Comp790-166: Computational Biology

Lecture 21

March 30, 2022

Good Morning Question

1. What was the intuition about the type of features SLICER wanted to include for the trajectory inference problem?
2. **Multiomics Datasets:** What are some examples of problems that you have worked on for merging multiple biological modalities?
3. What would you propose as a simple strategy to integrate multiple modalities for a set of patients?

Today

- Finding a joint embedding using Grassmann manifold and Rayleigh Ritz
- Multiomics Factor Analysis (MOFA)

Intermission for Announcements

- Next homework to be assigned $\sim$ April 6. Now is a good time to work on projects!
- Don't forget about reading summaries

Multiomics Integration TCGA Example

The focus on merging multiple datasets was inspired by The Cancer Genome Atlas, an effort to profile large patient cohorts of patients with various cancer types, with several modalities.

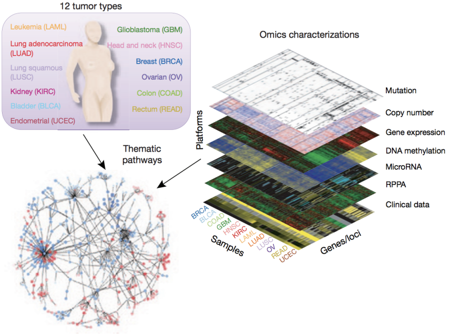

Figure: from TCGA, Nature Genetics. 2013.

Notation and Problem Formulation

- Consider M types of omics data measurements $\{\mathbf{X}^m\}_{m=1}^{M}$ from the same set of N patients.
- For a modality, m, there are p_m measured features and the dimensions of the data matrix are therefore $p_m \times N$
- We will let G^m be the graph for modality m

Comment

Before we had node2vec, we just used nice theorems from linear algebra!
:D (graph embedding for old people)

Overview of Subspace Merging

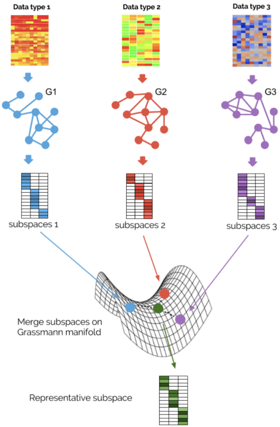

Figure: from Ding *et al.* Bioinformatics. 2019.

Build a Similarity Graph Between Patients in Each Modality

Use our ‘favorite‘ rule for calculating edge weights as,

$$S_{ij}^{m} = \exp\left(-\frac{\left\|\mathbf{x}_i^m - \mathbf{x}_j^m\right\|^2}{2t^2}\right), i = 1, \ldots, N, j = 1, \ldots . N$$

From here, retain the top k edges for each node based on S_{ij} and use W_{ij} for the notation of the edge weights retained, such that, $W_{ij}^{m} = S_{ij}^{m}$

Connection to Some GSP Conversation from a Few Weeks Ago

We already talked about the total variation of a signal in terms of the Graph Laplacian, or the variation of a signal around neighbors as,

$$\mathbf{x}^T \mathbf{L} \mathbf{x} = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{ij} (x_i - x_j)^2 \tag{1}$$

Pause for Rayleigh Ritz Theorem

Let $\mathbf{A}$ be a square, symmetric matrix, $N \times N$ matrix with eigenvalues, $\lambda_1 \leq \lambda_2 \cdots \leq \lambda_n$ and corresponding eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_n\}$. Then define

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}. \tag{2}$$

Then the minimum value of $R_{\mathbf{A}}(\mathbf{x})$ is λ_1 and it's taken for $\mathbf{x} = \mathbf{v}_1$

Matrix Extension

We will be seeing a lot on the form of $\mathbf{X}^T\mathbf{L}\mathbf{X}$. We can talk about the trace of that matrix product as the distance in vectors of adjacent nodes.

$$\text{trace}(\mathbf{X}^T\mathbf{L}\mathbf{X}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}||\mathbf{x}_i - \mathbf{x}_j|| \tag{3}$$

An extension of Rayleigh Ritz says that the minimum k-dimension matrix $\mathbf{X}$ of trace($\mathbf{X}^T\mathbf{L}\mathbf{X}$) is $\lambda_1 + \lambda_2. + \cdots + \lambda_k$ and corresponds to the first k eigenvectors of $\mathbf{L}$.

Specify Optimization Problem in terms of Normalized Graph Laplacian

$$\mathbf{L}^m = \mathbf{D}^{m^{-\frac{1}{2}}} \left(\mathbf{D}^m - \mathbf{W}^m\right) \mathbf{D}^{m^{-\frac{1}{2}}}$$

Written out this gives us,

$$L_{i,j}^{\mathrm{sym}} := \begin{cases} 1 & \text{if } i = j \text{ and } \deg\left(v_i\right) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg\left(v_j\right)}} & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise.} \end{cases}$$

Writing Down the Objective Function

The goal is to specify a $\mathbf{U}^m$ for each modality. The optimal graph embedding in k dimensions can written as,

$$\min_{\mathbf{U}^m \in \mathbb{R}^{N \times k}} \operatorname{tr}\left(\mathbf{U}^{m\prime} \mathbf{L}^m \mathbf{U}^m\right), \quad \text{s.t. } \mathbf{U}^{m\prime} \mathbf{U}^m = I$$

- It turns out the solution is the first k eigenvectors of the Graph Laplacian $\mathbf{L}^m$ by the Rayleigh–Ritz theorem

Merging Subspaces on a Grassmann Manifold

- With the subspace representations $\mathbf{U}_{m=1}^{M}$ from each data type, these will be merged on a Grassmann manifold
- A Grassmann manifold is defined as a set of linear subspaces of a euclidean space.
- To merge all $\mathbf{U}^m$, we seek to define an integrative subspace, $\text{span}(\mathbf{U}^m)$ that should also preserve connectivity in each G^m.

Defining a Projection Distance Between The Integrative Subspace and Individual Modality Subspaces

$$
\begin{aligned}
d^2_{\text{proj}}\ \left(\mathbf{U}, \{\mathbf{U}^m\}_{m=1}^M\right) &= \sum_{m=1}^{M} d^2_{\text{proj}}\ (\mathbf{U}, \mathbf{U}^m) \\
&= \sum_{m=1}^{M} \left[k - \operatorname{tr}\left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime}\right)\right] \\
&= kM - \sum_{i=1}^{M} \operatorname{tr}\left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime}\right)
\end{aligned}
$$

The subspace, **U** that minimizes this is close to all individual subspaces, $\{\mathbf{U}^m\}_{i=1}^M$

Optimization Problem for Multiple Subspaces

The optimization problem for merging multiple subspaces finally can be written as,

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times k}} \sum_{m=1}^{M} \operatorname{tr}\left(\mathbf{U}'\mathbf{L}^{m}\mathbf{U}\right) + \alpha\left[kM - \sum_{m=1}^{M} \operatorname{tr}\left(\mathbf{U}\mathbf{U}'\mathbf{U}^{m}\mathbf{U}^{m\prime}\right)\right], \quad \text{s.t. } \mathbf{U}'\mathbf{U} = I$$

The authors showed that this simplifies to,

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times k}} \operatorname{tr}\left[\mathbf{U}'\left(\sum_{i=1}^{M}\mathbf{L}^{m} - \alpha\sum_{m=1}^{M}\mathbf{U}^{m}\mathbf{U}^{m\prime}\right)\mathbf{U}\right], \quad \text{s.t. } \mathbf{U}'\mathbf{U} = I$$

Rayleigh Ritz Again....

$$\min_{\mathbf{U}\in\mathbb{R}^{n\times k}} \operatorname{tr}\left[\mathbf{U}'\left(\sum_{i=1}^{M}\mathbf{L}^m - \alpha\sum_{m=1}^{M}\mathbf{U}^m\mathbf{U}^{m\prime}\right)\mathbf{U}\right], \quad \text{s.t. } \mathbf{U}'\mathbf{U} = I$$

Hopefully you recognize the form of the objective. We can define a new matrix, $\mathbf{L}_{\text{mod}}$ and again the first k eigenvectors are the optimal solution. Or,

$$\mathbf{L}_{mod} = \sum_{m=1}^{M}\mathbf{L}^m - \alpha\sum_{m=1}^{M}\mathbf{U}^m\mathbf{U}^{m\prime}$$

Clustering on Merged Subspace

When you cluster on the merged subspace, you get groups with different prognostic interpretations.

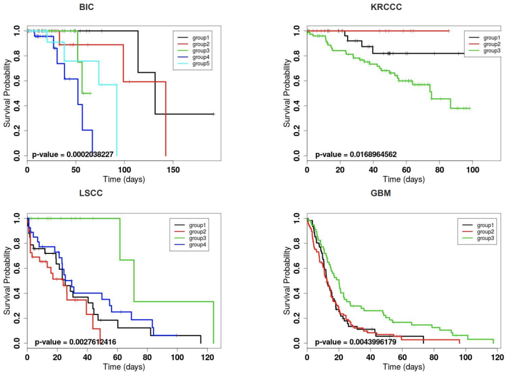

Figure: from Ding *et al.* Bioinformatics. 2018.

Another View : Between Patient Similarity

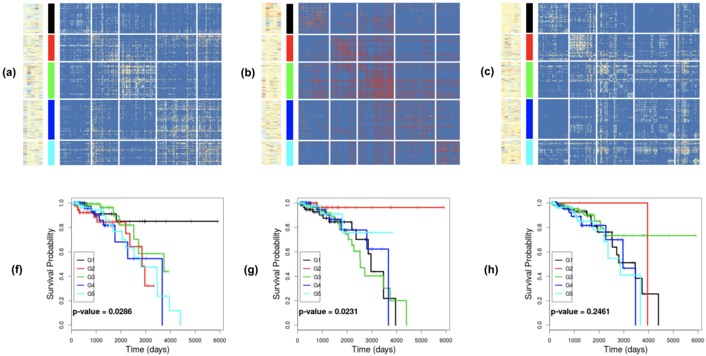

Figure: from Ding *et al.* Bioinformatics. 2018. Here we are viewing adjacency matrices between patients under different breast cancer subtypes based on features from mRNA, miRNA, and methylation.

And when you integrate all modalities...

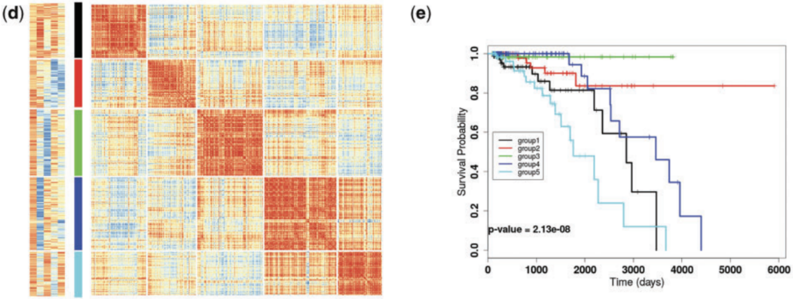

Figure: from Ding *et al.* Bioinformatics. 2018.

Same Story, Slightly Different Approach

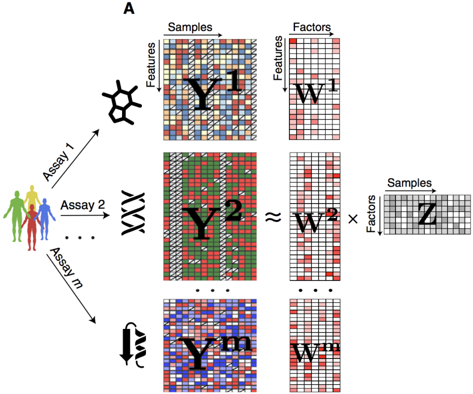

Figure: from Argelaguet *et al.* Molecular Systems Biology. 2018.

Complaints by MOFA Authors

The MOFA authors sought to develop a method that facilitates the following.

- **Interpretability:** The goal is to reconstruct the underlying ‘factors‘ that drive variation across samples in each modality.
- **Biological vs Technical Variation:** Should effectively untangle biological vs technical variation
- **Decomposing Variability:** Identify particular features and particular ‘factors‘ driving significant variance within and between modalities.

Notation and Goals

- Start with M data matrices, $\mathbf{Y}^1, \mathbf{Y}^2, \ldots, \mathbf{Y}^m$ on a set of N samples
- A Particular modality's matrix, $\mathbf{Y}_m$ has dimensions, $N \times D_m$

The goal is to write the input data matrix, $\mathbf{Y}^m$ as a product of a common factor matrix, $\mathbf{Z}$ (factors × samples) and a 'wight' matrix (e.g. one that relates features to factor space) as,

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

Unpacking...

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

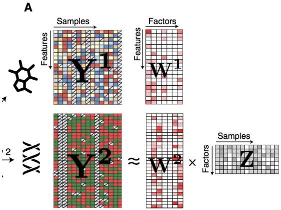

Figure: from Argelaguet *et al.* Molecular Systems Biology. 2018.

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‘).

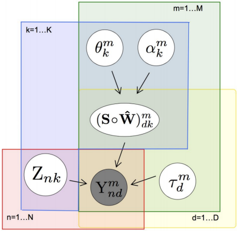

Figure: from Argelaguet *et al.* Molecular Systems Biology. 2018. As always, gray nodes are observed variables and white notes are the unobserved variables inferred by the model.

Starts Simple Enough....

The the observed value of feature *d* in sample *n* from modality *m* is modeled as,

$$y_{nd}^{m} \sim \mathcal{N}(y_{nd}^{m} \mid \mathbf{z}_{n,:}\mathbf{w}_{d,:}^{mT}, 1/\gamma_{d}^{m})$$

- $\mathbf{w}_{d,:}^{m}$ gives the *d*-th row of $\mathbf{W}^{m}$
- $\mathbf{z}_{n,:}$ is the *n*-th row of the latent factor matrix, $\mathbf{Z}$.

Regularization in $\mathbf{W}^m$

In computing each $\mathbf{W}^m$ there two desired kinds of sparsity :

- View and factor-wise sparsity
- Feature-wise sparsity

The intuition is that $\mathbf{w}^m_{:,k}$ is shrunk to 0 if factor k does not drive any variation in view m.

In practice...

Sparsity is enforced for each **W** through appropriate priors. Specifically, they model **W** as a product of a Gaussian random variable, $\hat{w}$, and a Bernoulli random variable, s.

$$\mathbf{W} = \mathbf{S}\hat{\mathbf{W}}$$

- $s^m_{dk} \sim \text{Bernoulli}(s^m_{d,k} \mid \theta^m_k)$
- $\hat{W}^m_{dk} \sim \mathcal{N}(0, 1/\alpha_k)$

Thinking More About Sparsity

Let's think about what happens with bernoulli probabilities:

- A value of θ_k^m close to 0 implies that most of the weights of factor k in view m tend towards 0.
- A θ_k^m close to 1 alternatively implies that most of the weights are non-zero (e.g. a non-sparse factor)

Further, there are priors for $\alpha_k^m \sim$ Beta and $\sigma_k^m \sim \Gamma$