Nonlinear Mapping and Distance Geometry

Alain Franc, Pierre Blanchard, Olivier Coulaud

Nonlinear Mapping and Distance Geometry

Alain Franc, Pierre Blanchard , Olivier Coulaud

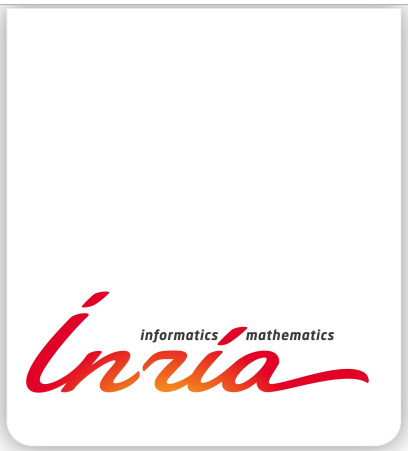
RESEARCH REPORT

N° 9210

October 2018

Project-Teams HiePACS & Pleiade

ISSN 0249-6399 ISRN INRIA/RR--9210--FR+ENG



Nonlinear Mapping and Distance Geometry

Alain Franc *†, Pierre Blanchard * ‡, Olivier Coulaud ‡

Project-Teams HiePACS & Pleiade

Research Report n° 9210 — October 2018 — 14 pages

Abstract: Distance Geometry Problem (DGP) and Nonlinear Mapping (NLM) are two well established questions: Distance Geometry Problem is about finding a Euclidean realization of an incomplete set of distances in a Euclidean space, whereas Nonlinear Mapping is a weighted Least Square Scaling (LSS) method. We show how all these methods (LSS, NLM, DGP) can be assembled in a common framework, being each identified as an instance of an optimization problem with a choice of a weight matrix. We study the continuity between the solutions (which are point clouds) when the weight matrix varies, and the compactness of the set of solutions (after centering). We finally study a numerical example, showing that solving the optimization problem is far from being simple and that the numerical solution for a given procedure may be trapped in a local minimum.

Key-words: Distance Geometry; Nonlinear Mapping, Discrete Metric Space, Least Square Scaling; optimization

* BIOGECO, INRA, Univ. Bordeaux, 33610 Cestas, France
† Pleiade team - INRIA Bordeaux-Sud-Ouest, France
‡ HiePACS team, Inria Bordeaux-Sud-Ouest, France

RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Géométrie sur les distances et meilleure image euclidienne avec distances pondérées

Résumé : Les domaines de géométrie sur les distances (distance geometry) et de recherche de meilleure image euclidienne avec distances pondérées (nonlinear mapping) sont deux domaines classiques : il s'agit pour le premier de construire une isométrie d'un espace métrique discret vers un nuage de points dans un espace euclidien, ne connaissant qu'une partie des distances, et pour le second de construire un nuage avec la meilleure approximation des distances, avec pondération. Nous montrons comment ces méthodes peuvent être rassemblée en une même famille, chacune représentant un choix de pondérations dans un problème d'optimisation. On étudie la continuité entre ces solutions (qui sont des nuages de points), et la compacité des ensembles de solutions (après centrage). On étudie également un exemple numérique, montrant cependant que le problème d'optimisation est loin d'être simple, et que la procédure d'optimisation peut facilement être piégée dans un minimum local.

Mots-clés : Géométrie sur les distances; Espaces métriques discrets; meilleure image euclidienne avec distances pondérées; optimisation

Contents

1 Introduction

Let us have a set V of n objects, and distances between every pair of them. The distance between object i and j is denoted $d(i,j)$, or d_{ij}. This defines a metric space (V,d) with $V = [\![1,n]\!]$. Among metric spaces, Euclidean spaces play a special role, because they establish a link with geometry. Moreover, the geometry of Euclidean spaces is well understood. Hence, even if many other metric spaces exist, like Riemanian manifolds with distances along a geodesic or graphs with shortest distance between vertices, many efforts have been devoted to specify those metric spaces for which an isometry exists with a Euclidean space, and if such an isometry exists, to build it. In such a case, any metric problem in V can be translated into a problem in Euclidean geometry, and, when lucky, solved. For example, supervised learning by discriminant analysis in a discrete metric space can be translated into the same problem in a Euclidean space and solved by Support Vector Machine approaches (see [2]).

The conditions for existence of an isometry between a discrete metric space and a subset of a Euclidean space are known, and are given by classical multidimensional scaling, proposed in [18] (see [1, 12] for classical presentation of MDS, and [5] for a recent presentation). If there is an isometry $i : i \longmapsto x_i$ between (V,d) and a subset of n points in a Euclidean space $\mathbb{R}^k$, then the Gram matrix of vectors $(x_i)_i$ is definite positive. It is known that the Gram matrix can be computed from pairwise distances only. A set of points $X = (x_i)_i$ such that

$$\forall\, i,j \in V, \quad \|x_i - x_j\| = d_{ij} \tag{1}$$

can be computed from the Singular Value Decomposition of the Gram matrix as a second step. If the Gram matrix has non positive eigenvalues*, there is no isometry on any subset of ℓ^2, regardless of the dimension. In such a case, for a given dimension k, one defines the cost of a map $i \longmapsto x_i$ by

$$\phi = \sum_{i<j} \left(\|x_i - x_j\| - d_{ij}\right)^2 \tag{2}$$

*In such a case, strictly speaking, the matrix built from the pairwise distances is not a Gram matrix.

If there is an isometry between (V, d) and a subset of $\mathbb{R}^k$, then there is a map for which $\phi = 0$. If not, Least Square Scaling (LSS) finds a map with minimal cost for a given dimension k by solving

Given	a discrete metric space (V, d)
	a dimension k
find	a map $i \in V \longmapsto x_i \in \mathbb{R}^k$
such that	ϕ is minimal

Least square scaling has been pioneered in [7]. See as well [1] for a presentation and a comparison with classical MDS[†].

In some situations, one is interested in relative error/distorsion between the distances d_{ij} and $\|x_i - x_j\|$. Therefore, Sammon has developed in [15] what he called Non Linear Mapping (NLM) in which each term in the cost function is weighted by the inverse of the distance:

$$\phi = \sum_{i<j} \frac{\left(\|x_i - x_j\| - d_{ij}\right)^2}{d_{ij}} \tag{3}$$

(Sammon introduces a normalizing constant c, that we do not mention here). It is natural to extend this towards

$$\phi = \sum_{i<j} \omega(d_{ij}) \left(\|x_i - x_j\| - d_{ij}\right)^2 \tag{4}$$

where $\omega \ : \ d \longmapsto \omega(d)$ is a weight function, which can be d^{-1} as in Sammon's seminal paper, or other classical maps such as $\exp -\beta d$ or d^{-z}. Nonlinear mapping is minimizing the cost function over all possible point sets. It is clear that a solution is not unique, and is given up to an isometry in a Euclidean space, i.e. the composition of a translation, a rotation and a reflection.

Let us now consider a slightly different situation, where there exists an unknown point set V of n points in a Euclidean space $\mathbb{R}^k$ and where distances are known for a subset only of pairs of points. Here, it is known as part of the problem that the distances are taken between points living in a Euclidean space, whereas in NLM, such an hypothesis is not required. Let us consider the graph $G = (V, E)$ where E is the set of pairs $(i, j) \in V^2$ for which $d(i, j)$ is known. The aim is to find a mapping

$$\begin{array}{rcl} x \ : \ V & \longrightarrow & \mathbb{R}^k \\ i & \longrightarrow & x_i \end{array} \tag{5}$$

such that

$$\forall\, (i, j) \in E, \quad \|x_i - x_j\| = d_{ij} \tag{6}$$

This is known as Distance Geometry Problem (DGP, see [11], equation (1.1)). A recent and thorough survey of this problem with historical background on how it grew over decades and in different guises is [11]. See also [14]. Here, the dimension k is given, and the distances are known accurately. In real world problems, such as determination of protein structures, the distances are known up to a given precision only. DGP has been studied as well as $k-$embeddability problem for graphs. It has been proved in [16] that the $1-$embeddability problem is NP-complete and the $k-$embeddability problem is NP-hard for $k > 1$.

[†]It is unfortunate that least-square scaling has been proposed with the same name MDS than classical MDS. However, classical texts like [1, 5, 12] are clear on this matter and agree on setting the vocabulary.

A link between both problems has been established in [13], where the DGP problem is recast as finding a map

$$\begin{array}{rcl} x \,:\, V & \longrightarrow & \mathbb{R}^k \\ i & \longrightarrow & x_i \end{array}$$

such that

$$\phi(X) = \sum_{i,j \in S} \omega_{ij} \left(\|x_i - x_j\|^2 - d_{ij}^2 \right)^2 \tag{7}$$

is minimal, where X is the $n \times k$ matrix with x_i in row i and ω_{ij} are weights. Let us note different choices for the exponents of the quantities to be compared: $\left(\|x_i - x_j\|^2 - d_{ij}^2 \right)^2$ in equation (7) and $\left(\|x_i - x_j\| - d_{ij} \right)^2$ in equation (4). Clearly, if DGP has a solution, the minimum of ϕ is zero, and a set $(x_1, \ldots, x_n)$ of rows of X where $\phi(X) = 0$ is a solution of the DGP problem. A known difficulty is that ϕ has in general many local minima. The technique used in [13] is to progressively smooth ϕ by a convolution (see [11, section 3.2.2] for a general presentation of smoothing-based methods and of DGSOL which implements it).

If all weights are equal to 1 in (4), one recovers least-square scaling (see [1]). If one has

$$\left\{ \begin{array}{lcl} (i,j) \in E & \Rightarrow & \omega_{ij} = 1 \\ (i,j) \notin E & \Rightarrow & \omega_{ij} = 0 \end{array} \right.$$

one recovers DGP. There is a sort of continuity between least-square scaling, nonlinear mapping and DGP when the weights vary smoothly from 1 to 0 on pairs of items outside E.

Here, we study whether this continuity can be given a sound basis in a relevant topology, and whether it can be translated into a continuity of numerical solutions between NLM and DGP when one or a set of parameters vary.

2 Continuity between LSS, NLM and DGP

Let (V, d) be a discrete metric space, with $|V| = n$. We denote $d(i,j) = d_{ij}$ with $i, j \in V$. Let $\omega \,:\, d \longmapsto \omega(d) \geq 0$ be a weight function on distances. (V, d) being known, this yields a $n \times n$ weight matrix Ω of general term ω_{ij} with $\omega_{ij} = \omega(d_{ij})$ enabling to run a nonlinear mapping of (V, d) in a Euclidean space $\mathbb{R}^k$, where $x_i \in \mathbb{R}^k$ is the image of $i \in V$. The set of points $(x_1, \ldots, x_n)$ is denoted as a $n \times k$ matrix X, with x_i being row i. The set of real $n \times k$ matrices is denoted $\mathcal{M}(n,k)$. We define

$$\phi(X, \Omega) = \sum_{i,j=1}^{n} \omega_{ij} \left(\|x_i - x_j\|^2 - d_{ij}^2 \right)^2 \tag{8}$$

and consider the NLM problem

$$\left| \begin{array}{ll} \text{Given} & \text{a metric space } (V, d) \\ & \text{a weight matrix } \Omega \\ & \text{a dimension } k \\ \text{find} & \text{a point cloud } X \in \mathcal{M}(n,k) \\ \text{such that} & \phi(X, \Omega) \text{ is minimal} \end{array} \right. \tag{9}$$

conditioned by Ω. This problem encompasses the problems mentioned above. For example, one can see that

- if $\Omega = \mathbf{1}_n$, i.e. the $n \times n$ matrix with ones only, or $\omega(d) = 1\,\forall d$, problem (9) is least-square scaling (see [1])
- if $\Omega = \mathbf{0}_n$, i.e. the $n \times n$ matrix with zeros only, or $\omega(d) = 0\,\forall d$, any point cloud X is a solution
- if Ω is a symmetric boolean matrix, i.e. $\omega_{ij} \in \{0, 1\}$, problem (9) is DGP.

Let $G = (V, E)$ be the graph such that $(i, j) \in E$ if $\omega_{ij} = 1$. Then, in DGP vocabulary, X is a *realization* of G. (see [11, sec. 1.1.4]). Let us mention that here and in the following $\omega(0)$ is undetermined, and one can select $\omega(0) = 0$ or $\omega(0) = 1$. This can be summarized as follows:

Ω	minimum of $\phi \geq 0$	minimum of $\phi = 0$
$\omega(d) = 1$	Least square scaling	Isometry with a Euclidean space = Classical MDS
$\omega(d) \geq 0$	Nonlinear mapping	
$\omega(d_{ij}) = 1$ if $(i, j) \in E$		Distance Geometry Problem

This raises the question of the continuity of the solutions of (9) depending on Ω. The definition of continuity is not straightforward as there is a set of matrices X which are solutions of (9). We begin by an intuitive notion of continuity, and make it rigorous in section 3. A solution is said continuous at Ω if, X being a solution at Ω, whatever the neighborhood $\mathcal{N}_{\mathrm{x}}$ of X, there exists a neighborhood of Ω such that for each Ω' in it there is a solution in $\mathcal{N}_{\mathrm{x}}$, or[‡]

$$\forall\, \epsilon > 0, \quad \exists\, \eta > 0 \,:\, |\Omega' - \Omega| < \eta \quad \Longrightarrow \quad |X' - X| < \epsilon$$

One can observe that the solution is not continuous at $\Omega = \mathbf{0}_n$. Indeed, any point cloud X' is a solution for $\Omega = \mathbf{0}_n$. Let us take an $\Omega \neq \mathbf{0}_n$ which has a "nice" solution, i.e. the set of solutions is an orbit of the group of isometries acting on $\mathbb{R}^k$, and denote X a solution. Whatever $\eta > 0$,

$$\phi(X, \eta\Omega) = \eta\phi(X, \Omega)$$

and for any $\eta > 0$, the set of solutions for $\eta\Omega$ still is the set of solutions for Ω. Let us take a point cloud X' distant from any X in this set of solutions, i.e. $|X' - X| > C$ for some $C > 0$ whatever the point cloud X in the set of solutions for Ω. X' is a solution for $\eta = 0$. Whatever the neighborhood of $\mathbf{0}_n$, there is a η such that $\eta\Omega$ is in this neighborhood. And, for this $\eta\Omega$, there is no point cloud in the neighborhood of X' that is a solution for $\eta\Omega$. Then, the solution of (9) is not continuous at $\Omega = \mathbf{0}_n$. Let us note that X and X' are incongruent in the sense of [11, sect. 1.1.4.3], as a congruency class is an orbit of the action of the group of isometries in $\mathbb{R}^k$.

Our objective is to study the continuity of the solution of (9), when Ω varies within the space of matrices with non negative elements. The motivation for this is that each problem has a specific approach to build a solution:

- DGP builds it explicitly (often with an optimization scheme)
- LSS or NLM uses an optimization scheme

We study here whether a continuity between solutions of NLM, when a parameter varies in a family $(\Omega_a)_a$ and the limit corresponds to a DGP problem (some distances have zero weight), can lead to a numerical solution of DGP. On the other hand, efficient optimization schemes have been derived for solving DGP which are close to some used for NLM (DC programming, see [8] and references 6 & 7 therein).

[‡]Rigorously, one should replace $|X' - X| < \epsilon$ by: there exists an X' in the set of solutions for Ω' such that $|X' - X| < \epsilon$.

3 A topology on the set of solutions

Let us define

$$\psi(\Omega) = \{X \in \mathcal{M}(n,k) \ : \ \phi(X,\Omega) \quad \text{is minimal}\}$$

For a given Ω, if $m = \min_X\{\phi(X,\Omega)\}$, we have

$$\psi(\Omega) = \phi^{-1}(m)$$

We will define a topology in two spaces, in which: (i) an element is a point cloud $X \in \psi(\Omega)$ and (ii) an element is a set of point clouds. The latter will enable to define the neighborhood of a set $\psi(\Omega)$ for a given Ω and study continuity of ψ. If the topology is associated to a distance, this will read

$$\forall\, \epsilon > 0, \quad \exists\, \eta > 0 \ : \ |\Omega' - \Omega| < \eta \quad \Longrightarrow \quad d(\psi(\Omega'), d(\Omega)) < \epsilon$$

For this, we must define a distance $d(\psi(\Omega'), d(\Omega))$. A distance can be defined between compact subsets of a metric set: the Hausdorff distance (see e.g. [6, p. 34]). Let us briefly recall its definition. Let A, B be two closed sets in a metric space where the distance between points is denoted d. One defines

$$\delta(A,B) = \max_{x\in A}\left\{\min_{y\in B} d(x,y)\right\}$$

It is easy to show that $\delta(A,B) = 0 \Leftrightarrow A = B$. But it is not symmetric. So one defines

$$d_{\text{H}}(A,B) = \max\{\delta(A,B), \delta(B,A\}$$

This is defined if A and B are compact subsets, and the triangular identity is fulfilled. We denote d_{H} the Hausdorff distance between A and B. We first define a distance between two points clouds X and X' as the Hausdorff distance between them. For this to hold, X and X' must be compact. As they are obviously closed, they must be bounded. Therefore, we show a first lemma which gives the condition under which $X \in \psi(\Omega)$ is bounded. Let $G = (V,E)$ be the graph associated to Ω and defined by

$$(i,j) \in E \quad \Longleftrightarrow \quad \omega_{ij} > 0$$

Then

Lemma 1. *The set $X \in \psi(\Omega)$ is bounded if, and only if, G is connected.*

Proof. • G connected $\Rightarrow$ X is bounded. We show first that $\|x_i - x_j\|$ is bounded for any pair (x_i, x_j) if $(i,j) \in E$. If $m = \min_X\ \phi(X,\Omega)$, we have

$$\phi(X,\Omega) = \sum_{i\sim j} \omega_{ij}\left(\|x_i - x_j\|^2 - d_{ij}^2\right)^2 = m$$

Hence,

$$\forall\, i \sim j, \quad \omega_{ij}\left(\|x_i - x_j\|^2 - d_{ij}^2\right)^2 \leq m$$

and

$$\|x_i - x_j\|^2 \leq d_{ij}^2 + \frac{m}{\omega_{ij}}$$

Let $\delta^2 = \max\ d_{ij}^2$ and $\omega = \min\ \omega_{ij}$. Then,

$$\forall i \sim j, \quad \|x_i - x_j\|^2 \leq \delta^2 + \frac{m}{\omega}$$

Now, let i, j with $i \nsim j$. As G is connected, there is a path $ia_1 \dots a_p j$ linking i with j. By triangular inequality, $\|x_i - x_j\| \leq \sum_{k=0}^{p} \|x_{a_k} - x_{a_{k+1}}\|$ (with $a_0 = i$ and $a_{p+1} = j$). As G is finite, it has a diameter D, and $\|x_i - x_j\| \leq D\left(\delta^2 + \frac{m}{\omega}\right)$. If there is a M such that for any pair i, j $\|x_i - x_j\| \leq M$ and $\sum_i x_i = 0$, then X is bounded.

• G not connected $\Rightarrow$ X is not bounded. If G is not connected, there are at least two subsets A, B of vertices without connection between A and B. We have

$$\phi(X, \Omega) = \sum_{i,j \in A} \omega_{ij} \left(\|x_i - x_j\|^2 - d_{ij}^2\right)^2 + \sum_{k,m \in B} \omega_{km} \left(\|x_k - x_m\|^2 - d_{km}^2\right)^2$$

Let us now set

$$x_i' = x_i + \frac{h}{|A|}, \qquad x_j' = x_j - \frac{h}{|B|}$$

We still have $\sum_i x_i' + \sum_j x_j' = 0$ and, $\forall h$, $\phi(X', \Omega) = \phi(X, \Omega) = m$. If $h \to +\infty$, X is unbounded

• As a consequence, X is bounded if, and only if, G is connected. □

Then, the set of point clouds X solution of (9) for a weight matrix with a connected associated graph is a metric space, with Hausdorff distance. We then consider the subsets $\psi(\Omega)$, running over the matrices Ω fulfilling connectedness condition. We call X a solution of (9), and $\psi(\Omega)$ the set of solutions. $\psi(\Omega)$ is closed as it is the pre-image of $\{m\}$, the minimum of ϕ. But it is not bounded. Indeed, $\psi(\Omega)$ is invariant by any isometry in $\mathbb{R}^k$. A translation is an isometry, and the distance between two solutions X, X' such that $X' = h + X$ with $h \in \mathbb{R}^k$ is $\|h\|$. Then, $\psi(\Omega)$ is unbounded. Therefore, we impose on each solution $X \in \psi(\Omega)$ that it is centered, and consider

$$\psi_{\text{C}}(\Omega) = \left\{ X \in \psi(\Omega) \ : \ \sum_i x_i = 0 \right\}$$

Then $\psi_{\text{C}}(\Omega)$ is bounded. Being a closed and bounded subset of a metric space, it is a compact set. We can use the Hausdorff distance d_{H} between $\psi_{\text{C}}(\Omega)$ and $\psi_{\text{C}}(\Omega')$, and continuity of the solution of (9) is defined as

$$\forall\, \epsilon > 0, \quad \exists\, \eta > 0 \ : \ |\Omega' - \Omega| < \eta \quad \Longrightarrow \quad d_{\text{H}}(\psi(\Omega'), d(\Omega)) < \epsilon \tag{10}$$

It is reasonable to impose that for any pair of points $0 \leq \omega_{ij} \leq 1$. The set of such weight matrices is homeomorphic to the hypercube $[0, 1]^N$ with $N = \frac{n(n-1)}{2}$. The weight matrices associated to DGP are boolean. They correspond to vertices of the hypercube.

4 Continuity and rigidity

Let us denote by $\mathcal{W}$ the set of weight matrices Ω such that the associated graph $G = (V, E)$ is connected, and by $\mathcal{X}$ the set of all centered point clouds X with n points. Then, $\psi(\Omega)$ is a compact subset of $\mathcal{X}$.

Let $\Omega \in \mathcal{W}$ be the weight matrix of a DGP, i.e. there is a connected graph $G = (V, E)$ such that $\omega(i, j) = 1$ if $(i, j) \in E$ and $\omega(i, j) = 0$ if $(i, j) \notin E$. We show here that the continuity of ψ at Ω is linked with the rigidity of the framework associated to the graph G. A framework is a set of n points in $\mathbb{R}^k$ solution of (6). A framework is *rigid* if it is defined up to an isometry (i.e.

the set of known distances is sufficient to derive all other distances in a unique way). Otherwise, it is said *flexible* (see [11, sect. 1.1.4] for the definitions, and section 4.2 of same reference for a thorough discussion of those notions). We show here

Lemma 2. *The set of weight matrices for which any realization in $\mathbb{R}^k$ is a rigid framework is not closed.*

Proof. For this, it suffices to exhibit a sequence $(\Omega_\eta)_\eta$ for which for any $\eta > 0$ DGP solution is a rigid framework which converges to a weight matrix $\Omega = \Omega_0$ for which the solution is flexible. Let us consider $n = 3$ and $k = 2$ (i.e. 3 points in $\mathbb{R}^2$) and

$$D = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & \sqrt{2} \\ 1 & \sqrt{2} & 0 \end{pmatrix} \qquad \text{and} \qquad \Omega_\eta = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & \eta \\ 1 & \eta & 1 \end{pmatrix}, \qquad 0 \leq \eta \leq 1$$

where D is the pairwise distance matrix between points. A realization is

$$X' = \begin{pmatrix} x = & 0 & 0 \\ y = & 0 & 1 \\ z = & 1 & 0 \end{pmatrix}$$

If $\eta > 0$, the realization is defined up to an isometry: the lengths of the edges of the triangle made by X' are known, and this fixes the triangle up to an isometry. Whereas if $\eta = 0$, $\psi(\Omega_0)$ is the set of points $x, y, z \in \mathbb{R}^2$ such that $\|x - y\| = \|x - z\| = 1$, which is isomorphic to $\mathbb{O} \times \mathbb{O}$ where $\mathbb{O}$ is the group of rotations in $\mathbb{R}^2$ (y and z are on the circle of center x and radius 1). At $\eta = 0$, Ω_η is flexible. □

As a consequence, ψ is not continuous at $\Omega = \Omega_0$. Indeed,

$$X = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ -1 & 0 \end{pmatrix} \in \psi(\Omega)$$

and there is a constant $C > 0$ (the calculation is easy, and omitted here) such that

$$\forall\, \eta > 0, \quad \forall\, X' \in \psi(\Omega_\eta), \quad d(X, X') > C$$

whereas $|\Omega - \Omega_\eta| = \eta$. X and X' are incongruent.

5 Convergence to a Heaviside function

We first succinctly give a motivation for the next short case-study of a given family of weight matrices. Molecular based taxonomy consists in assigning specimens to species based on similarities between some relevant DNA sequences, called markers [4]. Distances between sequences are computed with classical algorithms (see [3]). This dictionary between morphological based and molecular based taxonomy works up to a given threshold of distance (beyond this threshold, computed genetic distances are highly likely to be blurred). Thus, as a simplified setting, distances up to a given threshold only are taken into account. We have a set of pairwise distances, and we wish to build an Euclidean image of it that is as accurate as possible, ignoring the distances beyond a given threshold. Provided a dimension k is given, this can be formalized as a DGP problem, where in (V, d) distances below a given threshold θ only are given, i.e. G is defined by

$$E = \{(i, j) \,:\, d(i, j) \leq \theta\}$$

The weights are given by a Heaviside function $w(d) = H(\theta - d)$. Such a function has a discontinuity at $d = \theta$. We can consider the weight function

$$\omega_{a,\theta}(d) = \frac{1 - \tanh(a(d-\theta))}{2} \tag{11}$$

as well because distances are blurred progressively. We have

$$\lim_{a\to\infty} \omega_{a,\theta}(d) = H(\theta - d)$$

with the topology of uniform convergence

$$\forall\, \epsilon > 0, \quad \exists\, \alpha \in \mathbb{R}^+ \,:\, \forall\, d > 0, \quad \forall\, a > \alpha, \quad |\omega_{a,\theta}(d) - H(\theta - d)| < \epsilon$$

Let us assume that θ is fixed. We are given a converging family of weight matrices $(\Omega_a)_{a\in\mathbb{R}^+}$ satisfying

$$\Omega_\infty = \lim_{a\to\infty} \Omega_a$$

Then, if ψ is continuous at Ω_∞, one can define

$$\psi(\Omega_\infty) = \lim_{a\to\infty} \psi(\Omega_a)$$

As Ω_∞ is the weight matrix of a DGP, i.e. there is a graph $G = (V, E)$ such that $w_{\infty,ij} = 1$ for $(i, j) \in E$ and $w_{\infty,ij} = 0$ for $(i, j) \notin E$, this means that the solution of DGP can be found as the limit of a family of solutions of NLM, when $a \to \infty$.

Let us make a few remarks on possible frameworks to address such a problem. We call *Euclidean image* a solution of problem (9).

- the dimension k is not fixed by the problem itself. Hence, this problem can be rephrased as EMDCP.
- It is however interesting to have a Euclidean image in a low dimensional space where shapes of point clouds can be better studied (e.g. $(k \leq 3)$)
- In order to build such a solution, we can select a large k, and project it in an optimal way in a low dimensional space (by Principal Component Analysis)
- The choice to ignore distances beyond a given threshold is similar to the Isomap procedure, which is a manifold learning technique using graph based distances to approximate geodesic distances on a manifold [9, 17]. A link between Isomap and the solution of a DGP has recently been made in [10].

Here, we fix a dimension k (in our example, $k = 2$) and study the convergence of NLM solutions to a solution of DGP. We focus on the numerical stability of such an approach. Indeed, for a given Ω, the manifold $\mathscr{M}$ in $\mathbb{R}^{n\times k} \times \mathbb{R}$ defined by

$$\mathscr{M} = \left\{(X, z) \in \mathbb{R}^{n\times k} \times \mathbb{R} \quad \text{s.t.} \quad z = \phi(X, \Omega)\right\}$$

is far from having a unique minimum on z. The set $(\phi^{-1}(m), m)$ when m is the minimum of ϕ is a minimal submanifold (all its points have a minimum "elevation" z). It is likely that many other local minima exist, as well as "flat valleys". We study here whether the ill behavior of $\mathscr{M}$ may hamper to use the continuity of the solution between NLM and DGP as a way to obtain a numerical solution to DGP. We mention that there exist several efficient methods to solve DGP by numerical optimization, like Difference on Convex functions programming [8], Distance Continuation [13] or Isomap-based heuristics [10]. We are interested here in the continuity between solutions of NLM and DGP on continuous families of weight matrices Ω.

6 Numerical optimization schemes

We have implemented a numerical optimization scheme for solving problem (9) with function (11) for $k = 2$, with two methods known to be efficient for finding global minima: `BFGS` and `basin hopping` [19] in package `scipy.optimize`. As distances are unreliable beyond a given threshold, we have adopted the framework of isomap-based heuristics [10]. We have built a data set in silico, that we wish to recover (it should be the solution of DGP and NLM problems). The objective is to test the continuity of the numerical solution. The dataset consists of $n = 100$ points in $\mathbb{R}^2$ randomly (uniform law) distributed within the ring delimited by circles $r = 0.8$ and $r = 1$ centered at origin (see FIGURE 1 top left). We have selected the weight function defined in equation (11) with different values of stiffness coefficient a and threshold θ. The numerical results for various values of a and θ are presented in TABLE 1 for both BFGS and Basin Hoping. Each cell contains the value of the cost function for the result of the procedure for one value of a, one of θ and one method. The starting point for NLM optimization phase could be the result of Multidimensional Scaling. But this requires the knowledge of all distances. In our case, we wish to avoid the use of distances larger than a given threshold θ (even if they have been measured). Therefore, the starting point is the point cloud built on distances as computed by Isomap on the partial distance matrix of distances $d \leq \theta$. This yields the following observations:

- The cost function for `Isomap + BFGS` is always equal to or lower than the cost function for `Isomap + basin hoping`. For this type of problem, `Isomap + BFGS` is recommended (we have tested other methods, results not shown, like simulated annealing, for which results were worse).
- The cost function decreases when θ increases for a given a, and is less sensitive to θ. If θ is too small, there is a significative probability that the graph of partial distanes is not connected. When θ and a are low, the optimization step may not converge.

A picture of the datasets obtained by each method (`Isomap`, `Isomap + BFGS`, `Isomap + basin hopping`) is displayed in FIGURE 1. The eye can recognize a deformed ring in the bottom right graph, namely the best reconstruction with basin hopping. The optimization scheme has been trapped on this pattern. One is tempted to twist the outer small loop to recover a shape close to a ring. The fact that this twist (ouwards like here, or inwards in some other simulations) of a fraction of the ring is a trap can be heuristically understood: the main discrepancy between exact distances and reconstructed distances is for those points which have been twisted outward. They are much closer to the points on the opposite on the ring on the reconstruction than in the initial data set (top left). However, the stiffness of the decrease of the weights for $d \geq 1$ lets the cost function ϕ be nearly insensitive to those discrepancies. The role of stiffness a is then more important than the role of threshold θ as the weight of pairs of points separated by a large distance (like $(\|x_i - x_j\| \geq 1)$ is annihilated by a very low weight $(\omega_{a,\theta}(1) = 6.7\,10^{-3}$ for $a = 5$ and $\theta = 0.75)$. Careful observation of many simulations lead to the observation that similar low cost function values may correspond to very different geometric settings for the solution.

7 Conclusion

We have set a framework to study continuity of the solution of (9) when the weight matrix Ω varies. The main point to address is that a solution to (9) is never a single point cloud, but always a set of point clouds, a union of orbits of the action of the group of isometries on $\mathbb{R}^k$. We have exhibited an example where the solution is not continuous at a weight matrix where the realization of the solution is not rigid. We expect that it is continuous when realizations

are rigid, but this has not been shown here. This is deferred to further work on study of the topological structure of the solutions in relation with the weight matrices. We have linked DGP and NLM in a common framework using this continuity. We have studied whether the continuity of the solution can serve as a basis for ensuring the continuity of numerical solutions when a parameter varies in the weight matrix. Therefore, we have built a simple *in silico* dataset, and derived a procedure with the output of Isomap as initial point for optimization step in NLM, with two optimization schemes (BFGS and Basin Hopping). We have produced good hints to show convergence of NLM solution to DGP solution (a situation when $a \to \infty$). We have shown as well that different geometric settings of the solution may correspond to similar very low values of the cost function. It is likely that this is due to the complicated shape of the manifold $\mathscr{M}$ of cost function as function of coordinates of a point cloud, and motivates further studies.

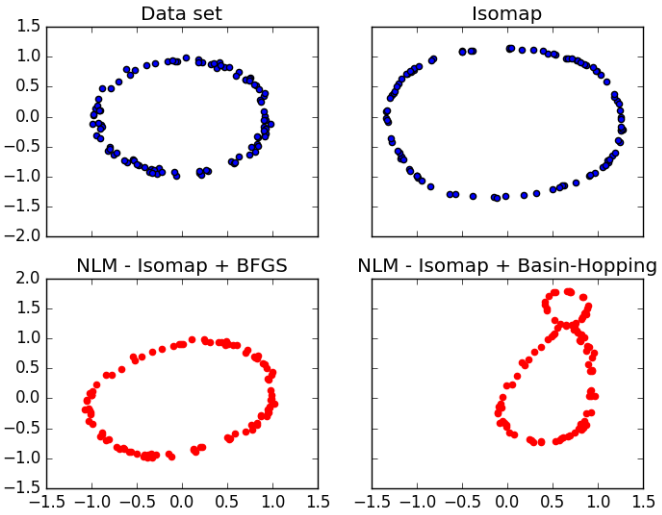


Figure 1: Results of numerical simulations for reconstructing the point cloud knowing pairwise distances lower than a given threshold only. Parameters: $\theta = 0.5$ and $a = 10$ (see text). Cost function for BFGS is $\phi = 1.018$ and for Basin Hopping $\phi = 2.211$. Top left: Simulated dataset. Top right: dataset reconstructed with standard Isomap procedure. Bottom: dataset reconstructed with BFGS optimization scheme (left) or Basin Hopping (right) with top right dataset as initial values.

	$a = 5$		$a = 10$		$a = 20$		$a = 50$	
	BFGS	BH	BFGS	BH	BFGS	BH	BFGS	BH
$\theta = 1$	94.82	94.82	0.848	10.82	$9.45\,10^{-3}$	9.4510^{-3}	1.8310^{-3}	1.8310^{-3}
$\theta = 0.5$	55.42	55.42	1.018	2.212	$1.58\,10^{-4}$	1.0310^{-2}	2.8510^{-5}	4.1910^{-3}
$\theta = 0.25$			0.672	0.672	$1.9\,10^{-4}$	8.6010^{-4}	2.1210^{-5}	1.2410^{-4}

Table 1: Result of nonlinear mapping: cost function for the reconstruction of an *in silico* dataset, for various values of parameters a and θ, and two methods for optimization phase in NLM: BFGS (for BFGS method) and BH (for Basin Hoping method). The starting point of optimization is the point cloud built by standard Isomap procedure.

References

[1] Cox, T., Cox, M.A.A.: Multidimensional Scaling - Second edition, Monographs on Statistics and Applied Probability, vol. 88. Chapman & al. (2001)

[2] Cristianini, N., Shawe-Taylor, J.: An introduction to Support Vector Machines and other Kernel-based learning methods. Cambridge University Press (2000)

[3] Gusfield, D.: Algorithms on strings, trees, and sequences. Cambridge University Press, Cambridge, UK (1997)

[4] Hillis, D.M., Moritz, C., Mable, B.: Molecular Systematics. Sinauer, Sunderland, Mass. (1996)

[5] Izenman, A.J.: Modern Multivariate Statistical Techniques. Springer, NY (2008)

[6] Krantz, S.G., Parks, H.R.: Geometric integration theory. Birkaüser (2008)

[7] Kruskal, J.B.: Non Metric Multidimensional Scaling: a numerical approach. Psychometrika 29(2), 115–129 (1964)

[8] Le Thi, H.A., Pham Dinh, T.: DC programming Approaches for Distance Geometry Problems, chap. 13 in Distance Geometry, Theory, Methods and Applications, Mucherino & al. (Ed.), pp. 225–290. Springer (2013)

[9] Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, NY (2007)

[10] Liberti, L., D'Ambrosio, C.: The Isomap Algorithm in Distance Geometry. In: Iliopoulos, C.S., Pissis, S.P., Puglisi, S.J., Raman, R. (eds.) 16th International Symposium on Experimental Algorithms (SEA 2017). pp. 5:1–5:13. Leibniz International Proceedings in Informatics (2017)

[11] Liberti, L., Lavor, C., Maculan, N., Mucherino, A.: Euclidean distance geometry and applications. SIAM review 56(1), 3–69 (2014)

[12] Mardia, K.V., Kent, J., Bibby, J.M.: Multivariate Analysis. Probability and Mathematical Statistics, Academic Press, (1979)

[13] Moré, J.J., Wu, Z.: Global continuation for Distance geometry Problems. SIAM J. Optim. 7(3), 814–836 (1997)

[14] Mucherino, A., Lavor, C., Liberti, L., Maculan, N. (eds.): Distance Geometry. Springer (2013)

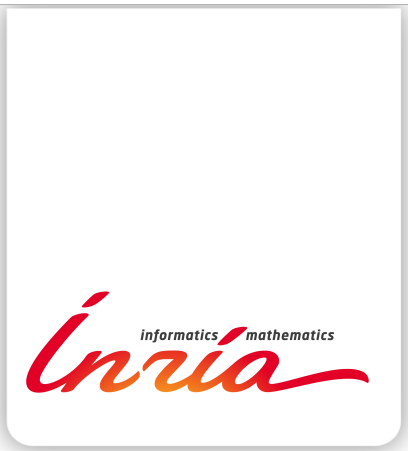
[15] Sammon, J.W.: A nonlinear mapping algorithm for data structure analysis. IEEE Transactions on Computers 18(5), 401–409 (1969)

[16] Saxe, J.B.: Embeddability of weighted graphs in k-space is strongly NP-hard. In: 17th Allerton Conference in Communication Control and Computing. pp. 480–489 (1979)

[17] Tennenbaum, J.B., de Silva B., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)

[18] Torgerson, W.S.: Multidimensional Scaling: I. Theory and Method. Psychometrika 17(4), 401–419 (1952)

[19] Wales, D.J., Doye, J.P.K.: Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. Journal of Physical Chemistry A 101(5111-5116) (1997)



RESEARCH CENTRE
BORDEAUX – SUD-OUEST

200 avenue de la Vieille Tour
33405 Talence Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399