A soft introduction to Spectral Graph theory

Haitao Mao

Michigan State University

September 25th, 2022

Outline

Lesson introduction

What can we learn from this course:

- What problem can the Spectral Graph Theory solve?
- What are the key matrices in spectral graph theory?
- How does the spectral graph theory connect with the eigenvalue?
- An application: Graph drawing problem.

What can graph theory solve?

Problems

Problems listed in Prof. Teng's book Chap 2.4

- Significant Nodes: Ranking and Centrality
- Coherent Groups: Clustering and Communities
- Interplay between Networks and Dynamic Processes
- Multiple Networks: Composition and Similarity

Significant Nodes: Ranking and Centrality

Identifying nodes of relevance and significance. e.g.:

> *Which nodes are the most* **significant** *nodes in a network or a sub-network? How quickly can we identify them?*

Significance could be measured either *numerically*, or by *ranking* the nodes.

> *Network centrality is a form of "dimensionality reduction" from "high dimensional" network data to "low dimensional" centrality measures or rankings.*

e.g. PageRank

Coherent Groups: Clustering and Communities

Identifying groups with significant structural properties.
Fundamental questions include:

- What are the significant clusters in a data set?
- How fast can we identify one, uniformly sample one, or enumerate all significant groups?
- How should we evaluate the consistency of a clustering or community-identification scheme?
- What desirable properties should clustering or community identification schemes satisfy?

Interplay between Networks and Dynamic Processes

Understanding the interplay between dynamic processes and their underlying networks.

A given social network can be part of different dynamic processes (e.g. epidemic spreading, viral marketing), which can potentially affect the relations between nodes. Fundamental questions include:

- How should we model the interaction between network nodes in a given dynamic process?
- How should we characterize node significance and group coherence with respect to a dynamic process?
- How fast can we identify influential nodes and significant communities?

Multiple Networks: Composition and Similarity

To understand multiple networks instead of individual networks.

- network composition, e.g. multi-layer social network, multi-view graphs
- network similarity
  - similarity between two different networks
  - construct a sparser network that approximates a known one

Key matrices in Spectral Graph Theory

Graph Definition

$G = (V, E)$ (Friendship graphs, Network graphs, Circuit graphs, Protein-Protein Interaction graphs, etc.)

- G: a graph/network
- V: its vertex/node set
- E: its edge set (pair of vertices); edges have weight 1 by default, could assign other weights optionally.

By default (unless otherwise specified), a graph to be discussed will be:

- undirected (unordered vertices pairs in E)
- simple (having no loops or multiple edges)
- finite (V and E being finite sets)

Matrices for Graphs

Why we care about matrices?
Given a vector $\mathbf{x} \in \mathbb{R}^n$ and a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$

- $\mathbf{M}$ could be an operator: $\mathbf{Mx} \in \mathbb{R}^n$
- $\mathbf{M}$ could be used to define a quadratic form: $\mathbf{x}^T\mathbf{Mx} \in \mathbb{R}$ (here it has to be $m == n$)

Matrices for Graphs: adjacency matrix

Adjacency matrix $\mathbf{M}_G$ of $G = (V, E)$:

$$\mathbf{M}_G(a, b) = \begin{cases} 1 & \text{if } (a, b) \in E \\ 0 & \text{otherwise} \end{cases}$$

- most natural matrix to associate with a graph
- least "useful" (means directly useful, but useful in terms of generating other matrices)

  This statement is made because it is only a **spreadsheet**, *neither a natural* **operator** *or a natural* **quadratic form**.

Matrices for Graphs: diffusion on adjacent matrix

Diffusion operator $\mathbf{D}_G$ of $G = (V, E)$ is a diagonal matrix, probably the most natural operator associated with G:

$$\mathbf{D}_G(a, a) = \mathbf{d}(a)$$

where $\mathbf{d}(a)$ is the degree of vertex a.

- unweighted case: number of edges attached to it
- weighted case: weighted degree

$$\mathbf{d} \stackrel{\text{def}}{=} \mathbf{M}_G \mathbb{1}$$

Matrices for Graphs: random-walk Markov matrix

There is a linear operator $\mathbf{W}_G$ defined as:

$$\mathbf{W}_G = \mathbf{M}_G \mathbf{D}_G^{-1}$$

regarded as an operator denoting the *changes* of the graph between time steps.
Recall that diffusion operator $\mathbf{D}_G$ is a diagonal matrix, $\mathbf{W}_G$ is merely a rescaling of $\mathbf{M}_G$ if the graph is *regular* [1].

With vector $\mathbf{p} \in \mathbb{R}^n$ denoting the values of n vertices (called "*distribution of how much stuff*" in the textbook), the distribution of stuff at each vertex will be $\mathbf{W}_G \mathbf{p}$.

[1] Regular graph's vertices have the same degree.

Matrices for Graphs: intuition for random-walk Markov matrix

This matrix is called a random-walk Markov matrix:

$$\mathbf{W}_G = \mathbf{M}_G \mathbf{D}_G^{-1}$$

The next time step is:

$$\mathbf{W}_G \mathbf{p} = \mathbf{M}_G \mathbf{D}_G^{-1} \mathbf{p}$$

Think about the case where $\mathbf{p}$ is a one-hot vector δ_a where only $\delta_a(a) = 1$ and all other elements are 0.

$$\mathbf{W}_G \delta_a = \mathbf{M}_G \mathbf{D}_G^{-1} \delta_a = \mathbf{M}_G (\mathbf{D}_G^{-1} \delta_a)$$

We find the vector $\mathbf{D}_G^{-1} \delta_a$ has value $1/\mathbf{d}(a)$ at vertex a and 0 everywhere else; $\mathbf{M}_G \mathbf{D}_G^{-1} \delta_a$ has value $1/\mathbf{d}(a)$ at all a's **neighbors** and 0 otherwise.

Matrices for Graphs: Markov Matrix

A commonly-seen form of $\mathbf{W}_G$ is sometimes more convenient:

$$\widetilde{\mathbf{W}}_G = \mathbf{I}/2 + \mathbf{W}_G/2$$

describing a *lazy random walk* (1/2 chance stay, 1/2 chance go).

> *One of the purposes of spectral theory is to understand what happens when a linear operator like* $\mathbf{W}_G$ *is repeatedly applied.*

That is why it is called a random walk Markov matrix.

Matrices for Graphs: Markov Matrix

$$\mathbf{W}_G = \mathbf{M}_G \mathbf{D}_G^{-1}$$

has each column summing up to 1. $\mathbf{W}_G(a, b)$, the value on the a^{th} row b^{th} column, is $\mathbf{d}(b)$ if $(a, b) \in E$ else 0.

In fact, what $\mathbf{W}_G\mathbf{p}$ resulting in is a "random walk" based on the neighbors' degree.

$\mathbf{W}_G^T\mathbf{p}$ will be the random walk based on the degree of each node itself. (An example in the upcoming page.) It could be computed as:

$$\mathbf{W}_G^T = \mathbf{D}_G^{-1} \mathbf{M}_G$$

Matrices for Graphs: Markov Matrix Example

An example:

$$\mathbf{M}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{D}_G = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{D}_G^{-1} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{W}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 0 & 0 \end{bmatrix}$$

$$\mathbf{W}_G\mathbf{p} = \begin{bmatrix} p_2 + p_3 \\ p_1/2 \\ p_1/2 \end{bmatrix} \qquad \mathbf{W}_G^T\mathbf{p} = \begin{bmatrix} (p_2 + p_3)/2 \\ p_1 \\ p_1 \end{bmatrix}$$

Matrices for Graphs: Laplacian Matrix

Laplacian matrix $\mathbf{L}_G$, the most natural quadratic form associated with the graph G:

$$\mathbf{L}_G \stackrel{\text{def}}{=} \mathbf{D}_G - \mathbf{M}_G$$

Given a vector $\mathbf{x} \in \mathbb{R}^n$, who could also be viewed as a *function* over the vertices, we have:

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b)\in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

representing the Laplacian quadratic form of a weighted graph ($w_{a,b}$ is the weight of edge (a, b)), could be used to measures the smoothness of $\mathbf{x}$ (it would be small if $\mathbf{x}$ is not changing drastically over any edge).

Matrices for Graphs: Laplacian Matrix Example

An example ($w_{a,b} = 1$):

$$\mathbf{M}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{D}_G = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\mathbf{L}_G = \mathbf{D}_G - \mathbf{M}_G = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

$$\begin{aligned} \mathbf{x}^T \mathbf{L}_G \mathbf{x} &= x_1(2x_1 - x_2 - x_3) + x_2(-x_1 + x_2) + x_3(-x_1 + x_3) \\ &= 2x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 - 2x_1x_3 = (x_1 - x_2)^2 + (x_1 - x_3)^2 \end{aligned}$$

Matrices for Graphs : Incidence Matrix

Incidence Matrix: $\mathbf{I}_G$, where each row corresponds to an edge, and columns to vertices indexes.

A row, corresponding to $(a, b) \in E$, sums up to 0, with only 2 non-zero elements: the a^{th} column being 1 and b^{th} being -1, or could be the opposite (a^{th} column -1 and b^{th} column 1).
Following the previous example:

$$\mathbf{M}_G = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{I}_G = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$$

In the case of weighted graph, ± 1 should be $\pm w_{a,b}$ instead.
There's very interesting relation:

$$\mathbf{L}_G = \mathbf{I}_G^T \mathbf{I}_G$$

Matrices for Graphs: Incident Matrix

Explanation on the reason why:

$$\mathbf{L}_G = \mathbf{I}_G^T \mathbf{I}_G$$

could be from the perspective that, $\mathbf{L}_G$ is associated with Hessian and $\mathbf{I}_G$ be associated with Jacobian.

Also note that the introduction of the Incidence Matrix immediately makes this proof obvious:

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b)\in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \mathbf{x}^T \mathbf{I}_G^T \mathbf{I}_G \mathbf{x} = \|\mathbf{I}_G \mathbf{x}\|^2 = \sum_{(a,b)\in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

Matrices for Graphs: Laplacian Normalization I

In practice we always use **normalized** Laplacian matrices. Intuitively, we want all diagonal entries to be 1. In a way, that is somewhat "regularize" of the matrix.

There are many ways of normalizing a Laplacian matrix. Two of them are:

- (symmetric)

$$\mathbf{L}_s = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$$

- (random walk)

$$\mathbf{L}_{rw} = \mathbf{L}\mathbf{D}^{-1} = (\mathbf{D} - \mathbf{M})\mathbf{D}^{-1} = \mathbf{I} - \mathbf{M}\mathbf{D}^{-1}$$

Matrices for Graphs: Laplacian Normalization II

$\mathbf{L}_s$ preserves every property of $\mathbf{L}$. Such as being positive semidefinite:

$$\mathbf{x}^T \mathbf{L}_s \mathbf{x} = \sum_{(a,b)\in E} w_{a,b} \left(\frac{\mathbf{x}(a)}{\sqrt{\mathbf{d}(a)}} - \frac{\mathbf{x}(b)}{\sqrt{\mathbf{d}(b)}} \right)^2$$

Recall that $\mathbf{M}\mathbf{D}^{-1}$ is the random walk Markov matrix $\mathbf{W}$. $\mathbf{L}_{rw} = \mathbf{I} - \mathbf{W}$. Therefore, $\mathbf{W}$ and $\mathbf{L}_{rw}$ have the same eigenvectors, while the corresponding eigenvalues sum up to 1:

$$\mathbf{A}\mathbf{x} = \mu\mathbf{x} \iff (\mathbf{A} - k\mathbf{I})\mathbf{x} = (\mu - k)\mathbf{x}$$

$$\mathbf{W}\boldsymbol{\psi} = \lambda\boldsymbol{\psi} \iff (\mathbf{I} - \mathbf{W})\boldsymbol{\psi} = (1 - \lambda)\boldsymbol{\psi}$$

Matrices for Graphs: Laplacian Normalization III

Additional comments on λ and $1 - \lambda$:

Sometimes, for $0 \leq \lambda \leq 1$, after some operations, such as multiplying the matrix (say, $\mathbf{A}$) for multiple times, small eigenvalues will become close to zero.

However, if we consider a trick:

$$\mathbf{I} - \mathbf{A}$$

the corresponding eigenvalue will be $0 \leq 1 - \lambda \leq 1$. After power iteration, the smallest eigenvalue becomes the largest.

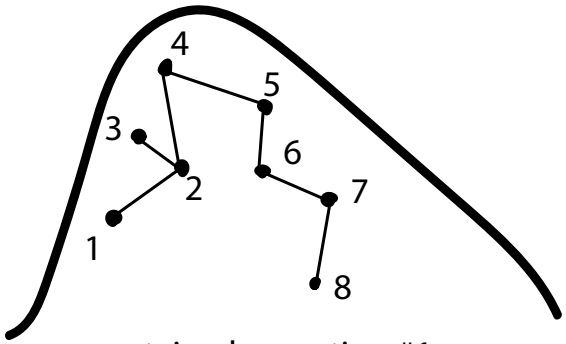
Intuitive Understanding of Graph Laplacian

Imagine that we are going to estimate the (absolute) height $\mathbf{h} \in \mathbb{R}^n$ of some selected points on a mountain. Let's say that there are n points to estimate in total.

Climbing up and down in the mountain, we have no clue what is its exact height, but we know k relative heights (e.g. relative height between vertices 1 and 2 is $\Delta_{1,2} = h_1 - h_2$). We denote the record of each relative height (the **edges**) as $\mathbf{m} \in \mathbb{R}^k$.

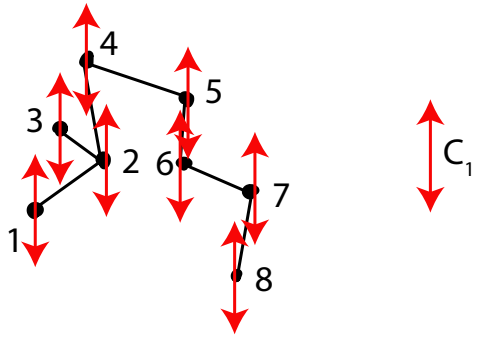
We denote the starting and ending of the nodes by an Incidence Matrix $\mathbf{I}_G \in \mathbb{R}^{k \times n}$.

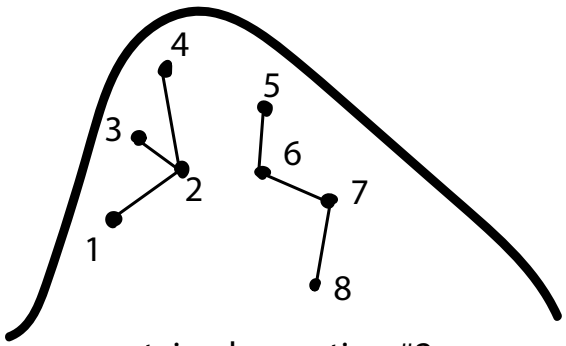
Illustration



mountain observation #1

#edge: k = 7
#node: n = 8
degree of freedom: 1





mountain observation #2

#edge: k = 6
#node: n = 8
degree of freedom: 2

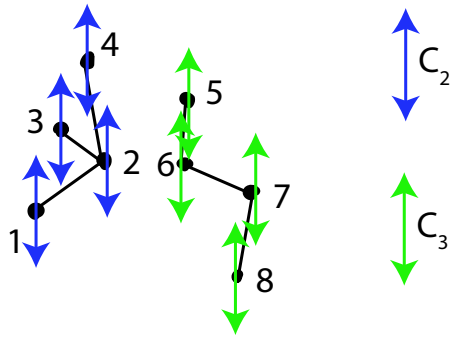


Figure: Illustration of the examples.

Illustration: mountain observation #1

$$\mathbf{m} = \mathbf{l}_G \mathbf{h}$$

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \\ m_7 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \end{bmatrix}$$

Illustration: mountain observation #2

$$\mathbf{m} = \mathbf{l}_G \mathbf{h}$$

$$\begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ m_5 \\ m_6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \end{bmatrix}$$

Problem

The problem is formally defined this way:

$$\mathbf{m} = \mathbf{l}_G \mathbf{h}$$

Knowing $\mathbf{m}, \mathbf{l}_G$, solving $\mathbf{h}$.

It is solved by minimizing over $\mathbf{h}$:

$$\|\mathbf{l}_G \mathbf{h} - \mathbf{m}\|^2$$

Recall that for any $\mathbf{Ax} = \mathbf{b}$ the solution is $\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$, since $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.

Solution

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

In this case, it means that:

$$\mathbf{I}_G^T\mathbf{I}_G\mathbf{h} = \mathbf{I}_G^T\mathbf{m}$$

Recall that the graph Laplacian $\mathbf{L}_G = \mathbf{I}_G^T\mathbf{I}_G$, therefore we have:

$$\mathbf{L}_G\mathbf{h} = \mathbf{I}_G^T\mathbf{m}$$

Just for convenience, we introduce a known value $\mathbf{b} = \mathbf{I}_G^T\mathbf{m} \in \mathbb{R}^n$.

$$\mathbf{L}_G\mathbf{h} = \mathbf{b}$$

Degree of Freedom

Now we consider the graph itself:

- #1: The graph is connected, but we will never know the **exact** absolute height of the mountain. Because whatever **h** value we result in, since we only know the nodes' relative height, it makes sense if we move the entire graph up and down alone the vertical direction. That is, after adding a constant value C_1 to every entry in **h**, we still result in a valid solution.
- #2: Similarly, this time we have 2 separate subgraphs, therefore, each subgraph could be moved up and down independently. Let's say that nodes in the two subgraphs can be shifted alone the vertical direction by C_2 and C_3 distance respectively.

This is why we say that the degree of freedom in case #1 is 1, and that in case #2 is 2.

Spectral Graph Theory & Eigenvalue

Spectral Theory

Theorem (The Spectral Theorem)

If M *is an n-by-n, real, symmetric matrix,* **then** *there exist real numbers* $\lambda_1, \ldots, \lambda_n$ *and n mutually orthogonal unit vectors* $\psi_1, \ldots, \psi_n$ *and such that* ψ_i *is an eigenvector of* **M** *of eigenvalue* λ_i*, for each* i*.*

If the matrix **M** is not symmetric, it might not have n eigenvalues. And, even if it has n eigenvalues, their eigenvectors will not be orthogonal (linearly independent). Many studies will no longer apply to it when the matrix is not symmetric.

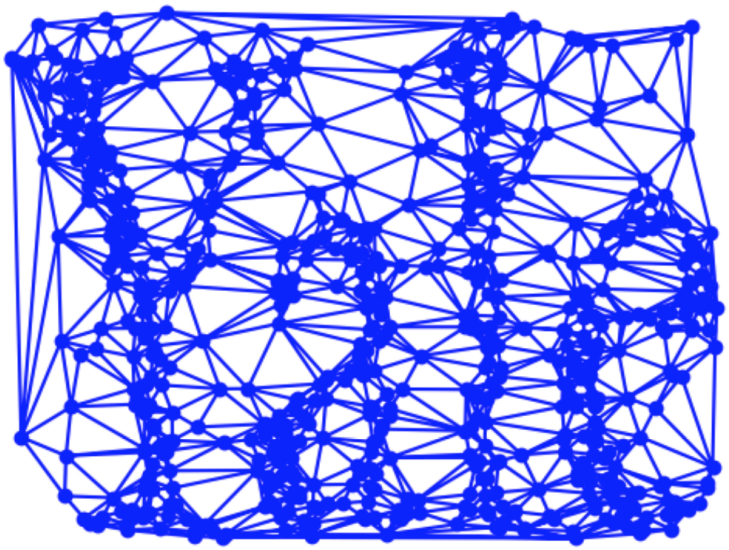
λ and μ

In this textbook, eigenvalues are sometimes denoted as λ and sometimes denoted as μ.

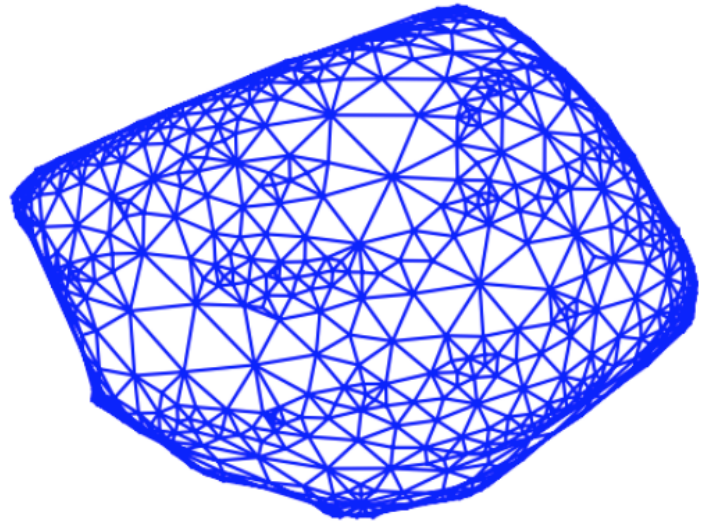
To my observation, they tend to use λ when the eigenvalues are ordered from the smallest to the largest, and μ when ordered from the largest to the smallest.

e.g., in the later chapters we'll see: eigenvalues of the adjacency matrix is denoted as μ (recall that we use λ for Laplacian's eigenvalues) and $\mu_1 \geq \mu_2 \cdots \geq \mu_n$. This is to make μ_i corresponds to λ_i.

Eigenvalue: Examples



(a) The original points sampled from Yale logo, with coordinates omitted and transformed into graph.



(b) Plot of vertices at $(\psi_2(a), \psi_3(a))$ coordinate.

Figure: An example showing the use of eigenvectors. More examples are listed in the textbook, Chap 1.

Example: Why Eigenvectors as Coordinates

Intuitively, using eigenvalues and eigenvectors could be regarded as mapping the nodes onto sine and cosine function curves.

The sine and cosine functions generally preserve the distances between a pair of nodes, but for some disturbance brought by the periods (can have the same value again at another point). However, the use of multiple sets of eigenvalue-eigenvectors, could be viewed as having multiple frequencies to measure.

Therefore, a pair of nodes that is far away might seem to be close measured by sine or cosine value on a certain frequency, but won't be always close to each other under different frequencies.

Example: Why Eigenvectors as Coordinates (*)

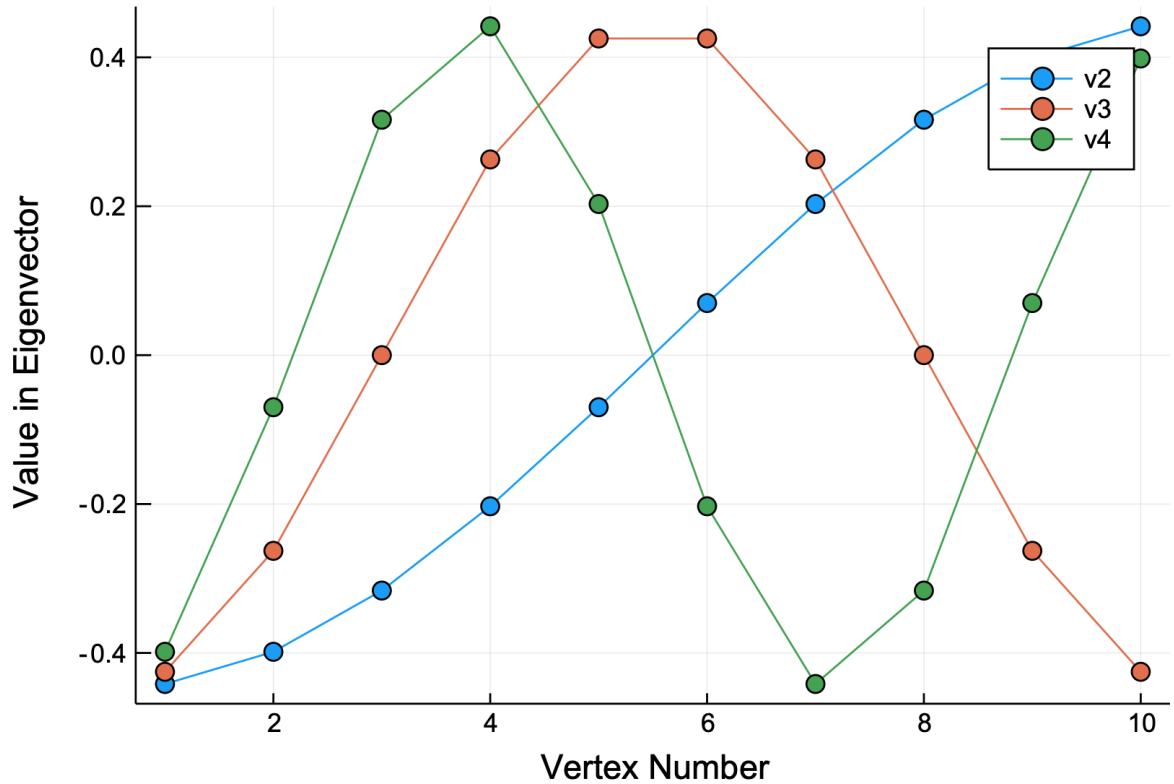


Figure: Plot of a length-4 path graph's (i.e. only $(i, i+1)$ are edges) Laplacian's eigenvectors $\mathbf{v}_2$, $\mathbf{v}_3$, $\mathbf{v}_4$, where $\lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \lambda_4$.

Connecting Eigenvalues and optimization problem

One reason why we are interested in **eigenvalues** of matrices is that, they arise as the solution to natural **optimization** problems.

The formal statement of this is given by the **Courant-Fischer Theorem**. And this Theorem could be proved by the **Spectral Theorem**.

The Courant-Fischer Theorem

It has various other names: the min-max theorem, variational theorem, Courant–Fischer–Weyl min-max principle.

> *It gives a variational characterization of eigenvalues of* **compact Hermitian operators** *on* **Hilbert spaces**.

- In the real-number field, a Hermitian matrix means a symmetric matrix.
- The real numbers $\mathbb{R}^n$ with $\langle \mathbf{u}, \mathbf{v} \rangle$ defined as the vector dot product of $\mathbf{u}$ and $\mathbf{v}$ is a typical finite-dimensional Hilbert space. [2]

[2] https://mathworld.wolfram.com/HilbertSpace.html

The Courant-Fischer Theorem

Theorem (2.0.1 Courant-Fischer Theorem)

Let $\mathbf{M}$ *be a symmetric matrix with eigenvalues* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. *Then,*

$$\mu_k = \max_{\substack{\mathcal{S}\subseteq\mathbb{R}^n \\ \dim(\mathcal{S})=k}} \min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \min_{\substack{\mathcal{T}\subseteq\mathbb{R}^n \\ \dim(\mathcal{T})=n-k+1}} \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

where the maximization and minimization are over subspaces $\mathcal{S}$ *and* $\mathcal{T}$ *of* $\mathbb{R}^n$.

The Courant-Fischer Theorem: Proof I

Using the Spectral Theorem to prove the Courant-Fischer Theorem.

> **Theorem (1.3.1 The Spectral Theorem)**
>
> **If M** *is an n-by-n, real, symmetric matrix,* **then** *there exist real numbers* $\lambda_1, \ldots, \lambda_n$ *and n mutually orthogonal unit vectors* $\psi_1, \ldots, \psi_n$ *and such that* ψ_i *is an eigenvector of* **M** *of eigenvalue* λ_i*, for each* i*.*

Main Steps:

- expanding a vector **x** in the basis of eigenvectors of **M**
- use the properties of eigenvalues and eigenvectors to prove it

The Courant-Fischer Theorem: Proof II

$\mathbf{M} \in \mathbb{R}^{n \times n}$: a symmetric matrix, with eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. The corresponding **orthogonal** eigenvectors are $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_n$.
Then we may write $\mathbf{x} \in \mathbb{R}^n$ as:

$$\mathbf{x} = \sum_i c_i \boldsymbol{\psi}_i, \qquad c_i = \boldsymbol{\psi}_i^T \mathbf{x}$$

Why $\mathbf{x}$ can be expanded in this way? (Intuitively obvious, but we need a mathematical explanation.)

The Courant-Fischer Theorem: Proof III

Let $\mathbf{\Psi}$ be a matrix whose columns are $\{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_n\}$ — orthogonal vectors. By definition, $\mathbf{\Psi}$ is an orthogonal matrix.

$$\mathbf{\Psi}\mathbf{\Psi}^T = \mathbf{\Psi}^T\mathbf{\Psi} = \boldsymbol{I}$$

Therefore we have:

$$\sum_i c_i\boldsymbol{\psi}_i = \sum_i \boldsymbol{\psi}_i c_i = \sum_i \boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\mathbf{x} = \Big(\sum_i \boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\Big)\mathbf{x} = \mathbf{\Psi}\mathbf{\Psi}^T\mathbf{x} = \mathbf{x}$$

and thus, since $\boldsymbol{\psi}_i^T\boldsymbol{\psi}_j = 1$ when $i = j$ and 0 otherwise,

$$\mathbf{x}^T\mathbf{x} = \Big(\sum_i c_i\boldsymbol{\psi}_i\Big)^T\Big(\sum_i c_i\boldsymbol{\psi}_i\Big) = \sum_{i,j} c_i^2\boldsymbol{\psi}_i^T\boldsymbol{\psi}_j = \sum_{i=1}^{n} c_i^2$$

The Courant-Fischer Theorem: Proof IV

Let's revisit the theorem to prove (Now we have $\mathbf{x}^T\mathbf{x}$, to prove it we need to consider $\mathbf{x}^T\mathbf{M}\mathbf{x}$):

Theorem (2.0.1 Courant-Fischer Theorem)

Let $\mathbf{M}$ *be a symmetric matrix with eigenvalues* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. *Then,*

$$\mu_k = \max_{\substack{\mathcal{S}\subseteq\mathbb{R}^n \\ \dim(\mathcal{S})=k}} \min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \min_{\substack{\mathcal{T}\subseteq\mathbb{R}^n \\ \dim(\mathcal{T})=n-k+1}} \max_{\substack{\mathbf{x}\in\mathcal{T} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

where the maximization and minimization are over subspaces $\mathcal{S}$ *and* $\mathcal{T}$ *of* $\mathbb{R}^n$.

The Courant-Fischer Theorem: Proof V

In the textbook, Lemma 2.1.1 suggests that, in the previous example, for any $\mathbf{x} = \sum_i c_i \boldsymbol{\psi}_i$:

$$\mathbf{x}^T \mathbf{M} \mathbf{x} = \sum_{i=1}^{n} c_i^2 \mu_i$$

Again, $\boldsymbol{\psi}_i^T \boldsymbol{\psi}_j = 1$ when $i = j$ and 0 otherwise, also because $\mathbf{M}\boldsymbol{\psi}_i = \mu_i \boldsymbol{\psi}_i$,

$$\begin{aligned}
\mathbf{x}^T \mathbf{M} \mathbf{x} &= \Big(\sum_i c_i \boldsymbol{\psi}_i \Big)^T \mathbf{M} \Big(\sum_i c_i \boldsymbol{\psi}_i \Big) \\
&= \Big(\sum_i c_i \boldsymbol{\psi}_i \Big)^T \Big(\sum_i c_i \mu_i \boldsymbol{\psi}_i \Big) \\
&= \sum_{i,j} c_i^2 \mu_i \boldsymbol{\psi}_i^T \boldsymbol{\psi}_j \\
&= \sum_i c_i^2 \mu_i
\end{aligned}$$

The Courant-Fischer Theorem: Proof VI

Take a look again:

Theorem (2.0.1 Courant-Fischer Theorem)

Let $\mathbf{M}$ *be a symmetric matrix with eigenvalues* $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$. *Then,*

$$\mu_k = \max_{\substack{\mathcal{S} \subseteq \mathbb{R}^n \\ \dim(\mathcal{S}) = k}} \min_{\substack{\mathbf{x} \in \mathcal{S} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \min_{\substack{\mathcal{T} \subseteq \mathbb{R}^n \\ \dim(\mathcal{T}) = n-k+1}} \max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}}$$

where the maximization and minimization are over subspaces $\mathcal{S}$ *and* $\mathcal{T}$ *of* $\mathbb{R}^n$.

The Courant-Fischer Theorem: Proof VII

We need the value of $\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$. In particular, we care about μ_k and subspace $\mathcal{S}$ where $\dim(\mathcal{S}) = k$. Also recall that we put $\{\mu_i\}_{i=1}^n$ in the **non-increasing** order.

$$\mathbf{x} = \sum_i^k c_i \boldsymbol{\psi}_i, \qquad c_i = \boldsymbol{\psi}_i^T \mathbf{x}$$

$$\frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \frac{\sum_i^k c_i^2 \mu_i}{\sum_i^k c_i^2} \geq \frac{\sum_i^k c_i^2 \mu_k}{\sum_i^k c_i^2} = \mu_k$$

Therefore,

$$\min_{\substack{\mathbf{x}\in\mathcal{S} \\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \geq \mu_k$$

The Courant-Fischer Theorem: Proof VIII

To prove the theorem, we also need to show that for all subspace $\mathcal{S} \subseteq \mathbb{R}^n$ where $\dim(\mathcal{S}) = k$,

$$\min_{\substack{\mathbf{x} \in \mathcal{S} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} \leq \mu_k$$

For this part we bring up the subspace $\mathcal{T}$ of dimension $n - k + 1$, whose basis vectors are $\psi_k, \ldots, \psi_n$. Similarly, for $\mathbf{x} \in \mathcal{T}$, we have:

$$\max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\mathbf{x}^T \mathbf{M} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \max_{\substack{\mathbf{x} \in \mathcal{T} \\ \mathbf{x} \neq \mathbf{0}}} \frac{\sum_k^n c_i^2 \mu_i}{\sum_k^n c_i^2} \leq \frac{\sum_k^n c_i^2 \mu_k}{\sum_k^n c_i^2} = \mu_k$$

The Courant-Fischer Theorem: Proof IX

Every subspace $\mathcal{S}$ of dimension k has an intersection with $\mathcal{T}$ (dimension $n-k+1$), the intersection has dimension at least 1 ($(n-k+1)+k=n+1$).

$$\min_{\substack{\mathbf{x}\in\mathcal{S}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \min_{\substack{\mathbf{x}\in\mathcal{S}\cap\mathcal{T}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \max_{\substack{\mathbf{x}\in\mathcal{S}\cap\mathcal{T}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \max_{\substack{\mathbf{x}\in\mathcal{T}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\mathbf{x}^T\mathbf{M}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \mu_k$$

The theorem is proved this way.

An application: Graph drawing problem

Graph Laplacian

Recall that weighted undirected graph $G = (V, E, w)$, with positive weigh $w : E \to \mathbb{R}^+$, is defined this way:

$$\mathbf{L}_G \stackrel{\text{def}}{=} \mathbf{D}_G - \mathbf{M}_G, \qquad \mathbf{D}_G = \sum_b w_{a,b}$$

where $\mathbf{D}_G$ is the diffusion matrix, $\mathbf{M}_G$ is the adjacency matrix.

Given a vector $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{x}^T \mathbf{L}_G \mathbf{x} = \sum_{(a,b) \in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

Hall's Idea on Graph Drawing

Hall's idea on graph drawing suggests that we choose the first coordinates of the n vertices as $\mathbf{x} \in \mathbb{R}^n$ that minimizes:

$$\mathbf{x}^T \mathbf{L} \mathbf{x} = \sum_{(a,b) \in E} w_{a,b} \big(\mathbf{x}(a) - \mathbf{x}(b)\big)^2$$

To avoid degenerating to $\mathbf{0}$, we have restriction:

$$\|\mathbf{x}\|^2 = \sum_{a \in V} \mathbf{x}(a)^2 = 1$$

To avoid degenerating to $\mathbb{1}/\sqrt{n}$, Hall suggested another constraint:

$$\mathbb{1}^T \mathbf{x} = \sum_{a \in V} \mathbf{x}(a) = 0$$

When there are multiple sets of coordinates, say $\mathbf{x}$ and $\mathbf{y}$; we require $\mathbf{x}^T \mathbf{y} = 0$, to avoid cases such as $\mathbf{x} = \mathbf{y} = \psi_2$.

Hall's Idea on Graph Drawing

We will minimize the sum of the squares of the lengths of the edges in the embedding. e.g. 2-D case:

$$\sum_{(a,b)\in E} \left\| \begin{bmatrix} \mathbf{x}(a) \\ \mathbf{y}(a) \end{bmatrix} - \begin{bmatrix} \mathbf{x}(b) \\ \mathbf{y}(b) \end{bmatrix} \right\|^2$$

$$= \sum_{(a,b)\in E} (\mathbf{x}(a) - \mathbf{x}(b))^2 + (\mathbf{y}(a) - \mathbf{y}(b))^2$$

$$= \mathbf{x}^T \mathbf{L}\mathbf{x} + \mathbf{y}^T \mathbf{L}\mathbf{y}$$

is the objective we want to minimize.

Properties to Prove

Here are some of the very interesting properties of a graph that we would like to prove.

- If and only if the graph is connected, there is only one eigenvalue of its Laplacian equals to zero.
- When mapping each vertex to a set of coordinates, choosing the coordinates to be the eigenvectors of the graph Laplacian is optimal.

Property #1

Lemma

Let $G = (V, E)$ be a graph, and let $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ be the eigenvalues of its Laplacian matrix, $\mathbf{L}$. Then, $\lambda_2 > 0$ if and only if G is connected.

Property #1: Proof I

First of all, there exists eigenvalue $\mathbf{0}$, because the all-one vector $\mathbb{1}$ satisfies:

$$\mathbf{L}\mathbb{1} = \mathbf{0}$$

To prove, if we view the Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{M}$ as an operator ($\mathbf{D} = \sum_{(a,b)\in E} w_{a,b}$), for each $\mathbf{x}$ we have its a^{th} entry of $\mathbf{Lx}$ being:

$$(\mathbf{Lx})(a) = d(a)\mathbf{x}(a) - \sum_{(a,b)\in E} w_{a,b}\mathbf{x}(b) = \sum_{(a,b)\in E} w_{a,b}\big(\mathbf{x}(a) - \mathbf{x}(b)\big)$$

It infers that $\mathbb{1}$ is an eigenvector corresponds to eigenvalue 0. Therefore, $\lambda_1 = 0$.

Next, we show that $\lambda_2 = 0$ if G is disconnected.

Property #1: Proof II

If G is disconnected, then we can split it into two graphs G_1 and G_2. Because we can safely reorder the vertices of a graph, we can have:

$$\mathbf{L} = \begin{bmatrix} \mathbf{L}_{G_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_{G_2} \end{bmatrix}$$

It has at least 2 orthogonal eigenvectors of eigenvalue zero:

$$\begin{bmatrix} \mathbf{0} \\ \mathbb{1} \end{bmatrix}, \text{ and } \begin{bmatrix} \mathbb{1} \\ \mathbf{0} \end{bmatrix}$$

Property #1: Proof III

On the other hand, for a eigenvector ψ of eigenvalue 0, $\mathbf{L}\psi = \mathbf{0}$,

$$\psi^T \mathbf{L} \psi = \sum_{(a,b) \in E} w_{a,b} \big(\psi(a) - \psi(b)\big)^2 = 0$$

For every pair of vertices (a, b) connected by an edge, we have $\psi(a) = \psi(b)$. In a connected graph, all vertices are directly or indirectly connected, and thus ψ must be a constant vector.

Contradiction found.

Therefore, G must be disconnected when $\lambda_2 = 0$.

Property #2

Theorem (3.2.1)

Let $\mathbf{L}$ *be a Laplacian matrix and let* $\mathbf{x}_1, \ldots, \mathbf{x}_k$ *be orthonormal* [a] *vectors that are all orthogonal to* $\mathbb{1}$*. Then*

$$\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{L} \mathbf{x}_i \geq \sum_{i=2}^{k+1} \lambda_i$$

and this inequality is tight only when $\mathbf{x}^T \psi_j = 0$ *for all* j *such that* $\lambda_j \geq \lambda_{k+1}$*.* λ_i *are the eigenvalues, the graph* G *is an undirected connected graph.*

[a] orthonormal = both orthogonal and normalized

Property #2: Proof I

We can order λ such that:

$$0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$$

As is proved before, $\lambda_1 = 0$ and because G is connected, ψ_1 is a constant vector.

Let $\mathbf{x}_{k+1} \ldots \mathbf{x}_n$ be vectors such that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is an orthogonal **basis**. It is done by choosing $\mathbf{x}_{k+1} \ldots \mathbf{x}_n$ to be an orthogonal basis of the space orthogonal to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Because they are orthogonal basis, (think of orthogonal matrix)

$$\sum_{j=1}^{n} (\psi_j^T \mathbf{x}_i)^2 = \sum_{j=1}^{n} (\mathbf{x}_i^T \psi_j)^2 = 1, \qquad i = 1, 2, \ldots, n$$

Property #2: Proof II

Because of that $\boldsymbol{\psi}_1^T \mathbf{x}_i \propto \mathbb{1}^T \mathbf{x}_i = 0$, and that $\sum_{j=1}^{n} (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 1$,

$$\sum_{j=2}^{n} (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 1$$

Previously, $\mathbf{x}^T \mathbf{M} \mathbf{x} = \sum_i c_i^2 \mu_i$, $c_i = \boldsymbol{\psi}_i^T \mathbf{x}$, $\mathbf{x} = \sum_i c_i \boldsymbol{\psi}_i$. Here,

$$\begin{aligned}
\mathbf{x}_i^T \mathbf{L} \mathbf{x}_i &= \sum_{j=2}^{n} \lambda_j (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = \lambda_{k+1} + \sum_{j=2}^{n} (\lambda_j - \lambda_{k+1}) (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \\
&\geq \lambda_{k+1} + \sum_{j=2}^{k+1} (\lambda_j - \lambda_{k+1}) (\boldsymbol{\psi}_j^T \mathbf{x}_i)^2
\end{aligned}$$

It is tight only when $\boldsymbol{\psi}_j^T \mathbf{x}_i = 0$ for $\lambda_j \geq \lambda_{k+1}$.

Property #2: Proof III

$$\lambda_{k+1} + \sum_{j=2}^{n} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \geq \lambda_{k+1} + \sum_{j=2}^{k+1} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2$$

Quick proof of when the above inequality is tight:

$$\lambda_{k+1} + \sum_{j=2}^{n} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = \lambda_{k+1} + \sum_{j=2}^{k+1} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2$$

$$\sum_{j=k+2}^{n} (\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 0$$

That is $\boldsymbol{\psi}_j^T \mathbf{x}_i = 0$ for $j > k+1$. When $j > k+1$, $\lambda_j \geq \lambda_{k+1}$.

Property #2: Proof IV

To prove the Theorem 3.2.1, we sum up over i:

$$\begin{aligned}\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{L} \mathbf{x}_i &\geq k\lambda_{k+1} + \sum_{i=1}^{k}\sum_{j=2}^{k+1}(\lambda_j - \lambda_{k+1})(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \\ &= k\lambda_{k+1} + \sum_{j=2}^{k+1}(\lambda_j - \lambda_{k+1})\sum_{i=1}^{k}(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \\ &\geq k\lambda_{k+1} + \sum_{j=2}^{k+1}(\lambda_j - \lambda_{k+1}) = \sum_{j=2}^{k+1}\lambda_j\end{aligned}$$

because: $\lambda_j - \lambda_{k+1} \leq 0$, and, $\sum_{i=1}^{k}(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 \leq \sum_{i=1}^{n}(\boldsymbol{\psi}_j^T \mathbf{x}_i)^2 = 1$.

Conclusion

The two properties are saying that:

- Eigenvalues of graphs Laplacian can easily reveal the graph's connectivity. The amount of eigenvalue 0 is exactly the amount of independent components in a graph. For a connected graph, only $\lambda_1 = 0$, $\lambda_2 > 0$. If the graph is disconnected, $\lambda_2 = 0$. If the graph contains 3 disconnected subgraphs, $\lambda_3 = 0$. etc.
- When visualizing a graph, using its eigenvectors (ψ_1 excluded) as vertices' coordinates, will be an optimal choice.