Lecture Notes on
The Laplacian Matrix

Laurenz Wiskott & Fabian Schönfeld
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany, EU

17 December 2018 (still under development)

Contents

LECTURE 1/2

1 Introduction

2 Intuition
   2.1 Heat diffusion analogy of the Laplacian eigenmaps
   2.2 Heat diffusion analogy of spectral clustering
   2.3 Heat diffusion equation for connected heat reservoirs
   2.4 Laplacian matrix
   2.5 Solution of the heat diffusion equation

3 Formalism
   3.1 Simple graphs
   3.2 Matrix representation
   3.3 Optimization problem
   3.4 Associated eigenvalue problem
   3.5 Symmetric normalized Laplacian matrix
   3.6 Random walk normalized Laplacian matrix
   3.7 Mathematical properties

LECTURE 2/2

© 2017–2018 Laurenz Wiskott (homepage https://www.ini.rub.de/PEOPLE/wiskott/). Do not distribute these lecture notes! This version is only for the personal use of my students. Important text is set in dark red boldface; ♦ marks important formulas or items worth remembering; ◇ marks less important formulas or items, which I also discuss in the lecture; ♠ marks sections that I typically skip during my lectures. More teaching material is available at https://www.ini.rub.de/PEOPLE/wiskott/Teaching/Material/.
## 4 Algorithms

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Similarity graphs</td>
<td>14</td>
</tr>
<tr>
<td>4.2</td>
<td>Laplacian eigenmaps (LEM)</td>
<td>14</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Motivation</td>
<td>14</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Objective</td>
<td>15</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Algorithm</td>
<td>15</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Sample applications</td>
<td>15</td>
</tr>
<tr>
<td>4.3</td>
<td>Locality preserving projections (LPP)</td>
<td>17</td>
</tr>
<tr>
<td>4.3.1</td>
<td>Linear LPP</td>
<td>17</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Sample application</td>
<td>18</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Nonlinear LPP</td>
<td>18</td>
</tr>
<tr>
<td>4.4</td>
<td>Spectral clustering</td>
<td>19</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Objective</td>
<td>19</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Algorithm</td>
<td>19</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Sample application</td>
<td>20</td>
</tr>
</tbody>
</table>

### Requirements:

I assume the student can already ...

- apply basic concepts from linear algebra, such as vector, matrix, matrix product, inverse matrix.
- solve an ordinary eigenvalue equation in linear algebra and explain intuitively what eigenvalues and eigenvectors are.
- relate the eigenvalues and eigenvectors of a symmetric matrix to the solutions of the minimization/maximization problem of the corresponding quadratic form.
- interpret a system of linear differential equations with constant coefficients.

### Learning objectives:

The learning objective of this unit is that the student can ...

- define basic notions of graph theory, namely graph, node, edge, and simple graph (Sec. 3.1).
- explain matrix representations of graphs, namely adjacency matrix, degree matrix, and Laplacian matrix (Sec. 3.2).
- reproduce and interpret the generalized eigenvalue equation (43) of the Laplacian matrix and weighted degree matrix and describe how it relates to the optimization problem of Laplacian eigenmaps and spectral clustering (Eqs. 44–46).
- summarize and motivate mathematical properties (6,7,9,11) (Sec. 3.7) of the eigenvalues and eigenvectors of the generalized eigenvalue equation (43).
- explain how a similarity graph can be generated from a set of data points (Sec. 4.1).
- explain how the Laplacian eigenmaps (LEM) algorithm (Sec. 4.2) and spectral clustering (Sec. 4.4) work.
- name a limitation of LEM and sketch how locality preserving projections (LPP) overcome it (Sec. 4.3).

Please also work on the exercises and read the solutions to deepen these concepts.
LEcTUEr 1/2

1 Introduction

Many problems in machine learning can be expressed by means of a graph with nodes representing training samples and edges representing the relationship between samples in terms of similarity, temporal proximity, or label information. Graphs can in turn be represented by matrices. A special example is the Laplacian matrix, which allows us to assign each node a value that varies only little between strongly connected nodes and more between distant nodes. Such an assignment can be used to extract a useful feature representation, find a good embedding of data in a low dimensional space, or perform clustering on the original samples. In the following we first introduce the Laplacian matrix and then present a small number of algorithms designed around it.

2 Intuition

This section is meant to give an intuitive introduction into the Laplacian matrix, Laplacian eigenmaps, and spectral clustering. This section is not necessary to understand the remainder of the lecture notes but hopefully makes it easier. If you are short on time and rich in math and machine learning background, you might prefer to skip it.

The Laplacian matrix can be used to model heat diffusion in a graph. Its theory can thus be understood intuitively with the help of the heat diffusion analogy.

2.1 Heat diffusion analogy of the Laplacian eigenmaps

First consider a very simple heat diffusion analogy for nonlinear dimensionality reduction from 2D to 1D with the Laplacian eigenmap algorithm. Figure 1 (left) shows seven points in 2D, labeled A through G. Their position might not be very meaningful but we assume that we have some similarity function that induces relationships between these points. This results in a simple undirected graph with seven nodes and six edges in this example. We see already that the graph is a simple linear graph, a chain, but in high dimensions with many more nodes and a slightly more complicated structure, this might not be so obvious anymore.

The heat diffusion analogy now says that nodes are considered heat reservoirs and heat can diffuse from one node to neighboring nodes via the edges, but no heat gets lost or added. So, let us randomly initialize the nodes with arbitrary temperatures, Figure 1 (middle). What happens if we...
wait? Well, it is obvious that heat diffuses from warmer to colder nodes until temperature has balanced out completely. It is also obvious that local temperature differences balance out quickly, while global temperature differences between distant nodes (distant in terms of the graph connectivity) take more time to balance out. So if one measures the temperatures quite late in the process, one finds a distribution like the one shown in Figure 1 (right). One end of the chain is slightly warmer than the other end, and from one end to other there is a monotonic decrease of temperature. This is interesting, because if one now plots the seven points again, but now in a 1D space according to their temperature, one gets the plot in Figure 1 (bottom right). The points are nicely ordered by their position in the linear graph. This is much better for visualization and interpretation and possibly further processing of the points, since the position in space now really reflects similarity relations very well.

This is essentially how the Laplacian eigenmaps algorithm works, except that one does not really use heat diffusion but finds the resulting heat distribution analytically in a more efficient and robust way. It is also possible to map the points into a 2D or even higher-dimensional space by taking more than one heat diffusion mode into account.

2.2 Heat diffusion analogy of spectral clustering

For a heat diffusion analogy of spectral clustering consider a different connectivity of the graph, like the one shown in Figure 2 (left). The difference to the example above is that now the graph has two disconnected subgraphs. No heat can diffuse from one subgraph to the other. If one waits long enough, the temperature within each subgraph has completely balanced out, but the two subgraphs have different temperature, because there is no edge between them, Figure 2 (right). If one now plots the seven points in a 1D space according to their temperature, Figure 2 (bottom right), all points of one subgraph cluster at one value and the points of the other subgraph cluster at another value. Thus, in this space separating the two subgraphs is trivial.

This is essentially how spectral clustering works. In real data the clusters, i.e. subgraphs, might not be completely disconnected, but with some tricks one can also deal with that.

The graphs in Figures 1 and 2 are drawn in a way that the position of the nodes actually has no meaning at all. This is to emphasize that the edges are the only thing that matters for the result of Laplacian eigenmaps and spectral clustering. In real world examples, however, spatial proximity often plays an important role and edges are preferably inserted between neighboring data points.

2.3 Heat diffusion equation for connected heat reservoirs

How can we model heat diffusion mathematically, and how can we figure out the relevant temperature distributions analytically? Heat diffusion is a continuous process, so we need a differential equation (DE)
for it. Since we consider heat diffusion between a discrete set of heat reservoirs rather than on a continuous medium, the DE is a system of ordinary DEs and not a partial DE. It is linear, e.g. if you have twice as much heat, diffusion will be twice as strong. And it is homogeneous, because if there is no heat, then there is no diffusion. Thus we consider the following system of ordinary linear DEs

\[ \dot{h}(t) = -Lh(t) \]  \hfill (1)

\[ \begin{align*}
\dot{h}_1(t) & = -L_{11}h_1(t) - L_{12}h_2(t) - L_{13}h_3(t) \\
\dot{h}_2(t) & = -L_{21}h_1(t) - L_{22}h_2(t) - L_{23}h_3(t) \\
\dot{h}_3(t) & = -L_{31}h_1(t) - L_{32}h_2(t) - L_{33}h_3(t)
\end{align*} \]  \hfill (2) (3) (4)

spelled out for three heat reservoirs, where \( h(t) \) is a nonnegative vector representing the temperatures of the nodes as a function of time. \( L \) is a matrix representing the heat diffusion between the nodes, and it will be explained in a moment.

Readers not so familiar with differential equations might find it easier to consider the temporally discretized version of it,

\[ \frac{h(t+\Delta t) - h(t)}{\Delta t} = \dot{h}(t) \quad \text{(for } \Delta t \to 0) \]  \hfill (5)

\[ \begin{align*}
\frac{h(t+\Delta t) - h(t)}{\Delta t} & \overset{\text{(i)}}{=} -Lh(t) \\
\iff h(t+\Delta t) & = h(t) - \Delta t \left( Lh(t) \right) \\
& = \left( I - \Delta t L \right)h(t) \\
\iff h_1(t+\Delta t) & = h_1(t) - \Delta t(L_{11}h_1(t) + L_{12}h_2(t) + L_{13}h_3(t)) \\
\& \quad h_2(t+\Delta t) = h_2(t) - \Delta t(L_{21}h_1(t) + L_{22}h_2(t) + L_{23}h_3(t)) \\
\& \quad h_3(t+\Delta t) = h_3(t) - \Delta t(L_{31}h_1(t) + L_{32}h_2(t) + L_{33}h_3(t))
\end{align*} \]  \hfill (6) (7) (8) (9) (10) (11)

which is an approximation to the differential equation \( \dot{h}(t) = -Lh(t) \), which is exact for \( \Delta t \to 0 \).

### 2.4 Laplacian matrix

In either case, it is clear that \( L \) is responsible for any change of \( h \) and that the physics of the heat diffusion process imposes constraints on \( L \). If \( L = 0 \) then \( h(t) \) is constant, which would correspond to three disconnected nodes (= heat reservoirs) that do not exchange any heat. A negative \( L_{ij} \) indicates that \( h_i \) increases proportional to \( h_j \) with factor \(-L_{ij}\). A positive \( L_{ij} \) indicates that \( h_i \) decreases proportional to \( h_j \) with factor \(-L_{ij}\).

We want that no heat gets lost or added to the system, thus \( \sum_{k} L_{ik} = 0 \) must be fulfilled, as one can easily verify by setting \( \dot{h}_1(t) + \dot{h}_2(t) + \dot{h}_3(t) = 0 \) or \( h_1(t+\Delta t) + h_2(t+\Delta t) + h_3(t+\Delta t) = \text{const} \) for any values of \( h_1(t), h_2(t), \) and \( h_3(t) \). Since the heat one node gains must come from some other nodes, one can say that \(-L_{ij}h_j\) (with negative \( L_{ij}\)) indicates the amount of heat node \( i \) gains from node \( j \) for \( i \neq j \). The term \(-L_{jj}h_j\) (with positive \( L_{jj}\)) indicates how much heat node \( j \) looses to the other nodes.

If we consider the situation that all three nodes are connected and one node, say Node 1, is hot and the other two nodes are absolutely freezing, i.e. \( h_2 = h_3 = 0 \) (Kelvin not Celsius) then initially only \( L_{11}, L_{21}, \) and \( L_{31} \) matter. It is ituitively clear that in this situation heat diffuses from Node 1 to Nodes 2 and 3, i.e. \( h_1 \) decreases and \( h_2 \) as well as \( h_3 \) increase proportionally to \( h_1 \). This implies \( 0 < L_{11}, L_{31} < 0 \), indicating that Nodes 1 and 2 exchange heat and Nodes 2 and 3 gain heat from Node 1. If a connection would be absent, e.g. between Nodes 2 and 1, then no heat diffuses between these two nodes and the corresponding entry is zero, \( L_{21} = 0 \). If a node, let say Node 1, is not connected to any other node, then it cannot gain or loose heat at all, resulting in \( L_{11} = 0 \). Thus, by symmetry arguments we have \( 0 \leq L_{ii} \leq L_{ij} \leq 0, \forall j \neq i \).

Finally, it is clear that if two different nodes \( i \) and \( j \) have same temperature, \( h_i = h_j \), then the heat \(-L_{ij}h_j\) diffusing from node \( j \) to node \( i \) equals the heat \(-L_{ji}h_i\) diffusing from node \( i \) to node \( j \), because otherwise one node would spontaneously become warmer and the other cooler, which would allow us to build a perpetuum mobile. This implies \( L_{ij} = L_{ji} \). Please notice here that if two connected nodes have same temperature, it does not mean that no heat diffuses from one to the other, it only means that the heat flows cancel out each other.
If we summarize the insights above we find that

\[ L_{ij} = L_{ji} \quad (L \text{ is symmetric}) \quad (12) \]

\[ \sum_i L_{ij} = 0 \quad (\text{rows and columns add up to zero}) \quad (13) \]

\[ L_{ii} \geq 0 \quad (\text{diagonal elements are non-negative}) \quad (14) \]

\[ L_{ij} \leq 0 \quad \forall j \neq i \quad (\text{off-diagonal elements are non-positive}) \quad (15) \]

An example of a matrix with all these properties is

\[
L = \begin{pmatrix}
0.2 & -0.2 & 0 \\
-0.2 & 1.0 & -0.8 \\
0 & -0.8 & 0.8
\end{pmatrix}
\]  \quad (16)

The corresponding graph is shown in Figure 3.

### 2.5 Solution of the heat diffusion equation

Assume the eigenvectors \( y_\alpha \) and eigenvalues \( \gamma_\alpha \) of the Laplacian matrix are known with

\[ Ly_\alpha = \gamma_\alpha y_\alpha \quad (17) \]

ordered such that \( \gamma_1 \leq \gamma_2 \leq ... \leq \gamma_V \). It turns out that all eigenvalues are non-negative and from (13) follows directly that one can choose \( y_1 = (1, 1, ..., 1)^T \) with \( \gamma_1 = 0 \).

For the discretized version of the differential equation it is interesting to see that

\[
(8) = (I - \Delta t L) y_\alpha = I y_\alpha - \Delta t Ly_\alpha
\]

\[ =: P \]

\[ \overset{(17)}{=} y_\alpha - \Delta t \gamma_\alpha y_\alpha \quad (19) \]

\[ = (1 - \Delta t \gamma_\alpha) y_\alpha \quad (20) \]

Thus the \( y_\alpha \) are also eigenvectors of \( P \) but with eigenvalues \( \xi_\alpha \) with \( 1 = \xi_1 \geq \xi_2 \geq ... \geq \xi_V > 0 \) for small enough \( \Delta t \).

Because the Laplacian matrix is symmetric and real, the set of eigenvectors is complete, and any initial temperature vector \( h(t = 0) \) can be written as a linear combination of the eigenvectors

\[ h(t = 0) = \sum_\alpha \omega_\alpha y_\alpha \quad (21) \]

with some appropriate prefactors \( \omega_\alpha \).

From the theory of systems of homogeneous linear differential equations we know that the general solution of (1) for this \( h(t = 0) \) is

\[
\diamond \quad h(t) = \sum_\alpha \omega_\alpha \exp(-\gamma_\alpha t) y_\alpha
\]  \quad (22)

For those who prefer the discretized version of the differential equation one can show that

\[
h(t = N\Delta t) \overset{(8, 18)}{=} P^N h(0)
\]

\[ \overset{(21)}{=} P^N \sum_\alpha \omega_\alpha y_\alpha \quad (23) \]

\[ = \sum_\alpha \omega_\alpha P^N y_\alpha \quad (24) \]

\[ \overset{(20)}{=} \sum_\alpha \omega_\alpha \xi_\alpha^N y_\alpha \quad (25) \]
In either case, if one waits long enough, only the first eigenvectors with eigenvalue 0 will still contribute to \( h(t) \), and one can show that if the graph is connected only the contribution of \( y_1 \) survives indefinitely long, because \( \exp(-\gamma_1 t) = \exp(-0t) = 1 \) and \( \xi_N \rightarrow 1^N = 1 \) for any \( t \). The last eigenvector fading away is \( y_2 \), and that is exactly the vector we are interested in for the Laplacian eigenmaps algorithm, see Figure 1 (right).

If the graph is disconnected then it is intuitively clear that each subgraph balances out its heat over time, but there is no heat exchange between subgraphs. The corresponding Laplacian matrix becomes a block matrix with as many blocks on the diagonal as there are subgraphs. In the example above in Figure 2, there are two subgraphs, and because of the block structure of the Laplacian matrix and the fact that rows add up to zero, one can verify that the second eigenvector \( y_2 = (1/4, -1/3, 1/4, -1/3, 1/4, -1/3, 1/4)^T \) is constant within each subgraph and has eigenvalue \( \gamma_2 = 0 \). This again reflects the temperature distribution that remains if one waits for a long time, and that is exactly the vector we are interested in the spectral clustering algorithm, see Figure 2 (right).

In summary, the second eigenvector of the Laplacian matrix provides a nice 1D arrangement of the nodes of a similarity graph. In practice one often also uses the third and possibly the forth eigenvector to get visualizations in 2D or 3D, but that is not so easy to understand with this intuitive explanation.

3 Formalism

After the intuitive explanation we now consider Laplacian eigenmaps and spectral clustering more directly and more formally. For both algorithms data must first be represented as a graph. Nodes represent data samples and edges represent similarities between data samples. The samples could be anything, e.g. words, persons, or melodies, they need not be vectors in a vector space. We just need a non-negative function that measures similarity between two data samples. And this function does not even need to be consistent with a metric. We first introduce some notions from graph theory and then consider the optimization problem.

3.1 Simple graphs

A graph (D: Graph) \( G = (V, E) \) is a set of nodes (or vertices or points) (D: Knoten) \( V = \{v_1, ..., v_V\} \) and a set of edges (D: Kanten) \( E = \{e_1, ..., e_E\} \). An edge \( e_i \) connects two nodes \( v_j \) and \( v_k \) and is therefore defined by a pair of nodes. Edges may be directed (D: gerichtet), going from node \( v_j \) to node \( v_k \), indicated by \( e_i = (v_j, v_k) \). Edges may also be undirected (D: ungerichtet), in which case the order of the vertices does not matter and we can write \( e_i := \{v_j, v_k\} \) where the curly brackets imply that the order does not matter. Simple graphs (D: einfacher Graph, schlichter Graph) are undirected graphs without loops (D: Schleife, Schlinge), which are edges that connect a node with itself, and no parallel edges, which are edges that connect the same pair of nodes. Here we consider mainly simple graphs.


3.2 Matrix representation

Graphs can be conveniently represented by real matrices. The adjacency matrix (D: Adjazenzmatrix, Nachbarschaftsmatrix)) \( A = (A_{ij}) \) of an undirected graph is \( V \times V \) and defined as

\[
A_{ij} := \begin{cases} 
1 & \text{if } \{v_i, v_j\} \in E \\
0 & \text{otherwise}
\end{cases}
\]  

(27)

i.e. it has a one in entry \( A_{ij} \) if and only if nodes \( v_i \) and \( v_j \) are connected with each other. Matrix \( A \) is naturally symmetric, since the edges are not directed.

The degree matrix (D: Gradmatrix) \( D = (D_{ij}) \) of an undirected graph is a diagonal matrix, where the diagonal entries \( D_{ii} \) indicate the number of edges connected to node \( v_i \).
In context of the Laplacian matrix, we generalize these definitions to weighted graphs, where the edges are labeled with a real (positive) number indicating their weight \( W_{ij} \). If one simply replaces the 1 values in (27) by these weights, then \( A \) becomes the (edge) weight matrix \( W \), and the weighted degree matrix \( D = (D_{ij}) \) gets the sum over all weights of the edges converging on a node in their diagonal entries.

\[
\begin{align*}
D_{ii} := \sum_j W_{ij} &= \sum_j W_{ji} 
\end{align*}
\] (28)

\[\begin{array}{c}
\text{Figure 3: } \text{Example of a simple, weighted, undirected graph. Edges are numbered in blue, their weights are shown in black.}
\end{array}\]

Figure 3 shows a simple weighted graph. The weighted adjacency matrix, or weight matrix, of the undirected graph is

\[
W = \begin{pmatrix}
    v_1 & v_2 & v_3 \\
    v_1 & 0   & 0.2 \\
    v_2 & 0.2 & 0.8 \\
    v_3 & 0   & 0.8
\end{pmatrix}
\] (29)

The weighted degree matrix of the undirected graph is

\[
D = \begin{pmatrix}
    v_1 & v_2 & v_3 \\
    v_1 & 0   & 0   \\
    v_2 & 0   & 1.0 \\
    v_3 & 0   & 0.8
\end{pmatrix}
\] (30)

The Laplacian matrix \( L \) is defined as the difference between weighted degree matrix \( D \) and weight matrix \( W \)

\[
L = D - W.
\] (31)

It is easy to verify that it has all the properties (12–15) derived in Section 2.4 from the heat diffusion analogy. The Laplacian matrix for the example above is

\[
L = \begin{pmatrix}
    0.2 & -0.2 & 0 \\
    -0.2 & 1.0 & -0.8 \\
    0 & -0.8 & 0.8
\end{pmatrix}
\] (32)

### 3.3 Optimization problem

The objective of Laplacian eigenmaps as well as spectral clustering is to assign similar values to similar nodes, i.e. strongly connected nodes, and dissimilar values to nodes that are not similar. Please note that this is a non-trivial operation, since similarity is a property of a pair of nodes, or an edge, while value is a property of a single node. It is not guaranteed that there is a good solution at all. Consider, for instance, three nodes \( A, B, \) and \( C \). If \( A \) and \( B \) are very similar as well as \( B \) and \( C \), but \( A \) and \( C \) are very dissimilar,
then there are no values that could reflect that. However, reasonable similarity measures usually do not lead to such conflicts, definitely not those inducing a proper metric. In any case, the objective is to

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_{ij} \\
\text{subject to} & \quad y^T 1 = 0 \quad \text{(zero mean)} \\
\text{and} & \quad y^T y = 1 \quad \text{(unit variance)} \\
\text{or subject to} & \quad y^T D 1 = 0 \quad \text{(weighted zero mean)} \\
\text{and} & \quad y^T D y = 1 \quad \text{(weighted unit variance)}
\end{align*}
\]

(33)

with \( y = (y_1, y_2, \ldots, y_V)^T \) and \( 1 = (1, 1, 1, \ldots, 1)^T \) indicating the one-vector. Objective (33) favors solutions where strongly connected nodes with a large edge weight \( W_{ij} \) have similar values \( y_i \) and \( y_j \). Constraints (34) and (35) in conjunction avoid the trivial constant solution, which implicitly guarantees that nodes that are not similar get assigned dissimilar values.

### 3.4 Associated eigenvalue problem

One can show that the normalized eigenvectors \( y_\alpha \) of the ordinary eigenvalue equation

\[
L y = \gamma y
\]

ordered by increasing (decreasing) eigenvalues \( \gamma_\alpha \) solve the optimization problem

\[
\begin{align*}
\text{minimize (maximize)} & \quad y^T_\alpha Ly_\alpha = \frac{1}{2} \sum_{i,j} (y_{\alpha,i} - y_{\alpha,j})^2 W_{ij} \\
\text{subject to} & \quad y^T_\alpha y_\alpha = 1 \quad \text{(unit norm)} \\
\text{and} & \quad y^T_\alpha y_\beta = 0 \quad \forall \beta < \alpha \quad \text{(order and orthogonality)}
\end{align*}
\]

(39)

(40)

(41)

where constraint (41) induces an order such that \( y_1 \) is the optimal solution without any orthogonality constraint (only the unit norm constraint), \( y_2 \) is the optimal solution with the additional constraint of being orthogonal to \( y_1 \), \( y_3 \) is the optimal solution with the additional constraint of being orthogonal to \( y_1 \) and \( y_2 \), etc. Constraints (40, 41) can be combined as \( y^T_\alpha y_\beta = \delta_\alpha \beta \forall \beta \leq \alpha \). Identity (39) is left to the reader as an exercise. The rest should be known, for instance from principal component analysis (Wiskott, 2016b).

Since

\[
y^T_\alpha Ly_\alpha \overset{(42)}{=} y^T_\alpha \gamma_\alpha y_\alpha = \gamma_\alpha y^T_\alpha y_\alpha \overset{(40)}{=} \gamma_\alpha
\]

the eigenvalues are the optimal values of the objective function.

In the algorithms below the constraint is usually \( u^T D u = 1 \) rather than \( u^T u = 1 \). Thus we note that the appropriately normalized eigenvectors \( u_\alpha \) of the generalized eigenvalue equation

\[
L u = \lambda D u
\]

ordered by increasing (decreasing) eigenvalues \( \lambda_\alpha \) solve the optimization problem

\[
\begin{align*}
\text{minimize (maximize)} & \quad u^T_\alpha Lu_\alpha = \frac{1}{2} \sum_{i,j} (u_{\alpha,i} - u_{\alpha,j})^2 W_{ij} \\
\text{subject to} & \quad u^T_\alpha Du_\alpha = 1 \quad \text{(weighted unit norm)} \\
\text{and} & \quad u^T_\alpha D u_\beta = 0 \quad \forall \beta < \alpha \quad \text{(order and weighted orthogonality)}
\end{align*}
\]

(43)

(44)

(45)

(46)

The derivation (42) does not hold here, since the eigenvectors must have weighted unit norm, not standard unit norm. But still we find analogously

\[
u^T_\alpha Lu_\alpha \overset{(43)}{=} u^T_\alpha \lambda_\alpha D u_\alpha = \lambda_\alpha u^T_\alpha D u_\alpha \overset{(45)}{=} \lambda_\alpha
\]

(47)
Thus, the eigenvalues are the value of the objective function for the different eigenvectors, it is intuitively clear that **eigenvectors with small eigenvalue are smooth in the sense that connected nodes tend to have similar values** while eigenvectors with large eigenvalue are more rugged, i.e. connected nodes tend to have different values.

Further reading: (Wikipedia, 2017c).

### 3.5 Symmetric normalized Laplacian matrix

For the algorithms below, we will consider the eigenvalues and -vectors of the generalized eigenvalue equation \( Lu = \lambda Du \). Since most of us are more familiar with the ordinary eigenvalue equation, it is interesting to note that one can convert the generalized eigenvalue equation into an ordinary one and back again. This allows us to transfer what we know about ordinary eigenvalue equations to the generalized ones.

First assume \( D_{ii} \neq 0 \) \( \forall i \) (0 \leq D_{ii} \) is true in any case) and define

\[
\begin{align*}
\hat{d} &:= (D_{11}, ..., D_{VV})^T \\
\hat{d} &:= (\sqrt{D_{11}}, ..., \sqrt{D_{VV}})^T \\
\hat{d} &:= (1/\sqrt{D_{11}}, ..., 1/\sqrt{D_{VV}})^T \\
\hat{D} &:= \text{diag}(\hat{d}) = \hat{D}^T \\
\n\end{align*}
\]

so that, for instance, \( \hat{D}D = D\hat{D} = I \) and \( D\hat{D} = D \).

Now we convert the generalized eigenvalue equation into an ordinary one.

\[
\begin{align*}
\hat{L}u &\overset{\triangle}{=} \hat{\lambda}u \\
\hat{L}u &\Leftrightarrow 0 = (\hat{L} - \hat{\lambda}I)u \\
\hat{L}u &\Leftrightarrow 0 = (\hat{L} - \hat{\lambda}I)\hat{u} \quad \text{(since } \hat{D} \text{ is invertible)} \\
\hat{L}u &\Leftrightarrow \hat{L}\hat{u} = \hat{\lambda}\hat{u} \quad \text{(60)}
\end{align*}
\]

with

\[
\begin{align*}
\hat{u} &\overset{\triangle}{=} \hat{D}u \\
\hat{u} &\Leftrightarrow u = \hat{D}\hat{u} \quad \text{(62)}
\end{align*}
\]

and the **symmetric normalized Laplacian matrix**

\[
\hat{L} := \hat{D}\hat{L}\hat{D} \quad \text{(63)}
\]

Thus, if and only if \( u \) is an eigenvector of the generalized eigenvalue equation with eigenvalue \( \lambda \), then \( \hat{u} \) is an eigenvector of the ordinary eigenvalue equation with same eigenvalue \( \lambda \). It is sometimes helpful to switch back and forth between these two views.

For the example above we find

\[
\begin{align*}
\hat{L} = \begin{pmatrix}
\frac{1}{\sqrt{0.2}} & \frac{1}{\sqrt{1.0}} & \frac{1}{\sqrt{0.8}} \\
\downarrow & \downarrow & \downarrow \\
0.2 & -0.2 & 0 \\
-0.2 & 1.0 & -0.8 \\
0 & -0.8 & 0.8 \\
\end{pmatrix}
\end{align*}
\]

\[
= \begin{pmatrix}
1.0 & -\sqrt{0.2} & 0 \\
-\sqrt{0.2} & 1.0 & -\sqrt{0.8} \\
0 & -\sqrt{0.8} & 1.0 \\
\end{pmatrix} \quad \text{(64)}
\]
where \( \div \sqrt{\cdot} \) indicates multiplication with \( D \) from the left along the rows and from the right along the columns. It is easy to see that \( \hat{L}_{ii} = 1 \) by construction, since \( DLD = D(D - W)D = (I - DW)D \) and \( DW \) has only zeroes on the diagonal. But the rows and columns do not add up to zero anymore.

The objective function related to the eigenvalue equation of the symmetric normalized Laplacian matrix is

\[
\hat{u}_\alpha^T \hat{L} \hat{u}_\alpha = \hat{u}_\alpha^T DLD \hat{u}_\alpha = \left( \hat{D}_\alpha \right)^T L \hat{D}_\alpha \quad \text{(since } D \text{ is diagonal)}
\]

\[
= \frac{1}{2} \sum_{i,j} (\hat{D}_\alpha)_{ij}^2 W_{ij} \quad \text{(since } D \text{ is diagonal)}
\]

(63, 64)

\[
\hat{L} \hat{u}_\alpha = \lambda \hat{u}_\alpha
\]

(65)

3.6 Random walk normalized Laplacian matrix

Another possibility to convert the generalized eigenvalue equation into an ordinary one is simply to multiply (43) from the left with the inverse of the weighted degree matrix.

\[
\hat{L} \hat{u}_\alpha = \lambda \hat{u}_\alpha
\]

(66)

\[
\hat{L} \approx \frac{1}{2} \sum_{i,j} (\hat{D}_\alpha)_{ij}^2 W_{ij}
\]

(67)

\[
\hat{L} = \hat{D}^{-1} \hat{L} \hat{D} \approx \frac{1}{2} \sum_{i,j} \left( \frac{\hat{u}_\alpha,i}{\sqrt{D_{ii}}} - \frac{\hat{u}_\alpha,j}{\sqrt{D_{jj}}} \right)^2 W_{ij}
\]

(68)

\[
\hat{L} \hat{u}_\alpha = \lambda \hat{u}_\alpha
\]

(69)

\[
\hat{D}^{-1} \hat{L} \approx \hat{D}
\]

(70)

\[
\hat{L} \approx \hat{D}^{-1} \hat{L}
\]

(71)

\[\hat{L} := \hat{D}^{-1} \hat{L} \] is the random walk normalized Laplacian matrix and has the same eigenvalues and eigenvectors as the generalized eigenvalue equation of the Laplacian matrix. Its main disadvantage is that it is non-symmetric.

For the example above we find

\[
\hat{L} \approx \begin{pmatrix}
\div 0.2 & 0.2 & -0.2 & 0 \\
\div 1.0 & -0.2 & 1.0 & -0.8 \\
\div 0.8 & 0 & -0.8 & 0.8
\end{pmatrix}
\]

\[
= \begin{pmatrix}
1.0 & -1.0 & 0 \\
-0.2 & 1.0 & -0.8 \\
0 & -1.0 & 1.0
\end{pmatrix}
\]

(72)

where \( \div \) indicates multiplication with \( \hat{D}^{-1} \) from the left along the rows. Notice that \( \hat{L} \approx 1 \) and that the rows, but not the columns, add up to zero. \( P := I - \hat{L} \approx \) is a right stochastic matrix (Wikipedia, 2017e), which can be interpreted as a transition matrix for a random walk between the nodes of the graph. Therefore the name. We are not sure how useful this intuition is, since the right stochastic matrix has to be multiplied from the right, in order to simulate a random walk, but in the eigenvalue equation \( \hat{L} \approx \) is multiplied from the left.

In what follows we focus on \( \hat{L} \) rather than \( \hat{L} \approx \), because the non-symmetry makes the latter more difficult to deal with.

LECTURE 2/2

3.7 Mathematical properties

The Laplacian matrix appears in a multitude of different algorithms, three of which will be discussed in this lecture: Laplacian eigenmaps (LEM), locality preserving projections (LPP), and spectral clustering. When using the Laplacian matrix in an algorithm, we are usually interested in its eigenvectors and eigenvalues. The set of eigenvalues of a matrix is referred to as its spectrum.

The Laplacian matrix, its eigenvectors, and its spectrum have the following properties:
1. **L and \( \hat{L} \) are both symmetric (and real).** The symmetry of \( L \) follows directly from equation (31) since \( D \) is diagonal and \( W \) is symmetric. The symmetry of \( \hat{L} \) follows from equation (63) and the symmetry of \( L \). See (32) and (64) for the example above.

2. **L and \( \hat{L} \) each have a complete set of orthogonal eigenvectors \( y_\alpha \) and \( \hat{u}_\alpha \), respectively, with real eigenvalues.** This is true for any real symmetric matrix, see Property (1).

3. **L and \( \hat{L} \) are both positive semi-definite.** For \( L \) this is shown in one of the exercises; for \( \hat{L} \) this follows from equation (63) and the fact that it holds for \( L \).

4. **L and \( \hat{L} \) have only non-negative eigenvalues.** This follows from Property (3). Note, however, that the eigenvalues of \( L \) and \( \hat{L} \) may be different. We indicate the eigenvalues of \( L \) by \( \gamma_\alpha \) and those of \( \hat{L} \) by \( \lambda_\alpha \).

5. \( \hat{L}\hat{u} = \lambda \hat{u} \) and \( Lu = \lambda Du \) have the same set of eigenvalues \( \lambda_\alpha \) and their eigenvectors are related by \( u_\alpha = \hat{D}\hat{u}_\alpha \Leftrightarrow \hat{u}_\alpha = D u_\alpha \). See Section 3.5.

6. The generalized eigenvalue equation \( Lu = \lambda Du \) has only non-negative eigenvalues \( \lambda_\alpha \) and a full set of eigenvectors \( u_\alpha \) that are orthogonal wrt the inner product \( u_\alpha^T Du_\beta \) for \( \alpha \neq \beta \). This follows from Properties (2,4) with Property (5), since \( \forall \alpha \neq \beta : 0 \Leftrightarrow u_\alpha^T \hat{u}_\beta \Leftrightarrow u_\alpha^T D \hat{u}_\beta = u_\alpha^T Du_\beta \).

7. \( \mathbf{1} := (1,1,\ldots,1)^T \) (the one-vector) is a solution of the ordinary eigenvalue equation \( Ly = \gamma y \) as well as the generalized eigenvalue equation \( Lu = \lambda Du \) with eigenvalue 0. This follows directly from the definition of \( L \) (31), since its rows sum up to zero, and because the two eigenvalue equations are identical for \( \gamma = \lambda = 0 \). We chose the appropriately normalized one-vector to be the first eigenvectors \( y_1 = \mathbf{1}/\sqrt{\mathbf{1}^T \mathbf{1}} \) and \( u_1 = \mathbf{1}/\sqrt{\mathbf{1}^T \hat{D} \mathbf{1}} \) with \( \gamma_1 = \lambda_1 = 0 \).

8. \( \hat{d} \), see (49), is a solution of the ordinary eigenvalue equation \( \hat{L}\hat{u} = \lambda \hat{u} \) with eigenvalue 0. This follows from Property (7) and equation (62) since \( \hat{D}\hat{d} = \mathbf{1} \Leftrightarrow u_1 \). We chose this ‘square-root degree-vector’ normalized to norm one to be the first eigenvector \( \hat{u}_1 = \hat{d}/\sqrt{\hat{d}^T \hat{d}} \) with \( \lambda_1 = 0 \).

9. Property (7) generalizes to several eigenvalues with eigenvalue 0 for disconnected graphs, see exercise. If a graph has \( N \) subgraphs that are intrinsically connected but not mutually, then \( L \) has \( N \) orthogonal eigenvectors with eigenvalue 0. Each of these eigenvectors has identical values within each of the connected subgraphs and possibly different values between subgraphs. Since it is possible to arbitrarily rotate a set of eigenvectors with identical eigenvalue and still get a set of eigenvectors, it is possible to chose the eigenvectors with eigenvalue 0 such that each one has the value 1 within a subgraph and value 0 on all other nodes. Such vectors are referred to as indicator vectors (Wikipedia, 2016).

10. If we do not perform the rotation mentioned in Property (9) to get indicator vectors, but rather choose the first eigenvector to be the one-vector, then all higher eigenvectors of the ordinary eigenvalue equation \( Ly = \gamma y \) have zero mean, since \( \forall \alpha \neq 1 : 0 \Leftrightarrow y_\alpha^T y_1 \Leftrightarrow y_\alpha^T 1 = \sum_j y_{\alpha j} \) by Properties (2,7).

11. Similarly, if the first eigenvector is the one-vector all higher eigenvectors of the generalized eigenvalue equation \( Lu = \lambda Du \) have weighted zero mean since \( \forall \alpha \neq 1 : 0 \Leftrightarrow u_\alpha^T Du_1 \Leftrightarrow u_\alpha^T D 1 = \sum_j u_{\alpha j} D_{jj} \) by Properties (6,7).

12. **The eigenvectors are solutions to the optimization problems and the eigenvalues are the values that the objective functions assume for the optimal solutions,** see Section 3.4. Equation (42) yields \( y_\alpha^T Ly_\alpha = \gamma_\alpha \), and \( \hat{u}_\alpha^T \hat{L}\hat{u}_\alpha = \lambda_\alpha \) holds analogously. For the generalized eigenvalue equation, we find (47) \( u_\alpha^T Lu_\alpha = \lambda_\alpha \).

### Table 1: Overview over different Laplacian matrices, eigenvalue equations, optimization problems, and solutions.

<table>
<thead>
<tr>
<th>Degree matrix: $D_{ij} := \delta_{ij} \sum_j W_{ij}$</th>
<th>Laplacian matrix: $L = D - W$</th>
<th>Is symmetric: $L = L^T$</th>
<th>Sym. norm. Lapl. matrix: $\hat{L} := DLD$</th>
<th>Is symmetric: $\hat{L} = \hat{L}^T$</th>
<th>Is positive semi-definite: $x^T \hat{L} x \geq 0 \forall x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equation: $Ly = \gamma y$</td>
<td>Optimization problem: minimize $y^T Ly = \gamma y$</td>
<td>Trivial first solution: $y_1 = 1/\sqrt{1^T 1}$ with $\gamma_1 = 0$</td>
<td>Ordinary eigenvalue equation: $Ly = \gamma y$</td>
<td>Optimization problem: minimize $y^T Ly = \gamma y$</td>
<td>Objective function value: $y^T Ly = \gamma y$</td>
</tr>
<tr>
<td>Equation: $u^T Lu = \lambda u$</td>
<td>Optimization problem: minimize $u^T Lu = \lambda u$</td>
<td>Trivial first solution: $u_1 = 1/\sqrt{1^T D 1}$ with $\lambda_1 = 0$</td>
<td>Ordinary eigenvalue equation: $\hat{L} \hat{u} = \lambda \hat{u}$</td>
<td>Optimization problem: minimize $\hat{u}^T \hat{L} \hat{u} = \lambda \hat{u}$</td>
<td>Objective function value: $\hat{u}^T \hat{L} \hat{u} = \lambda \hat{u}$</td>
</tr>
<tr>
<td>Equation: $\hat{u}^T \hat{L} \hat{u} = \lambda \hat{u}$</td>
<td>Optimization problem: minimize $\hat{u}^T \hat{L} \hat{u} = \lambda \hat{u}$</td>
<td>Trivial first solution: $\hat{u}_1 = d/\sqrt{d^T d}$ with $\lambda_1 = 0$</td>
<td>Ordinary eigenvalue equation: $\hat{L} \hat{u} = \lambda \hat{u}$</td>
<td>Optimization problem: minimize $\hat{u}^T \hat{L} \hat{u} = \lambda \hat{u}$</td>
<td>Objective function value: $\hat{u}^T \hat{L} \hat{u} = \lambda \hat{u}$</td>
</tr>
</tbody>
</table>

\[
D := \text{diag}(1/\sqrt{D_{11}}, \ldots, 1/\sqrt{D_{VV}})
\]

\[
\hat{L} = DLD
\]

\[
\hat{u}_1 = d/\sqrt{d^T d}
\]

\[
\lambda_1 = 0
\]

\[
u_1 = D \hat{u}_1
\]

\[
\alpha = \lambda_1
\]

\[
u_1 = D \hat{u}_1
\]
4 Algorithms

4.1 Similarity graphs

The algorithms presented in the following are all based on the properties of the Laplacian matrix discussed above. In order to take advantage of the Laplacian matrix, though, any input data first has to be represented as a graph, commonly referred to as a similarity graph: A simple graph where the nodes represent individual data samples and edge weights denote the similarity (or distance) between two connected nodes, i.e. data samples. Appropriate similarity metrics depend on the problem and can be as simple as the Euclidean or Manhattan distance between two points.

There are different ways to construct a similarity graph, depending on the problem at hand (e.g. Belkin and Niyogi, 2003, Sec. 2; He and Niyogi, 2004, Sec. 2.2; Von Luxburg, 2007, Sec. 2). Three common methods are $\epsilon$-neighborhood, $k$-nearest neighbors, and fully connected graphs:

- $\epsilon$-neighborhood: Two nodes are connected if the distance between them is smaller than a given threshold $\epsilon$. Often $\epsilon$ is chosen so small that the distance values within an $\epsilon$-neighborhood do not carry much useful information. In this case edges are often weighted binary, i.e., with 1 or 0 depending on whether the data samples in question are close enough or not, respectively.

- $k$-nearest neighbors: Node $v_i$ is connected to $v_j$ if $v_j$ is among the $k$ nearest neighbors of $v_i$. Note that this neighborhood relation is not symmetric and yields a directed graph, thus some cleanup is required. To arrive at a simple graph instead we take each unilateral edge that has no mirrored counterpart and either remove it or keep it and set it as bilateral. Removal results in a graph where each node has at most $k$ neighbors (mutual $k$-nearest neighbor graph), while setting bilateral results in a graph where each node has at least $k$ neighbors ($k$-nearest neighbor graph). All edges are weighted by the similarity between the two nodes they connect. Binary weighting, as in the preceding method, is more dangerous here, because it cannot be guaranteed that connected nodes are close to each other.

- Fully connected: To construct a fully connected graph each data sample is simply connected to all others. In this case, using binary weights renders the graph entirely meaningless. A fully connected graph always requires weighting the edges with a similarity function (e.g. a Gaussian similarity function $s(x_i, x_j) = w_{ij} = \exp(-||x_i - x_j||^2/(2\sigma^2))$ where $\sigma$ defines the extent of local neighborhoods).

4.2 Laplacian eigenmaps (LEM)

4.2.1 Motivation

Many algorithms work only on vectorial data and are limited in the dimensionality they can process efficiently. This causes problems if one has data that is either not vectorial, such as text, or too high dimensional, such as images. If one can define a similarity function on the data, yielding a scalar similarity value for each pair of data samples, the Laplacian eigenmaps algorithm can provide a low-dimensional vectorial embedding of the data that tends to preserve similarity relationships and allows to apply other algorithms to the data that would not be applicable directly (Belkin and Niyogi, 2003). Laplacian eigenmaps are also very good for 2- or 3-dimensional visualization of data.

Example: Imagine a drone hovering through the air while equipped with a downward facing camera. Using the high dimensional pictures from its camera, we could, in theory, precisely compute the drone’s current position and elevation. Unfortunately, the space of all possible high dimensional images is effectively intractable. Luckily though, we are merely interested in a small subset of this space, namely only those images the drone’s camera can actually produce in a particular environment. And while each data point of this vastly smaller subset still is of the original, high dimensionality, it can be fully described by six dimensions alone: the position and orientation of the drone in 3D space. Laplacian eigenmaps can be used to find a low dimensional embedding of the images that still permits extracting positional and orientation information.
4.2.2 Objective

The objective of the Laplacian eigenmaps algorithm is to find a mapping from a set of \( V \) arbitrary data samples (do not need to be vectors, but there must be a similarity function) to a set of \( n \)-dimensional data points \( \{y_1, ..., y_V\} \) such that samples with high similarity preserve their proximity in the mapped space. For \( n = 1 \), i.e. mapping into only a 1-dimensional space, this objective translates into minimizing

\[
\sum_{i,j} (y_i - y_j)^2 W_{ij}
\]

where the \( y_i \) are the values assigned to the samples and \( W_{ij} \) indicates the similarity between two samples. We have already seen above how this optimization problem is solved by the second eigenvector of the Laplacian matrix, (39) or (44) depending on the constraint. Each additional eigenvector provides us with a mapping of all data samples into a new, 1-dimensional space that is orthogonal (meaning the values are uncorrelated) to those provided by the other eigenvectors. The quality of the mapping of each eigenvector is given by its associated eigenvalue, which directly relates to the actual value of sum (73). The best \( n \)-dimensional mapping is thus given by the first \( n \) eigenvectors of the Laplacian matrix with smallest eigenvalues (excluding the first one).

We now have all the required components to formulate the Laplacian eigenmaps algorithm.

4.2.3 Algorithm

Laplacian eigenmaps algorithm (Belkin and Niyogi, 2003)

1. Given a set of \( V \) data samples construct a similarity graph \( G \) according to one of the methods described in Section 4.1.
2. Construct the \( V \times V \) weight matrix \( W \), degree matrix \( D \) (28) and Laplacian matrix \( L \) (31) for \( G \).
3. Compute the \( V \) eigenvectors \( u_i \) of the generalized eigenvalue problem

\[
\lambda u = \nabla L u
\]

ordered by ascending eigenvalues.
4. An \( M \)-dimensional representation of the data sample \( i \) is now given by \( (u_{2,i}, ..., u_{M+1,i})^T \).

4.2.4 Sample applications

Figure 4 shows a toy example of dimensionality reduction of 1000 images of size \( 40 \times 40 \) with either a vertical or a horizontal bar (Belkin and Niyogi, 2002). One can clearly see how the images with the horizontal bar are separate from the images with the vertical bar. It would be interesting to see a three dimensional Laplacian eigenmap, because presumably the red and blue points would each form a square manifold representing \( x \)- and \( y \)-position. A projection onto the first two principal components is shown for comparison.

Figures 5 and 6 show an application of Laplacian eigenmaps to a set of 300 frequently used words (Belkin and Niyogi, 2003). Each word was represented by a 600-dimensional vector indicating how often any of the other words was found to the left or to the right of the considered word. Similarity was defined based on these 600-dimensional vectors. Zooming into Figure 6 shows that grammatically closely related words are grouped together.

Further reading: (Belkin and Niyogi, 2003).
Figure 4: Dimensionality reduction of 1000 40×40 images with either a vertical or a horizontal bar (plotted together but distinguished by color for illustrative purposes, the original is in greyscale). Left: Two input images superimposed, one with a horizontal bar (red) one with a vertical bar (blue). Middle: Result of LEM. Right: Result of PCA for comparison.

Figure 5: Dimensionality reduction for 300 frequently used words from their word context data.
4.3 Locality preserving projections (LPP)

4.3.1 Linear LPP

Laplacian eigenmaps have the disadvantage that they only provide values for the data used during training. There is no straightforward way to process new data. This can be changed if the nodes \( v_i \) are data points in Euclidean space \( v_i = x_i \) and the values of the vectors \( u_\alpha \) are approximated by linear functions in the data points (He and Niyogi, 2004). Since the values of the nodes are now computed with a linear function rather than assigned freely, new data can be processed by applying the same linear function. On the training data the linear function yields the values of the nodes as follows

\[
\begin{align*}
\hat{u}_{\alpha, i} &= x_i^T z_\alpha \\
\iff u_\alpha &= X^T z_\alpha \\
\text{with data } X := (x_1, x_2, ..., x_V)
\end{align*}
\]

The vectors \( z_\alpha \) are the variables to be optimized. Inserting this in (44) and the corresponding constraints (45,46) yields

\[
\begin{align*}
\text{minimize/maximize } u_\alpha^T L u_\alpha &= z_\alpha^T X L X^T z_\alpha = z_\alpha^T L' z_\alpha \\
\text{subject to } 1 = u_\alpha^T D u_\alpha &= z_\alpha^T X D X^T z_\alpha = z_\alpha^T D' z_\alpha \\
\text{and } 0 = u_\alpha^T D u_\beta &= z_\alpha^T X D X^T z_\beta = z_\alpha^T D' z_\beta \quad \forall \beta < \alpha
\end{align*}
\]

This optimization problem can again be solved through a generalized eigenvalue problem, much like the original one. Notice, however, that the eigenvalues are not necessarily identical to those of the original eigenvalue problem, because \( u_\alpha \in \mathbb{R}^V \) is not free but constrained to be a linear function in the \( x_i \in \mathbb{R}^N \). Notice also that this problem is not of the dimensionality of the number of data points \( V \) as before but
only of the dimension $N$ of the data points, which is usually much smaller and, consequently makes this approximation more computationally efficient. For instance, if you have 100 data points in 3D, the problem is 3-dimensional not 100-dimensional as for the LEM algorithm. The main advantage, however, is that new data points $x_j$ can easily be embedded in the low-dimensional space by applying the linear functions $x_j^T z_\alpha$. Performing Laplacian eigenmaps with this linear approximation is referred to as *locality preserving projections* (LPP).

### 4.3.2 Sample application

An application of LPP to face images of a single person are shown in Figure 7 (He and Niyogi, 2004). Even though the mapping is only linear, LPP still captures some prominent variations and orders the images nicely in 2D. The person looks to the left (or right) at the top (or bottom) of the plot, and it smiles on the right side while it makes faces on the left.

![Figure 7](URL)

**Figure 7:** Dimensionality reduction of face images of a single person down to two dimensions with linear LPP. Face images in the plot indicate what some points stand for and the line of faces at the bottom corresponds to the line of data points on the right.

### 4.3.3 Nonlinear LPP

LPP can be generalized to nonlinear functions by adding a nonlinear expansion (Wiskott, 2017) prior to the algorithm. Assume $f(x)$ is such a nonlinear expansion from $\mathbb{R}^N \rightarrow \mathbb{R}^P$ with $N \ll P$, then
one can define
\[ u_{\alpha,i} = f(x_i)^T z_{\alpha} \tag{81} \]
\[ \iff \quad u_{\alpha} = F^T z_{\alpha} \tag{82} \]
with \[ F := (f(x_1), f(x_2), ..., f(x_V)) \tag{83} \]
and then run the algorithm as before. Notice that now \( z_{\alpha} \in \mathbb{R}^p \) rather than \( \mathbb{R}^N \). This method can also be kernelized (Wiskott, 2007).

Further reading: (He and Niyogi, 2004).

4.4 Spectral clustering

4.4.1 Objective

Spectral clustering is an umbrella term for a number of algorithms that use the eigenvectors of the Laplacian matrix to perform clustering on a given set of data points. In particular, spectral clustering is often used in image processing to identify connected parts of a given image and, ideally, identify the extent of the individual components of an image, a process called image segmentation.

As illustrated intuitively in Figure 2 the eigenvectors of the Laplacian matrix place the nodes of connected subgraphs at the same location. This also holds for the eigenvectors of the generalized eigenvalue problem, and this also holds approximately if the subgraphs are not completely separate from each other. Given this representation it is much easier than on the original data to cluster the nodes with some standard clustering algorithm (Wiskott, 2016a).

Remember that for \( k \) intrinsically connected but mutually disconnected subgraphs, i.e. clusters, there are exactly \( k \) eigenvectors with constant values on each of the clusters. For extracting \( k \) clusters one would therefore use the first \( k \) eigenvectors.

4.4.2 Algorithm

Normalized spectral clustering algorithm (Ng et al., 2002)

1. Given a set of \( N \) data samples construct a similarity graph \( G \) according to one of the methods described in Section 4.1. For instance, when performing segmentation on a single image, each pixel becomes a node of the graph with similarity between nodes usually being a function of color and spatial distance.

2. Compute the weight matrix \( W \), degree matrix \( D \) (28), and Laplacian matrix \( L \) (31) for \( G \).

3. Compute the first \( k \) eigenvectors of the generalized eigenvalue problem
\[ Lu = \lambda Du. \tag{84} \]
ordered by ascending eigenvalue. Let matrix \( U \) contain the eigenvectors \( u_1, ..., u_n \) as its column vectors.

4. Compute matrix \( T \) from \( U \) by normalizing the row vectors to 1:
\[ T_{ij} = U_{ij} / \left( \sum_k U_{ik}^2 \right)^{1/2}. \tag{85} \]
A \( k \)-dimensional representation \( y_i \) of data sample \( i \) is now given by the \( i \)-th row vector of \( T \).

5. Perform the \( k \)-means algorithm on the set of mapped data points \( \{ y_1, ..., y_N \} \) to partition the data into \( k \) clusters.
4.4.3 Sample application

Figure 8 shows an example of applying spectral clustering to an old data set collected by Edgar Anderson (Wikipedia, 2017d). He measured length and width of the sepal (D: Kelchblatt) and petal (D: Kronblatt) from 50 exemplars of three types of iris. One species is well separated from the other two, which in turn are hard to distinguish just based on these measurements. Spectral clustering performs fairly well on this task as one can see by comparing ground truth on the left with the clustering result on the right.

![Iris Data (red=setosa,green=versicolor,blue=virginica)](image1)

Figure 8: Spectral clustering on iris (the plant, not the eye) data. Left: length and width of the sepal (D: Kelchblatt) and petal (D: Kronblatt) from 50 exemplars of three types of iris as indicated by the three colors. Right: Result of spectral clustering on the 150 four-dimensional data points.

Further reading: (Von Luxburg, 2007), an excellent tutorial on spectral clustering.

References


Notes


2Belkin & Niyogi, 2003, Neur. Comp., Fig. 4, https://pdfs.semanticscholar.org/989a/f45f8242b96cecb91d48b85620e7322e4aa7.pdf

3Belkin & Niyogi, 2003, Neur. Comp., Fig. 5, https://pdfs.semanticscholar.org/989a/f45f8242b96cecb91d48b85620e7322e4aa7.pdf

4He and Niyogi, 2004, NIPS, Fig. 3, http://papers.nips.cc/paper/2359-locality-preserving-projections.pdf

5Nicoguaro, 2016, Wikimedia, © CC BY 4.0, https://commons.wikimedia.org/wiki/File:Iris_dataset_scatterplot.svg