Distance Geometry-Matrix Completion

Christos Konaxis

Algs in Struct BioInfo 2010

Outline

Measure diffrence of matched sets

Def. Root Mean Square Deviation

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|^2},$$

where $x_i, y_i \in \mathbb{R}^3$ are (C_α) atom coordinates in SAME coordinate frame.

$X = [x_1, \ldots, x_n]$, $Y = [y_1, \ldots, y_n]$, then

$$RMSD(X, Y) = \frac{1}{\sqrt{n}}|X - Y|_F, \text{ where } |M|_F^2 = \sum_{i,j} M_{ij}^2 = \mathrm{tr}(M^T M),$$

is the Frobenious metric, and $\mathrm{tr}(A) = \sum_i A_{ii}$ is the trace of matrix $A = [A_{ij}]$.

Optimal alignment of matched sets

Translate to common origin by subtracting centroid

$$x_c = \frac{1}{n} \sum_i x_i.$$

Rotate to optimal alignment by rotation matrix $Q : Q^T = I$.
Also should have $\det Q = 1$.
Exists deterministic linear algebra algorithm [Kabsch].

Overall cost= $O(n^3)$, but least-squares approximation in $O(n)$.

Optimal rotation

Assume common centroid:

$$RMSD(X, Y) = \min_Q |Y - XQ|_F, \quad Q^T Q = I_3.$$

$$|Y - XQ|_F = \mathrm{tr}(Y^T Y) + \mathrm{tr}(X^T X) - 2\mathrm{tr}(Q^T X^T Y),$$

so must maximize $\mathrm{tr}(Q^T X^T Y)$.

Consider SVD:

$$X^T Y = U\Sigma V^T, \quad U^T U = V^T V = I, \ \Sigma = \mathrm{diag}[s_i] : s_1 \geq \cdots \geq s_3.$$

Then $\mathrm{tr}(Q^T X^T Y) = \mathrm{tr}(Q^T U\Sigma V^T) = \mathrm{tr}(V^T Q^T U\Sigma) \leq \mathrm{tr}(\Sigma)$,

because $T = V^T Q^T U$ orthonormal $\Rightarrow |T_{ij}| \leq 1$. Hence maximum at

$$T = I \Leftrightarrow Q = UV^T.$$

If $\det T = -1$ then just negate T_{33}.

Introduction

- Nuclear Magnetic Resonance (NMR) and Nuclear Overhauser Effect (NOE) spectroscopy provide approximate inter-atomic distances for molecular structures as large as 5.000 atoms.
- The distances measured by NMR and NOESY experiments (usually a small subset of all possible pairs) must be converted into a 3D structure consistent with the measurements.
- In general the distances are imprecisely measured: for each distance d_{ij} we have $l_{ij} \leq d_{ij} \leq u_{ij}$.

- The Distance Geometry Method is based on the foundational work of Cayley (1841) and Menger (1928) who showed how convexity and other basic geometric properties could be defined in terms of distances between pairs of points.
- The problem can be reduced to the completion of a partial matrix M satisfying certain prorperties.

Distance matrices

- **Definition (and structure)**: A distance matrix D is square, $D_{ii} = 0$, $D_{ij} = D_{ji} \geq 0$.
- **Definition.** A distance matrix D is euclidean and embeddable in $\mathbb{R}^k$ iff

$$\exists \text{ points } p_i \in \mathbb{R}^k \ : \ D_{ij} = \frac{1}{2} \text{ dist}(p_i - p_j)^2.$$

Embeddable matrices in $\mathbb{R}^3$ correspond to 3D conformations.

Embedding matrices in $\mathbb{R}^k$

- **Thm** [Schoenberg'35,Blumenthal'53] Take border (Cayley-Menger) matrix

$$B = \begin{bmatrix} 0 & 1 \cdots 1 \\ 1 & \\ \vdots & D \\ 1 & \end{bmatrix}.$$

  Then, D embeds in $\mathbb{R}^k$ iff $\text{rank}(B) \leq k+2$.

- **Cor.** A distance matrix D expresses a 3D conformation iff $\text{rank}(B) = 5$

Cyclohexane's distance matrix

$$
\begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \end{array}
\begin{array}{c}
\begin{array}{ccccccc} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \end{array} \\
\left[\begin{array}{ccccccc}
0 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & u & c & x_{14} & c & u \\
1 & u & 0 & u & c & x_{25} & c \\
1 & c & u & 0 & u & c & x_{36} \\
1 & x_{14} & c & u & 0 & u & c \\
1 & c & x_{25} & c & u & 0 & u \\
1 & u & c & x_{36} & c & u & 0
\end{array}\right]
\end{array}
$$

Known: $u \simeq 1.526$ (adjacent), $\phi \simeq 110.4^{\circ} \Rightarrow c \simeq 2.285$ (triangle).

Rank condition $(= 5)$ equivalent to the vanishing of all 6×6 minors.
This yields a 3×3 system of quadratic polynomials in the x_{14}, x_{25}, x_{36}.

If all c, u same, then 2 isolated conformations, one 1-dim set.

If the c, u perturbed, then ≤ 16 solutions $\in \mathbb{R}$ [Emiris-Mourrain].

Points from distances

- Given distance matrix D, compute coordinate matrix X.

$$D = \begin{pmatrix} 0 & |v_1|^2 & |v_2|^2 & \dots & |v_n|^2 \\ |v_1|^2 & 0 & |v_1 - v_2|^2 & \dots & |v_1 - v_n|^2 \\ |v_2|^2 & |v_2 - v_1|^2 & 0 & \dots & |v_2 - v_n|^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ |v_n|^2 & |v_n - v_1|^2 & |v_n - v_2|^2 & \dots & 0 \end{pmatrix}$$

- Find Gram matrix $G = \begin{pmatrix} v_1 v_1 & v_1 v_2 & \dots & v_1 v_n \\ v_2 v_1 & v_2 v_2 & \dots & v_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n v_1 & v_n v_2 & \dots & v_n v_n \end{pmatrix}$
- Elements of G computed using: $2v_i v_j = |v_i|^2 + |v_j|^2 - |v_i - v_j|^2$, or
  1. Subtract the first row of D from each row.
  2. Subtract the first column from each column.
  3. Delete the first row and column. Result: $-2G$.

Points from distances

- Given G, find $n \times 3$ matrix X s.t. $G = X^T X$ using eigenvectors matrix V (s.t. $V^T V = I$), and eigenvalues diagonal matrix E.
- Then, $GV = EV$. Since G is symmetric and comes from a 3D distance matrix, it has 3 non-zero real eigenvalues with real eigenvectors.
- Construct a diagonal matrix $\sqrt{E}$ whose entries on the diagonal are the square roots of the entries of E.
- Then, $G = V\sqrt{E}\sqrt{E}V^T = X^T X$, where $X := \sqrt{E}V^T$.

Positive (semi)definite matrices

Def. An $n \times n$ real matrix M is positive (semi)definite if

$$x^T M x > 0 \quad (x^T M x \geq 0) \quad \forall\, x \neq 0.$$

Denoted $M \succ 0$, $M \succeq 0$.

Examples: $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ $\begin{bmatrix} 3 & -1 & -2 \\ -2 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$

Lem. [Sylvester].
$M \succ 0 \Leftrightarrow \det M_i > 0, \forall\ i \times i$ upper-left minor $M_i,\ i = 1, \ldots, n$.

Cor. $M \succ 0 \Rightarrow \det M > 0$.

If $M \succ 0$ $(M \succeq 0)$, then all its eigenvalues are positive (non-negative).

Embeding via rank

Thm. $\{p_i\}$ embed in $\mathbb{R}^3$ (and not $\mathbb{R}^2$) iff $D \succeq 0$ & $\mathrm{rk}D = 3$.

[$\Rightarrow$]
$\forall y \in \mathbb{R}^n$, $y^T P P^T y = (y^T P)(y^T P)^T \geq 0$: positive semidefinite.
$\mathrm{rk}(AB) = \min\{\mathrm{rk}A, \mathrm{rk}B\} \Rightarrow \mathrm{rk}D = \mathrm{rk}P$.
embed $\Rightarrow \exists p_a, p_b, p_c$ linearly independent $\Rightarrow \mathrm{rk}P = 3$.

[$\Leftarrow$]
Singular Value Decomposition (SVD): symmetric real $D = USU^T$,
where $U^T U = UU^T = I$, $S = \mathrm{diag}[s_1, \ldots, s_n]$, $s_i = |\text{eigenvalue}|^2 \geq 0$.

$\mathrm{rk}D = 3 \Rightarrow S = [s_1, s_2, s_3, 0, \cdots]$, $\sqrt{S} = [\sqrt{s_i}]$, $D = U\sqrt{S} \cdot \sqrt{S}U^T$.
$P := U\sqrt{S}$ is $n \times 3$: defines n points $p_i \in \mathbb{R}^3$.

Bound smoothing

- For any three points i, j, k in $\mathbb{R}^3$ the triangle inequality holds:

$$|d_{ik} - d_{jk}| \leq d_{ij} \leq d_{ik} + d_{jk}.$$

- Meausered distances:

$$l_{ij} \leq d_{ij} \leq u_{ij}$$
$$l_{ik} \leq d_{ik} \leq u_{ik}$$
$$l_{jk} \leq d_{jk} \leq u_{jk}$$

- Improved upper bound: $\bar{u}_{ij} = min\{u_{ij}, u_{ik} + u_{jk}\}$
- Improved lower bound: $\bar{l}_{ij} = max\{l_{ij}, l_{ik} - u_{jk}, l_{jk} - u_{ik}\}$

- The tightened upper bounds can be computed independently of the lower.
- The tightened upper bounds further improve the tightened lower bounds: $\bar{u}_{ik} \leq u_{ik}$, $\bar{u}_{jk} \leq u_{jk}$, $\bar{l}_{ij} = max\{l_{ij}, l_{ik} - u_{jk}, l_{jk} - u_{ik}\}$, so $\bar{l}_{ij} = max\{l_{ij}, l_{ik} - \bar{u}_{jk}, l_{jk} - \bar{u}_{ik}\}$
- If $\bar{l}_{ij} > \bar{u}_{ij}$ (e.g. when an upper bound is too low) we have a triangle inequality violation.

Matrix completion problems

- **Problem:** Given a partial matrix M, can M be completed to a positive semidefinite matrix (PSD), positive definite matrix (PD), Euclidean distance matrix (EDM)?

- *EDM* is reduced to *PSD*:
  If $D = (d_{ij})$, $d_{ii} = 0$, is a symmetric $n \times n$ matrix, define $(n-1) \times (n-1)$ symmetric matrix $X = (x_{ij})$, where

$$x_{ij} := \frac{1}{2}(d_{in} + d_{jn} - d_{ij}), \quad \forall i, j = 1, \ldots, n-1.$$

  D is a distance matrix iff X is positive semidefinite.
  Moreover $v_i \in \mathbb{R}^k$ iff $rank(X) \leq k$.

- It is not known if PSD is in NP.
- PSD can be solved with an arbitrary precision in polynomial time (interior point, elipsoid method).
- We will examine polynomial instances of PSD.
- If a matrix contains a fully determined line or column then its completion problem reduces to the completion of a smaller matrix.

- Assumptions: M = Hermitian ($M^* = M$),
  all diagonal entries of M are specified (for positive semidefinite matrices), moreover they are equal to zero (for distance matrices).
- If m_{ij} is specified, then $m_{ji} = m_{ij}^*$ is also specified.
- For every $n \times n$ matrix M we define the graph (pattern of M) $G = ([1, n], E)$ with vertices $[1, \ldots, n]$. There is an edge ij between vertices i, j if entry m_{ij} is specified.

Partial PSD matrices

- **Def.** A matrix M is partial positive semidefinite (partial-PSD) if every principal specified submatrix of M is positive semidefinite.
- **Lem.** If the incomplete matrix M has a PSD (PD, distance matrix) completion, then M is partial-PSD (partial-PD, partial-distance matrix).
- Hence we have necessary (but not sufficient) conditions for the existance of a completion of M.
  Counter-example: Partial-PSD but $\nexists$ PSD-completion:

$$\begin{bmatrix} 1 & 1 & ? & 0 \\ & 1 & 1 & ? \\ & & 1 & 1 \\ & & & 1 \end{bmatrix}$$

Chordal graphs

- A chord of a circuit C of a graph G is an edge of G which is not in C but which joins two vertices of C. A graph is chordal if every circuit of length ≥ 4 has at least one chord.
- Th. Every partial positive semidefinite matrix M with pattern G has a positive semidefinite completion iff G is chordal.
- Constructive proof; can be turned into a polynomial time algorithm.
- Th. PSD can be solved in polynomial time if the matrix has a chordal pattern. (bit model).