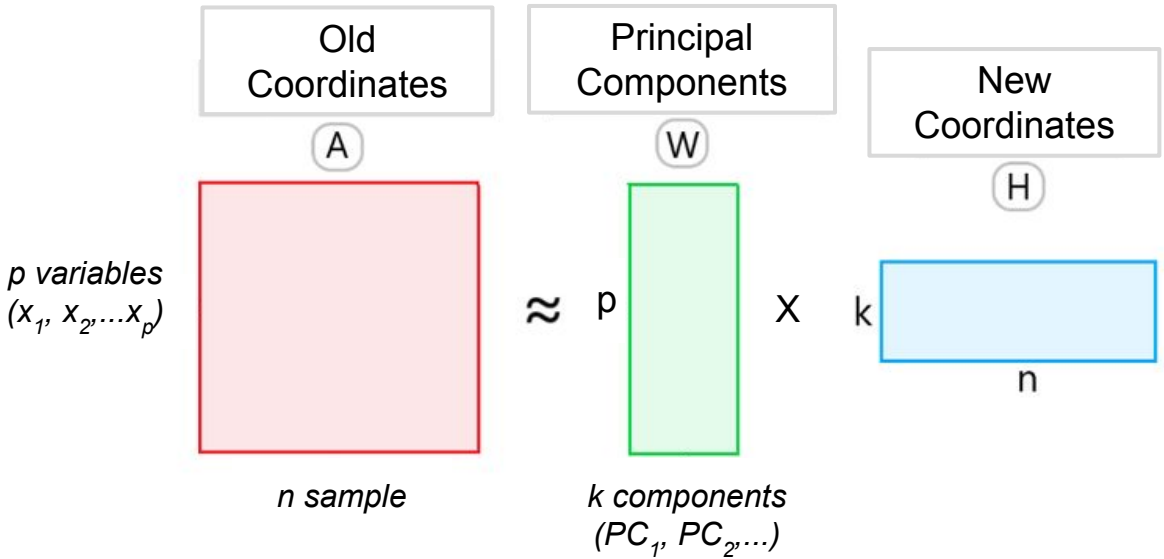
Useful for dimensionality reduction (k features in the W matrix) and feature extraction (the H matrix)

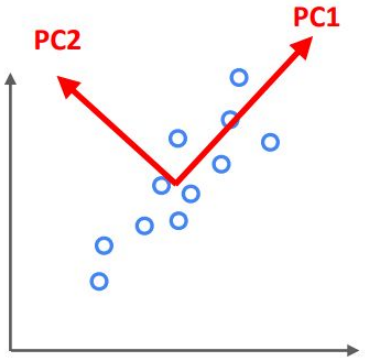
Example: image processing with Non-Negative Matrix Factorization (W ≥ 0 and H ≥ 0)



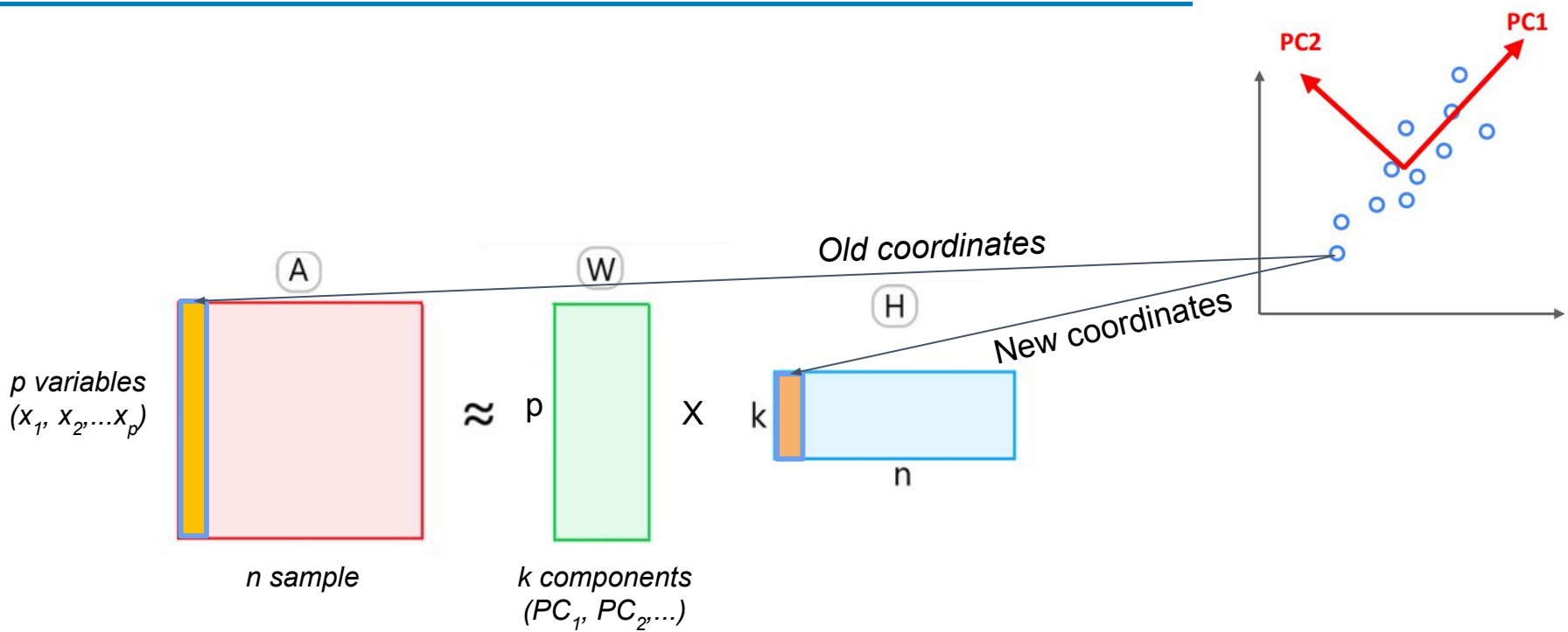
→ PCA can be formulated as an approximation of matrix factorization

Matrix factorization, PCA

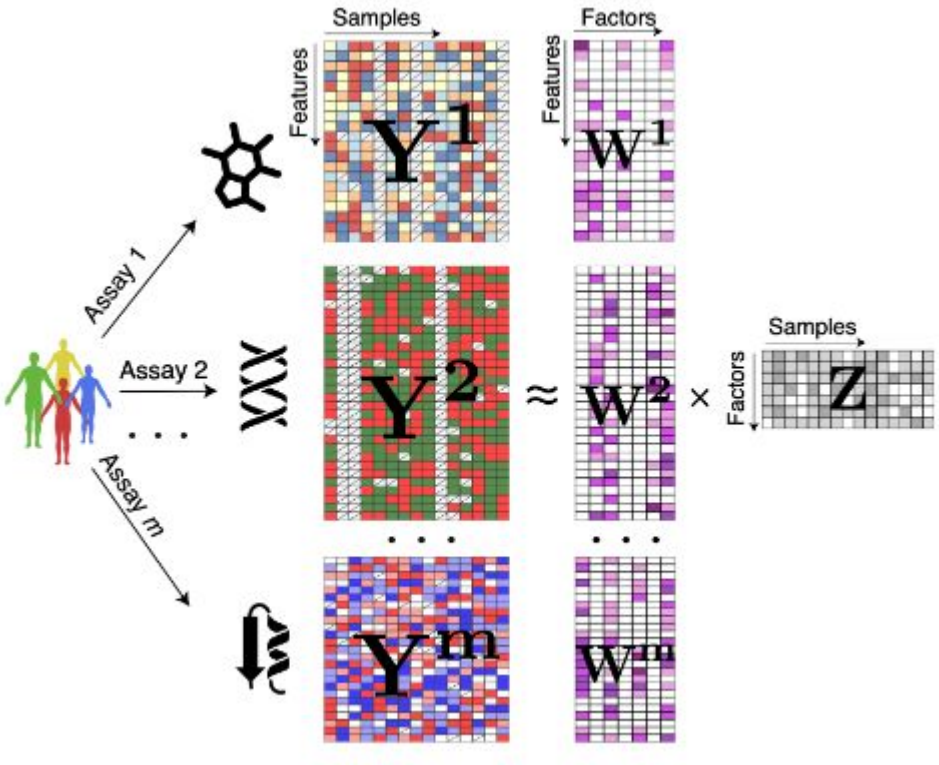




Matrix factorization, PCA



MOFA: PCA generalization



$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m$$

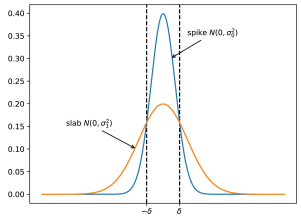
- *m* views / omic sources
- share the Z matrix between views
- 2 levels of sparsity on W^m :
  - view and factor-wise
    → active/inactive factors in *a view*
  - feature-wise
    → sparse biological phenomenon
- Y^m and ε^m can follow different models :
  - Gaussian (continuous)
  - Poisson (natural/count)
  - Bernoulli (binary)

MOFA: sparsity priors

- Automatic Relevance Determination prior
  - view and factor-wise prior following Bernoulli distribution

  -> all factors are not active in each view (view-specific biological phenomenon)

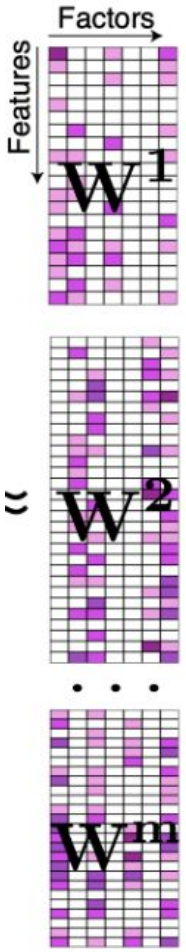
- Spike-and-slab sparsity prior



  - feature-wise prior in each view *m* for a given factor *k following 'normal' distribution*

  -> small number of features with active weights (sparse biological phenomenon)

$$W = S\widehat{W}, \text{where } s_{dk}^m \sim \text{Ber}\ (\theta_k^m) \text{ and } \widehat{W}_{dk}^m \sim N(0, 1/\alpha_k^m).$$



MOFA: how it works?

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m$$

- Parameter estimation through variational Bayesian inference

$$p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \prod_{m=1}^{M}\prod_{n=1}^{N}\prod_{d=1}^{D_m}\mathcal{N}\left(y_{nd}^m \,|\, \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right)$$
$$\prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K}\mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right)\mathrm{Ber}(s_{d,k}^m|\theta_k^m)$$
$$\prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}\left(z_{nk} \,|\, 0, 1\right)$$
$$\prod_{m=1}^{M}\prod_{k=1}^{K}\mathrm{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right)$$
$$\prod_{m=1}^{M}\prod_{k=1}^{K}\mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right)$$
$$\prod_{m=1}^{M}\prod_{d=1}^{D_m}\mathcal{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right).$$

posterior distribution of unobserved data X, P(X/Y), is approximated by $q(\mathbf{X}) = \prod_i q(\mathbf{X}_i)$

$$q(\mathbf{Z}, \mathbf{S}, \hat{\mathbf{W}}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\theta}) = q(\mathbf{Z})q(\boldsymbol{\alpha})q(\boldsymbol{\theta})q(\boldsymbol{\tau})q(\mathbf{S}, \hat{\mathbf{W}})$$

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m$$

- Parameter estimation through variational Bayesian inference
- Evidence Lower Bound (ELBO)
  - the true log marginal likelihood $\log p(\mathbf{Y})$ is lower bound by the ELBO $\mathcal{L}(\mathbf{X})$

$$\begin{aligned}\mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X}) \Big(\log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y}) \Big) d\mathbf{X} \\ &= \log p(\mathbf{Y}) - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) \\ &\leq \log p(\mathbf{Y})\end{aligned}$$

  - the objective is to optimise $\mathcal{L}(\mathbf{X})$ with respect to the distribution q(X)

MOFA: how it works?

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m$$

- Parameter estimation through variational Bayesian inference
- Evidence Lower Bound (ELBO)
- Iterative estimation process similar to the Expectation-Maximization (EM) algorithm
  - each unobserved variable is updated one by one considering the others

$$q(\mathbf{Z}) = \prod_{k=1}^{K}\prod_{n=1}^{N} q(z_{nk}) = \prod_{k=1}^{K}\prod_{n=1}^{N} \mathcal{N}\left(z_{nk} \,|\, \mu_{z_{nk}}, \sigma_{z_{nk}}\right)$$

$$q(\hat{\mathbf{W}}, \mathbf{S}) = \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} q(\hat{w}_{dk}^m, s_{dk}^m) = \prod_{m=1}^{M}\prod_{d=1}^{D_m}\prod_{k=1}^{K} q(\hat{w}_{dk}^m | s_{dk}^m) q(s_{dk}^m)$$

MOFA: how it works?

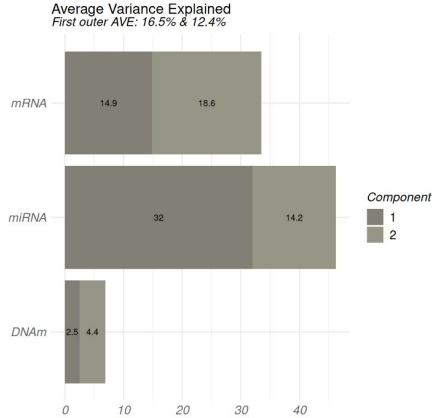
$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m$$

- Parameter estimation through variational Bayesian inference
- Evidence Lower Bound (ELBO)
- Iterative estimation process similar to the Expectation-Maximization (EM) algorithm
- Iteration stop when ELBO change is small enough
- Automatically drop factors with low variance explained…

MOFA results

Variance decomposition by factors for MDD dataset

→ percentage of variance explained by each factor for each data modality



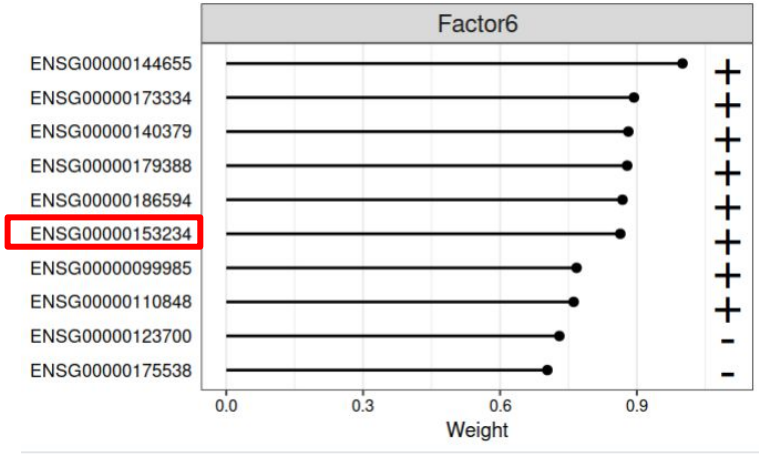
with RGCCA

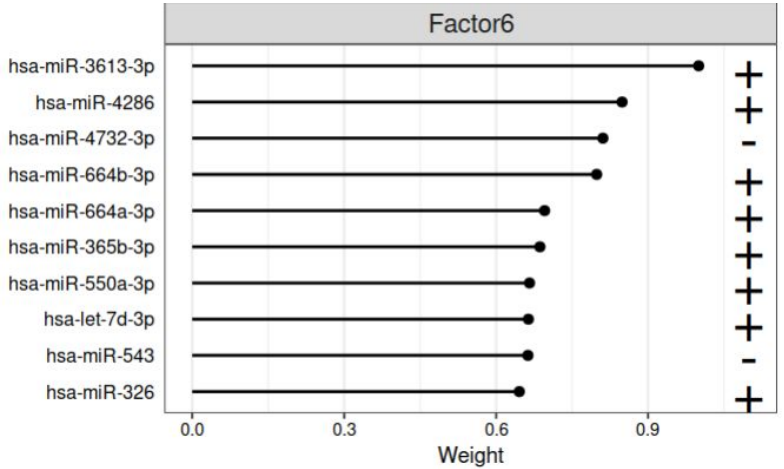
- Factor 1 captures a source of variability that is present mainly in the miRNA view
- Factor 2 captures variation that is present across all data modalities but mainly in mRNA.

MOFA results on W matrices

Feature weights by factor for each view/omic (ie a W^m column)

→ weights provide a signed score (association measure) for each feature for a given factor (below mRNA and miRNA for Factor 6, associated with Sample_Group variable)



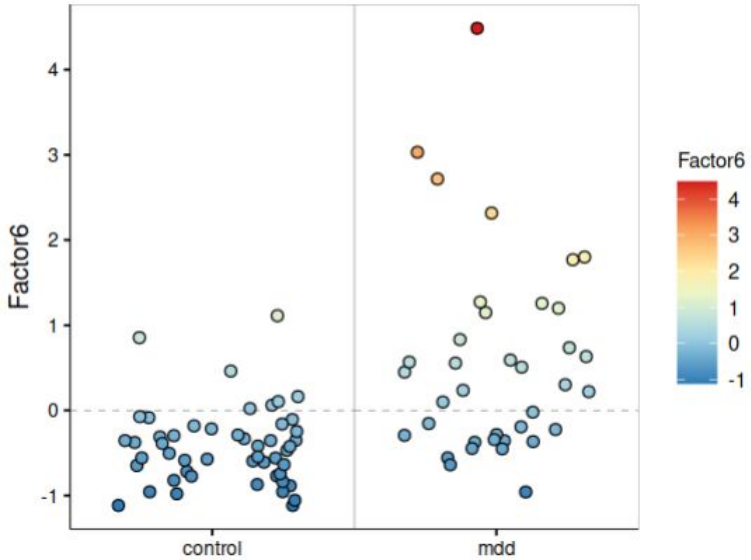


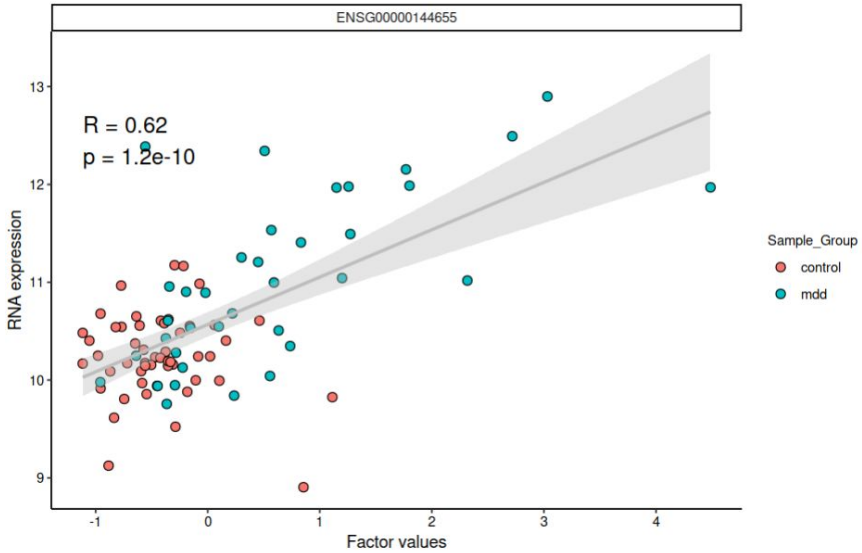
Identified as significant in Component 1 of supervised RGCCA (the most associated to Sample_Group variable)

MOFA results on Z matrices

Factor values regarding known groups of samples (ie a Z row)

→ detect association between a factor and a specific variable/feature





- Separation between control and MDD patients shows association with Factor 6
- Expression of CSRNP1 gene (ENSG00000144655) is also associated with Factor 6 (and MDD status)

MOFA characteristics

- Choice of **k** (number of factors)
  → inactive factors can be eliminated through a user defined threshold for explained variance

- Random initialization : no guarantee of optimal solution
  → run MOFA several times (~10 times) with different initialisations (solved in MOFA+ thanks to PCA)
  → keep the model with the highest ELBO for downstream analysis

- Missing value
  → no need for imputation, missing values are ignored in the model thanks to probabilistic modelling

- Data pre-processing
  → no need as long as indicated distributions are respected (eg. Gaussian) → to check ++

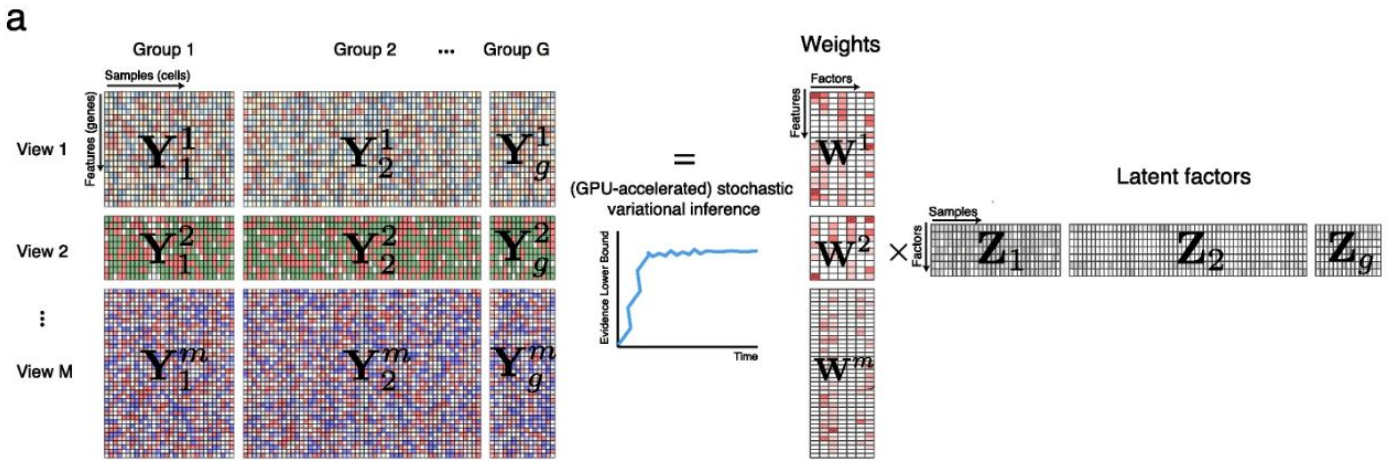
MOFA limits

- Different views but on the same samples
- Mainly linear relationships are captured
- Assumes independence between features in the model (and samples)
- Unbalanced modalities sensibility

MOFA+ extension for groups

$$\mathrm{Y}_{\mathrm{gm}} = \mathrm{Z}_{\mathrm{g}} \mathrm{W}_{\mathrm{m}}^{T} + \epsilon_{\mathrm{gm}}$$

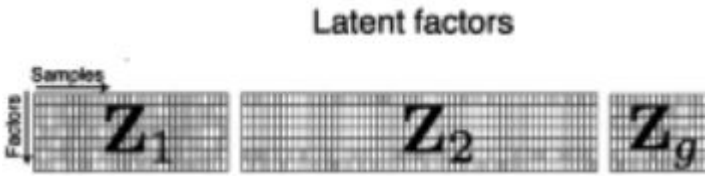
- Define sample groups (batch, experiments, conditions)
- Group-wise prior applied to the Z matrix
- Stochastic variational inference framework (GPU accelerated computations)



- Additional prior on Z matrix with ARD + spike-and-slab (as for W matrix)

  → all factors are not active in each group (group-specific biological phenomenon)

  → small number of samples with active factor (sparse biological phenomenon)



- Stochastic Variational Inference
  - fast approximation of the gradient with a random subset of the data (batch)
  - gradient ascent step size ajusted at each iteration
  - efficient only when the number of samples >>> the number of features
  - additional parameters : batch size and step size (start and forgetting)

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \nabla F\left(\mathbf{x}^{(t)}\right) \qquad \rho^{(t)} = \frac{\tau}{(1+\kappa t)^{3/4}}$$

MOFA+ advises

- use Gaussian likelihood when possible (instead of Poisson or Bernoulli)
- use appropriate normalization to fit with available likelihoods

  e.g. for single cell rna-seq: size factor normalization + VST
- consider feature selection procedure before MOFA+ (*e.g. highly variable features*)
  - to limit view imbalance
  - to speedup the model
- use regression before MOFA+ to remove undesired technical sources of variation
- groups should not be defined as for a differential analysis
  - identify shared and exclusive source of variability in the data
  - larger groups → more complexity → more factors