Math 190 **Project 2** **Due: Friday 4/28 at 11:59 p.m. EST**

Instruction: The project is on distance geometry. The total points that one can score is 100. In addition to the solution of the problems, include with your submission a brief report that summarizes in non-technical terms the problem, the method, merits of the method, limitations of the method and limitations of the model if one was to use the proposed method in a practical setting. Include a printout of your code with your project submission. You should submit the project on Gradescope. I hope you enjoy the problem!

1 Distance geometry

Assume that we have n points $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n$ in $\mathcal{R}^r$. The Euclidean distance between $\mathbf{p}_i$ and $\mathbf{p}_j$ can be computed as follows

$$d_{i,j} = ||\mathbf{p}_i - \mathbf{p}_j||_2 = \sqrt{\sum_{k=1}^{r} (\mathbf{p}_i - \mathbf{p}_j)_k^2}.$$

If we are given all the points, ignoring for now computational complexity, computing all the pairwise distances is theoretically straightforward. We can just apply the above formula. In this project, we consider the inverse problem. In the inverse problem, we are given some of the pairwise distances and the problem is to recover the points. This is known as the distance geometry problem [Liberti and Lavor, 2017]. The problem finds applications in many tasks such as structure prediction, sensor localization and machine learning. Few notations are in order. Here on, $\mathbf{D} \in \mathcal{R}^{n \times n}$ denotes the squared distance matrix with $D_{i,j} = d_{i,j}^2$. The set $\Omega \subset \{(i,j) | 1 \leq i, j \leq n \ \& \ i \neq j\}$ is the set that consists all the (i,j) indices of the known entries of the squared distance matrix. To give a concrete example, let $n = 3, r = 2$ and we only know the distance between $\mathbf{p}_1$ and $\mathbf{p}_3$. In this case, $\Omega = \{(1,3), (3,1)\}$.

(a) Prove that $\mathbf{D}_{i,j} = \mathbf{p}_i^T \mathbf{p}_i + \mathbf{p}_j^T \mathbf{p}_j - 2\mathbf{p}_i^T \mathbf{p}_j$.

(b) Let $\mathbf{P} \in \mathcal{R}^{r \times n}$ be the matrix whose columns are the points. The matrix $\mathbf{X} = \mathbf{P}^T \mathbf{P}$ is known as the Gram matrix. Prove that the Gram matrix is symmetric and positive semidefinite.

(c) Let rank$(\mathbf{P}) = r$. Prove that the rank of the Gram matrix is r.

(d) Let the eigendecomposition of $\mathbf{X}$ be $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ where the eigenvalues of $\mathbf{D}$ are arranged in decreasing order. Consider the following optimization problem

$$\min_{\mathbf{P} \in \mathcal{R}^{r \times n}} ||\mathbf{X} - \mathbf{P}^T \mathbf{P}||_F^2.$$

Prove that the optimal solution is $\mathbf{P}^* = \mathbf{D}_r^{\frac{1}{2}} \mathbf{U}_r^T$ where $\mathbf{U}_r$ contains the first r eigenvectors corresponding to the largest r eigenvalues and $\mathbf{D}_r \in \mathcal{R}^{r \times r}$ is a diagonal matrix of the first r largest eigenvalues.

(e) Using (a), prove that the Gram matrix and the square distance matrix can be related in the following way

$$\mathbf{D} = \mathbf{1}\text{diag}(\mathbf{X})^T + \text{diag}(\mathbf{X})\mathbf{1}^T - 2\mathbf{X},$$

where $\mathbf{1} \in \mathcal{R}^n$ is a vector of ones and diag$(\cdot)$ is a column vector of the diagonal entries of the matrix in consideration.

(f) Prove that the relation in (d) can be written as follows

$$D_{i,j} = X_{i,i} + X_{j,j} - 2X_{i,j}$$

(g) When $n \gg r$, $\mathbf{X}$ is an $n \times n$ matrix which has low rank. With that, we consider the following optimization problem to recover $\mathbf{X}$. Note once $\mathbf{X}$ is recovered, we can find $\mathbf{P}$ by employing eigendecomposition.

$$\min_{\mathbf{X} \succeq \mathbf{0}} \ \text{rank}(\mathbf{X}) \ \text{ such that } \ X_{i,i} + X_{j,j} - 2X_{i,j} = D_{i,j} \quad (i,j) \in \Omega,$$

where $\mathbf{X} \succeq \mathbf{0}$ denotes that $\mathbf{X}$ is symmetric and positive semidefinite. Show that this is not a convex optimization problem.

(h) A common approach to obtain a convex problem for distance geometry is by replacing the rank function with the trace. Under some conditions, it is known that trace minimization promotes low rank solutions [Tasissa and Lai, 2018]. With that, consider the following optimization problem

$$\min_{\mathbf{X} \succeq \mathbf{0}} \ \text{trace}(\mathbf{X}) \ \text{ such that } \ X_{i,i} + X_{j,j} - 2X_{i,j} = D_{i,j} \quad (i,j) \in \Omega, \tag{1}$$

Prove that the optimization problem is convex.

(i) In this problem, we apply the trace minimization problem in (1) for structure prediction problem. Our target structure is the protein 1PTQ where the structure is obtained from `https://www.rcsb.org/structure/1PTQ`. 1PTQ has 402 atoms. Our goal is to recover the points using only 4% of the distance matrix. Download `omega.mat`, `points.mat` and `dist.mat` from the `project 2` folder on Canvas. `omega.mat` contains a binary matrix. If the (i,j) entry of the matrix is 1, it indicates that we know the distance between $\mathbf{p}_i$ and $\mathbf{p}_j$. The `points.mat` contains the true points which we will use to check how well the algorithm performs. `dist.mat` contains the squared distance matrix. if you use Python, download `omega.csv`, `points.csv` and `dist.csv`. Implement the algorithm in (1) to recover the optimal $\mathbf{X}^*$. Using Procrustes analysis in MATLAB or Python, show a scatter plot of the aligned $\mathbf{P}^*$ with $\mathbf{P}$.

References

[Liberti and Lavor, 2017] Liberti, L. and Lavor, C. (2017). *Euclidean distance geometry*, volume 133. Springer.

[Tasissa and Lai, 2018] Tasissa, A. and Lai, R. (2018). Exact reconstruction of euclidean distance geometry problem using low-rank matrix completion. *IEEE Transactions on Information Theory*, 65(5):3124–3144.