Distance Shrinkage and Euclidean Embedding via Regularized Kernel Estimation

Luwan Zhang*, Grace Wahba† and Ming Yuan‡§

*‡Morgridge Institute for Research and *†‡Department of Statistics
University of Wisconsin-Madison

(August 4, 2015)

*Research supported in part by NSF Career Award DMS-1321692 and FRG Award DMS-1265202.

†Research supported in part by NIH Grants EY09946, 1U54AI117924-01 and NSF Grant DMS1308877.

‡Research supported in part by NSF Career Award DMS-1321692 and FRG Award DMS-1265202, and NIH Grant 1U54AI117924-01.

§Address for correspondence: Department of Statistics, University of Wisconsin-Madison, 1300 University Avenue, Madison, WI 53706.

Abstract

Although recovering an Euclidean distance matrix from noisy observations is a common problem in practice, how well this could be done remains largely unknown. To fill in this void, we study a simple distance matrix estimate based upon the so-called regularized kernel estimate. We show that such an estimate can be characterized as simply applying a constant amount of shrinkage to all observed pairwise distances. This fact allows us to establish risk bounds for the estimate implying that the true distances can be estimated consistently in an average sense as the number of objects increases. In addition, such a characterization suggests an efficient algorithm to compute the distance matrix estimator, as an alternative to the usual second order cone programming known not to scale well for large problems. Numerical experiments and an application in visualizing the diversity of Vpu protein sequences from a recent HIV-1 study further demonstrate the practical merits of the proposed method.

Key words: Embedding, Euclidean distance matrix, kernel, multidimensional scaling, regularization, shrinkage, trace norm.

1 Introduction

The problem of recovering an Euclidean distance matrix from noisy or imperfect observations of pairwise (dis)similarity scores between a set of objects arises naturally in many different contexts. It allows us to map objects from an arbitrary domain to Euclidean spaces, and therefore makes them amenable for subsequent statistical analyses, and also provides tools for visualization. Consider, for example, evaluating (dis)similarity between molecular sequences. A standard approach is through sequence alignment and measuring the (dis)similarity between a pair of sequences using their corresponding alignment score (see, Durbin et al., 1998). Although encoding invaluable insights into the relationship between sequences, it is well known that these scores do not correspond directly to a distance metric in the respective sequence space and therefore cannot be employed in kernel based learning methods. Similarly, there are also numerous other instances where it is possible to derive similarity or dissimilarity scores for pairs of objects from expert knowledge or other information, which, if successfully converted into positive semi-definite kernels or Euclidean distances, could allow themselves to play an important role in a myriads of statistical and computational analyses (e.g., Schölkopf and Smola, 2002; Székely, Rizzo and Bakirov, 2007).

A canonical example where this type of problem occurs is multidimensional scaling which aims to place each object in a low dimensional Euclidean space such that the between-object distances are preserved as well as possible. As such it also forms the basis for several other more recent approaches to nonlinear dimension reduction and manifold learning. See, Schölkopf (1998), Tenanbaum, De Silva and Langford (2000), Lu et al. (2005), Venna and Kaski (2006), Weinberger et al. (2007), Chen and Buja (2009, 2013) among others. Despite the popularity of multidimensional scaling, very little is known about to what extent the distances among the embedded points could faithfully reflect the true pairwise distances when observed with noises; and it is largely used only as an exploratory tool for initial data analysis.

Another example where it is of interest to reconstruct an Euclidean distance matrix is the determination of molecular structures using nuclear magnetic resonance (NMR, for short) spectroscopy, a technique pioneered by Nobel laureate Kurt Wüthrich (see, e.g., Wüthrich, 1986). As demonstrated by Wüthrich, distances between atoms could be inferred from chemical shifts measured by NMR spectroscopy. These distances obviously need to conform to a three dimensional Euclidean space yet experimental data on distances are inevitably noisy and as a result, the observed distances may not translate directly into locations of these atoms in a stable structure. Therefore, this becomes a problem of recovering an Euclidean distance matrix in 3D from noisy observations of pairwise distances. Similar problems also occur in graph realization and Euclidean representation of graphs where the goal is to embed the vertex set of a graph in an Euclidean space in such a fashion that the distance between two embedded vertexes matches their corresponding edge weight (see, e.g., Pouzet, 1979). While an exact embedding of a graph is typically of very high dimension, it is useful in some applications to instead seek approximate yet low dimensional embeddings (see, e.g., Roy, 2010).

More specifically, let $\{O_i : i = 1, 2, \ldots, n\}$ be a collection of objects from domain $\mathcal{O}$ which could be the coordinates of atoms in the case of molecular structure determination using NMR spectroscopy, or the vertex set of a graph in the case of graph realization. Let Ω be a subset of $\{(i, j) : 1 \leq i, j \leq n\}$, and $\{x_{ij} : (i, j) \in \Omega\}$ be the observed dissimilarity scores between them such that

$$x_{ij} = d_{ij} + \varepsilon_{ij}, \qquad (i, j) \in \Omega,$$

where ε_{ij}s are the measurement errors and $D = (d_{ij})_{1 \leq i,j \leq n}$ is a so-called Euclidean distance

matrix in that there exist points $p_1, \ldots, p_n \in \mathbb{R}^k$ for some $k \in \mathbb{N}$ such that

$$d_{ij} = \|p_i - p_j\|^2, \qquad 1 \le i < j \le n; \tag{1}$$

see, e.g., Darrotto (2013). Here $\|\cdot\|$ stands for the usual Euclidean distance. Our goal is to estimate the Euclidean distance matrix D from $(x_{ij})_{(i,j)\in\Omega}$. In many applications, all pairwise dissimilarity scores are observable, that is $\Omega = \{(i,j) : 1 \le i < j \le n\}$. In these cases, we can more conveniently write the observed scores in a matrix form $X = (x_{ij})_{1\le i,j\le n}$ where we adopt the convention that $x_{ji} = x_{ij}$ and $x_{ii} = 0$. To fix ideas, in the rest of the paper, we focus primarily on this setting of complete observations, with the exception of Section 4 where we discuss specifically how the methodology could handle the more general situations in a seamless fashion.

In the light of (1), D can be identified with the points p_is, which suggests an embedding of O_is in $\mathbb{R}^k$. Obviously, if O_is can be embedded in the Euclidean space of a particular dimension, then it is also possible to embed them in a higher dimensional Euclidean space. We refer to the smallest k in which such an embedding is possible as the embedding dimension of D, denoted by $\dim(D)$. As is clear from the aforementioned examples, oftentimes, either the true Euclidean distance matrix D itself is of low embedding dimension; or we are interested in an approximation of D that allows for a low dimensional embedding. Such is the case, for example, for molecular structure determination where the the embedding dimension of the true distance matrix D is necessarily three. Similarly, for multidimensional scaling or graph realization, we typically are interested in mapping objects in two or three dimensions.

Recall that

$$d_{ij} = p_i^\top p_i + p_j^\top p_j - 2p_i^\top p_j,$$

which relates D to the so-called kernel (or Gram) matrix $K = (p_i^\top p_j)_{1\le i,j\le n}$. Furthermore, it is also clear that the embedding dimension $\dim(D)$ equals to $\mathrm{rank}(K)$. Motivated by this correspondence between an Euclidean distance matrix and a kernel matrix, we consider estimating D by $\widehat{D} = (\widehat{d}_{ij})_{1\le i,j\le n}$ where

$$\widehat{d}_{ij} = \left\langle \widehat{K}, (e_i - e_j)(e_i - e_j)^\top \right\rangle = \widehat{k}_{ii} + \widehat{k}_{jj} - 2\widehat{k}_{ij}. \tag{2}$$

Here $\langle A, B\rangle = \mathrm{trace}(A^\top B)$, e_i is the ith column vector of the identity matrix, and $\widehat{K} = (\widehat{k}_{ij})_{1\le i,j\le n}$ is the the so-called regularized kernel estimate; see, e.g., Lu et al. (2005) and Weinberger et al. (2007). More specifically,

$$\widehat{K} = \underset{M\succeq 0}{\mathrm{argmin}} \left\{ \sum_{(i,j)\in\Omega} \left(x_{ij} - \left\langle M, (e_i - e_j)(e_i - e_j)^\top \right\rangle\right)^2 + \lambda_n \mathrm{trace}(M) \right\}, \tag{3}$$

where $\lambda_n \geq 0$ is a tuning parameter that balances the tradeoff between goodness-of-fit and the preference towards an estimate with smaller trace norm. Hereafter, we write $M \succeq 0$ to indicate that a matrix M is positive semi-definite. The trace norm penalty used in defining $\widehat{K}$ encourages low-rankness of the estimated kernel matrix and hence low embedding dimension of $\widehat{D}$. See, e.g., Lu et al. (2005), Yuan et al. (2007), Negahban and Wainwright (2011), Rohde and Tsybakov (2011), and Lu, Monteiro and Yuan (2012) among many others for similar use of this type of penalty. The goal of the current article is to study the operating characteristics and statistical performance of the estimate $\widehat{D}$ defined by (2).

A fundamental difficulty in understanding the behavior of the proposed distance matrix estimate $\widehat{D}$ comes from the simple observation that a kernel is not identifiable given pairwise distances alone, even without noise, as the latter is preserved under translation while the former is not. Therefore, it is not clear what exactly $\widehat{K}$ is estimating, and subsequently what the relationship between $\widehat{D}$ and D is. To address this challenge, we introduce a notion of minimum trace kernel to resolve the ambiguity associated with kernel estimation. Understanding of this concept allows us to more directly and explicitly characterize $\widehat{D}$ as first applying a constant amount of shrinkage to all observed distances; and then projecting the shrunken distances to an Euclidean distance matrix. Because the distance between a pair of points shrinks when they are projected onto a linear subspace, this characterization offers a geometrical explanation to the ability of $\widehat{D}$ to induce low dimensional embeddings. In addition, this direct characterization of $\widehat{D}$ also suggests an efficient way to compute it using a version of Dykstra's alternating projection algorithm thanks to the special geometric structure of $\mathcal{D}_n$, the set of $n \times n$ Euclidean distance matrices. See, e.g., Glunt et al. (1990). Obviation of semidefinite programming, and more generally second order cone programmings computational expense is the principal advantage of this alternating projection technique. Furthermore, based on this explicit characterization, we establish statistical risk bounds for the discrepancy $\widehat{D} - D$ and show that the true distances can be recovered consistently in average if D allows for (approximate) low dimensional embeddings.

The rest of the paper is organized as follows. In Section 2, we discuss in details the shrinkage effect of the estimate $\widehat{D}$ by exploiting the duality between a kernel matrix and an Euclidean distance matrix. Taking advantage of our explicit characterization of $\widehat{D}$ and the geometry of the convex cone of Euclidean distance matrices, Section 3 establishes risk bounds for $\widehat{D}$ and Section 4 describes how $\widehat{D}$ can be computed using an efficient alternating projection algorithm. The merits of $\widehat{D}$ is further illustrated via numerical examples, both simulated and real, in Section 5. All proofs are relegated to the Appendix.

2 Distance Shrinkage

In this section, we show that there is a one-to-one correspondence between an Euclidean distance matrix and a so-called minimum trace kernel; and exploit this duality explicitly to characterize $\widehat{D}$.

2.1 Minimum Trace Kernels

Despite the popularity of regularized kernel estimate $\widehat{K}$, rather little is known about its statistical performance. This is perhaps in a certain sense inevitable because a kernel is not identifiable given pairwise distances alone. To resolve this ambiguity, we introduce the concept of minimum trace kernel, and show that $\widehat{K}$ is targeting at the unique minimum trace kernel associated with the true Euclidean distance matrix.

Recall that any $n \times n$ positive semidefinite matrix K can be identified with a set of points $p_1, \ldots, p_n \in \mathbb{R}^k$ for some $k \in \mathbb{N}$ such that $K = PP^\top$ where $P = (p_1, \ldots, p_n)^\top$. At the same time, these points can also be associated with an $n \times n$ Euclidean distance matrix $D = (d_{ij})_{1 \le i,j \le n}$ where

$$d_{ij} = \|p_i - p_j\|^2, \qquad 1 \le i < j \le n.$$

Obviously,

$$d_{ij} = \langle K, B_{ij} \rangle,$$

where

$$B_{ij} = (e_i - e_j)(e_i - e_j)^\top.$$

It is clear that any positive semi-definite matrix M can be a kernel matrix and therefore translated uniquely into a distance matrix. In other words,

$$\mathscr{T}(M) = \mathrm{diag}(M)\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(M)^\top - 2M = (m_{ii} + m_{jj} - 2m_{ij})_{1 \le i,j \le n}$$

is a surjective map from the set $\mathcal{S}_n$ of $n \times n$ positive semi-definite matrices to $\mathcal{D}_n$. Hereafter, we write $\mathbf{1}$ as a vector of ones of conformable dimension. The map $\mathscr{T}$, however, is not injective because, geometrically, translation of the embedding points results in different kernel matrix yet the distance matrix remains unchanged. As a result, it may not be meaningful, in general, to consider reconstruction of a kernel matrix from dissimilarity scores alone.

It turns out that one can easily avoid such an ambiguity by requiring the embeddings to be centered in that $P^\top \mathbf{1} = \mathbf{0}$ where $\mathbf{0}$ is a vector of zeros of conformable dimension. We note that even with the centering, the embeddings as represented by P for any given

Euclidean distance matrix still may not be unique as distances are invariant to rigid motions. However, their corresponding kernel matrix, as the following result shows, is indeed uniquely defined. Moreover the kernel matrix can be characterized as having the smallest trace among all kernels that correspond to the same distance matrix, hence will be referred to as the minimum trace kernel.

Theorem 1. *Let D be an $n \times n$ distance matrix. Then the preimage of D under $\mathscr{T}$*

$$\mathcal{M}(D) = \{M \in \mathcal{S}_n : \mathscr{T}(M) = D\}$$

is convex; and $-JDJ/2$ is the unique solution to following convex program

$$\operatorname*{argmin}_{M \in \mathcal{M}(D)} \operatorname{trace}(M),$$

where $J = I - (\mathbf{1}\mathbf{1}^\top/n)$. In addition, if $p_1, \ldots, p_n \in \mathbb{R}^n$ is an embedding of D such that $p_1 + \ldots + p_n = \mathbf{0}$, then $PP^\top = -JDJ/2$, where $P = (p_1, \ldots, p_n)^\top$.

In the light of Theorem 1, $\mathscr{T}$ is bijective when restricted to the set of minimum trace kernels:

$$\mathcal{K} = \{M \succeq 0 : \operatorname{trace}(M) \leq \operatorname{trace}(A), \quad \forall A \in \mathcal{M}(\mathscr{T}(M))\}.$$

and its inverse is $\mathscr{R}(M) = -JMJ/2$ as a map from distance matrices to kernels with minimum trace. From this viewpoint, the regularized kernel estimate $\widehat{K}$ intends to estimate $\mathscr{R}(D)$ instead of the original data-generating kernel. In addition, it is clear that

Proposition 2. *For any $\lambda_n > 0$, the regularized kernel estimate $\widehat{K}$ as defined in (3) is a minimum trace kernel. In addition, any embedding $\widehat{P}$ of $\widehat{K}$, that is $\widehat{K} = \widehat{P}\widehat{P}^\top$, is necessarily centered so that $\widehat{P}^\top \mathbf{1} = \mathbf{0}$.*

The relationships among the data-generating kernel K, D, $\mathscr{R}(D)$, regularized kernel estimate $\widehat{K}$ as defined by (3), and the distance matrix estimate $\widehat{D}$ as defined by (2) can be described by Figure 1.

2.2 Distance Shrinkage

We now study the properties of the proposed distance matrix estimate given by (2). Recall that, in the case of complete observation, the regularized kernel estimate $\widehat{K}$ is given by

$$\widehat{K} = \operatorname*{argmin}_{M \succeq 0} \left\{ \frac{1}{2} \|X - \mathscr{T}(M)\|_{\mathrm{F}}^2 + \lambda_n \operatorname{trace}(M) \right\}, \tag{4}$$

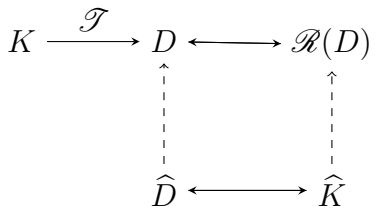


Figure 1: Relationships among K, D, $\mathscr{R}(D)$, $\widehat{K}$ and $\widehat{D}$: the true distance matrix D is determined by the data-generating kernel K; there is a one-to-one correspondence between D and the minimum trace kernel $\mathscr{R}(D)$. Similarly, there is a one-to-one correspondence between $\widehat{D}$ and $\widehat{K}$ which are estimate of D and $\mathscr{R}(D)$ respectively.

where $\|\cdot\|_{\mathrm{F}}$ stands for the usual matrix Frobenius norm. It turns out that, following Theorem 1, $\widehat{D} = \mathscr{T}(\widehat{K})$ actually allows for a more explicit and concise expression.

To this end, observe that the set $\mathcal{D}_n$ of $n \times n$ Euclidean distance matrices is a closed convex cone (Schönberg, 1935; Young and Householder, 1938). Let $\mathscr{P}_{\mathcal{D}_n}$ denote the projection to $\mathcal{D}_n$ in that

$$\mathscr{P}_{\mathcal{D}_n}(A) = \operatorname*{argmin}_{M \in \mathcal{D}_n} \|A - M\|_{\mathrm{F}}^2.$$

for $A \in \mathbb{R}^{n \times n}$. Then

Theorem 3. *Let $\widehat{D}$ be defined by (2) with the regularized kernel estimate $\widehat{K}$ given by (4). Then*

$$\widehat{D} = \mathscr{P}_{\mathcal{D}_n}\left(X - \frac{\lambda_n}{2n} D_0\right)$$

where D_0 is an Euclidean distance matrix whose diagonal elements are zero and off-diagonal entries are ones.

Theorem 3 characterizes $\widehat{D}$ as the projection of $X - (\lambda_n/2n)D_0$ to an Euclidean distance matrix. Therefore, it can be computed as soon as we can evaluate the projection onto the closed convex set $\mathcal{D}_n$. As shown in Section 4, this could be done efficiently using an alternating projection algorithm thanks to the geometric structure of $\mathcal{D}_n$. In addition, subtraction of $(\lambda_n/2n)D_0$ from X amounts to applying a constant shrinkage to all observed pairwise distances. Geometrically, distance shrinkage can be the result of projecting points in an Euclidean space onto a lower dimensional linear subspace, and therefore encourages low dimensional embeddings. We now look at the specific example when $n = 3$ to further illustrate such an effect.

In the special case of $n = 3$ points, the projection to Euclidean distance matrices can be computed analytically. Let

$$X = \begin{bmatrix} 0 & x_{12} & x_{13} \\ x_{12} & 0 & x_{23} \\ x_{13} & x_{23} & 0 \end{bmatrix}$$

be the observed distance matrix. We now determine the embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X - \eta D_0)$.

Let

$$Q = \frac{1}{3+\sqrt{3}} \begin{bmatrix} 2+\sqrt{3} & -1 & -(1+\sqrt{3}) \\ -1 & 2+\sqrt{3} & -(1+\sqrt{3}) \\ -(1+\sqrt{3}) & -(1+\sqrt{3}) & -(1+\sqrt{3}) \end{bmatrix}$$

be a 3×3 Householder matrix. Then, for a 3×3 symmetric hollow matrix X,

$$QXQ = \begin{bmatrix} -\frac{1}{3}x_{12} - \frac{1+\sqrt{3}}{3}x_{13} + \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{23} & \frac{2}{3}x_{12} - \frac{1}{3}x_{13} - \frac{1}{3}x_{23} & * \\ \frac{2}{3}x_{12} - \frac{1}{3}x_{13} - \frac{1}{3}x_{23} & -\frac{1}{3}x_{12} + \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{13} - \frac{1+\sqrt{3}}{3}x_{23} & * \\ * & * & * \end{bmatrix},$$

where we only give the 2×2 leading principle matrix of QXQ and leave the other entries unspecified. As shown by Hayden and Wells (1988), the minimal embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X)$ can be determined by the eigenvalues of the principle matrix.

More specifically, denote by

$$\tilde{D}(X) = \begin{bmatrix} \frac{1}{3}x_{12} + \frac{1+\sqrt{3}}{3}x_{13} - \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{23} & -\frac{2}{3}x_{12} + \frac{1}{3}x_{13} + \frac{1}{3}x_{23} \\ -\frac{2}{3}x_{12} + \frac{1}{3}x_{13} + \frac{1}{3}x_{23} & \frac{1}{3}x_{12} - \frac{1+\sqrt{3}}{6+3\sqrt{3}}x_{13} + \frac{1+\sqrt{3}}{3}x_{23} \end{bmatrix},$$

and

$$\tilde{D}(X) = U \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} U^\top$$

its eigenvalue decomposition. Write

$$\Delta_x := \sqrt{2[(x_{12} - x_{13})^2 + (x_{12} - x_{23})^2 + (x_{13} - x_{23})^2]}. \tag{5}$$

Then, it can be calculated that

$$\alpha_1 = \frac{(x_{12} + x_{13} + x_{23}) + \Delta_x}{3}, \qquad \text{and} \qquad \alpha_2 = \frac{(x_{12} + x_{13} + x_{23}) - \Delta_x}{3}. \tag{6}$$

In the light of Theorem 6.1 of Glunt et al. (1990), we have

Proposition 4.

$$\dim(\mathscr{P}_{\mathcal{D}_3}(X)) = \begin{cases} 2 & \text{if } x_{12}+x_{13}+x_{23} > \Delta_x \\ 1 & \text{if } -\frac{1}{2}\Delta_x < x_{12}+x_{13}+x_{23} \leq \Delta_x \\ 0 & \text{otherwise} \end{cases},$$

where Δ_x *is given by (5), and* $\dim(\mathscr{P}_{\mathcal{D}_3}(X)) = 0$ *means* $\mathscr{P}_{\mathcal{D}_3}(X) = \mathbf{0}$.

To appreciate the effect of distance shrinkage, consider the case when $\mathscr{P}_{\mathcal{D}_3}(X)$ has a minimum embedding dimension of two. By Proposition 4, this is equivalent to assuming $\alpha_2 > 0$. Observe that

$$\tilde{D}(X - \eta D_0) = \tilde{D}(X) - \eta I_2.$$

The eigenvalues of $\tilde{D}(X - \eta D_0)$ are therefore $\alpha_1 - \eta$ and $\alpha_2 - \eta$ where $\alpha_1 \geq \alpha_2$ are the eigenvalues of $\tilde{D}(X)$ as given by (6). This indicates that, by applying sufficient amount of distance shrinkage, we can reduce the minimum embedding dimension as illustrated in Figure 2.

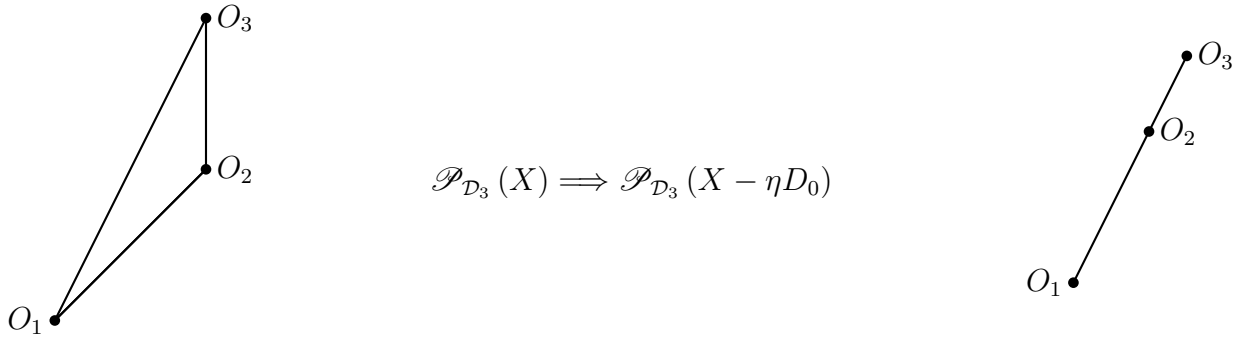


Figure 2: Effect of distance shrinkage when $n = 3$.

More specifically,

- If
$$\frac{1}{3}(x_{12}+x_{13}+x_{23}) - \frac{\Delta_x}{3} \leq \eta < \frac{1}{3}(x_{12}+x_{13}+x_{23}) + \frac{2\Delta_x}{3},$$
then the minimum embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X - \eta D_0)$ is one.

- If
$$\eta \geq \frac{1}{3}(x_{12}+x_{13}+x_{23}) + \frac{2\Delta_x}{3},$$
then the minimum embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X - \eta D_0)$ is zero;

3 Estimation Risk

The previous section provides an explicit characterization of the proposed distance matrix estimate $\widehat{D}$ as a distance shrinkage estimator. We now take advantage this characterization to establish statistical risk bounds for $\widehat{D}$.

3.1 Estimation Error for Distance Matrix

A natural measure of the quality of a distance matrix estimate $\tilde{D}$ is the averaged squared error of all pairwise distances:

$$L(\tilde{D}, D) := \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \left(\tilde{d}_{ij} - d_{ij} \right)^2 .$$

It is clear that when both $\tilde{D}$ and D are $n \times n$ Euclidean distance matrices,

$$L(\tilde{D}, D) = \frac{1}{n(n-1)} \|\tilde{D} - D\|_{\mathrm{F}}^2 .$$

For convenience, we shall now consider bounding $\|\widehat{D} - D\|_{\mathrm{F}}^2$. Taking advantage of the characterization of $\widehat{D}$ as a projection onto the set of $n \times n$ Euclidean distance matrices, we can derive the following oracle inequality.

Theorem 5. *Let $\widehat{D}$ be defined by (2). Then for any λ_n such that $\lambda_n \ge 2\|X - D\|$,*

$$\|\widehat{D} - D\|_{\mathrm{F}}^2 \le \inf_{M \in \mathcal{D}_n} \left\{ \|M - D\|_{\mathrm{F}}^2 + \frac{9}{4} \lambda_n^2 (\dim(M) + 1) \right\},$$

where $\|\cdot\|$ stands for the matrix spectral norm.

Theorem 5 gives a deterministic upper bound for the error of $\widehat{D}$, $\|\widehat{D} - D\|_{\mathrm{F}}^2$ in comparison with that of an arbitrary approximation to D. More specifically, let $\tilde{D}$ be the closest Euclidean distance matrix with embedding dimension r to D, in terms of Frobenius norm. Then Theorem 5 implies that with sufficiently large tuning parameter λ_n,

$$L(\widehat{D}, D) \le L(\tilde{D}, D) + \frac{Cr\lambda_n^2}{n^2},$$

for some constant $C > 0$. In particular, if D itself is embedding dimension r, then

$$L(\widehat{D}, D) \le \frac{Cr\lambda_n^2}{n^2}.$$

More explicit bounds for the estimation error can be derived from this general result. Consider, for example, the case when the observed pairwise distances are the true distances subject to additive noise:

$$x_{ij} = d_{ij} + \varepsilon_{ij}, \qquad 1 \le i < j \le n, \tag{7}$$

where the measurement errors ε_{ij}s are independent with mean $\mathbb{E}(\varepsilon_{ij}) = 0$ and variance $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$. Assume that the distributions of measurement errors have light tails such that

$$\mathbb{E}(\varepsilon_{ij})^{2m} \le (c_0 m)^m, \qquad \forall m \in \mathbb{N} \tag{8}$$

for some constant $c_0 > 0$. Then the spectral norm of $X - D$ satisfies

$$\|X - D\| = 2\sigma\left(\sqrt{n} + O_p(n^{-1/6})\right).$$

See, e.g., Sinai and Soshnikov (1998). Thus,

Corollary 6. *Let $\widehat{D}$ be defined by (2). Under the model given by (7) and (8), if $\lambda_n = 4\sigma(n^{1/2} + 1)$, then with probability tending to one,*

$$\|\widehat{D} - D\|_{\mathrm{F}}^2 \le \inf_{M \in \mathcal{D}_n} \left\{\|M - D\|_{\mathrm{F}}^2 + 36n\sigma^2(\dim(M) + 1)\right\},$$

as $n \to \infty$. In particular, if $\dim(D) = r$, then with probability tending to one,

$$\|\widehat{D} - D\|_{\mathrm{F}}^2 \le 36n\sigma^2(r + 1).$$

In other words, under the model given by (7) and (8),

$$L(\widehat{D}, D) \le L(\tilde{D}, D) + \frac{Cr\sigma^2}{n},$$

for some constant $C > 0$, where as before, $\tilde{D}$ is the closest Euclidean distance matrix to D with embedding dimension r. In particular, if D itself is embedding dimension r, then

$$L(\widehat{D}, D) \le \frac{Cr\sigma^2}{n}.$$

3.2 Low Dimensional Approximation

As mentioned before, in some applications, the chief goal may not be to recover D itself but rather its embedding in a prescribed dimension. This is true, in particular, for multidimensional scaling and graph realization where we are often interested in embedding a distance

matrix in $\mathbb{R}^2$ or $\mathbb{R}^3$. Following the classical multidimensional scaling, a parameter of interest in these cases is

$$D_r := \operatorname*{argmin}_{M \in \mathcal{D}_n(r)} \|J(D - M)J\|_{\mathrm{F}}^2,$$

where $\mathcal{D}_n(r)$ is the set of all $n \times n$ Euclidean distance matrices of embedding dimension at most r. An obvious estimate of D_r can be derived by replacing D with $\widehat{D}$:

$$\widehat{D}_r := \operatorname*{argmin}_{M \in \mathcal{D}_n(r)} \|J(\widehat{D} - M)J\|_{\mathrm{F}}^2. \tag{9}$$

Similar to the classical multidimensional scaling, the estimate $\widehat{D}_r$ can be computed more explicitly as follows. Let $\widehat{K}$ be the regularized kernel estimate corresponding to $\widehat{D}$, and $\widehat{K} = U\Gamma U^\top$ be its eigenvalue decomposition with $\Gamma = \mathrm{diag}(\gamma_1, \gamma_2, \ldots)$ and $\gamma_1 \geq \gamma_2 \geq \ldots$. Then $\widehat{D}_r = \mathscr{T}(\widehat{K}_r)$ where $\widehat{K}_r = U\mathrm{diag}(\gamma_1, \ldots, \gamma_r, 0, \ldots)U^\top$.

The risk bounds we derived for $\widehat{D}$ can also be translated into that for $\widehat{D}_r$. More specifically,

Corollary 7. *Let $\widehat{D}_r$ be defined by (9) where $\widehat{D}$ is given by (2) with $\lambda_n \geq 2\|X - D\|$. Then there exists a numerical constant $C > 0$ such that*

$$\|J(\widehat{D}_r - D)J\|_{\mathrm{F}}^2 \leq C \left(\min_{M \in \mathcal{D}_n(r)} \|J(D - M)J\|_{\mathrm{F}}^2 + \lambda_n^2 r \right),$$

In particular, under the model given by (7) and (8), if $\lambda_n = 4\sigma(n^{1/2}+1)$, then with probability tending to one,

$$\|J(\widehat{D}_r - D)J\|_{\mathrm{F}}^2 \leq C \left(\min_{M \in \mathcal{D}_n(r)} \|J(D - M)J\|_{\mathrm{F}}^2 + nr\sigma^2 \right).$$

4 Computation

It is not hard to see that the optimization problem involved in defining the regularized kernel estimate can be formulated as a second order cone program (see, e.g., Lu et al. 2005; Yuan et al., 2007). This class of optimization problems can be readily solved using generic solvers such as `SDPT3` (Toh, Todd and Tutuncu, 1999; Tutuncu, Toh and Todd, 2003). Although in principle, these problems can be solved in polynomial time, on the practical side, the solvers are known not to scale well to large problems. Instead of starting from the regularized kernel estimate, as shown in Section 3, $\widehat{D}$ can be directly computed as a projection onto the set of Euclidean distance matrices. Taking advantage of this direct characterization and the particular geometric structure of the closed convex cone $\mathcal{D}_n$, we can devise a more efficient algorithm to compute $\widehat{D}$.

4.1 Alternating Projection

We shall adopt, in particular, an alternating projection algorithm introduced by Dykstra (1983). Dykstra's algorithm is a refinement of the von Neumann alternating projection algorithm specifically designed to compute projection onto the intersection of two closed convex sets by constructing a sequence of projections to the two sets alternatively.

Data: x.
Result: Projection of x onto the intersection of two closed convex set $\mathcal{C}_1$ and $\mathcal{C}_2$.
Initialization: $x_0 = x$, $p_0 = 0$, $q_0 = 0$, $k = 0$;
repeat
  $s_k \leftarrow \mathscr{P}_{\mathcal{C}_1}(x_k + p_k)$;
  $p_{k+1} \leftarrow x_k + p_k - s_k$;
  $x_{k+1} \leftarrow \mathscr{P}_{\mathcal{C}_2}(s_k + q_k)$;
  $q_{k+1} \leftarrow s_k + q_k - x_{k+1}$;
  $k \leftarrow k + 1$;
until a certain convergence criterion is met;
return x_{k+1}.

Algorithm 1: Dykstra's alternating projection algorithm: $\mathscr{P}_{\mathcal{C}_1}$ and $\mathscr{P}_{\mathcal{C}_2}$ are the projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively.

The main idea of Dykstra's algorithm can be illustrated by Figure 3 where the projection of a point onto the intersection of two half-spaces is computed. The alternating projection algorithms, albeit simple, are very powerful and have found numerous applications in practice. It is also known that, under mild regularity conditions, the algorithm converges linearly regardless of the initial point. Interested readers are referred to the monograph by Escalante and Raydan (2011) for further details.

Now consider evaluating $\widehat{D}$ which is the projection of $X - \eta_n D_0$ onto $\mathcal{D}_n$. Observe that $\mathcal{D}_n$ is the intersection of two closed convex cones:

$$\mathcal{C}_1 = \{M \in \mathbb{R}^{n\times n} : JMJ \preceq 0\},$$

and

$$\mathcal{C}_2 = \{M \in \mathbb{R}^{n\times n} : \mathrm{diag}(M) = \mathbf{0}\}.$$

Dykstra's alternating projection algorithm can then be readily applied with input $X - \eta_n D_0$. The use of alternating projection algorithms is motivated by the fact that although $\mathscr{P}_{\mathcal{C}_1\cap\mathcal{C}_2}$

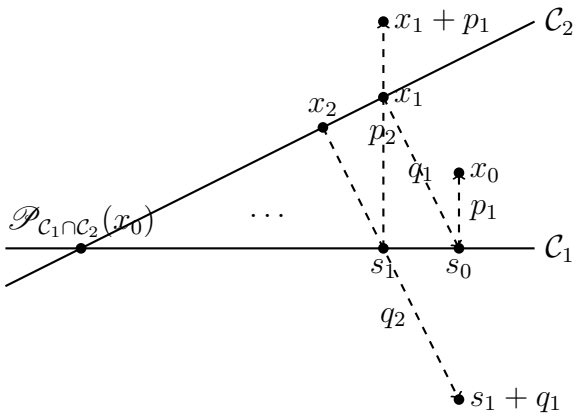


Figure 3: Illustration of alternating projection algorithm.

is difficult to evaluate, projections to $\mathcal{C}_1$ and $\mathcal{C}_2$ actually have explicit form and are easy to compute.

More specifically, for any symmetric matrix $A \in \mathbb{R}^{n \times n}$, let $\bar{A}_{11}$ be the $(n-1)$th leading principle submatrix of its Householder transform QAQ where $Q = I - vv^\top/n$ and $v = [1, \ldots, 1, 1 + \sqrt{n}]^\top$. In other words,

$$A = Q \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} Q$$

Let $\bar{A}_{11} = U \Gamma U^\top$ be its eigenvalue decomposition. Then

$$\mathscr{P}_{\mathcal{C}_1}(A) = Q \begin{bmatrix} U \Gamma^+ U^\top & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} Q$$

where $\Gamma^+ = \mathrm{diag}(\max\{\gamma_{ii}, 0\})$. See Hayden and Wells (1988). On the other hand, it is clear that $\mathscr{P}_{\mathcal{C}_2}(A)$ simply replaces all diagonal entries of A with zeros.

4.2 Dealing with Missing Data

We have thus far focused on the case when all pairwise distances are observable. Although this is true in many applications, there are also situations where some of the distances may not be available. Missing data can be conveniently handled within our framework through a combination of the alternating projection and EM algorithm.

More specifically, recall that $\Omega \subset \{(i,j) : 1 \le i < j \le n\}$ is the set of entries observed in

X. As the complete data case, we proceed to estimate D by $\widehat{D}^{\Omega} = \mathscr{T}(\widehat{K}^{\Omega})$ where

$$\widehat{K}^{\Omega} = \underset{M \succeq 0}{\operatorname{argmin}} \left\{ \sum_{(i,j)\in\Omega} \left(x_{ij} - \left\langle M, (e_i - e_j)(e_i - e_j)^{\top} \right\rangle \right)^2 + \lambda_n \operatorname{trace}(M) \right\}.$$

Here we use the superscript Ω to signify the dependence on the set Ω of the observed entries. Unlike the case without missing data, $\widehat{D}^{\Omega}$ in general can not be characterized as a projection of $X_{\Omega} = (x_{ij})_{(i,j)\in\Omega}$ onto the set of Euclidean distance matrices. To address this difficulty, we iterate between an E step where the missing observations are imputed using the current estimate of the pairwise distances, and an M step where we can appeal to the alternating projection algorithm on the observed distances along with those imputed in the E step.

Data: $X_{\Omega} = (x_{ij})_{(i,j)\in\Omega}$, $\eta_n \geq 0$
Result: $\widehat{D}^{\Omega}$
Initialization: initialize x_{ij} for $i < j$ and $(i,j) \notin \Omega$, and let $X = X^{\top} = (x_{ij})_{1\leq i,j\leq n}$ where $x_{ii} = 0$; $k = 0$, and $X^{(0)} = X$;
repeat
  M Step – $D^{(k+1)} = \mathscr{P}_{\mathcal{D}_n}(X^{(k)} - \eta_n D_0)$;
  E Step – $x_{ij}^{(k+1)} = x_{ij}$ if $(i,j) \in \Omega$, 0 if $i = j$, and $d_{ij}^{(k+1)}$ otherwise ;
until a certain convergence criterion is met;
$\widehat{D}^{\Omega} \leftarrow D^{(k+1)}$;
return $\widehat{D}^{\Omega}$.

Algorithm 2: EM algorithm to handle missing data.

4.3 Tuning

The ability to handle missing data also facilitates the tuning of λ_n or equivalently η_n. Clearly, the performance of the proposed method depends on the choice of the tuning parameter. In some cases, we want to embed data into an Euclidean space of a fixed dimensionality. For example, the atoms of a protein have to live in a three dimensional space. To this end, we can experiment with different values of the tuning parameter and use the one corresponding to the desired embedding dimension. Our experience suggests this strategy works fairly well in numerical experiments and the performance of the resulting estimate is also fairly stable for a broad range of tuning parameter choices. In many other situations, however, a more objective choice of tuning parameter may become desirable. A common strategy to address

this is through cross-validation, which can be done effectively using the algorithm presented before.

To do cross-validation, we first randomly divide the entries of X into T mutually exclusive subsets: $\Omega^{(1)}, \Omega^{(2)}, \ldots, \Omega^{(T)}$, for some fixed T, so that

$$\Omega^{(1)} \cup \Omega^{(2)} \cup \ldots \cup \Omega^{(T)} = \{(i,j) : 1 \leq i < j \leq n\}.$$

In particular, the choice of $T = 5$ or 10 is often advocated in practice (see, e.g., Hastie, Tibshirani and Friedman, 2009). For each $t = 1, \ldots, T$, we can then apply the algorithm given in the previous subsection to compute the distance shrinkage estimate with a given tuning parameter η_n based on partial observations:

$$X_{-\Omega^{(t)}} := \{X_{ij} : 1 \leq i < j \leq n,\ (i,j) \notin \Omega^{(t)}\}.$$

Denote by $\widehat{D}^{(t),\eta_n}$ $(t = 1, \ldots, T)$ the resulting estimates. We evaluate the suitability of a tuning parameter η_n by its cross validation score:

$$\mathrm{CV}(\eta_n) = \frac{1}{T}\sum_{t=1}^{T}\left[\sum_{(i,j)\in\Omega^{(t)}}\left(X_{ij} - \widehat{D}_{ij}^{(t),\eta_n}\right)^2\right].$$

The same procedure can be repeated for a sequence of different values of η_n, and the one associated with the smallest cross valuation score will be selected to the final choice. The distance shrinkage estimate based on this choice of the tuning parameter is then computed based on all observations to yield the final estimate.

5 Numerical Examples

To illustrate the practical merits of the proposed methods and the efficacy of the algorithm, we conducted several numerical experiments.

5.1 Sequence Variation of Vpu Protein Sequences

The current work was motivated in part by a recent study on the variation of Vpu (HIV-1 virus protein U) protein sequences and their relationship to preservation of tetherin and CD4 counter-activities (Pickering et al., 2014). Viruses are known for their fast mutation and therefore an important task is to understand the diversity within a viral population. Of particular interest in this study is a Vpu sequence repertoire derived from actively replicating

plasma virus from 14 HIV-1-infected individuals. Following standard MACS criteria, five of these individuals can be classified as Long-term nonprogressors, five as rapid progressors, and four as normal progressors, according to how long the progression from seroconversion to AIDS takes. A total of 304 unique amino acid sequences were obtained from this study.

We first performed pairwise alignment between these amino acid sequences using various BLOSUM substitution matrices. The results using different substitution matrices are fairly similar; and to fix ideas, we shall report here analysis based on the BLOSUM62 matrix. These pairwise similarity scores $\{s_{ij} : 1 \le i \le j \le n\}$ are converted into dissimilarity scores:

$$x_{ij} = s_{ii} + s_{jj} - 2s_{ij}, \qquad \forall 1 \le i < j \le n.$$

As mentioned earlier, $X = (x_{ij})_{1 \le i,j \le n}$ is not an Euclidean distance matrix. To this end, we first applied the classical multidimensional scaling to X. The three dimensional embedding is given in the top left panel of Figure 4. The amino acid sequences derived from the same individuals are represented by the same symbol and color. Different colors correspond to the three different classes of disease progression: long-term nonprogressors are represented in red, normal in green, and rapid progressors in purple. For comparison, we also computed $\widehat{D}$ with various choices of the tuning parameters. Similar to the observations made by Lu et al. (2005), the corresponding embeddings are qualitatively similar for a wide range of choices of λ_n. A typical one is given in the top right panel of Figure 4. It is clear that both embeddings share a lot of similarities. For example, sequences derived from the same individual are more similar as they tend to cluster together. The key difference, however, is that the embedding corresponding to $\widehat{D}$ suggests an outlying sequence. We went back to the original pairwise dissimilarity scores and identified the sequence as derived from a rapid progressor. It is fairly clear from the original scores that this sequence is different from the others. The minimum dissimilarity score from the particular sequence to any other sequence is 245 whereas the largest score between any other pair of sequences is 215. The histogram of the scores between the sequence and other sequences, or among other sequences are given in the bottom panel of Figure 4.

Given these observations, we now consider the analysis with the outlying sequence removed. To gain insight, we consider different choices of λ_n to visually inspect the Euclidean embeddings given by the proposed distance shrinkage. The embeddings given in Figure 5 correspond to λ_n equals 4000, 8000, 12000 and 16000 respectively. These embedding are qualitatively similar.

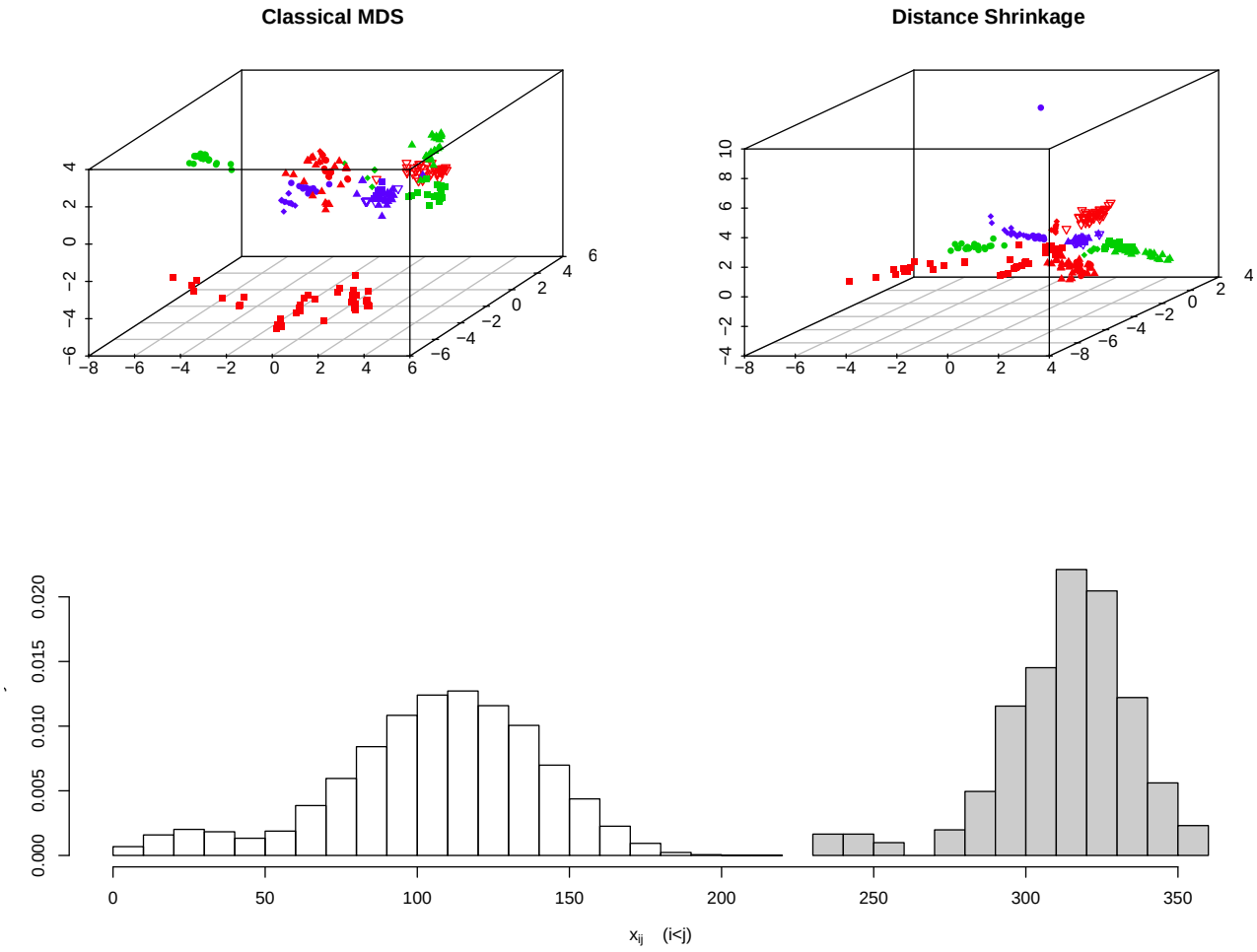


Figure 4: Three dimensional embedding for 304 amino acid sequences: the top panels are embeddings from classical multidimensional scaling and distance shrinkage respectively. The histogram of the pairwise dissimilarity scores is given in the bottom panel. The shaded histogram corresponds to those scores between the outlying sequence and the other sequences.

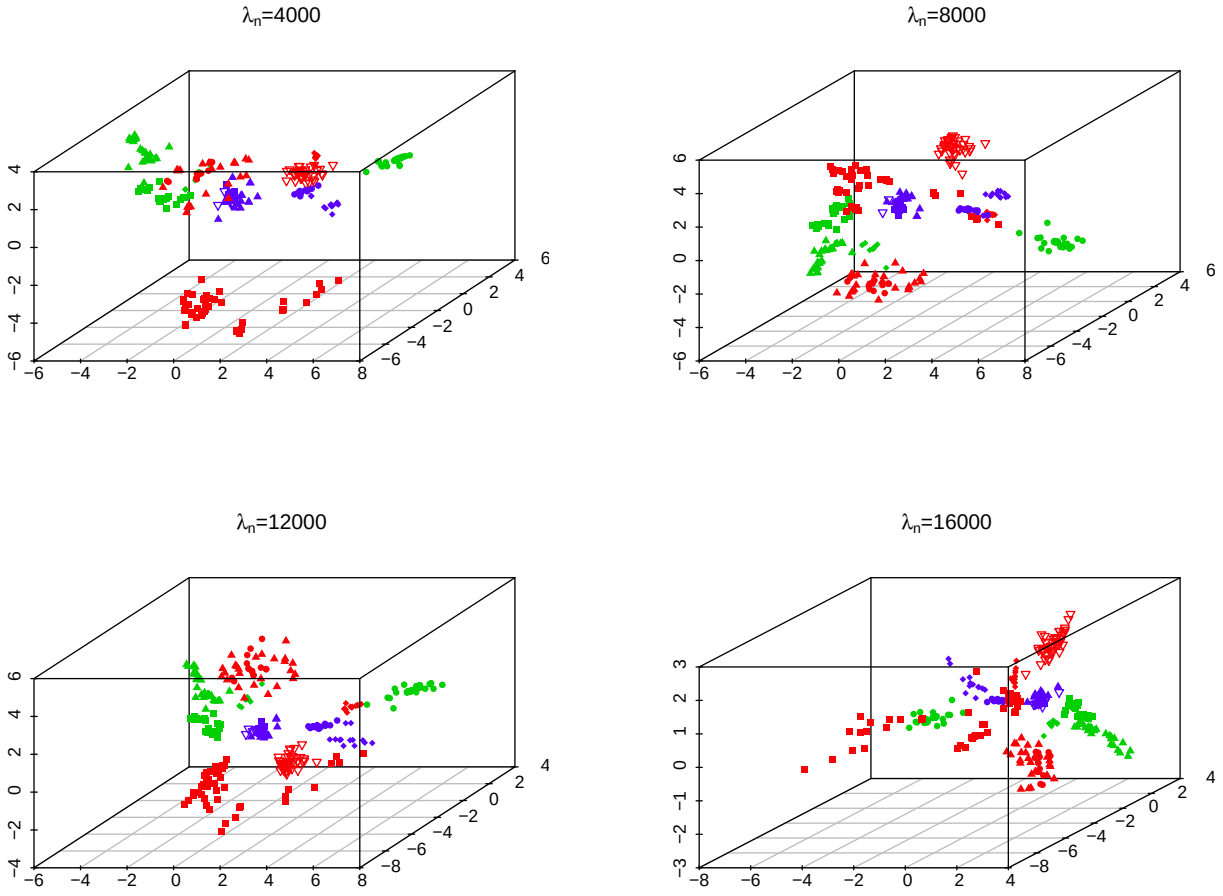


Figure 5: Euclidean embedding of 303 amino acid sequences via distance shrinkage: the outlying sequence was removed from the original data and each panel corresponds to different choice of λ_n.

5.2 Simulated Examples

To further compare the proposed distance shrinkage approach with the classical multidimensional scaling, we carried out several sets of simulation studies. For illustration purposes, we took the setup of the molecular conformation problem discussed earlier. In particular, we considered the problem of protein folding, a process of a random coil conformed to a physically stable three-dimensional structure equipped with some unique characteristics and functions.

We started by extracting the existing data on the 3D structure of the channel-forming trans-membrane domain of Vpu protein from HIV-1 mentioned before. The data obtained from protein data bank (symbol: 1PJE) contains the 3D coordinates of a total of $n = 91$ atoms. The exact Euclidean distance matrix D was then calculated from these coordinates. We note that in this case the embedding dimension is known to be three. We generated observations x_{ij} by adding an measurement error $\varepsilon_{ij} \sim N(0, \sigma^2)$ for $1 \leq i < j \leq n$. We considered three different values of σ^2 – 0.05, 0.25 and 0.5 respectively, representing relatively high, medium and low signal to noise ratio. For each value of σ^2, we simulated one hundred datasets and computed for each dataset the Euclidean distance matrix corresponding to the classical multidimensional scaling and the distance shrinkage. We evaluated the performance of each method by the Kruskal's stress defined as $\|\widehat{D} - D\|_{\mathrm{F}}/\|D\|_{\mathrm{F}}$. The results are summarized by Table 1.

Signal-to-Noise Ratio	Method	Mean	Standard error
High	Distance Shrinkage	0.010	2.0e-04
	Classical MDS	0.078	9.3e-04
Medium	Distance Shrinkage	0.024	4.8e-04
	Classical MDS	0.185	2.5e-03
Low	Distance Shrinkage	0.035	8.4e-04
	Classical MDS	0.301	3.9e-03

Table 1: Kruskal's stress for 1PJE data with measurement error.

To better appreciate the difference between the two methods, Figure 6 gives the ribbon plot of the protein backbone structure corresponding to the true Euclidean distance matrix and the estimated ones from a typical simulation run with different signal to noise ratios. It is noteworthy that the improvement of the distance shrinkage over the classical

multidimensional scaling becomes more evident with higher level of noise.

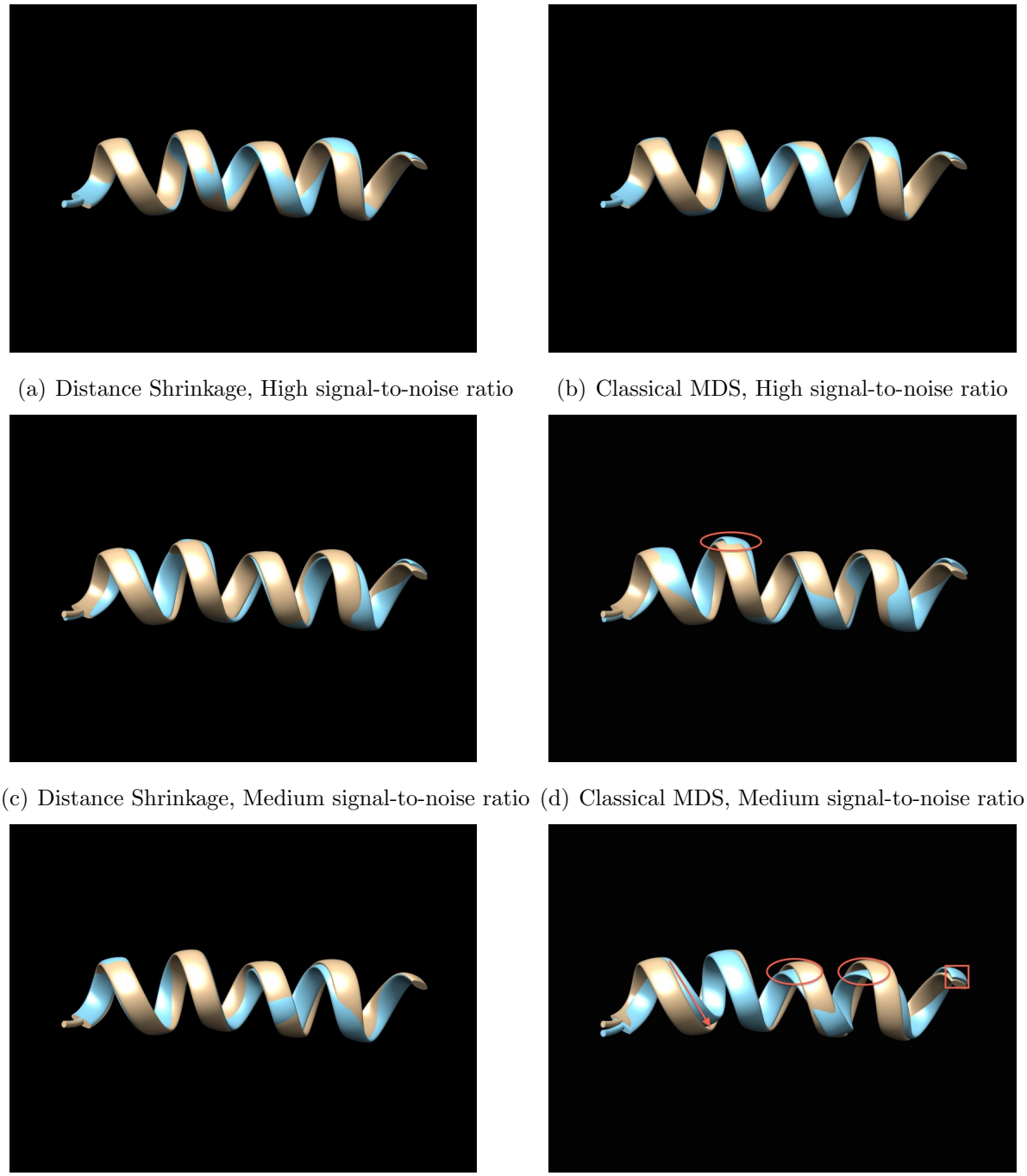
Our theoretical analysis suggests better performances for larger number of atoms. To further illustrate this effect of n, we repeated the previous experiment for HIV-1 virus protein U cytoplasmic domain (protein data bank symbol: 2K7Y) consisting of $n = 671$ atoms. We simulated data in the same fashion as before and the Kruskal stress, based on one hundred simulated dataset for each value of σ^2, is reported in Table 2. The performance compares favorable with that for 1PJE.

Signal-to-Noise Ratio	Method	mean	standard error
High	Distance Shrinkage	1.66e-04	2.70e-07
	Classical MDS	3.2e-03	4.84e-06
Medium	Distance Shrinkage	8.32e-04	1.48e-06
	Classical MDS	1.61e-02	2.45e-05
Low	Distance Shrinkage	1.7e-03	3.05e-06
	Classical MDS	3.22e-02	5.28e-05

Table 2: Kruskal's stress for 2K7Y data with measurement error.

To demonstrate the efficacy of cross-validation as a tuning method, we give in Figure 7 the true Kruskal stress as a function of the tuning parameter λ along with the five fold cross validation scores for a typical simulated dataset under each of the three levels of signal-to-noise ratio. These plots were generated by computing the distance matrix estimate for a series of values for the tuning parameter. It is clear from these plots that the tuning parameter selected by the cross validation is fairly close to optimal choice that minimizes the true Kruskal stress.

In the next set of simulation, we assess the effect of missing data for the proposed distance shrinkage estimate. Similar to before, we take the 3D coordinates data from protein data bank for five different proteins with different number of atoms. Pairwise distances were first computed for each of the protein. To mimic the typical NMR experiments, we assume that the larger distances are missing. In particular, we consider cases where the top 50%, 25% or 10% of the distances are unobservable. For those observed distances, independent Gaussian measurement errors with mean 0 and variance 0.5 were added. We ran the proposed distance shrinkage estimate on the simulated data. We experimented a range of tuning parameter choices and the performance is fairly similar. The results are summarized in the following table. As expected, the method performs better with the amount of missing data reduces.



(a) Distance Shrinkage, High signal-to-noise ratio

(b) Classical MDS, High signal-to-noise ratio

(c) Distance Shrinkage, Medium signal-to-noise ratio

(d) Classical MDS, Medium signal-to-noise ratio

(e) Distance Shrinkage, Low signal-to-noise ratio

(f) Classical MDS, Low signal-to-noise ratio

Figure 6: Ribbon plot of 1PJE protein back structure: the true structure is represented in gold whereas the structured corresponding to the estimated Euclidean distance matrix is given in blue. The left panels are for the distance shrinkage estimate whereas the right panels are for the the classical multidimensional scaling. Particular regions where the distance shrinkage shows visible improvement is circled out in red in the right panels.

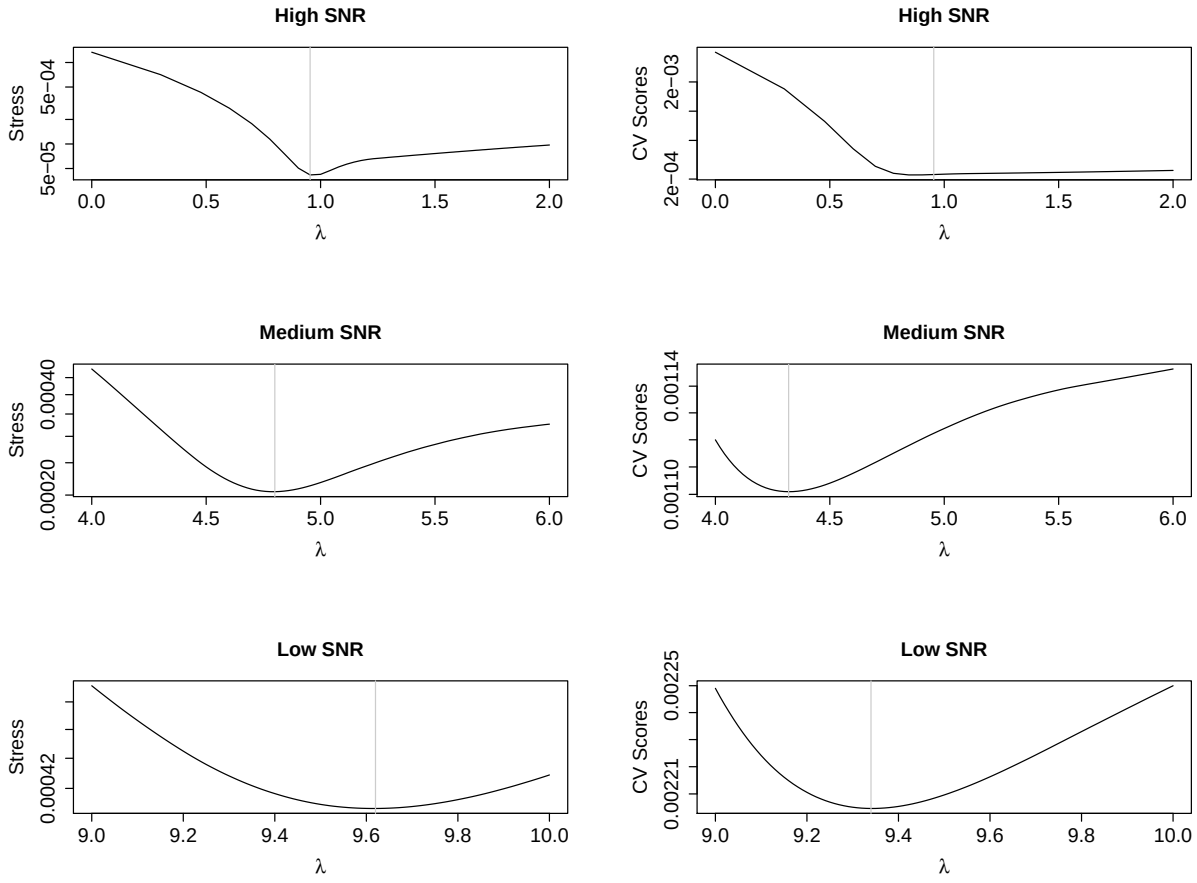


Figure 7: Comparison of Kruskal stress and cross-validation scores for simulated 2K7Y data. The right column gives plots of the Krusal stress as a function of the tuning parameter λ for different signal-to-noise ratios, and the left column gives plots of the cross-validation scores. In each panel, the minimizing tuning parameter is marked with the grey vertical line.

The distance shrinkage estimate works reasonably well even with 10% of missing data.

Table 3: Effect of Missing Data

PDB ID	# of Atoms	Kruskal's Stress		
		50% Missing	25% Missing	10% Missing
1PTQ	402	.57	.35	.18
1HOE	558	.56	.33	.15
1PHT	811	.56	.34	.17
1AX8	1003	.57	.36	.18

Finally, to further demonstrate the robustness of the approach to non-Gaussian measurement error, we generated pairwise distance scores between the 671 atoms following Gamma distributions:

$$x_{ij} \sim \text{Ga}(d_{ij}, 1), \qquad \forall 1 \leq i < j \leq 671,$$

so that both the mean and variance of x_{ij} are d_{ij}, where d_{ij} is the true squared distance between the ith and jth atoms. We again applied both classical multidimensional scaling and distance shrinkage to estimate the true distance matrix and reconstruct the 3D folding structure. The result from a typical simulated dataset is given in Figure 8.

References

[1] Chen, L. and Buja, A. (2009), Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis, *Journal of the American Statistical Association*, **104**, 209-219.

[2] Chen, L. and Buja, A. (2013), Stress functions for nonlinear dimension reduction, proximity analysis, and graph drawing, *Journal of Machine Learning Research*, **14**, 1145-1173.

[3] Darrotto, J. (2013), *Convex Optimization and Euclidean Distance Geometry*, Palo Alto: Meboo.

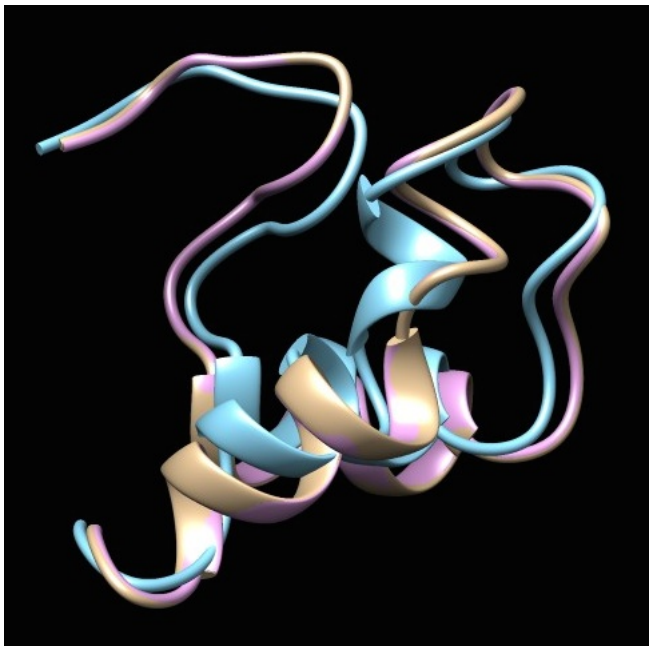


Figure 8: Ribbon plot of 2K7Y protein back structure: the true structure, and the structures corresponding to the classical multidimensional scaling and the distance shrinkage estimate are represented in gold, blue and pink respectively.

[4] Durbin, R., Eddy, S. Krogh, A. and Mitchison, G. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge: Cambridge University Press.

[5] Dykstra, R. (1983), An algorithm for restricted least squares regression, *Journal of the American Statistical Association*, **78**, 837-842.

[6] Escalante, R. and Raydan, M. (2011), *Alternating Projection Methods*, Philadelphia: Society for Industrial and Applied Mathematics.

[7] Glunt, W., Hayden, T., Hong, S. and Wells, J. (1990), An alternating projection algorithm for computing the nearest Euclidean distance matrix, *SIAM Journal of Matrix Analysis and Applications*, **11**, 589-600.

[8] Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning*, New York: Springer.

[9] Hyden, T.L. and Wells, J. (1988), Approximation by matrices positive semidefinite on a subspace, *Linear Algebra and Its Applications*, **109**, 115-130.

[10] Lu, F., Keles, S., Wright, S. and Wahba, G. (2005), Framework for kernel regularization with application to protein clustering, *Proceedings of the National Academy of Sciences*, **102**, 12332-12337.

[11] Lu, Z., Monteiro, R. and Yuan, M. (2012), Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression, *Mathematical Programming*, **131**, 163-194.

[12] Negahban S., Wainwright M. (2011), Estimation of (near) low-rank matrices with noise and high-dimensional scaling, *The Annals of Statistics*, **39(2)**, 1069-1097.

[13] Pickering, S., Hué, S., Kim, E., Reddy, S., Wolinsky, S. and Neil, S. (2014), Preservation of Tetherin and CD4 counter-activities in circulating Vpu alleles despite extensive sequence variation within HIV-1 infected individuals, *PLOS Pathogens*, **10(1)**, e1003895.

[14] Pouzet M. (1979), Note sur le probléme de Ulam, *Journal of Combinatorial Theory, Series B*, **27(3)**, 231-236.

[15] Rohde A., Tsybakov A. (2011), Estimation of high-dimensional low-rank matrices, *The Annals of Statistics*, **39(2)** 887-930.

[16] Roy, A. (2010), Minimal Euclidean representations of graphs, *Discrete Mathematics*, **310(4)**, 727-733.

[17] Schölkopf, B. and Smola, A. (1998), Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10** 1299-1319.

[18] Schölkopf, B. and Smola, A. (2002), *Learning with Kernels*, Cambridge: MIT Press.

[19] Schoenberg, I.J. (1935), Remarks to Maurice Frechet article "Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert", *Annals of Mathematics*, **38**, 724-732.

[20] Sinai, Y. and Soshnikov, A. (1998), A refinement of Wigners semi-circle law in a neighborhood of the spectrum edge for random symmetric matrices, *Functional Analysis and Its Applications*, **32(2)**, 114-131.

[21] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007), Measuring and testing independence by correlation of distances, *The Annals of Statistics*, **35**, 2769-2794.

[22] Tenenbaum J., De Silva V., Langford J. (2000), A global geometric framework for nonlinear dimensionality reduction, *Science*, **290(5500)**, 2319-2323.

[23] Toh, K.C., Todd, M.J. and Tutuncu, R.H. (1999), SDPT3 – a Matlab software package for semidefinite programming, *Optimization Methods and Software*, **11**, 545-581.

[24] Tutuncu, R.H., Toh, K.C. and Todd, M.J. (2003), Solving semidefinite-quadratic-linear programs using SDPT3, *Mathematical Programming Series B*, **95**, 189-217.

[25] Venna, J. and Kaski, S. (2006), Local multidimensional scaling, *Neural Networks*, **19**, 889-899.

[26] Wüthrich, K. (1986), *NMR of Proteins and Nucleic Acids*, New York: John Wiley Sons, Inc..

[27] Young, G. and Householder, A.S. (1938), Discussion of a set of points in terms of their mutual distances, *Psychometrika*, **3**, 19-22.

[28] Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007), Dimension reduction and coefficient estimation in multivariate linear regression, *Journal of the Royal Statistical Society: Series B*, **69**, 329-346.

Appendix – Proofs

Proof of Theorem 1. Denote by $M_0 = -JDJ/2$. We first show that $M_0 \in \mathcal{M}(D)$. Note first that

$$J(e_i - e_j) = (e_i - e_j).$$

Therefore,

$$\langle M_0, B_{ij} \rangle = -\frac{1}{2}(e_i - e_j)^\top JDJ(e_i - e_j) = -\frac{1}{2}(e_i - e_j)^\top D(e_i - e_j) = d_{ij},$$

where in the last equality follows from the facts that D is symmetric and $\mathrm{diag}(D) = \mathbf{0}$. Together with the fact that $M_0 \succeq 0$ (Schönberg, 1935; Young and Householder, 1938), this implies that $M_0 \in \mathcal{M}(D)$.

Next, we show that for any $M \in \mathcal{M}(D)$, $\mathrm{trace}(M_0) \leq \mathrm{trace}(M)$. To this end, observe that

$$D = \mathscr{T}(M) = \mathrm{diag}(M)\mathbf{1}^\top + \mathbf{1}\mathrm{diag}(M)^\top - 2M.$$

Then

$$\begin{aligned}
\mathrm{trace}(M_0) &= \mathrm{trace}(-JDJ/2) \\
&= \frac{1}{2}\mathrm{trace}\left[\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\right)\right] \\
&= \frac{1}{2}\mathrm{trace}\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right) \\
&\quad -\frac{1}{n}\mathbf{1}^\top\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\mathbf{1} \\
&\quad +\frac{1}{2n^2}\mathrm{trace}\left[\mathbf{1}\mathbf{1}^\top\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\mathbf{1}\mathbf{1}^\top\right] \\
&= -\frac{1}{2n}\mathbf{1}^\top\left(2M - \mathrm{diag}(M)\mathbf{1}^\top - \mathbf{1}\mathrm{diag}(M)^\top\right)\mathbf{1} \\
&= \mathrm{trace}(M) - \frac{1}{n}\mathbf{1}^\top M\mathbf{1}.
\end{aligned}$$

The positive semi-definiteness of M ensures that $\mathbf{1}^\top M\mathbf{1} \geq 0$, which implies that M_0 has the minimum trace in $\mathcal{M}(D)$. We now show it is also the only one.

Assume the contrary that there exists an $M \in \mathcal{M}(D)$ such that $M \neq M_0$ yet $\mathrm{trace}(M) = \mathrm{trace}(M_0)$. Following the previous calculation, we have $\mathbf{1}^\top M\mathbf{1} = 0$. Recall that $M \succeq 0$. The fact that $\mathbf{1}^\top M\mathbf{1} = 0$ necessarily implies that $\mathbf{1} \in \ker(M)$. As a result, $M = JMJ$, and

$$M - M_0 = J(M - M_0)J.$$

On the other hand,

$$\langle M, B_{ij}\rangle = \langle M_0, B_{ij}\rangle = d_{ij}, \quad \forall i < j.$$

Therefore,

$$\langle J(M - M_0)J, B_{ij}\rangle = \langle M - M_0, B_{ij}\rangle = 0, \quad \forall i < j.$$

It is not hard to see that

$$\{B_{ij} : i < j\} \cup \{e_i e_i^\top : 1 \le i \le n\}$$

forms a basis of the collection of $n \times n$ symmetric matrices. In other words, there exists α_{ij} $(1 \le i \le j)$ such that

$$M - M_0 = \sum_{1\le i<j\le n} \alpha_{ij} B_{ij} + \sum_{i=1}^{n-1} \alpha_{ii} e_i e_i.$$

Recall that $\mathbf{1} \in \ker(M) \cap \ker(M_0)$. Hence

$$(M - M_0)\mathbf{1} = [\alpha_{11}, \ldots, \alpha_{nn}]^\top = \mathbf{0}.$$

In other words,

$$M - M_0 = \sum_{1\le i<j\le n} \alpha_{ij} B_{ij}.$$

Thus

$$\|M - M_0\|_{\mathrm{F}}^2 = \|J(M - M_0)J\|_{\mathrm{F}}^2 = \sum_{1\le i<j\le n} \alpha_{ij}\langle J(M - M_0)J, B_{ij}\rangle = 0.$$

This obviously contradicts with the assumption that $M \neq M_0$.

The second statement follows from the same argument. Note that $PP^\top \in \mathcal{M}(D)$. Because the embedding points are centered, we have $\mathbf{1}^\top PP^\top \mathbf{1} = 0$. The previous argument then suggests that $PP^\top = M_0$. □

Proof of Theorem 3. Following Theorem 1, $\widehat{D} = \mathscr{T}(\widehat{K})$ can be equivalently expressed as

$$\widehat{D} = \underset{M\in\mathcal{D}_n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|X - M\|_{\mathrm{F}}^2 + \lambda_n \operatorname{trace}\left(-\frac{1}{2}JMJ\right)\right\}. \tag{10}$$

Recall that $J = I - (\mathbf{1}\mathbf{1}^\top/n)$. Observe that $D_0 = (n-1)I - nJ$. Therefore, for any $M \in \mathcal{D}_n$,

$$\begin{aligned}
\left\|\left(X - \frac{\lambda_n}{2n}D_0\right) - M\right\|_{\mathrm{F}}^2 &= \|X - M\|_{\mathrm{F}}^2 + \frac{\lambda_n}{n}\langle M, D_0\rangle + (\text{terms not involving } M) \\
&= \|X - M\|_{\mathrm{F}}^2 + \frac{\lambda_n}{n}\langle M, (n-1)I - nJ\rangle + (\text{terms not involving } M) \\
&= \|X - M\|_{\mathrm{F}}^2 - \lambda_n\langle M, J\rangle + (\text{terms not involving } M),
\end{aligned}$$

where the last equality follows from the fact that any distance matrix is hollow, e.g., its diagonals are zeros, hence $\langle M, I\rangle = 0$. Because J is idempotent,

$$\langle M, J\rangle = \langle M, J^2\rangle = \mathrm{trace}(JMJ).$$

Therefore,

$$\begin{aligned}\mathscr{P}_{\mathcal{D}_n}\left(X - \frac{\lambda_n}{2n}D_0\right) &= \underset{M\in\mathcal{D}_n}{\mathrm{argmin}}\left\{\frac{1}{2}\|X-M\|_{\mathrm{F}}^2 - \frac{\lambda_n}{2}\mathrm{trace}(JMJ)\right\} \\ &= \underset{M\in\mathcal{D}_n}{\mathrm{argmin}}\left\{\frac{1}{2}\|X-M\|_{\mathrm{F}}^2 + \lambda_n\mathrm{trace}\left(-\frac{1}{2}JMJ\right)\right\},\end{aligned}$$

which, in the light of (10), implies the desired statement. □

Proof of Theorem 5. By Theorem 3, $\widehat{D} = \mathscr{P}_{\mathcal{D}_n}(X - (\lambda_n/2n)D_0)$. Write $\eta_n = \lambda_n/(2n)$ for simplicity. Recall that for any $M \in \mathbb{R}^{n\times n}$, its projection to the closed convex set $\mathcal{D}_n$, $\mathscr{P}_{\mathcal{D}_n}(M)$, can be characterized by the so-called Kolmogorov criterion:

$$\langle A - \mathscr{P}_{\mathcal{D}_n}(M), M - \mathscr{P}_{\mathcal{D}_n}(M)\rangle \le 0, \qquad \forall A \in \mathcal{D}_n.$$

See, e.g., Escalante and Raydan (2011). In particular, taking $M = X - \eta_n D_0$ yields

$$\langle A - \widehat{D}, D - \widehat{D}\rangle \le \langle X - D - \eta_n D_0, \widehat{D} - A\rangle.$$

A classical result in distance geometry by Schönberg (1935) indicates that a distance matrix is conditionally negative semi-definite on the set

$$\mathcal{X}_n = \{x \in \mathbb{R}^n : x^\top \mathbf{1} = 0\},$$

that is, $x^\top M x \le 0$ for any $x \in \mathcal{X}_n$. See also Young and Householder (1938). In other words, if $M \in \mathcal{D}_n$, then the so-called Schönberg transform JMJ is negative semi-definite where, as before, $J = I - (\mathbf{1}\mathbf{1}^\top/n)$.

Let V be the eigenvectors of JAJ, and $V_\perp$ be an orthonormal basis of the orthogonal complement of the linear subspace spanned by $\{\mathbf{1}\}$ and V. Then $[\mathbf{1}/\sqrt{n}, V, V_\perp]$ forms an orthonormal basis of $\mathbb{R}^n$. Then for any symmetric matrix M, write

$$M = \mathscr{P}_0 M + \mathscr{P}_1 M,$$

where

$$\mathscr{P}_1 M = V_\perp V_\perp^\top M V_\perp V_\perp^\top$$

and

$$\begin{aligned}\mathscr{P}_0 M &= M - \mathscr{P}_1 M = [\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top M[\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top \\ &\quad + V_\perp V_\perp^\top M[\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top + [\mathbf{1}/\sqrt{n}, V][\mathbf{1}/\sqrt{n}, V]^\top M V_\perp V_\perp^\top.\end{aligned}$$

Therefore,

$$\begin{aligned}\langle X - D, \widehat{D} - A\rangle &= \langle \mathscr{P}_0(X - D), \mathscr{P}_0(\widehat{D} - A)\rangle + \langle \mathscr{P}_1(X - D), \mathscr{P}_1(\widehat{D} - A)\rangle \\ &= \langle \mathscr{P}_0(X - D), \mathscr{P}_0(\widehat{D} - A)\rangle + \langle \mathscr{P}_1(X - D), \mathscr{P}_1\widehat{D}\rangle \\ &\leq \|\mathscr{P}_0(X - D)\|\|\mathscr{P}_0(\widehat{D} - A)\|_* + \|\mathscr{P}_1(X - D)\|\|\mathscr{P}_1\widehat{D}\|_*\end{aligned}$$

where in the last inequality we used the fact that for any matrices $M_1, M_2 \in \mathbb{R}^{n\times n}$,

$$\langle M_1, M_2\rangle \leq \|M_1\|\|M_2\|_*,$$

and $\|\cdot\|$ and $\|\cdot\|_*$ represent the matrix spectral and nuclear norm respectively. It is clear that

$$\|\mathscr{P}_1(X - D)\| \leq \|X - D\|,$$

and

$$\|\mathscr{P}_0(X - D)\| \leq 2\|X - D\|.$$

Then,

$$\langle X - D, \widehat{D} - A\rangle \leq \|X - D\|\left(2\|\mathscr{P}_0(\widehat{D} - A)\|_* + \|\mathscr{P}_1\widehat{D}\|_*\right)$$

On the other hand, recall that both D and $\widehat{D}$ are hollow and $D_0 = (n-1)I - nJ$. Thus,

$$\begin{aligned}\langle D_0, \widehat{D} - A\rangle &= n\langle A - \widehat{D}, J\rangle \\ &= n\mathrm{trace}(J(A - \widehat{D})J) \\ &= -n\mathrm{trace}(VV^\top(\widehat{D} - A)VV^\top) - n\mathrm{trace}(\mathscr{P}_1\widehat{D}) \\ &= -n\mathrm{trace}(VV^\top(\widehat{D} - A)VV^\top) + n\|\mathscr{P}_1\widehat{D}\|_* \\ &\geq -n\|VV^\top(\widehat{D} - A)VV^\top\|_* + n\|\mathscr{P}_1\widehat{D}\|_* \\ &\geq -n\|\mathscr{P}_0(\widehat{D} - A)\|_* + n\|\mathscr{P}_1\widehat{D}\|_*,\end{aligned}$$

where the last equality follows from the fact that $\mathscr{P}_1\widehat{D}$ is negative semi-definite.

Taking $n\eta_n \geq \|X - D\|$ yields that

$$\langle X - D - \lambda_n D_0, \widehat{D} - A\rangle \leq 3n\eta_n\|\mathscr{P}_0(\widehat{D} - A)\|_*.$$

Note that, by Cauchy-Schwartz inequality, for any $M \in \mathbb{R}^{n\times n}$

$$\|M\|_* \leq \sqrt{\mathrm{rank}(M)}\|M\|_{\mathrm{F}}.$$

Therefore,

$$\begin{aligned}
\|\mathscr{P}_0(\widehat{D}-A)\|_* &\leq \sqrt{\mathrm{rank}(JAJ)+1}\|\mathscr{P}_0(\widehat{D}-A)\|_{\mathrm{F}} \\
&\leq \sqrt{\mathrm{rank}(JAJ)+1}\|\widehat{D}-A\|_{\mathrm{F}} \\
&= \sqrt{\dim(A)+1}\|\widehat{D}-A\|_{\mathrm{F}},
\end{aligned}$$

where the last equality follows from the fact that for any Euclidean distance matrix A, $\dim(A) = \mathrm{rank}(JAJ)$. See, e.g., Schönberg (1935) and Young and Householder (1938). As a result,

$$\langle A-\widehat{D}, D-\widehat{D}\rangle \leq 3n\eta_n\sqrt{\dim(A)+1}\|\widehat{D}-A\|_{\mathrm{F}}.$$

Simple algebraic manipulations show that

$$\langle A-\widehat{D}, D-\widehat{D}\rangle = \frac{1}{2}\left(\|\widehat{D}-D\|_{\mathrm{F}}^2 + \|\widehat{D}-A\|_{\mathrm{F}}^2 - \|A-D\|_{\mathrm{F}}^2\right).$$

Thus,

$$\|\widehat{D}-D\|_{\mathrm{F}}^2 + \|\widehat{D}-A\|_{\mathrm{F}}^2 \leq \|A-D\|_{\mathrm{F}}^2 + 6n\eta_n\sqrt{\dim(A)+1}\|\widehat{D}-A\|_{\mathrm{F}},$$

which implies that

$$\begin{aligned}
\|\widehat{D}-D\|_{\mathrm{F}}^2 &\leq \|A-D\|_{\mathrm{F}}^2 + 6n\eta_n\sqrt{\dim(A)+1}\|\widehat{D}-A\|_{\mathrm{F}} - \|\widehat{D}-A\|_{\mathrm{F}}^2 \\
&= \|A-D\|_{\mathrm{F}}^2 + 9n^2\eta_n^2(\dim(A)+1) - \left(\|\widehat{D}-A\|_{\mathrm{F}} - 3n\eta_n\sqrt{\dim(A)+1}\right)^2 \\
&\leq \|A-D\|_{\mathrm{F}}^2 + 9n^2\eta_n^2(\dim(A)+1).
\end{aligned}$$

This completes the proof. □

Proof of Corollary 7. Observe first that

$$\widehat{D}_r = \underset{M\in\mathcal{D}_r}{\mathrm{argmin}}\,\|J(M-\widehat{D})J\|_{\mathrm{F}}^2.$$

Therefore,

$$\begin{aligned}
\|J(\widehat{D}_r-D)J\|_{\mathrm{F}}^2 &\leq 2\|J(\widehat{D}_r-\widehat{D})J\|_{\mathrm{F}}^2 + 2\|J(\widehat{D}-D)J\|_{\mathrm{F}}^2 \\
&\leq 2\|J(D_r-\widehat{D})J\|_{\mathrm{F}}^2 + 2\|\widehat{D}-D\|_{\mathrm{F}}^2 \\
&\leq 4\|J(D_r-D)J\|_{\mathrm{F}}^2 + 4\|J(\widehat{D}-D)J\|_{\mathrm{F}}^2 + 2\|\widehat{D}-D\|_{\mathrm{F}}^2 \\
&\leq 4\min_{M\in\mathcal{D}_n(r)}\|J(D-M)J\|_{\mathrm{F}}^2 + 6\|\widehat{D}-D\|_{\mathrm{F}}^2
\end{aligned}$$

On the other hand, taking $M = D_r$ in Theorem 5 yields

$$\begin{aligned}\frac{1}{n^2}\|\widehat{D} - D\|_{\mathrm{F}}^2 &\leq \frac{1}{n^2}\|D_r - D\|_{\mathrm{F}}^2 + 9\eta_n^2(r+1) \\ &= \frac{1}{n^2}\min_{M\in\mathcal{D}_n(r)}\|J(D-M)J\|_{\mathrm{F}}^2 + 9\eta_n^2(r+1),\end{aligned}$$

where, as before, $\eta_n = \lambda_n/2n$. Therefore,

$$\frac{1}{n^2}\|J(\widehat{D}_r - D)J\|_{\mathrm{F}}^2 \leq \frac{10}{n^2}\min_{M\in\mathcal{D}_n(r)}\|J(D-M)J\|_{\mathrm{F}}^2 + 54\eta_n^2(r+1),$$

which completes the proof. □