A least squares approach for the Discretizable Distance Geometry Problem with inexact distances

Douglas S. Gonçalves

Department of Mathematics
Universidade Federal de Santa Catarina

Distance Geometry Theory and Applications

DIMACS - New Jersey - July, 2016

Partially supported by CNPq.

Distance Geometry problem

Definition (DGP)

Given a simple weighted undirected graph $G(V, E, d)$, $d : E \to \mathbb{R}_+$, and a positive integer K, is there a map $x : V \to \mathbb{R}^K$ such that the constraints

$$\|x_i - x_j\|^2 = d_{ij}^2, \quad \forall \{i, j\} \in E$$

are satisfied ?

Discretizable Distance Geometry problem

Definition(DDGP)

A DGP is said discretizable if there exists a vertex order $\{v_1, v_2, \ldots, v_N\}$ ensuring that:

(a) $G[\{v_1, v_2, \ldots, v_K\}]$ is a clique;

(b) For each $i > K$:

i) $\{v_j, v_i\} \in E$, for $j = i - K, \ldots, i - 2, i - 1$,

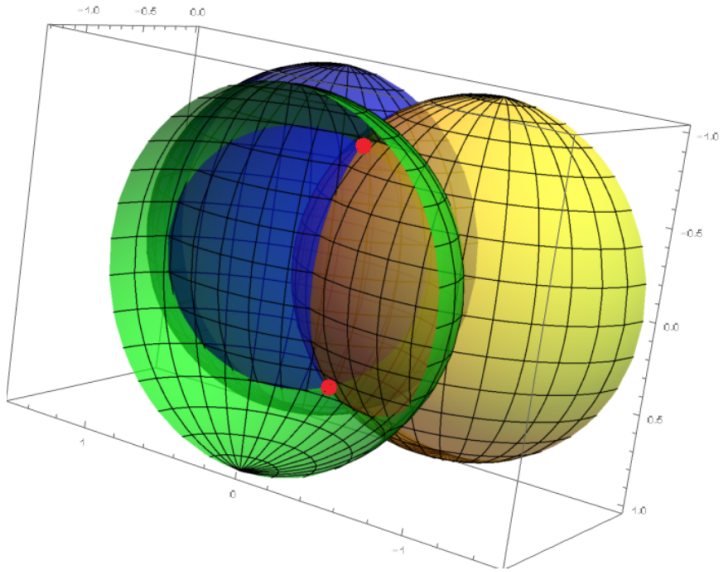
ii) $\mathbb{V}^2(\Delta(\{v_{i-K}, \ldots, v_{i-1}\})) > 0$.

* The definition ensures that the underlying graph is a chain of $(K + 1)$-cliques.

Exact distances: a branch-and-prune approach

By DDGP assumptions we have that coordinates x_i for each vertex v_i are obtained by intersecting K spheres:

$$\begin{aligned} \|x_{i-1} - x_i\|^2 &= d^2_{i-1,i} \\ \|x_{i-2} - x_i\|^2 &= d^2_{i-2,i} \\ &\vdots \\ \|x_{i-K} - x_i\|^2 &= d^2_{i-K,i} \end{aligned}$$



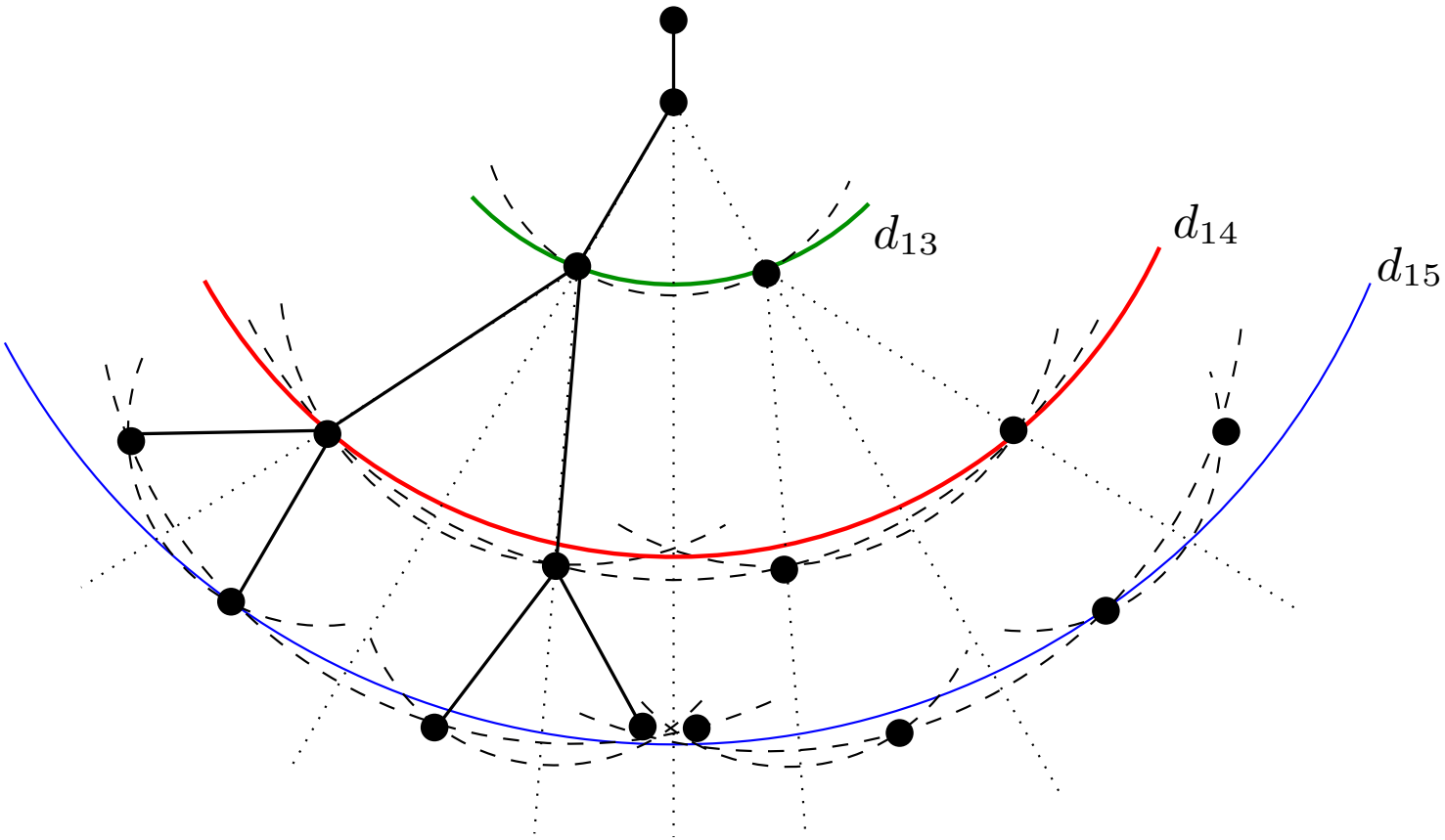
which leads to at most 2 candidate positions**(branching)**.

Pruning: Direct Distance Feasibility(DDF)

$$|\|x_h - x_i\| - d_{hi}| < \epsilon, \quad \forall h : \{h, i\} \in E \text{ and } h < i - K$$

(Lavor et al., **Comp. Optim. App., 52**, 2012)

Exact distances: search tree



(Liberti et al., **Discrete App. Math., 165**, 2014)

Exact distances: symmetries and other properties

- Search space has the structure of a binary tree (with 2^{N-K} leaf nodes)
- If pruning distances appear frequently enough it is possible to efficiently explore the search space
- The number of solutions is a power of 2
- Due to the symmetries in the DDGP search tree, it suffices to find the 1st solution: the others can be constructed by partial reflections

(Liberti et al., **SIAM Review, 56**, 2014)

DDGP with noisy distances

Consider that exact distances d_{ij}^2 are disturbed by a small noise δ_{ij}

$$\tilde{d}_{ij}^2 = d_{ij}^2 + \delta_{ij},$$

with $|\delta_{ij}| \leq \delta$, such that

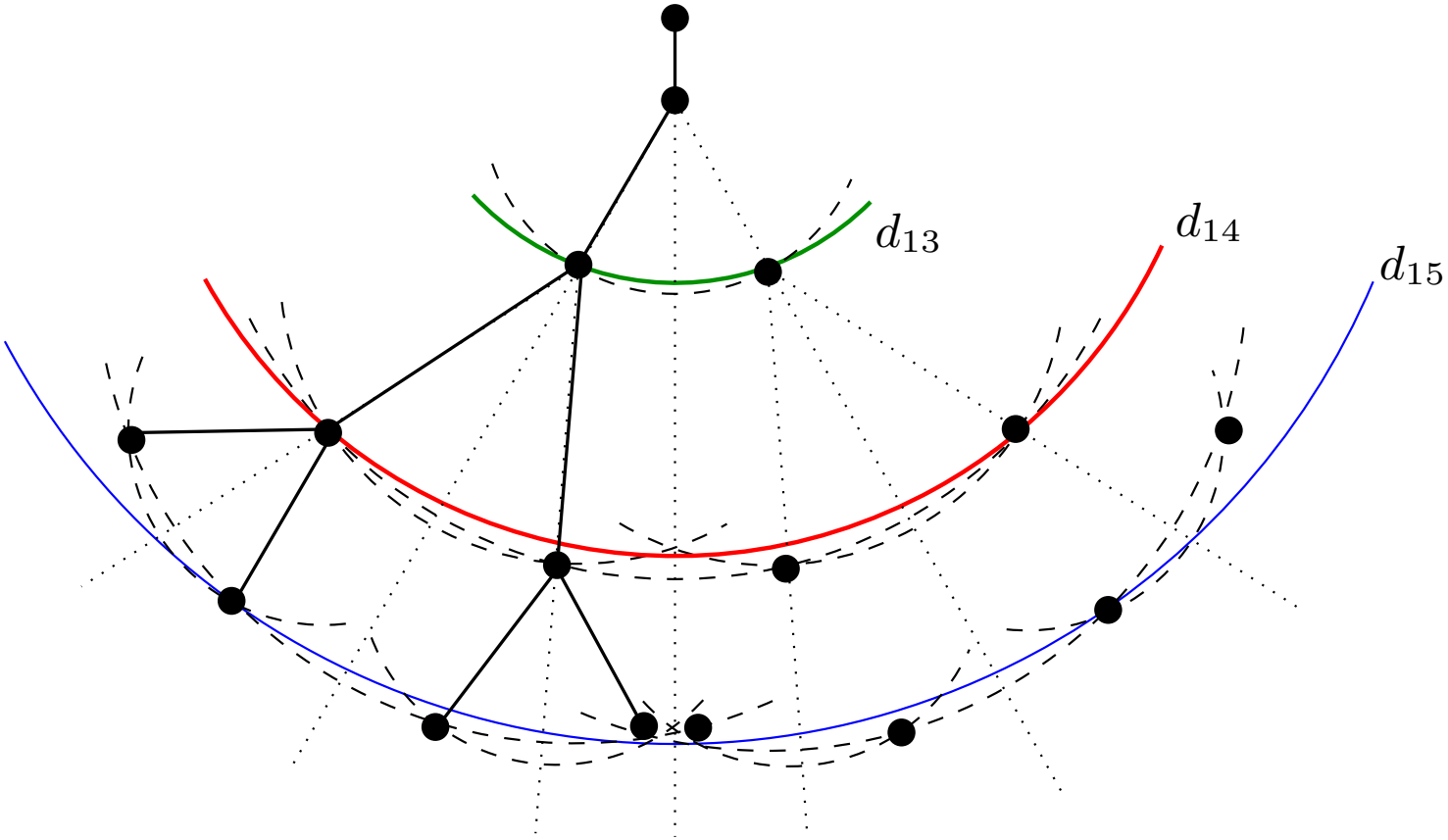
$$\|\delta \mathbf{d}\| \leq \sqrt{m}\,\delta.$$

Problem: find approximate solutions of

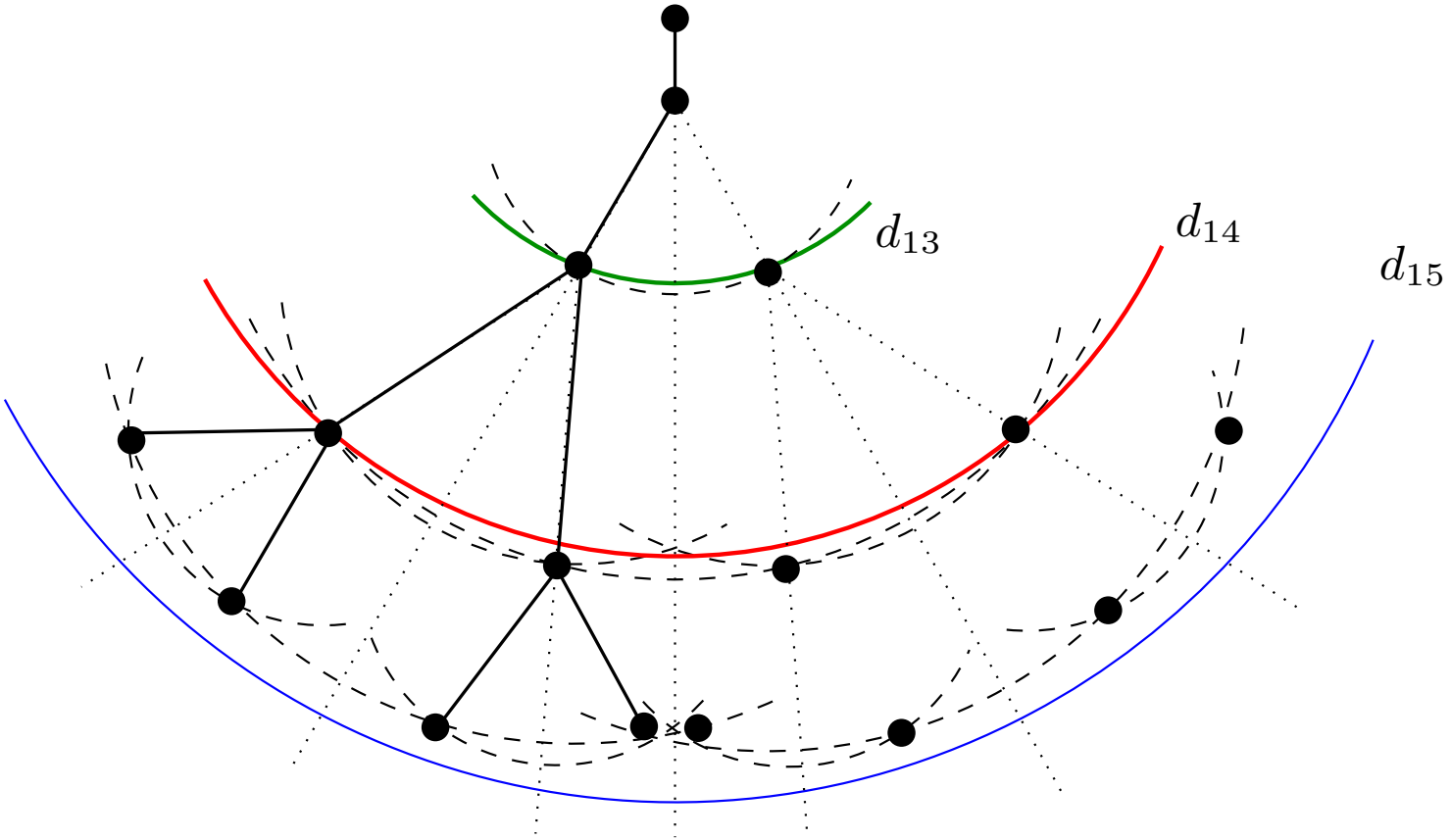
$$\|x_i - x_j\|^2 - \tilde{d}_{ij}^2 = 0, \quad \forall \{i,j\} \in E$$

Aim: extend the BP approach for DDGP with noisy distances

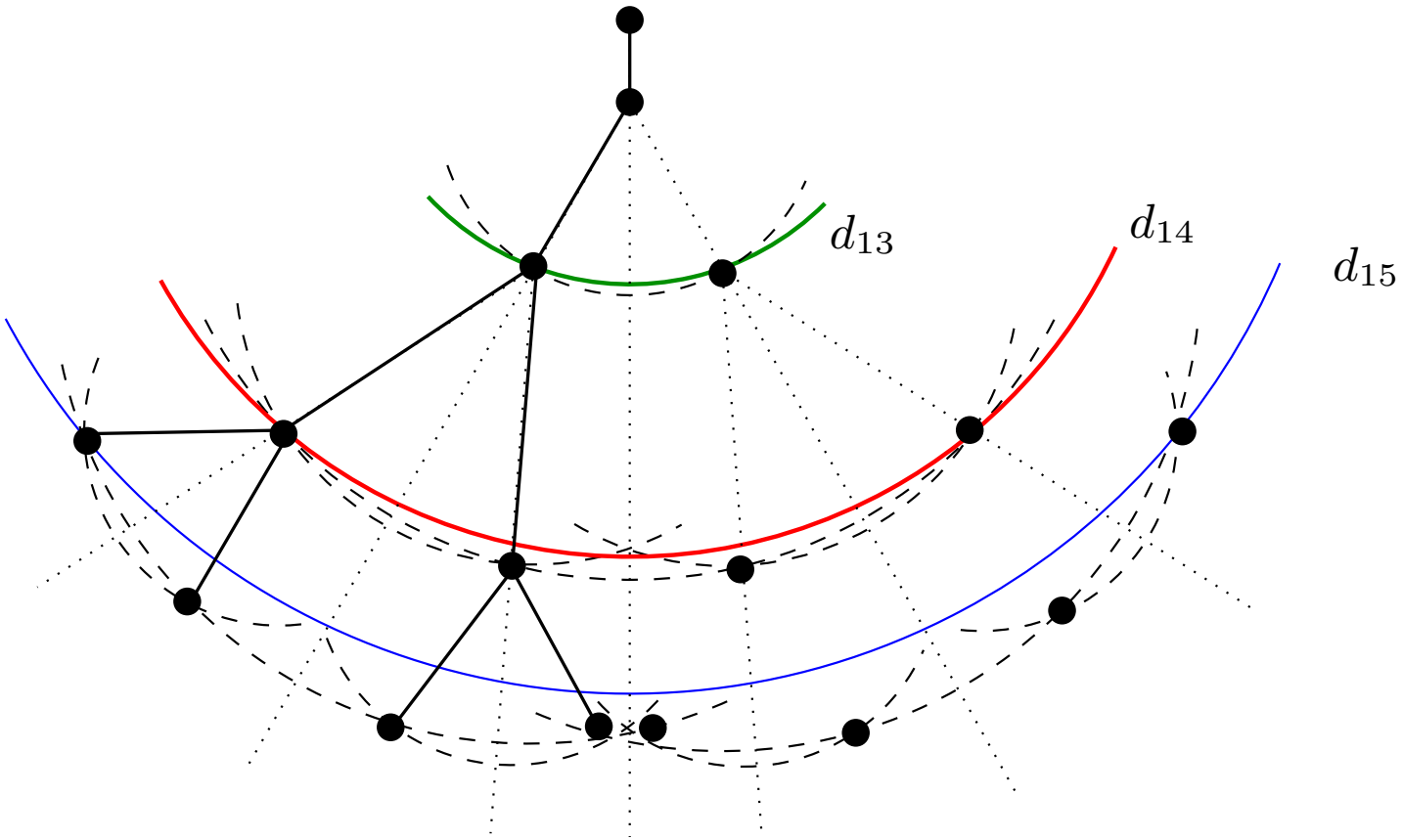
Noisy distances



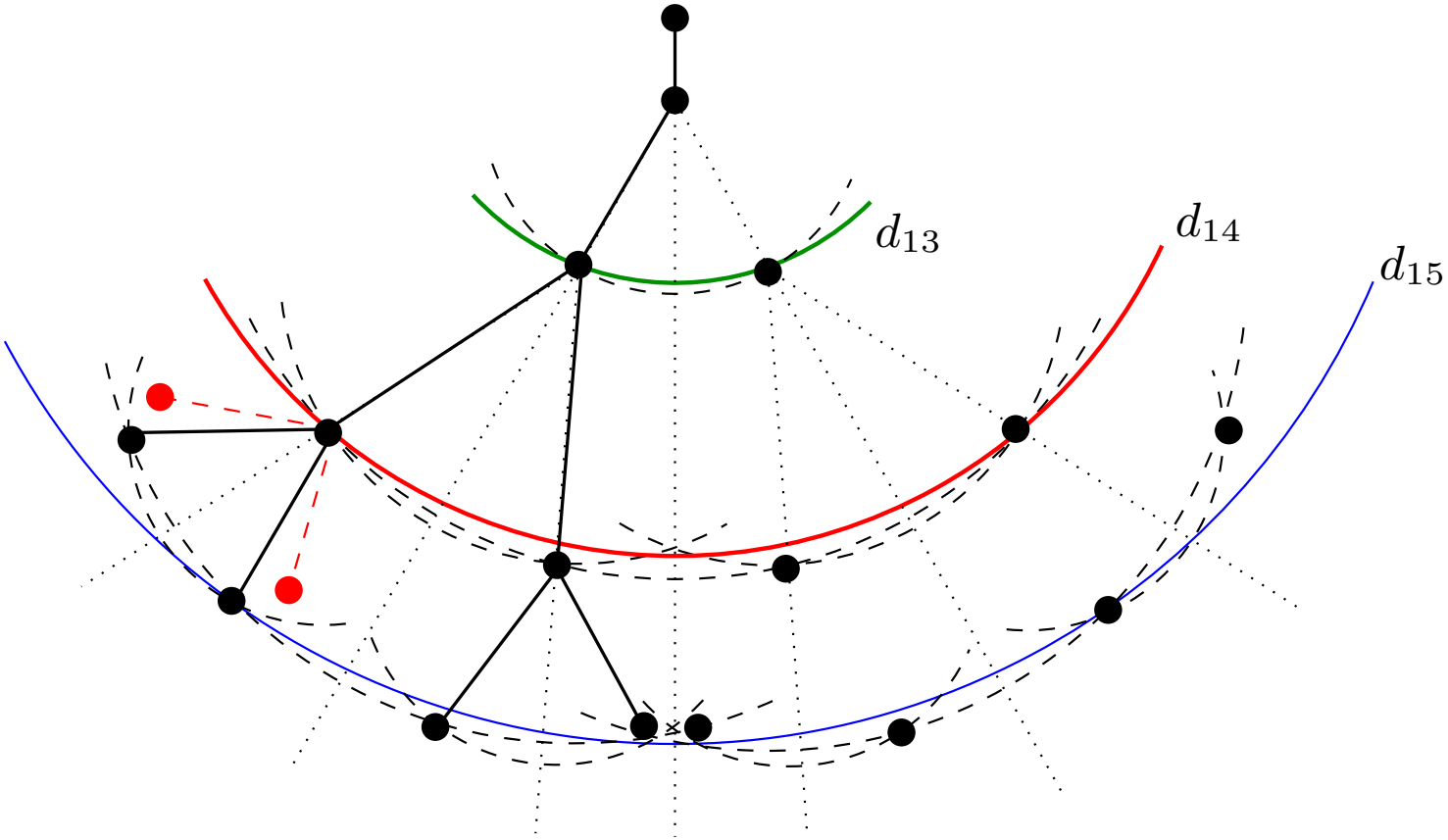
Noisy distances



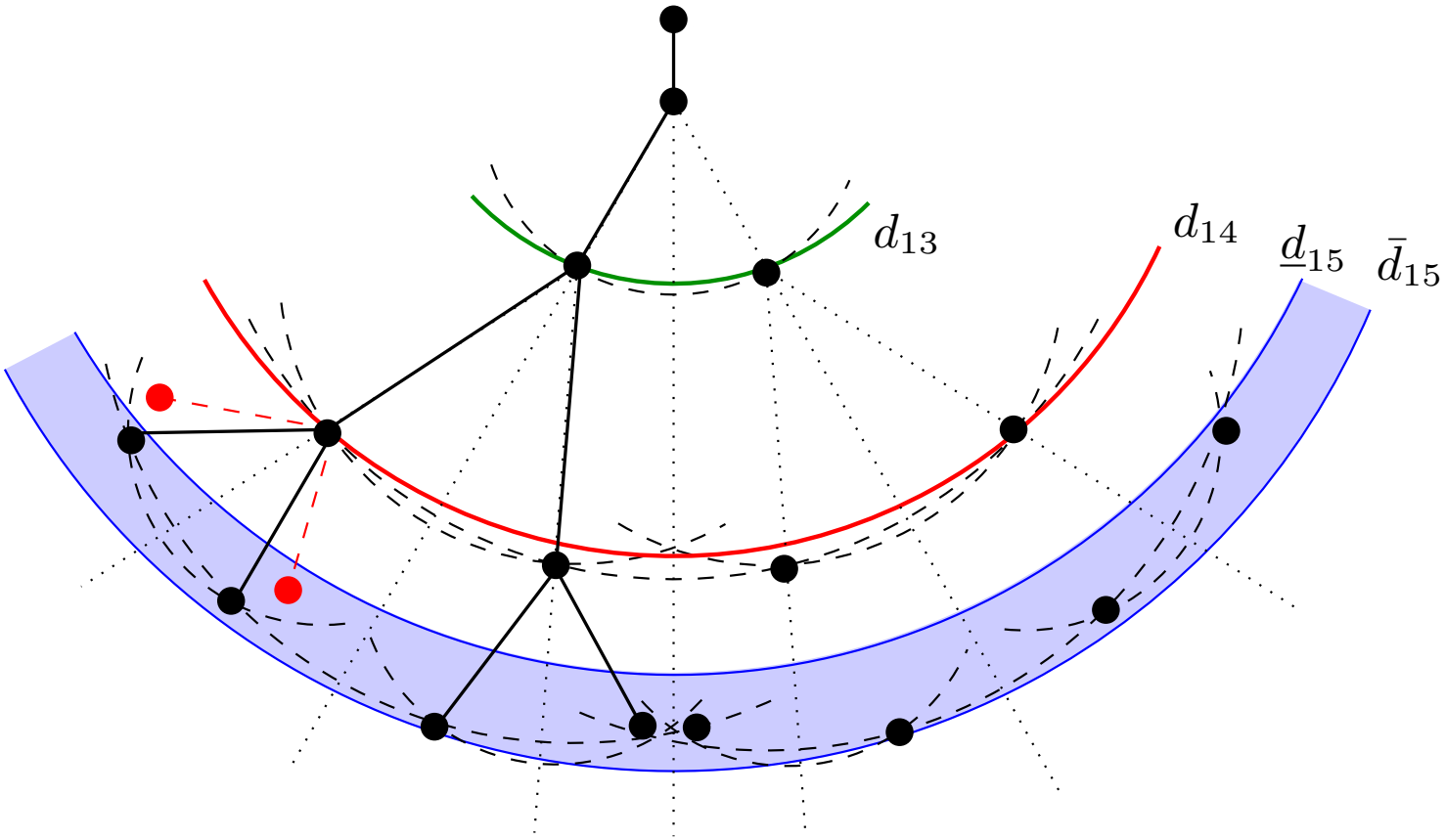
Noisy distances



Noisy distances



Noisy distances



Least-squares, SVD and candidate positions

Theorem (Low rank approximation)

If $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ are the nonzero singular values of $A \in \mathbb{R}^{n\times n}$ and $A = U\Sigma V^\top$, then for each $K < r$, the distance from A to the closest matrix of rank K is

$$\sigma_{K+1} = \min_{\mathit{rank}(B)=K} \|A - B\|_2,$$

achieved at $B = \sum_{i=1}^{K} \sigma_i u_i v_i^\top$.

Corollary:

$$\sum_{i=K+1}^{n} \sigma_i^2 = \min_{\mathsf{rank}(B)=K} \|A - B\|_F^2.$$

(Golub and Van Loan, **Matrix Computations**, 1996)

Candidate positions: 1st candidate

- $\tilde{D}_i$: reduced(complete) distance matrix related to $\{v_{i-K}, \ldots, v_{i-1}, v_i\}$
- $X_i \in \mathbb{R}^{K\times(K+1)}$, $X_i = [x_{i-K} \quad \ldots \quad x_{i-1} \quad x_i]$
- $H = I_n - \frac{1}{n}ee^\top$: centering matrix, $\tilde{G}_i = -\frac{1}{2}H\tilde{D}_iH$: Gram matrix

If $\tilde{G}_i = U\tilde{\Sigma}U^\top$, then

$$\bar{G}_i = \arg\min_{\text{rank}(G)=K} \|G - \tilde{G}_i\|_2 = \sum_{k=1}^{K} \tilde{\sigma}_k u_k u_k^\top,$$

and, since $\bar{G}_i = \bar{X}_i^\top \bar{X}_i$, candidate positions are given by:

$$\bar{X}_i = (\tilde{\Sigma}(1:K, 1:K))^{1/2}(U(:, 1:K))^\top$$

(Sit et al., **Bull. Math. Bio., 71**, 2009)

Orthogonal Procrustes

The first K vectors $X = [\bar{x}_{i-K} \quad \dots \quad \bar{x}_{i-1}]$ are used to transform the coordinates of $\bar{x}_i$ back to the original reference system: $Y = [x_{i-K} \quad \dots \quad x_{i-1}]$ (already placed)

After centering $X_c = X(I - \frac{1}{n}ee^\top)$, $Y_c = Y(I - \frac{1}{n}ee^\top)$, find Q such that

$$\min_{Q^\top Q = I} \|QX_c - Y_c\|_F^2.$$

Given $Y_c X_c^\top = U\Sigma V^\top$, we have $Q = UV^\top$

$$x_i' \leftarrow Q\bar{x}_i + \mathbf{t},$$

where $\mathbf{t} = \frac{1}{n}Ye - Q\frac{1}{n}Xe = y_c - Qx_c$.

(Dokmanic et al., **IEEE Signal Proces., 32**, 2015)

Reflection: 2nd candidate

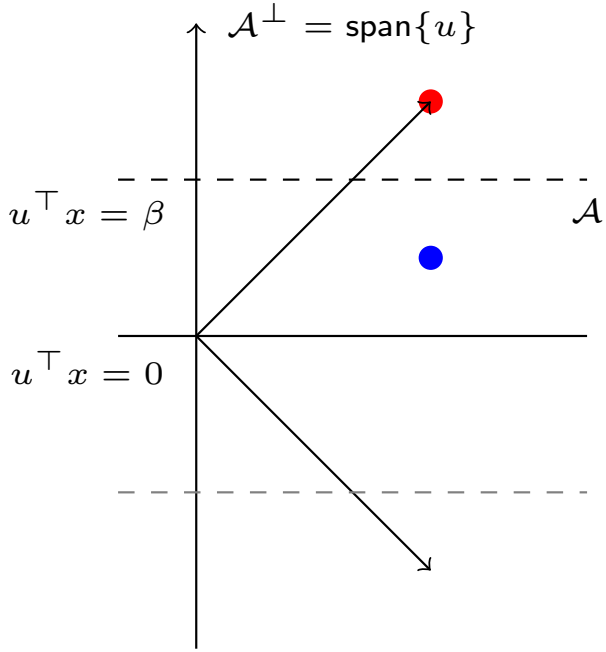
From the assumptions of DDGP, the set $\{x_{i-K}, \ldots, x_{i-1}\}$ is affinely independent, generating an affine subspace $\mathcal{A}$ of dimension $K-1$.

Let u be a unit vector orthogonal to $\mathcal{A}$. Then the points in $\mathcal{A}$ satisfy

$$u^\top x = \beta,$$

and the reflection of x_i through that hyperplane is given by

$$x_i'' = (I - 2uu^\top)x_i + 2\beta u$$



Consistency

Let D_i, $\tilde{D}_i$ and $\bar{D}_i$ be the true, disturbed and approximated reduced distance matrices, respectively, and G_i, $\tilde{G}_i$ and $\bar{G}_i$ their associated Gram matrix.

As

$$\|G_i - \tilde{G}_i\|_2 = \frac{1}{2}\|D_i - \tilde{D}_i\|_2 = \frac{1}{2}\|E_i\|_2 \leq \frac{1}{2}\|E_i\|_F \leq \frac{1}{2}\sqrt{\frac{n(n-1)}{2}}\delta,$$

we have that

$$\tilde{\sigma}_{K+1} = \|\bar{G}_i - \tilde{G}_i\|_2 \leq \|G_i - \tilde{G}_i\|_2 \leq \frac{1}{2}\sqrt{\frac{n(n-1)}{2}}\delta.$$

Therefore $\tilde{\sigma}_{K+1} \to 0$ as $\delta \to 0$, implying $\|\bar{G}_i - \tilde{G}_i\| \to 0$.

But when $\delta \to 0$, $\tilde{G}_i \to G_i$, thus $\bar{G}_i \to G_i$.

Pruning devices: DDF criterion

Direct Distance Feasibility: for all $j < i - K \ : \ \{j, i\} \in E$

$$\left| \|x_i - x_j\|^2 - \tilde{d}_{ij}^2 \right| \leq \varepsilon_1.$$

How to choose ε_1 ?

Let $\tilde{\mathbf{d}}$ be the vector with components $\tilde{d}_{ij}^2$. Choose ε_1 such that

$$\mathsf{MDE}(\mathbf{x}(\varepsilon_1); \tilde{\mathbf{d}}) \leq \tau \|\delta \mathbf{d}\|,$$

where $\tau \geq 1$, $\mathbf{x}(\varepsilon_1)$ is the first solution found by BP and

$$\mathsf{MDE}(\mathbf{x}; \mathbf{d}) = \frac{1}{|E|} \sum_{\{i,j\} \in E} \frac{|\|x_i - x_j\| - d_{ij}|}{d_{ij}}.$$

Rigidity and noisy distances

Let $\mathbf{x} \in \mathbb{R}^{KN}$ be a realization of $G(V, E)$, $R \in \mathbb{R}^{|E| \times KN}$ be the rigidity matrix of $(G, \mathbf{x})$ and $\tilde{\mathbf{x}}$ the solution of

$$\min_{\mathbf{x}} \frac{1}{2} \sum_{\{i,j\} \in E} \left(\|x_i - x_j\|^2 - \tilde{d}_{ij}^2 \right)^2.$$

Define $\delta\mathbf{x} = \tilde{\mathbf{x}} - \mathbf{x}$ and $\delta\mathbf{d}$ the vector with entries $\delta_{ij} = \tilde{d}_{ij}^2 - d_{ij}^2$. From the first order Taylor approximation, we have

$$R\delta\mathbf{x} = \frac{1}{2}\delta\mathbf{d}.$$

Thus

$$\delta\mathbf{x} = \frac{1}{2}R^{\dagger}\delta\mathbf{d}.$$

and

$$\|\delta\mathbf{x}\| = \frac{1}{2}\|R^{\dagger}\|\|\delta\mathbf{d}\| = \frac{1}{2\sigma_r}\|\delta\mathbf{d}\|.$$

(Anderson et al., **SIAM J. Discrete Math., 24**, 2010)

Pruning devices: a relaxed DDF criterion

Thus, for the solution $\tilde{\mathbf{x}}$ of the perturbed NLSP, we have

$$\begin{aligned}
\left|\|\tilde{x}_i - \tilde{x}_j\|^2 - \tilde{d}^2\right| &\approx \left|2(x_i - x_j)^\top(\delta x_i - \delta x_j) - \delta_{ij}\right| \\
&\leq 2\|x_i - x_j\|\|\delta x_i - \delta x_j\| + |\delta_{ij}| \\
&\leq 2\left(\max_{ij} d_{ij}\right) 2\|\delta\mathbf{x}\| + \delta \leq 2\left(\max_{ij} d_{ij}\right)\frac{\|\delta\mathbf{d}\|}{\sigma_r} + \delta \\
&\leq \left(2(\max_{ij} d_{ij})\,\frac{\sqrt{m}}{\sigma_r} + 1\right)\delta.
\end{aligned}$$

Therefore, we demand that the approximate solution $\bar{\mathbf{x}}$ satisfies:

$$\left|\|\bar{x}_i - \bar{x}_j\|^2 - \tilde{d}^2\right| \leq \overbrace{\gamma\left(2(\max_{ij} \tilde{d}_{ij})\,\sqrt{m}\,c_1 + 1\right)\delta}^{\approx\varepsilon_1},$$

where $\gamma > 1$ and c_1 is an estimate for $1/\sigma_r$.

Pruning devices: Singular value ratio

Let $\hat{D}_i$ be the matrix of square distances related to v_i and its predecessors(neighbors v_j of v_i such that $j < i$).

Missing entries of $\hat{D}_i$ are obtained from already computed positions x_j, $j < i$.

Let

$$\hat{G}_i = -\frac{1}{2} H \hat{D}_i H = U \Sigma V^\top$$

A wrong choice of previous candidate positions may forbids the distances in $\hat{D}_i$ to lead to a realization in $\mathbb{R}^K$. Thus, we consider the ratio

$$\rho = \frac{\sum_{k=1}^{K} \hat{\sigma}_k}{\sum_{k=1}^{n} \hat{\sigma}_k},$$

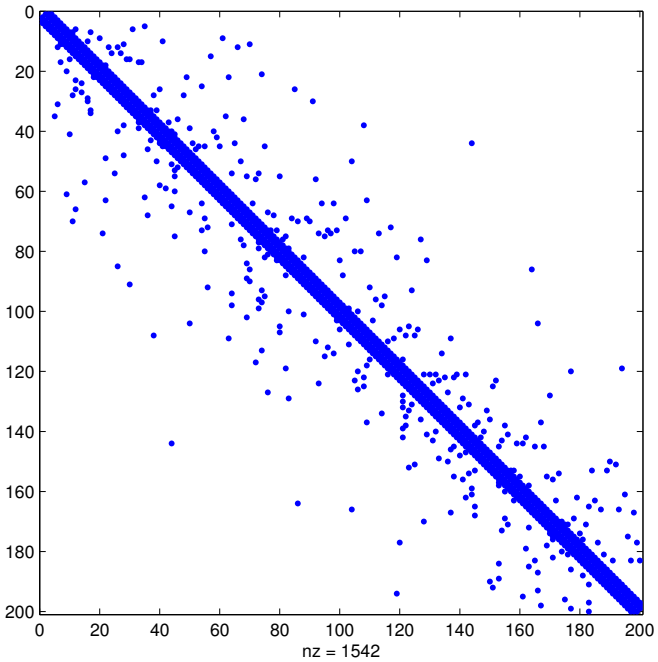
and the current tree path is pruned whenever: $(1 - \rho) > \varepsilon_2$.

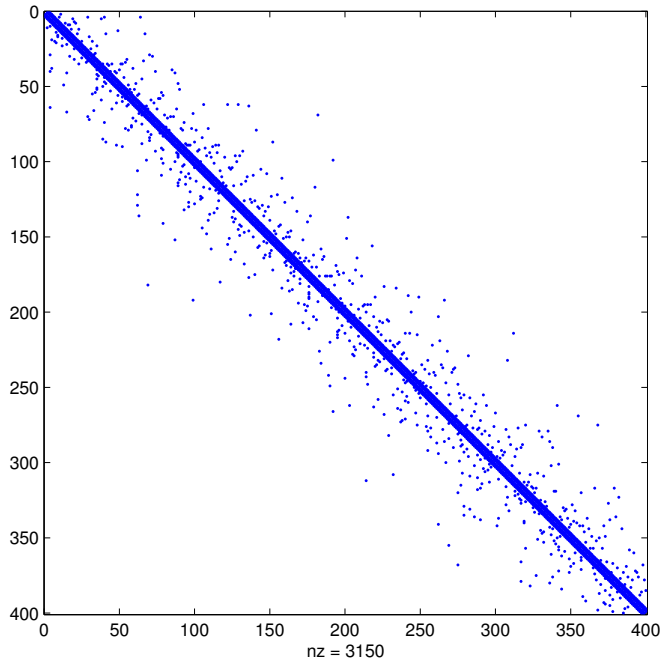
Algorithm

BP$(i, n, D, K, \varepsilon_1, \varepsilon_2, \delta)$
if $(i > n)$ **then**
  print current conformation *// one solution is found*
else
  if $(p_i > K)$ **then**
    Obtain $\hat{D}_i$ of order $p_i + 1$ and its SVD. If $(1 - \rho) > \varepsilon_2$, prune.
  end if
  // 1st candidate
  Set $\tilde{D}_i = \text{dist}(\{v_{i-K}, \ldots v_{i-1}, v_i\})$ and $\tilde{G}_i = -(1/2)H\tilde{D}_iH$;
  Find the $K \times (K+1)$ matrix $\bar{X}_i$ minimizing $\|X_i^\top X_i - \tilde{G}_i\|$;
  Transform $\bar{x}_i$ back to the original coordinate system: x_i'.
  if (x_i' is feasible) **then**
    BP$(i + 1, n, D, K, \varepsilon_1, \varepsilon_2, \delta)$
  end if
  // 2nd candidate
  Reflect x_i' around the hyperplane defined by $\{x_{i-K}, \ldots, x_{i-1}\}$: x_i''
  if (x_i'' is feasible) **then**
    BP$(i + 1, n, D, K, \varepsilon_1, \varepsilon_2, \delta)$
  end if
end if

Numerical experiments - I (Random points)

- Random points in $\mathbb{R}^3$ whose coordinates are drawn from $N(0, \Delta)$
- Discretization distances are kept
- At most one $\{j, i\} \in E$ with $j < i - 3$, for each i
- $\tilde{d}_{ij}^2 = d_{ij}^2 + \delta_{ij}$, where $|\delta_{ij}| < \delta$





Numerical experiments - I (Random points)

$\Delta = 10$			$\delta = 10^{-8}$					$\delta = 10^{-6}$					$\delta = 10^{-4}$				
$\|V\|$	$\|E\|$	dens.	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD
100	367	0.071	0.001	0.15	2	5e-10	2e-7	0.01	0.19	2	3e-8	1e-5	0.25	0.24	2	4e-6	1e-3
300	1171	0.026	0.001	1.92	4	1e-09	7e-7	0.05	2.00	4	2e-7	2e-4	0.50	2.28	4	7e-6	5e-3
500	1972	0.015	0.001	1.39	2	4e-10	5e-7	0.01	1.40	2	4e-8	3e-5	0.25	1.43	2	2e-6	1e-3
700	2765	0.011	0.001	8.80	4	3e-10	5e-7	0.01	9.07	4	4e-8	5e-5	0.50	9.16	4	4e-6	3e-3
900	3571	0.008	0.001	12.93	4	1e-09	1e-6	0.01	12.83	4	4e-8	6e-5	7.50	62.75	32	1e-5	1e-2

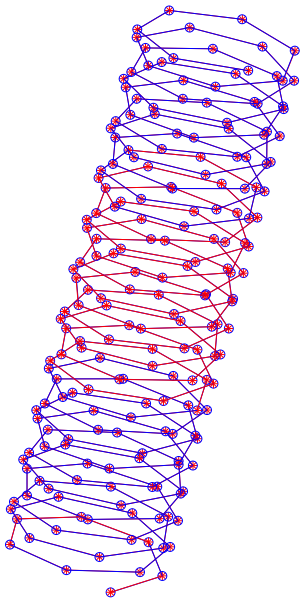
$\Delta = 1$			$\delta = 10^{-8}$					$\delta = 10^{-6}$					$\delta = 10^{-4}$				
$\|V\|$	$\|E\|$	dens.	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD
100	377	0.076	0.001	0.21	4	6e-8	1e-6	0.01	0.22	4	5e-6	2e-4	0.25	0.88	16	3e-4	1e-2
300	1170	0.026	0.001	0.39	2	6e-8	9e-6	0.01	0.40	2	4e-6	7e-4	0.20	1.50	16	2e-4	1e-2
500	1974	0.015	0.001	2.02	4	3e-8	3e-6	0.01	2.06	4	3e-6	3e-4	0.20	5.08	16	2e-4	1e-2
700	2774	0.011	0.001	2.30	2	4e-8	5e-6	0.02	3.12	4	6e-6	1e-3	0.65	32.10	256*	1e-3	7e-2
900	3575	0.008	0.002	3.34	2	3e-7	1e-4	0.02	3.35	2	7e-6	1e-3	1.20	18.77	256*	1e-3	2e-1

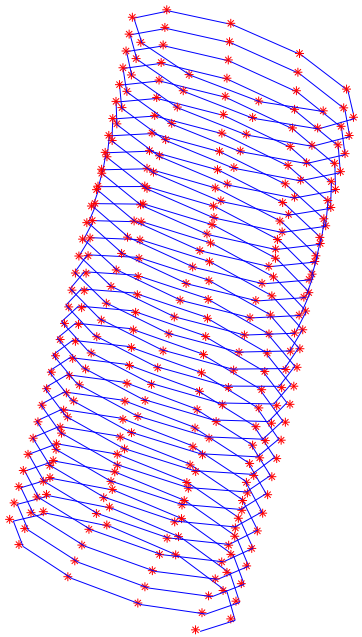
Numerical experiments - II (Helices)

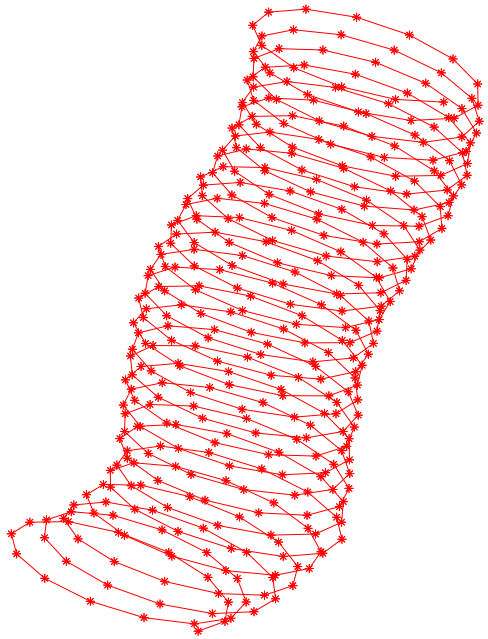
N points uniformly distributed over the helix:

$$\mathbf{x}(t) = 4\cos 3t\,\mathbf{i} + 4\sin 3t\,\mathbf{j} + 2t\,\mathbf{k}, \qquad t \in [0, 20\pi]$$

		$\delta = 10^{-6}$					$\delta = 10^{-4}$				
$\|V\|$	$\|E\|$	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD
100	370	0.001	0.31	4	7e-9	8e-7	0.01	0.30	4	7e-7	2e-4
200	769	0.01	1.47	2	3e-7	1e-4	1.00	1.95	2	4e-5	1e-3
300	1171	0.07	0.83	2	3e-6	2e-3	6.00	2.98	4	4e-4	1e-1
400	1567	0.25	1.31	4	2e-5	6e-3	25.00	31.45	256*	3e-3	0.98

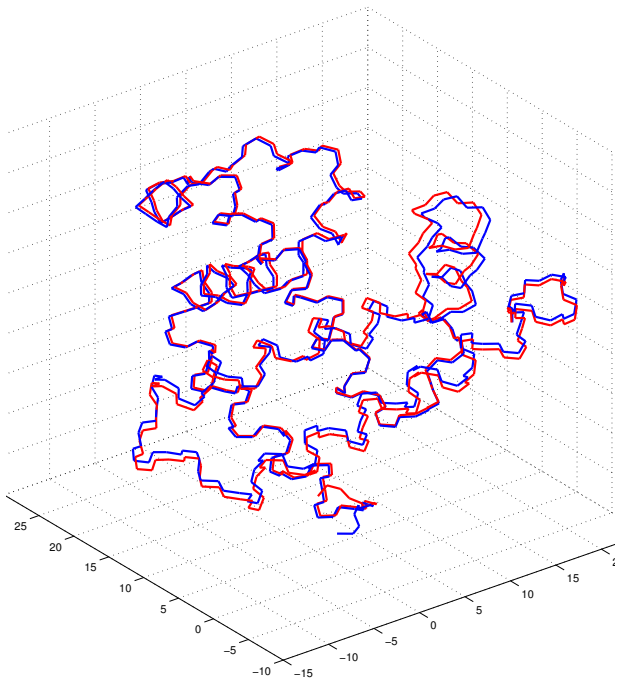






Numerical experiments - III (Small proteins)

- Artificial instances from PDB data
- Sequence of backbone atoms: N-C_α-C
- All distances among four consecutive atoms or distances < 6Å
- Random noise added to exact distances: $\tilde{d}_{ij}^2 = d_{ij}^2 + U[-\delta, \delta]$



			$\delta = 10^{-8}$					$\delta = 10^{-6}$					$\delta = 10^{-4}$				
PDB	$\|V\|$	dens.	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD	ε_1	t(s)	$\|S\|$	MDE	RMSD
2erl	122	0.10	0.001	0.09	2	4e-8	6e-6	0.02	0.15	2	8e-6	9e-4	0.30	0.41	8	2e-4	2e-2
1crn	138	0.09	0.001	0.16	2	1e-7	9e-6	0.02	0.16	2	1e-5	9e-4	0.12	2.33	8	4e-4	2e-2
1hoe	222	0.05	0.001	0.24	2	8e-8	2e-6	0.07	0.58	4	5e-5	1e-3	0.35	300*	18	6e-4	5e-2
1a70	291	0.04	0.003	0.37	2	1e-6	9e-5	0.04	12.08	8	6e-5	3e-3	9.99*	300*	-	-	-
1poa	354	0.03	0.001	0.44	2	2e-7	3e-5	0.11	2.67	2	3e-5	7e-3	9.99*	300*	-	-	-
1mbn	459	0.03	0.003	1.97	16	3e-7	4e-4	0.09	18.39	192	1e-5	8e-3	0.42	74.18	256*	1e-4	0.65

Final remarks and future works

- Extension of the BP approach to handle DDGP with noisy distances
- Approximate solutions can be obtained when the noise is small enough
- DDF is clearly sensitive to noise: it is difficult to set up the tolerance
- Long pruning distances should be treated differently
- The problem gets harder as the "condition number" of the rigidity matrix increases
- Additional pruning devices should be integrated for specific applications
- Devise a sharp bound for $\delta\mathbf{x}$ generated by BP

References


A. Sit, Z. Wu, Y. Yuan. *A geometric buildup algorithm for the solution of the distance geometry problem using least-squares approximation*. Bulletin of mathematical biology, 71, 1914-–1933, 2009.


B. D. O. Anderson, I. Shames, G. Mao, B. Fidan, *Formal theory of noisy sensor network localization*, SIAM J. Discrete Math., 24, 684–698, 2010.


C. Lavor, L. Liberti, N. Maculan, A. Mucherino, *The discretizable molecular distance geometry problem.* Computational Optimization and Applications, 52, 115–146, 2012.


L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean distance geometry and applications*. SIAM Review 56, 3-69, 2014.


I. Dokmanic, R. Parhizkar, J. Ranieri, M. Vetterli, *Euclidean Distance Matrices: Essential Theory, Algorithms and Applications*, IEEE Signal Processing Magazine, 32, 12–30, 2015.

References

- A. Sit, Z. Wu, Y. Yuan. *A geometric buildup algorithm for the solution of the distance geometry problem using least-squares approximation*. Bulletin of mathematical biology, 71, 1914-–1933, 2009.
- B. D. O. Anderson, I. Shames, G. Mao, B. Fidan, *Formal theory of noisy sensor network localization*, SIAM J. Discrete Math., 24, 684–698, 2010.
- C. Lavor, L. Liberti, N. Maculan, A. Mucherino, *The discretizable molecular distance geometry problem.* Computational Optimization and Applications, 52, 115–146, 2012.
- L. Liberti, C. Lavor, N. Maculan, A. Mucherino, *Euclidean distance geometry and applications*. SIAM Review 56, 3-69, 2014.
- I. Dokmanic, R. Parhizkar, J. Ranieri, M. Vetterli, *Euclidean Distance Matrices: Essential Theory, Algorithms and Applications*, IEEE Signal Processing Magazine, 32, 12–30, 2015.

Thanks for your attention!