Comp790-166: Computational Biology

Lecture 8

February 11, 2021

Announcements

- No class on Tuesday- Wellness day!
- Homework due in 1 week!

Topics for Today

- Finish meld
  - A couple more GSP basics
  - Graph Fourier Transform
  - Low-pass filtering
  - Meld's low pass filter

MELD Overview Recap

Compute an enhanced experimental signal (EES) that explains how prototypical a cell is for a particular experimental condition.

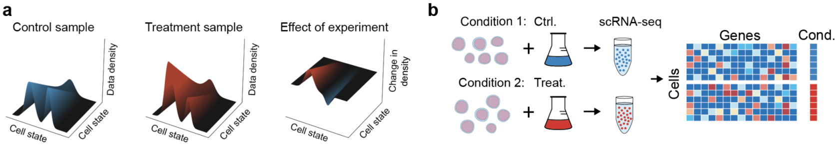


Figure: Burkhardt *et al.*, Nature Biotechnology. 2021

General Overview of the Steps of MELD

- Build a graph between cells based on gene or protein expression measurements
- **Graph Signals:** Experimental label (a binary indicator) is used to label each cell according to experimental condition
- Using GSP techniques, MELD filters biological and technical noise to look at how much the experimental signal of a cell matches the true experimental label. This quantifies how prototypical each cell is in its condition.
- Relate back to cell-types and features that differ between experimental conditions

EES represents the enhanced experimental signal, in comparison to RES, which was the raw, binary signal.

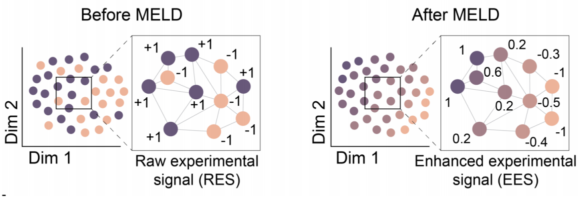


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021

Sources of Noise

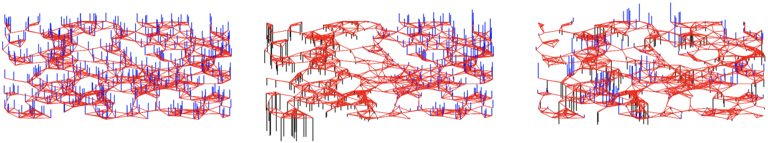
- Cells with similar feature measurements are said to be in the same state (biologically)
- **High Frequency Noise** : High frequency noise is when the labels of neighboring cells are rapidly fluctuating.
- Graph Fourier Transform is used to study the frequency of a signal over an irregular domain, like a graph.

What is GFT (on a high level?)

- Explain frequency content of the experimental labels (aka graph signal) as a weighted sum of the eigenvectors of the Graph Laplacian
- The eigenvectors of the Graph Laplacian comprise the **Graph Fourier Basis and can help to decouple high and low frequency signals**

Zero Crossings

The signs of the entries in the between nodes connected in the graph tend to be different more for the eigenvectors corresponding to higher eigenvalues



$\mathbf{u}_0$ $\mathbf{u}_1$ $\mathbf{u}_{50}$

Figure: from GSP Review https://arxiv.org/abs/1211.0053

Example

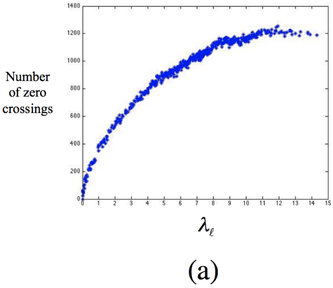


Figure: from `https://arxiv.org/abs/1211.0053`

- Higher eigenvalues (x-axis) correspond to higher frequency eigenvectors.

Local Variation of a Signal

The local variation of a signal or the sum of differences around a node can be written as,

$$
\begin{aligned}
(\mathcal{L}\mathbf{f})(i) &= ([\mathbf{D} - \mathbf{A}]\mathbf{f})(i) && (1)\\
&= d(i)\mathbf{f}(i) - \sum_j A_{ij}\mathbf{f}(j) && (2)\\
&= \sum_j A_{ij}(\mathbf{f}(i) - \mathbf{f}(j)) && (3)
\end{aligned}
$$

Local Variation Leads to Total Variation

The total variation of a signal on a graph is defined as follows and is also known as the Laplacian Quadratic Form

$$
\begin{aligned}
TV(\mathbf{f}) &= \sum_{i,j} A_{ij}(\mathbf{f}(i) - \mathbf{f}(j))^2 && (4) \\
&= \mathbf{f}^T \mathcal{L} \mathbf{f} && (5)
\end{aligned}
$$

- Note here I have been assuming that we have an unweighted graph, but you could certainly substitute A_{ij} with a weighted version, W_{ij}

Getting to Graph Fourier Basis

- We can look at eigenvectors, $\mathbf{\Psi} = [\psi_1, \psi_2, \ldots, \psi_N]$ of $\mathcal{L}$
- and eigenvalues, $\mathbf{\Lambda} = [0 = \lambda_1 \leq \cdots \leq \lambda_N]$ of $\mathcal{L}$

The Graph Fourier Transform of a Signal

The Graph Fourier Transform ($\hat{\mathbf{f}}$) of a signal, $\mathbf{f}$ can be written as,

$$\hat{f}(\lambda_\ell) = \sum_i f(i)\psi_\ell^T(i) =< \mathbf{f}, \psi_\ell > \tag{6}$$

Said otherwise in matrix form as,

$$\hat{\mathbf{f}} = \boldsymbol{\Psi}^T \mathbf{f} \tag{7}$$

GFT Will Be Used to Filter

- A filter on the graph will take in a signal and attenuate it according to a frequency response function.
- **Low-Pass Filter:** We filter or preserve only frequencies corresponding to eigenvalues below some threshold, λ_k. So, consider frequencies λ_b, with $\lambda_b < \lambda_k$
- **High-Pass Filters**: Preserve only frequencies corresponding to eigenvalues above some threshold, λ_k. So, consider frequencies λ_b, with $\lambda_b \geq \lambda_{k+1}$

A Simple Low-Pass Filter

Define some filter h as,

$$h : [0, \max(\mathbf{\Lambda})] \to [0, 1] \tag{8}$$

Assuming the cutoff is λ_k,
$h(x) > 0$, for $x < \lambda_k$ and $h(x) = 0$, otherwise

Defining Notation

To match notation from the MELD paper, define $h(\mathbf{\Lambda})$ as a diagonal matrix of eigenvalues with the filter applied.

Filtering a Signal Based on GFT

Based on what we computed with GFT, the filtered signal, $\hat{f_{filt}}$ can be computed as,

$$\hat{\mathbf{f}}_{filt} = h(\mathbf{\Lambda})\hat{\mathbf{f}} \tag{9}$$

Incorporating these ideas into meld

- Low frequency components are thought to be where the true signal comes from (e.g. cell states that can differentiate groups)
- Define a latent variable $\mathbf{z}$ that describes the biological process that differs between the two conditions

An optimization problem can be defined for low pass filtering as,

$$\mathbf{y} = \underset{\mathbf{z}}{\operatorname{argmin}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\mathbf{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\mathbf{b}} \tag{10}$$

Unpacking

y is the EES or Enhanced Experimental Signal

$$\mathbf{y} = \operatorname*{argmin}_{\mathbf{z}} \underbrace{\|\mathbf{x} - \mathbf{z}\|_2^2}_{\mathbf{a}} + \underbrace{\beta \mathbf{z}^T \mathcal{L} \mathbf{z}}_{\mathbf{b}} \tag{11}$$

- The Laplacian Regularization (term b) acts as a low-pass filter for an input graph signal, **x**
- (**a**) Term a represents reconstruction between **x** and **z**
- (**b**) Term b represents Laplacian regularization or a measure of smoothness on the graph. Recall this looks a lot like total variation.

$$\beta \mathbf{z}^T \mathcal{L} \mathbf{z} = \beta \sum_{i,j} A_{ij} (\mathbf{z}(i) - \mathbf{z}(j))^2 \tag{12}$$

Introducing the MELD Filter

They adjust the filer a bit as follows. The following allows also for a flexible notion of figure order, ρ,

$$\mathbf{y} = \underset{z}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{z}\|_2^2 + \mathbf{z}^T \mathcal{L}_* \mathbf{z}$$
$$\text{where } \mathcal{L}_* = [\beta \mathcal{L} - \alpha \mathbf{I}]^\rho \tag{13}$$

Takeaway

They show that their Laplacian Regularization is a filter with the following frequency response,

$$h_{\mathrm{MELD}}(\lambda) = \frac{1}{1 + (\beta\lambda - \alpha)^{\rho}} \tag{14}$$

This was a lot to unpack. I recommend staring at the details (if you are interested) in
`https://www.biorxiv.org/content/10.1101/532846v1.full.pdf`

Filter Variety

Here are some experiments showing what parameters on the MELD filter will do to the frequency response, $h(\lambda)$.

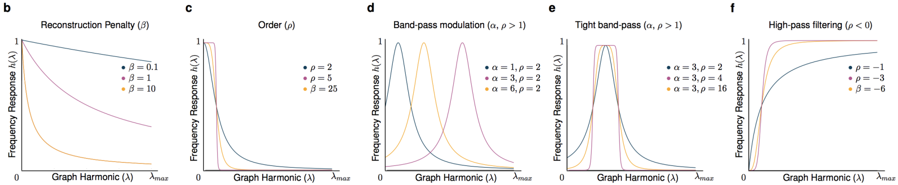


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021. Negative values of ρ, for example, can produce a high-pass filter.

Reminder : What we Do with a Filter

Given GFT, $\hat{\mathbf{f}}$, our filtered signal is computed as

$$\hat{\mathbf{f}}_{filt} = h(\mathbf{\Lambda})\hat{\mathbf{f}} \tag{15}$$

- As a reminder, $h(\mathbf{\Lambda})$ applied $h()$ to each eigenvalue

Meld Results

Computing the EES cleans up some of the noise and helps to better identify prototypical cells in each experimental condition.

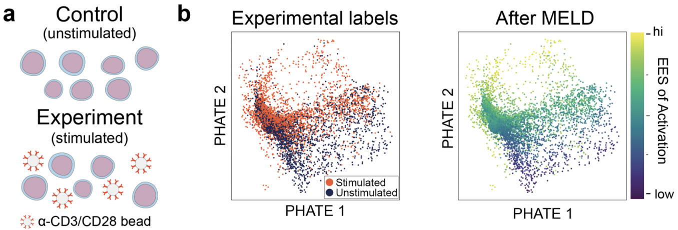


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

Gene Expression Profiles Based on RES and EES

You can look at the gene expression profiles of cells with similar EES scores.

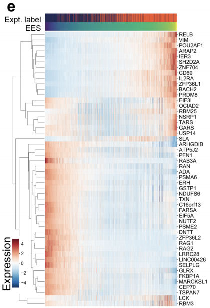


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

Zooming in on High and Low Frequency Regions

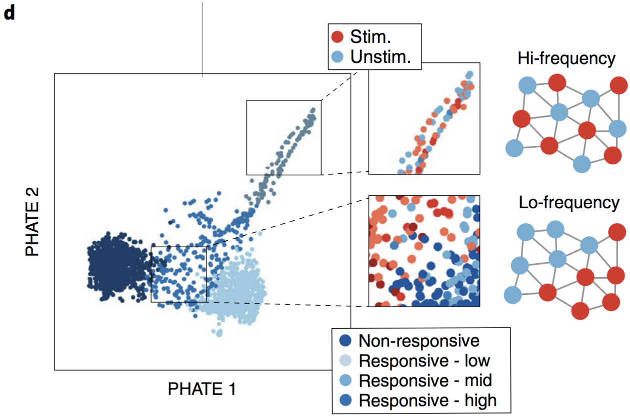


Figure: from Burkhardt *et al.*, Nature Biotechnology. 2021.

What's Coming up Next?

- MELD defines clusters of cells based on both features measured per cells and frequency information
- We will soon start some papers focused on differential analysis of cell populations.
- Such approach is ‘univariate‘ in the sense that a bunch of individual things are being tested for differences.