ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2016)
78, *Part* 4, *pp.* 849–867

Distance shrinkage and Euclidean embedding via regularized kernel estimation

Luwan Zhang, Grace Wahba and Ming Yuan

University of Wisconsin—Madison, USA

[Received September 2014. Revised July 2015]

Summary. Although recovering a 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 on 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 which is known not to scale well for large problems. Numerical experiments and an application in visualizing the diversity of Vpu protein sequences from a recent study of human immunodeficiency virus type 1 further demonstrate the practical merits of the method proposed.

Keywords: Embedding; Euclidean distance matrix; Kernel; Multi-dimensional scaling; Regularization; Shrinkage; Trace norm

1. Introduction

The problem of recovering a 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 it 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 by 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 semidefinite kernels or Euclidean distances, could allow themselves to play an important role in a myriad of statistical and computational analyses (e.g. Schölkopf and Smola (2002) and Székely *et al.* (2007)).

A canonical example where this type of problem occurs is multi-dimensional scaling which aims to place each object in a low dimensional Euclidean space such that the between-object dis-

Address for correspondence: Ming Yuan, Department of Statistics, University of Wisconsin—Madison, 1300 University Avenue, Madison, WI 53706, USA.
E-mail: ming.mingyuan@gmail.com

1369–7412/16/78849

tances are preserved as well as possible. As such it also forms the basis for several other more recent approaches to non-linear dimension reduction and manifold learning. See Schölkopf (1998), Tenenbaum *et al.* (2000), Lu *et al.* (2005), Venna and Kaski (2006), Weinberger *et al.* (2007) and Chen and Buja (2009, 2013) among others. Despite the popularity of multi-dimensional scaling, very little is known about to what extent the distances between the embedded points could faithfully reflect the true pairwise distances when observed with noise, and it is largely used only as an exploratory tool for initial data analysis.

Another example where it is of interest to reconstruct a Euclidean distance matrix is the determination of molecular structures by using nuclear magnetic resonance spectroscopy, which is a technique that was pioneered by Nobel laureate Kurt Wüthrich (see for example, Wüthrich (1986)). As demonstrated by Wüthrich, distances between atoms could be inferred from chemical shifts measured by nuclear magnetic resonance 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 distances observed may not translate directly into locations of these atoms in a stable structure. Therefore, this becomes a problem of recovering a Euclidean distance matrix in three dimensions 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 a Euclidean space in such a fashion that the distance between two embedded vertices matches their corresponding edge weight (see, for example, Pouzet (1979)). Although an exact embedding of a graph is typically of very high dimension, it is useful in some applications to seek approximate yet low dimensional embeddings instead (see, for example, 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 co-ordinates of atoms in the case of molecular structure determination by using nuclear magnetic resonance spectroscopy, or the vertex set of a graph in the case of graph realization. Let Ω be a subset of $\{(i, j) : 1 \leqslant i, j \leqslant n\}$, 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 $D = (d_{ij})_{1 \leqslant i, j \leqslant n}$ is a so-called Euclidean distance matrix in that there are points $p_1, \ldots, p_n \in \mathbb{R}^k$ for some $k \in \mathbb{N}$ such that

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

see, for example, 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, i.e. $\Omega = \{(i, j) : 1 \leqslant i < j \leqslant n\}$. In these cases, we can more conveniently write the observed scores in a matrix form $X = (x_{ij})_{1 \leqslant i, j \leqslant n}$ 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 result (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, which is 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 embedding dimension of the true distance matrix D is necessarily 3. Similarly, for multi-dimensional scaling or graph realization, we typically are interested in mapping objects in two or three dimensions.

Recall that

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

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

$$\hat{d}_{ij} = \langle \hat{K}, (e_i - e_j)(e_i - e_j)^{\mathrm{T}} \rangle = \hat{k}_{ii} + \hat{k}_{jj} - 2\hat{k}_{ij}. \tag{2}$$

Here $\langle A, B \rangle = \mathrm{tr}(A^{\mathrm{T}} B)$, e_i is the ith column vector of the identity matrix and $\hat{K} = (\hat{k}_{ij})_{1 \leqslant i,j \leqslant n}$ is the the so-called regularized kernel estimate; see, for example, Lu *et al.* (2005) and Weinberger *et al.* (2007). More specifically,

$$\hat{K} = \underset{M \succeq 0}{\arg\min} \Big[\sum_{(i,j) \in \Omega} \{x_{ij} - \langle M, (e_i - e_j)(e_i - e_j)^{\mathrm{T}} \rangle\}^2 + \lambda_n \mathrm{tr}(M) \Big], \tag{3}$$

where $\lambda_n \geqslant 0$ is a tuning parameter that balances the trade-off 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 semidefinite. The trace norm penalty that is used in defining $\hat{K}$ encourages low rankness of the estimated kernel matrix and hence low embedding dimension of $\hat{D}$. See, for example, Lu *et al.* (2005), Yuan *et al.* (2007), Negahban and Wainwright (2011), Rohde and Tsybakov (2011) and Lu *et al.* (2012) among many others for similar use of this type of penalty. The goal of the current paper is to study the operating characteristics and statistical performance of the estimate $\hat{D}$ defined by equation (2).

A fundamental difficulty in understanding the behaviour of the proposed distance matrix estimate $\hat{D}$ comes from the simple observation that a kernel is not identifiable given pairwise distances alone, even without noise, as the distance matrix is preserved under translation whereas the kernel matrix is not. Therefore, it is not clear what exactly $\hat{K}$ is estimating, and subsequently what the relationship between $\hat{D}$ and D is. To address this challenge, we introduce a notion of minimum trace kernel to resolve the ambiguity that is associated with kernel estimation. Understanding this concept allows us to characterize $\hat{D}$ more directly and explicitly as first applying a constant amount of shrinkage to all observed distances, and then projecting the shrunken distances to a 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 $\hat{D}$ to induce low dimensional embeddings. In addition, this direct characterization of $\hat{D}$ also suggests an efficient way to compute it by 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, for example, Glunt *et al.* (1990). Obviation of semidefinite programming, and more generally second-order cone programming's computational expense, is the principal advantage of this alternating projection technique. Furthermore, on the basis of this explicit characterization, we establish statistical risk bounds for the discrepancy $\hat{D} - D$ 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 detail the shrinkage effect of the estimate $\hat{D}$ by exploiting the duality between a kernel matrix and a Euclidean distance matrix. Taking advantage of our explicit characterization of $\hat{D}$ and the geometry of the convex cone of Euclidean distance matrices, Section 3 establishes risk bounds for $\hat{D}$ and Section 4 describes how $\hat{D}$ can be computed by using an efficient alternating projection algorithm. The merits of $\hat{D}$ are further illustrated via numerical examples, both simulated and real, in Section 5. All proofs are relegated to the on-line appendix.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

http://wileyonlinelibrary.com/journal/rss-datasets

2. Distance shrinkage

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

2.1. Minimum trace kernels

Despite the popularity of regularized kernel estimate $\hat{K}$, 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 a minimum trace kernel and show that $\hat{K}$ 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 $K = PP^{\mathrm{T}}$ where $P = (p_1, \ldots, p_n)^{\mathrm{T}}$. At the same time, these points can also be associated with an $n \times n$ Euclidean distance matrix $D = (d_{ij})_{1 \leqslant i,j \leqslant n}$ where

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

Obviously,

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

where

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

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

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

is a surjective map from the set $\mathcal{S}_n$ of $n \times n$ positive semidefinite matrices to $\mathcal{D}_n$. Hereafter, we write $\mathbf{1}$ as a vector of 1s of conformable dimension. The map $\mathscr{T}$, however, is not injective because, geometrically, translation of the embedding points results in a 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 we can easily avoid such an ambiguity by requiring the embeddings to be centred in that $P^{\mathrm{T}}\mathbf{1} = \mathbf{0}$ where $\mathbf{0}$ is a vector of 0s of conformable dimension. We note that, even with the centring, 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, and hence it 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 the convex program

$$\underset{M \in \mathcal{M}(D)}{\arg\min}\ \operatorname{tr}(M),$$

where $J = I - (\mathbf{1}\mathbf{1}^{\mathrm{T}}/n)$. 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 $PP^{\mathrm{T}} = -JDJ/2$, where $P = (p_1, \ldots, p_n)^{\mathrm{T}}$.

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

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

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 $\hat{K}$ intends to estimate $\mathscr{R}(D)$ instead of the original data-generating kernel. In addition, the following proposition is clear.

Proposition 1. For any $\lambda_n > 0$, the regularized kernel estimate $\hat{K}$ as defined in equation (3) is a minimum trace kernel. In addition, any embedding $\hat{P}$ of $\hat{K}$, i.e. $\hat{K} = \hat{P}\hat{P}^{\mathrm{T}}$, is necessarily centred so $\hat{P}^{\mathrm{T}}\mathbf{1} = \mathbf{0}$.

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

2.2. Distance shrinkage

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

$$\hat{K} = \underset{M \succeq 0}{\arg\min}\{\tfrac{1}{2}\|X - \mathscr{T}(M)\|_{\mathrm{F}}^2 + \lambda_n \operatorname{tr}(M)\}, \tag{4}$$

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

For this, 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 $\mathscr{P}q_{\mathcal{D}_n}$ denote the projection to $\mathcal{D}_n$ in that

$$\begin{tikzcd}
K \arrow[r, "\mathscr{T}"] & D \arrow[r, leftrightarrow] & \mathscr{R}(D) \\
& \widehat{D} \arrow[u, dashed] \arrow[r, leftrightarrow] & \widehat{K} \arrow[u, dashed]
\end{tikzcd}$$

Fig. 1. Relationships between K, D, $\mathscr{R}(D)$, $\hat{K}$ and $\hat{D}$: 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 $\hat{D}$ and $\hat{K}$ which are estimates of D and $\mathscr{R}(D)$ respectively

$$\mathscr{P}_{\mathcal{D}_n}(A) = \underset{M\in\mathcal{D}_n}{\arg\min} \|A-M\|_{\mathrm{F}}^2$$

for $A \in \mathbb{R}^{n\times n}$. Then we have the following theorem.

Theorem 2. Let $\hat{D}$ be defined by expression (2) with the regularized kernel estimate $\hat{K}$ given by equation (4). Then

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

where D_0 is a Euclidean distance matrix whose diagonal elements are 0 and off-diagonal entries are 1s.

Theorem 2 characterizes $\hat{D}$ as the projection of $X-(\lambda_n/2n)D_0$ onto a 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 by 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 a 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 illustrate such an effect further.

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

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

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

Let

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

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

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

where we give only the 2×2 leading principal 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 principal matrix.

More specifically, denote by

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

and

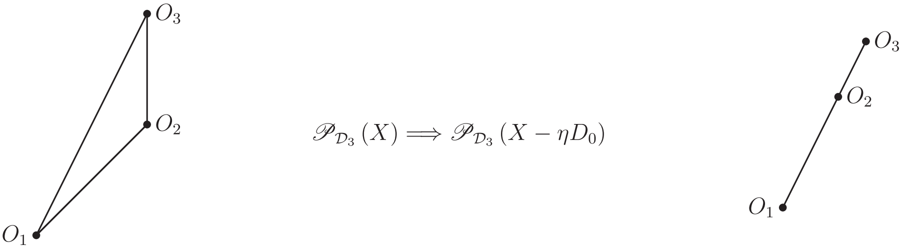

O_3
O_2
O_1
$\mathscr{P}_{\mathcal{D}_3}(X) \Longrightarrow \mathscr{P}_{\mathcal{D}_3}(X-\eta D_0)$
O_3
O_2
O_1

Fig. 2. Effect of distance shrinkage when $n=3$

$$\tilde{D}(X)=U\begin{pmatrix}\alpha_1 & 0\\ 0 & \alpha_2\end{pmatrix}U^{\mathrm{T}}$$

its eigenvalue decomposition. Write

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

Then, it can be calculated that

$$\begin{aligned}\alpha_1&=\frac{x_{12}+x_{13}+x_{23}+\Delta_x}{3},\\ \alpha_2&=\frac{x_{12}+x_{13}+x_{23}-\Delta_x}{3}.\end{aligned} \tag{6}$$

In the light of theorem 6.1 of Glunt *et al.* (1990), we have the following proposition.

Proposition 2.

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

where Δ_x is given by equation (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 2. By proposition 2, this is equivalent to assuming that $\alpha_2>0$. 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 $\alpha_1\geqslant\alpha_2$ are the eigenvalues of $\tilde{D}(X)$ as given by expression (6). This indicates that, by applying a sufficient amount of distance shrinkage, we can reduce the minimum embedding dimension as illustrated in Fig. 2.

More specifically,

(a) if

$$\frac{1}{3}(x_{12}+x_{13}+x_{23})-\frac{\Delta_x}{3}\leqslant\eta<\frac{1}{3}(x_{12}+x_{13}+x_{23})+\frac{2\Delta_x}{3},$$

then the minimum embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X-\eta D_0)$ is 1 and,

(b) if

$$\eta\geqslant\frac{1}{3}(x_{12}+x_{13}+x_{23})+\frac{2\Delta_x}{3},$$

then the minimum embedding dimension of $\mathscr{P}_{\mathcal{D}_3}(X-\eta D_0)$ is 0.

3. Estimation risk

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

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:

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

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

Theorem 3. Let $\hat{D}$ be defined by expression (2). Then, for any λ_n such that $\lambda_n \geqslant 2\|X - D\|$,

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

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

Theorem 3 gives a deterministic upper bound for the error of $\hat{D}$, $\|\hat{D} - D\|_{\mathrm{F}}^2$, 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 3 implies that, with sufficiently large tuning parameter λ_n,

$$L(\hat{D}, D) \leqslant L(\tilde{D}, D) + \frac{C r \lambda_n^2}{n^2},$$

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

$$L(\hat{D}, D) \leqslant \frac{C r \lambda_n^2}{n^2}.$$

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:

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

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

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

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

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

See, for example, Sinai and Soshnikov (1998). Thus we have the following corollary.

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

$$\|\hat{D} - D\|_{\mathrm{F}}^2 \leqslant \inf_{M \in \mathcal{D}_n} [\|M - D\|_{\mathrm{F}}^2 + 36n\sigma^2\{\dim(M)+1\}],$$

as $n \to \infty$. In particular, if $\dim(D) = r$, then, with probability tending to 1,

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

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

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

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

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

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 multi-dimensional 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 multi-dimensional scaling, a parameter of interest in these cases is

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

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 $\hat{D}$:

$$\hat{D}_r := \underset{M \in \mathcal{D}_n(r)}{\arg\min} \|J(\hat{D} - M)J\|_{\mathrm{F}}^2. \tag{9}$$

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

The risk bounds that we derived for $\hat{D}$ can also be translated into that for $\hat{D}_r$. More specifically, we have the following corollary.

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

$$\|J(\hat{D}_r - D)J\|_{\mathrm{F}}^2 \leqslant C\{\min_{M \in \mathcal{D}_n(r)} \|J(D - M)J\|_{\mathrm{F}}^2 + \lambda_n^2 r\}.$$

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

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

4. Computation

It is not difficult to see that the optimization problem that is involved in defining the regularized kernel estimate can be formulated as a second-order cone program (see, for example, Lu *et al.* (2005) and Yuan *et al.* (2007)). This class of optimization problems can be readily solved by using generic solvers such as SDPT3 (Toh *et al.*, 1999; Tutuncu *et al.*, 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, $\hat{D}$ 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 $\hat{D}$.

4.1. Alternating projection

We shall adopt, in particular, an alternating projection algorithm that was introduced by Dykstra (1983). Dykstra's algorithm (Table 1) 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 alternately.

The main idea of Dykstra's algorithm can be illustrated by Fig. 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 Escalante and Raydan (2011) for further details.

Now consider evaluating $\hat{D}$ 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 $\mathscr{P}_{\mathcal{C}_1 \cap \mathcal{C}_2}$ 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 principal submatrix of its Householder transform QAQ where $Q = I - vv^{\mathrm{T}}/n$ and $v = (1, \ldots, 1, 1 + \sqrt{n})^{\mathrm{T}}$. In other words,

Table 1. Dykstra's alternating projection algorithm†

```
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 $x_{k+1}$
```

†$\mathscr{P}_{\mathcal{C}_1}$ and $\mathscr{P}_{\mathcal{C}_2}$ are the projections onto $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively.

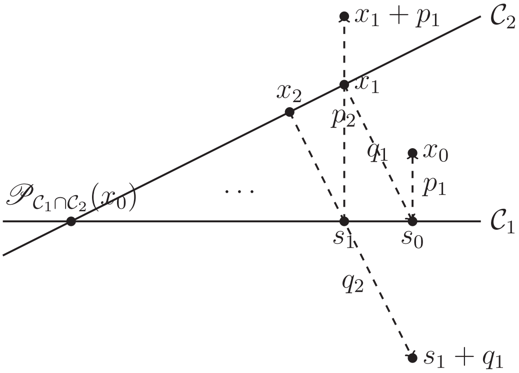


Fig. 3. Illustration of the alternating projection algorithm

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

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

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

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

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 $\Omega \subset \{(i, j) : 1 \leqslant i < j \leqslant n\}$ is the set of entries observed in X. As the complete-data case, we proceed to estimate D by $\hat{D}^{\Omega} = \mathscr{T}(\hat{K}^{\Omega})$ where

$$\hat{K}^{\Omega} = \underset{M \succeq 0}{\arg\min}[\sum_{(i,j)\in\Omega} \{x_{ij} - \langle M, (e_i - e_j)(e_i - e_j)^{\mathrm{T}}\rangle\}^2 + \lambda_n \mathrm{tr}(M)].$$

Here we use the superscript Ω to signify the dependence on the set Ω of the observed entries. Unlike the case without missing data, $\hat{D}^{\Omega}$ in general cannot 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 by 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 (Table 2).

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 method proposed depends on the choice of the tuning parameter. In some cases, we want to embed data into a Euclidean space of a fixed dimensionality. For example, the atoms of a protein must live in a three-dimensional space. For this, we can experiment with different values of the tuning parameter and use the value corresponding to the desired embedding dimension. Our experience suggests that this strategy works fairly well

Table 2. EM algorithm to handle missing data

```
Data: $X_\Omega = (x_{ij})_{(i,j)\in\Omega}$, $\eta_n \geqslant 0$
Result: $\hat{D}^\Omega$
Initialization: initialize $x_{ij}$ for $i<j$ and $(i,j)\notin\Omega$, and let $X = X^{\mathrm{T}} = (x_{ij})_{1\leqslant i,j\leqslant n}$
    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
$\hat{D}^\Omega \leftarrow D^{(k+1)}$;
return $\hat{D}^\Omega$
```

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 by using the algorithm in Table 2.

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

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

In particular, the choice of $T=5$ or $T=10$ is often advocated in practice (see, for example, Hastie *et al.* (2009)). For each $t=1,\ldots,T$, we can then apply the algorithm given in Table 2 to compute the distance shrinkage estimate with a given tuning parameter η_n based on partial observations:

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

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

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

The same procedure can be repeated for a sequence of different values of η_n, and the sequence that is 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 on the basis of all observations to yield the final estimate.

5. Numerical examples

To illustrate the practical merits of the methods proposed 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 (human immunodeficiency virus (HIV) type 1 virus protein U) protein sequences and their relationship to

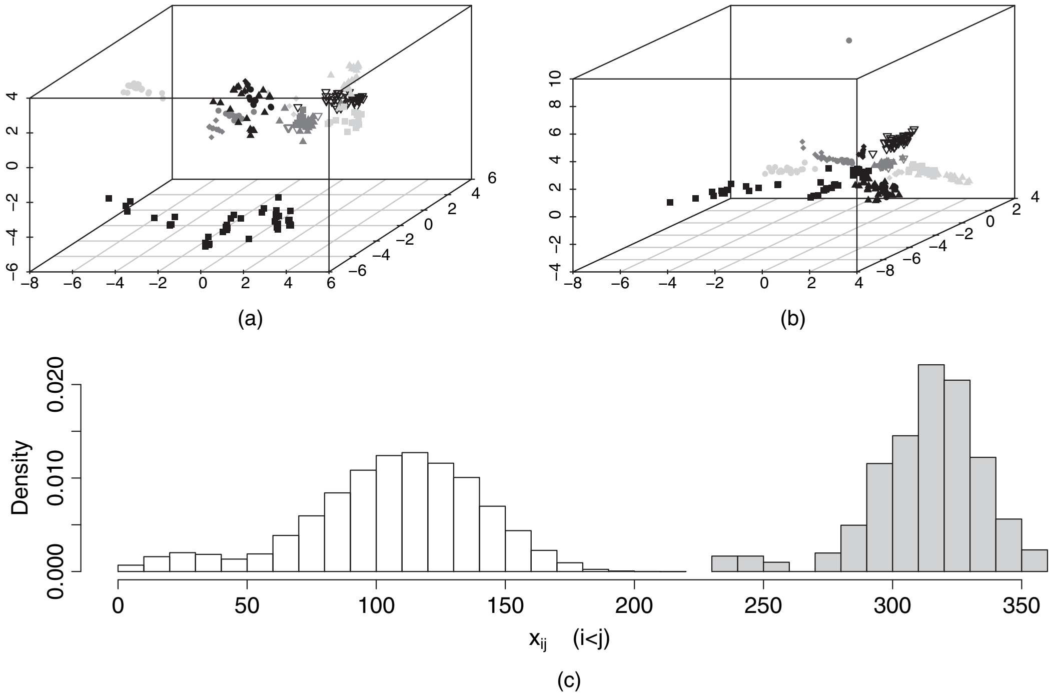


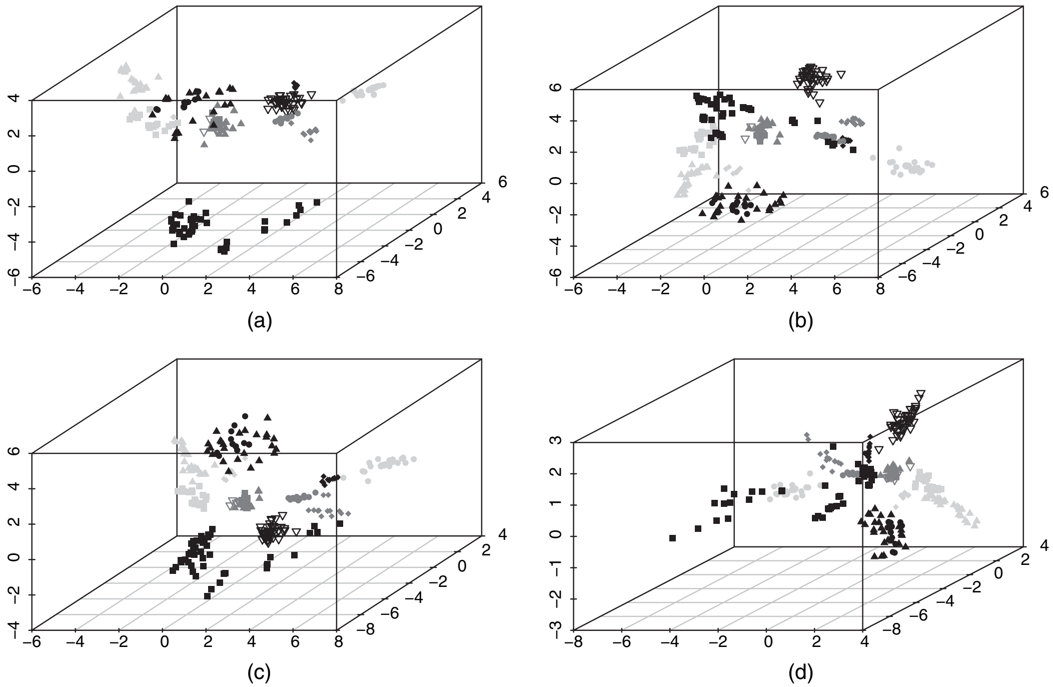
Fig. 4. Three-dimensional embedding for 304 amino acid sequences: (a) classical multi-dimensional scaling embedding; (b) distance shrinkage embedding; (c) pairwise dissimilarity scores (■, scores between the outlying sequence and the other sequences)

preservation of tetherin and CD4 cell counteractivities (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 type 1 infected individuals. Following standard Multicenter Cohort Study criteria (Thio *et al.*, 2002), five of these individuals can be classified as long-term non-progressors, five as rapid progressors and four as normal progressors, according to how long the progression from seroconversion to acquired immune deficiency syndrome takes. A total of 304 unique amino acid sequences were obtained from this study.

We first performed pairwise alignment between these amino acid sequences by using various blocks substitution matrices. The results by using different substitution matrices are fairly similar, and, to fix ideas, we shall report here analysis based on the blocks substitution matrix 62. These pairwise similarity scores $\{s_{ij} : 1 \leqslant i \leqslant j \leqslant n\}$ are converted into dissimilarity scores:

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

As mentioned earlier, $X = (x_{ij})_{1 \leqslant i,j \leqslant n}$ is not a Euclidean distance matrix. For this, we first applied classical multi-dimensional scaling to X. The three-dimensional embedding is given in Fig. 4(a). The amino acid sequences derived from the same individuals are represented by the same symbol and colour. Different colours correspond to the three different classes of disease progression: long-term non-progressors are represented in black, normal in dark grey and rapid progressors in light grey. For comparison, we also computed $\hat{D}$ with various choices of the tuning parameters. Similarly to the observations that were made by Lu *et al.* (2005), the corresponding embeddings are qualitatively similar for a wide range of choices of λ_n. A typical choice is given



(a)

(b)

(c)

(d)

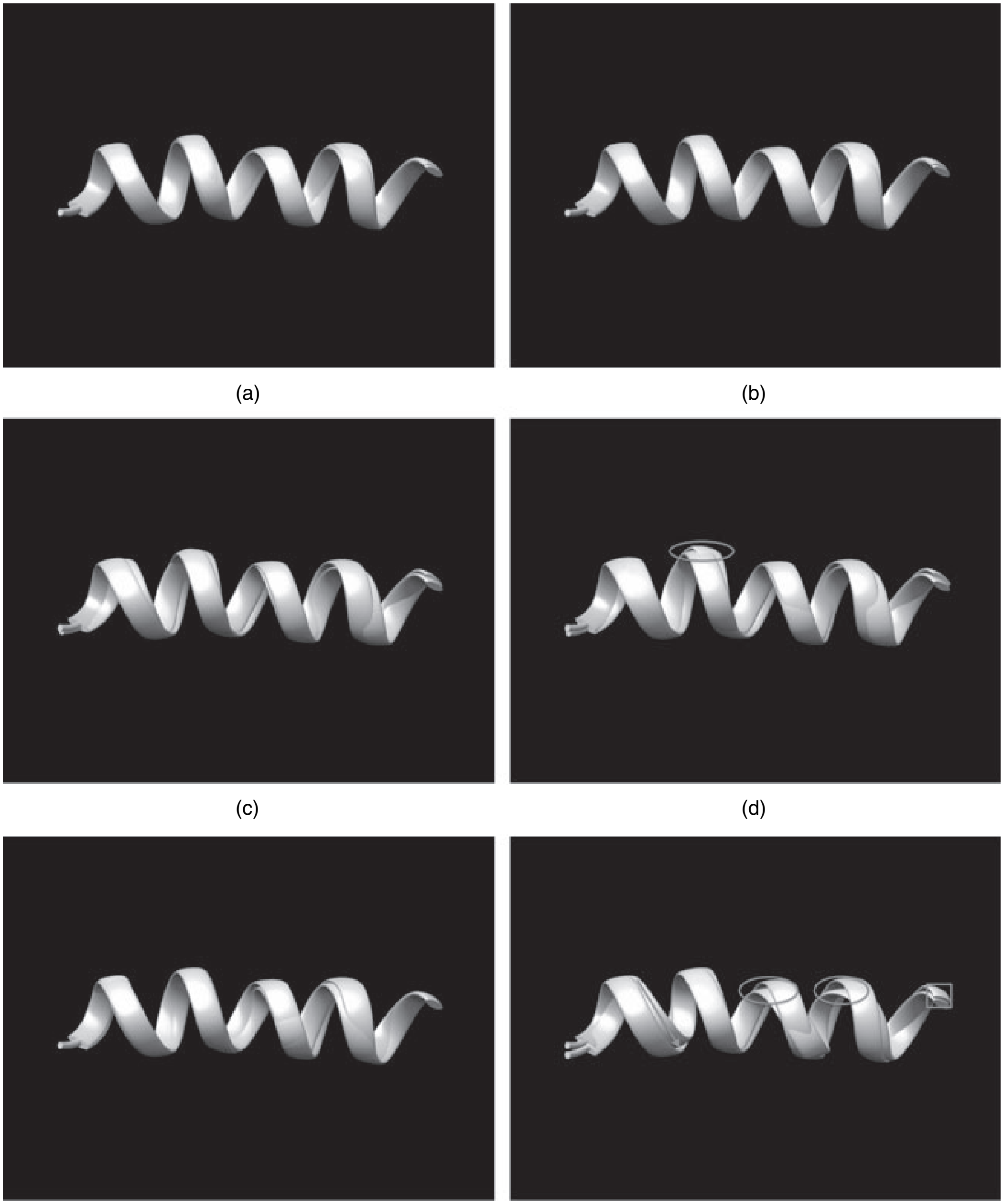
Fig. 5. Euclidean embedding of 303 amino acid sequences via distance shrinkage (the outlying sequence was removed from the original data): (a) $\lambda_n = 4000$; (b) $\lambda_n = 8000$; (c) $\lambda_n = 12000$; (d) $\lambda_n = 16000$

in Fig. 4(b). It is clear that both embeddings share many similarities. For example, sequences that are derived from the same individual are more similar as they tend to cluster together. The key difference, however, is that the embedding corresponding to $\hat{D}$ 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, is given in Fig. 4(c).

Given these observations, we now consider the analysis with the outlying sequence removed. To gain insight, we consider different choices of λ_n to inspect visually the Euclidean embeddings

Table 3. Kruskal's stress for 1PJE data with measurement error

Signal-to-noise ratio	*Method*	*Mean*	*Standard error*
High	Distance shrinkage	0.010	2.0×10^{-4}
	Classical multi-dimensional scaling	0.078	9.3×10^{-4}
Medium	Distance shrinkage	0.024	4.8×10^{-4}
	Classical multi-dimensional scaling	0.185	2.5×10^{-3}
Low	Distance shrinkage	0.035	8.4×10^{-4}
	Classical multi-dimensional scaling	0.301	3.9×10^{-3}



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 6. Ribbon plots of 1PJE protein back structure (■, true structure; ▪, structure corresponding to the estimated Euclidean distance matrix; ⬭, particular regions where the distance shrinkage shows visible improvement): (a) distance shrinkage, high signal-to-noise ratio; (b) classical multi-dimensional scaling, high signal-to-noise ratio; (c) distance shrinkage, medium signal-to-noise ratio; (d) classical multi-dimensional scaling, medium signal-to-noise ratio; (e) distance shrinkage, low signal-to-noise ratio; (f) classical multi-dimensional scaling, low signal-to-noise ratio

Table 4. Kruskal's stress for 2K7Y data with measurement error

Signal-to-noise ratio	*Method*	*Mean*	*Standard error*
High	Distance shrinkage	1.66×10^{-4}	2.70×10^{-7}
	Classical multi-dimensional scaling	3.2×10^{-3}	4.84×10^{-6}
Medium	Distance shrinkage	8.32×10^{-4}	1.48×10^{-6}
	Classical multi-dimensional scaling	1.61×10^{-2}	2.45×10^{-5}
Low	Distance shrinkage	1.7×10^{-3}	3.05×10^{-6}
	Classical multi-dimensional scaling	3.22×10^{-2}	5.28×10^{-5}

given by the proposed distance shrinkage. The embeddings that are given in Fig. 5 correspond to $\lambda_n = 4000$, 8000, 12000, 16000. These embedding are qualitatively similar.

5.2. Simulated examples

To compare the proposed distance shrinkage approach with classical multi-dimensional scaling further, we carried out several sets of simulation studies. For illustration, we took the set-up of the molecular conformation problem that was 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 three-dimensional structure of the channel-forming transmembrane domain of Vpu protein from HIV type 1 mentioned before. The data that were obtained from the protein databank (symbol 1PJE) contain the three-dimensional co-ordinates of a total of $n = 91$ atoms. The exact Euclidean distance matrix D was then calculated from these co-ordinates. We note that in this case the embedding dimension is known to be 3. We generated observations x_{ij} by adding measurement error $\varepsilon_{ij} \sim N(0, \sigma^2)$ for $1 \leqslant i < j \leqslant n$. We considered three different values of σ^2: 0.05, 0.25 and 0.5, representing relatively high, medium and low signal-to-noise ratio respectively. For each value of σ^2, we simulated 100 data sets and computed for each data set the Euclidean distance matrix corresponding to classical multi-dimensional scaling and distance shrinkage. We evaluated the performance of each method by the Kruskal stress defined as $\|\hat{D} - D\|_{\mathrm{F}} / \|D\|_{\mathrm{F}}$. The results are summarized by Table 3.

To appreciate better the difference between the two methods, Fig. 6 gives the ribbon plot of the protein backbone structure corresponding to the true Euclidean distance matrix and the estimated distances from a typical simulation run with different signal-to-noise ratios. It is noteworthy that the improvement of distance shrinkage over classical multi-dimensional scaling becomes more evident with higher level of noise.

Our theoretical analysis suggests better performances for larger numbers of atoms. To illustrate this effect of n further, we repeated the previous experiment for HIV type 1 virus protein U cytoplasmic domain (protein databank symbol 2K7Y) consisting of $n = 671$ atoms. We simulated data in the same fashion as before and the Kruskal stress, based on 100 simulated data sets for each value of σ^2, is reported in Table 4. The performance compares favourably with that for 1PJE.

To demonstrate the efficacy of cross-validation as a tuning method, we give in Fig. 7 the true Kruskal stress as a function of the tuning parameter λ along with the fivefold cross-validation scores for a typical simulated data set 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

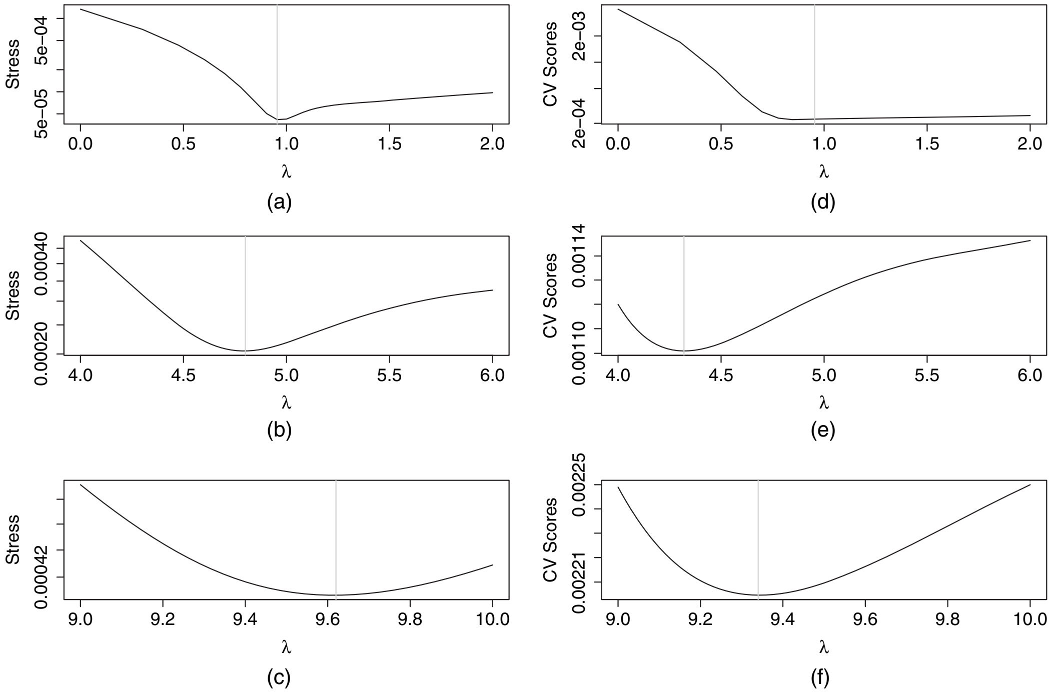


Fig. 7. Comparison of (a)–(c) Kruskal stress and (d)–(f) cross-validation scores for simulated 2K7Y protein data (|, minimizing tuning parameter): (a), (d) high signal-to-noise ratio; (b), (e) medium signal-to-noise ratio; (c), (f) low signal-to-noise ratio

Table 5. Effect of missing data

Protein databank identification	*Number of atoms*	*Kruskal's stress*		
		50% missing	*25% missing*	*10% missing*
1PTQ	402	0.57	0.35	0.18
1HOE	558	0.56	0.33	0.15
1PHT	811	0.56	0.34	0.17
1AX8	1003	0.57	0.36	0.18

the tuning parameter. It is clear from these plots that the tuning parameter that is selected by cross-validation is fairly close to the optimal choice that minimizes the true Kruskal stress.

In the next set of simulations, we assess the effect of missing data for the distance shrinkage estimate proposed. Similarly to before, we take the three-dimensional co-ordinates data from the protein databank for five different proteins with different numbers of atoms. Pairwise distances were first computed for each of the proteins. To mimic the typical nuclear magnetic resonance 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 with a range

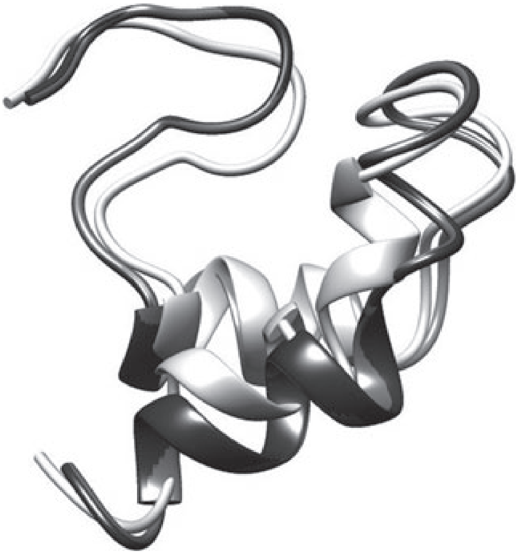


Fig. 8. Ribbon plot of 2K7Y protein back structure: ■, true structure; ░, structure corresponding to the classical multi-dimensional scaling; ▒, structure corresponding to distance shrinkage

of tuning parameter choices and the performance is fairly similar. The results are summarized in Table 5. As expected, the method performs better as the amount of missing data reduces. The distance shrinkage estimate works reasonably well even with 10% of missing data.

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

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

so that both the mean and the 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 multi-dimensional scaling and distance shrinkage to estimate the true distance matrix and reconstruct the three-dimensional folding structure. The result from a typical simulated data set is given in Fig. 8.

Acknowledgements

The research of Luwan Zhang was supported in part by National Science Foundation career award DMS-1321692 and Focused Research Groups award DMS-1265202. The research of Grace Wahba was supported in part by National Institutes of Health grants EY09946 and 1U54AI117924-01 and National Science Foundation grant DMS1308877. The research of Ming Yuan was supported in part by National Science Foundation career award DMS-1321692 and Focused Research Groups award DMS-1265202 and National Institutes of Health grant 1U54AI117924-01.

References

Chen, L. and Buja, A. (2009) Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *J. Am. Statist. Ass.*, **104**, 209–219.

Chen, L. and Buja, A. (2013) Stress functions for nonlinear dimension reduction, proximity analysis, and graph drawing. *J. Mach. Learn. Res.*, **14**, 1145–1173.

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

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

Dykstra, R. (1983) An algorithm for restricted least squares regression. *J. Am. Statist. Ass.*, **78**, 837–842.

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

Glunt, W., Hayden, T., Hong, S. and Wells, J. (1990) An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM J. Matr. Anal. Appl.*, **11**, 589–600.

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

Hayden, T. L. and Wells, J. (1988) Approximation by matrices positive semidefinite on a subspace. *Lin. Alg. Appl.*, **109**, 115–130.

Lu, F., Keles, S., Wright, S. and Wahba, G. (2005) Framework for kernel regularization with application to protein clustering. *Proc. Natn. Acad. Sci. USA*, **102**, 12332–12337.

Lu, Z., Monteiro, R. and Yuan, M. (2012) Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. *Math. Program.*, **131**, 163–194.

Negahban, S. and Wainwright, M. (2011) Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, **39**, 1069–1097.

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 Path.*, **10**, no. 1, article e1003895.

Pouzet, M. (1979) Note sur le probléme de Ulam. *J. Combin. Theor.* B, **27**, 231–236.

Rohde, A. and Tsybakov, A. (2011) Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, **39**, 887–930.

Roy, A. (2010) Minimal Euclidean representations of graphs. *Discr. Math.*, **310**, 727–733.

Schölkopf, B. and Smola, A. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neurl Computn*, **10**, 1299–1319.

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

Schölkopf, B., Smola, A. and Müller, K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neurl Computn*, **10**, 1299–1319.

Schönberg, 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". *Ann. Math.*, **38**, 724–732.

Sinai, Y. and Soshnikov, A. (1998) A refinement of Wigners semi-circle law in a neighborhood of the spectrum edge for random symmetric matrices. *Functnl Anal. Appl.*, **32**, 114–131.

Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007) Measuring and testing independence by correlation of distances. *Ann. Statist.*, **35**, 2769–2794.

Tenenbaum, J., De Silva, V. and Langford, J. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Thio, C. L., Seaberg, E. C., Skolasky, Jr, R., Phair, J., Visscher, B., Muñoz, A., Thomas, D. L. and the Multicenter AIDS Cohort Study (2002) HIV-1, hepatitis B virus, and risk of liver-related mortality in the Multicenter Cohort Study (MACS). *Lancet*, **360**, 1921–1926.

Toh, K. C., Todd, M. J. and Tutuncu, R. H. (1999) SDPT3—a Matlab software package for semidefinite programming. *Optimizn Meth. Softwr.*, **11**, 545–581.

Tutuncu, R. H., Toh, K. C. and Todd, M. J. (2003) Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Program.* B, **95**, 189–217.

Venna, J. and Kaski, S. (2006) Local multidimensional scaling. *Neurl Netwrks*, **19**, 889–899.

Weinberger, K. Q. (2007) Metric learning with convex optimization.*PhD Thesis*. University of Pennsylvania, Philadelphia.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. New York: Wiley.

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

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007) Dimension reduction and coefficient estimation in multivariate linear regression. *J. R. Statist. Soc.* B, **69**, 329–346.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to "Distance shrinkage and Euclidean embedding via regularized kernel estimation"'.