# Graphical Spectral Features

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

# "Roughly connected" sets of data points

Another variant of principal components ideas concerns "spectral features" of a dataset built on thinking of data cases as corresponding to vertices on a **graph**. This material has emphases in common with a local version of multi-dimensional scaling treated in the notes, and can sometimes provide a way to separate "unconventional" but distinct structures of data points in $\Re^p$. The basic motivation is to not necessarily look for "convex" groups of points in $p$-space, but rather for "roughly connected"/ "contiguous" sets of points of *any* shape in $p$-space.

# Adjacencies

Begin with $N$ vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ in $\Re^p$. Consider weights $w_{ij} = w\left(\|\mathbf{x}_i - \mathbf{x}_j\|\right)$ for a decreasing function $w : [0, \infty) \longrightarrow [0, 1]$ and use them to define **similarities/adjacencies** $s_{ij}$. (For example, we might use $w(d) = \exp\left(-d^2/c\right)$ for some $c > 0$.) Similarities can be exactly $s_{ij} = w_{ij}$, but can be even more "locally" defined as follows. For fixed $k$ consider the symmetric set of index pairs

$$\mathcal{N}_k = \left\{ (i,j) \,\middle|\, \begin{array}{c} \text{the number of } j' \text{ with } w_{ij'} > w_{ij} \text{ is less than } k \\ \text{or the number of } i' \text{ with } w_{i'j} > w_{ij} \text{ is less than } k \end{array} \right\}$$

(an index pair is in the set if one of the items is in the $k$-nearest neighbor neighborhood of the other). One might then define $s_{ij} = w_{ij} I\left[(i,j) \in \mathcal{N}_k\right]$.

# Adjacency matrix and node degrees

We'll call the matrix

$$\mathbf{S} = \left(s_{ij}\right)_{\substack{i=1,\ldots,N \\ j=1,\ldots,N}}$$

the **adjacency matrix**, and use the notations

$$g_i = \sum_{j=1}^{N} s_{ij} \quad \text{and} \quad \mathbf{G} = \mathbf{diag}\left(g_1, g_2, \ldots, g_N\right)$$

It is common to think of the points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ in $\Re^p$ as nodes/vertices on a graph, with edges between nodes weighted by similarities $s_{ij}$, and the $g_i$ so-called **node degrees**, i.e. sums of weights of the edges connected to nodes $i$. In such thinking, $s_{ij} = 0$ indicates that there is no "edge" between case $i$ and case $j$.

# Graph Laplacians

The matrix

$$\mathbf{L} = \mathbf{G} - \mathbf{S}$$

is called the (unnormalized) **graph Laplacian**, and one standardized (with respect to the node degrees) version of this is

$$\tilde{\mathbf{L}} = \mathbf{G}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{G}^{-1}\mathbf{S}$$

and a second (*symmetric*) standardized version is

$$\mathbf{L}^{*} = \mathbf{G}^{-1/2}\mathbf{L}\mathbf{G}^{-1/2} = \mathbf{I} - \mathbf{G}^{-1/2}\mathbf{S}\mathbf{G}^{-1/2}$$

# Representation of a quadratic form

Note that for any vector $\mathbf{u}$,

$$\begin{aligned}
\mathbf{u}'\mathbf{L}\mathbf{u} &= \sum_{i=1}^{N} g_i u_i^2 - \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j s_{ij} \\
&= \frac{1}{2}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij} u_i^2 + \sum_{j=1}^{N}\sum_{i=1}^{N} s_{ij} u_j^2\right) - \sum_{i=1}^{N}\sum_{j=1}^{N} u_i u_j s_{ij} \\
&= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}\left(u_i - u_j\right)^2
\end{aligned}$$

so that the $N \times N$ symmetric $\mathbf{L}$ is nonnegative definite.

# Spectral/eigen analyses

Consider the spectral/eigen decomposition of $\mathbf{L}$ and focus on the *small* eigenvalues. For $\mathbf{v}_1, \ldots, \mathbf{v}_m$ eigenvectors corresponding to the 2nd through $(m+1)$st smallest non-zero eigenvalues (since $\mathbf{L1} = \mathbf{0}$ there is an uninteresting 0 eigenvalue), let

$$\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_m)$$

These are "graphical spectral features" and one might think of cases with **similar rows of V** as "alike." And small eigenvalues are associated with linear combinations of columns of $\mathbf{L}$ that are close to 0.

# Graphical Spectral Features

Why should this work to identify connected structures in a training set? For $\mathbf{v}_l$ a column of $\mathbf{V}$ that is a eigenvector of $\mathbf{L}$ corresponding to a small eigenvalue $\lambda_l$, will have

$$\lambda_l = \mathbf{v}_l' \mathbf{L} \mathbf{v}_l = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} s_{ij} \left( v_{li} - v_{lj} \right)^2 \approx 0$$

and points $\mathbf{x}_i$ and $\mathbf{x}_j$ with large adjacencies must have similar corresponding coordinates of the eigenvectors.

HTF essentially argue that the number of "0 or nearly 0" eigenvalues of $\mathbf{L}$ is indicative of the number of connected structures in the original $N$ data vectors. A series of points could be (in sequence) close to successive elements of the sequence but have very small adjacencies for points separated in the sequence. "Structures" by this methodology need NOT be "clumps" of points, but could be serpentine "chains" of points in $\Re^p$.

# A second version

A second version of this is easily built on the **symmetric normalized Laplacian**, $\mathbf{L}^*$. Its eigenvalues are nonnegative and it has a 0 eigenvalue. Let $\lambda_1^* \leq \cdots \leq \lambda_m^*$ be the 2nd through $(m+1)$st smallest eigenvalues and $\mathbf{v}_1^*, \ldots, \mathbf{v}_m^*$ be corresponding eigenvectors. Then for $\lambda_l^*$ such a small non-negative eigenvalue,

$$\lambda_l^* = \mathbf{v}_l^{*\prime}\mathbf{L}^*\mathbf{v}_l^* = \mathbf{v}_l^{*\prime}\left(\mathbf{G}^{-1/2}\mathbf{L}\mathbf{G}^{-1/2}\right)\mathbf{v}_l^* = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}\left(\frac{v_{li}^*}{\sqrt{g_i}} - \frac{v_{lj}^*}{\sqrt{g_j}}\right)^2 \approx 0$$

and points $\mathbf{x}_i$ and $\mathbf{x}_j$ with large adjacencies must have similar corresponding coordinates of the vector $\mathbf{G}^{-1/2}\mathbf{v}_l^*$. So one might treat vectors $\mathbf{G}^{-1/2}\mathbf{v}_l^*$ (or perhaps normalized versions of them) as a second version of $m$ graphical features.

# Markov chain motivation

It is also easy to see that

$$\mathbf{P} \equiv \mathbf{G}^{-1}\mathbf{S}$$

is a stochastic matrix and thus specifying an $N$-state stationary Markov Chain. It is plausible that the standardized graph Laplacian $\widetilde{\mathbf{L}} = \mathbf{I} - \mathbf{P}$ identifies groups of states such that transition by such a chain between the groups is relatively infrequent (the MC more typically moves within groups).