Estimating missing data in Euclidean Distance Matrices using low rank completion techniques

• N. Moreira† • C. Lavor† • C. Torezzan§

† Department of Applied Mathematics, University of Campinas, Campinas - SP, 13083-970

§ School of Applied Sciences, University of Campinas, Limeira - SP, 13484-350

Euclidean distances are found in many real-world applications. For example, in molecular conformation problem, some pairs of atoms are connected and the distances between them can be estimated using nuclear magnetic resonance experiments [10]. Other interesting examples and references can be found in [3, 6, 9].

Due to several reasons, in many situations we cannot access all the pairwise distances and, therefore, it becomes necessary to estimate those missing data. This problem is known as the *Matrix Completion Problem* [1].

In many applications, the matrices to be completed are low-rank ones, that is, the minimum between the number of rows and the number of columns is much larger than the rank of the matrix. For instance, in molecular conformation problem, we may be interested in the coordinates of thousands of atoms (points in $\mathbb{R}^3$). This leads us to a Euclidean distance matrix with rank at most 5, which is very low comparing to the number points [3].

Let $X = [x_1\ x_2\ \ldots\ x_n]\,,\ \ x_i \in \mathbb{R}^k$, be a matrix whose columns represent n points in a Euclidean $k-$dimensional space. The distance-square (d_{ij}) between any two points in $\mathbb{R}^k$ is defined by:

$$d_{ij} = \|x_i - x_j\|_2^2 \triangleq \langle x_i - x_j\,,\, x_i - x_j\rangle = x_i^T x_i - 2x_i^T x_j + x_j^T x_j = \|x_i\|_2^2 + \|x_j\|_2^2 - 2x_i^T x_j. \quad (1)$$

The matrix $D = [d_{ij}] \in \mathbb{R}^{n\times n}$, whose elements represent the square of the distances between n points in $\mathbb{R}^k$ is called *Euclidean Distance Matrix*, denoted by $edm(X)$.

Motivated by the fact that the rank of the $edm(X)$ is at most $k+2$ and based on models presented in [1, 8], we modeled the problem of estimating missing data in EDM as the following optimization problem:

$$\begin{array}{ll} \min\limits_{\hat{D}} & \dfrac{1}{2}\displaystyle\sum_{(i,j)\in\Omega}(d_{ij} - \hat{d}_{ij})^2 + \lambda\|\hat{D}\|_* \\ \textit{subject to} & \operatorname{rank}(\hat{D}) = k+2, \end{array} \quad (2)$$

where Ω denotes the indices of observed entries, $\lambda \geq 0$ is a regularization parameter controlling the nuclear norm of the minimizer.

Since there is not a straightforward formula for $\operatorname{rank}(\hat{D})$ as a function of the variables $\hat{d}_{ij}$, we propose a heuristic approach to solve (2) based on a modification

in the Soft-impute Algorithm [8]. The resulting method is called the *Fixed-Rank Soft-Impute* and is summarized in Algorithm 1 [7].

Algorithm 1 - Fixed-Rank Soft-Impute

- Inputs: a matrix with missing data, $D \in \mathbb{R}^{n \times n}$, the rank r of the target matrix X and a tolerance $\epsilon > 0$.

1. Inicialize $X^{old} = 0$.
2. Choose a initial value to λ ($\lambda = \lambda_0$).
3. For $i = 1$ to K do:
   (a) Calcule $X^{new} \rightarrow S_\lambda(P_\Omega(D) + P_\Omega^\perp(X^{old}))$.
   (b) Set $\lambda = \beta\sigma_{r+1}$, with $\beta \in (0, 1)$.
   (c) If $\frac{\|X^{new} - X^{old}\|_F^2}{\|X^{old}\|_F^2} < \epsilon$, exit.
   (d) Do $X^{old} \leftarrow X^{new}$.

- Output: X^{new}.

In Algorithm 1, σ_{r+1} is the $(r+1)$th singular value of the matrix $P_\Omega(D) + P_\Omega^\perp(X^{old})$, X^{old} is the most recent approximation, $S_\lambda(X) = U\Sigma_\lambda V^T$, with $\Sigma_\lambda = diag[(\sigma_1 - \lambda)_+, \ldots, (\sigma_r - \lambda)_+]$ ($U\Sigma V^T$ denotes the singular value decomposition of X), $\Sigma = diag[\sigma_1, \ldots, \sigma_r]$, $t_+ = \max(t, 0)$, $r = rank(X)$, $\|\cdot\|_F$ is the Frobenius norm $\left(\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}|a_{ij}|^2}\right)$ and

$$P_\Omega(X)(i, j) = \begin{cases} X_{i,j}, & \text{if} \quad (i, j) \in \Omega \\ 0, & \text{if} \quad (i, j) \notin \Omega \end{cases}.$$

The tests performed were based on a random generation of distance matrices and random deletion of a percentage p of their elements. The original matrix D was maintained for comparison purposes. We used the relative error, defined by $er = \frac{\|D - \hat{D}\|_F^2}{\|D\|_F^2}$, where $||A||_F$ represents the Frobenius norm of matrix A. We also compute the maximum error (max_err) between two corespondent elements, i.e. $max\_err = max|d_{ij} - \hat{d}_{ij}|$, which is a very rigorous criteria of convergence.

Table 1 shows numerical results of tests performed with EMD's obtained from random generation. We varied the number of points generated, the space dimension, the percentage of deletion, and we fixed the tolerance (relative error) equal to 10^{-8}.

# Points (n)	Dimension (d)	Rank (r)	Deletion (%)	# Iteration	Maximum error
500	10	12	50	61	2.76×10^{-7}
1000	3	5	70	82	7.11×10^{-8}
2000	10	12	70	86	2.45×10^{-7}
5000	50	52	50	44	1.27×10^{-6}
5000	200	202	50	74	8.32×10^{-6}
5000	50	52	80	193	1.61×10^{-6}
5000	3	5	90	201	1.24×10^{-4}
10000	100	102	80	187	3.40×10^{-6}

Tab. 1: Performance of Algorithm 1 varying some parameters such as the number of points generated, the space dimension, and the percentage of random deletion. For every experiment we fixed the tolerance (relative error) equal to 10^{-8}.

According to these results, we can see that Algorithm 1 was able to recover missing data with very high accuracy. Additionally, Table 1 shows us that we can recover matrices with not so small rank. We recovered, for instance, matrices with rank equal to 100 and 200 with a very good accuracy.

References

[1] E. J. Candès and B. Recht. *Exact Matrix Completion via Convex Optimization.* Found. of Comput. Math., 9 717-772, 2008.

[2] E. J. Candès and T. Tao. *The Power of Convex Relaxation: Near-Optimal Matrix Completion.* IEEE Transactions on Information Theory 56(5), 2053-2080. 2010.

[3] I. Dokmanić, R. Parhizkar, J. Ranieri and M. Vetterli. *Euclidean Distance Matrices*, IEEE Signal Processing Magazine (Volume: 32, Issue: 6, Nov. 2015).

[4] J. F. Cai, E. J. Candès and Z. Shen. *A Singular Value Thresholding Algorithm for Matrix Completion.* SIAM J. on Optimization 20(4), 1956-1982, 2008.

[5] J. Leeuw. Multidimensional scaling. International Encyclopedia of the Social & Behavioral Sciences, pages 1351213519. Elsevier, 2004.

[6] L. Liberti, C. Lavor, N. Maculan and A. Mucherino. *Euclidean Distance Geometry and Applications*, SIAM Rev. 56(1), 369 (2014).

[7] N. Moreira, L. T. Duarte, C. Lavor, and C. Torezzan. *A novel low-rank matrix completion approach to estimate missing entries in Euclidean distance matrices*, Submited to Journal of Global Optimization, 2017.

[8] R. Mazumder, T. Hastie and R. Tibshirani. *Spectral Regularization Algorithms for Learning Large Incomplete Matrices*, Journal of Machine Learning Research 11 (2010), 2287-2322.

[9] S. Billinge, P. Duxbury, D. Gonçalves, C. Lavor, and A. Mucherino. *Assigned and unassigned distance geometry: applications to biological molecules and nanostructures.* 4OR, 14:337-376, 2016.

[10] T. F. Havel and K. Wüthrich. *An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformations in solution*, J. Mol. Biol., vol. 182, no. 2, pp. 281294, 1985.