Euclidean Distance Matrix Corruption-Reversal Algorithm

Brian Chen [1] Yihan Shen [1] Andrew Yang [2] Kevin Zhang [1]

[1]Department of Computer Science, Columbia University

[2]Department of Applied Physics, Columbia University

Introduction

Distance matrices are ubiquitous in fields ranging from crystallography in applied physics to sensor network localization. For example, when calculating the positions of atoms in a crystalline structure, X-ray diffraction [1] techniques can yield a pair distribution function [2] that gives information on the pairwise distance between atoms, but experimental error can lead to noise in the distance measurements [3]. Similarly, one can consider a sensor network problem, where the distances between two sensors can only be measured if the two sensors are within a certain distance of each other. Such limitations can manifest as random noise corruption, masking of large entries, and missing distance entries altogether.

Embeddability Conditions

Theorem 1 The Gram matrix $G = -\frac{1}{2}HDH$ is symmetric PSD iff D is a distance matrix. The centering matrix is $H = I - (1/n)\mathbb{1}\mathbb{1}^T$.

Proposition 1 Let α_i be the i-th (zero-indexed) largest eigenvalue of D and β_i be the i-th smallest eigenvalue of G. For $i > 0$, $\alpha_i + 2\beta_i < 0$.

Lemma 1 A matrix D is conditionally negative definite iff D is a distance matrix.

Proposition 2 A matrix M that is conditionally negative definite and has all entries $m_{ij} > 0$ satisfies $M^{\circ p}$ PSD for any $p < 0$.

Theorem 2 Let λ be the second largest eigenvalue of a distance matrix D. For $0 < \epsilon \leq \lambda$, $(D + \epsilon I)^{\circ p}$ is PSD for all $p < 0$. *Proof*: For all $p < 0$, a positive CND matrix A satisfies $A^{\circ p}$ is PSD [4].

Theorem 3 If D is a distance matrix, $D^{\circ 1/p}$ is PSD for $p > 1$.

Theorem 4 Given a matrix D embeddable in $\mathbb{R}^k$, define the $n-1 \times n-1$ matrix D' element-wise by $d'_{ij} = -(d_{i,j} + d_{i+1,j+1} - d_{i,j+1} - d_{i+1,j})$. D is a distance matrix iff D' is PSD. Furthermore, $\text{rank}(D), \text{rank}(D') \leq k + 2$.

Theorem 5 Given D, ϵ, p, the matrices G, $(D + \epsilon I)^{\circ p}$, $D^{\circ 1/p}$, and D' can all be computed in $O(n^2)$.

Embeddability Preservation

Theorem 11 Given a n-point dataset $x_i \in \mathbb{R}^k$ and another n-point dataset $y_i \in \mathbb{R}^k$, let $\delta_i = y_i - x_i$. Define $T \equiv D_y - D_x$ to be the distortion between distance matrices. Let G_y be the Gram matrix computed with D_y and $\rho(G_y)$ its spectral radius.

If we only have access to D_y and know $\delta = \max_i ||\delta_i||_2$, then we can bound $\max\{-4R^2, 4\delta^2 - 8\delta R\} \leq t_{ij} \leq 4\delta^2 + 8R\delta$, where $R = \sqrt{k(\rho(G_y) + \delta^2)/2}$.

Theorem 12 Let A be a PSD matrix. Let the matrix N^{ij} for $i \neq j$ be defined by $[N^{ij}]_{k\ell} = \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}$. Then, $A + \delta N^{ij}$ is PSD if $0 \leq \delta \leq \lambda_0/(2n-2)$, where λ_0 is the smallest eigenvalue of A.

Theorem 13 Let the matrix N^{ij} for $i \neq j$ be defined by $[N^{ij}]_{k\ell} = \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}$. Then, given a distance matrix D and its Gram matrix G, $D + \delta N^{ij}$ is Euclidean embeddable if $-2\lambda_1 \leq \delta \leq (2n/(2-n))\lambda_1$, where λ_1 is the second smallest eigenvalue of G.

Theorem 14 Define $M_D^i \in \mathbb{R}^{n\times n}$ by $[M_D^i]_{jk} = \delta_{ij}(1-\delta_{jk})m^2 + \delta_{ik}(1-\delta_{jk})m^2$. Define $M_G^i \in \mathbb{R}^{n\times n}$ by $[M_G^i]_{jk} = \delta_{ij}\delta_{ik}(m(1-1/n))^2 + (m/n)^2\delta_{jk}(1-\delta_{ij}\delta_{ik})$. Let $D \in \mathbb{R}^{n\times n}$ be a distance matrix and $G = -(1/2)HDH$ be the corresponding Gram matrix. If $\text{rank}(G) < n$, then $G + M_G^i = -(1/2)H(D + M_D^i)H$, and $\text{rank}(G + M_G^i) = \text{rank}(G) + 1$.

Problem Setup

Consider an $n \times n$ Euclidean distance matrix D embeddable in $\mathbb{R}^k$. Suppose a set of corruptions is applied to D. We are interested in the following questions:

1. Is it possible to identify the location of the corruption in D' without having to perform eigendecomposition? Are there algorithms faster than $O(n^3)$?
2. Once we have identified the location of the corrupted distance(s), how can we find D?
3. What kinds of corrputions to a distance matrix can still maintain Euclidean embeddability?

Definition 1 Given n distinct points $x_i \in \mathbb{R}^k$, define the distance matrix D entry-wise as $d_{ij} = ||x_i - x_j||_2^2$. We say D is embeddable in $\mathbb{R}^{k'}$ for $k' \leq k$ if there exists some $y_i \in \mathbb{R}^{k'}$ s.t. $d_{ij} = ||y_i - y_j||_2^2$.

Single-Entry Gaussian Noise Correction

Theorem 6. Let X and y be defined as above and suppose D is an EDM. The embedding dimension of D and X are equal if and only if $2D_{1n} = y^T X^+ y$.

Theorem 7. Suppose X is a $n \times n$ PSD matrix and let $v_1, \ldots, v_k$ be the k eigenvalues with non-zero eigenvalue. Then, $XX^+y = y$ if and only if $y \in \text{span}(v_1, \ldots, v_k)$.

Theorem 8. Let X and y be defined as above. Then, D is an EDM with the same embedding dimension as X if and only if 1) X is PSD, 2) $y^T X^+ y = 2D_{1n}$, 3) $y = Vw$, where $w \in \mathbb{R}^k$ and $V \in \mathbb{R}^{(n-1)\times k}$ has columns which are the k non-zero eigenvectors of X. Moreover, $D_{in} = D_{ni} = D_{1i} + \frac{1}{2}y^T X^+ y - y_i$.

Definition 2. Given $X \in \mathbb{R}^{(n-1)\times(n-1)}$, $y \in \mathbb{R}^{n-1}$ and $D_{1n} = d \in \mathbb{R}$ satisfying the conditions in Lemma 5, define $\text{Gram}(X, y, d)$ to be the Gram matrix

$$\text{Gram}(X, y, d) = \begin{pmatrix} X & y \\ y^T & 2d \end{pmatrix},$$

and $\text{EDM}(X, y, d)$ to be the EDM D satisfying $-L^T DL = \text{Gram}(X, y, d)$.

Theorem 9. Let D be an underlying $n \times n$ EDM and X be defined as above for D. Suppose $\tilde{D}$ is a noisy realization of D where the n-th row and column are corrupted by some additive Gaussian noise $\mathcal{N}(0, \sigma^2)$ with fixed variance.

Let (λ_i, v_i) be non-zero eigenvalue-vector pairs of X. Define $\Gamma = \text{Diag}(\sqrt{2\lambda_1}, \ldots, \sqrt{2\lambda_k})$ and $M = V\Gamma$. Furthermore, define $b \in \mathbb{R}^{(n-1)}$ where $b_i = \tilde{D}_{1i} - \tilde{D}_{in}$. Then, the maximum likelihood estimate of D is $\text{EDM}(X, Mz, \|z\|^2)$, where

$$z = \arg\min_{x \in \mathbb{R}^k} \left\| b + \|x\|^2 \mathbf{1}_{n-1} - Mx \right\|^2$$

Theorem 10. Let $\tilde{D}$ be a noisy realization of an underlying EDM D corrupted in the same manner as in Theorem 7. Suppose we know exactly the distance between point 1 and n (i.e. $\tilde{D}_{1n} = D_{1n}$). Then, the maximum likelihood estimate of D is equal to $\text{EDM}(X, Mz, \|z\|^2)$, where

$$z = (\Gamma^2 + \lambda I)^{-1} M^T c,$$

and λ is the unique solution to

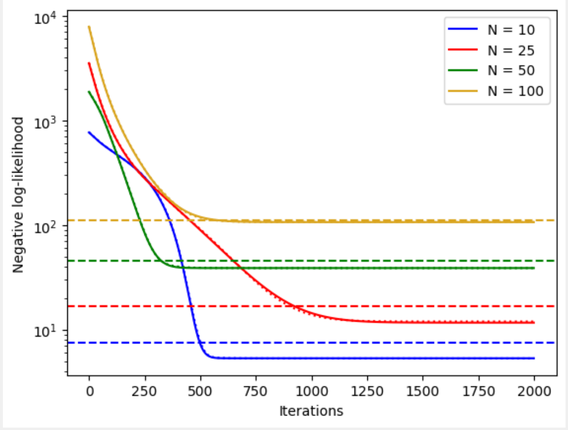
$$\sum_{i=1}^{k} \left[\frac{(M^T c)_i}{2\lambda_i + \lambda} \right]^2 = \tilde{D}_{1n}.$$

Prior Work

Previous approaches to the EDM completion problem rely on $O(n^3)$ techniques such as semidefinite relaxation, multidimensional unfolding, and rank alternation. Recent proposals improve these methods to handle multiple observations of noisy and incomplete distance matrices [5]. Such techniques operate by either formatting the EDM completion problem as an $O(n^3)$ convex optimization problem (SMR), performing singular value decomposition on the point set to satisfy certain rank constraints based on the dimensionality of the point set (rank alternation), and so on.

Another problem of interest is the unassigned distance geometry problem (UDGP), where a histogram of distances (distance list) is given rather than a matrix [6]. Algorithms including LIGA and TRIBOND are able to assign labels to atoms given embeddable distance lists, but have not been extensively limit tested on experimental (e.g. noisy) ones [6]. The LIGA and TRIBOND algorithms also have intermediate steps involving the construction of structure with distance matrices; these steps can be improved by our corruption reversal techniques.

Convergence of Gaussian Noise Reduction



Convergence of the algorithm outlined in Theorem 10. Distance matrices are $N \times N$ and Gaussian noise $\mathcal{N}(0, 1)$ is added to the last row and column.

References

[1] Khadija El Bourakadi, Rachid Bouhfid, and Abou el Kacem Qaiss. Chapter 2 - characterization techniques for hybrid nanocomposites based on cellulose nanocrystals/nanofibrils and nanoparticles. In Denis Rodrigue, Abou el Kacem Qaiss, and Rachid Bouhfid, editors, *Cellulose Nanocrystal/Nanoparticles Hybrid Nanocomposites*, Woodhead Publishing Series in Composites Science and Engineering, pages 27–64. Woodhead Publishing, 2021. ISBN 978-0-12-822906-4. doi: https://doi.org/10.1016/B978-0-12-822906-4.00010-4. URL `https://www.sciencedirect.com/science/article/pii/B9780128229064000104`.

[2] Christopher L. Farrow and Simon J. L. Billinge. Relationship between the atomic pair distribution function and small-angle scattering: implications for modeling of nanoparticles. *Acta Crystallographica Section A*, 65(3):232–239, May 2009. doi: 10.1107/S0108767309009714. URL `https://doi.org/10.1107/S0108767309009714`.

[3] Long V. Le, Jeroen A. Deijkers, Young D. Kim, Haydn N. G. Wadley, and David E. Aspnes. Noise reduction and peak detection in x-ray diffraction data by linear and nonlinear methods. *Journal of Vacuum Science & Technology B*, 41(4):044004, 06 2023. ISSN 2166-2746. doi: 10.1116/6.0002526. URL `https://doi.org/10.1116/6.0002526`.

[4] Rajendra Bhatia and Tanvi Jain. Mean matrices and conditional negativity, Jul 2024. URL `https://journals.uwyo.edu/index.php/ela/index`.

[5] Sai Sumanth Natva and Santosh Nannuru. Denoising and completion of euclidean distance matrix from multiple observations. In *2024 National Conference on Communications (NCC)*, pages 1–6. IEEE, 2024.

[6] P.M. Duxbury, L. Granlund, S.R. Gujarathi, P. Juhas, and S.J.L. Billinge. The unassigned distance geometry problem. *Discrete Applied Mathematics*, 204:117–132, 2016. ISSN 0166-218X. doi: https: