Comp790-166: Computational Biology

Lecture 13

March 23, 2021

Announcements

- Thanks for your great project presentations
- CS Faculty comp bio candidate talk at 11am today (immediately after class meeting). Please attend if you are interested!

Today

- Finish up Grassmann Subspace Merging
- Dealing with Longitudinal Multiomodal Data
- Predicting cognitive scores in a longitudinal dataset of AD patients (integrating clinical, biological and imaging data)
- Start ADMM (classic optimization...)

The Cancer Genome Atlas (TCGA) – Reminder!

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.
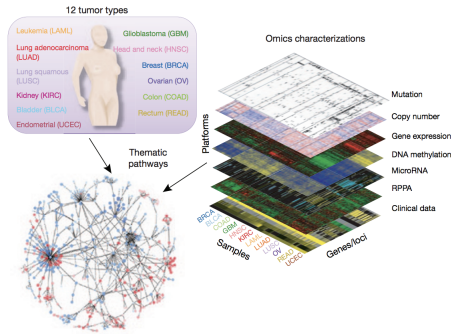

Figure: from TCGA, Nature Genetics. 2013.

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 $p_m \times N$

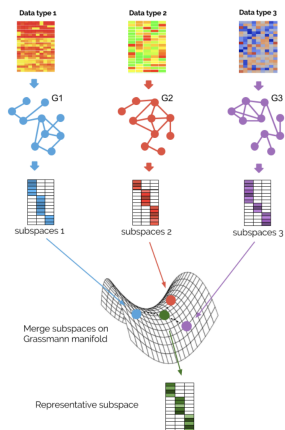- We will let G^m be the graph for modality m

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

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,

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

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 k-dimensional $\mathbf{X}$ that minimizes $\text{trace}(\mathbf{X}^T \mathbf{L} \mathbf{X})$ corresponds to the first k eigenvectors of $\mathbf{L}$.

Specify Optimization Problem in terms of Normalized Graph Laplacian

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

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,

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

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,

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

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 an extension of 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

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.

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

$$d_{\text{proj}}^2 \left(\mathbf{U}, \{\mathbf{U}^m\}_{m=1}^M \right) = \sum_{m=1}^M d_{\text{proj}}^2 \left(\mathbf{U}, \mathbf{U}^m \right)$$

$$= \sum_{m=1}^M \left[k - \text{tr} \left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime} \right) \right]$$

$$= kM - \sum_{i=1}^M \text{tr} \left(\mathbf{U}\mathbf{U}'\mathbf{U}^m\mathbf{U}^{m\prime} \right)$$

The subspace, $\mathbf{U}$ 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,

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

Optimization Problem for Multiple Subspaces

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

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

The authors showed that this simplifies to,

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

Rayleigh Ritz Again....

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

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

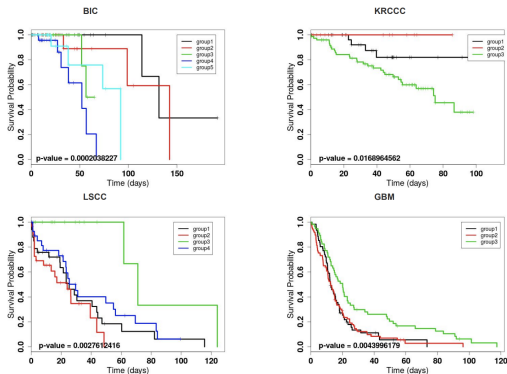When you cluster on the merged subspace, you get groups with different prognostic interpretations.

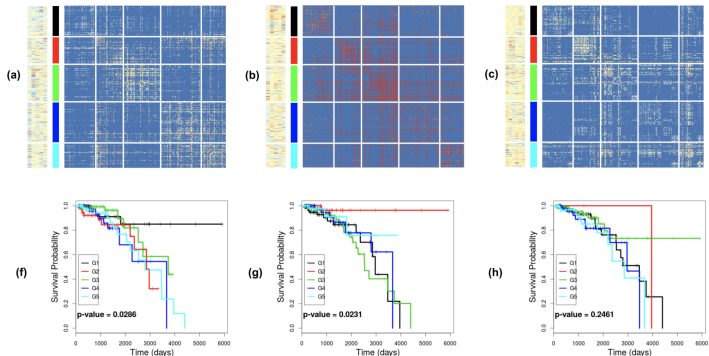# Another View : Between Patient Similarity

Figure: from Ding *et al.* Bioinformatics. 2018. Here we are viewing adjacency matrices between patients, based on all features jointly.

Transitioning to Other Multimodal Data Integration Problems in Biology

- Everything becomes more interesting when you add a time axis

Transitioning to Other Multimodal Data Integration Problems in Biology

- 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

Transitioning to Other Multimodal Data Integration Problems in Biology

- 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

Transitioning to Other Multimodal Data Integration Problems in Biology

- 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.

Transitioning to Other Multimodal Data Integration Problems in Biology

- 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)

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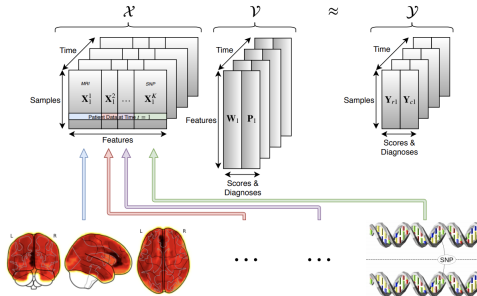http://adni.loni.usc.edu/data-samples/data-types/

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 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.

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)

- **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 $\times$ features $\times$ timepoints.

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 $\times$ features $\times$ timepoints.

- Note that each $\mathbf{X}_t$ can be broken down across the K modalities as, $\{\mathbf{X}_t\}_{j=1}^{K}$

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 $\times$ features $\times$ 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, $\mathcal{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_T\} \in \mathbb{R}^{n \times c \times T}$

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 $\times$ features $\times$ 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).

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 $\times$ features $\times$ 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 ($\mathbf{W}$s) and classification ($\mathbf{P}$s) across the T timepoints.
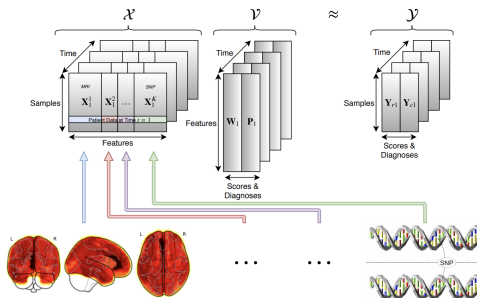
Figure: from Brand, Wang, emphet al. PacSym Biocomputing. 2020.

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.

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.

- $\mathcal{R}(\mathcal{V})$ is a regularization function applied to the matrix unfolded from the tensor, $\mathcal{V} \rightarrow \mathbf{V}^{d \times cT}$. 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 as,

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

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

To associate image and genomic features to cognitive scores and diagnoses over time, apply an $\ell_{2,1}$ norm 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 $\mathbf{V}$ corresponding to modality j as,

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

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,

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

- 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.

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.

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,

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