Maximum feasible subsystems of distance geometry constraints

Maurizio Bruglieri[1], Roberto Cordone[2], Leo Liberti[3]

[1] Politecnico di Milano, Milan, Italy
maurizio.bruglieri@polimi.it
[2] Universitá degli Studi di Milano, Milan, Italy
roberto.cordone@unimi.it
[3] CNRS LIX - Ecole Polytechnique, Palaiseau, France
liberti@lix.polytechnique.fr

Abstract

In this work we discuss mathematical programming formulations for satisfying the maximum number of distance geometry constraints with minimum error.

Keywords : *protein folding, experimental error formulation, systematic error formulation.*

1 Introduction

We discuss an interesting hybrid of two problems: the Maximum Feasible Subsystem (MaxFS) [1] and the Distance Geometry Problem (DGP) [4], and its application to the problem of determining the spatial conformation of proteins from distance data derived from Nuclear Magnetic Resonance (NMR) experiments.

The MaxFS is as follows: given a system of constraints, generally of the form

$$\forall i \in I \qquad g_i^L \leq g_i(x) \leq g_i^U, \tag{1}$$

with $g_i : \mathbb{R}^n \to \mathbb{R}$, determine a subset $S \subseteq I$ of maximum cardinality such that the subsystem of constraints of Eq. (1) indexed by S is feasible.

The (Euclidean) DGP is as follows: given an integer $K > 0$ and a simple connected edge-weighted graph $G = (V, E, d)$, where $d : E \to \mathbb{R}_+$, determine whether there exists a *realization* $x : V \to \mathbb{R}^K$ such that:

$$\forall \{i, j\} \in E \qquad \|x_i - x_j\|_2 = d_{ij}. \tag{2}$$

There are many applications of the DGP and even more variants [4]. The one we are specially interested in is the *interval* DGP (*i*DGP), which replaces $d : E \to \mathbb{R}_+$ with the interval weight function $d : E \to \mathbb{IR}_+$ such that $d(\{i, j\}) = [L_{ij}, U_{ij}]$. Specifically, Eq. (2) becomes

$$\forall \{i, j\} \in E \qquad L_{ij} \leq \|x_i - x_j\|_2 \leq U_{ij}. \tag{3}$$

From now on, we shall assume all norms are Euclidean unless stated otherwise.

We are now in the position of stating the main problem discussed in this paper.

> Maximum Feasible Subsystem of Distance Geometry constraints ($MaxFS_{DGP}$).
> Given an integer $K > 0$ and a simple connected edge-weighted graph $G = (V, E, d)$ with $d : E \to \mathbb{IR}_+$, determine the maximum cardinality subset $S \subseteq E$ inducing a connected subgraph of G, such that there exists a realization $x : V[E] \to \mathbb{R}^K$ satisfying
>
> $$\forall \{i, j\} \in S \qquad L_{ij} \leq \|x_i - x_j\|_2 \leq U_{ij}. \tag{4}$$

The $MaxFS_{DGP}$ is motivated by a specific application of the DGP, namely the determination of the shape of proteins given some of their inter-atomic distances. In principle, NMR can determine all inter-atomic distances in a given protein up to a certain length threshold (somewhere between 5Å and 6Å). In practice, reality is fuzzier than this. First, we note that proteins rarely crystallize (so X-ray crystallography does not help), but usually live in a solution. Secondly, proteins vibrate, but we assume that they do not (this is called the "molecular rigidity assumption"). NMR experiments actually help determine a probability distribution over triplets (atom label, atom label, distance value); this distribution is used to imperfectly reconstruct the weighted graph G that is the actual input to the DGP, often using a simulated annealing approach. According to [2], this process induces two types of errors: *experimental errors* (due to the rigidity assumption), and *systematic errors* (due to the imperfect reconstruction). Specifically, The experimental errors are accommodated by the interval bounds on the iDGP. The systematic errors are described in [2] as consisting of a certain proportion of completely wrong distances. This induces sets of constraints in (3) that are likely to be infeasible. It is clear that this situation is well addressed by the $MaxFS_{DGP}$.

2 Formulations

In this section we present and discuss several Mathematical Programming (MP) formulations related to experimental errors, systematic errors, and trade-offs between the two. We remark that ℓ_2 norm terms are always squared in order to remove the square root: in terms of MP literature, working with Polynomial Programming (PP) offers more opportunities than with general Nonlinear Programming (NLP).

2.1 Experimental errors

Experimental errors are usually addressed by minimizing the infeasibilities w.r.t. Eq. (2) or Eq. (3). A commonly seen box-constrained formulation targeting the DGP is:

$$\min_{x \in [x^L, x^U]} \sum_{\{i,j\} \in E} (\|x_i - x_j\|^2 - d_{ij}^2)^2, \tag{5}$$

where x^L, x^U are given lower and upper bounds for the decision variable $n \times K$ matrix $x = (x_1, \ldots, x_n)$. Eq. (5) was tested computationally in e.g. [3]. An equivalent formulation for the iDGP replaces each term $\|x_i - x_j\|^2 - d_{ij}^2$ with

$$\max(0, L_{ij}^2 - \|x_i - x_j\|^2) + \max(0, \|x_i - x_j\|^2 - U_{ij}^2),$$

yielding

$$\left.\begin{array}{rrcl} \min & \sum_{\{i,j\} \in E} (s_{ij} + t_{ij}) & & \\ \forall \{i,j\} \in E & L_{ij}^2 - \|x_i - x_j\|^2 & \leq & s_{ij} \\ \forall \{i,j\} \in E & \|x_i - x_j\|^2 - U_{ij}^2 & \leq & t_{ij} \\ \forall \{i,j\} \in E & s_{ij}, t_{ij} & \geq & 0 \\ & x^L \leq x & \leq & x^U. \end{array}\right\} \tag{6}$$

2.2 Systematic errors

The $MaxFS_{DGP}$ can be formulated in a natural way, using big-M techniques [1] as follows:

$$\left.\begin{array}{rrcl} \max & \sum_{\{i,j\} \in E} y_{ij} & & \\ \forall \{i,j\} \in E & d_{ij}^2 - M(1 - y_{ij}) \leq & \|x_i - x_j\|^2 & \leq d_{ij}^2 + M(1 - y_{ij}) \\ & y & \in & \{0,1\}^m. \end{array}\right\} \tag{7}$$

We point out that a valid value of M exists for any instance of $MaxFS_{DGP}$.

Proposition 1 *If $M = R^2$, where $R = \sum_{\{i,j\}\in E} d_{ij}$, then the optimal solution of Eq. (7) solves the $MaxFS_{DGP}$.*

Proof : First, we claim that any feasible DGP instance can be realized in a sphere of radius R. A cycle graph C on $V = \{1, 2, \ldots, n\}$ with $E = \{\{1,2\},\{2,3\},\ldots,\{n-1,m\},\{1,n\}$ with $d_{1n} = \sum_{\{i,j\}\in E} d_{ij}$ can be realized on a straight segment of length $R = d_{1n}$ embedded in any Euclidean space [5]; if this segment is centered about the origin it belongs by construction to the sphere $R\mathbb{S}^{K-1}$. Any other biconnected graph on n vertices will have more cycles than C, and hence will induce realizations in $\mathbb{R}^K$ having segments shorter than R when projected on any coordinate axis. Connected but non-biconnected graphs are the same as trees: the tree yielding a realization with longest segment projection on any coordinate axis is the path on n vertices realized as a segment of length R; again, by centering the segment it is easy to see that the path can be realized in a sphere of radius R.

Lastly, we simply note that the above claim also shows that the maximum possible distance between two vertices i, j in a realization is R. This shows that if a $MaxFS_{DGP}$ instance has a solution with a certain support vector y^* for the maximum cardinality set of feasible constraints, then setting $y = y^*$ in Eq. (7) will induce a valid realization x^* of the subgraph consisting of the edges $\{i, j\}$ for which $y^*_{ij} = 1$, and vice versa. □

In practice, segment realizations are extremely rare, and therefore M can be tightened w.r.t. Prop. 1. We remark that bounds on M can also be inferred from x^L, x^U, if they are given; and, conversely, that $[x^L, x^U]$ can be set to $[-M, M]$ if the application field does not explicitly provide them.

2.3 Bi-objective formulation

As explained in the introduction, the $MaxFS_{DGP}$ requires a the trade-off between the experimental and the systematic error. Such a trade-off is modeled by the following bi-objective formulation based on: binary variables y_{ij} equal to 1 if the constraint corresponding to the edge $\{i, j\}$ is deactivated, 0 otherwise; continuous variables s_{ij} representing the experimental error on edge $\{i, j\}$ and continuous auxiliary variables r_{ij} to model the behavior of variables y_{ij}. In particular, the last constraint forces $y_{ij} = 0$ if $r_{ij} > 0$, while it becomes redundant if $r_{ij} = 0$. This way, variables s_{ij} model the experimental error only for the distance geometry constraints that are not deactivated.

$$\left.\begin{array}{rrcl} \max & \sum\limits_{\{i,j\}\in E} y_{ij} & & \\ \min & \sum\limits_{\{i,j\}\in E} s^2_{ij} & & \\ \forall\{i,j\}\in E & \|x_i - x_j\|^2 & = & d^2_{ij} y_{ij} + r_{ij} + s_{ij} \\ \forall\{i,j\}\in E & y_{ij} & \leq & 1 - \frac{r_{ij}}{M_{ij}} \\ & x & \in & \mathbb{R}^{nK} \\ & y & \in & \{0,1\}^m \\ & r & \in & \mathbb{R}^m_+ \\ & s & \in & \mathbb{R}^m. \end{array}\right\} \quad (8)$$

where

$$M_{ij} = max_{x_i\in[x^L_i,x^U_i],x_j\in[x^L_j,x^U_j]}\|x_i - x_j\|^2$$

If x^L, x^U are not provided, we can exploit again Proposition 1 and set $M_{ij} = R^2$.

2.4 Minimization of experimental error with systematic error cardinality constraint

Solving bi-objective formulations does not yield "a solution", in general, but a whole set of Pareto-optimal solutions, which might in principle be infinite. A possible way to obtain them

is by ϵ-constraint method which consists in turning one of the objectives into a constraint bounded by an arbitrary value, which is then changed iteratively. Since the first objective is discrete, it is more advantageous to apply such a method to it rather than the second one. This way, the first objective of formulation (8) is turned in a constraint ensuring that at most p distance geometry constraints can be violated. The resulting formulation is the following one:

$$\left.\begin{array}{rrcl} \min & \sum\limits_{\{i,j\}\in E} s_{ij}^2 & & \\ \forall\{i,j\}\in E & \|x_i - x_j\|^2 & = & d_{ij}^2 y_{ij} + r_{ij} + s_{ij} \\ \forall\{i,j\}\in E & y_{ij} & \leq & 1 - \frac{r_{ij}}{M_{ij}} \\ & \sum\limits_{\{i,j\}\in E} y_{ij} & \geq & m - p \\ & x & \in & \mathbb{R}^{nK} \\ & y & \in & \{0,1\}^m \\ & r & \in & \mathbb{R}_+^m \\ & s & \in & \mathbb{R}^m. \end{array}\right\} \tag{9}$$

3 Conclusion and perspectives

All the described formulations have been implemented in AMPL and solved by the state-of-the-art solver BARON. Preliminary results show that the new formulation proposed in Section 2.4 allows to obtain a strong reduction of the experimental error compared to that obtained with the original formulation described in Section 2.1 where the systematic error component was not considered. Future works, concern the development of smarter solution methods (e.g., branch-and-bound based on a Semidefinite Programming relaxation or heuristic methods) since the results provided by the solver strongly depend on the starting guessed value of the decision variables, in particular of x.

References

[1] E. Amaldi, M. Bruglieri, and G. Casale. A two-phase relaxation-based heuristic for the maximum feasible subsystem problem. *Computers and Operations Research*, 35:1465–1482, 2008.

[2] B. Berger, J. Kleinberg, and T. Leighton. Reconstructing a three-dimensional model with arbitrary errors. *Journal of the ACM*, 46(2):212–235, 1999.

[3] C. Lavor, L. Liberti, and N. Maculan. *Computational experience with the molecular distance geometry problem.* Global Optimization: Scientific and Engineering Case Studies, Springer, Berlin„ 2006.

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

[5] J. Saxe. Embeddability of weighted graphs in k-space is strongly **NP**-hard. *Proceedings of 17th Allerton Conference in Communications, Control and Computing*, pages 480–489, 1979.