Integer programming for the search of a discretization order in distance geometry problems

Jeremy Omer[1], Antonio Mucherino[2]

[1] INSA Rennes and IRMAR, Rennes, France
jomer@insa-rennes.fr
[2] IRISA, University of Rennes 1, Rennes, France
antonio.mucherino@irisa.fr

Keywords : *distance geometry, graph embedding, integer programming, vertex order.*

1 The distance geometry problem

Given a simple weighted undirected graph $G = (V, E, d)$, where $d : E \mapsto \mathbb{R}^+$, and a dimension $K \in \mathbb{N}$, the distance geometry problem (DGP) consists in finding a set of points of $\{x_i \in \mathbb{R}^K\}_{i \in V}$ such that

$$\|x_i - x_j\|_2 = d(i,j),\ \forall (i,j) \in E.$$

The DGP is NP-hard, and has natural applications in sensor network localization ($K = 2$) and in the identification of protein configurations ($K = 3$) [1].

DGP can be modeled as an unconstrained nonconvex and nonlinear program:

$$\min \sum_{(i,j)\in E} \left| \|x_i - x_j\|_2^2 - d(i,j)^2 \right|$$

Finding a local optimum is generally not interesting unless its value is zero. Global optimization techniques thus need to be applied, which involves a strong limitation on the size of the graphs that can be treated.

To overcome this issue, Liberti et al. [2] have noticed that in most instances of the cited applications, the graph G admits a *discretization order*, i.e. a numbering of the vertices such that:

- the first K vertices form a clique;
- the following vertices are adjacent to at least K vertices among those with a smaller number.

Assuming some independency property on the distance function, the intersection of K spheres of $\mathbb{R}^K$ consists of at most two points. So the existence of a discretization order means that we can enumerate the set of possible positions for each vertex in a tree structure. Every layer of the tree corresponds to the set of potential positions for one vertex of G, and each solution of the DGP corresponds to one leaf of the tree. Branch-and-prune algorithms (BP) have been proposed for an efficient exploration of this tree.

A greedy algorithm has been designed to find discretization orders [3]. In its basic version, the main idea is to start with a $K-$clique and then add the vertices with most neighbors among the already numbered vertices. However, no formal definition of the search for an optimal discretization order has been given yet. In this work, we propose a formulation based on an upper bound of the number of nodes in the BP tree, and we provide an integer programming model (IP) for its exact solution.

2 Formulation as an integer program

In some DGP applications, it is of interest to find all the solutions of the problem. In order to do so, a solution algorithm needs to explore the entire search tree. We say that a discretization order is optimal if, by definition, it gives rise to a search tree containing a minimum number of nodes, while allowing the enumeration of the complete set of solutions of the DGP. Since the number of nodes of the search tree cannot be known without building the tree, our approach is based on the computation of an upper bound on the number of nodes. We thus develop a model to search for a discretization order that minimizes this upper bound.

The model relies on the following observations (a vertex u can play the role of *reference* for v if $(u, v) \in E$):

1. a vertex with K references has at most two possible positions;
2. a vertex with $K + 1$ references or more has at most one possible position;
3. one half of the BP nodes are pruned when a) current vertex has $K+1$ references or more, and b) one of its references has exactly K references.

The upper bound can then be formulated using a simple induction on the number of BP nodes at level $k + 1$ of the tree as a function of the number of BP nodes at level k.

We then model the problem that consists of finding the order that minimizes the upper bound on the number of BP nodes as an IP. This IP involves variables that indicate:

- if i is a reference for j $(\forall(i, j) \in E)$
- if i occupies rank k in the order $(\forall i \in V, k = 1, \ldots, |V|)$
- if i has more than three references $(\forall i \in V)$
- if level k of the BP tree has at most 2^p nodes, $(\forall k = 1, \ldots, |V|, p \leq k)$

Assuming that the number of edges can be bounded by $C|V|$, where C is a constant valid for all the instances, we obtain a model with $(O)(|V|^2)$ variables and constraints.

3 Experimental results and perspectives

The IP has been solved on a benchmark of protein configuration instances. Preliminary tests validate that the optimal value of the IP is a good indicator of the number of BP tree nodes explored during the enumeration of all the solutions. However, the problem remains highly combinatorial, so instances with more than 100 vertices cannot be solved in reasonable time.

In future research, we will look into the decomposition of the problem based on a systematic analysis of the symmetries of the problem. We also plan to extend the formulation to handle the case where some distances should lie into given intervals.

References

[1] Liberti, Leo, Lavor, Carlile, Maculan, Nelson and Mucherino, Antonio *Euclidean Distance Geometry and Applications.* SIAM Review, vol. 56, no. 1, pp. 3-69, 2014

[2] Liberti, Leo, Lavor, Carlile and Maculan, Nelson *A Branch-and-Prune algorithm for the Molecular Distance Geometry Problem.* International Transactions in Operational Research, 2008

[3] Gonçalves, Douglas S. and Mucherino, Antonio *Optimal partial discretization orders for discretisable distance geometry.* International Transactions in Operational Research, 23(5), 947–967, 2016.