LAPLACIANS, GIFI AND RELATIONAL DATABASES

GEORGE MICHAILIDIS AND JAN DE LEEUW

1. **Laplacians of Graphs**

Suppose we have collected data of N objects on J categorical variables with k_j categories per variable. The data set can be represented by a bypartite graph $H = (V, E)$ with vertex set $V = V_1 \cup V_2$ and edge set E, where V_1 corresponds to the set of the object vertices and V_2 to the set of category vertices. The degree of every vertex $v \in V_1$ is J, while the vertices $v \in V_2$ have varying degrees.

Let G denote the $N \times K$ superindicator matrix ($K = \sum_{j=1}^{J} k_j$). Then, the *adjacency* matrix of the graph is given by

$$A = \begin{pmatrix} 0 & G \\ G' & 0 \end{pmatrix}$$

Let $\tilde{D}$ be the diagonal matrix with elements $\tilde{D}(v, v),\ v = 1, \dots, N, N+1, \dots, N+K$ corresponding to the degree of vertex $v \in V$. For the bypartite graph H at hand $\tilde{D}$ is given by

$$\tilde{D} = \begin{pmatrix} JI_N & 0 \\ 0 & D \end{pmatrix}$$

where $D = \text{diag}(G'G)$.

The matrix $L(H) = \tilde{D} - A$ is called the *Laplacian* matrix of graph H, also known as the *admittance* or *Kirchoff* matrix [1]. It should be noted that $L(H)$ acts naturally on the vector space $\ell^2(V(H))$. Note that $Lu = 0$, where u is a vector comprised of ones. Hence, the Laplacian for our bipartite graph is given by

$$L(H) = \begin{pmatrix} JI_N & -G \\ -G' & D \end{pmatrix}$$

We mention next some properties of the Laplacian (for proofs see [2]) for a graph H of order n.

1) If H is a (weighted) graph with all weights being non-negative, then:
   a) $L(H)$ is positive semidefinite. This follows from the expression for the inner product $< L(H)z, z >= \sum_{uv\in E}(z(u) - z(v))^2$, with $z(u), z(v) \in \ell^2(V(H))$.
   b) $L(H)$ has only real eigenvalues.
   c) Its smallest eigenvalue $\lambda_1 = 0$ with corresponding eigenvector $u = (1, 1, \ldots, 1)'$.
   d) The second smallest eigenvalue $\lambda_2 > 0$ if and only if H is connected.
2) The following bounds on the eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ have been established:
   a) $\lambda_2 \leq \frac{n}{n-1}\min\{\tilde{d}(v),\ v \in V\}$.
   b) $\frac{n}{n-1}\max\{\tilde{d}(v),\ v \in V\} \leq \lambda_n \leq \max\{\tilde{d}(v) + \tilde{d}(u),\ uv \in E\}$.

Suppose we view a graph as a kinematic system of vertices joined by elastic springs, corresponding to its edges, vibrating in the line or in the plane. Then the goal becomes to minimize the kinetic energy of the system given by

$$\min_z \sigma(z) = \sum_{uv\in E} \big(z(u) - z(v)\big)^2 = z'Lz, \tag{1.1}$$

subject to the constraint $z'u = 0$ (Z centered) and $z'z = 1$ (z orthonormal) for $z \in \mathbb{R}$. The variational characterization of the second smallest eigevalue states that λ_2 is the required minimum and z corresponds to its eigenvector. For another solution orthogonal to the first one we can take the eigenvector corresponding to the third smallest eigenvalue, and so on and so forth. But the eigenvectors of the Laplacian offer a way of embedding the graph H in the line, the plane etc.

2. **Gifi and the Laplacian**

Let $z = (x, y)$ with x an $N \times 1$ vector corresponding to the object vertices and y a $K \times 1$ vector corresponding to the category vertices. Then, (1.1) can be written for our bipartite graph H as

$$\min_{z=(x,y)} \sigma(z)z'Lz = Jx'x - 2x'Gy + y'Dy = \sum_{j=1}^{J} \text{SSQ}\ \big(x - G_j y_j\big), \tag{2.1}$$

where $y = (y_1', y_2', \ldots, y_J')'$ and G_j is the $N \times k_j$ indicator matrix of variable j. Hence, it can be seen that the Laplacian gives rise to the Gifi loss function, the only difference being in the proportionality constant $1/J$, since the Gifi loss function corresponds to the average star loss, while the Laplacian to the sum of the star losses. However, the main difference between the two approaches is the normalizations used. In the Gifi system only the scores of the object vertices are normalized, while in the Laplacian approach the scores of both the object and category vertices are normalized simultaneously (does this give rise to the spring algorithm of Eades?).

Remark 2.1. The *direct sum* of two graphs $H_1 = (V_1, E_1)$ and $H_2 = (V_2, E_2)$ ($V_1 \cap V_2 = \emptyset$) is the graph $H = (V, E)$ with $V = V_1 \cup V_2$ and $E = E_1 \cup E_2$. The adjacency matrix of H is given by $A = A_1 \bigoplus A_2$ and consequently the Laplacian by $L = L_1 \bigoplus L_2$. This observation together with (2.1) shows that multilevel homogeneity analysis deals essentially with direct sums of I bipartite graphs.

3. **Relational Databases**

A relational database is comprised of a number of tables (data matrices) that have a Gifi type data structure. Let us take an example. Suppose we have two sets of objects: advisors and students. Each set of objects is characterized by a number of attributes (variables). So, this gives rise to two bipartite graphs. However, students and advisors are also related, since each student has at least an advisor and each advisor supervises potentially many students. This indicates that there is an edge set connecting the two sets of object vertices. Let $H_1 = (V_1, E_1)$ and $H_2 = (V_2, E_2)$ be the two bipartite graphs with $V_1 \cap V_2 = \emptyset$. In this case there exists a third edge set E_{12} with edges connecting vertices in V_1 with vertices in V_2. Let G_1 be the $N_1 \times K_1$ superindicator matrix of H_1, G_2 the $N_2 \times K_2$ superindicator matrix of H_2 and B the $N_1 \times N_2$ matrix that indicates the adjacent vertices of V_1 and V_2. Thus, the adjacency matrix of the whole graph is given by

$$A = \begin{pmatrix} 0 & G_1 & B & 0 \\ G_1' & 0 & 0 & 0 \\ B' & 0 & 0 & G_2 \\ 0 & 0 & G_2' & 0 \end{pmatrix}$$

and

$$\tilde{D} = \begin{pmatrix} J_1 I_{N_1} + \text{diag}(BB') & 0 & 0 & 0 \\ 0 & D_1 & 0 & 0 \\ 0 & 0 & J_2 I_{N_2} + \text{diag}(B'B) & 0 \\ 0 & 0 & 0 & D_2 \end{pmatrix}$$

where $D_i = \text{diag}(G_i' G_i),\ i = 1, 2$. Let us concentrate for a while to the symmetric matrices BB' and $B'B$. The matrix BB' records the *co-membership* relation for the objects vertices of graph H_1; that is, indicates the number of object vertices of graph H_2 jointly shared by each pair of object vertices of graph H_1. Moreover, the diagonal elements give the degree (with respect to graph H_2) of the object vertices of graph H_1 (analogously for the matrix $B'B$). In terms of the example, the non-diagonal elements of the matrix BB' indicate the number of students each pair of advisors share together, while the diagonal elements indicate how many students each advisor supervises. We can then define the Laplacian of the relational database as $L = \tilde{D} - A$, and using the results of section 1 we can embed it in the line, the plane etc.

The next question to ask is what type of loss function does this Laplacian give rise to. Let $x_i,\ y_i,\ i=1,2$ be the scores of the object and category vertices, respectively, for the two data matrices, so that $z=(x_1',y_1',x_2',y_2')'$. Then, some algebra gives that

$$\begin{aligned} z'Lz &= J_1x_1'x_1 - 2x_1'G_1y_1 + y'1D_1y_1 \\ &+ J_2x_2'x_2 - 2x_2'G_2y_2 + y'2D_2y_2 \\ &+ x_1'\text{diag}(BB')x_1 + x_2'\text{diag}(B'B)x_2 - 2x_1'Bx_2 \\ &= \sum_{i=1}^{2}\sum_{j=1}^{J_i}\text{SSQ}\big(x_i - G_{ji}y_{ji}\big) + x_1'\text{diag}(BB')x_1 + x_2'\text{diag}(B'B)x_2 - 2x_1'Bx_2 \end{aligned} \tag{3.1}$$

Ignoring for the time being the possible normalizations that can be used let us derive the optimal scores for the object and category vertices.

For fixed x's we have

$$y_{ji} = D_{ji}^{-1}G_{ji}'x_i,\ i=1,2,\ j=1,\dots,J_i, \tag{3.2}$$

while for fixed y's we get

$$x_1 = \big(J_1I_{N_1} + \text{diag}(BB')\big)^{-1}\big(G_1y_1 + Bx_2\big) \tag{3.3}$$

$$x_2 = \big(J_2I_{N_2} + \text{diag}(B'B)\big)^{-1}\big(G_2y_2 + B'x_1\big). \tag{3.4}$$

Equation (3.2) shows that the centroid principle continues to hold for the category scores. However, the optimal object scores depend on each other, something imposed by the structure of the graph at hand.

Let us turn our attention to the Laplacian of the bipartite graph with adjacency matrix B and edge set E_{12}; that is the bipartite graph with vertex set comprised of the object vertices of graphs H_1 and H_2. Its Laplacian is given by

$$L_{\text{objects}} = \begin{pmatrix} \text{diag}(BB') & -B \\ -B' & \text{diag}(BB') \end{pmatrix}$$

A close look at L_{objects} reveals its similarity (with the exception of the minus sign on matrix B) to the Burt table of a special contingency table (comprised of 1's and 0's). The structure of this Laplacian also indicates that in (3.1) three Laplacians ($L_1,\ L_2$ and L_{objects}) are involved.

Rewriting (3.1) after substituting the optimal y_{ji}'s given by (3.2) and we get

$$z'Lz = J_1x_1'x_1 - x_1'G_1D_1^{-1}G_1'x_1 \tag{3.5}$$

$$= J_2x_2'x_2 - x_2'G_2D_2^{-1}G_2'x_2 \tag{3.6}$$

$$+ x_1'\text{diag}(BB')x_1 + x_2'\text{diag}(B'B)x_2 + x_1'Bx_2 \tag{3.7}$$

$$= x'\big(C + L_{\text{objects}}\big)x - x'GD^{-1}G'x \tag{3.8}$$

where $x = (x_1, x_2)$, and

$$C = \begin{pmatrix} J_1 I_{N_1} & 0 \\ 0 & J_2 I_{N_2} \end{pmatrix} G = \begin{pmatrix} G_1 & 0 \\ 0 & G_2 \end{pmatrix} D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

Suppose we use the following normalizations $u'x = 0$ (object scores centered) and $x'\big(C + L_{\text{objects}}\big)x = I$. Then, minimizing (3.5) is equivalent to maximizing $x'GD^{-1}G'x$ subject to the normalization constraints. But the centering constraint can be incorporated by maximizing $x'OGD^{-1}G'Ox$ subject to the constraint $x'O\big(C + L_{\text{objects}}\big)Ox$, where $O = I - uu'/u'u$ is a centering operator. But the latter problem corresponds to a generalized eigenvalue problem.

3.1. **Example:** We analyze the data given in the Appendix. Table 1 gives the adjacency matrix B matrix between 10 objects labeled $A1 - A10$ and 2 objects labeled $B1 - B10$. Tables 3 and 5 give the attribute variables a, b, c, d with 2,2,3,3 categories respectively that characterize the two sets of objects. Finally, tables 2, 4 and 6 give the diagonal elements of the matrices BB' and $B'B$ and the univariate marginals of the 4 variables, respectively. The combined object and category points plot is given in Figures 1 and 2.

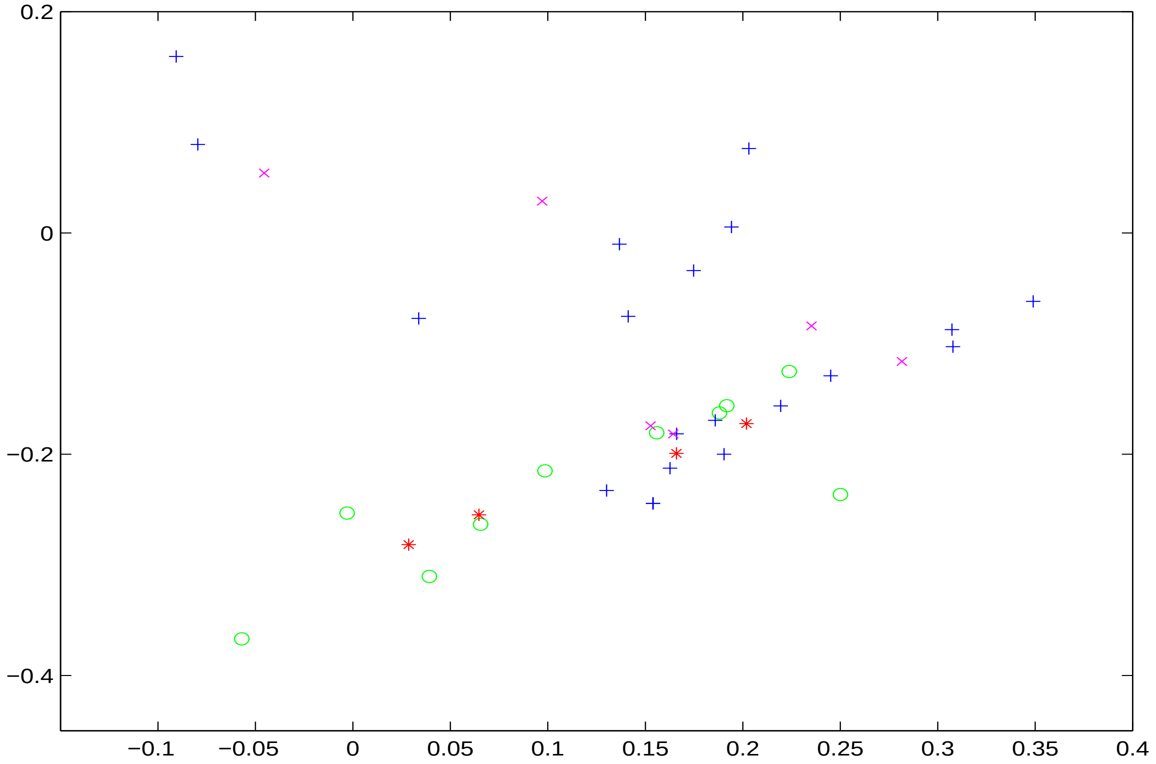

FIGURE 1. Object and category points plot (o=first set of object points, +=second set of object points, *=variable points of first set, x=variable points of second set

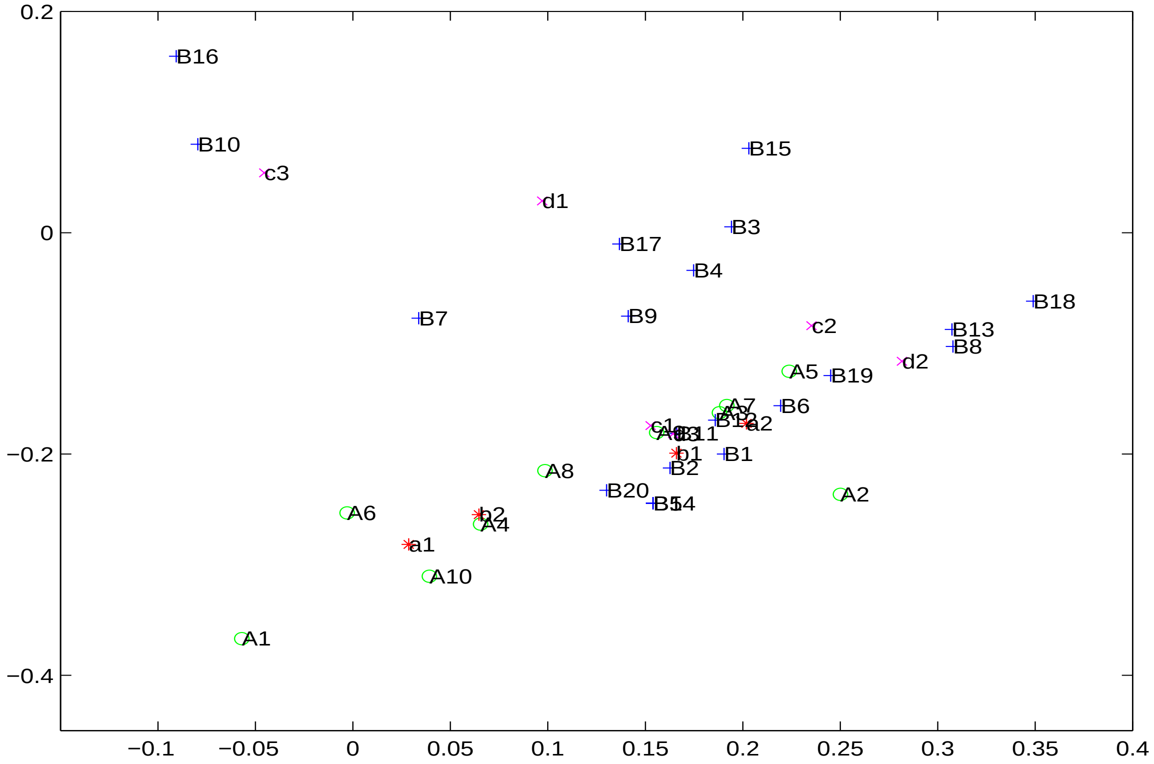

FIGURE 2. Object and category points plot (labeled)

It can be seen that the category points are the centroid of the object points the belong to that category (known from (3.2)). Also, objects that do not interact ’too much’ with other objects are in general in the periphery of the plot (e.g. objects A1, A2, B10, B16, B13, B18 etc). On the other hand, objects that interact a lot are closer together and in the center of the graph (e.g. B2, A5, A7, A3, B20). It can be seen from the graph that objects B16 and B10 belong to categories c3 and d1 (this similarity in the variables profile brings them close together), but don’t interact too much with the A-type of objects (degrees 2 and 3 respectively). On the other hand, objects A5 and A7 belong to categories a2 and b1 (so they have to be close together) and also interact with many B-type objects (in particular, both are connected to objects B1, B3, B6, B9, B19, B20). In general, this procedure tries to balance interactions between the sets of objects and similarities in their attribute profiles (which becomes clear when examining (3.3)).

4. **Extensions**

Suppose that we want to examine the bipartite graph given by (V_1, E_2), for example the advisors and the student attributes, that is what type of students characterize a certain advisor. Similarly, we can examine the bipartite graph (V_2, E_2). In this case the degree of the object vertices is not J_i anymore as was the case for the simple bipartite graphs of section 1. Moreover, there are multiple edges between an object and a category, which gives rise to a weighted graph and therefore to a weighted Laplacian. The Laplacian of the graph (V_1, E_2) is given by

$$L_1^2 = \begin{pmatrix} J_1 \text{diag}(BB') & -BG_2 \\ -(BG_2)' & D_2 \end{pmatrix}$$

while that of the graph (V_2, E_1) by

$$L_2^1 = \begin{pmatrix} J_1 \text{diag}(B'B) & -B'G_1 \\ -(B'G_1)' & D_1 \end{pmatrix}$$

It should also be noted that using the concept of the Laplacian, generalizing to the case of I data tables is a fairly straightforward exercise. Let $B_{ii'}$ denote the adjacency matrix of the set of objects i and the set of objects i'. Then, some algebra shows that we want to maximize $x'GD^{-1}G'x$, subject to the normalization constraints $u'x = 0$ and $x'\big(C + \tilde{L}\big)x$, where $C = \bigoplus_{i=1}^{I} J_i I_{N_i}$ and $\tilde{L} = \bigoplus_{i \neq i'} L_{ii'}$, and where

$$L_{ii'} = \begin{pmatrix} \text{diag}(B_{ii'}B_{ii'}') & -B_{ii'} \\ -B_{ii'}' & \text{diag}(B_{ii'}'B_{ii'}) \end{pmatrix}$$

Jan, the problem with this procedure is that the interactions between the various sets of objects are absorbed in the normalization constraint. What do you think?

5. **Appendix**

The adjacency matrix of objects is taken from a real data set. On the other hand, the attribute variables of both sets of objects are cooked up.

References

[1] Cvetkovic, D.M., Doob, M. and Sachs, H. (1995), *Spectra of Graphs*, (3rd edition), Heidelberg: Barth

[2] Mohar, B. (1991), ”The Laplacian Spectrum of Graphs,” *Graph Theory, Combinatorics and Applications*, Alavi et al. (eds), New York: Wiley

Department of Engineering-Economic Systems & Operations Research, Stanford University, Stanford, CA 94305-4023

E-mail address: gmichail@leland.stanford.edu

Interdivisional Program in Statistics, University of California, Los Angeles, Los Angeles, CA 90095-1554

E-mail address: deleeuw@stat.ucla.edu

	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
A1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
A2	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1
A3	1	0	1	1	0	1	0	0	1	0	1	0	1	0	0	0	1	1	1	1
A4	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	1
A5	1	0	1	1	0	1	0	1	1	0	1	0	1	0	1	0	1	1	1	1
A6	0	1	0	0	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	1
A7	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	0	0	0	1	1
A8	0	1	0	0	0	1	1	0	1	1	1	0	1	0	0	0	0	0	0	1
A9	0	0	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
A10	0	0	0	1	1	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1

TABLE 1. Adjacency matrix B of objects

Objects	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Degree	3	3	11	5	13	6	14	8	8	6
Objects	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
Degree	3	5	4	5	2	6	4	4	5	3
Objects	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20
Degree	8	2	3	2	2	2	2	2	3	10

TABLE 2. Degrees of objects or elements of the matrices diagBB' and diag$B'B$

	Variables	
	a	b
A1	1	2
A2	2	1
A3	2	2
A4	1	1
A5	2	1
A6	1	2
A7	2	1
A8	1	1
A9	2	2
A10	1	2

TABLE 3. Attribute variables of first set of objects

Categories	Frequencies
a1	5
a2	5
b1	5
b2	5

TABLE 4. Category frequencies of attribute variables for first set of objects

	Variables	
	c	d
B1	1	3
B2	1	2
B3	2	1
B4	2	1
B5	1	3
B6	2	3
B7	3	3
B8	2	2
B9	1	1
B10	3	1
B11	2	3
B12	2	3
B13	2	2
B14	1	3
B15	2	1
B16	3	1
B17	1	1
B18	2	2
B19	2	3
B20	1	3

TABLE 5. Attribute variables of second set of objects

Categories	Frequencies
c1	7
c2	10
c3	3
d1	7
d2	4
d3	9

TABLE 6. Category frequencies of attribute variables for second set of objects