Comp790-166: Computational Biology

Lecture 22

April 12, 2023

Today

- Finish up multiomics factor analysis with MOFA and MOFA+
- Convex optimization to prime a data integration strategy with ADMM approach

Quick Announcements

- Homework 2 is now available online and due **Friday April 21.** → `https://github.com/natalies-teaching/CompBio2023/tree/main/Homework2_2023`
- Sign up for final presentations in the google doc `https://docs.google.com/document/d/1x9mIJCZAkeogAhmGlpqJkXwuoAKlB_0gewV1LpXZDoU/edit`.

Review Questions

1. What are the two matrices $\mathbf{Y}$ is factorized into with MOFA?
2. Which are the latent variables and which are the observed variables in the MOFA approach?

MOFA Review

Assuming there are k factors and given

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

,

- $\mathbf{Z} \in \mathbb{R}^{N \times K}$ relate the original samples to the factors
- $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ relates the original features to the factors

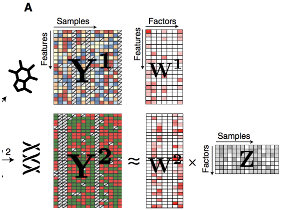


Figure: from Argelaguet *et al.* Molecular Systems Biology. 2018.

Overview of the Model

This is latent variable model and feature dependencies are attempted to be explained in terms of k latent classes (or ‘factors‘).

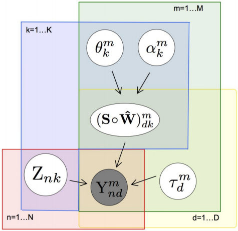


Figure: from Argelaguet *et al.* Molecular Systems Biology. 2018. As always, gray nodes are observed variables and white notes are the unobserved variables inferred by the model.

Results in Dataset 1 : CLL dataset

Find data here

`https://rdrr.io/github/bioFAM/MOFAdata/man/CLL_data.html`

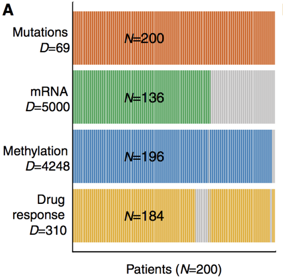


Figure: from Argelaguet *et al.* Molecular Systems Biology. 2018. Modalities and present/missing features for each patient.

Visualization of Samples

Samples can be projected into two dimensions based on the first two factors inferred in the matrix, $\mathbf{Z}$

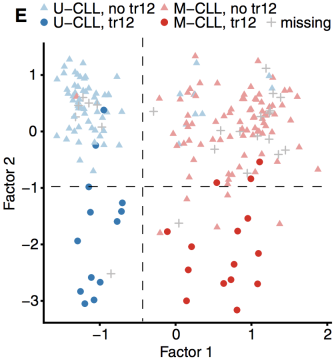


Figure: colors denote something related to the status of the tumor. Shape indicates chromosome 12 trisomy status.

Jointly Predicting Clinical Outcome in Contrast to a Single Type of Data

Experiment :

- Predict time to treatment for $N = 174$ patients using multivariate Cox regression trained using the 10 factors from MOFA
- Compare this performance on the prediction accuracy on individual modalities.
  - In this case, reduce each individual modalities to the top 10 PCs.

Results Predicting Time to Treatment

Note that the Y-axis is simply a statistic reflecting goodness of fit between true and predicted times to treatment (Use top PCs instead of regular feature space).

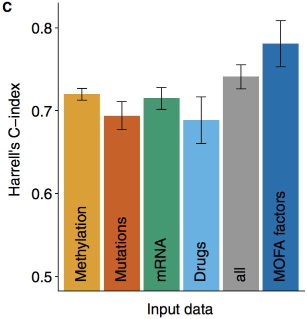


Figure: from Fig. 4 of Argelaguet *et al.* Molecular Systems Biology. 2018.

Instead of using PC Representations for Each Modality

Instead of using PC representations for each modality, the authors also compared to prediction with all features from each individual modality. Again, predicting time to treatment.

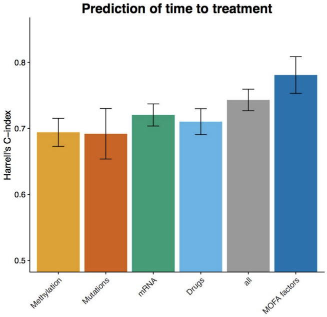


Figure: from Supp. Figure 20. Using all features still does not beat performance based on MOFA factors.

Certain Combinations of Modalities Can Turn out to be More Effective

Applying MOFA with individual modalities held out reveals sometimes certain combinations of modalities is more effective than just throwing everything in all together!

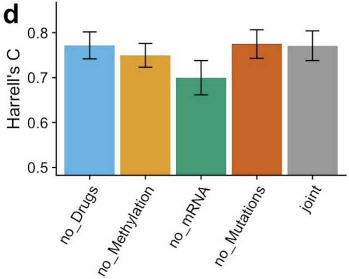


Figure: from Supplementary Figure 8.

Survival Analysis According to Particular Factors

Survival probability vs time-to-treatment for the individual MOFA factors

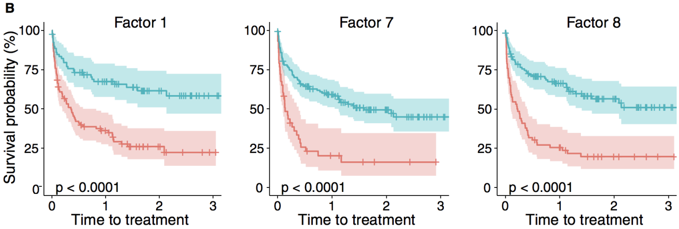


Figure: from Fig. 4. We can see distinct patterns based on survival probability against predicted time to next treatment.

Dataset 2 : Single-Cell Data for a Differentiation Trajectory

Data from a single-cell multiomics dataset. This is applied to 87 mouse embryonic stem calls, with 16 that were cultured in '2i1' media, giving them a naive pluripotent state. The remaining cells were serum grown and hence committed to a differentiation trajectory.

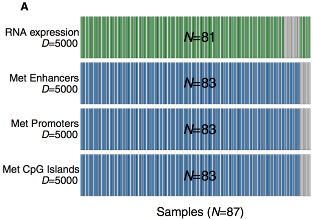


Figure: from Fig. 5. Distribution of samples profiled in each modality.

Visualizing Cells Based on **Z**

Using **Z** to plot cells based on the first two factors, the cells separate according to how they were cultured (which also corresponds to differentiation status!)

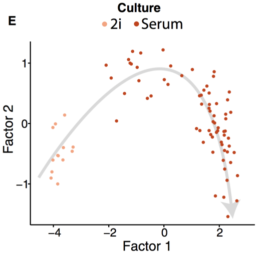


Figure: from Fig. 5.

Inferred Factors on Genes and How They Relate to Pluripotency and Differentiation

Using the $\mathbf{W}^m s$ for the mRNA data, the authors compared the ranks of genes involved in pluripotency and differentiation, respectively to the loadings.

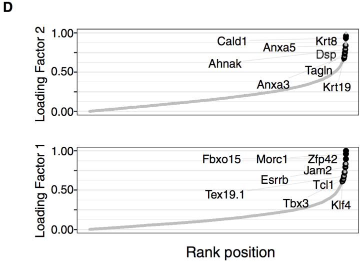


Figure: from Fig. 5. Pluripotency genes (top) vs differentiation related genes (bottom)

An Extension, MOFA+

A trivial extension....almost the same, except now we learn a **Z** per group.

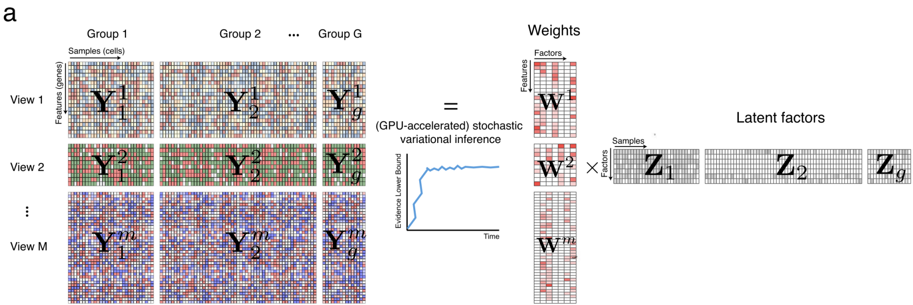


Figure: from Argalaguet *et al.* Genome Biology. 2020. Views are modalities. Groups are some partitioning of the samples or cells.

Group Brainstorming Activity

- Come up with a multiomics experiment where you would have difference groups.
- What would be your views and what would be your groups?
- How would you use the factor information?

Example

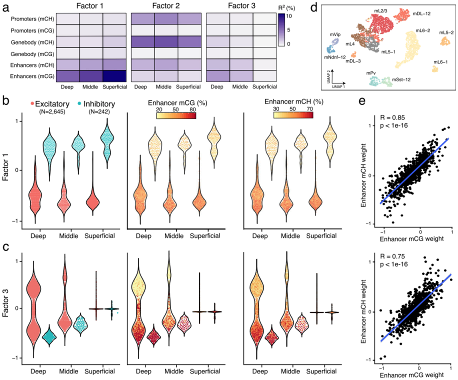


Figure: Context dependent DNA methulation signature associated with cellular diversity in the mammalian cortex. Views are genomic content (DNA methylation signatures), groups are cortical layer.

Defining Y_{gm}

Similar to what we have seen with the regular MOFA,

$$Y_{gm} = Z_g W_m^T + \epsilon_{gm}$$

- Y_{gm} is the matrix of observations for the mth modality and gth group
- W_m is the weight matrix for the mth modality
- Z_g is the factor matrix for the gth group
- ϵ_{gm} is the residual noise for the mth modality in the gth group

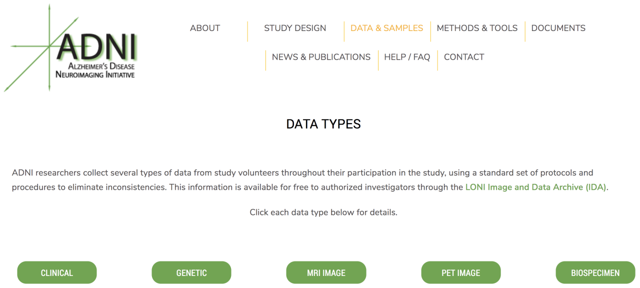
Summary

- MOFA for decomposing a sample x feature matrix to (feature x factor)(factor x sample)
- MOFA$+$ for doing this calculation per group

Adding a time axis into the integration problem

- Everything becomes more interesting when you add a time axis
- One interesting example is examining the interplay between some continuous measure and diagnosis over time
- **Neuroscience Application :** Integrating multiple modalities (imaging, genomic, clinical data)
  - Continuous outcomes: score on cognitive tests
  - Classification or Patient Diagnosis : Alzheimer's Disease or not.
  - The performance on cognitive tasks tends to also relate to the diagnosis.

ADNI Data (Multimodal Brain Imagining + Biomarkers + Genetic + Clinical Data)



DATA TYPES

ADNI researchers collect several types of data from study volunteers throughout their participation in the study, using a standard set of protocols and procedures to eliminate inconsistencies. This information is available for free to authorized investigators through the LONI Image and Data Archive (IDA).

Click each data type below for details.

Figure: Access ADNI data
http://adni.loni.usc.edu/data-samples/data-types/

A Joint Model of Cognitive Scores and Diagnosis

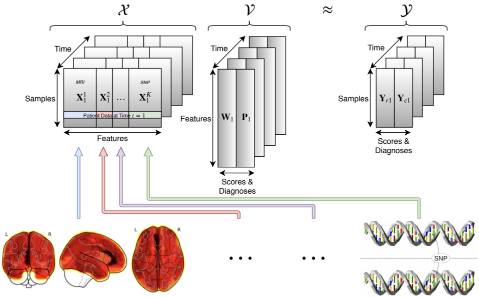


Figure: from Brand, Wang, *et al.* PacSym Biocomputing. 2020.

A Joint Model of Cognitive Scores and Diagnosis, Continued

- The assumption that there is a relationship between the classification and regression tasks
- A patient that performs poorly on cognitive tests is more likely to be diagnosed with AD.
- The overall goal is to find biomarkers (e.g. biological features) that are predictive or explain these patterns

Notation and Problem Formulation

- **Input Features:** $\mathcal{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_T\} \in \mathbb{R}^{n \times d \times T}$ This represents patients × features × timepoints.
- Note that each $\mathbf{X}_t$ can be broken down across the K modalities as, $\{\mathbf{X}_t\}_{j=1}^{K}$
- The output diagnoses and cognitive scores are represented by another tensor, $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T\} \in \mathbb{R}^{n \times c \times T}$
- Each $\mathbf{Y}_t = [\mathbf{Y}_{rt}, \mathbf{Y}_{ct}]$ represents concatenated response variables for regression (r) and classification (c).
- The goal is to learn a tensor, $\mathcal{V} = \{[\mathbf{W}_1, \mathbf{P}_1], [\mathbf{W}_2, \mathbf{P}_2], [\mathbf{W}_T, \mathbf{P}_T]\}$ which represents the coefficients for each feature for regression (**W**s) and classification (**P**s) across the T timepoints.

Take a Look at the Fig. 1 Again...

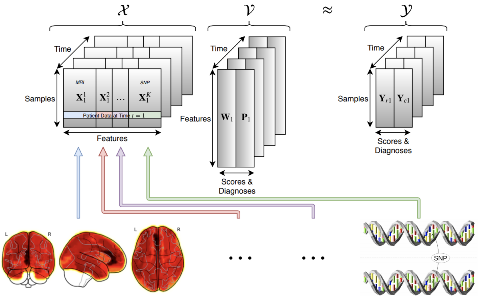


Figure: from Brand, Wang, emphet al. PacSym Biocomputing. 2020.

A Regularized Joint Learning Model

$$\min_{\mathcal{W},\mathcal{P}} \mathcal{L}_r(\mathcal{W}) + \mathcal{L}_c(\mathcal{P}) + \mathcal{R}(\mathcal{V})$$

- Here, $\mathcal{L}_r(\mathcal{W})$ and $\mathcal{L}_c(\mathcal{P})$ are the loss functions for the regression and classification tasks, respectively.
- Regression and classification coefficient matrices are $\mathbf{W}_t \in \mathbb{R}^{d\times c_r}$ and $\mathbf{P}_t \in \mathbb{R}^{d\times c_c}$
- $\mathcal{R}(\mathcal{V})$ is a regularization function applied to the matrix unfolded from the tensor, $\mathcal{V} \to \mathbf{V}^{d\times cT}$. Here, $\mathbf{V}^{d\times cT}$ is constructed by taking the $(\mathbf{W}_t, \mathbf{P}_t)$ matrix pairs and joining by the columns.

Regularization, $\mathcal{R}(\mathcal{V})$

To associate image and genomic features to cognitive scores and diagnoses over time, apply an $\ell_{2,1}$ norm to unfolded coefficient matrix as,

$$\mathbf{V} : ||\mathbf{V}||_{2,1} = \sum_{d=1}^{d} ||\mathbf{v}^{i}||_2$$

Next, to capture the impact of each modality (e.g. MRI, SNP, etc), the authors use the group ℓ_1-norm (G_1 norm) on the rows of **V** corresponding to modality j as,

$$||\mathbf{V}||_{G_1} = \sum_{j=1}^{K} ||\mathbf{V}^{j}||_2$$

Regularization, $\mathcal{R}(\mathcal{V}$, Continued

Finally, to account for inter-modal relationships (or relatedness of features within a modality to cognitive outcome), they use trace norm regularization of **V** as,

$$\mathbf{V} : ||\mathbf{V}||_* = \sum \sigma_i(\mathbf{V})$$

. Here, $\sigma_i(\mathbf{V})$ are the singular values of **V**

Objective

Incorporating the three regularizations, the objective can be specified as follows,

$$\min_{\mathbf{V}} J = \sum_{t=1}^{T} \left[\|\mathbf{X}_t \mathbf{W}_t - \mathbf{Y}_{rt}\|_F^2 \right] + \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{k=1}^{c_c} \left[\left(1 - \left(\mathbf{x}^{it} \mathbf{p}_{kt} + b_{kt}\right) y_{ikt}\right)_+ \right]$$
$$+ \gamma_1 \|\mathbf{V}\|_{2,1} + \gamma_2 \|\mathbf{V}\|_{G_1} + \gamma_3 \|\mathbf{V}\|_* \quad ,$$

- The second term is the loss of $c_c \times T$ one-vs-all multi-class SVM
- $y_{ikt} \in \{-1, 1\}$ is the class label associated with the i-th patient at time t
- b_{kt} is the bias associated with the $(k \times t)$-th SVM
- $(\cdot)_+$ is defined as $(a)_+ = \max(0, a)$

A Coming Attraction for the Optimization: ADMM

- ADMM stands for alternating direction method of multipliers.
- The basic idea is to break up a big problem into sub-problems.
- The complicated objective here will be solved using multi-block ADMM, an extension of regular ADMM as,

$$\begin{aligned} \min_{x_i} \quad & f_1(x_1) + f_2(x_2) + \cdots + f_K(x_K) \\ \text{subject to} \quad & \mathbf{E}_1 x_1 + \mathbf{E}_2 x_2 + \cdots + \mathbf{E}_K x_K = c \end{aligned}$$