GRAPH DRAWING BY PULLING

JAN DE LEEUW AND GEORGE MICHAILIDIS

Contents

1. Graphs and the Adjacency Matrix

Consider a graph $G = (V, E)$, where V is the set of the n vertices and E the set of edges. The set of edges can be represented in matrix form through the *adjacency matrix* $W = \{w_{ij}\},\ i, j = 1, \ldots, n$. Thus vertices $i,\ j \in G$ are connected if and only if $w_{ij} > 0$, otherwise $w_{ij} = 0$. In case all w_{ij}'s are 0 or 1 we have a *simple* graph, otherwise a *weighted graph*. Graphs are used to model complex systems (e.g. transportation networks, VLSI layouts, molecules etc) and to visualize relationships (e.g. social networks). In Michailides and de Leeuw [1999] and de Leeuw and Michailides [1999] the problem of representing datasets with categorical variables as bipartite graphs is considered.

1.1. **Graph Drawing.** In many instances the right picture is the key to understanding. The various ways of visualizing a graph provide different insights, and often hidden relationships and interesting patterns are revealed. An increasing body of literature is considering the problem of how to draw a graph (see for instance Tammasia's book on Graph Drawing, the proceedings of the annual conference on Graph Drawing etc). Also, several problems in distance geometry and in graph theory have their origin in the problem of graph drawing in higher dimensional spaces. The problem of graph drawing is closely related to the problem of *graph embedding* in a desirable space (going back to Steinitz [1922] and definitely the pioneering work of Tutte [1963]).

During the past few years the study of graphs from a geometric perspective has been emerging. Geometric models of graphs can be classified as either (i) topological models, (ii) adjacency models, or (iii) metric models. The *topological* approach is mainly concerned with graph planarity and embeddability of graphs on other 2 dimensional manifolds [Kelmans, 1993], as well as 3 dimensional embeddings mostly in the context of knot theory [Welsh, 1993]. In an adjacency model, the geometry respects only the relation of adjacency/nonadjacency of vertices as captured by the matrix W. A prime example for this approach is the celebrated Koebe-Andreev-Thurston theorem (**?** and references therein), that every planar graph is the contact graph of openly disjoint planar discs. Higher dimensional results in the same line look at embeddings in the n dimensional Euclidean sphere (e.g. Frankl and Maehara [1988], Reiterman et al. [1989]). Orthogonal embeddings are investigated in Lovász [1979], embeddings to $n-1$ dimensional convex simplices in Lovász et al. [1989] and to lattices, and cubes in Graham and Winkler [1985] and Winkler [1983] respectively. The graph drawing community in computer science has paid particular attention to this approach due to its obvious importance in VLSI design. In the third approach graph metrics play a role. The most commonly used metric is the one where the distance between vertices i and j is given by the number of edges in the shortest path from i to j. This turns the graph G into a finite dimensional metric space, and there is a long line of research dealing with embeddings of metric spaces into normed spaces (e.g. Schoenberg [1937, 1938], Blumenthal [1978] and more recently Johnson et al. [1987], Bourgain [1985]) and the question of the minimum dimension Witsenhausen [1986].

The latter approach brings us to multidimensional scaling (MDS), a popular technique in multidimensional data analysis, where in the metric case the goal is to embed a finite number of semi-distances (known as *dissimilarities*) to a low dimensional Euclidean space (see Gower [1984, 1985]). For categorical multivariate datasets, due to their special structure, the technique of multiple correspondence analysis has also be used.

In this paper we look at the general problem of graph drawing and embedding into $\mathbb{R}^p$ by adopting primarily the adjacency model, through objective functions that measure the quality of the resulting embedding. However, we also make the necessary connections to the MDS approach and therefore to the metric approach. We study extensively certain graph structures such as bipartite and multipartite graphs, trees and more generally planar graphs, that are often encountered in multivariate data analysis.

1.2. **Pull Function.** Suppose $Z = \{z_{is}\}$ is a $n \times p$ *configuration*, which we use to make a picture of the graph in $\mathbb{R}^p$. For *optimal graph drawing* we define objective functions, that must be optimized over Z. Our first proposal is

$$\mathbf{pull}_\phi(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}\phi(d_{ij}(Z)), \tag{1}$$

where $d_{ij}(Z)$ is Euclidean distance. For the time being, we will suppose the weights w_{ij} are non-negative and ϕ is an increasing function. Thus we minimize the weighted sum of the transformed distances between the points that are connected in the graph.

Important special cases are

$$\mathbf{pull}_s(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}d_{ij}^s(Z), \tag{2}$$

with $s \geq 0$.

1.3. **Normalization.** Minimizing pull without further restrictions usually does not make much sense. We can minimize pull by collapsing all points in the origin of the space. All distances become zero.

There are two ways out of this dilemma. The first one, and the one we study in this paper, is to do a constrained optimization over *normalized* configurations, where the normalization serves the function of pushing points apart.

Consider, for example, the $\mathbf{pull_s}$ function. Let E^c denote the complement of the edge set E; that is all the pairs of vertices in G that are *not* connected. Then, the problem becomes to minimizing $\mathbf{pull_s}$ subject to a *push* constraint of the form

$$\sum_{i,j \in S} z_{ij}^s \geq |S|^{1+1/p},\ \forall S \subset E^c. \tag{3}$$

We can think of other such "spreading" constraints as well.

We can normalize by using linear inequalities etc, and solve a QP problem. We can normalize by $\mathbf{tr}\ Z'Z = 1$ (but not if we use $\mathbf{pull}_2$). We can Tutte normalize. We can use partitioned normalization, especially if the graph is bipartite.

2. Relation with Other Problems

2.1. **The Location Problem.** In his dissertation Heiser [1981] uses the location problem to introduced a wide class of scaling and unfolding techniques.

In the classical location problem, we know the coordinates of a number of *facilities* z_i in $\mathbb{R}^p$ and we try to locate a *new facility* x, such that

$$\mathbf{pull}_1(x) = \sum_{i=1}^{n} w_i d(x, z_i) \tag{4}$$

is minimized. It is not necessarily true that the distance d is Euclidean. In fact, a great deal of attention is given to non-Euclidean distances such as the City Block or ℓ_1 metric. It is also not necessarily true that this *minisum* formulation is the most natural one. In some cases *minimax* is a more direct translation of what we want to obtain. In that case

$$\mathbf{pull}_1(x) = \max_{i=1}^{n} w_i d(x, z_i) \tag{5}$$

must be minimized.

An obvious generalization is to locate more than one server in the space, but this leads to various complications, because we do not only want clients and servers to be close, we may also want servers to be relatively far apart. For instance, if we are locating m servers, we may want to minimize

$$\mathbf{pull}_{1,min}(X) = \sum_{i=1}^{n} w_i \min_{j=1}^{m} d(x_j, z_i) \tag{6}$$

This minimizes the sum of the distances of the clients to the closest server. If we are locating toilets in a campground, for instance, this seems to be the appropriate criterion. In other cases it may make sense to minimize

$$\mathbf{pull}_{1,max}(X) = \sum_{i=1}^{n} w_i \max_{j=1}^{m} d(x_j, z_i), \tag{7}$$

for instance if each client must always visit all servers. Another example, mentioned in Francis et al. [1992], is.

There is another way to define these loss functions. Let

$$\mathbf{pull}_1(X, W) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} d(x_j, z_i). \tag{8}$$

Then

$$\mathbf{pull}_{1,min}(X) = \min\{\mathbf{pull}_1(X, W) | \sum_{j=1}^{m} w_{ij} = w_i, w_{ij} \geq 0\}, \tag{9a}$$

$$\mathbf{pull}_{1,max}(X) = \max\{\mathbf{pull}_1(X, W) | \sum_{j=1}^{m} w_{ij} = w_i, w_{ij} \geq 0\}. \tag{9b}$$

In Francis et al. [1992, Chapter 6] the multifacility location problem is defined as

$$\mathbf{pull}_1(X) = \sum_{i=1}^n \sum_{j=1}^n v_{ij} d(x_i, x_j) + \sum_{i=1}^n \sum_{j=1}^m w_{ij} d(x_j, z_i) \tag{10}$$

2.2. **The QA/TSP Problem.** Heiser also discusses in this context the *quadratic assignment problem*. In this case *all* points are in fixed locations. It uses the loss function

$$\mathbf{pull}_1(\Pi) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} d_{\pi(i)\pi(j)}(Z), \tag{11}$$

2.3. **Multidimensional Scaling.**

3. Using Squared Distance

Let us first study the case in which $\phi(d) = \frac{1}{2} d_{ij}^2$. The following notation is convenient. $d_{ij}^2(Z) = \mathbf{tr}\ Z' A_{ij} Z$ with $A_{ij} = (e_i - e_j)(e_i - e_j)'$ and with the e_i unit vectors.

Define the matrix V, where

$$V = \sum_{i=1}^n \sum_{j=1}^n w_{ij} A_{ij}. \tag{12}$$

Thus V has the negative values $-w_{ij}$ as its off-diagonal elements, and the row-sums (or column-sums) of W as its diagonal elements. Then

$$\mathbf{pull}_2(Z) = \mathbf{tr}\ Z' V Z. \tag{13}$$

This must be minimized under some normalization condition on Z.

3.1. **Normalization Conditions.** The results in the previous section suggest that using the normalization condition $\mathbf{tr}\ Z'Z = 1$ is natural. Unfortunately, this does not work. For $\mathbf{pull}_2$, for instance, we find the stationary equations $VZ = \lambda Z$, which implies that all columns of Z are proportional to the eigenvector corresponding with the smallest non-zero eigenvalue of V. Thus the optimal Z is of rank one. In order to prevent this from happening, we can choose other scalar normalizations such as $\mathbf{det}\ Z'Z = 1$. Or we can choose $Z'Z = I$. These basically all result in Z being equal to the p eigenvectors corresponding to the p smallest nonzero eigenvalues of V.

3.2. **The Laplacian Connection.** Some algebra shows that $V = \frac{1}{2}L$, where $L = D - W$, with D being a diagonal matrix containing the degrees of the vertices in V. Define $\mathcal{L} = T^{-1/2}LT^{-1/2}$. This is the Laplacian of the graph G, an object of intense study over the last 20 years (starting with Fiedler in 1973). Notice that trace$\mathcal{L} = n$. In the literature the vectors $Z(:, i)$ are also known as Fiedler vectors. We discuss next some properties of $\mathcal{L}$.

1. $\lambda_1 = 0$ withe corresponding eigenvector $T^{-1/2}u$ with u comprised of all ones.
2. $\lambda_2 > 0$ iff the graph is connected.
3. $\lambda_n = 2$ iff the graph is bipartite.
4. For the complete graph K_n, the eigenvalues are 0, and $n/(n-1)$ with multiplicity $n-1$.
5. For the complete bipartite graph K_{n_1,n_2}, the eigenvalues are 0, 1 with multiplicity $n_1 + n_2 - 2$ and 2.
6. For the star graph on n vertices, the eigenvalues are 0, 1 with multiplicity $n-2$ and 2.
7. For the n dimensional hypercube on 2^n vertices, the eigenvalues are $2k/n$, with multiplicity $\binom{n}{k}$ for $k = 0, 1, \ldots, n$

In general, the second eigenvalue λ_2 provides a lot of information for the underlying graph. The larger its value is the more connected its components are and therefore the harder to split it; thus implying that clustering a dataset with a large λ_2 is hard. However, the converse is not true, since a highly connected graph with a single isolated vertex will necessarily have $\lambda_2 = 0$.

3.3. **Partitioned Normalization.** Suppose Z is partitioned as

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix} \tag{14}$$

where X is normalized in some way, and Y is free. Then

$$\mathbf{pull}_2(X, Y) = \mathbf{tr}\ X'V_{11}X + 2\mathbf{tr}\ X'V_{12}Y + \mathbf{tr}\ Y'V_{22}Y, \tag{15}$$

and thus

$$\mathbf{pull}_2(X, \star) = \min_Y \sigma(X, Y) = \mathbf{tr}\ X'\{V_{11} - V_{12}V_{22}^{-1}V_{21}\}X. \tag{16}$$

and this quadratic must still be minimized over normalized X.

An example is the normalization proposed by Tutte [1963]. In the *Tutte Normalization* X is simply fixed at some value, and the optimal Y is chosen as above. Clearly if the normalization actually fixes X, there is not need to minimize (16). The choice of which points to fix can be somewhat arbitrary, however, and it has a great deal of influence on the solution (ref to festschrift, ref perhaps to Goldstein and Healy in Biometrika).

3.4. Bipartite Graphs.

If

$$W = \begin{bmatrix} 0 & A \\ A' & 0 \end{bmatrix}, \tag{17}$$

then

$$V = \begin{bmatrix} D & -A \\ -A' & E \end{bmatrix}, \tag{18}$$

with D and E diagonal matrices with the row sums and column sums of A. Thus

$$\mathbf{pull}_2(X, Y) = \mathbf{tr}\ X'DX - 2\mathbf{tr}\ X'AY + \mathbf{tr}\ Y'EY. \tag{19}$$

Thus the minimum over X for fixed Y is

$$X = D^{-1}AY, \tag{20a}$$

and the minimum over Y for fixed X is

$$Y = E^{-1}A'X. \tag{20b}$$

These are the two *centroid principles*. Reference to chapter in Dekker Book. They are familiar from discussion of correspondence analysis, but we see that they apply much more generally.

4. Majorization

4.1. General Theory.

The algorithms proposed in this paper are all of the majorization type. In a majorization algorithm we want to maximize $\phi(\theta)$ over $\theta \in \Theta$. Suppose $\psi(\theta, \xi)$ on $\Theta \times \Theta$, which we call the *majorization function*, satisfies

$$\phi(\theta) \geq \psi(\theta, \xi) \text{ for all } \theta, \xi \in \Theta, \tag{21a}$$

$$\phi(\theta) = \psi(\theta, \theta) \text{ for all } \theta \in \Theta. \tag{21b}$$

Thus, for a fixed ξ, $\psi(\bullet, \xi)$ is below ϕ, and it touches ϕ in the point $(\xi, \phi(\xi))$.

There are two key theorems associated with these definitions.

Theorem 4.1. *If ϕ attains its maximum on Θ at $\hat{\theta}$, then $\psi(\bullet, \hat{\theta})$ also attains its maximum on Θ at $\hat{\theta}$.*

Proof. Suppose $\psi(\tilde{\theta}, \hat{\theta}) > \psi(\hat{\theta}, \hat{\theta})$ for some $\tilde{\theta} \in \Theta$. Then, by (21a) and (21b), $\phi(\tilde{\theta}) \geq \psi(\tilde{\theta}, \hat{\theta}) > \psi(\hat{\theta}, \hat{\theta}) = \phi(\hat{\theta})$, which contradicts the definition of $\hat{\theta}$ as the maximizer of ϕ on Θ. □

Theorem 4.2. *If $\tilde{\theta} \in \Theta$ and $\hat{\theta}$ maximizes $\psi(\bullet, \tilde{\theta})$ over Θ, then $\phi(\hat{\theta}) \geq \phi(\tilde{\theta})$.*

Proof. By (21a) we have $\phi(\hat{\theta}) \geq \psi(\hat{\theta}, \tilde{\theta})$. By the definition of $\hat{\theta}$ we have $\psi(\hat{\theta}, \tilde{\theta}) \geq \psi(\tilde{\theta}, \tilde{\theta})$. And by (21b) we have $\psi(\tilde{\theta}, \tilde{\theta}) = \phi(\tilde{\theta})$. Combining the three results gives the desired conclusion. □

If ϕ is bounded above on Θ, then the algorithm generates a bounded increasing sequence of function values, thus it converges. Some mild continuity considerations are needed to actually show that the sequence of θ values converges as well. See de Leeuw [1990], or for a general discussion the book by Zangwill [1969].

Example . This is an artificial example, chosen because of its simplicity. Consider $\phi(\omega) = \omega^4 - 10\omega^2$. Because $\omega^2 \geq \xi^2 + 2\xi(\omega - \xi) = 2\xi\omega - \xi^2$ we see that $\psi(\omega, \xi) = \omega^4 - 20\xi\omega + 10\xi^2$ is a suitable majorization function. The majorization algorithm is $\omega^+ = \sqrt[3]{5\omega}$.

The algorithm is illustrated in Figure 4.1. We start with $\omega^{(0)} = 5$. Then $\psi(\omega, 5)$ is the dashed function. It is minimized at $\omega^{(1)} \approx 2.92$, where $\psi(\omega^{(1)}, 5) \approx 30.70$, and $\phi(\omega^{(1)}) \approx -12.56$. We then majorize by using the dotted function $\psi(\omega, \omega^{(1)})$, which has its minimum at about 2.44, equal to about -21.79. The corresponding value of ϕ at this point is about -24.1. Thus we are rapidly getting close to the local minimum at $\sqrt{5}$, with value 25. The linear convergence rate at this point is $\frac{1}{3}$.

Theorem 4.3. *This implies that $\mathcal{D}_2(\omega, \omega) = 0$ for all ω, and consequently $\mathcal{D}_{12} = -\mathcal{D}_{22}$. Thus $\mathcal{M} = -\mathcal{D}_{11}^{-1}\mathcal{D}_{12}$.*

4.2. Weiszfeld's Algorithm for the Weber Probkem.

4.3. Majorizing the minimum of a number of convex functions.

Suppose f_s are convex, with $s = 1, \ldots, p$. Define

$$g_0 = \sum_{s=1}^{p} \lambda_s f_s, \lambda_s \in \mathbb{R}$$

$$g_r = \sum_{s=1}^{p} f_s - f_r,$$

and also

$$h_0 = \min_{s=1}^{p} f_s,$$

$$h_1 = \max_{s=1}^{p} g_s.$$

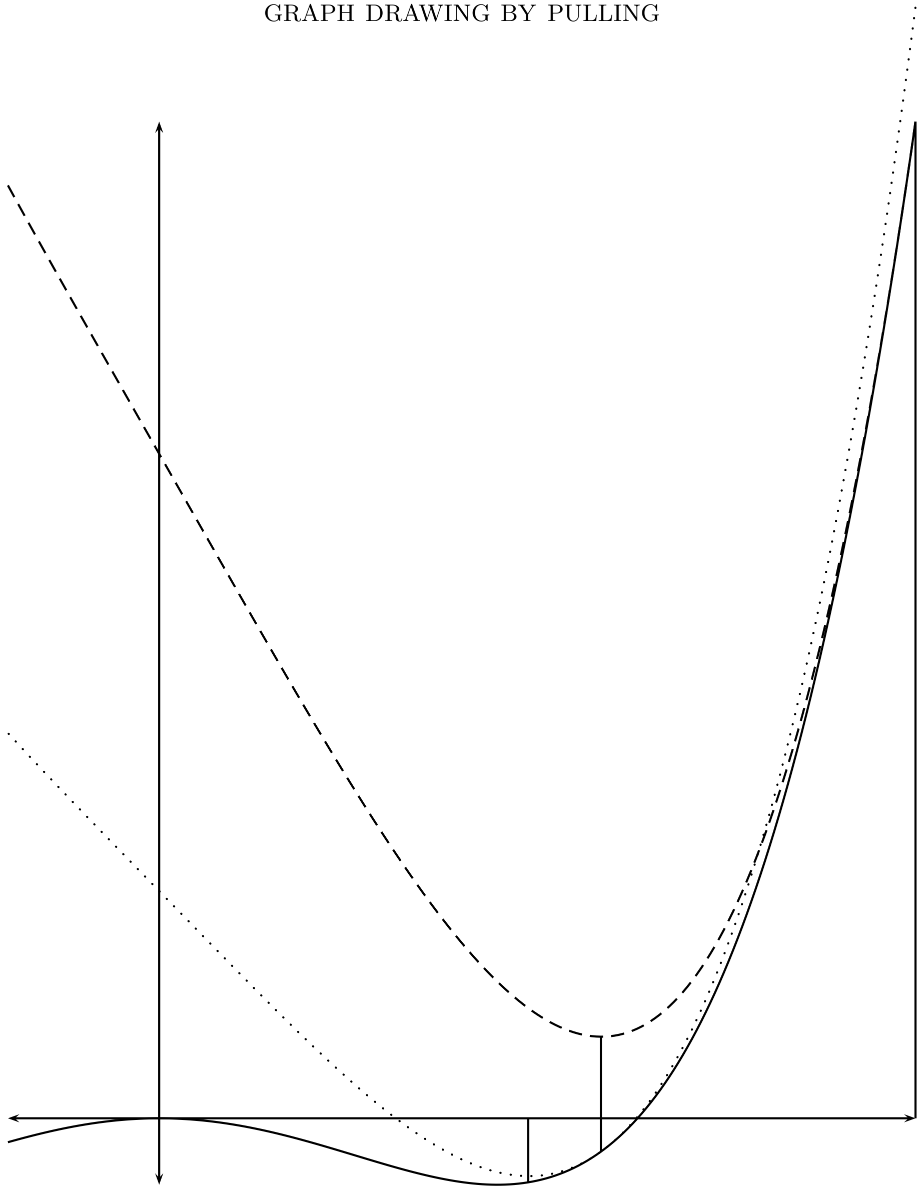

Figure 1. Majorization

Clearly, all off $g_0, \ldots, g_p$ are convex, and so is h_1. Also, trivially,

$$h_0 = g_0 - h_1.$$

This represents the minimum of the f_s as a d.c. (difference of convex functions) function. Thus we can majorize h_0 by a convex function, using

$$h_0(x) \leq g_0(x) - h_1(y) - \mathcal{D}h_1(y)(x-y) = \\ g_0(x) - h_1(y) - \mathcal{D}g_{r(y)}(y)(x-y),$$

where $r(y)$ is such that

$$g_{r(y)}(y) = h_1(y).$$

This result can be applied in a straightforward way to the multifacility location problem, which is minimization of

$$\sigma(x_1, \ldots, x_p) = \sum_{i=1}^{n} w_i \min_{s=1}^{p} \|z_i - x_s\|,$$

where the z_i are existing facilities and the x_s are new facilities (to be located).

Minimizing differences of convex functions constitutes an important subset of the field of global optimization. The class of d.c functions is very rich as the next result shows:

Lemma 4.4. *Let $f_s,\ s = 0, \ldots, p$ be d.c. functions. Then, the following are also d.c. functions.*

1. $g_0 = \sum_{s=1}^{p} \lambda_s f_s, \lambda_s \in \mathbb{R}$,
2. $h_0 = \min_{s=1}^{p} f_s$,
3. $h_1 = \max_{s=1}^{p} f_s$,
4. $|f_0|,\ f_0^+ = \max\{0,\ f_0\},\ f_0^- = \min\{0,\ f_0\}$.

Proof:

1. It is a straightforward consequence of basic properties of convex and concave functions.
2. Write $f_s = \alpha_s - \beta_s,\ s = 1, \ldots, p$. Then,

$$h_0 = \max_{s=1}^{p}\{\alpha_s + \sum_{q=1, q\neq s}^{p} \beta_q\} - \sum_{q=1}^{p} \beta_q, \tag{22}$$

   which is a d.c decomposition, since the sum and the maximum of finitely many convex functions are convex. Similar steps, establish the result for h_1.

3. $|f_0|$ can be written as $|f_0| = 2\max\{\alpha_0,\ \beta_0\} - (\alpha_0 + \beta_0)$, which is a d.c decomposition. Similarly, $f_0^+ = \max\{\alpha_0,\ \beta_0\} - \beta_0$ and $f_0^- = \alpha_0 - \max\{\alpha_0,\ \beta_0\}$, are d.c decompositions of f_0^+ and f_0^- respectively.

■

A d.c function $f:\ \mathbb{R}^n \rightarrow \mathbb{R}$ is called *locally* d.c if for every $x_0 \in \mathbb{R}^n$, there exists a neighborhood $N = N(x_0, \epsilon)$ of x_0 and convex functions

α_n, β_N such that $f(x) = \alpha_N(x) - \beta_N(x)$ for all $x \in N$. An important theorem proved in **?** states that every locally d.c function is d.c. This allows the characterization of d.c functions through their partial derivatives.

Lemma 4.5. *Every function f with continuous second partial derivatives everywhere is d.c.*

Proof: Since f has continuous second partial derivatives everywhere, the Hessian $\mathcal{D}^2 f$ is bounded on the closed neighborood $N(x_0, \epsilon) = \{x \in \mathbb{R}^n : ||x - x_0||_2 < \epsilon\}$, $\epsilon > 0$ and all $x_0 \in \mathbb{R}^n$. Then, there exist a real number $\delta > 0$ such that $f(x) + \delta||x||_2^2$ is convex on $N(x_), \epsilon)$. So, f has a d.c decomposition on $N(x_0, \epsilon)$, namely $f(x) = (f(x) + \delta||x||_2^2) - \delta||x||_2^2$ and hence due to the theorem is d.c. ∎

5. Pulling in General

We examine next three classes of functions $\phi(Z)$ of particular interest to use.

5.1. Concave Functions.

Suppose ϕ is concave. Often, we have in addition that it is increasing and passes through the origin, but these last two conditions are not really necessary for our development. Examples are $f(x) = x/(1+x)$, $f(x) = x^s$, $s \in (0, 1]$, the logistic function $f(x) = e^x/(1 + e^x)$, the logarithm $f(x) = \log x$, etc.

Because $\phi(x)$ is a concave function the following inequality holds.

$$\phi(d_{ij}(Z)) \leq \phi(d_{ij}(Y)) + \phi'(d_{ij}(Y))(d_{ij}(Z) - d_{ij}(Y)). \tag{23}$$

Hence, we iteratively minimize

$$\mathbf{pull}_\phi(Z) \leq \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\phi'(d_{ij}(Y))d_{ij}(Z), \tag{24}$$

which is a linear function in Z. But this corresponds to $\phi(x) = x$, which is treated in the next subsection. If we consider squared distances instead we get

$$\mathbf{pull}_\phi(Z) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\phi'(d_{ij}^2(Y))d_{ij}^2(Z), \tag{25}$$

which is a quadratic function in Z, and so in one iteration we minimize $Z'B(Y)Z$, where

$$B(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_{ij}}{\phi'(d_{ij}^2(Y))} A_{ij}. \tag{26}$$

Lemma 5.1. *Suppose $\phi(x)$ is concave and increasing. Let $\psi(x) = \phi(\sqrt{x})$. Then $\psi(x)$ is concave and increasing.*

Proof. By definition $\psi(\lambda x + (1-\lambda)y) = \phi(\sqrt{\lambda x + (1-\lambda)y})$. But the square root in concave, and thus $\sqrt{\lambda x + (1-\lambda)y} \geq \lambda\sqrt{x} + (1-\lambda)\sqrt{y}$, and because ϕ is increasing $\psi(\lambda x + (1-\lambda)y) \geq \phi(\lambda\sqrt{x} + (1-\lambda)\sqrt{y})$. Finally, because ϕ is concave, $\psi(\lambda x + (1-\lambda)y) \geq \lambda\psi(x) + (1-\lambda)\psi(y)$. □

This means that if ϕ is a squasher, and thus

$$\mathbf{pull}_2(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}\phi(d_{ij}^2(Z))$$

is concave with a linear majorizer, then ψ is a squasher too, while

$$\mathbf{pull}_1(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}\psi(d_{ij}^2(Z)) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}\phi(d_{ij}(Z))$$

is concave too with a linear majorizer. If not increasing

5.2. **Functions with bounded second derivatives.** This class of functions satisfies $\phi'(x) \geq 0$ and $0 \geq \phi''(x) \leq M$ for all $x \geq 0$, and for some $M \geq 0$, so we have increasing functions with bounded second derivatives. Examples of such functions include Huber's and the bi-weight functions Verboon [1994]. Under these conditions $\phi(x)$ satisfies

$$\phi(d_{ij}(Z)) \leq \phi(d_{ij}(Y)) + \phi'(d_{ij}(Y))(d_{ij}(Z) - d_{ij}(Y)) + \frac{1}{2}M(d_{ij}(Z) - d_{ij}(Y))^2, \tag{27}$$

which after defining $\eta(d_{ij}(Y)) = d_{ij}(Z) - M^{-1}\phi'(d_{ij}(Y))$ can be written as

$$\phi(d_{ij}(Z)) \leq \phi(d_{ij}(Y)) - \frac{1}{2}M^{-1}\big(\phi'(d_{ij}(Y)\big) + \frac{1}{2}\big(d_{ij}(Z) - \eta(d_{ij}(Y))\big)^2, \tag{28}$$

which shows that we are dealing with a quadratic majorizing function. In this case we can follow similar steps as in the SMACOF algorithm Suppose $X_1, \ldots, X_m$ are disjoint subsets of $\mathbb{R}^p$, and $f_i : X_i \Rightarrow \mathbb{R}$ are real valued functions. Define

$$f(x) = \sum_{i=1}^m \delta_i(x) f_i(x),$$

where $\delta_i(x)$ is the indicator of X_i.

Theorem 5.2. *Suppose $g_i(x,y)$ majorizes $f_i(x)$ on X_i, i.e.*

$$f_i(x) \leq g_i(x,y) \text{ for all } x \in X_i \text{ and all } y \in \mathbb{R}^p,$$
$$f_i(x) = g_i(x,x) \text{ for all } x \in X_i.$$

Then

$$g(x,y) = \sum_{i=1}^m \delta_i(x) g_i(x,y)$$

majorizes $f(x)$ on $X = \cup_{j=1}^m X_i$.

Proof. Suppose $x \in X_i$. Then $f(x) = f_i(x) \leq g_i(x,y) = g(x,y)$ for all $y \in \mathbb{R}^p$. Also $f(x) = f_i(x) = g_i(x,x) = g(x,x)$. □

5.3. **Convex Functions with Slow Growth Rates.**

Lemma 5.3. *Let $h : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be continuous and strictly increasing with $h(0) = 0$, let k be the inverse of h, and define $H(x) = \int_0^x h(y)dy$ and $K(x) = \int_0^x k(y)dy$. Then, for all $a, b \in \mathbb{R}_+$*

$$ab \leq H(a) + K(b),$$

with equality if, and only if, $b = h(a)$.

Proof. This is it.

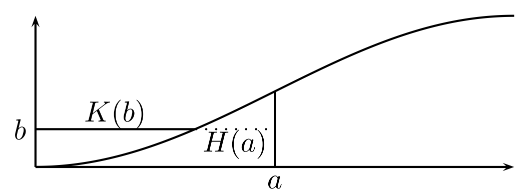

FIGURE 2. Proof of Young's Inequality

□

We are interested in minimizing

$$\textbf{pull}_\phi(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} \phi(d_{ij}(Z)). \tag{29}$$

or

$$\textbf{pull}_\phi(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} \phi(d_{ij}^2(Z)). \tag{30}$$

In this case a straightforward application of Young's inequality shows that

$$\phi(d_{ij}(Z)) = d_{ij}^s(Z) \leq \frac{2-s}{2} d_{ij}^s(Y) + \frac{2}{s d_{ij}^{2-s}(Y)} d_{ij}^2(Z), \tag{31}$$

which implies that we can construct a quadratic majorizing function. We have to minimize

$$\mathbf{pull}_\phi(Z) \leq \sum_{i=1}^n \sum_{j=1}^n w_{ij}\big(\frac{2-s}{2} d_{ij}^s(Y) + \frac{2}{s d_{ij}^{2-s}(Y)} d_{ij}^2(Z)\big), \tag{32}$$

which shows that in an interation we minimize $\mathbf{tr} Z'B(Y)Z$ over normalized Z, where

$$B(Y) = \sum_{i=1}^n \sum_{j=1}^n \frac{w_{ij}}{d_{ij}^{2-s}(Y)} A_{ij}. \tag{33}$$

6. Applications of Majorization

6.1. **Distance without the square.** A particularly interesting pull function is

$$\mathbf{pull}_1(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} d_{ij}(Z). \tag{34}$$

This minimization problem can be easily solved by using majorization based on the AM/GM inequality.

$$d_{ij}(Z) \leq \frac{1}{2} \frac{1}{d_{ij}(Y)} (d_{ij}^2(Z) + d_{ij}^2(Y)), \tag{35}$$

and thus

$$\mathbf{pull}_1(Z) \leq \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{w_{ij}}{d_{ij}(Y)} (d_{ij}^2(Z) + d_{ij}^2(Y)) = \frac{1}{2}\{\mathbf{tr}\ Z'B(Y)Z + \mathbf{tr}\ Y'B(Y)Y\},$$

where

$$B(Y) = \sum_{i=1}^n \sum_{j=1}^n \frac{w_{ij}}{d_{ij}(Y)} A_{ij}. \tag{36}$$

Thus in an iteration we minimize $\mathbf{tr}\ Z'B(Z^{previous})Z$ over normalized Z.

6.2. **Logarithm of Distance.** Suppose

$$\mathbf{pull}_{log}(Z) = \sum_{i=1}^n \sum_{j=1}^n w_{ij} \log d_{ij}(Z)$$

Again we minimize this by using majorization. A first possibility is to use

$$\log d_{ij}(Z) \leq \log d_{ij}(Y) + \frac{1}{d_{ij}(Y)}(d_{ij}(Z) - d_{ij}(Y)).$$

This implies that we iteratively minimize

$$\mathbf{pull}_1(Z) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{ij}}{d_{ij}(Y)}d_{ij}(Z).$$

by the methods explained for $\mathbf{pull}_1$.

It may be more convenient to use

$$\log d_{ij}^2(Z) \leq \log d_{ij}^2(Y) + \frac{1}{d_{ij}^2(Y)}(d_{ij}^2(Z) - d_{ij}^2(Y)),$$

which we can also write as

$$\log d_{ij}(Z) \leq \log d_{ij}(Y) + \frac{1}{2d_{ij}^2(Y)}(d_{ij}^2(Z) - d_{ij}^2(Y)).$$

This amounts to minimizing

$$\mathbf{pull}_2(Z) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{ij}}{d_{ij}^2(Y)}d_{ij}^2(Z),$$

in each iteration and this is a quadratic problem of the form $\mathbf{tr}\ Z'H(Y)Z$, where

$$H(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{w_{ij}}{d_{ij}^2(Y)}A_{ij}.$$

Observe that using this second majorization gives a less precise approximation than the first, and consequently may lead to slower convergence.

7. Examples

Not sure what to put in.

References

L.M. Blumenthal. *Theory and Applications of Distance Geometry.* Chelsea, New York, 1978. Second edition.

I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling.* Springer-Verlag, 1997.

J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

J. de Leeuw. Multivariate analysis with optimal scaling. In S. Das Gupta and J. Sethuraman, editors, *Progress in Multivariate Analysis*, Calcutta, India, 1990. Indian Statistical Institute.

J. de Leeuw and W. J. Heiser. Theory of multidimensional scaling. In P.R. Krishnaiah, editor, *Handbook of statistics, volume II*. North Holland Publishing Company, Amsterdam, The Netherlands, 1980.

J. de Leeuw and G. Michailides. Graph layout techniques and multidimensional data analysis. In T. Bruss and L. Lecam, editors, *Festschrift for Thomas S. Ferguson*. IMS, 1999.

R.L. Francis, L.F. McGinnis Jr., and J.A. White. *Facility Layout and Location: An Analytical Approach*. Prentice Hall, 1992. Second Edition.

P. Frankl and H. Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44:355–362, 1988.

J.C. Gower. Distance matrices and their euclidean approximation. In E. Diday et al., editor, *Data Analysis and Informatics, III*, Amsterdam, The Netherlands, 1984. North Holland Publishing Company.

J.C. Gower. Properties of euclidean and non-euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.

R.L. Graham and P.M. Winkler. On isometric embeddings of graphs. *Transactions of the American Mathematical Society*, 288:527–536, 1985.

W.J. Heiser. *Unfolding Analysis of Proximity Data*. PhD thesis, University of Leiden, 1981.

W.B. Johnson, J. Lindenstrauss, and G. Schechtman. On lipschitz embedding of finite metric spaces in low dimensional normed spaces. In J. Lindenstrauss and J. Milman, editors, *Geometrical Aspects of Functional Analysis*. Springer, 1987. Lecture Notes in Mathematics 1267.

A.K. Kelmans. Graph planarity and related topics. *Contemporary Mathematics*, 147:635–667, 1993.

L. Lovász. On the shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25:1–7, 1979.

L. Lovász, M. Saks, and A. Schrijver. Orthogonal representations and connectivity of graphs. *Linear Algebra and its Applications*, 114-115: 439–454, 1989.

G. Michailides and J. de Leeuw. The Gifi system for nonlinear multivariate analysis. *Statistical Science*, 13:307–336, 1999.

J. Reiterman, V. Rodl, and E. Sinajova. Geometrical embeddings of graphs. *Discrete Mathematics*, 1989.

I.J. Schoenberg. On certain metric spaces arising from euclidean spaces by a change of metric and their imbedding in hilbert space. *Annals of Mathematics*, 38:787–793, 1937.

I.J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44:522–536, 1938.

E. Steinitz. Polyheder unad raumeinteilungen. *Enzyklopedie der Mathematische Wissenschaften*, 3:1–139, 1922.

W.T. Tutte. How to draw a graph. *Proceedings of the London Mathematical Society*, 13:743–767, 1963.

P. Verboon. *A Robust Approach to Nonlinear Multivariate Analysis.* PhD thesis, University of Leiden, 1994. Also published by DSWO Press.

D.J.A. Welsh. *Complexity: Knots, Colourings and Counting.* Cambridge University press, 1993. LMS Lecture Note Series 186.

P.M. Winkler. Proof of the squashed cube conjecture. *Combinatorica*, 3:135–139, 1983.

H.S. Witsenhausen. Minimum dimension embedding of finite metric spaces. *Journal of Combinatorial Theory, Series A*, 42:184–199, 1986.

W.I. Zangwill. *Nonlinear Programming. A Unified Approach.* Prentice-Hall, 1969.