Noname manuscript No.
(will be inserted by the editor)

Convex Optimization Learning of Faithful Euclidean Distance Representations in Nonlinear Dimensionality Reduction

Chao Ding · **Hou-Duo Qi**

Received: date / Accepted: date

Abstract Classical multidimensional scaling only works well when the noisy distances observed in a high dimensional space can be faithfully represented by Euclidean distances in a low dimensional space. Advanced models such as Maximum Variance Unfolding (MVU) and Minimum Volume Embedding (MVE) use Semi-Definite Programming (SDP) to reconstruct such faithful representations. While those SDP models are capable of producing high quality configuration numerically, they suffer two major drawbacks. One is that there exist no theoretically guaranteed bounds on the quality of the configuration. The other is that they are slow in computation when the data points are beyond moderate size. In this paper, we propose a convex optimization model of Euclidean distance matrices. We establish a non-asymptotic error bound for the random graph model with sub-Gaussian noise, and prove that our model produces a matrix estimator of high accuracy when the order of the uniform sample size is roughly the degree of freedom of a low-rank matrix up to a logarithmic factor. Our results partially explain why MVU and MVE often work well. Moreover, the convex optimization model can be efficiently solved by a recently proposed 3-block alternating direction method of multipliers. Numerical experiments show that the model can produce configurations of high quality on large data points that the SDP approach would struggle to cope with.

Keywords Euclidean distance matrix · convex matrix optimization · multidimensional scaling · nonlinear dimensionality reduction · low-rank matrix · error bounds · random graph models

Mathematics Subject Classification (2010) 49M45 · 90C25 · 90C33

1 Introduction

The chief purpose of this paper is to find a complete set of faithful Euclidean distance representations in a low-dimensional space from a partial set of noisy distances, which are supposedly observed in a higher dimensional space. The proposed model and method thus belong to the vast field of nonlinear dimensionality reduction. Our model is strongly inspired by several high-profile Semi-Definite Programming (SDP) models, which aim to achieve a similar purpose, but suffer two major drawbacks: (i) theoretical guarantees yet to

This version: May 10, 2016. This work is supported by Engineering and Physical Science Research Council (UK) project EP/K007645/1.

C. Ding
Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, P.R. China.
E-mail: dingchao@amss.ac.cn

H.D. Qi
School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK.
E-mail: hdqi@soton.ac.uk

be developed for the quality of recovered distances from those SDP models and (ii) the slow computational convergence, which severely limits their practical applications even when the data points are of moderate size. Our distinctive approach is to use convex optimization of Euclidean Distance Matrices (EDM) to resolve those issues. In particular, we are able to establish theoretical error bounds of the obtained Euclidean distances from the true distances under the assumption of uniform sampling, which has been widely used in modelling social networks. Moreover, the resulting optimization problem can be efficiently solved by a 3-block alternating direction method of multipliers. In the following, we will first use social network to illustrate how initial distance information is gathered and why the uniform sampling is a good model in understanding them. We then briefly discuss several SDP models in nonlinear dimensionality reduction and survey relevant error-bound results from matrix completion literature. They are included in the first three subsections below and collectively serve as a solid motivation for the current study. We finally summarize our main contributions with notation used in this paper.

1.1 Distances in Social Network and Their Embedding

The study of structural patterns of social network from the ties (relationships) that connect social actors is one of the most important research topics in social network analysis [59]. To this end, measurements on the actor-to-actor relationships (kinship, social roles, etc.) are collected or observed by different methods (questionnaires, direct observation, etc.) and the measurements on the relational information are referred as the network composition. The measurement data usually can be presented as an $n \times n$ measurement matrix, where the n rows and the n columns both refer to the studied actors. Each entry of these matrices indicates the social relationship measurement (e.g., presence/absence or similarity/dissimilarity) between the row and column actors. In this paper, we are only concerned with symmetric relationships, i.e., the relationship from actor i to actor j is the same as that from actor j to actor i. Furthermore, there exist standard ways to convert the measured relationships into Euclidean distances, see [18, Sect. 1.3.5] and [8, Chp. 6].

However, it is important to note that in practice, only partial relationship information can be observed, which means that the measurement matrix is usually incomplete and noisy. The observation processes are often assumed to follow certain random graph model. One simple but widely used model is the Bernoulli random graph model [52,20]. Let n labelled vertices be given. The Bernoulli random graph is obtained by connecting each pair of vertices independently with the common probability p and it reproduces well some principal features of the real-world social network such as the "small-world" effect [40,19]. Other properties such as the degree distribution and the connectivity can be found in e.g., [7,27]. For more details on the Bernoulli as well as other random models, one may refer to the review paper [42] and references therein. In this paper, we mainly focus on the Bernoulli random graph model. Consequently, the observed measurement matrix follows the uniform sampling rule which will be described in Sect. 2.3.

In order to examine the structural patterns of a social network, the produced images (e.g., embedding in 2 or 3 dimensional space for visualization) should preserve the structural patterns as much as possible, as highlighted by [23], "*the points in a visual image should be located so the observed strengths of the inter-actor ties are preserved*." In other words, the designed dimensional reduction algorithm has to assure that the embedding Euclidean distances between points (nodes) fit in the best possible way the observed distances in a social space. Therefore, the problem now reduces to whether one can effectively find the best approximation in a low dimensional space to the true social measurement matrix, which is incomplete and noisy. The classical Multidimensional Scaling (cMDS) (see Sect. 2.1) provides one of the most often used embedding methods. However, cMDS alone is often not adequate to produce satisfactory embedding, as rightly observed in several high-profile embedding methods in manifold learning.

1.2 Embedding Methods in Manifold Learning

The cMDS and its variants have found many applications in data dimension reduction and have been well documented in the monographs [18,8]. When the distance matrix (or dissimilarity measurement matrix) is

close to a true EDM with the targeted embedding dimension, cMDS often works very well. Otherwise, a large proportion of unexplained variance has to be cut off or it may even yield negative variances, resulting in what is called embedding in a pseudo-Euclidean space and hence creating the problem of unconventional interpretation of the actual embedding (see e.g., [46]).

cMDS has recently motivated a number of high-profile numerical methods, which all try to alleviate the issue mentioned above. For example, the ISOMAP of [54] proposes to use the shortest path distances to approximate the EDM on a low-dimensional manifold. The Maximum Variance Unfolding (MVU) of [61] through SDP aims for maximizing the total variance and the Minimum Volume Embedding (MVE) of [51] also aims for a similar purpose by maximizing the eigen gap of the Gram matrix of the embedding points in a low-dimensional space. The need for such methods comes from the fact that the initial distances either are in stochastic nature (e.g., containing noises) or cannot be measured (e.g., missing values). The idea of MVU has also been used in the refinement step of the celebrated SDP method for sensor network localization problems [6].

It was shown in [54,4] that ISOMAP enjoys the elegant theory that the shortest path distances (or graph distances) can accurately estimate the true geodesic distances with a high probability if the finite points are chosen randomly from a compact and convex submanifold following a Poisson distribution with a high density, and the pairwise distances are obtained by the k-nearest neighbor rule or the unit ball rule (see Sect. 2.3 for the definitions). However, for MVU and MVE, there exist no theoretical guarantee as to how good the obtained Euclidean distances are. At this point, it is important to highlight two observations. (i) The shortest path distance or the distance by the k-nearest neighbor or the unit-ball rule is often not suitable in deriving distances in social network. This point has been emphasized in the recent study on E-mail social network by [10]. (ii) MVU and MVE models only depend on the initial distances and do not depend on any particular ways in obtaining them. They then rely on SDP to calculate the best fit distances. From this point of view, they can be applied to social network embedding. This is also pointed out in [10]. Due to the space limitation, we are not able to review other leading methods in manifold learning, but refer to [12, Chp. 4] for a guide.

Inspired by their numerical success, our model will inherit the good features of both MVU and MVE. Moreover, we are able to derive theoretical results in guaranteeing the quality of the obtained Euclidean distances. Our results are the type of error bounds, which have attracted growing attention recently. We review the relevant results below.

1.3 Error Bounds in Low-Rank Matrix Completion and Approximation

As mentioned in the preceding section, our research has been strongly influenced by the group of researches that are related to the MVU and MVE models, which have natural geometric interpretations and use SDP as their major tool. Their excellent performance in data reduction calls for theoretical justification. Our model also enjoys a similar geometric interpretation, but departs from the two models in that we deal with EDM directly rather than reformulating it as SDP. This key departure puts our model in the category of matrix approximation problems, which have attracted much attention recently from machine learning community and motivated our research.

The most popular approach to recovering a low-rank matrix solution of a linear system is via the nuclear norm minimization [38,22]. What makes this approach more exciting and important is that it has a theoretically guaranteed recoverability (recoverable with a high probability). The first such a theoretical result was obtained by [48] by employing the Restricted Isometric Property (RIP). However, for the matrix completion problem the sample operator does not satisfy the RIP (see e.g., [13]). For the noiseless case, [14] proved that a low-rank matrix can be fully recovered with high probability provided that a small number of its noiseless observations are uniformly sampled. See [15] for an improved bound and [26] for the optimal bound on the sample number. We also refer to [47] for a short and intelligible analysis of the recoverability of the matrix completion problem.

The matrix completion with noisy observations was studied by [13]. Recently, the noisy case was further studied by several groups of researchers including [34], [41] and [32], under different settings. In particular, the matrix completion problem with fixed basis coefficients was studied by [39], who proposed a rank-

corrected procedure to generate an estimator using the nuclear semi-norm and established the corresponding non-asymmetric recovery bounds.

Very recently, [28] proposed a SDP model for the problem of (sensor network) localization from an incomplete set of noisy Euclidean distances. Using the fact that the squared Euclidean distances can be represented by elements from a positive semidefinite matrix:

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j\rangle = X_{ii} + X_{jj} - 2X_{ij},$$

where $\mathbf{x}_i \in \Re^d$ are embedding points and X defined by $X_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ is the Gram matrix of those embedding points, the SDP model aims to minimize $\mathrm{Tr}(X)$ (the nuclear norm of X). Equivalently, the objective is to minimize the total variance $\sum \|\mathbf{x}_i\|^2$ of the embedding points. This objective obviously contradicts the main idea of MVU and MVE, which aim to make the total variance as large as possible. It is important to point out that making the variance as big as possible seems to be indispensable for SDP to produce high quality of localization. This has been numerically demonstrated in [6].

The impressive result in [28] roughly states that the obtained error bound reads as $O((nr^d)^5 \frac{\Delta}{r^4})$ containing an undesirable term $(nr^d)^5$, where r is the radius used in the unit ball rule, d is the embedding dimension, Δ is the bound on the measurement noise and n is the number of embedding points. As pointed out by [28] that the numerical performance suggested the error seems to be bounded by $O(\frac{\Delta}{r^4})$, which does not match the derived theoretical bound. This result also shows tremendous technical difficulties one may have to face in deriving similar bounds for EDM recovery.

To summarize, most existing error bounds are derived from the nuclear norm minimization. When translating to the Euclidean distance learning problem, minimizing the nuclear norm is equivalent to minimizing the variance of the embedding points, which contradicts the main idea of MVU and MVE in making the variance as large as possible. Hence, the excellent progress in matrix completion/approximation does not straightforwardly imply useful bounds about the Euclidean distance learning in a low-dimensional space. Actually one may face huge difficulty barriers in such extension. In this paper, we propose a convex optimization model to learn faithful Euclidean distances in a low-dimensional space. We derive theoretically guaranteed bounds in the spirit of matrix approximation and therefore provide a solid theoretical foundation in using the model. We briefly describe the main contributions below.

1.4 Main Contributions

This paper makes two major contributions to the field of nonlinear dimensionality reduction. One is on building a convex optimization model with guaranteed error bounds and the other is on a computational method.

(a) Building a convex optimization model and its error bounds. Our departing point from the existing SDP models is to treat EDM (vs positive semidefinite matrix in SDP) as a primary object. The total variance of the desired embedding points in a low-dimensional space can be quantitatively measured through the so-called EDM score. The higher the EDM score is, the more the variance is explained in the embedding. Therefore, both MVU and MVE can be regarded as EDM score driven models. Moreover, MVE, being a nonconvex optimization model, is more aggressive in driving the EDM score up. However, MVU, being a convex optimization model, is more computationally appealing. Our convex optimization model strikes a balance between the two models in the sense that it inherits the appealing features from both.

What makes our model more important is that it yields guaranteed non-asymptotic error bounds under the uniform sampling rule. More precisely, we show in Thm. 1 that for the unknown $n \times n$ Euclidean distance matrix with the embedding dimension r and under mild conditions, the average estimation error is controlled by $Crn\log(n)/m$ with high probability, where m is the sample size and C is a constant independent of n, r and m. It follows from this error bound that our model will produce an estimator with high accuracy as long as the sample size is of the order of $rn\log(n)$, which is roughly the degree of freedom of a symmetric hollow matrix with rank r up to a logarithmic factor in the matrix size. It is worth to point out that with special choices of model parameters, our model reduces to MVU and covers the subproblems solved by MVE. Moreover, our theoretical result corresponding to those specific model parameters explains why under the uniform sampling

rule, the MVE often leads to configurations of higher quality than the MVU. To our knowledge, it is the first such theoretical result that shed lights on the MVE model. There are some theoretical results on the asymptotic behavior of MVU obtained recently in [2,45]. However, these results are different from ours in the sense that they are only true when the number of the points is sufficiently large.

(b) An efficient computational method. Treating EDM as a primary object not only benefits us in deriving the error-bound results, but also leads to an efficient numerical method. It allows us to apply a recently proposed convergent 3-block alternating direction method of multipliers (ADMM) [3] even for problems with a few thousands of data points. Previously, the original models of both MVU and MVE have numerical difficulties when the data points are beyond 1000. They may even have difficulties with a few hundreds of points when their corresponding slack models are to be solved. In order to increase the scalability of MVU, some algorithms are proposed in [62]. Most recently, Chen et al. [16] derive a novel variant of MVU: the Maximum Variance Correction (MVC), which greatly improves its scalability. However, for some social network applications, the quality of the embedding graph form MVC is questionable, probably because there is no theoretical guarantee on the embedding accuracy. For instance, as shown in Sect. 6, for US airport network (1572 nodes) and Political blogs (1222 nodes), MVC embedding failed to capture any important features in the two networks, although it is much faster in computing time.

Moreover, We are also able to develop theoretically optimal estimates of the model parameters. This gives a good indication how we should set the parameter values in our implementation. Numerical results both on social networks and the benchmark test problems in manifold learning show that our method can fast produce embeddings of high quality.

1.5 Organization and Notation

The paper is organized as follows. Sect. 2 provides necessary background with a purpose to cast the MVU and MVE models as EDM-score driven models. This viewpoint will greatly benefit us in understanding our model, which is described in Sect. 3 with more detailed interpretation. We report our error bound results in Sect. 4. Sect. 5 contains the theoretical optimal estimates of the model parameters as well as a convergent 3-block ADMM algorithm. We report our extensive numerical experiments in Sect. 6 and conclude the paper in Sect. 7.

Notation. Let $\mathbb{S}^n$ be the space of $n \times n$ real symmetric matrices with the trace inner product $\langle X, Y \rangle := \text{trace}(XY)$ for $X, Y \in \mathbb{S}^n$ and its induced Frobenius norm $\|\cdot\|$. Denote $\mathbb{S}^n_+$ the symmetric positive semidefinite matrix cone. We also write $X \succeq 0$ whenever $X \in \mathbb{S}^n_+$. We use $I \in \mathbb{S}^n$ to represent the identity matrix and $\mathbf{1} \in \Re^n$ to represent the vector of all ones. Column vectors are denoted by lower case letters in boldface, such as $\mathbf{x} \in \Re^n$. Let $\mathbf{e}_i \in \Re^n$, $i = 1, \ldots, n$ be the column vector with the i-th entry being one and the others being zero. For a given $X \in \mathbb{S}^n$, we let $\text{diag}(X) \in \Re^n$ denote the vector formed from the diagonal of X. Below are some other notations to be used in this paper:

- For any $Z \in \Re^{m \times n}$, we denote by Z_{ij} the (i, j)-th entry of Z. We use $\mathbb{O}^n$ to denote the set of all n by n orthogonal matrices.
- For any $Z \in \Re^{m \times n}$, we use $\mathbf{z}_j$ to represent the j-th column of Z, $j = 1, \ldots, n$. Let $\mathscr{J} \subseteq \{1, \ldots, n\}$ be an index set. We use $Z_{\mathscr{J}}$ to denote the sub-matrix of Z obtained by removing all the columns of Z not in $\mathscr{J}$.
- Let $\mathscr{I} \subseteq \{1, \ldots, m\}$ and $\mathscr{J} \subseteq \{1, \ldots, n\}$ be two index sets. For any $Z \in \Re^{m \times n}$, we use $Z_{\mathscr{I}\mathscr{J}}$ to denote the $|\mathscr{I}| \times |\mathscr{J}|$ sub-matrix of Z obtained by removing all the rows of Z not in $\mathscr{I}$ and all the columns of Z not in $\mathscr{J}$.
- We use "$\circ$" to denote the Hadamard product between matrices, i.e., for any two matrices X and Y in $\Re^{m \times n}$ the (i, j)-th entry of $Z := X \circ Y \in \Re^{m \times n}$ is $Z_{ij} = X_{ij} Y_{ij}$.
- For any $Z \in \Re^{m \times n}$, let $\|Z\|_2$ be the spectral norm of Z, i.e., the largest singular value of Z, and $\|Z\|_*$ be the nuclear norm of Z, i.e., the sum of singular values of Z. The infinity norm of Z is denoted by $\|Z\|_\infty$.

2 Background

This section contains three short parts. We first give a brief review of cMDS, only summarizing some of the key results that we are going to use. We then describe the MVU and MVE models, which are closely related to ours. Finally, we explain three most commonly used distance-sampling rules.

2.1 cMDS

cMDS has been well documented in [18,8]. In particular, Section 3 of [46] explains when it works. Below we only summarize its key results for our future use. A $n \times n$ matrix D is called Euclidean distance matrix (EDM) if there exist points $\mathbf{p}_1, \ldots, \mathbf{p}_n$ in $\Re^r$ such that $D_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|^2$ for $i, j = 1, \ldots, n$, where $\Re^r$ is called the embedding space and r is the embedding dimension when it is the smallest such r.

An alternative definition of EDM that does not involve any embedding points $\{\mathbf{p}_i\}$ can be described as follows. Let $\mathbb{S}_h^n$ be the hollow subspace of $\mathbb{S}^n$, i.e., $\mathbb{S}_h^n := \{X \in \mathbb{S}^n \mid \mathrm{diag}(X) = 0\}$. Define the almost positive semidefinite cone $\mathbb{K}_+^n$ by

$$\mathbb{K}_+^n := \left\{ A \in \mathbb{S}^n \mid \mathbf{x}^T A \mathbf{x} \geq 0,\ \mathbf{x} \in \mathbf{1}^\perp \right\} = \{ A \in \mathbb{S}^n \mid JAJ \succeq 0 \}, \tag{1}$$

where $\mathbf{1}^\perp := \{\mathbf{x} \in \Re^n \mid \mathbf{1}^T \mathbf{x} = 0\}$ and $J := I - \mathbf{1}\mathbf{1}^T/n$ is known as the centering matrix. It is well-known [50, 63] that $D \in \mathbb{S}^n$ is EDM if and only if $-D \in \mathbb{S}_h^n \cap \mathbb{K}_+^n$. Moreover, the embedding dimension is determined by the rank of the doubly centered matrix JDJ, i.e., $r = \mathrm{rank}(JDJ)$.

Since $-JDJ$ is positive semidefinite, its spectral decomposition can be written as

$$-\frac{1}{2} JDJ = P \mathrm{diag}(\lambda_1, \ldots, \lambda_n) P^T,$$

where $P^T P = I$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ are the eigenvalues in nonincreasing order. Since $\mathrm{rank}(JDJ) = r$, we must have $\lambda_i = 0$ for all $i \geq (r+1)$. Let P_1 be the submatrix consisting of the first r columns (eigenvectors) in P. One set of the embedding points are

$$\begin{pmatrix} \mathbf{p}_1^T \\ \vdots \\ \mathbf{p}_n^T \end{pmatrix} = P_1 \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r}). \tag{2}$$

cMDS is built upon the above result. Suppose a pre-distance matrix D (i.e., $D \in \mathbb{S}_h^n$ and $D \geq 0$) is known. It computes the embedding points by (2). Empirical evidences have shown that if the first r eigenvalues are positive and the absolute values of the remaining eigenvalues (they may be negative as D may not be a true EDM) are small, then cMDS often works well. Otherwise, it may produce misleading embedding points. For example, there are examples that show that ISOMAP might cut off too many eigenvalues, hence failing to produce satisfactory embedding (see e.g., Teapots data example in [61]). Both MVU and MVE models aim to avoid such situation.

The EDM score has been widely used to interpret the percentage of the total variance being explained by the embedding from leading eigenvalues. The EDM score of the leading k eigenvalues is defined by

$$\mathrm{EDMscore}(k) := \sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{n} \lambda_i, \qquad k = 1, 2, \ldots, n.$$

It is only well defined when D is a true EDM. The justification of using EDM scores is deeply rooted in the classic work of [25], who showed that cMDS is a method of principal component analysis, but working with EDMs instead of correlation matrices.

The centering matrix J plays an important role in our analysis. It is the orthogonal projection onto the subspace $\mathbf{1}^{\perp}$ and hence $J^2 = J$. Moreover, we have the following. Let $\mathbb{S}^n_c$ be the geometric center subspace in $\mathbb{S}^n$:

$$\mathbb{S}^n_c := \{Y \in \mathbb{S}^n \mid Y\mathbf{1} = 0\}. \tag{3}$$

Let $\mathscr{P}_{\mathbb{S}^n_c}(X)$ denote the orthogonal projection onto $\mathbb{S}^n_c$. Then we have $\mathscr{P}_{\mathbb{S}^n_c}(X) = JXJ$. That is, the doubly centered matrix JXJ, when viewed as a linear transformation of X, is the orthogonal projection of X onto $\mathbb{S}^n_c$. Therefore, we have

$$\langle JXJ,\, X - JXJ\rangle = 0. \tag{4}$$

It is also easy to verify the following result.

Lemma 1 *For any $X \in \mathbb{S}^n_h$, we have $X - JXJ = \frac{1}{2}\left(\mathrm{diag}(-JXJ)\,\mathbf{1}^T + \mathbf{1}\,\mathrm{diag}(-JXJ)^T\right)$.*

2.2 MVU and MVE Models

The input of MVU and MVE models is a set of partially observed distances $\left\{d^2_{ij} :\ (i,j) \in \Omega_0\right\}$ and $\Omega_0 \subseteq \Omega := \{(i,j) :\ 1 \le i < j \le n\}$. Let $\{\mathbf{p}_i\}^n_{i=1}$ denote the desired embedding points in $\Re^r$. They should have the following properties. The pairwise distances should be faithful to the observed ones. That is,

$$\|\mathbf{p}_i - \mathbf{p}_j\|^2 \approx d^2_{ij} \qquad \forall\ (i,j) \in \Omega_0 \tag{5}$$

and those points should be geometrically centered in order to remove the translational degree of freedom from the embedding:

$$\sum_{i=1}^{n} \mathbf{p}_i = 0. \tag{6}$$

Let $K = V^T V$ be the Gram matrix of the embedding points, where $V \in \Re^{r\times n}$ is a matrix whose columns are the vectors $\mathbf{p}_i$, $i = 1,\ldots,n$. Then the conditions in (5) and (6) are translated to

$$K_{ii} - 2K_{ij} + K_{jj} \approx d^2_{ij} \qquad \forall\ (i,j) \in \Omega_0 \quad \text{and} \quad \langle \mathbf{1}\mathbf{1}^T,\, K\rangle = 0.$$

To encourage the dimension reduction, MVU argues that the variance, which is $\mathrm{Tr}(K)$, should be maximized. In summary, the slack model (or the least square penalty model) of MVU takes the following form:

$$\begin{array}{ll} \max & \langle I,\, K\rangle - \nu \sum_{(i,j)\in\Omega_0} \left(K_{ii} - 2K_{ij} + K_{jj} - d^2_{ij}\right)^2 \\ \text{s.t.} & \langle \mathbf{1}\mathbf{1}^T,\, K\rangle = 0 \quad \text{and} \quad K \succeq 0, \end{array} \tag{7}$$

where $\nu > 0$ is the penalty parameter that balances the trade-off between maximizing variance and preserving the observed distances. See also [62,53] for more variants of this problem.

The resulting EDM $D \in \mathbb{S}^n$ from the optimal solution of (7) is defined to be $D_{ij} = K_{ii} - 2K_{ij} + K_{jj}$ and it satisfies $K = -0.5JDJ$. Empirical evidence shows that the EDM scores of the first few leading eigenvalues of K are often large enough to explain high percentage of the total variance.

MVE seeks to improve the EDM scores in a more aggressive way. Suppose the targeted embedding dimension is r. MVE tries to maximize the eigen gap between the leading r eigenvalues of K and the remaining eigenvalues. This gives rise to

$$\begin{array}{ll} \max & \sum_{i=1}^{r} \lambda_i(K) - \sum_{i=r+1}^{n} \lambda_i(K) \\ \text{s.t.} & K_{ii} - 2K_{ij} + K_{jj} \approx d^2_{ij} \qquad \forall\ (i,j) \in \Omega_0 \\ & \langle \mathbf{1}\mathbf{1}^T,\, K\rangle = 0 \quad \text{and} \quad K \succeq 0. \end{array}$$

There are a few standard ways in dealing with the constraints corresponding to $(i,j) \in \Omega_0$. We are interested in the MVE slack model:

$$\begin{array}{ll} \max & \sum_{i=1}^{r} \lambda_i(K) - \sum\limits_{i=r+1}^{n} \lambda_i(K) - \nu \sum\limits_{(i,j)\in\Omega_0} \left(K_{ii} - 2K_{ij} + K_{jj} - d_{ij}^2\right)^2 \\ \text{s.t.} & \langle \mathbf{1}\mathbf{1}^T, K\rangle = 0 \quad \text{and} \quad K \succeq 0, \end{array} \tag{8}$$

where $\nu > 0$. The MVE model (8) often yields higher EDM scores than the MVU model (7). However, (7) is a SDP problem while (8) is nonconvex, which can be solved by a sequential SDP method (see [51]).

2.3 Distance Sampling Rules

In this part, we describe how the observed distances indexed by Ω_0 are selected in practice. We assume that those distances are sampled from unknown true Euclidean distances $\overline{d}_{ij}$ in the following fashion.

$$d_{ij} = \overline{d}_{ij} + \eta\,\xi_{ij}, \qquad (i,j) \in \Omega_0, \tag{9}$$

where ξ_{ij} are i.i.d. noise variables with $\mathbb{E}(\xi) = 0$, $\mathbb{E}(\xi^2) = 1$ and $\eta > 0$ is a noise magnitude control factor. We note that in (9) it is the true Euclidean distance $\overline{d}_{ij}$ (rather than its squared quantity) that is being sampled. There are three commonly used rules to select Ω_0.

(i) **Uniform sampling rule**. The elements are independently and identically sampled from Ω with the common probability $1/|\Omega|$.
(ii) k **nearest neighbors** (k**-NN) rule**. For each i, $(i,j) \in \Omega_0$ if and only if d_{ij} belongs to the first k smallest distances in $\{d_{i\ell} : i \neq \ell = 1,\ldots,n\}$.
(iii) **Unit ball rule**. For a given radius $\varepsilon > 0$, $(i,j) \in \Omega_0$ if and only if $d_{ij} \leq \varepsilon$.

The k-NN and the unit ball rules are often used in low-dimensional manifold learning in order to preserve the local structure of the embedding points, while the uniform sampling rule is often employed in some other dimensionality reductions including embedding social network in a low-dimensional space.

3 A Convex Optimization Model for Distance Learning

Both MVU and MVE are trusted distance learning models in the following sense. They both produce a Euclidean distance matrix, which is faithful to the observed distances and they both encourage high EDM scores from the first few leading eigenvalues. However, it still remains a difficult (theoretical) task to quantify how good the resulting embedding is. In this part, we will propose a new learning model, which inherit the good properties of MVU and MVE. Moreover, we are able to quantify the embedding quality by deriving error bounds of the resulting solutions under the uniform sampling rule. Below, we first describe our model, followed by detailed interpretation.

3.1 Model Description

In order to facilitate the description of our model and to set the platform for our subsequent analysis, we write the sampling model (9) as an observation model. Define two matrices $\overline{D}$ and $\overline{D}^{(1/2)}$ respectively by $\overline{D} := \left(\overline{d}_{ij}^2\right)$ and $\overline{D}^{(1/2)} := \left(\overline{d}_{ij}\right)$. Assume that there exists a constant $b > 0$ such that $\|\overline{D}\|_\infty \leq b$. A sampled basis matrix X has the following form:

$$X := \frac{1}{2}(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T) \qquad \text{for some } (i,j) \in \Omega.$$

For each $(i,j) \in \Omega_0$, there exists a corresponding sampling basis matrix. We number them as $X_1, \ldots, X_m$.

Define the corresponding observation operator $\mathscr{O} : \mathbb{S}^n \to \Re^m$ by

$$\mathscr{O}(A) := (\langle X_1, A\rangle, \ldots, \langle X_m, A\rangle)^T, \quad A \in \mathbb{S}^n. \tag{10}$$

That is, $\mathscr{O}(A)$ samples all the elements A_{ij} specified by $(i,j) \in \Omega_0$. Let $\mathscr{O}^* : \Re^m \to \mathbb{S}^n$ be its adjoint, i.e., $\mathscr{O}^*(\mathbf{z}) = \sum_{l=1}^m z_l X_l, \quad z \in \Re^m$. Thus, the sampling model (9) can be re-written as the following compact form

$$\mathbf{y} = \mathscr{O}(\overline{D}^{(1/2)}) + \eta\xi, \tag{11}$$

where $\mathbf{y} = (y_1, \ldots, y_m)^T$ and $\xi = (\xi_1, \ldots, \xi_m)^T$ are the observation vector and the noise vector, respectively.

Since $-J\overline{D}J \in \mathbb{S}^n_+$, we may assume that it has the following singular value decomposition (SVD):

$$-J\overline{D}J = \overline{P}\mathrm{Diag}(\overline{\lambda})\overline{P}^T, \tag{12}$$

where $\overline{P} \in \mathbb{O}^n$ is an orthogonal matrix, $\overline{\lambda} = (\overline{\lambda}_1, \overline{\lambda}_2, \ldots, \overline{\lambda}_n)^T \in \Re^n$ is the vector of the eigenvalues of $-J\overline{D}J$ arranged in nondecreasing order, i.e., $\overline{\lambda}_1 \geq \overline{\lambda}_2 \geq \ldots \geq \overline{\lambda}_n \geq 0$.

Suppose that $\widetilde{D}$ is a given initial estimator of the unknown matrix $\overline{D}$, and it has the following singular value decomposition $-J\widetilde{D}J = \widetilde{P}\mathrm{Diag}(\widetilde{\lambda})\widetilde{P}^T$, where $\widetilde{P} \in \mathbb{O}^n$. In this paper, we always assume the embedding dimension $r := \mathrm{rank}(J\overline{D}J) \geq 1$. Thus, for any given orthogonal matrix $P \in \mathbb{O}^n$, we write $P = [P_1 \;\; P_2]$ with $P_1 \in \Re^{n\times r}$ and $P_2 \in \Re^{n\times(n-r)}$. For the given parameters $\rho_1 > 0$ and $\rho_2 \geq 0$, we consider the following convex optimization problem

$$\begin{array}{ll} \min & \dfrac{1}{2m}\|\mathbf{y}\circ\mathbf{y} - \mathscr{O}(D)\|^2 + \rho_1\Big(\langle I, -JDJ\rangle - \rho_2\langle\Theta, -JDJ\rangle\Big) \\ \text{s.t.} & D \in \mathbb{S}^n_h, \quad -D \in \mathbb{K}^n_+, \quad \|D\|_\infty \leq b, \end{array} \tag{13}$$

where $\Theta := \widetilde{P}_1\widetilde{P}_1^T$. This problem has EDM as its variable and this is in contrast to MVU, MVE and other learning models (e.g., [28]) where they all use SDPs. The use of EDMs greatly benefit us in deriving the error bounds in the next section. Our model (13) tries to accomplish three tasks as we explain below.

3.2 Model Interpretation

The three tasks that model (13) tries to accomplish correspond to the three terms in the objective function. The first (quadratic) term is nothing but $\sum_{(i,j)\in\Omega_0}(d_{ij}^2 - D_{ij})^2$ corresponding to the quadratic terms in the slack models (7) and (8). Minimizing this term (i.e, least-squares) is essentially to find an EDM D that minimizes the error rising from the sampling model (11).

The second term $\langle I, -JDJ\rangle$ is actually the nuclear norm of $(-JDJ)$. Recall that in cMDS, the embedding points in (2) come from the spectral decomposition of $(-JDJ)$. Minimizing this term means to find the smallest embedding dimension. However, as argued in both MVU and MVE models, minimizing the nuclear norm is against the principal idea of maximizing variance. Therefore, to alleviate this conflict, we need the third term $-\langle\widetilde{P}_1\widetilde{P}_1^T, -JDJ\rangle$.

In order to motivate the third term, let us consider an extreme case. Suppose the initial EDM $\widetilde{D}$ is close enough to D in the sense that the leading eigenspaces respectively spanned by $\{\widetilde{P}_1\}$ and by $\{P_1\}$ coincide. That is $\widetilde{P}_1\widetilde{P}_1^T = P_1P_1^T$. Then, $\langle\widetilde{P}_1\widetilde{P}_1^T, -JDJ\rangle = \sum_{i=1}^r \lambda_i =: t$. Hence, minimizing the third term is essentially maximizing the leading eigenvalues of $(-JDJ)$. Over the optimization process, the third term is likely to push the quantity t up, and the second term (nuclear norm) forces the remaining eigenvalues $s := \sum_{i=r+1}^n \lambda_i$ down. The consequence is that the EDM score

$$\mathrm{EDMscore}(r) = f(t,s) := \frac{t}{t+s}$$

gets higher. This is because

$$f(t_2,s_2) > f(t_1,s_1) \qquad \forall\, t_2 > t_1 \quad \text{and} \quad s_2 < s_1.$$

Therefore, the EDM scores can be controlled by controlling the penalty parameters ρ_1 and ρ_2. The above heuristic observation is in agreement with our extensive numerical experiments.

It is easy to see that Model (13) reduces to the nuclear norm penalized least squares (NNPLS) model if $\rho_2 = 0$[1] and the MVU model (with the bounded constraints) if $\rho_2 = 2$ and $\Theta = I$. Meanwhile, let $\rho_2 = 2$ and $\widetilde{D}$ to be one of the iterates in the MVE SDP subproblems (with the bounded constraints). The combined term $\langle I, -JDJ\rangle - 2\langle \widetilde{P}_1\widetilde{P}_1^T, -JDJ\rangle$ is just the objective function in the MVE SDP subproblem. In other words, MVE keeps updating $\widetilde{D}$ by solving the SDP subproblems. Therefore, Model (13) covers both MVU and MVE models as special cases. the error-bound results (see the remark after Thm. 1 and Prop. 5) obtained in next section will partially explain why under the uniform sampling rule, our model often leads to higher quality than NNPLS, MVU and MVE.

Before we go on to derive our promised error-bound results, we summarize the key points for our model (13). It is EDM based rather than SDP based as in the most existing research. The use of EDM enables us to establish the error-bound results in the next section. It inherits the nice properties in MVU and MVE models. We will also show that this model can be efficiently solved.

4 Error Bounds Under Uniform Sampling Rule

The derivation of the error bounds below, though seemingly complicated, has become standard in matrix completion literature. We will refer to the exact references whenever similar results (using similar proof techniques) have appeared before. For those who are just interested in what the error bounds mean to our problem, they can jump to the end of the section (after Thm. 1) for more interpretation.

Suppose that $X_1,\ldots,X_m$ are m independent and identically distributed (i.i.d.) random observations over Ω with the common[2] probability $1/|\Omega|$, i.e., for any $1 \le i < j \le n$,

$$\mathbb{P}\left(X_l = \frac{1}{2}(\mathbf{e}_i\mathbf{e}_j^T + \mathbf{e}_j\mathbf{e}_i^T)\right) = \frac{1}{|\Omega|}, \quad l = 1,\ldots,m.$$

Thus, for any $A \in \mathbb{S}_h^n$, we have

$$\mathbb{E}\left(\langle A,X\rangle^2\right) = \frac{1}{2|\Omega|}\|A\|^2. \tag{14}$$

Moreover, we assume that the i.i.d. noise variables in (9) have the bounded fourth moment, i.e., there exists a constant $\gamma > 0$ such that $\mathbb{E}(\xi^4) \le \gamma$.

Let $\overline{D}$ be the unknown true EDM. Suppose that the positive semidefinite matrix $-J\overline{D}J$ has the singular value decomposition (12) and $\overline{P} = [\overline{P}_1, \overline{P}_2]$ with $\overline{P}_1 \in \Re^{n\times r}$. We define the generalized geometric center subspace in $\mathbb{S}^n$ by (compare to (3)) $T := \left\{Y \in \mathbb{S}^n \mid Y\overline{P}_1 = 0\right\}$. Let $T^\perp$ be its orthogonal subspace. The orthogonal projections to the two subspaces can hence be calculated respectively by

$$\mathscr{P}_T(A) := \overline{P}_2\overline{P}_2^T A\overline{P}_2\overline{P}_2^T \quad \text{and} \quad \mathscr{P}_{T^\perp}(A) := \overline{P}_1\overline{P}_1^T A + A\overline{P}_1\overline{P}_1^T - \overline{P}_1\overline{P}_1^T A\overline{P}_1\overline{P}_1^T.$$

It is clear that we have the following orthogonal decomposition

$$A = \mathscr{P}_T(A) + \mathscr{P}_{T^\perp}(A) \quad \text{and} \quad \langle \mathscr{P}_T(A), \mathscr{P}_{T^\perp}(B)\rangle = 0 \quad \forall A,B \in \mathbb{S}^n. \tag{15}$$

Moreover, we know from the definition of $\mathscr{P}_T$ that for any $A \in \mathbb{S}^n$, $\mathscr{P}_{T^\perp}(A) = \overline{P}_1\overline{P}_1^T A + \overline{P}_2\overline{P}_2^T A\overline{P}_1\overline{P}_1^T$, which implies that $\mathrm{rank}(\mathscr{P}_{T^\perp}(A)) \le 2r$. This yields for any $A \in \mathbb{S}^n$

$$\|\mathscr{P}_{T^\perp}(A)\|_* \le \sqrt{2r}\|A\|. \tag{16}$$

[1] In this case, Model (13) can be regarded as the counterpart of the model proposed in [28].

[2] This assumption can be replaced by any positive probability $p_{ij} > 0$. But it would complicate the notation used.

For given $\rho_2 \geq 0$, define

$$\alpha(\rho_2) := \frac{1}{\sqrt{2r}}\|\overline{P}_1\overline{P}_1^T - \rho_2\Theta\|. \tag{17}$$

Let $\zeta := (\zeta_1, \ldots, \zeta_m)^T$ be the random vector defined by

$$\zeta = 2\mathscr{O}(\overline{D}^{(1/2)}) \circ \xi + \eta(\xi \circ \xi). \tag{18}$$

The non-commutative Bernstein inequality provides the probability bounds of the difference between the sum of independent random matrices and its mean under the spectral norm (see e.g., [47,56,26]). The following Bernstein inequality is taken from [41, Lemma 7], where the independent random matrices are bounded under the spectral norm or bounded under the ψ_1 Orlicz norm of random variables, i.e.,

$$\|x\|_{\psi_1} := \inf\{t > 0 \mid \mathbb{E}\exp(|x|/t) \leq e\},$$

where the constant e is the base of the natural logarithm.

Lemma 2 *Let $Z_1, \ldots, Z_m \in \mathbb{S}^n$ be independent random symmetric matrices with mean zero. Suppose that there exists $M > 0$, for all l, $\|Z_l\|_2 \leq M$ or $\big\|\|Z_l\|_2\big\|_{\psi_1} \leq M$. Denote $\sigma^2 := \|\mathbb{E}(Z_l^2)\|_2$. Then, we have for any $t > 0$,*

$$\mathbb{P}\Big(\big\|\frac{1}{m}\sum_{l=1}^m Z_l\big\|_2 \geq t\Big) \leq 2n\max\left\{\exp\left(-\frac{mt^2}{4\sigma^2}\right), \exp\left(-\frac{mt}{2M}\right)\right\}.$$

Now we are ready to study the error bounds of the model (13). It is worth to note that the optimal solution of the convex optimization problem (13) always exists, since the feasible set is nonempty and compact. Denote an optimal solution of (13) by D^*. The following result represents the first major step to derive our ultimate bound result. It contains two bounds. The first bound (19) is on the norm-squared distance between D^* and $\overline{D}$ under the observation operator $\mathscr{O}$. The second bound (20) is about the nuclear norm of $D^* - \overline{D}$. Both bounds are in terms of the Frobenius norm of $D^* - \overline{D}$.

Proposition 1 *Let $\zeta = (\zeta_1, \ldots, \zeta_m)^T$ be the random vector defined in* (18) *and $\kappa > 1$ be given. Suppose that $\rho_1 \geq \kappa\eta\big\|\frac{1}{m}\mathscr{O}^*(\zeta)\big\|_2$ and $\rho_2 \geq 0$, where $\mathscr{O}^*$ is the adjoint operator of $\mathscr{O}$. Then, we have*

$$\frac{1}{2m}\|\mathscr{O}(D^* - \overline{D})\|^2 \leq \left(\alpha(\rho_2) + \frac{2}{\kappa}\right)\rho_1\sqrt{2r}\|D^* - \overline{D}\| \tag{19}$$

and

$$\|D^* - \overline{D}\|_* \leq \frac{\kappa}{\kappa - 1}\left(\alpha(\rho_2) + 2\right)\sqrt{2r}\|D^* - \overline{D}\|. \tag{20}$$

Proof For any $D \in \mathbb{S}^n$, we know from (11) that

$$\begin{aligned}
\frac{1}{2m}\|\mathbf{y} \circ \mathbf{y} - \mathscr{O}(D)\|^2 &= \frac{1}{2m}\big\|\mathscr{O}(\overline{D}^{1/2}) \circ \mathscr{O}(\overline{D}^{1/2}) + 2\eta\mathscr{O}(\overline{D}^{1/2}) \circ \xi + \eta^2\xi \circ \xi - \mathscr{O}(D)\big\|^2 \\
&= \frac{1}{2m}\|\mathscr{O}(\overline{D}) + 2\eta\mathscr{O}(\overline{D}^{1/2}) \circ \xi + \eta^2\xi \circ \xi - \mathscr{O}(D)\|^2 = \frac{1}{2m}\|\mathscr{O}(D - \overline{D}) - \eta\zeta\|^2 \\
&= \frac{1}{2m}\|\mathscr{O}(D - \overline{D})\|^2 - \frac{\eta}{m}\langle\mathscr{O}(D - \overline{D}), \zeta\rangle + \frac{\eta^2}{2m}\|\zeta\|^2.
\end{aligned} \tag{21}$$

In particular, we have $\frac{1}{2m}\|\mathbf{y} \circ \mathbf{y} - \mathscr{O}(\overline{D})\|^2 = \frac{\eta^2}{2m}\|\zeta\|^2$. Since D^* is the optimal solution of (13) and $\overline{D}$ is also feasible, we obtain that

$$\frac{1}{2m}\|\mathbf{y} \circ \mathbf{y} - \mathscr{O}(D^*)\|^2 \leq \frac{1}{2m}\|\mathbf{y} \circ \mathbf{y} - \mathscr{O}(\overline{D})\|^2 + \rho_1\left[\langle I, -J(\overline{D} - D^*)J\rangle - \rho_2\langle\Theta, -J(\overline{D} - D^*)J\rangle\right]$$

Therefore, we know from (21) that

$$\frac{1}{2m}\|\mathscr{O}(D^*-\overline{D})\|^2 \le \frac{\eta}{m}\langle \mathscr{O}(D^*-\overline{D}),\zeta\rangle + \rho_1\left[-\langle I,-J(D^*-\overline{D})J\rangle + \rho_2\langle\Theta,-J(D^*-\overline{D})J\rangle\right]. \tag{22}$$

For the first term of the right hand side of (22), we have

$$\begin{aligned}\frac{\eta}{m}\langle \mathscr{O}(D^*-\overline{D}),\zeta\rangle &= \frac{\eta}{m}\langle D^*-\overline{D},\mathscr{O}^*(\zeta)\rangle \le \eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2\|D^*-\overline{D}\|_*\\ &= \eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2\|D^*-\overline{D}-J(D^*-\overline{D})J+J(D^*-\overline{D})J\|_*\\ &\le \eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2\left(\|D^*-\overline{D}-J(D^*-\overline{D})J\|_*+\|-J(D^*-\overline{D})J\|_*\right).\end{aligned} \tag{23}$$

By noting that $D^*, \overline{D}\in\mathbb{S}^n_h$, we know from Lemma 1 that the rank of $D^*-\overline{D}-J(D^*-\overline{D})J$ is no more than 2, which implies $\|D^*-\overline{D}-J(D^*-\overline{D})J\|_* \le \sqrt{2}\|D^*-\overline{D}-J(D^*-\overline{D})J\|$. Moreover, it follows from (4) that $\langle J(D^*-\overline{D})J, D^*-\overline{D}-J(D^*-\overline{D})J\rangle = 0$, which implies

$$\|D^*-\overline{D}\|^2 = \|D^*-\overline{D}-J(D^*-\overline{D})J\|^2 + \|J(D^*-\overline{D})J\|^2. \tag{24}$$

Thus, we have

$$\|D^*-\overline{D}-J(D^*-\overline{D})J\|_* \le \sqrt{2}\|D^*-\overline{D}\|. \tag{25}$$

By noting that $\mathscr{P}_T(-J(D^*-\overline{D})J)+\mathscr{P}_{T^\perp}(-J(D^*-\overline{D})J) = -J(D^*-\overline{D})J$, we know from (23) and (25) that

$$\begin{aligned}\frac{\eta}{m}\langle \mathscr{O}(D^*-\overline{D}),\zeta\rangle \le \left\|\frac{\eta}{m}\mathscr{O}^*(\zeta)\right\|_2\Big(&\sqrt{2}\|D^*-\overline{D}\|+\|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_*\\ &+\|\mathscr{P}_{T^\perp}(-J(D^*-\overline{D})J)\|_*\Big).\end{aligned} \tag{26}$$

Meanwhile, since for any $A\in\mathbb{S}^n$, $\|\mathscr{P}_T(A)\|_* = \|\overline{P}_2^TA\overline{P}_2\|_*$, we know from the directional derivative formula of the nuclear norm [60, Thm. 1] that

$$\begin{aligned}\|-JD^*J\|_*-\|-J\overline{D}J\|_* &\ge \langle \overline{P}_1\overline{P}_1^T,-J(D^*-\overline{D})J\rangle + \|\overline{P}_2^T(-J(D^*-\overline{D})J)\overline{P}_2\|_*\\ &= \langle \overline{P}_1\overline{P}_1^T,-J(D^*-\overline{D})J\rangle + \|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_*.\end{aligned}$$

Thus, since $-JD^*J,\ -J\overline{D}J\in\mathbb{S}^n_+$, we have $-\langle I,-J(D^*-\overline{D})J\rangle = -(\|-JD^*J\|_*-\|-J\overline{D}J\|_*)$, which implies that

$$\begin{aligned}&-\langle I,-J(D^*-\overline{D})J\rangle+\rho_2\langle\Theta,-J(D^*-\overline{D})J\rangle\\ \le\ &-\langle \overline{P}_1\overline{P}_1^T,-J(D^*-\overline{D})J\rangle - \|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_*+\rho_2\langle\Theta,-J(D^*-\overline{D})J\rangle.\end{aligned}$$

By using the decomposition (15) and the notations defined in (17), we conclude from (24) that

$$\begin{aligned}&-\langle I,-J(D^*-\overline{D})J\rangle+\rho_2\langle\Theta,-J(D^*-\overline{D})J\rangle \le -\langle \overline{P}_1\overline{P}_1^T-\rho_2\Theta,-J(D^*-\overline{D})J\rangle - \|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_*\\ \le\ &\|\overline{P}_1\overline{P}_1^T-\rho_2\Theta\|\|J(D^*-\overline{D})J\| - \|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_* \le \alpha(\rho_2)\sqrt{2r}\|D^*-\overline{D}\| - \|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_*.\end{aligned}$$

Thus, together with (26), we know from (22) that

$$\begin{aligned}\frac{1}{2m}\|\mathscr{O}(D^*-\overline{D})\|^2 \le\ &\left(\sqrt{2}\eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2+\sqrt{2r}\rho_1\alpha(\rho_2)\right)\|D^*-\overline{D}\|\\ +\ \eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2&\|\mathscr{P}_{T^\perp}(-J(D^*-\overline{D})J)\|_*-\left(\rho_1-\eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2\right)\|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_*.\end{aligned} \tag{27}$$

Since $\eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2 \leq \frac{\rho_1}{\kappa}$ and $\kappa > 1$, we know from (16) and (24) that

$$\begin{aligned}\frac{1}{2m}\|\mathscr{O}(D^*-\overline{D})\|^2 &\leq \left(\frac{1}{\kappa}\sqrt{2}+\alpha(\rho_2)\sqrt{2r}\right)\rho_1\|D^*-\overline{D}\| + \frac{1}{\kappa}\sqrt{2r}\rho_1\|D^*-\overline{D}\| - \left(1-\frac{1}{\kappa}\right)\rho_1\|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_* \\ &\leq \left(\frac{1}{\kappa}(\sqrt{2}+\sqrt{2r})+\alpha(\rho_2)\sqrt{2r}\right)\rho_1\|D^*-\overline{D}\| - \frac{\kappa-1}{\kappa}\rho_1\|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_* \qquad (28)\\ &\leq \left(\frac{1}{\kappa}(\sqrt{2}+\sqrt{2r})+\alpha(\rho_2)\sqrt{2r}\right)\rho_1\|D^*-\overline{D}\|. \qquad (29)\end{aligned}$$

Since $r \geq 1$, the desired inequality (19) follows from (29), directly.

Next we shall show that (20) also holds. By (28), we have

$$\|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_* \leq \frac{\kappa}{\kappa-1}\left(\frac{\sqrt{2}}{\kappa}+\left(\alpha(\rho_2)+\frac{1}{\kappa}\right)\sqrt{2r}\right)\|D^*-\overline{D}\|.$$

Therefore, by combining with (25) and (16), we know from the decomposition (15) that

$$\begin{aligned}\|D^*-\overline{D}\|_* &\leq \|D^*-\overline{D}-J(D^*-\overline{D})J\|_* + \|\mathscr{P}_{T^\perp}(-J(D^*-\overline{D})J)\|_* + \|\mathscr{P}_T(-J(D^*-\overline{D})J)\|_* \\ &\leq (\sqrt{2}+\sqrt{2r})\|D^*-\overline{D}\| + \frac{\kappa}{\kappa-1}\left(\frac{\sqrt{2}}{\kappa}+\left(\alpha(\rho_2)+\frac{1}{\kappa}\right)\sqrt{2r}\right)\|D^*-\overline{D}\|.\end{aligned}$$

Finally, since $r \geq 1$, we conclude that

$$\begin{aligned}\|D^*-\overline{D}\|_* &\leq \frac{\kappa}{\kappa-1}\sqrt{2}\|D^*-\overline{D}\| + \frac{\kappa}{\kappa-1}\left(\alpha(\rho_2)+1\right)\sqrt{2r}\|D^*-\overline{D}\| \\ &\leq \frac{\kappa}{\kappa-1}\left(\alpha(\rho_2)+2\right)\sqrt{2r}\|D^*-\overline{D}\|.\end{aligned}$$

This completes the proof. □

The second major technical result below shows that the sampling operator $\mathscr{O}$ satisfies the following restricted strong convexity [41] in the set $\mathscr{C}(\tau)$ for any $\tau > 0$, where

$$\mathscr{C}(\tau) := \left\{A \in \mathbb{S}_h^n \mid \|A\|_\infty = \frac{1}{\sqrt{2}},\ \|A\|_* \leq \sqrt{\tau}\|A\|,\ \mathbb{E}(\langle A,X\rangle^2) \geq \sqrt{\frac{256\log(2n)}{m\log(2)}}\right\}.$$

Lemma 3 *Let $\tau > 0$ be given. Suppose that $m > C_1 n\log(2n)$, where $C_1 > 1$ is a constant. Then, there exists a constant $C_2 > 0$ such that for any $A \in \mathscr{C}(\tau)$, the following inequality holds with probability at least $1-1/n$.*

$$\frac{1}{m}\|\mathscr{O}(A)\|^2 \geq \frac{1}{2}\mathbb{E}\left(\langle A,X\rangle^2\right) - 256C_2\tau|\Omega|\frac{\log(2n)}{nm}.$$

Proof. Firstly, we shall show that for any $A \in \mathscr{C}(\tau)$, the following inequality holds with probability at least $1-1/n$,

$$\frac{1}{m}\|\mathscr{O}(A)\|^2 \geq \frac{1}{2}\mathbb{E}\left(\langle A,X\rangle^2\right) - 256\tau|\Omega|\left(\mathbb{E}\left(\left\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\right\|_2\right)\right)^2,$$

where $\varepsilon = (\varepsilon_1,\ldots,\varepsilon_m)^T \in \Re^m$ with $\{\varepsilon_1,\ldots,\varepsilon_m\}$ is an i.i.d. Rademacher sequence, i.e., a sequence of i.i.d. Bernoulli random variables taking the values 1 and -1 with probability $1/2$. This part of proof is similar with that of Lemma 12 in [32] (see also [39, Lemma 2]). However, we include the proof here for completion.

Denote $\Sigma := 256r|\Omega|\left(\mathbb{E}\left(\left\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\right\|_2\right)\right)^2$. We will show that the probability of the following "bad" events is small

$$\mathscr{B} := \left\{\exists A \in \mathscr{C}(\tau) \text{ such that } \left|\frac{1}{m}\|\mathscr{O}(A)\|^2 - \mathbb{E}\left(\langle A,X\rangle^2\right)\right| > \frac{1}{2}\mathbb{E}\left(\langle A,X\rangle^2\right) + \Sigma\right\}.$$

It is clear that the events interested are included in $\mathscr{B}$. Next, we will use a standard peeling argument to estimate the probability of $\mathscr{B}$. For any $\nu > 0$, we have

$$\mathscr{C}(\tau) \subseteq \bigcup_{k=1}^{\infty} \left\{ A \in \mathscr{C}(\tau) \mid 2^{k-1}\nu \leq \mathbb{E}\left(\langle A, X\rangle^2\right) \leq 2^k \nu \right\}.$$

Thus, if the event $\mathscr{B}$ holds for some $A \in \mathscr{C}(\tau)$, then there exists some $k \in \mathbb{N}$ such that $2^k\nu \geq \mathbb{E}\left(\langle A, X\rangle^2\right) \geq 2^{k-1}\nu$. Therefore, we have

$$\left| \frac{1}{m}\|\mathscr{O}(A)\|^2 - \mathbb{E}\left(\langle A, X\rangle^2\right) \right| > \frac{1}{2}2^{k-1}\nu + \Sigma = 2^{k-2}\nu + \Sigma.$$

This implies that $\mathscr{B} \subseteq \bigcup_{k=1}^{\infty} \mathscr{B}_k$, where for each k,

$$\mathscr{B}_k := \left\{ \exists A \in \mathscr{C}(\tau) \text{ such that } \left| \frac{1}{m}\|\mathscr{O}(A)\|^2 - \mathbb{E}\left(\langle A, X\rangle^2\right) \right| > 2^{k-2}\nu + \Sigma,\ \mathbb{E}\left(\langle A, X\rangle^2\right) \leq 2^k \nu \right\}.$$

We shall estimated the probability of each $\mathscr{B}_k$. For any given $\Upsilon > 0$, define the set

$$\mathscr{C}(\tau; \Upsilon) := \left\{ A \in \mathscr{C}(\tau) \mid \mathbb{E}\left(\langle A, X\rangle^2\right) \leq \Upsilon \right\}.$$

For any given $\Upsilon > 0$, denote $Z_\Upsilon := \sup_{A \in \mathscr{C}(\tau;\Upsilon)} \left| \frac{1}{m}\|\mathscr{O}(A)\|^2 - \mathbb{E}\left(\langle A, X\rangle^2\right) \right|$. We know from (10), the definition of the observation operator $\mathscr{O}$, that

$$\frac{1}{m}\|\mathscr{O}(A)\|^2 - \mathbb{E}\left(\langle A, X\rangle^2\right) = \frac{1}{m}\sum_{l=1}^{m} \langle A, X_l\rangle^2 - \mathbb{E}\left(\langle A, X\rangle^2\right).$$

Meanwhile, since $\|A\|_\infty = 1/\sqrt{2}$, we have for each $l \in \{1, \ldots, m\}$, $\left| \langle A, X_l\rangle^2 - \mathbb{E}\left(\langle A, X\rangle^2\right) \right| \leq 2\|A\|_\infty^2 = 1$. Thus, it follows from Massart's concentration inequality [11, Thm. 14.2] that

$$\mathbb{P}\left(Z_\Upsilon \geq \mathbb{E}(Z_\Upsilon) + \frac{\Upsilon}{8} \right) \leq \exp\left(\frac{-m\Upsilon^2}{512} \right). \tag{30}$$

By applying the standard Rademacher symmetrization [33, Thm. 2.1], we obtain that

$$\mathbb{E}(Z_\Upsilon) = \mathbb{E}\left(\sup_{A \in \mathscr{C}(\tau;\Upsilon)} \left| \frac{1}{m}\sum_{l=1}^{m} \langle A, X_l\rangle^2 - \mathbb{E}\left(\langle A, X\rangle^2\right) \right| \right) \leq 2\mathbb{E}\left(\sup_{A \in \mathscr{C}(\tau;\Upsilon)} \left| \frac{1}{m}\sum_{l=1}^{m} \varepsilon_l \langle A, X_l\rangle^2 \right| \right),$$

where $\{\varepsilon_1, \ldots, \varepsilon_m\}$ is an i.i.d. Rademacher sequence. Again, since $\|A\|_\infty = 1/\sqrt{2}$, we know that $|\langle A, X_i\rangle| \leq \|A\|_\infty < 1$. Thus, it follows from the contraction inequality (see e.g., [36, Thm. 4.12]) that

$$\mathbb{E}(Z_\Upsilon) \leq 8\mathbb{E}\left(\sup_{A \in \mathscr{C}(\tau;\Upsilon)} \left| \frac{1}{m}\sum_{l=1}^{m} \varepsilon_l \langle A, X_l\rangle \right| \right) = 8\mathbb{E}\left(\sup_{A \in \mathscr{C}(\tau;\Upsilon)} \left| \langle \frac{1}{m}\mathscr{O}^*(\varepsilon), A\rangle \right| \right) \leq 8\mathbb{E}\left(\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\| \right) \left(\sup_{A \in \mathscr{C}(\tau;\Upsilon)} \|A\|_* \right).$$

For any $A \in \mathscr{C}(\tau; \Upsilon)$, we have

$$\|A\|_* \leq \sqrt{\tau}\|A\| = \sqrt{2\tau|\Omega|\mathbb{E}\left(\langle A, X\rangle^2\right)} \leq \sqrt{2\tau|\Omega|\Upsilon}.$$

Thus, we obtain that

$$\mathbb{E}(Z_\Upsilon) + \frac{\Upsilon}{8} \leq 8\mathbb{E}\left(\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\| \right) \left(\sup_{A \in \mathscr{C}(\tau;\Upsilon)} \|A\|_* \right) + \frac{\Upsilon}{8} \leq 8\mathbb{E}\left(\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\| \right) \sqrt{2\tau|\Omega|\Upsilon} + \frac{\Upsilon}{8}.$$

Since $256\tau|\Omega|\left(\mathbb{E}\left(\frac{1}{m}\mathscr{O}^*(\varepsilon)\right)\right)^2+\frac{\Upsilon}{8}\geq 8\mathbb{E}\left(\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\|\right)\sqrt{2\tau|\Omega|\Upsilon}$, we have

$$\mathbb{E}(Z_\Upsilon)+\frac{\Upsilon}{8}\leq 256\tau|\Omega|\left(\mathbb{E}\left(\frac{1}{m}\mathscr{O}^*(\varepsilon)\right)\right)^2+\frac{\Upsilon}{4}.$$

It follows from (30) that

$$\mathbb{P}\Big(Z_\Upsilon\geq\frac{\Upsilon}{4}+256\tau|\Omega|\big(\mathbb{E}\big(\frac{1}{m}\mathscr{O}^*(\varepsilon)\big)\big)^2\Big)\leq\mathbb{P}\left(Z_\Upsilon\geq\mathbb{E}(Z_\Upsilon)+\frac{\Upsilon}{8}\right)\leq\exp\left(\frac{-m\Upsilon^2}{512}\right).$$

By choosing $\Upsilon=2^k\nu$, it is easy to seen that for each k, if the even $\mathscr{B}_k$ occurs, then $Z_\Upsilon\geq\frac{\Upsilon}{4}+256\tau|\Omega|\big(\mathbb{E}\big(\frac{1}{m}\mathscr{O}^*(\varepsilon)\big)\big)^2$, which implies that

$$\mathbb{P}(\mathscr{B}_k)\leq\mathbb{P}\Big(Z_\Upsilon\geq\frac{\Upsilon}{4}+256\tau|\Omega|\big(\mathbb{E}\big(\frac{1}{m}\mathscr{O}^*(\varepsilon)\big)\big)^2\Big)\leq\exp\left(\frac{-4^k\nu^2m}{512}\right).$$

By noting that $\log(x)<x$ for any $x>1$, we conclude that

$$\mathbb{P}(\mathscr{B})\leq\sum_{k=1}^{\infty}\mathbb{P}(\mathscr{B}_k)\leq\sum_{k=1}^{\infty}\exp\left(\frac{-4^k\nu^2m}{512}\right)<\sum_{k=1}^{\infty}\exp\left(\frac{-\log(4)k\nu^2m}{512}\right)=\frac{\exp\left(\frac{-\log(2)k\nu^2m}{256}\right)}{1-\exp\left(\frac{-\log(2)k\nu^2m}{256}\right)}.$$

Choosing $\nu=\sqrt{\dfrac{256\log(2n)}{m\log(2)}}$, it yields $\mathbb{P}(\mathscr{B})\leq 1/(2n-1)\leq 1/n$.

Finally, the lemma then follows if we prove that for $m>C_1n\log n$ with $C_1>1$, there exists a constant $C_1'>0$ such that

$$\mathbb{E}\Big(\big\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\big\|_2\Big)\leq C_1'\sqrt{\frac{\log(2n)}{mn}}. \tag{31}$$

The following proof is similar with that of Lemma 7 [31] (see e.g., [32, Lemma 6]). We include it again for the seek of completeness. Denote $Z_l:=\varepsilon_lX_l$, $l=1,\ldots,m$. Since $\{\varepsilon_1,\ldots,\varepsilon_m\}$ is an i.i.d. Rademacher sequence, we have $\|Z_l\|_2=1/2$ for all l. Moreover,

$$\|\mathbb{E}(Z_l^2)\|_2=\|\mathbb{E}(\varepsilon_l^2X_l^2)\|_2=\|\mathbb{E}(X_l^2)\|_2=\frac{1}{4|\Omega|}\big\|\sum_{1\leq i<j\leq n}(\mathbf{e}_i\mathbf{e}_j^T+\mathbf{e}_j\mathbf{e}_i^T)^2\big\|_2=\frac{1}{4|\Omega|}(n-1)=\frac{1}{2n}.$$

By applying the Bernstein inequality (Lemma 2), we obtain the following tail bound for any $t>0$,

$$\mathbb{P}\Big(\big\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\big\|_2\geq t\Big)\leq 2n\max\left\{\exp\left(-\frac{nmt^2}{2}\right),\exp(-mt)\right\}. \tag{32}$$

By Hölder's inequality, we have

$$\begin{aligned}
&\mathbb{E}\Big(\big\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\big\|_2\Big)\leq\left(\mathbb{E}\Big(\big\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\big\|_2^{2\log(2n)}\Big)\right)^{\frac{1}{2\log(2n)}}=\left(\int_0^\infty\mathbb{P}\Big(\big\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\big\|_2\geq t^{\frac{1}{2\log(2n)}}\Big)dt\right)^{\frac{1}{2\log(2n)}}\\
&\leq\left(2n\int_0^\infty\exp\left(-\frac{1}{2}nmt^{\frac{1}{\log(2n)}}\right)dt+2n\int_0^\infty\exp\left(-mt^{\frac{1}{2\log(2n)}}\right)dt\right)^{\frac{1}{2\log(2n)}}\\
&=e^{1/2}\left(\log(2n)(\frac{nm}{2})^{-\log(2n)}\Gamma(\log(2n))+2\log(2n)m^{-2\log(2n)}\Gamma(2\log(2n))\right)^{\frac{1}{2\log(2n)}}.
\end{aligned}\tag{33}$$

Since for $x\geq 2$, $\Gamma(x)\leq(x/2)^{x-1}$, we obtain from (33) that for $n\geq 4$,

$$\mathbb{E}\Big(\big\|\frac{1}{m}\mathscr{O}^*(\varepsilon)\big\|_2\Big)\leq e^{1/2}\left(2\left(\sqrt{\frac{\log(2n)}{nm}}\right)^{2\log(2n)}+2\left(\frac{\log(2n)}{m}\right)^{2\log(2n)}\right)^{\frac{1}{2\log(2n)}}. \tag{34}$$

Since $m > C_1 n\log(2n)$ and $C_1 > 1$, we have $\sqrt{\frac{\log(2n)}{nm}} > \frac{\sqrt{C_1}\log(2n)}{m} > \frac{\log(2n)}{m}$. Let $C_1' = e^{1/2}2^{1/\log 2}$. It follows from (34) that the inequality (31) holds. □

Next, combining Proposition 1 and Lemma 3 leads to the following result.

Proposition 2 *Let $\kappa > 1$ be given. Suppose that $\rho_1 \geq \kappa\eta\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2$ and $\rho_2 \geq 0$. Furthermore, assume that $m > C_1 n\log(2n)$ for some constant $C_1 > 1$. Then, there exists a constant $C_3 > 0$ such that with probability at least $1 - 1/n$,*

$$\frac{\|D^* - \overline{D}\|^2}{|\Omega|} \leq C_3 \max\left\{r|\Omega|\Big(\big(\alpha(\rho_2) + \frac{2}{\kappa}\big)^2\rho_1^2 + \big(\frac{\kappa}{\kappa-1}\big)^2\big(\alpha(\rho_2)+2\big)^2 b^2\frac{\log(2n)}{nm}\Big), b^2\sqrt{\frac{\log(2n)}{m}}\right\}.$$

Proof. Since $\|\overline{D}\|_\infty \leq b$, we know that $\|D^* - \overline{D}\|_\infty \leq 2b$. Consider the following two cases.

Case 1: If $\mathbb{E}\left(\langle D^* - \overline{D}, X\rangle^2\right) < 8b^2\sqrt{\frac{256\log(2n)}{m\log(2)}}$, then we know from (14) that

$$\frac{\|D^* - \overline{D}\|^2}{|\Omega|} < 16b^2\sqrt{\frac{256\log(2n)}{m\log(2)}} = 16b^2\sqrt{\frac{256}{\log(2)}}\sqrt{\frac{\log(2n)}{m}}.$$

Case 2: If $\mathbb{E}\left(\langle D^* - \overline{D}, X\rangle^2\right) \geq 8b^2\sqrt{\frac{256\log(2n)}{m\log(2)}}$, then we know from (20) that $(D^* - \overline{D})/\sqrt{2}\|D^* - \overline{D}\|_\infty \in \mathscr{C}(\tau)$ with $\tau = 2r(\frac{\kappa}{\kappa-1})^2\left(\alpha(\rho_2)+2\right)^2$. Thus, it follows from Lemma 3 that there exists a constant $C_2' > 0$ such that with probability at least $1 - 1/n$,

$$\frac{1}{2}\mathbb{E}\left(\langle D^* - \overline{D}, X\rangle^2\right) \leq \frac{1}{m}\|\mathscr{O}(D^* - \overline{D})\|^2 + 2048C_2'b^2\tau|\Omega|\frac{\log(2n)}{nm}.$$

Thus, we know from (14) and (19) in Proposition 1 that

$$\begin{aligned}\frac{\|D^* - \overline{D}\|^2}{2|\Omega|} = \mathbb{E}\left(\langle D^* - \overline{D}, X\rangle^2\right) &\leq \frac{2}{m}\|\mathscr{O}(D^* - \overline{D})\|^2 + 4096C_2'b^2\tau|\Omega|\frac{\log(2n)}{nm}\\ &\leq 4\sqrt{2r}\left(\alpha(\rho_2) + \frac{2}{\kappa}\right)\rho_1\|D^* - \overline{D}\| + 4096C_2'b^2\tau|\Omega|\frac{\log(2n)}{nm}\\ &\leq \frac{\|D^* - \overline{D}\|^2}{4|\Omega|} + 32r|\Omega|\left(\alpha(\rho_2) + \frac{2}{\kappa}\right)^2\rho_1^2 + 4096C_2'b^2\tau|\Omega|\frac{\log(2n)}{nm}.\end{aligned}$$

By substituting τ, we obtain that there exists a constant $C_3' > 0$ such that

$$\frac{\|D^* - \overline{D}\|^2}{|\Omega|} \leq C_3'r|\Omega|\left(\left(\alpha(\rho_2) + \frac{2}{\kappa}\right)^2\rho_1^2 + \left(\frac{\kappa}{\kappa-1}\right)^2\left(\alpha(\rho_2)+2\right)^2 b^2\frac{\log(2n)}{nm}\right).$$

The result then follows by combining these two cases. □

This bound depends on the model parameters ρ_1 and ρ_2. In order to establish an explicit error bound, we need to estimate ρ_1 (ρ_2 will be estimated later), which depends on the quantity $\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2$, where $\zeta = (\zeta_1, \ldots, \zeta_m)^T \in \Re^m$ with ζ_l, $l = 1, \ldots, m$ are i.i.d. random variables given by (18). To this end, from now on, we always assume that the i.i.d. random noises ξ_l, $l = 1, \ldots, m$ in the sampling model (11) satisfy the following sub-Gaussian tail condition.

Assumption 3 *There exist positive constants K_1 and K_2 such that for all $t > 0$,*

$$\mathbb{P}\left(|\xi_l| \geq t\right) \leq K_1\exp\left(-t^2/K_2\right).$$

By applying the Bernstein inequality (Lemma 2), we have

Proposition 4 *Let $\zeta = (\zeta_1, \ldots, \zeta_m)^T$ be the random vector defined in* (18)*. Assume that the noise magnitude control factor satisfies $\eta < \omega := \|\mathscr{O}(\overline{D}^{(1/2)})\|_\infty$. Suppose that there exists $C_1 > 1$ such that $m > C_1 n \log(n)$. Then, there exists a constant $C_3 > 0$ such that with probability at least $1 - 1/n$,*

$$\left\| \frac{1}{m} \mathscr{O}^*(\zeta) \right\|_2 \leq C_3 \omega \sqrt{\frac{\log(2n)}{nm}}. \tag{35}$$

Proof. From (18), the definition of ζ, we know that

$$\left\| \frac{1}{m} \mathscr{O}^*(\zeta) \right\|_2 \leq 2\omega \left\| \frac{1}{m} \mathscr{O}^*(\xi) \right\|_2 + \eta \left\| \frac{1}{m} \mathscr{O}^*(\xi \circ \xi) \right\|_2,$$

where $\omega := \left\| \mathscr{O}(\overline{D}^{(1/2)}) \right\|_\infty$. Therefore, for any given $t_1, t_2 > 0$, we have

$$\mathbb{P}\left(\left\| \frac{1}{m} \mathscr{O}^*(\zeta) \right\|_2 \geq 2\omega t_1 + \eta t_2 \right) \leq \mathbb{P}\left(\left\| \frac{1}{m} \mathscr{O}^*(\xi) \right\|_2 \geq t_1 \right) + \mathbb{P}\left(\left\| \frac{1}{m} \mathscr{O}^*(\xi \circ \xi) \right\|_2 \geq t_2 \right). \tag{36}$$

Recall that $\dfrac{1}{m} \mathscr{O}^*(\xi) = \dfrac{1}{m} \sum\limits_{l=1}^{m} \xi_l X_l$. Denote $Z_l := \xi_l X_l$, $l = 1, \ldots, m$. Since $\mathbb{E}(\xi_l) = 0$ and ξ_l and X_l are independent, we have $\mathbb{E}(Z_l) = 0$ for all l. Also, we have

$$\|Z_l\|_2 \leq \|Z_l\| = |\xi_l|, \quad l = 1, \ldots, m,$$

which implies that $\|\|Z_l\|_2\|_{\psi_1} \leq \|\xi_l\|_{\psi_1}$. Since ξ_l is sub-Gaussian, we know that there exists a constant $M_1 > 0$ such that $\|\xi_l\|_{\psi_1} \leq M_1$, $l = 1, \ldots, m$ (see e.g., [58, Section 5.2.3]). Meanwhile, for each l, it follows from $\mathbb{E}(\xi_l^2) = 1$, (14) and $|\Omega| = n(n-1)/2$ that

$$\|\mathbb{E}(Z_l^2)\|_2 = \|\mathbb{E}(\xi_l^2 X_l^2)\|_2 = \|\mathbb{E}(X_l^2)\|_2 = \frac{1}{4|\Omega|} \Big\| \sum_{1 \leq i < j \leq n} (\mathbf{e}_i \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_i^T)^2 \Big\|_2 = \frac{1}{4|\Omega|}(n-1) = \frac{1}{2n}.$$

For $\dfrac{1}{m} \mathscr{O}^*(\xi \circ \xi) = \dfrac{1}{m} \sum\limits_{l=1}^{m} \xi_l^2 X_l$, denote $Y_l := \xi_l^2 X_l - \mathbb{E}(X_l)$, $l = 1, \ldots, m$, where

$$\mathbb{E}(X_l) = \frac{1}{2|\Omega|} \sum_{1 \leq i < j \leq n} (\mathbf{e}_i \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_i^T) = \frac{1}{2|\Omega|}(\mathbf{1}\mathbf{1}^T - I).$$

It is clear that for each l, $\|\mathbb{E}(X_l)\| = 1$ and $\|\mathbb{E}(X_l)\|_2 = 1/n$. Therefore, since $\mathbb{E}(\xi_l^2) = 1$, we know that $\mathbb{E}(Y_l) = 0$ for all l. Moreover, we have

$$\|Y_l\|_2 = \|\xi_l^2 X_l - \mathbb{E}(X_l)\|_2 \leq \|\xi_l^2 X_l - \mathbb{E}(X_l)\| \leq \xi_l^2 + \|\mathbb{E}(X_l)\| = \xi_l^2 + 1.$$

Thus, we have $\|\|Y_l\|_2\|_{\psi_1} \leq \left\|\xi_l^2\right\|_{\psi_1} + 1$. From [58, Lemma 5.14], we know that the random variable ξ_l is sub-Gaussian if and only if ξ_l^2 is sub-exponential, which implies there exists $M_2 > 0$ such that $\|\xi_l^2\|_{\psi_1} \leq M_2$ [58, see e.g., Section 5.2.3 and 5.2.4]. Therefore, $\|\|Y_l\|_2\|_{\psi_1} \leq M_2 + 1$. Meanwhile, we have

$$\begin{aligned} \|\mathbb{E}(Y_l^2)\|_2 &= \left\|\mathbb{E}\left((\xi_l^2 X_l - \mathbb{E}(X_l))(\xi_l^2 X_l - \mathbb{E}(X_l))\right)\right\|_2 = \left\|\mathbb{E}\left(\xi_l^4 X_l^2\right) - \mathbb{E}(X_l)\mathbb{E}(X_l)\right\|_2 \\ &\leq \left\|\mathbb{E}\left(\xi_l^4 X_l^2\right)\right\|_2 + \|\mathbb{E}(X_l)\mathbb{E}(X_l)\|_2 = \left\|\mathbb{E}\left(\xi_l^4 X_l^2\right)\right\|_2 + \|\mathbb{E}(X_l)\|_2^2 = \frac{\gamma}{2n} + \frac{1}{n^2}. \end{aligned}$$

Therefore, for the sufficiently large n, we always have $\|\mathbb{E}(Y_l^2)\|_2 \le \gamma/n$. Denote $M_3 = \max\{M_1, M_2+1\}$ and $C_3' = \max\{1/2, \gamma\}$. We know from Lemma 2 that for any given $t_1, t_2 > 0$

$$\mathbb{P}\left(\left\|\frac{1}{m}\mathscr{O}^*(\xi)\right\|_2 \ge t_1\right) \le 2n\max\left\{\exp\left(-\frac{nmt_1^2}{4C_3'}\right), \exp\left(-\frac{mt_1}{2M_3}\right)\right\} \tag{37}$$

and

$$\mathbb{P}\left(\left\|\frac{1}{m}\mathscr{O}^*(\xi\circ\xi)\right\|_2 \ge t_2\right) \le 2n\max\left\{\exp\left(-\frac{nmt_2^2}{4C_3'}\right), \exp\left(-\frac{mt_2}{2M_3}\right)\right\}. \tag{38}$$

By choosing $t_1 = 2\sqrt{2}\sqrt{\dfrac{C_3'\log(2n)}{nm}}$ and $t_2 = \omega t_1/\eta$, we know from $m > C_1 n\log(2n)$ (for the sufficiently large C_1) that the first terms of the right hand sides of (37) and (38) both dominate the second terms, respectively. Thus, since $\eta < \omega$, we have

$$\mathbb{P}\left(\left\|\frac{1}{m}\mathscr{O}^*(\xi)\right\|_2 \ge t_1\right) \le \frac{1}{2n} \quad \text{and} \quad \mathbb{P}\left(\left\|\frac{1}{m}\mathscr{O}^*(\xi\circ\xi)\right\|_2 \ge t_2\right) \le \frac{1}{2n}.$$

Finally, it follows from (36) that

$$\mathbb{P}\left(\left\|\frac{1}{m}\mathscr{O}^*(\zeta)\right\|_2 \ge 12\omega\sqrt{\frac{C_3'\log(2n)}{nm}}\right) \le \frac{1}{n}.$$

The proof is completed. □

This result suggests that ρ_1 can take the particular value:

$$\rho_1 = \kappa\eta\omega C_3\sqrt{\frac{\log(2n)}{mn}}, \tag{39}$$

where $\kappa > 1$. Our final step is to combine Proposition 2 and Proposition 4 to get the following error bound.

Theorem 1 *Suppose that the noise magnitude control factor satisfies $\eta < \omega = \|\mathscr{O}(\overline{D}^{(1/2)})\|_\infty$. Assume the sample size m satisfies $m > C_1 n\log(2n)$ for some constant $C_1 > 1$. For any given $\kappa > 1$, let ρ_1 be given by* (39) *and $\rho_2 \ge 0$. Then, there exists a constant $C_4 > 0$ such that with probability at least $1 - 2/n$,*

$$\frac{\|D^* - \overline{D}\|^2}{|\Omega|} \le C_4\Big(\big(\kappa\alpha(\rho_2)+2\big)^2\eta^2\omega^2 + \frac{\kappa^2}{(\kappa-1)^2}\big(\alpha(\rho_2)+2\big)^2 b^2\Big)\frac{r|\Omega|\log(2n)}{nm}. \tag{40}$$

For MVU, since $\rho_2 = 2$ and $\Theta = I$, by (17), we have $(\alpha_{MVU})^2 = \frac{1}{2r}\|\overline{P}_1\overline{P}_1^T - 2I\|^2 \ge \frac{1}{2}$. For MVE and our EDM models, since $\Theta = \widetilde{P}_1\widetilde{P}_1^T$, the only remaining unknown parameter in (40) is ρ_2 though $\alpha(\rho_2)$. It follows from (17) that

$$(\alpha(\rho_2))^2 = \frac{1}{2r}\left(\|\overline{P}_1\overline{P}_1^T\|^2 - 2\rho_2\langle\overline{P}_1\overline{P}_1^T, \widetilde{P}_1\widetilde{P}_1^T\rangle + \rho_2^2\|\widetilde{P}_1\widetilde{P}_1^T\|^2\right). \tag{41}$$

Since $\|\overline{P}_1\overline{P}_1^T\|^2 = \|\widetilde{P}_1\widetilde{P}_1^T\|^2 = r$ and $\langle\overline{P}_1\overline{P}_1^T, \widetilde{P}_1\widetilde{P}_1^T\rangle \ge 0$, we can bound $\alpha(\rho_2)$ by $(\alpha(\rho_2))^2 \le \frac{1}{2}(1+\rho_2^2)$. This bound suggest that $\rho_2 = 0$ (corresponding to the nuclear norm minimization) would lead to a lower bound than MVU. In fact, the best choice ρ_2^* for ρ_2 is when it minimizes the right-hand side bound in (40) and is given by (42) in Subsect. 5.1, where we will show that $\rho_2 = 1$ is a better choice than both $\rho_2 = 0$ and $\rho_2 = 2$.

The major message from Thm. 1 is as follows. We know that if the true Euclidean distance matrix $\overline{D}$ is bounded, and the noises are small (less than the true distances), in order to control the estimation error, we only need samples with the size m of the order $r(n-1)\log(2n)/2$, since $|\Omega| = n(n-1)/2$. Note that, $r = \mathrm{rank}(J\overline{D}J)$ is usually small (2 or 3). Therefore, the sample size m is much smaller than $n(n-1)/2$, the total number of the off-diagonal entries. Moreover, since the degree[3] of freedom of n by n symmetric hollow matrix with

[3] We know from Lemma 1 that the rank of the true EDM $\mathrm{rank}(\overline{D}) = O(r)$.

rank r is $n(r-1)-r(r-1)/2$, the sample size m is close to the degree of freedom if the matrix size n is large enough. However, we emphasize that one cannot obtain exact recovery from the bound (40) even without noise, i.e., $\eta = 0$. As mentioned in [41], this phenomenon is unavoidable due to lack of identifiability. For instance, consider the EDM $\overline{D}$ and the perturbed EDM $\widetilde{D} = \overline{D} + \varepsilon \mathbf{e}_1 \mathbf{e}_1^T$. Thus, with high probability, $\mathscr{O}(D^*) = \mathscr{O}(\widetilde{D})$, which implies that it is impossible to distinguish two EDMs even if they are noiseless. If one is interested only in exact recovery in the noiseless setting, some addition assumptions such as the matrix incoherence conditions are necessary.

We would like to mention a relevant result by Keshavan et al. [29], who proposed their `OptSpace` algorithm for matrix completion problem. For the Gaussian noise and squared matrices case, the corresponding error bound in [29] reads as

$$\|D^* - \overline{D}\| \leq C\kappa(\overline{D})^2 \eta \sqrt{\frac{rn}{m}},$$

with high probability, where $C > 0$ is a constant and $\kappa(\overline{D})$ is the condition number of the true unknown matrix $\overline{D}$. It seems that the resulting bound is stronger than ours for the case of the matrix completion problem. However, since the condition number for a matrix with rank larger than one can be arbitrarily large, the bound is not necessarily stronger than that proved in Thm. 1.

Finally, we also want to compare our error bound result in Thm. 1 with the result obtained in [57, Section 7]. The results obtained in [57] is for the sensor network localization where some location points are fixed as anchors. This makes the corresponding analysis completely different. Moreover, roughly speaking, the estimation error of the second-order cone relaxation is bounded by the square root of the distance error, which is a function of estimator (see [57, Proposition 7.2]). This means that the right-hand side of the error bound obtained by [57] depends on the resulting estimator. However, the error bound proved in Thm. 1 only depends on the initial input data of problems.

5 Model Parameter Estimation and the Algorithm

In general, the choice of model parameters can be tailored to a particular application. A very useful property about our model (13) is that we can derive a theoretical estimate, which serves as a guideline for the choice of the model parameters in our implementation. In particular, we set ρ_1 by (39) and prove that $\rho_2 = 1$ is a better choice than both the case $\rho_2 = 0$ (corresponding to the nuclear norm minimization) and $\rho_2 = 2$ (MVE model). The first part of this section is to study the optimal choice of ρ_2 and the second part briefly introduces a convergent 3-block alternating direction method of multipliers (ADMM) algorithm, which is particularly suitable to our model.

5.1 Optimal Estimate of ρ_2

It is easy to see from the inequality (40) that in order to reduce the estimation error, the best choice ρ_2^* of ρ_2 is the minimum of $\alpha(\rho_2)$. We obtain from (41) that $\rho_2^* \geq 0$ and

$$\rho_2^* = \frac{1}{r}\langle \overline{P}_1 \overline{P}_1^T, \widetilde{P}_1 \widetilde{P}_1^T \rangle = 1 + \frac{1}{r}\langle \overline{P}_1 \overline{P}_1^T, \widetilde{P}_1 \widetilde{P}_1^T - \overline{P}_1 \overline{P}_1^T \rangle. \tag{42}$$

The key technique that we are going to use to estimate ρ_2^* is the Löwner operator. We express both the terms $\widetilde{P}_1 \widetilde{P}_1^T$ and $\overline{P}_1 \overline{P}_1^T$ as the values from the operator. We then show that the Löwner operator admits a first-order approximation, which will indicate the magnitude of ρ_2^*. The technique is extensively used by [39]. We briefly describe it below.

Denote $\delta := \|\widetilde{D} - \overline{D}\|$. Assume that $\delta < \overline{\lambda}_r/2$. Define the scalar function $\phi : \Re \to \Re$ by

$$\phi(x) = \begin{cases} 1 & \text{if } x \geq \overline{\lambda}_r - \delta, \\ \dfrac{x-\delta}{\overline{\lambda}_r - 2\delta} & \text{if } \delta \leq x \leq \overline{\lambda}_r - \delta, \\ 0 & \text{if } x \leq \delta. \end{cases} \tag{43}$$

Let $\Phi : \mathbb{S}^n \to \mathbb{S}^n$ be the corresponding Löwner operator with respect to ϕ, i.e.,

$$\Phi(A) = P\mathrm{Diag}(\phi(\lambda_1(A)), \ldots, \phi(\lambda_n(A)))P^T, \quad A \in \mathbb{S}^n, \tag{44}$$

where $P \in \mathbb{O}^n$ comes from the eigenvalue decomposition

$$A = P\mathrm{Diag}(\lambda_1(A), \ldots, \lambda_n(A))P^T.$$

Immediately we have $\Phi(-J\overline{D}J) = \overline{P}_1\overline{P}_1^T$. We show it is also true for $\widetilde{D}$.

It follows the perturbation result of Weyl for eigenvalues of symmetric matrices [5, p. 63] that

$$\|\overline{\lambda}_i - \widetilde{\lambda}_i\| \leq \|J(\overline{D} - \widetilde{D})J\| \leq \|\overline{D} - \widetilde{D}\|, \qquad i = 1, \ldots, n.$$

We must have

$$\widetilde{\lambda}_i \geq \overline{\lambda}_r - \delta \quad \text{for } i = 1, \ldots, r \quad \text{and} \quad \widetilde{\lambda}_i \leq \delta \quad \text{for } i = r+1, \ldots, n.$$

We therefore have $\Phi(-J\widetilde{D}J) = \widetilde{P}_1\widetilde{P}_1^T$.

As a matter of fact, the scalar function defined by (43) is twice continuously differentiable (actually, ϕ is analytic) on $(-\infty, \delta) \cup (\overline{\lambda}_r - \delta, \infty)$. Therefore, we know from [5, Exercise V.3.9] that Φ is twice continuously differentiable near $-J\overline{D}J$ (actually, Φ is analytic near $-J\overline{D}J$). Therefore, under the condition that $\delta < \overline{\lambda}_r/2$, we have by the derivative formula of the Löwner operator (see e.g., [5, Thm. V.3.3]) that

$$\begin{aligned} \widetilde{P}_1\widetilde{P}_1^T - \overline{P}_1\overline{P}_1^T = \Phi(-J\widetilde{D}J) - \Phi(-J\overline{D}J) &= \Phi'(-J\overline{D}J)(-JHJ) + O(\|-JHJ\|^2) \\ &= \overline{P}\left[\overline{W} \circ (\overline{P}^T(-JHJ)\overline{P})\right]\overline{P}^T + O(\|H\|^2), \end{aligned}$$

where $H := \widetilde{D} - \overline{D}$ and $\overline{W} \in \mathbb{S}^n$ is given by

$$(\overline{W})_{ij} := \begin{cases} \dfrac{1}{\overline{\lambda}_i} & \text{if } 1 \leq i \leq r \text{ and } r+1 \leq j \leq n, \\ \dfrac{1}{\overline{\lambda}_j} & \text{if } r+1 \leq i \leq n \text{ and } 1 \leq j \leq r, \\ 0 & \text{otherwise,} \end{cases} \qquad i, j \in \{1, \ldots, n\}.$$

We note that the leading $r \times r$ block of $\overline{W}$ is 0, which implies $\langle \overline{P}_1\overline{P}_1^T, \overline{P}\left[\overline{W} \circ (\overline{P}^T(-JHJ)\overline{P})\right]\overline{P}^T \rangle = 0$. Therefore, we know from (42) that if $\widetilde{D}$ is sufficiently close to $\overline{D}$, $\rho_2^* = 1 + O(\|H\|^2)$.

This shows that $\rho_2 = 1$ is nearly optimal if the initial estimator $\widetilde{D}$ is close to $\overline{D}$. We will show that in terms of the estimation errors the choice $\rho_2 = 1$ is always better than the nuclear norm penalized least squares model ($\rho_2 = 0$) and the minimum volume embedding model ($\rho_2 = 2$).

Proposition 5 *If* $\|\widetilde{D} - \overline{D}\| < \overline{\lambda}_r/2$, *then* $\alpha(1) < \min\{\alpha(0), \alpha(2)\}$.

Proof. By Ky Fan's inequality [21], we know that $\langle \overline{P}_1\overline{P}_1^T, \widetilde{P}_1\widetilde{P}_1^T\rangle \le r$. From (41), we have

$$\alpha^2(2) = \frac{1}{2r}(5r - 4\langle \overline{P}_1\overline{P}_1^T, \widetilde{P}_1\widetilde{P}_1^T\rangle) \ge \frac{1}{2r}(5r-4r) = \frac{1}{2} = \alpha^2(0).$$

Therefore, we only need to show that $\alpha(1) = \frac{1}{\sqrt{2r}}\|\widetilde{P}_1\widetilde{P}_1^T - \overline{P}_1\overline{P}_1^T\| < \frac{1}{\sqrt{2}} = \alpha(0)$. The rest of the proof is similar to that of [39, Thm. 3]. Let $\mathscr{N}_\delta := \{D \in \mathbb{S}^n \mid \|D-\overline{D}\| \le \delta\}$, where $\delta = \|\widetilde{D}-\overline{D}\|$. For any $D \in \mathscr{N}_\delta$, we have

$$|\lambda_i(-JDJ) - \lambda_i(-J\overline{D}J)| = |\lambda_i(-JDJ) - \overline{\lambda}_i| \le \|-JDJ + J\overline{D}J\| \le \|D-\overline{D}\| \le \delta, \quad i = 1,\ldots,n.$$

Moreover, it follows from $\delta < \overline{\lambda}_r/2$ that for any $D \in \mathscr{N}_\delta$, $\lambda_r(-JDJ) \ge \overline{\lambda}_r - \delta > \overline{\lambda}_r/2 > \delta \ge \lambda_{r+1}(-JDJ)$. Therefore, for any $D \in \mathscr{N}_\delta$, we have $\Phi(-JDJ) = P_1P_1^T$, where $P = [P_1\ \ P_2] \in \mathbb{O}^n$ satisfies $-JDJ = P\mathrm{Diag}(\lambda(-JDJ))P^T$ with $P_1 \in \Re^{n\times r}$ and $P_2 \in \Re^{n\times(n-r)}$. Moreover, Φ defined by (44) is continuously differentiable over $\mathscr{N}_\delta$. Thus, we know from the mean value theorem that

$$\widetilde{P}_1\widetilde{P}_1^T - \overline{P}_1\overline{P}_1^T = \Phi(-J\widetilde{D}J) - \Phi(-J\overline{D}J) = \int_0^1 \Phi'(-JD_tJ)(-J\widetilde{D}J + J\overline{D}J)\mathrm{d}t, \tag{45}$$

where $D_t := \overline{D} + t(\widetilde{D}-\overline{D})$.

For any $D \in \mathscr{N}_\delta$, we know from the derivative formula of the Löwner operator that for any $H \in \mathbb{S}^n$, $\Phi'(-JDJ)H = P[\Omega \circ (P^THP)]P^T$, where $\Omega \in \mathbb{S}^n$ is given by

$$(\Omega)_{ij} := \begin{cases} \dfrac{1}{\lambda_i(-JDJ) - \lambda_j(-JDJ)} & \text{if } 1 \le i \le r \text{ and } r+1 \le j \le n, \\ \dfrac{-1}{\lambda_i(-JDJ) - \lambda_j(-JDJ)} & \text{if } r+1 \le i \le n \text{ and } 1 \le j \le r, \\ 0 & \text{otherwise,} \end{cases}$$

which implies that

$$\|\Phi'(-JDJ)H\| \le \frac{\|H\|}{\lambda_r(-JDJ) - \lambda_{r+1}(-JDJ)}.$$

This, together with (45) yields

$$\|\widetilde{P}_1\widetilde{P}_1^T - \overline{P}_1\overline{P}_1^T\| \le \int_0^1 \|\Phi'(-JD_tJ)(-J\widetilde{D}J + J\overline{D}J)\|\mathrm{d}t \le \int_0^1 \frac{\|\widetilde{D}-\overline{D}\|}{\lambda_r(-JD_tJ) - \lambda_{r+1}(-JD_tJ)}\mathrm{d}t.$$

By Ky Fan's inequality, we know that

$$(\lambda_r(-JD_tJ) - \overline{\lambda}_r)^2 + \lambda_{r+1}^2(-JD_tJ) \le \|\lambda(-JD_tJ) - \lambda(-J\overline{D}J)\|^2 \le \|-JD_tJ + J\overline{D}J\|^2 \le \|D_t - \overline{D}\|^2 = t^2\delta^2.$$

It can be checked directly that $\lambda_r(-JD_tJ) - \overline{\lambda}_r - \lambda_{r+1}(-JD_tJ) \ge -\sqrt{2}t\delta$, which implies that

$$\lambda_r(-JD_tJ) - \lambda_{r+1}(-JD_tJ) \ge \overline{\lambda}_r + \lambda_r(-JD_tJ) - \overline{\lambda}_r - \lambda_{r+1}(-JD_tJ) \ge \overline{\lambda}_r - \sqrt{2}t\delta.$$

Thus, $\|\widetilde{P}_1\widetilde{P}_1^T - \overline{P}_1\overline{P}_1^T\| \le \int_0^1 \frac{\delta}{\overline{\lambda}_r - \sqrt{2}t\delta}\mathrm{d}t = -\frac{1}{\sqrt{2}}\log\left(1 - \frac{\sqrt{2}\delta}{\overline{\lambda}_r}\right)$. Since $r \ge 1$, we know that

$$\delta/\overline{\lambda}_r < 1/2 < 0.5351 < \frac{1}{\sqrt{2}}\left(1 - \exp\left(-\sqrt{2r}\right)\right),$$

which implies that $\frac{1}{\sqrt{r}}\|\widetilde{P}_1\widetilde{P}_1^T - \overline{P}_1\overline{P}_1^T\| < 1$. Therefore, the proof is completed. □

5.2 A convergent 3-block ADMM algorithm

Without loss of generality, we consider the following convex quadratic problem

$$\begin{array}{l} \min\ \frac{1}{2}\|\mathscr{A}(X)-\mathbf{a}\|^2+\langle C,X\rangle \\ \text{s.t. } \mathscr{B}(X)=\mathbf{c},\ \ X\in\mathbb{K}^n_+,\ \ \|X\|_\infty\le b, \end{array} \tag{46}$$

where $\mathbb{K}^n_+$ is the almost positive semidefinite cone defined by (1), $X,C\in\mathbb{S}^n$, $\mathbf{a}\in\Re^m$, $\mathbf{c}\in\Re^k$, $b>0$, and $\mathscr{A}:\mathbb{S}^n\to\Re^m$, $\mathscr{B}:\mathbb{S}^n\to\Re^k$ are two given linear operators. By setting $\mathscr{A}\equiv\mathscr{O}$, $\mathscr{B}\equiv\mathrm{diag}(\cdot)$, $\mathbf{a}\equiv-(\mathbf{y}\circ\mathbf{y})\in\Re^m$, $\mathbf{c}\equiv 0\in\Re^n$ and $C\equiv m\rho_1 J(I-\rho_2\widetilde{P}_1\widetilde{P}_1^T)J$, one can easily verify that (46) is equivalent to the trusted distance learning model (13).

The problem (46) can be solved by an efficient 3-block ADMM method [3], which is inspired by the recent work of Li et al. [37] for general convex quadratic programming. By introducing a new variable $\mathbf{t}=\mathscr{A}(X)-\mathbf{a}$ and a slack variable $W\in\mathbb{S}^n$, we can rewrite (46) as the following equivalent form:

$$\begin{array}{l} \min\ \frac{1}{2}\|\mathbf{t}\|^2+\langle C,X\rangle+\delta_{\mathbb{K}^n_+}(X)+\delta_{\mathbb{B}^\infty_b}(W) \\ \text{s.t. } \mathscr{A}(X)-\mathbf{t}=\mathbf{a},\ \ \mathscr{B}(X)=\mathbf{c},\ \ X=W, \end{array} \tag{47}$$

where $\mathbb{B}^\infty_b:=\{X\in\mathbb{S}^n\mid\|X\|_\infty\le b\}$ and for any given set $\mathcal{F}$, $\delta_{\mathcal{F}}$ is the indicator function over $\mathcal{F}$. Moreover, the corresponding Lagrangian dual problem is given by

$$\begin{array}{l} \max\ \frac{1}{2}\|\mathbf{y}_1\|^2+\langle\mathbf{a},\mathbf{y}_1\rangle+\langle\mathbf{c},\mathbf{y}_2\rangle+\delta_{(\mathbb{K}^n_+)^*}(S)-\delta^*_{\mathbb{B}^\infty_b}(-Z) \\ \text{s.t. } Z+\mathscr{A}^*\mathbf{y}_1+\mathscr{B}^*\mathbf{y}_2-S=C, \end{array} \tag{48}$$

where $((\mathbf{y}_1,\mathbf{y}_2),Z,S)\in\Re^{m+k}\times\mathbb{S}^n\times\mathbb{S}^n$ are dual variables grouped in 3-block format, $(\mathbb{K}^n_+)^*$ is the dual cone of $\mathbb{K}^n_+$ and $\delta^*_{\mathbb{B}^\infty_b}$ is the support function of $\mathbb{B}^\infty_b$. The details of the convergent 3-block ADMM algorithm can be found from [3, Section IV (C)]. We omit the details here for simplicity.

6 Numerical Experiments

In this section, we demonstrate the effectiveness of the proposed EDM Embedding (EDME) model (13) by testing on some real world examples. The examples are in two categories: one is of the social network visualization problem, whose initial link observation can be modelled by uniform random graphs. The other is from manifold learning, whose initial distances are obtained by the k-NN rule. The known physical features of those problems enable us to evaluate how good EDME is when compared to other models such as ISOMAP and MVU. It appears that EDME is capable of generating configurations of very high quality both in terms of extracting those physical features and of higher EDM scores. The test also raises an open question whether our theoretical results can be extended to this case where the k-NN rule is used.

For comparison purpose, we also report the performance of MVU and ISOMAP for most cases. The SDP solver used is the state-of-art SDPT3 package, which allows us to test problems of large data sets. We did not compare with MVE as it solves a sequence of SDPs and consequently it is too slow for our tested problems. Details on this and other implementation issues can be found in Subsection 6.3.

6.1 Social Networks

Two real-world networks arising from the different applications are used to demonstrate the quality of our new estimator from EDME.

(SN1) US airport network. In this example, we try to visualize the social network of the US airport network from the data of 2010 [43]. There are $n=1572$ airports under consideration. The number of the passengers transported from the i-th airport to the j-th airport in 2010 is recorded and denoted by C_{ij}. Therefore, the

social distance between two cities can be measured by the passenger numbers. The social distances (or dissimilarities) between users are computed from the communication counts. It is natural to assume that larger communication count implies smaller social distance. Without loss of generality, we employ the widely used Jaccard dissimilarity [30] to measure the social distance of users:

$$D_{ij} = \sqrt{1 - \frac{C_{ij}}{\sum_k C_{ik} + \sum_k C_{jk} - C_{ij}}} \quad \text{if } C_{ij} \neq 0. \tag{49}$$

The observed distance matrix is also incomplete, and only very few entrances are observed ($< 1.4\%$). The two dimensional embeddings obtained by the MVU and EDME methods are shown in Figure 1. The ten busiest US airports by total passenger traffic in 2010 are indicated by the red circles. Note that there are a large number of passengers transporting between them, which means the corresponding social distances among them should be relatively small. Thus, it is reasonable to expect that the embedding points of these top ten airports cluster around the zero point. Both MVU and EDME methods are able to show this important feature. The details on the numerical performance of MVU and EDME on this example are reported in Table 1.

A close look reveals more interesting location clusters among the 10 cities (see the inserted graphs of the enlarged locations in both MVU and EDME embeddings). From the EDME embedding, we can observe that these ten airports are naturally separated into four groups: Group 1 $= \{1(\texttt{ATL}), 2(\texttt{ORD}), 7(\texttt{IAH})\}$; Group 2 $= \{6(\texttt{JFK})\}$; Group 3 $= \{8(\texttt{LAS}), 10(\texttt{PHX}\})$; and Group 4 $= \{3(\texttt{LAX}), 4(\texttt{DFW}), 5(\texttt{DEN}), 9(\texttt{SFO})\}$. This has an interesting geographical meaning. For example, Group 1 corresponds to three southeast US cities: `Atlanta`, `Orlando` and `Houston`; Group 2 corresponds to one big east-coast city: `New York`; Group 3 has two closed related southwest cities: `Las Vegas` and `Phoenix`; Group 4 are four west cities: `Los Angeles`, `Dallas`, `Denver` and `San Francisco`. Also, Group 1 & 2 are east cities and Group 3 & 4 are west ones. However, by MVU, we can only obtain two groups: one consists of the east cities: $\{\texttt{ATL}, \texttt{ORD}, \texttt{IAH}, \texttt{JFK}\}$, and another consists of the west ones: $\{\texttt{DFW}, \texttt{LAS}, \texttt{PHX}, \texttt{LAX}, \texttt{DEN}, \texttt{SFO}\}$. Furthermore, it can be seen from the eigenvalue spectrum in Figure 1 that the MVU only captured 74.3% variance in the top two leading eigenvectors, while the EDME method captured all the variance in the two dimensional space. We also apply the MVC package [16] to this example. The corresponding parameters are set as follows `MVCiter=5`, `perpatch=200`, `init='glMVU'`, `outdim=2`. MVC only needs 373.66 seconds to produce an approximate solution, which significantly reduces the computational time of the original MVU. However, it can be observed from Figure 2a that it failed to capture the important geographical feature mentioned above.

(SN2) Political blogs [1] collected the data including links, citations and posts on the 1940 political blogs around the 2004 US presidential election period. These blogs are classified as two parts: 758 left-leaning blogs and 732 right-leaning blogs. In this paper, we will use the data on the links between the blogs, which can be found from [24] to visualize the corresponding social network. Similar to the communication network, we use (49) to measure the social distance of blogs. Without loss of generality, the 718 isolated blogs are removed from the original data, which means that we consider the remaining $n = 1222$ blogs with 586 left-leanings and 636 right-leanings. The social networks obtained by the MVU and the EDME are presented in Figure 3. From the results, we can see clearly that the embedding points generated by the MVU are concentrated near the zero point, and the rank of the corresponding Gram matrix is much higher than 2, which is 1135. However, our EDME method is able to capture all variance of the data in the two dimensions, providing a more accurate lower dimensional embedding. In fact, the embedding points in the visualizing network obtained by the EDME are naturally separated into two groups: the left-leaning blogs (the blue circles) and the right-leaning ones (the red circles). MVC package is also tested for this example. All parameters are chosen as the same for the previous example. Again, we can see that the computational cost is significantly reduced by MVC, which only needs 28.40 seconds even faster than EDME. However, it can be seen from Figure 2b that all left-leaning and right-leaning blogs are mixed. From now on we will not test MVC package anymore.

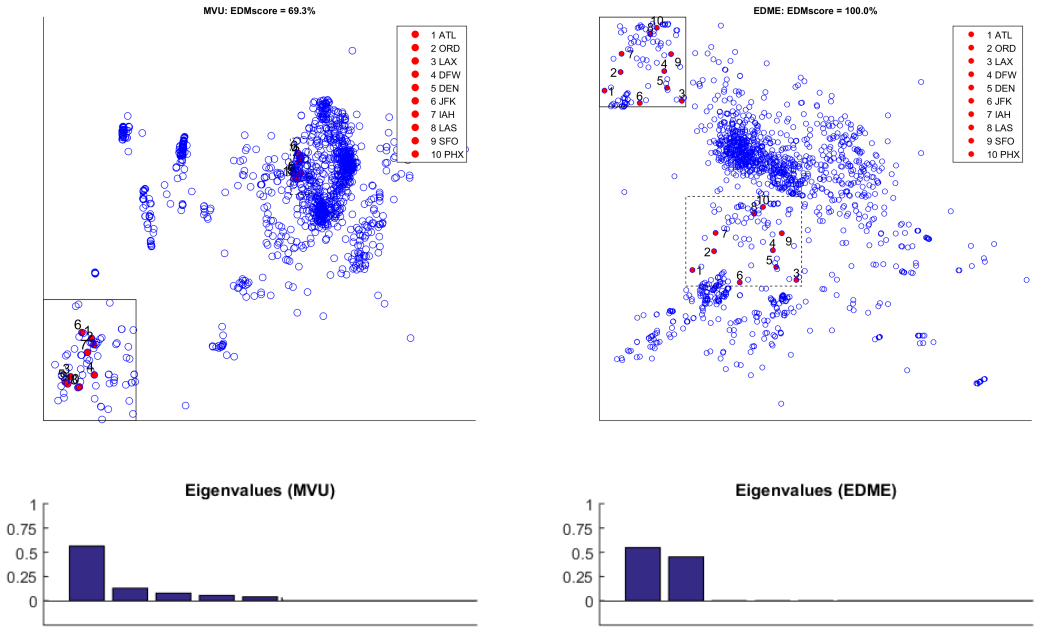
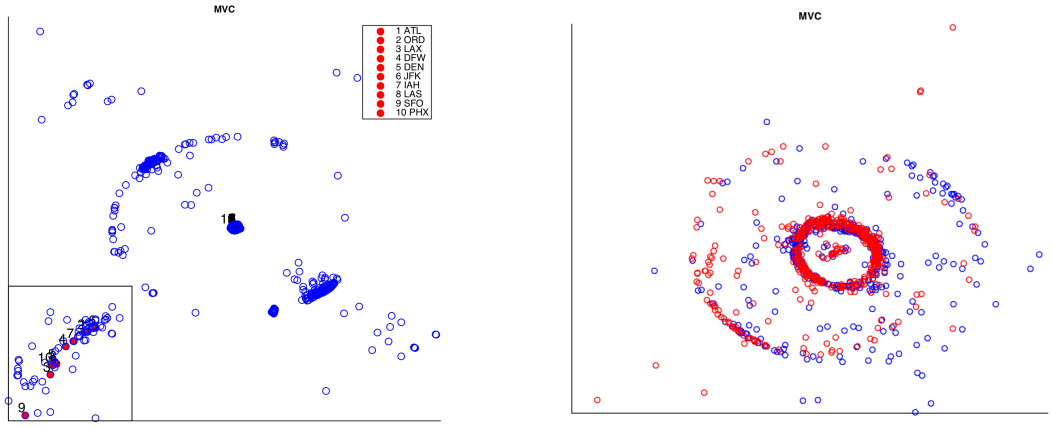


Fig. 1: The embedding networks of USairport2010



(a) USairport2010 networks by MVC

(b) Political blogs networks by MVC

Fig. 2: MVC embedding for USairport2010 and Political blogs. Both embeddings failed to capture the important features shown by MVU and EDME.

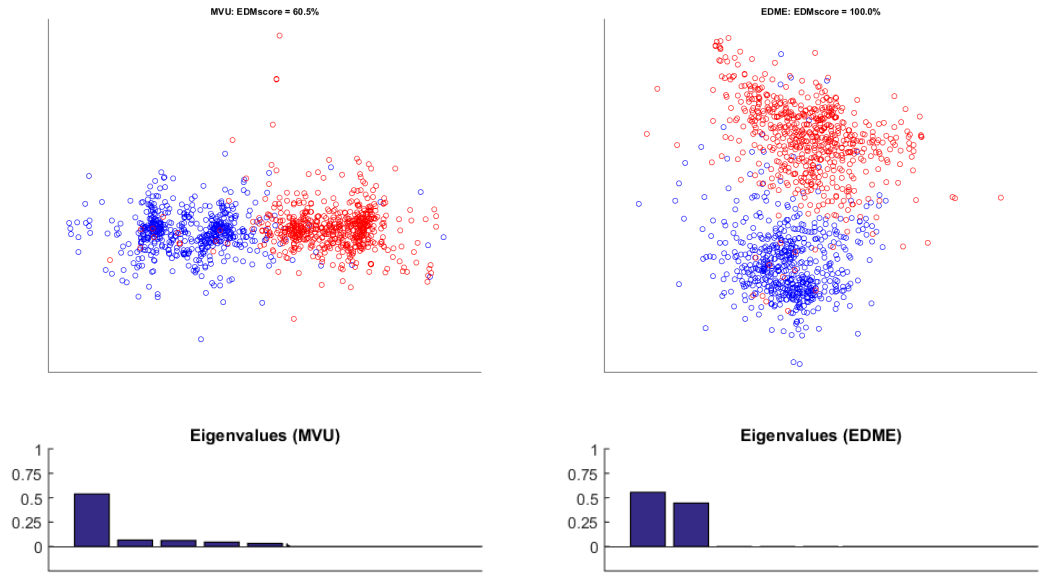


Fig. 3: The political blogs: the blue color represents the left-leaning blogs and the red for the right-leaning blogs.

6.2 Manifold learning

In this subsection, we test two widely used data sets in manifold learning. The initial distances used are generated by the k-NN rule. We describe them below with our findings for MVU, EDME and the celebrated manifold learning algorithm ISOMAP.

(ML1) Data of Face698. In this example, we try to represent the high dimensional face image data [54] in a low dimension space. There are $n = 698$ images (64 pixel by 64 pixel) of faces with the different poses (up-down and left-right) and different light directions. Therefore, it is natural to expect that these high dimensional input data lie in the three dimensional space parametrized by the face poses and the light directions and that the equal importance of the three features can be sufficiently captured. Similar to the previous example, we use $k = 5$ to generate a connected graph. Both MVU and EDME methods successfully represent the data in the desired three dimensional space and their embedding results of the MVU and EDME are similar. For simplicity only the result of the EDME is shown in Figure 4. However, the Gram matrix learned by the ISOMAP has more than three nonzero eigenvalues. This is shown in the corresponding eigenvalue spectrum in Figure 4. Furthermore, for the ISOMAP, if we only compute the two-dimension embedding, then we only capture a smaller percentage of the total variance. It is interesting to observe that EDME is the only model that treats the three features equally important (the three leading eigenvalues are roughly equal). Moreover, the EDME model performs much better than MVU in terms of the numerical efficiency. See Table 1 for more details.

(ML2) The digits Data The data is from the MNIST database [35]. We first consider the data set of digit "1", which includes $n = 1135$ 8-bit grayscale images of "1". Each image has 28×28 pixels, which is represented as 784 dimensional vector. We note that the two most important features of "1"s are the slant and the line thickness. Therefore, the embedding results are naturally expected to lie in the two dimensional space parametrized by these two major features. In this example, we set $k = 6$. Figure 5 shows the two dimensional embeddings computed by ISOMAP, MVU and EDME. It can be clearly seen that EDME significantly outperforms the other two methods. In particular, EDME is able to accurately represent the data in the two dimensional space and

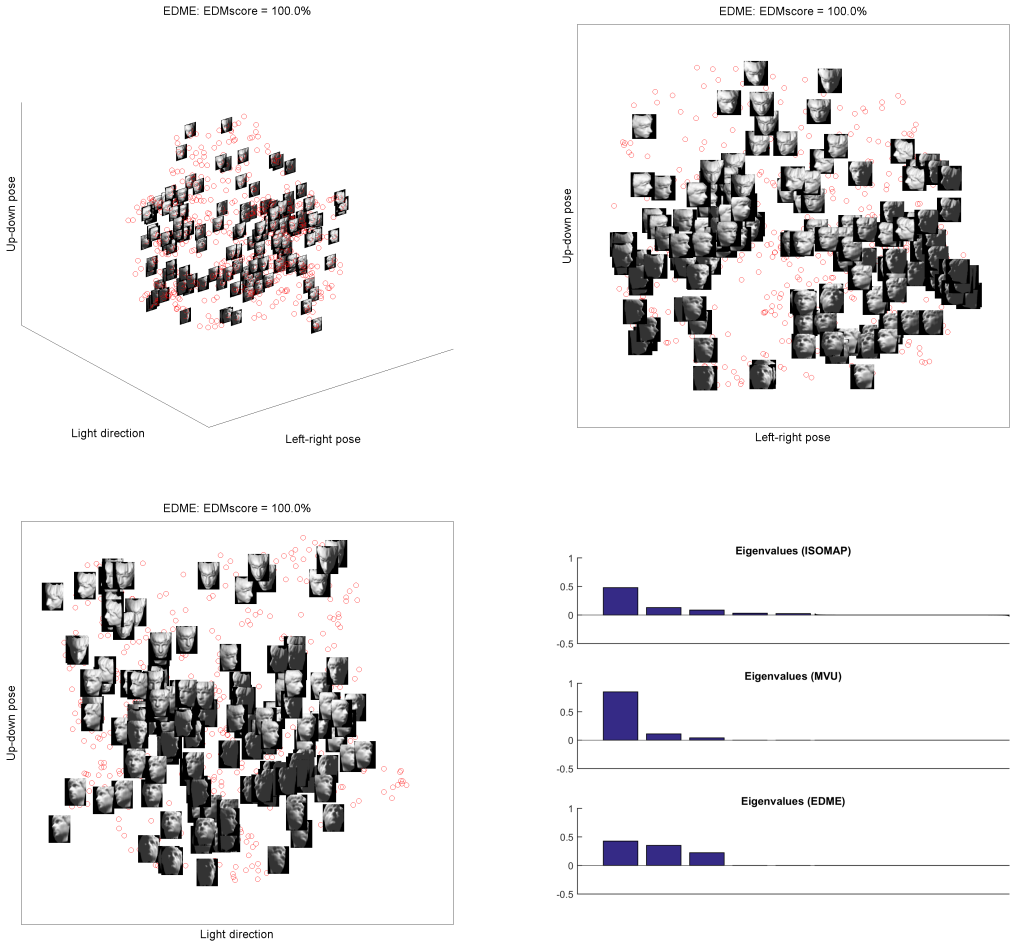


Fig. 4: Face698

captures the correct features. However, MVU returns an almost one dimensional embedding and only captures one of the major features, i.e., the slant of "1"s. For the ISOMAP, it only captures a small percentage of the total variance. Moreover, our method also outperforms the nuclear norm penalized least squares (NNPLS) model (see Figure 6). As mentioned before, the nuclear norm penalty approach has one key drawback, i.e., the "crowding phenomenon" of the embedding points (the total variance among the given data is reduced). Therefore, the resulting embeddings fail to capture two important features of "1"s.

6.3 Numerical performance

We tested the ISOMAP, the MVU and our proposed EDME methods in MATLAB 8.5.0.197613 (R2015a), and the numerical experiments are run in MATLAB under a Windows 10 64-bit system on an Intel 4 Cores i7 3.60GHz CPU with 8GB memory.

Besides the examples mentioned before, the following examples are also tested: the Enron email dataset [17], the facebook-like social network [44], the Madrid train bombing data [9] (downloaded from [24]), the

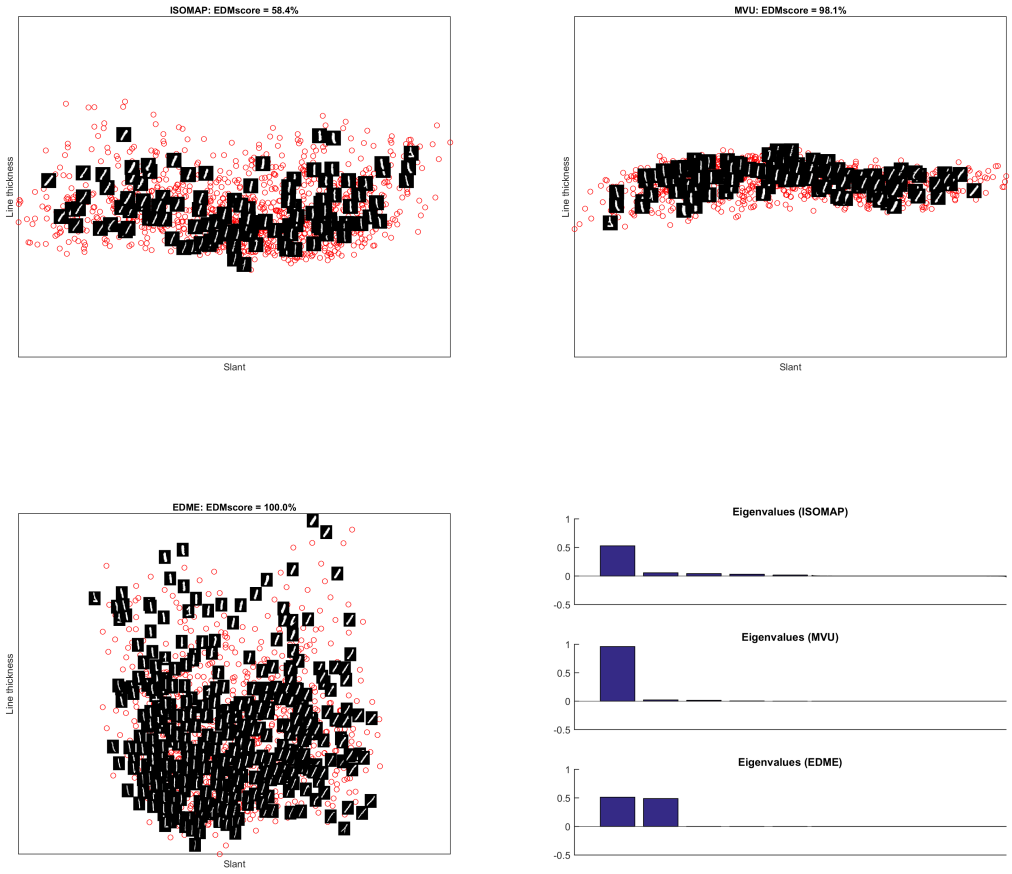


Fig. 5: Digit 1

teapots data [61], the digits "1" and "9" and the Frey face images data [49]. To save space, we do not include the actual embedding graphs for these examples, but just report the numerical performance in Table 1.

In our numerical experiments, we use the SDPT3 [55], a Matlab software package for semidefinite-quadratic-linear programming, to solve the corresponding SDP problem of the original MVU model. The termination tolerance of the SDPT3 is tol $= 10^{-3}$. For our EDME model, we terminate the ADMM algorithm if the following condition obtained from the general optimality conditions (KKT conditions) of (47) and (48) is met, i.e.,

$$R := \max\{R_p, R_d, R_Z, R_{C_1}, R_{C_2}\} \leq \text{tol},$$

where $R_p = \|(\mathscr{A}(X) - \mathbf{t} - \mathbf{a}, \mathscr{B}(X) - \mathbf{c})\|/(1 + \|(\mathbf{a}; \mathbf{c})\|)$, $R_d = (Z + \mathscr{A}^*\mathbf{y}_1 + \mathscr{B}^*\mathbf{y}_2 - S - C)/(1 + \|C\|)$, $R_Z = \|X + \Pi_{\mathbb{B}_b^\infty}(X + Z)\|/(1 + \|X\| + \|Z\|)$, $R_{C_1} = |\langle S, X\rangle|/(1 + \|S\| + \|X\|)$ and $R_{C_2} = \|X - \Pi_{\mathbb{K}_+^n}(X)\|/(1 + \|X\|)$. Clearly, R_p measures the violation of primal feasibility; R_D measures the violation of the equation constraint in the dual problem (48); R_Z measures the violation of X belonging to $\mathbb{B}_b^\infty$; R_{C_1} measures the complementarity condition between S and X; and R_{C_2} measures the violation of X belonging to $\mathbb{K}_+^n$. The tolerance is also set at tol $= 10^{-3}$. The details on the numerical performance of the MVU and EDME methods can be found from Table 1, where we report the EDM scores from the leading two eigenvalues and cpu time in seconds.

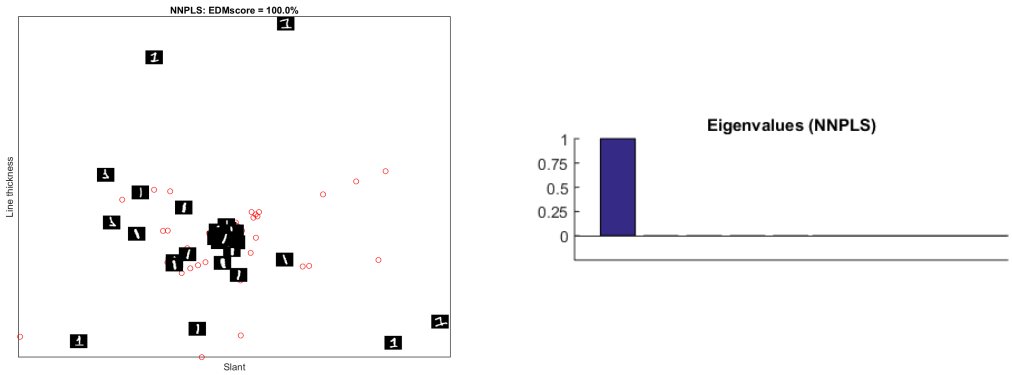


Fig. 6: NNPLS for Digit 1.

Problems		MVU (SDPT3)			EDME		
	n/edges	relgap	EDMscore	cpu(s)	*R*	EDMscore	cpu(s)
Enron	182/2097	3.25e-04	48.1%	5.46	9.92e-04	100%	1.05
Facebook-like	1893/13835	4.40e-04	20.6%	624.18	9.78e-04	100%	165.34
TrainBombing	64/243	9.38e-04	91.8%	0.49	9.99e-04	100%	0.56
USairport2010	1572/17214	7.30e-04	69.3%	31587.08	9.98e-04	100%	80.53
Blogs	1222/16714	4.71e-04	60.5%	12173.31	8.40e-04	100%	83.66
	k/*n*/edges	relgap	EDMscore	cpu(s)	*R*	EDMscore	cpu(s)
Teapots400	5/400/1050	8.45e-04	100%	3.44	9.75e-04	100%	3.77
Face698	5/698/2164	2.96e-04	100%	14.25	9.94e-04	100%	29.73
Digit1	6/1135/4885	7.49e-04	98.1%	68.85	9.95e-04	100%	39.62
Digits19	6/1000/4394	6.57e-04	94.0%	51.02	9.83e-04	100%	27.66
FreyFace	5/1965/6925	9.48e-04	86.2%	214.41	8.72e-04	100%	187.56

Table 1: Numerical performance comparison of the MVU and the EDME

We observe that the performance of EDME is outstanding in terms of numerical efficiency. Taking USairport2010 as example, MVU used about 10 hours while EDME only used about 80 seconds. For the examples in manifold learning, the gap between the two models are not as severe as for the social network examples. The main reason is that the initial guess obtained by ISOMAP is a very good estimator that can roughly capture the low-dimensional features in manifold learning. However, it fails to capture meaningful features for the social network examples. This echoes the comment made in [10] that the shortest path distance is not suitable to measure the distances in social networks. We also like to point out that for all tested problems, EDME captured nearly 100% variance and it treats the local features equally important in terms of the leading eigenvalues being of the same magnitude.

7 Conclusions

The paper aimed to explain a mysterious situation regarding the SDP methodology to reconstruct faithful Euclidean distances in a low-dimensional space from incomplete set of noisy distances. The SDP models can construct numerical configurations of high quality, but they lack theoretical backups in terms of bounding errors. We took a completely different approach that heavily makes use of Euclidean Distance Matrix instead of positive semidefinite matrix in SDP models. This led to a convex optimization that inherits the nice features of MVU and MVE models. More importantly, we were able to derive error-bound results under the uniform sampling rule. The optimization problem can also be efficiently solved by the proposed algorithm. Numerical

results in both social networks and manifold leading showed that our model can capture low-dimensional features and treats them equally important.

Given that our model worked very well for the manifold learning examples, an interesting question regarding this approach is whether the theoretical error-bound results can be extended to the case where the distances are obtained by the k-NN rule. It seems very difficult if we follow the technical proofs in this paper. It also seems that the approach of [28] would lead to some interesting (but very technical) results. We plan to investigate those issues in future.

8 Acknowledgements

We would like to thank the referees as well as the associate editor for their constructive comments that have helped to improve the quality of the paper.

References

1. Adamic, A.A., Glance, N.: The political blogosphere and the 2004 US election: divided they blog. In Proceedings of the 3rd international workshop on Link Discovery (2005).
2. Arias-Castro, E., Pelletier, B.: On the convergence of maximum variance unfolding. J. Machine Learn. Res. **14**, 1747–1770 (2013).
3. Bai, S.H., Qi, H.-D.: Tackling the flip ambiguity in wireless sensor network localization and beyond. Preprint available from: `http://www.personal.soton.ac.uk/hdqi/REPORTS/EDMSNL.pdf` (2015).
4. Bernstein, M., De Silva, V., Langford, J.C., Tenenbaum, J.B.: Graph approximations to geodesics on embedded manifolds. available from: `http://isomap.stanford.edu/BdSLT.pdf`, Stanford University (2000).
5. Bhatia, R.: Matrix Analysis, Springer-Verlag, New York, 1997.
6. Biswas, P., Liang, T.-C., Toh, K.-C., Ye, Y., Wang, T.C.: Semidefinite programming approaches for sensor network localization with noisy distance measurements. IEEE Trans. Auto. Sci. Eng. **3**, 360–371 (2006).
7. Bollobás, B.: Random Graphs. Cambridge University Press, 2001.
8. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling, Springer, 2005.
9. Brian, V.: Connecting the dots. Ameri. Scientist **95**, 400–404 (2006).
10. Budka, M., Juszczyszyn, K., Musial, K., Musial, A.: Molecular model of dynamic social network based on e-mail communication. Social Network Anal. Mining **3**, 543–563 (2013).
11. Bühlmann, P. and Van De Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, 2011.
12. Burges, C.J.C.: Dimension Reduction: A Guided Tour. Foundation Trend in Machine Learning **2**, 275–365 (2009).
13. Candès, E.J., Plan, Y.: Matrix Completion With Noise, Proceedings of the IEEE **98**, 925–936 (2010).
14. Candès, E.J., Recht, B.: Exact matrix completion via convex optimization. Found. Comp. Math. **9**, 717–772 (2008).
15. Candès, E.J., Tao, T.: The power of convex relaxation: Near-optimal matrix completion. IEEE Trans. Infor. Theory **56**, 2053–2080 (2010).
16. Chen, W., Chen, Y., Weinberger, K.Q.: Maximum variance correction with application to A* search. In Proc. the 30th Inter. Conf. Machine Learn. (ICML-13), 302–310 (2013).
17. Cohen, W.W., William, W.: Enron email dataset, 2009.
18. Cox, T.F., Cox, M.A.A.: Multidimensional Scaling, 2nd Ed, Chapman and Hall/CRC, 2001.
19. de Sola Pool, I., Kochen, M.: Contacts and influence. Social Networks **1**, 5–51 (1979).
20. Erdős, P., Rényi, A.: On random graphs. Publicationes Mathematicae Debrecen **6**, 290–297 (1959).
21. Fan, K.: On a theorem of Weyl concerning eigenvalues of linear transformations I. Proc. Nat. Aca. Sci. **35**, 652–655 (1949).
22. Fazel, M.: Matrix Rank Minimization with Applications. PhD Thesis, Stanford University, 2002.
23. Freeman, L.C.: Graphic techniques for exploring social network data. Models and Methods in Social Network Analysis, 248–269 (2005).
24. Freeman, L.C.: Freeman Datasets, available from: `http://moreno.ss.uci.edu/data.html`, 2010.
25. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika **53**, 325–338 (1966).
26. Gross, D.: Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Info. Theory **57**, 1548–1566 (2011).
27. Janson, S., Luczak, T., Rucinski, A.: Random Graphs, John Wiley & Sons, 2011.
28. Javanmard, A., Montanari, A.: Localization from incomplete noisy distance measurements. Found. Comp. Math. **13**, 297–345 (2013).
29. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from noisy entries. J. Machine. Learn. Res. **11**, 2057–2078 (2010).
30. Klavans, R., Boyack, K.W.: Identifying a better measure of relatedness for mapping science. J. the Amer. Soc. for Inf. Sci. and Tech. **57**, 251–263 (2006).
31. Klopp, O.: Rank penalized estimators for high-dimensional matrices. Elec. J. of Stat. **5**, 1161–1183 (2011).

32. Klopp, O.: Noisy low-rank matrix completion with general sampling distribution. Bernoulli **20**, 282–303 (2014).
33. Koltchinskii, V.: Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. In: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008, Vol. 2033, Springer, 2011.
34. Koltchinskii, V., Lounici, K., Tsybakov, A.B.: Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Annals of Statistics **39**, 2302–2329 (2011).
35. LeCun, Y., Cortes, C. and Burges, C.J.C.: MNIST, available from: `http://yann.lecun.com/exdb/mnist/`, 1998.
36. Ledoux, M. and Talagrand, M.: Probability in Banach Spaces: Isoperimetry and Processes, Springer, 1991.
37. Li, X., Sun, D.F. and Toh, K.-C.: A schur complement based semiproximal ADMM for convex quadratic conic programming and extensions. Math. Prog. **155**, 333–373 (2016).
38. Mesbahi, M.: *On the rank minimization problem and its control applications*, Systems & control letters 33 (1998) 31–36.
39. Miao, W., Pan, S., Sun, D.F.: *A rank-corrected procedure for matrix completion with fixed basis coefficients*. Math. Prog. (2016) DOI: 10.1007/s10107-015-0961-7.
40. Milgram, S.: The small world problem. Psychology today **2**, 60–67 (1967).
41. Negahban, S., Wainwright, M.J.: Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. J. Machine Learn. Res. **13**, 1665–1697 (2012).
42. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. **45**, 167–256 (2003).
43. Opsahl, T.: US Airport 2010. available from: `http://toreopsahl.com/datasets/\#usairports`, 2011.
44. Opsahl, T., Panzarasa, P.: Clustering in weighted networks. Social Networks **31**, 155–163 (2009).
45. Paprotny, A., Garcke, J.: On a connection between maximum variance unfolding, shortest path problems and isomap. Int. Conf. on Arti. Intel. and Stat. 859–867 (2012).
46. Pękalska, E., Paclík, P., Duin, P.W.: A generalized kernel approach to dissimilarity-based classification, J. Machine Learn. Res. **2**, 175–211 (2002).
47. Recht, B.: A simpler approach to matrix completion. J. Machine Learn. Res. **12**, 3413–3430 (2011).
48. Recht, B., Fazel, M., Parrilo, P.A.: Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**, 471–501 (2010).
49. Roweis, S.T., Saul, L.K.: Frey Face, available from: `http://www.cs.nyu.edu/~roweis/data.html`, 2000.
50. Schoenberg, I.J.: Remarks to Maurice Fréchet's article "Sur la définition axiomatque d'une classe d'espaces vectoriels distanciés applicbles vectoriellement sur l'espace de Hilbet". Ann. Math. **36**, 724–732 (1935).
51. Shaw, B., Jebara, T.: Minimum volume embedding. Int. Conf. on Arti. Intel. and Stat., 460–467 (2007).
52. Solomonoff, R., Rapoport, A.: Connectivity of random nets. Bull. of Math. Biophy. **13**, 107–117 (1951).
53. Sun, J., Boyd, S., Xiao, L., Diaconis, P.: The fastest mixing Markov process on a graph and a connection to a maximum variance unfolding problem. SIAM Rev. **48**, 681–699 (2006).
54. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000).
55. Toh, K.C., Todd, M.J., Tütüncü, R.H.: SDPT3 – a MATLAB software package for semidefinite programming, version 1.3. Optim. Methods and Software **11**, 545–581 (1999).
56. Tropp, J.A.: User-friendly tail bounds for sums of random matrices. Found. of Comp. Math. **12**, 389–434 (2012).
57. Tseng, P.: Second-order cone programming relaxation of sensor network localization. SIAM J. Optim. **18**, 156–185 (2007).
58. Vershynin, R.: Introduction to the non-asymptotic analysis of random matrices. In: Compressed sensing: theory and applications, (eds) by Eldar, Y.C., Kutyniok, G.: Cambridge University Press, 2012.
59. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, Cambridge University Press, 1994.
60. Watson, G.A.: Characterization of the subdifferential of some matrix norms. Linear Alg. Appl. **170**, 33–45 (1992).
61. Weinberger, K.Q., Saul, L.K.: Unsupervised learning of image manifolds by semidefinite programming. Int. J. of Computer Vision **70**, 77–90 (2006).
62. Weinberger, K.Q., Sha, F., Zhu, Q., Saul, L.K.: Graph Laplacian regularization for large-scale semidefinite programming. Advances in Neural Information Processing Systems **19**, 1489—1496 (2007).
63. Young, G., Householder, A.S.: Discussion of a set of points in terms of their mutual distances. Psychometrika **3**, 19–22 (1938).