Journal of Machine Learning Research 21 (2020) 1-37 Submitted 10/18; Published 1/20

Perturbation Bounds for Procrustes, Classical Scaling, and Trilateration, with Applications to Manifold Learning

Ery Arias-Castro EARIASCA@UCSD.EDU
Department of Mathematics
University of California
San Diego, CA 92093, USA

Adel Javanmard AJAVANMA@USC.EDU
Department of Data Sciences and Operations
Marshall School of Business
University of Southern California
Los Angeles, CA 90089 , USA

Bruno Pelletier BRUNO.PELLETIER@MATH.CNRS.FR
Département de Mathématiques
IRMAR - UMR CNRS 6625
Université Rennes II

Editor: Miguel Carreira-Perpinan

Abstract

One of the common tasks in unsupervised learning is dimensionality reduction, where the goal is to find meaningful low-dimensional structures hidden in high-dimensional data. Sometimes referred to as manifold learning, this problem is closely related to the problem of localization, which aims at embedding a weighted graph into a low-dimensional Euclidean space. Several methods have been proposed for localization, and also manifold learning. Nonetheless, the robustness property of most of them is little understood. In this paper, we obtain perturbation bounds for classical scaling and trilateration, which are then applied to derive performance bounds for Isomap, Landmark Isomap, and Maximum Variance Unfolding. A new perturbation bound for procrustes analysis plays a key role.

1. Introduction

Multidimensional scaling (MDS) can be defined as the task of embedding an itemset as points in a (typically) Euclidean space based on some dissimilarity information between the items in the set. Since its inception, dating back to the early 1950's if not earlier (Young, 2013), MDS has been one of the main tasks in the general area of multivariate analysis, a.k.a., unsupervised learning.

One of the main methods for MDS is called classical scaling, which consists in first double-centering the dissimilarity matrix and then performing an eigen-decomposition of the obtained matrix. This is arguably still the most popular variant, even today, decades after its introduction at the dawn of this literature. (For this reason, this method is often referred to as MDS, and we will do the same on occasion.) Despite its wide use, its perturbative properties remain little understood. The major contribution on this question dates back

to the late 1970's with the work of Sibson (1979), who performs a sensitivity analysis that resulted in a Taylor development for the classical scaling to the first nontrivial order. Going beyond Sibson (1979)'s work, our first contribution is to derive a bonafide perturbation bound for classical scaling (Theorem 1).

Classical scaling amounts to performing an eigen-decomposition of the dissimilarity matrix after double-centering. Only the top d eigenvectors are needed if an embedding in dimension d is desired. Using iterative methods such as the Lanczos algorithm, classical scaling can be implemented with a complexity of $O(dn^2)$, where n is the number of items (and therefore also the dimension of the dissimilarity matrix). In applications, particularly if the intent is visualization, the embedding dimension d tends to be small. Even then, the resulting complexity is quadratic in the number of items n to be embedded. There has been some effort in bringing this down to a complexity that is linear in the number of items. The main proposals (Faloutsos and Lin, 1995; Wang et al., 1999; de Silva and Tenenbaum, 2004) are discussed by Platt (2005), who explains that all these methods use a Nyström approximation. The procedure proposed by de Silva and Tenenbaum (2004), which they called landmark MDS (LMDS) and which according to Platt (2005) is the best performing methods among these three, works by selecting a small number of items, perhaps uniformly at random from the itemset, and embedding them via classical scaling. These items are used as landmark points to enable the embedding of the remaining items. The second phase consists in performing trilateration, which aims at computing the location of a point based on its distances to known (landmark) points. Note that this task is closely related to, but distinct, from triangulation, which is based on angles instead. If ℓ items are chosen as landmarks in the first step (out of n items in total), then the procedure has complexity $O(d\ell^2 + d\ell n)$. Since ℓ can in principle be chosen on the order of d, and $d \leq n$ always, the complexity is effectively $O(d^2 n)$, which is linear in the number of items. A good understanding of the robustness properties of LMDS necessitates a good understanding of the robustness properties of not only classical scaling (used to embed the landmark items), but also of trilateration (used to embed the remaining items). Our second contribution is a perturbation bound for trilateration (Theorem 2). There are several closely related method for trilateration, and we study on the method proposed by de Silva and Tenenbaum (2004), which is rather natural. We refer to this method simply as trilateration in the remaining of the paper.

de Silva and Tenenbaum (2004) build on the pioneering work of Sibson (1979) to derive a sensitivity analysis of classical scaling. They also derive a sensitivity analysis for their trilateration method following similar lines. In the present work, we instead obtain bonafide perturbation bounds, for procrustes analysis (Section 2), for classical scaling (Section 3), and for the same trilateration method (Section 4). In particular, our perturbation bounds for procrustes analysis and classical scaling appear to be new, which may be surprising as these methods have been in wide use for decades. (The main reason for deriving a perturbation bound for procrustes analysis is its use in deriving a perturbation bound for classical scaling, which was our main interest.) These results are applied in Section 5 to Isomap, Landmark Isomap, and also Maximum Variance Unfolding (MVU). These may be the first performance bounds of any algorithm for manifold learning in its 'isometric embedding' variant, even as various consistency results have been established for Isomap (Zha and Zhang, 2007), MVU (Arias-Castro and Pelletier, 2013), and a number of other methods (Donoho and

Grimes, 2003; Ye and Zhi, 2015; Giné and Koltchinskii, 2006; Smith et al., 2008; Belkin and Niyogi, 2008; von Luxburg et al., 2008; Singer, 2006; Hein et al., 2005; Coifman and Lafon, 2006). (As discussed in (Goldberg et al., 2008), Local Linear Embedding, Laplacian Eigenmaps, Hessian Eigenmaps, and Local Tangent Space Alignment, all require some form of normalization which make them inconsistent for the problem of isometric embedding.) In Section 7 we discuss the question of optimality in manifold learning and also the choice of landmarks. The main proofs are gathered in Section 8.

2. A perturbation bound for procrustes

The orthogonal procrustes problem is that of aligning two point sets (of same cardinality) using an orthogonal transformation. In formula, given two point sets, $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$ in $\mathbb{R}^d$, the task consists in solving

$$\min_{Q \in \mathcal{O}(d)} \sum_{i=1}^{m} \|y_i - Q x_i\|^2, \tag{1}$$

where $\mathcal{O}(d)$ denotes the orthogonal group of $\mathbb{R}^d$. (Here and elsewhere, when applied to a vector, $\|\cdot\|$ will denote the Euclidean norm.)

In matrix form, the problem can be posed as follows. Given matrices X and Y in $\mathbb{R}^{m \times d}$, solve

$$\min_{Q \in \mathcal{O}(d)} \|Y - XQ\|_2, \tag{2}$$

where $\|\cdot\|_2$ denotes the Frobenius norm (in the appropriate space of matrices). As stated, the problem is solved by choosing $Q = UV^\top$, where U and V are d-by-d orthogonal matrices obtained by a singular value decomposition of $X^\top Y = UDV^\top$, where D is the diagonal matrix with the singular values on its diagonal (Seber, 2004, Sec 5.6). Algorithm 1 describes the procedure.

Algorithm 1 Procrustes (Frobenius norm)

Input: point sets $x_1, \ldots, x_m$ and $y_1, \ldots, y_m$ in $\mathbb{R}^d$
Output: an orthogonal transformation Q of $\mathbb{R}^d$

1: store the point sets in $X = [x_1^\top \cdots x_m^\top]$ and $Y = [y_1^\top \cdots y_m^\top]$
2: compute $X^\top Y$ and its singular value decomposition $UDV^\top$
Return: the matrix $Q = UV^\top$

In matrix form, the problem can be easily stated using any other matrix norm in place of the Frobenius norm. There is no closed-form solution in general, even for the operator norm (as far as we know), although some computational strategies have been proposed for solving the problem numerically (Watson, 1994). In what follows, we consider an arbitrary Schatten norm. For a matrix $A \in \mathbb{R}^{m \times n}$, let $\|A\|_p$ denote the Schatten p-norm, where $p \in [1, \infty]$ is assumed fixed:

$$\|A\|_p \equiv \Big(\sum_{i \geq 1} \nu_i^p(A) \Big)^{1/p}, \tag{3}$$

with $\nu_1(A) \geq \nu_2(A) \geq \ldots \geq 0$ the singular values of A. Note that $\|\cdot\|_2$ coincides with the Frobenius norm. We also define $\|A\|_\infty$ to be the usual operator norm, i.e., the maximum singular value of A. Henceforth, we will also denote the operator norm by $\|\cdot\|$, on occasion. We denote by $A^\ddagger$ the pseudo-inverse of A (see Section 8.1). Henceforth, we also use the notation $a \wedge b = \min(a, b)$ for two numbers a, b.

Our first theorem is a perturbation bound for procrustes, where the distance between two configurations of points X and Y is bounded in terms of the distance between their Gram matrices $XX^\top$ and $YY^\top$.

Theorem 1 *Consider two tall matrices X and Y of same size, with X having full rank, and set $\varepsilon^2 = \|YY^\top - XX^\top\|_p$. Then, we have*

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_p \leq \begin{cases} \|X^\ddagger\|\varepsilon^2 + \left((1 - \|X^\ddagger\|^2\varepsilon^2)^{-1/2}\|X^\ddagger\|\varepsilon^2\right) \wedge \left(d^{1/2p}\varepsilon\right), & \text{if } \|X^\ddagger\|\varepsilon < 1\,, \\ \|X^\ddagger\|\varepsilon^2 + d^{1/2p}\varepsilon & \text{otherwise.} \end{cases} \tag{4}$$

Consequently, if $\|X^\ddagger\|\varepsilon \leq \frac{1}{\sqrt{2}}$, then

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_p \leq (1 + \sqrt{2})\|X^\ddagger\|\varepsilon^2. \tag{5}$$

The proof is in Section 8.2. Interestingly, to establish the upper bound we use an orthogonal matrix constructed from the singular value decomposition of $X^\ddagger Y$. This is true regardless of p, which may be surprising since a solution for the Frobenius norm (corresponding to the case where $p = 2$) is based on a singular value decomposition of $X^\top Y$ instead.

Also, let us stress that ε in the theorem statement, by definition, depends on the choice of p-norm.

Example 1 *(***Orthonormal matrices)** *The case where X and Y are orthonormal and of the same size is particularly simple, at least when $p = 2$ or $p = \infty$, based on what is already known in the literature. Indeed, from (Stewart and Sun, 1990, Sec II.4) we find that, in that case,*

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_p = \|2\sin(\tfrac{1}{2}\theta(X, Y))\|_p, \tag{6}$$

where $\theta(X, Y)$ is the diagonal matrix made of the principal angles between the subspaces defined by X and Y, and for a matrix A, $\sin(A)$ is understood entrywise. In addition,

$$\varepsilon^2 = \|YY^\top - XX^\top\|_p = \|\sin\theta(X, Y)\|_p. \tag{7}$$

Using the elementary inequality $\sqrt{2}\sin(\alpha/2) \leq \sin(\alpha) \leq 2\sin(\alpha/2)$, valid for $\alpha \in [0, \pi/2]$, we get

$$\varepsilon^2 \leq \min_{Q\in\mathcal{O}} \|Y - XQ\|_p \leq \sqrt{2}\varepsilon^2. \tag{8}$$

Note that, in this case, $\|X\| = \|X^\ddagger\| = 1$, and our bound (5) gives the upper bound $(1 + \sqrt{2})\varepsilon^2$, which is tight up to a factor of $1 + \frac{1}{\sqrt{2}}$.

Example 2 *The derived perturbation bound* (5) *includes the pseudo-inverse of the configuration,* $\|X^\ddagger\|$*. Nonetheless, the example of orthogonal matrices does not capture this factor because* $\|X^\ddagger\| = 1$ *in that case. To build further insight on our result in Theorem 1, we consider another example where* X *and* Y *share the same singular vectors. Namely* $X = U\Lambda V^\top$ *and* $Y = U\Theta V^\top$*, with* $U \in \mathbb{R}^{m\times d}$*,* $V \in \mathbb{R}^{d\times d}$ *orthonormal matrices, and* $\Lambda = \operatorname{diag}(\{\lambda_i\})_{i=1}^d$ *and* $\Theta = \operatorname{diag}(\{\theta_i\})_{i=1}^d$*. Consider the case of* $p = 2$*, and let* $X^\top Y = V(\Lambda\Theta)V^\top$ *be a singular value decomposition. Then by Algorithm 1, the optimal rotation is given by* $Q = I$*. We therefore have*

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_2 = \Big[\sum_{i\in[n]} (\theta_i - \lambda_i)^2\Big]^{1/2} = \Big[\sum_{i\in[n]} \Big(\frac{\theta_i^2 - \lambda_i^2}{\theta_i + \lambda_i}\Big)^2\Big]^{1/2} \tag{9}$$

$$\leq \frac{1}{(\min_{i\in[n]} |\lambda_i|)}\Big[\sum_{i\in[n]} \Big(\theta_i^2 - \lambda_i^2\Big)^2\Big]^{1/2} = \frac{1}{(\min_{i\in[n]} |\lambda_i|)}\|YY^\top - XX^\top\|_2$$

$$= \|X^\ddagger\|\varepsilon^2\,. \tag{10}$$

Let us stress that the above derivation applies only to this example, but it showcases the relevance of $\|X^\ddagger\|$ *in the bound.*

We next develop a lower bound for the following specific case. Let $D = \operatorname{diag}(1, 1, \ldots, \delta)$ *for arbitrary but fixed* $\delta \in [0, 1]$ *and let* $\Theta = \operatorname{diag}(1, 1, \ldots, \sqrt{\delta^2 + \varepsilon^2})$*. Then,* $\|X^\ddagger\| = 1/\delta$ *and* $\|XX^\top - YY^\top\|_2 = \varepsilon^2$*. By* (9) *we have*

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_2 = \Big[\sum_{i\in[n]} (\theta_i - \lambda_i)^2\Big]^{1/2} = \sqrt{\delta^2 + \varepsilon^2} - \delta = \delta\left(\sqrt{1 + \tfrac{\varepsilon^2}{\delta^2}} - 1\right). \tag{11}$$

Also, from the condition $\|X^\ddagger\|\varepsilon \leq \frac{1}{\sqrt{2}}$ *we have* $\frac{\varepsilon}{\delta} < \frac{1}{\sqrt{2}}$*. Using* $\sqrt{1 + x^2} - 1 \geq (\sqrt{6} - 2)x^2$*, which holds for* $x < \frac{1}{\sqrt{2}}$ *and substituting for* $\delta = 1/\|X^\ddagger\|$*, we obtain*

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_2 \geq (\sqrt{6} - 2)\|X^\ddagger\|\varepsilon^2 \tag{12}$$

From (10) *and* (12)*, we observe that the* $\|X^\ddagger\|$ *term appears both in the upper and the lower bounds of the procrustes error, which confirms its relevance.*

Remark 1 *We emphasize that the general bound in*(4) *does not require any restriction on* ε*. However, as it turns out, the result in* (5) *would be already enough for our purposes in the next sections and deriving our results in the context of manifold learning. Regarding the procrustes error bound in Theorem 1, we do conjecture that there is a smooth transition between a bound in* ε^2 *and a bound in* ε *as* $\|X^\ddagger\|$ *increases to infinity (and therefore* X *degenerates to a singular matrix). For instance, in Example 2, when* $\delta = \|X^\ddagger\|^{-1} \to 0$ *faster than* ε*, the lower bound* (11) *scales linearly in* ε*.*

It is worth noting that other types of perturbation analysis have been carried out for the procrustes problem. For example (Söderkvist, 1993) considers the procrustes problem over the class of rotation matrices, a subset of orthogonal matrices, and study how its solution

(optimal rotation) would be perturbed if both configurations were perturbed. In (Zha and Zhang, 2009), the authors study the perturbation of the null space of a similarity matrix from manifold learning, using the standard perturbation theory for invariant subspaces (Stewart and Sun, 1990).

3. A perturbation bound for classical scaling

In multidimensional scaling, we are given a matrix, $\Delta = (\Delta_{ij}) \in \mathbb{R}^{m\times m}$, storing the dissimilarities between a set of m items (which will remain abstract in this paper). A square matrix Δ is called dissimilarity matrix if it is symmetric, $\Delta_{ii} = 0$, and $\Delta_{ij} > 0$, for $i \neq j$. (Δ_{ij} gives the level of dissimilarity between items $i, j \in [m]$.) Given a positive integer d, we seek a configuration, meaning a set of points, $y_1, \cdots, y_m \in \mathbb{R}^d$, such that $\|y_i - y_j\|^2$ is close to Δ_{ij} over all $i, j \in [m]$. The itemset $[m]$ is thus embedded as $y_1, \ldots, y_m$ in the d-dimensional Euclidean space $\mathbb{R}^d$.

Algorithm 2 describes classical scaling, the first practical and the most prominent method for solving this problem. The method is widely attributed to Torgerson (1958) and Gower (Gower, 1966) and it is also known under the names Torgerson scaling and Torgerson-Gower scaling.

Algorithm 2 Classical Scaling

Input: dissimilarity matrix $\Delta \in \mathbb{R}^{m\times m}$, embedding dimension d
Output: set of points $y_1, \ldots, y_m \in \mathbb{R}^d$

1: compute the matrix $\Delta^c = -\frac{1}{2}H\Delta H$
2: let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ be the eigenvalues of Δ^c, with corresponding eigenvectors $u_1, \ldots, u_m$
3: compute $Y \in \mathbb{R}^{m\times d}$ as $Y = [\sqrt{\lambda_{1,+}}\, u_1, \ldots, \sqrt{\lambda_{d,+}}\, u_d]$
Return: the row vectors $y_1, \ldots, y_m$ of Y

In the description, $H = I - J/m$ is the centering matrix in dimension m, where I is the identity matrix and J is the matrix of ones. Further, we use the notation $a_+ = \max(a, 0)$ for a scalar a. The basic idea of classical scaling is to assume that the dissimilarities are Euclidean distances and then find coordinates that explain them.

For a general dissimilarity matrix Δ, the doubly centered matrix Δ^c may have negative eigenvalues and that is why in the construction of Y, we use the positive part of the eigenvalues. However, if Δ is an Euclidean dissimilarity matrix, namely $\Delta_{ij} = \|x_i - x_j\|^2$ for a set points $\{x_1, \ldots, x_m\}$ in some ambient Euclidean space, then Δ^c is a positive semi-definite matrix. This follows from the following identity relating a configuration X with the corresponding squared distance matrix Δ:

$$-\frac{1}{2}H\Delta H = HXX^TH\,. \tag{13}$$

Consider the situation where the dissimilarity matrix Δ is exactly realizable in dimension d, meaning that there is a set of points $y_1, \ldots, y_m$ such that $\Delta_{ij} = \|y_i - y_j\|^2$. It is worth noting that, in that case, the set of points that perfectly embed Δ in dimension d are rigid

transformations of each other. It is well-known that classical scaling provides such a set of points which happens to be centered at the origin (see Eq. 13).

We perform a perturbation analysis of classical scaling, by studying the effect of perturbing the dissimilarities on the embedding that the algorithm returns. This sort of analysis helps quantify the degree of robustness of a method to noise, and is particularly important in applications where the dissimilarities are observed with some degree of inaccuracy, which is the case in the context of manifold learning (Section 5.1).

Definition 1 *We say that $\Delta \in \mathbb{R}^{m\times m}$ is a d-Euclidean dissimilarity matrix if there exists a set of points $\{x_1, \ldots, x_m\} \in \mathbb{R}^d$ such that $\Delta_{ij} = \|x_i - x_j\|^2$.*

Recall that $\mathcal{O}$ denotes the orthogonal group of matrices in the appropriate Euclidean space (which will be clear from context).

Corollary 1 *Let $\Lambda, \Delta \in \mathbb{R}^{m\times m}$ denote two d-Euclidean dissimilarity matrices, with Δ corresponding to a centered and full rank configuration $Y \in \mathbb{R}^{m\times d}$. Set $\varepsilon^2 = \frac{1}{2}\|H(\Lambda - \Delta)H\|_p$. If it holds that $\|Y^\ddagger\|\varepsilon \leq \frac{1}{\sqrt{2}}$, then classical scaling with input dissimilarity matrix Λ and dimension d returns a centered configuration $Z \in \mathbb{R}^{m\times d}$ satisfying*

$$\min_{Q\in\mathcal{O}} \|Z - YQ\|_p \leq (1+\sqrt{2})\|Y^\ddagger\|\varepsilon^2. \tag{14}$$

We note that $\varepsilon^2 \leq \frac{1}{2}d^{2/p}\|\Lambda - \Delta\|_p$, after using the fact that $\|H\|_p = (d-1)^{1/p}$ since H has one zero eigenvalue and $d-1$ eigenvalues equal to one.

Proof We have

$$\|\Lambda^c - \Delta^c\|_p = \tfrac{1}{2}\|H(\Lambda - \Delta)H\|_p = \varepsilon^2. \tag{15}$$

Note that since Δ and Λ are both d-Euclidean dissimilarity matrices, using identity (13), the doubly centered matrices Δ^c and Λ^c are both positive semi-definite and of rank at most d. Indeed, since Y is full rank (rank d) and centered, then (13) implies that Δ^c is of rank d. Therefore, for the underlying configuration Y and the configuration Z, returned by classical scaling, we have $\Delta^c = YY^\top$ and $\Lambda^c = ZZ^\top$. We next simply apply Theorem 1, which we can do since Y has full rank by assumption, to conclude. ■

Remark 2 *The perturbation bound* (14) *is optimal in how it depends on ε. Indeed, suppose without loss of generality that $p = 2$. (All the Schatten norms are equivalent modulo constants that depend on d and p.) Consider a configuration Y with squared distance matrix Δ as in the statement, and define $\Lambda = (1+a)^2\Delta$, with $0 \leq a \leq 1$, as a perturbation of Δ. Then, it is easy to see that classical scaling with input dissimilarity matrix Λ returns $Z = (1+a)Y$. On the one hand, we have (Seber, 2004, Sec 5.6)*

$$\min_{Q\in\mathcal{O}} \|Z - YQ\|_2 = \|Z - Y\|_2 = a\|Y\|_2\,. \tag{16}$$

On the other hand,

$$\varepsilon^2 = \frac{1}{2}\|H(\Lambda - \Delta)H\|_p = \frac{1}{2}((1+a)^2 - 1)\|H\Delta H\|_p = ((1+a)^2 - 1)\|YY^\top\|_2\,. \tag{17}$$

Therefore, the right-hand side in (14) *can be bounded by* $3(1+\sqrt{2})a\|Y^\ddagger\|\|YY^\top\|_2$, *using that* $a \in [0,1]$. *We therefore conclude that the ratio of the left-hand side to the right-hand side in* (14) *is at least*

$$\frac{a\|Y\|_2}{3(1+\sqrt{2})a\|Y^\ddagger\|\|YY^\top\|_2} \geq \frac{1}{3(1+\sqrt{2})}(\|Y\|\|Y^\ddagger\|)^{-1}, \tag{18}$$

using the fact that $\|YY^\top\|_2 \leq \|Y\|\|Y\|_2$. *Therefore, our bound* (14) *is tight up to a multiplicative factor depending on the condition number of the configuration* Y.

Remark 3 *Condition* $\|Y^\ddagger\|\varepsilon \leq \frac{1}{\sqrt{2}}$ *in Corollary 1 is of crucial importance in that without it the dissimilarity matrix* Λ *may have rank less than* d. *In this case, the classical scaling (Algorithm 2) with input* Λ, *returns a configuration* Z *which contains zero columns and hence suffers a large procrustes error.*

We now translate this result in terms of point sets instead of matrices. For a centered point set $y_1, \ldots, y_m \in \mathbb{R}^d$, stored in the matrix $Y = [y_1 \cdots y_m]^\top \in \mathbb{R}^{m\times d}$, define its radius as the largest standard deviation along any direction in space (therefore corresponding to the square root of the top eigenvalue of the covariance matrix). We denote this by $\rho(Y)$ and note that

$$\rho(Y) = \|Y\|/\sqrt{m}. \tag{19}$$

We define its half-width as the smallest standard deviation along any direction in space (therefore corresponding to the square root of the bottom eigenvalue of the covariance matrix). We denote this by $\omega(Y)$ and note that it is strictly positive if and only if the point set $\{y_1, \ldots, y_m\}$ spans the whole space $\mathbb{R}^d$; in other words, the matrix $Y = [y_1 \cdots y_m]^\top \in \mathbb{R}^{m\times d}$ is of rank d. In this case

$$\omega(Y) = \|Y^\ddagger\|^{-1}/\sqrt{m}. \tag{20}$$

It is well-known that the half-width quantifies the best affine approximation to the point set, in the sense that

$$\omega(Y)^2 = \min_{\mathcal{L}} \frac{1}{m}\sum_{i\in[m]} \|y_i - P_{\mathcal{L}} y_i\|^2, \tag{21}$$

where the minimum is over all affine hyperplanes $\mathcal{L}$, and for a subspace $\mathcal{L}$, $P_{\mathcal{L}}$ denotes the orthogonal projection onto $\mathcal{L}$. We note that $\rho(Y)/\omega(Y) = \|Y\|\|Y^\ddagger\|$ is the aspect ratio of the point set.

Corollary 2 *Consider a centered point set* $y_1, \ldots, y_m \in \mathbb{R}^d$ *with radius* ρ, *and with half-width* ω, *and with pairwise dissimilarities* $\delta_{ij} = \|y_i - y_j\|^2$. *Consider another arbitrary set of numbers* $\{\lambda_{ij}\}$, *for* $1 \leq i, j \leq m$ *and set* $\eta^4 = \frac{1}{m^2}\sum_{i,j}(\lambda_{ij} - \delta_{ij})^2$. *If* $\eta/\omega \leq \frac{1}{\sqrt{2}}$, *then classical scaling with input dissimilarities* $\{\lambda_{ij}\}$ *and dimension* d *returns a point set* $z_1 \cdots z_m \in \mathbb{R}^d$ *satisfying*

$$\min_{Q\in\mathcal{O}} \left(\frac{1}{m}\sum_{i\in[m]} \|z_i - Qy_i\|^2\right)^{1/2} \leq \frac{\sqrt{d}(\rho/\omega + 2)}{\omega}\,\eta^2 \leq \frac{3\sqrt{d}\rho\,\eta^2}{\omega^2}. \tag{22}$$

This corollary follows from Theorem 1. We refer to Section 8.4 for its proof.

Remark 4 *In some applications, one might be interested in an approximate embedding, where the goal is to embed a large fraction (but not necessarily all) of the points with high accuracy. Note that bound* (22) *provides a non-trivial bound for this objective. Indeed, for any optimal Q for the left-hand side of* (22)*, and an arbitrary fixed $\delta > 0$, let $N(\delta) \equiv |\{i \in [m] : \|z_i - Qy_i\| > \delta\}|$ be the number of points that are not embedded within accuracy δ. Then,* (22) *implies that*

$$N(\delta)\delta^2 \leq \sum_{i\in[m]} \|z_i - Qy_i\|^2 \leq m\left(\frac{3\sqrt{d}\rho\,\eta^2}{\omega^2}\right)^2,$$

and hence

$$N(\delta) \leq m\left(\frac{3\sqrt{d}\rho\,\eta^2}{\delta\omega^2}\right)^2. \tag{23}$$

4. A perturbation bound for trilateration

The problem of trilateration is that of positioning a point, or set of points, based on its (or their) distances to a set of points, which in this context serve as landmarks. In detail, given a set of landmark points $y_1, \ldots, y_m \in \mathbb{R}^d$ and a set of dissimilarities $\tilde{\delta}_1, \ldots, \tilde{\delta}_m$, the goal is to find $\tilde{y} \in \mathbb{R}^d$ such that $\|\tilde{y} - y_i\|^2$ is close to $\tilde{\delta}_i$ over all $i \in [m]$. Algorithm 3 describes the trilateration method of de Silva and Tenenbaum (2004) simultaneously applied to multiple points to be located. The procedure is shown in (de Silva and Tenenbaum, 2004) to recover the position of points $\tilde{y}_1, \ldots, \tilde{y}_n$ exactly, when it is given the squared distances $\tilde{\delta}_{ij} = \|\tilde{y}_i - y_j\|^2$ as input and the landmark point set $\{y_1, \ldots, y_m\}$ spans $\mathbb{R}^d$. We provide a more succinct proof of this in the Section A.1.

Algorithm 3 Trilateration

Input: centered point set $y_1, \ldots, y_m \in \mathbb{R}^d$, dissimilarities $\tilde{\Delta} = (\tilde{\delta}_{ij}) \in \mathbb{R}^{n\times m}$
Output: points $\tilde{y}_1, \ldots, \tilde{y}_n \in \mathbb{R}^d$

1: compute $\bar{a} = \frac{1}{m}\sum_{i=1}^m a_i$, where $a_i = (\|\tilde{y}_i - y_1\|^2, \ldots, \|\tilde{y}_i - y_m\|^2)$
2: compute the pseudo-inverse $Y^\ddagger$ of $Y = [y_1 \cdots y_m]^\top$
3: compute $\tilde{Y}^\top = \frac{1}{2}Y^\ddagger(\bar{a}1^\top - \Delta^\top)$
Return: the row vectors of $\tilde{Y}$, denoted $\tilde{y}_1, \ldots, \tilde{y}_n \in \mathbb{R}^d$

We perturb both the dissimilarities and the landmark points, and qualitatively characterize how it will affect the returned positions by trilateration. (In principle, the perturbed point set need not have the same mean as the original point set, but we assume this is the case, for simplicity and because it suffices for our application of this result in Section 5.) For a configuration $Y = [y_1 \cdots y_m]^\top$, define its max-radius as

$$\rho_\infty(Y) = \max_{i\in[m]} \|y_i\|, \tag{24}$$

and note that $\rho(Y) \leq \rho_\infty(Y)$. We content ourselves with a bound in Frobenius norm.[1]

Theorem 2 *Consider a centered configuration $Y \in \mathbb{R}^{m\times d}$ that spans the whole space $\mathbb{R}^d$, and for a given configuration $\tilde{Y} \in \mathbb{R}^{n\times d}$, let $\tilde{\Delta} \in \mathbb{R}^{n\times m}$ denote the matrix of dissimilarities between $\tilde{Y}$ and Y, namely $\tilde{\Delta}_{ij} = \|\tilde{y}_i - y_j\|^2$. Let $Z \in \mathbb{R}^{m\times d}$ be another centered configuration that spans the whole space, and let $\tilde{\Lambda} \in \mathbb{R}^{n\times m}$ be an arbitrary matrix. Then, trilateration with inputs Z and $\tilde{\Lambda}$ returns $\tilde{Z} \in \mathbb{R}^{n\times d}$ satisfying*

$$\begin{aligned}\|\tilde{Z} - \tilde{Y}\|_2 \leq \tfrac{1}{2}\|Z^\ddagger\|\|\tilde{\Lambda} - \tilde{\Delta}\|_2 + 2\|\tilde{Y}\|\|Z^\ddagger\|\|Z - Y\|_2 \\ + 3\sqrt{m}(\rho_\infty(Y) + \rho_\infty(Z))\|Z^\ddagger\|\|Z - Y\|_2 + \|Y\|\|\tilde{Y}\|\|Z^\ddagger - Y^\ddagger\|_2\,. \quad (25)\end{aligned}$$

In the bound (25), we see that the first term captures the effect of the error in the dissimilar matrix, i.e., $\|\tilde{\Delta} - \tilde{\Lambda}\|$, while the other three terms reflect the impact of the error in the landmark positions, i.e, $\|Z - Y\|$. As we expect, we have a more accurate embedding as these two terms get smaller, and in particular, when $\tilde{\Delta} = \tilde{\Lambda}$ and $Y = Z$ (no error in the inputs), we have exact recovery, which corroborates our derivation in Section A.1.

Remark 5 *For a bound not involving the pseudo-inverse of Z – which may be difficult to interpret – we can upper bound the right-hand side of* (25) *using*

$$\rho_\infty(Z) \leq \rho_\infty(Y) + \rho_\infty(Z - Y), \quad \|Z^\ddagger\| \leq \|Y^\ddagger\| + \|Z^\ddagger - Y^\ddagger\|, \quad (26)$$

and

$$\|Z^\ddagger - Y^\ddagger\|_p \leq \frac{\sqrt{2}\|Y^\ddagger\|^2\|Z - Y\|_p}{(1 - \|Y^\ddagger\|\|Z - Y\|)_+^2}, \quad p \in \{2, \infty\}, \quad (27)$$

as per Lemma 2. Also, a simple application of Mirsky's inequality (50) *implies that, when Y spans the whole space then so does Z whenever $\|Y^\ddagger\|\|Z - Y\| < 1$.*

The proof is in Section 8.3. We now derive from this result another one in terms of point sets instead of matrices.

Corollary 3 *Consider a centered point set $y_1, \ldots, y_m \in \mathbb{R}^d$ with radius ρ, max-radius ρ_∞, and half-width $\omega > 0$. For a point set $\tilde{y}_1, \ldots, \tilde{y}_n \in \mathbb{R}^d$ with radius ζ, set $\tilde{\delta}_{ij} = \|\tilde{y}_i - y_j\|^2$. Also, let $z_1, \ldots, z_m \in \mathbb{R}^d$ denote another centered point set, and let $(\tilde{\lambda}_{ij})$ denote another arbitrary set of numbers for $1 \leq i \leq n$, $1 \leq j \leq m$. Set $\varepsilon = \max_{i\in[m]} \|z_i - y_i\|$ and $\eta^4 = \frac{1}{nm}\sum_{ij}(\tilde{\lambda}_{ij} - \tilde{\delta}_{ij})^2$. If $\varepsilon \leq \omega/2$, trilateration with inputs $z_1, \ldots, z_m$ and $(\tilde{\lambda}_{ij})$ returns $\tilde{z}_1, \ldots, \tilde{z}_n \in \mathbb{R}^d$ satisfying*

$$\left(\frac{1}{n}\sum_{i\in[n]} \|\tilde{z}_i - \tilde{y}_i\|^2\right)^{1/2} \leq C_0\left(\frac{\eta^2}{\omega} + \left[\frac{\rho\zeta}{\omega^2} + \frac{\sqrt{m}\rho_\infty}{\sqrt{n}\omega}\right]\varepsilon\right), \quad (28)$$

where C_0 is a universal constant.

Corollary 3 follows from Theorem 2 and its proof is given in Section 8.5.

1. All Schatten norms are equivalent here up to a multiplicative constant that depends on d, since the matrices that we consider have rank of order d.

Remark 6 *In the bound* (28)*, the terms ε and η respectively quantify the errors in the positions of landmarks and the error in the dissimilarities that are fed to the trilateration procedure. As we expect, smaller values of ε and η lead to a more accurate embedding of the points, and in the extreme situation where $\varepsilon = 0$ and $\eta = 0$, we can infer the positions of the points $\tilde{y}_i$ exactly. Also, the bound is reciprocal in the half-width of the landmark set, ω. This is also expected because a small ω means that the landmarks have small dispersion along some direction in $\mathbb{R}^d$ and hence the positions of other points cannot be well approximated along that direction. This can also be seen from Step 3 of the trilateration procedure. The quantity $\|Y^\ddagger\|$ measures the sensitivity of X to the dissimilarities Δ. Invoking* (20)*, $\omega = \|Y^\ddagger\|^{-1}/m$, and hence a small ω corresponds to large sensitivity, meaning that a small perturbation in Δ can lead to large errors in X. This is consistent with our bound as the error in Δ, i.e., η^2 appears by the scaling factor $1/\omega$.*

5. Applications to manifold learning

Consider a set of points in a possibly high-dimensional Euclidean space, that lie on a smooth Riemannian manifold. Isometric manifold learning (or embedding) is the problem of embedding these points into a lower-dimensional Euclidean space, and do as while preserving as much as possible the Riemannian metric. There are several variants of the problem under other names, such as nonlinear dimensionality reduction.

Remark 7 *Manifold learning is intimately related to the problem of embedding items with only partial dissimilarity information, which practically speaking means that some of the dissimilarities are missing. We refer to this problem as graph embedding below, although it is known under different names such as graph realization, graph drawing, and sensor localization. This connection is due to the fact that, in manifold learning, the short distances are nearly Euclidean, while the long distances are typically not. In fact, the two methods for manifold learning that we consider below can also be used for graph embedding. The first one, Isomap (Tenenbaum et al., 2000), coincides with MDS-MAP (Shang et al., 2003) (see also (Niculescu and Nath, 2003)), although the same method was suggested much earlier by Kruskal and Seery (1980); the second one, Maximum Variance Unfolding, was proposed as a method for graph embedding by the same authors (Weinberger et al., 2006), and is closely related to other graph embedding methods (Biswas et al., 2006; Javanmard and Montanari, 2013; So and Ye, 2007).*

5.1. A performance bound for (Landmark) Isomap

Isomap is a well-known method for manifold learning, suggested by Tenenbaum, de Silva, and Langford (2000). Algorithm 4 describes the method. (There, we use the notation $A^{\circ 2}$ to denote the matrix with entries A_{ij}^2.)

There are two main components to Isomap: 1) Form the r-ball neighborhood graph based on the data points and compute the shortest-path distances; 2) Pass the obtained distance matrix to classical scaling (together with the desired embedding dimension) to obtain an embedding. The algorithm is known to work well when the underlying manifold is isometric to a convex domain in $\mathbb{R}^d$. Indeed, assuming an infinite sample size, so that the

Algorithm 4 Isomap

Input: data points $x_1, \ldots, x_n \in \mathbb{R}^D$, embedding dimension d, neighborhood radius r
Output: embedding points $z_1, \ldots, z_n \in \mathbb{R}^d$

1: construct the graph on $[n]$ with edge weights $w_{ij} = \|x_i - x_j\| \, \mathbb{I}\{\|x_i - x_j\| \leq r\}$
2: compute the shortest-path distances in that graph $\Gamma = (\gamma_{ij})$
3: apply classical scaling with inputs $\Gamma^{\circ 2}$ and d, resulting in points $z_1, \ldots, z_n \in \mathbb{R}^d$
Return: the points $z_1, \ldots, z_n$

data points are in fact all the points of the manifold, as $r \to 0$, the shortest-path distances will converge to the geodesic distances on the manifold, and thus, in that asymptote (infinite sample size and infinitesimal radius), an isometric embedding in $\mathbb{R}^d$ is possible under the stated condition. We will assume that this condition, that the manifold is isometric to a convex subset of $\mathbb{R}^d$, holds.

In an effort to understand the performance of Isomap, Bernstein et al. (2000) study how well the shortest-path distances in the r-ball neighborhood graph approximate the actual geodesic distances. Before stating their result we need to state a definition.

Definition 2 *The* reach *of a subset $\mathcal{A}$ in some Euclidean space is the supremum over $t \geq 0$ such that, for any point x at distance at most t from $\mathcal{A}$, there is a unique point among those belonging to $\mathcal{A}$ that is closest to x. When $\mathcal{A}$ is a C^2 submanifold, its reach is known to bound its radius of curvature from below (Federer, 1959).*

Assume that the manifold $\mathcal{M}$ has reach at least $\tau > 0$, and the data points are sufficiently dense in that

$$\min_{i \in [n]} g_{\mathcal{M}}(x, x_i) \leq a, \quad \forall x \in \mathcal{M}, \tag{29}$$

where $g_{\mathcal{M}}$ denote the metric on $\mathcal{M}$ (induced by the surrounding Euclidean metric). If r is sufficiently small in that $r < \tau$, then Bernstein et al. (2000) show that

$$1 - c_0(r/\tau)^2 \leq \frac{\gamma_{ij}}{g_{ij}} \leq 1 + c_0(a/r), \quad \forall i, j \in [n], \tag{30}$$

where γ_{ij} is the graph distance, g_{ij} is the geodesic distance between x_i and x_j, and $c_0 \geq 1$ is a universal constant. (In fact, Bernstein et al. (2000) derive such a bound under the additional condition that $\mathcal{M}$ is geodesically convex, although the result can be generalized without much effort (Arias-Castro and Gouic, 2017).)

We are able to improve the upper bound in the restricted setting considered here, where the underlying manifold is assumed to be isometric to a convex domain.

Proposition 1 *In the present situation, there is a universal constant $c_1 \geq 1$ such that, if $a/r \leq 1/\sqrt{c_1}$,*

$$\frac{\gamma_{ij}}{g_{ij}} \leq 1 + c_1(a/r)^2, \quad \forall i, j \in [n]. \tag{31}$$

Thus, if we set

$$\xi = c_0(r/\tau)^2 \vee c_1(a/r)^2, \tag{32}$$

using the notation $a \vee b = \max(a, b)$, and it happens that $\xi < 1$, we have

$$1 - \xi \leq \frac{\gamma_{ij}}{g_{ij}} \leq 1 + \xi, \quad \forall i, j \in [n]. \tag{33}$$

Armed with our perturbation bound for classical scaling, we are able to complete the analysis of Isomap, obtaining the following performance bound.

Corollary 4 *In the present context, let $y_1, \ldots, y_n \in \mathbb{R}^d$ denote a possible (exact and centered) embedding of the data points $x_1, \ldots, x_n \in \mathcal{M}$, and let ρ and ω denote the max-radius and half-width of the embedded points, respectively. Let ξ be defined by Equation* (32)*. If $\xi \leq \frac{1}{24}(\rho/\omega)^{-2}$, then Isomap returns $z_1, \ldots, z_n \in \mathbb{R}^d$ satisfying*

$$\min_{Q \in \mathcal{O}} \left(\frac{1}{n} \sum_{i \in [n]} \|z_i - Q y_i\|^2 \right)^{1/2} \leq \frac{36\sqrt{d}\rho^3}{\omega^2} \xi. \tag{34}$$

Remark 8 *As we can see, the performance of Isomap degrades as ω gets smaller, which we already justified in Remark 6. Also the performance improves for smaller values of ξ. Recalling the definition of ξ in* (32)*, fixing r, a smaller ξ corresponds to a denser set of points on the manifold, i.e., a smaller a, and also a smaller reach, i.e., a smaller τ, which leads the graph distances to better approximate the geodesic distances.*

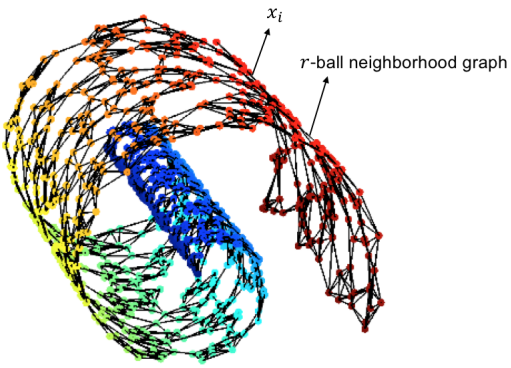
Proof Before we provide the proof, we refer to Figure 1 for a schematic representation of exact locations $y_i \in \mathbb{R}^d$, data points $x_i \in \mathcal{M}$, returned locations by Isomap $z_i \in \mathbb{R}^d$, as well as graph distances γ_{ij} and geodesic distances g_{ij}.

The proof itself is a simple consequence of Corollary 2. Indeed, with (33) it is straightforward to obtain (with η defined in Corollary 2 and γ_{ij} and g_{ij} as above),

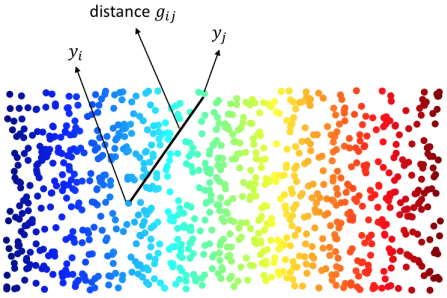
$$\eta^2 \leq \max_{i,j \in [n]} |\gamma_{ij}^2 - g_{ij}^2| \leq \max_{i,j \in [n]} (2\xi + \xi^2) g_{ij}^2 \leq (2\xi + \xi^2)(2\rho)^2 \leq 12\rho^2 \xi, \tag{35}$$

where in the last step we used the fact that $\xi < 1$ because $\xi \leq \frac{1}{24}(\rho/\omega)^{-2}$ by our assumption and $\omega \leq \rho$, by definition. In particular, η fulfills the conditions of Corollary 2 under the stated bound ξ, so we may conclude by applying that corollary and simplifying. ■

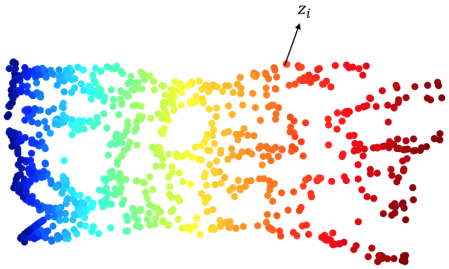
If $\mathcal{D}$ is a domain in $\mathbb{R}^d$ that is isometric to $\mathcal{M}$, then the radius of the embedded points (ρ above) can be bounded from above by the radius of $\mathcal{D}$, and under mild assumptions on the sampling, the half-width of the embedded points (ω above) can be bounded from below by a constant times the half-width of $\mathcal{D}$, in which case ρ and ω should be regarded as fixed. Similarly, τ should be considered as fixed, so that the bound is of order $O(r^2 \vee (a/r)^2)$, optimized at $r \asymp a^{1/2}$. If the points are well spread-out, for example if the points are sampled iid from the uniform distribution on the manifold, then a is on the order of $(\log(n)/n)^{1/d}$, and the bound (with optimal choice of radius) is $O((\log(n)/n)^{1/d})$.



(a) Data points $x_i \in \mathcal{M}$ and the r-ball neighborhood graph



(b) Exact embedding onto the low-dimensional space



(c) Returned locations by Isomap

Figure 1: Schematic representation of exact locations $y_i \in \mathbb{R}^d$, data points $x_i \in \mathcal{M}$, returned locations by Isomap $z_i \in \mathbb{R}^d$. Note that $g_{ij} = \|y_i - y_j\|$ is the geodesic distance between x_i and x_j because $\{y_i\}_{i=1}^n$ is an exact isometric embedding of data points $\{x_i\}_{i=1}^n$. Also the distances γ_{ij} are computed as shortest path distances between x_i and x_j on the r-ball neighborhood graph.

Landmark Isomap Because of the relatively high computational complexity of Isomap, and also of classical scaling, de Silva and Tenenbaum (2004, 2003) proposed a Nyström approximation (as explained in (Platt, 2005)). Seen as a method for MDS, it starts by embedding a small number of items, which effectively play the role of landmarks, and then embedding the remaining items by trilateration based on these landmarks. Seen as a method for manifold learning, the items are the points in space, and the dissimilarities are the squared graph distances, which are not provided and need to be computed. Algorithm 5 details the method in this context. The landmarks may be chosen at random from the data points, although other options are available, and we discuss some of them in Section 7.2.

With our work, we are able to provide a performance bound for Landmark Isomap.

Corollary 5 *Consider n data points $x_1, \ldots, x_n \in \mathcal{M}$, which has a possible (exact and centered) embedding in $\mathbb{R}^d$. Let $\mathcal{L}$ be a subset of the points ($|\mathcal{L}| = \ell$) with exact embedding $\{y_1, \ldots, y_\ell\}$ and denote the embedding of the other points by $\tilde{y}_1, \ldots, \tilde{y}_{n-\ell}$. Assume that $\{y_1, \ldots, y_m\}$ has half-width $\omega_* > 0$, the exact embedding $\{y_1, \ldots, y_\ell\} \cup \{\tilde{y}_1, \ldots, \tilde{y}_{n-\ell}\}$ has*

Algorithm 5 Landmark Isomap

Input: data points $x_1, \ldots, x_n \in \mathbb{R}^D$, embedding dimension d, neighborhood radius r, number of landmarks ℓ

Output: embedding points $\{z_i : i \in \mathcal{L}\} \cup \{\tilde{z}_i : i \notin \mathcal{L}\} \subseteq \mathbb{R}^d$ for a choice of $|\mathcal{L}| = \ell$ landmarks

1: construct the graph on $[n]$ with edge weights $w_{ij} = \|x_i - x_j\| \mathbb{I}\{\|x_i - x_j\| \leq r\}$
2: select $\mathcal{L} \subset [n]$ of size ℓ according to one of the methods in Section 7.2
3: compute the shortest-path distances in that graph $\Gamma = (\gamma_{ij})$ for $(i, j) \in [n] \times \mathcal{L}$
4: apply classical scaling with inputs $\Gamma^{\circ 2}_{\mathcal{L}\times\mathcal{L}}$ and d, resulting in (landmark) points $z_i, i \in \mathcal{L}$ in $\mathbb{R}^d$
5: for each $i \notin \mathcal{L}$, apply trilateration based on $\{z_j : j \in \mathcal{L}\}$ and $\Gamma^{\circ 2}_{i\times\mathcal{L}}$ to obtaining $\tilde{z}_i \in \mathbb{R}^d$

Return: the points $\{z_i : i \in \mathcal{L}\} \cup \{\tilde{z}_i : i \notin \mathcal{L}\}$.

maximum-radius ρ, and $\ell \leq (n/2) \wedge [(72\sqrt{d}\xi)^{-2}(\rho/\omega_*)^{-6}]$. Then the Landmark Isomap, with the choice of $\mathcal{L}$ as landmarks, returns $\{z_1, \ldots, z_\ell\} \cup \{\tilde{z}_1, \ldots, \tilde{z}_{n-\ell}\} \subseteq \mathbb{R}^d$ satisfying

$$\min_{Q\in\mathcal{O}} \left(\frac{1}{n} \sum_{i\in[\ell]} \|z_i - Qy_i\|^2 + \frac{1}{n} \sum_{i\in[n-\ell]} \|\tilde{z}_i - Q\tilde{y}_i\|^2 \right)^{1/2} \leq C_1 \frac{\rho^2}{\omega_*}, \tag{36}$$

where C_1 is a universal constant.

The result is a direct consequence of applying Corollary 4, which allows us to control the accuracy of embedding the landmarks using classical scaling, followed by applying Corollary 3, which allows us to control the accuracy of embedding using trilateration. The proof is given in Section 8.6. As we see our bound (36) on the embedding error improves when the half-width of the landmarks, ω_*, increases. We justified this observation in Remark 6: a higher half-width of the landmarks yields a better performance of the trilateration procedure. In Section 7.2, we use this observation to provide guidelines for choosing landmarks.

We note that for the set of (embedded) landmarks to have positive half-width, it is necessary that they span the whole space, which compels $\ell \geq d + 1$. In Section 7.2 we show that choosing the landmarks at random performs reasonably well in that, with probability approaching 1 very quickly as ℓ increases, their (embedded) half-width is at least half that of the entire (embedded) point set.

5.2. A performance bound for Maximum Variance Unfolding

Maximum Variance Unfolding is another well-known method for manifold learning, proposed by Weinberger and Saul (2006a,b). Algorithm 6 describes the method, which relies on solving a semidefinite relaxation. There is also an interpretation of MVU as a regularized shortest path solution (Paprotny and Garcke, 2012, Theorem 2).

Although MVU is broadly regarded to be more stable than Isomap, Arias-Castro and Pelletier (2013) show that it works as intended under the same conditions required by Isomap, namely, that the underlying manifold is geodesically convex. Under these conditions, in fact, under the same conditions as in Corollary 4, where in particular (33) is

Algorithm 6 Variance Unfolding (MVU)

Input: data points $x_1, \ldots, x_n \in \mathbb{R}^D$, embedding dimension d, neighborhood radius r
Output: embedded points $z_1, \ldots, z_n \in \mathbb{R}^d$

1: set $\gamma_{ij} = \|x_i - x_j\|$ if $\|x_i - x_j\| \leq r$, and $\gamma_{ij} = \infty$ otherwise
2: solve the following semidefinite program

$$\text{maximize} \sum_{i,j \in [n]} \|p_i - p_j\|^2 \text{ over } p_1, \ldots, p_n \in \mathbb{R}^D, \text{ subject to } \|p_i - p_j\| \leq \gamma_{ij}$$

3: center a solution set and embed it into $\mathbb{R}^d$ using principal component analysis
Return: the embedded point set, denoted by $z_1, \ldots, z_n$

assumed to hold with ξ sufficiently small, Paprotny and Garcke (2012) show that MVU returns an embedding, $z_1, \ldots, z_n \in \mathbb{R}^d$, with dissimilarity matrix $\Lambda = (\lambda_{ij})$, $\lambda_{ij} = \|z_i - z_j\|^2$, satisfying

$$|\Lambda - \Delta|_1 \leq 9\rho^2 n^2 \xi, \tag{37}$$

where $\Delta = (\delta_{ij})$, $\delta_{ij} = \|y_i - y_j\|^2$ (the correct underlying distances), and for a matrix $A = (a_{ij})$, $|A|_p^p = \sum_{i,j} |a_{ij}|^p$. Based on that, and on our work in Section 3, we are able to provide the following performance bound for MVU, which is similar to the bound we obtained for Isomap.

Corollary 6 *Let $y_1, \ldots, y_n \in \mathbb{R}^d$ denote a possible (exact and centered) embedding of the data points $x_1, \ldots, x_n \in \mathcal{M}$, and let ρ and ω denote the max-radius and half-width of the embedded points, respectively. Suppose that the neighborhood radius r is chosen so that the corresponding neighborhood graph on points $\{x_i\}_{i \in [n]}$ is connected. Let ξ be defined by Equation* (32)*. If $\xi \leq (12\sqrt{3})^{-1}(\rho/\omega)^{-2}$, then Maximum Variance Unfolding returns $z_1, \ldots, z_n \in \mathbb{R}^d$ satisfying*

$$\min_{Q \in \mathcal{O}} \left(\frac{1}{n} \sum_{i \in [n]} \|z_i - Q y_i\|^2 \right)^{1/2} \leq \frac{18\sqrt{3d}\rho^3}{\omega^2} \xi. \tag{38}$$

Proof As in (35), we have

$$|\Lambda - \Delta|_\infty = \max_{i,j \in [n]} |\gamma_{ij}^2 - g_{ij}^2| \leq 12\rho^2 \xi, \tag{39}$$

so that, in combination with (37), we have

$$\|\Lambda - \Delta\|_2 \leq |\Lambda - \Delta|_\infty^{1/2} |\Lambda - \Delta|_1^{1/2} \leq 6\sqrt{3} n \rho^2 \xi. \tag{40}$$

In particular, the conditions of Corollary 2 are met under the stated bound on ξ. Therefore, we may apply that corollary to conclude. ■

6. Numerical Experiments

Procrustes problem. We let $n = 100$, $d = 10$ and generate $X \in \mathbb{R}^{n\times d}$ as $X = UDV^\top$, where $U, V \in \mathbb{R}^{n\times d}$ are two random orthonormal matrices drawn independently from the Haar measure and D is a diagonal matrix of size d, with its diagonal entries chosen uniformly at random from $[0, 10\delta]$. We also generate $Z \in \mathbb{R}^{n\times d}$ via the same generative model as X and let $Y = aX + (1-a)Z$ for a changing values from zero to one. As a varies, we compute $\varepsilon^2 = \|YY^\top - XX^\top\|_2$ and then solve for the procrustes problem $\min_{Q\in\mathcal{O}} \|Y - XQ\|_2$ using Algorithm 1. Figure 2 plots $\|Y - XQ\|_2$ versus ϵ in the log-log scale, for different values of $\delta = 1, 2, \ldots, 5, 10$.

Firstly, we observe that the slope of the best fitted line to each curve is very close to 2, indicating that $\|Y - XQ\|_2$ scales as ε^2. Secondly, since the singular values of X (there are $d = 10$ of them) are drawn uniformly at random from $[0, 10\delta]$, we have that $\|X^\ddagger\|$ changes as $1/\delta$. As we observe from the plot, for fixed ε, the term $\|Y - XQ\|_2$ is monotone in $\|X^\ddagger\| \sim \delta^{-1}$. These observations are in good match with our theoretical bound in Theorem 1.

We next compare the procrustes error $\|Y - XQ\|_2$ with the proposed upper bounds (4) and (5) in Theorem 1. Recall that the upper bound (4) reads as

$$\min_{Q\in\mathcal{O}} \|Y - XQ\|_p \le \begin{cases} \|X^\ddagger\|\varepsilon^2 + \left((1 - \|X^\ddagger\|^2\varepsilon^2)^{-1/2}\|X^\ddagger\|\varepsilon^2\right) \wedge \left(d^{1/4}\varepsilon\right), & \text{if } \varepsilon\|X^\ddagger\| < 1\,, \\ \|X^\ddagger\|\varepsilon^2 + d^{1/4}\varepsilon & \text{otherwise.} \end{cases}$$

Under the same generative model for configurations X and $Y \in \mathbb{R}^{n\times d}$, with $\delta = 0.1$, Figure 3(a) plots the procrustes error along with the above upper bound in the log-log scale. The solid part of the red curve corresponds to the regime where $\varepsilon\|X^\ddagger\| < 1$ and the dashed part refers to the regime where $\varepsilon\|X^\ddagger\| > 1$. Likewise, we plot the upper bound (5) in black, which assumes $\varepsilon\|X^\ddagger\| < \frac{1}{\sqrt{2}}$. The part of this upper bound where this assumption is violated is plotted in dashed form. Figure 3(b) depicts the same curves in the regular (non-logarithmic) scale. In Figure 3(c), we show the ratio of the upper bounds over the computed procrustes error from the simulation.

Manifold learning algorithms. To evaluate the error rates obtained for manifold learning algorithms in Section 5, we carry out two numerical experiments.

For the first experiment, we consider the 'bending map' $\mathcal{B} : [-0.5, 0.5]^d \mapsto \mathbb{R}^{d+1}$, defined as

$$\mathcal{B}(t_1, t_2, \ldots, t_d) = [R\sin(t_1/R), t_2, \ldots, t_d, R(1 - \cos(t_1/R))]\,.$$

This map bends the d-dimensional hypercube in the $(d+1)$-dimensional space and the parameter R controls the degree of bending (with a large R corresponding to a small amount of bending), and thus controls the reach of the resulting submanifold of $\mathbb{R}^{d+1}$. See Figure 4a for an illustration.

We set $R = 0.2$ and generate n points $y_1, \ldots, y_n$ uniformly at random in the d-dimensional hypercube. The samples on the manifold are then given by $x_i = \mathcal{B}(y_i)$, for $i = 1, \ldots, n$. Since the points are well spread out on the manifold, the quantity a given by (29) is $O(\log(n)/n)^{1/d}$ and following our discussion after the proof of Corollary 4, our bound (34) is optimized at $r \asymp a^{1/2}$. With this choice of a, our bound (34) becomes of order $O((\log(n)/n)^{1/d})$. Follow-

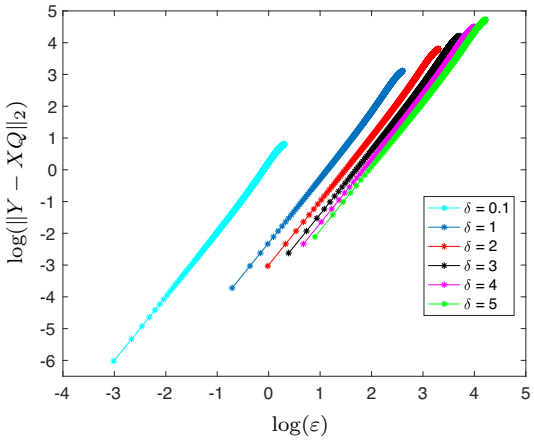


Figure 2: Procrustes error $\min_{Q\in\mathcal{O}} \|Y - XQ\|_2$ versus ϵ, in log-log scale and for different values of δ (i.e., different values of $\|X^{\ddagger}\|$).

ing this guideline, we let $r = 2(\log(n)/n)^{1/(2d)}$ and run Isomap (Algorithm 4) for $d = 2, 8, 15$ and $n = 100, 200, \ldots, 1000$.

Denoting by $z_1, \ldots, z_n \in \mathbb{R}^d$ the output of Isomap in $\mathbb{R}^d$, and $Z = [z_1, \ldots, z_n]^\top \in \mathbb{R}^{n\times d}$, $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^{n\times d}$, we compute the mismatch between the inferred locations Z and the original ones Y via our metric

$$\mathrm{d}(Y, Z) = \frac{1}{\sqrt{n}} \min_{Q\in\mathcal{O}} \|Z - YQ\|_2 = \min_{Q\in\mathcal{O}} \left(\frac{1}{n}\sum_{i\in[n]} \|z_i - Qy_i\|^2\right)^{1/2}.$$

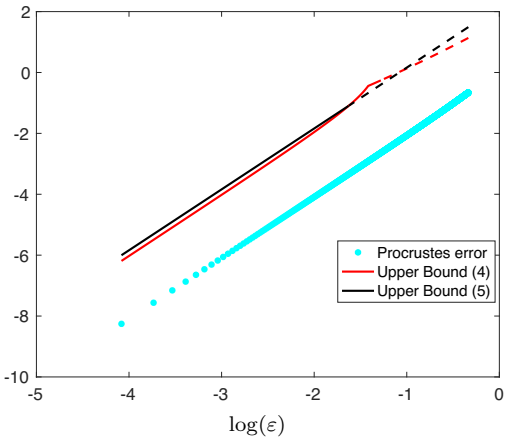
For each n, we run the experiment for 50 different realizations of the points in the hypercube and compute the average and the 95% confidence region of the the error $\mathrm{d}(Y, Z)$. Figure 4b reports the results for Isomap in a log-log scale, along with the best linear fits to the data points. The slopes of the best fitted lines are $-0.50, -0.14, -0.08$, for $d = 2, 8, 15$, which are close to the corresponding exponent $-\frac{1}{d}$ implied by our Corollary 4, namely, -0.50, -0.125, -0.067 (ignoring logarithmic factors).

Likewise, Figure 4c shows the error for Maximum Variance Unfolding (MVU) in the same experiment. As we see, MVU is achieving lower error rates than Isomap. Also the slopes of the best fitted lines are $-0.47, -0.12, -0.04$, for $d = 2, 8, 15$, which are in good agreement with our error rate ($O(\sqrt{d}n^{-1/d})$) in Corollary 6.

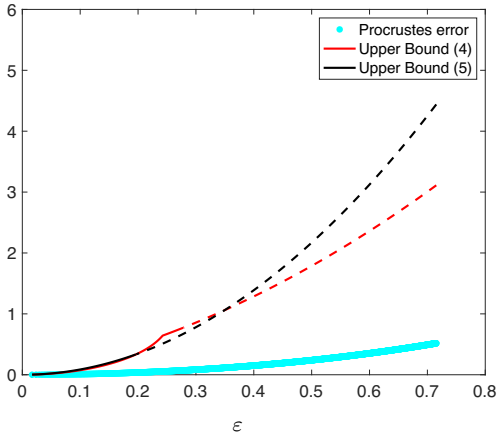
In the second experiment, we consider the Swiss Roll manifold, which is a prototypical example in manifold learning. Specifically we consider the mapping $\mathcal{T} : [-\frac{9\pi}{2}, \frac{15\pi}{2}] \times [-40, 40] \mapsto \mathbb{R}^3$, given by

$$\mathcal{T}(t_1, t_2) = [t_1\cos(t_1), t_2, t_1\sin(t_1)]\,. \tag{41}$$

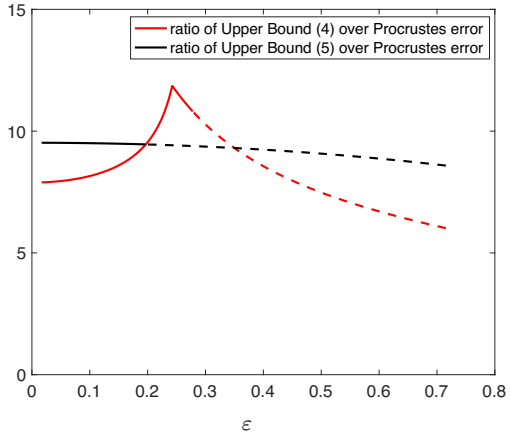
The range of this mapping is a Swiss Roll manifold (see Figure 5a for an illustration.) For this experiment, we consider non-uniform samples from the manifold as follows. For each n, we keep drawing points with first coordinate $\sim \mathsf{N}(1, \sigma^2)$ and the second coordinate $\sim \mathsf{N}(0, (10\sigma)^2)$, for a pre-determined value of σ. If the generated point falls in the rectangle



(a) log-log scale



(b) regular (non-logarithmic) scale

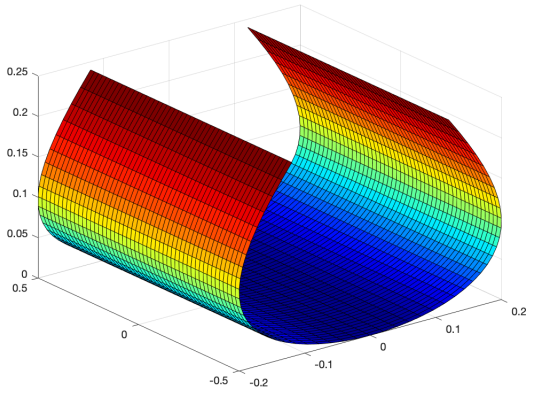


(c) ratio of upper bounds over the procrustes error

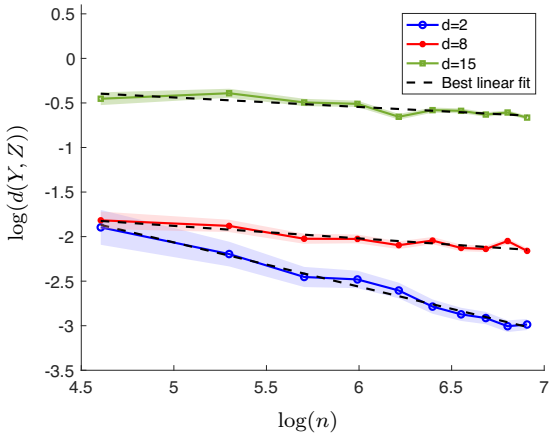
Figure 3: Comparison between the procrustes error $\|Y - XQ\|_2$ with the upper bound (4) and upper bound (5) for the described generative model for configurations X, Y with $\delta = 0.1$; (a) is in log-log scale, (b) is in regular scale; (c) plots the ratio of the upper bounds over the procrustes error. For the red curve (upper bound (4)) the solid part corresponds to the regime $\varepsilon\|X^\ddagger\| < 1$. For the black curve (upper bound (5)) the solid part corresponds to the regime where the assumption in deriving this bound, namely $\varepsilon\|X^\ddagger\| < \frac{1}{\sqrt{2}}$, holds.

$[-\frac{9\pi}{2}, \frac{15\pi}{2}] \times [-40, 40]$, we keep that otherwise reject it. We continue this procedure until we generate n points $y_1, \ldots, y_n$. The samples on the manifold are given by $x_i = \mathcal{T}(y_i)$. The parameter σ controls the dispersion of the samples on the manifold.

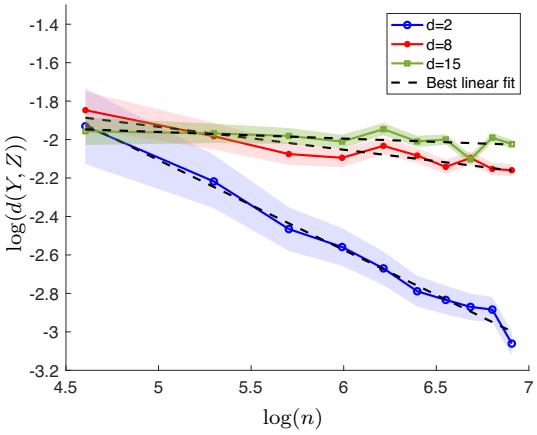
We run Isomap and MVU to infer the underlying positions y_i from the samples x_i on the manifold. For each $\sigma = 0.5, 1, 2$ and $n = 100, 200, \ldots, 1000$, we run the experiment 50 times and compute the average error $\mathrm{d}(Y, Z)$ and the 95% confidence region. The results are reported in Figure 5 in a log-log scale. As we see the error curves for both algorithms



(a) Bended square ($d = 2$, $R = 0.2$)



(b) Isomap



(c) Maximum Variance Unfolding

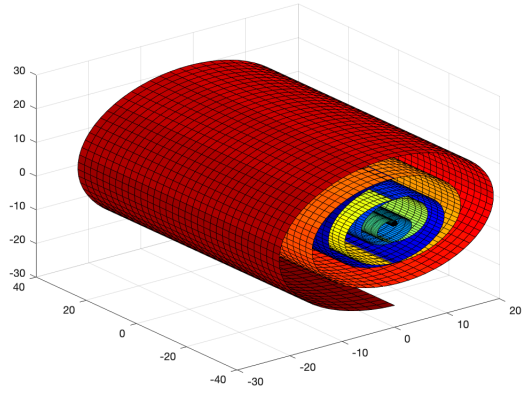
Figure 4: Performance of Isomap and MVU on the data points sampled from the bended hypercube of dimension d. Different curves correspond to different values of d. Each curve is plotted along with the corresponding best fitted line and the 95% confidence region.

scales as $\sim n^{-1/2}$ for various choice of σ, which again supports our theoretical error rates stated in Section 5.

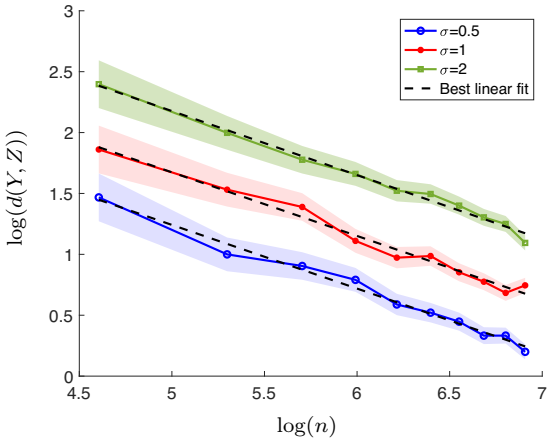
7. Discussion

7.1. Optimality considerations

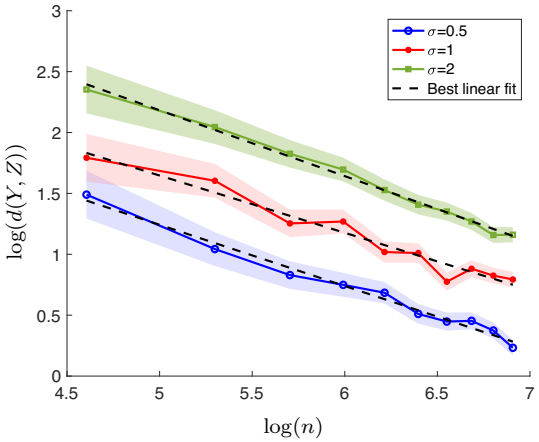
The performance bounds that we derive for Isomap and Maximum Variance Unfolding are the same up to a universal multiplicative constant. This may not be surprising as they are known to be closely related, since the work of Paprotny and Garcke (2012). Based on our analysis of classical scaling, we believe that the bound for Isomap is sharp up to a



(a) Swiss roll manifold



(b) Isomap



(c) Maximum Variance Unfolding

Figure 5: Performance of Isomap and MVU on the data points sampled non-uniformly from the Swiss Roll manifold. Different curves correspond to different values of σ, which controls the dispersion of the sampled points. Each curve is plotted along with the corresponding best fitted line and the 95% confidence region.

multiplicative constant. But one may wonder if Maximum Variance Unfolding, or a totally different method, can do strictly better.

This optimality problem can be formalized as follows:

Consider the class of isometries $\varphi : \mathcal{D} \to \mathcal{M} \subset \mathbb{R}^D$, one-to-one, such that its domain $\mathcal{D}$ is a convex subset of $\mathbb{R}^d$ with max-radius at most ρ_0 and half-width at least $\omega_0 > 0$, and its range $\mathcal{M}$ is a submanifold with reach at least $\tau_0 > 0$. To each such isometry φ, we associate the uniform distribution on its range $\mathcal{M}$, denoted by P_φ. We then assume that we are provided with iid samples of size n from P_φ, for some unknown isometry φ in that class. If the sample is denoted by $x_1, \ldots, x_n \in \mathcal{M}$, with $x_i = \varphi(y_i)$ for some $y_i \in \mathcal{D}$, the

goal is to recover $y_1, \ldots, y_n$ up to a rigid transformation, and the performance is measured in average squared error. Then, what is the optimal achievable performance?

Despite some closely related work on manifold estimation, in particular work of Genovese et al. (2012) and of Kim and Zhou (2015), we believe the problem remains open. Indeed, while in the setting in dimension $d = 1$ the two problems are particularly close, in dimension $d \geq 2$ the situation here appears more delicate here, as it relies on a good understanding of the interpolation of points by isometries.

7.2. Choosing landmarks

In this subsection we discuss the choice of landmarks. We consider the two methods originally proposed by de Silva and Tenenbaum (2004):

- Random. The landmarks are chosen uniformly at random from the data points.
- MaxMin. After choosing the first landmark uniformly at random from the data points, each new landmark is iteratively chosen from the data points to maximize the minimum distance to the existing landmarks.

(For both methods, de Silva and Tenenbaum (2004) recommend using different initializations.)

The first method is obviously less computationally intensive compared to the second method, but the hope in the more careful (and also more costly) selection of landmarks in the second method is that it would require fewer landmarks to be selected. In any case, de Silva and Tenenbaum (2004) observe that the random selection is typically good enough in practice, so we content ourselves with analyzing this method.

In view of our findings (Corollary 3, 4, 5, and 6), a good choice of landmarks is one that has large (embedded) half-width, ideally comparable to, or even larger than that of the entire dataset. In that light, the problem of selecting good landmarks is closely related, if not identical, to problem of selecting rows of a tall matrix in a way that leads to a submatrix with good condition number. In particular, several papers have established bounds for various ways of selecting the rows, some of them listed in (Holodnak and Ipsen, 2015, Tab 2). Here the situation is a little different in that the dissimilarity matrix is not directly available, but rather, rows (corresponding to landmarks) are revealed as they are selected.

The Random method, nonetheless, has been studied in the literature. Rather than fetch existing results, we provide a proof for the sake of completeness. As everyone else, we use random matrix concentration (Tropp, 2012). We establish a bound for a slightly different variant where the landmarks are selected with replacement, as it simplifies the analysis. Related work is summarized in (Holodnak and Ipsen, 2015, Tab 3), although for the special case where the data matrix (denoted Y earlier) has orthonormal columns (an example of paper working in this setting is (Ipsen and Wentworth, 2014)).

Proposition 2 *Suppose we select ℓ landmarks among n points in dimension d, with half-width ω and max-radius ρ_∞, according to the* Random *method, but with replacement. Then with probability at least $1 - 2(d+1)\exp[-\ell\omega^2/9\rho_\infty^2]$, the half-width of the selected landmarks is at least $\omega/2$.*

The proof of Proposition 2 is given in Section A.3. Thus, if $\ell \geq 9(\rho_\infty/\omega)^2 \log(2(d+1)/\delta)$, then with probability at least $1-\delta$ the landmark set has half-width at least $\omega/2$. Consequently, if the dataset is relatively well-conditioned in that its aspect ratio, ρ_∞/ω, is relatively small, then Random (with replacement) only requires the selection of a few landmarks in order to output a well-conditioned subset (with high probability).

8. Proofs

8.1. Preliminaries

We start by stating a number of lemmas pertaining to linear algebra and end the section with a result for a form of procrustes analysis, a well-known method for matching two sets of points in a Euclidean space.

Schatten norms For a matrix[2] A, we let $\nu_1(A) \geq \nu_2(A) \geq \cdots$ denote its singular values. Let $\|\cdot\|_p$ denote the following Schatten quasi-norm,

$$\|A\|_p \equiv \big(\nu_1(A)^p + \cdots + \nu_d(A)^p\big)^{1/p}, \tag{42}$$

which is a true norm when $p \in [1,\infty]$. When $p=2$ it corresponds to the Frobenius norm (which will also be denoted by $\|\cdot\|_2$) and when $p=\infty$ it corresponds to the usual operator norm (which will also be denoted by $\|\cdot\|$). We mention that each Schatten quasi-norm is unitary invariant, and satisfies

$$\|AB\|_p \leq \|A\|_\infty \|B\|_p, \tag{43}$$

for any matrices of compatible sizes, and it is sub-multiplicative if it is a norm ($p \geq 1$). In addition, $\|A\|_p = \|A^\top\|_p$ and $\|A\|_p = \|A^\top A\|_{p/2}^{1/2} = \|AA^\top\|_{p/2}^{1/2}$, due to the fact that

$$\|A\|_p^p = \sum_j \nu_j(A)^p = \sum_j \nu_j(A^\top A)^{p/2} = \|A^\top A\|_{p/2}^{p/2}. \tag{44}$$

and if A and B are positive semidefinite satisfying $A \preceq B$, where $\preceq$ denotes the Loewner order, then $\|A\|_p \leq \|B\|_p$. To see this, note that by definition $A \preceq B$ means $0 \preceq B-A$, and so $0 \leq v^\top(B-A)v$ for any vector v. Therefore, by using the variational principle of eigenvalues (min-max Courant-Fischer theorem) we have

$$\begin{aligned}\nu_j(A) &= \min_{V,\dim(V)=n-j+1} \max_{v\in V,\|v\|=1} v^\top A v \\ &\leq \min_{V,\dim(V)=n-j+1} \max_{v\in V,\|v\|=1} v^\top B v = \nu_j(B)\,,\end{aligned}$$

for all j. As a result, $\|A\|_p \leq \|B\|_p$. We refer the reader to (Bhatia, 2013) for more details on the Schatten norms and the Loewner ordering on positive semidefinite matrices.

Unless otherwise specified, p will be fixed in $[1,\infty]$. Note that, for any fixed matrix A, $\|A\|_p \leq \|A\|_q$ whenever $q \leq p$, and

$$\|A\|_p \to \|A\|_\infty, \quad p \to \infty. \tag{45}$$

2. All the matrices and vectors we consider are real, unless otherwise specified.

Moore-Penrose pseudo-inverse The Moore-Penrose pseudo-inverse of a matrix is defined as follows (Stewart and Sun, 1990, Thm III.1). Let A be a m-by-k matrix, where $m \geq k$, with singular value decomposition $A = UDV^\top$, where U is m-by-k orthogonal, V is k-by-k orthogonal, and D is k-by-k diagonal with diagonal entries $\nu_1 \geq \cdots \geq \nu_l > 0 = \cdots = 0$, so that the ν_j's are the nonzero singular values of A and A has rank l. The pseudo-inverse of A is defined as $A^\ddagger = VD^\ddagger U^\top$, where $D^\ddagger = \mathrm{diag}(\nu_1^{-1}, \ldots, \nu_l^{-1}, 0, \ldots, 0)$. If the matrix A is tall and full rank, then $A^\ddagger = (A^\top A)^{-1} A^\top$. In particular, if a matrix is square and non-singular, its pseudo-inverse coincides with its inverse.

Lemma 1 *Suppose that A is a tall matrix with full rank. Then $A^\ddagger$ is non-singular, and for any other matrix B of compatible size,*

$$\|B\|_p \leq \|A^\ddagger\|_\infty \|AB\|_p. \tag{46}$$

Proof This simply comes from the fact that $A^\ddagger A = I$ (since A is tall and full rank), so that

$$\|B\|_p = \|A^\ddagger AB\|_p \leq \|A^\ddagger\|_\infty \|AB\|_p, \tag{47}$$

by (43). ■

Lemma 2 *Let A and B be matrices of same size. Then, for $p \in \{2, \infty\}$,*

$$\|B^\ddagger - A^\ddagger\|_p \leq \frac{\sqrt{2}\|A^\ddagger\|^2 \|B - A\|_p}{(1 - \|A^\ddagger\| \|B - A\|)_+^2}. \tag{48}$$

Proof A result of Wedin (Stewart and Sun, 1990, Thm III.3.8) gives [3]

$$\|B^\ddagger - A^\ddagger\|_p \leq \sqrt{2}\left(\|B^\ddagger\| \vee \|A^\ddagger\|\right)^2 \|B - A\|_p, \quad p \in \{2, \infty\}. \tag{49}$$

Assuming B has exactly k nonzero singular values, using Mirsky's inequality (Stewart and Sun, 1990, Thm IV.4.11), namely

$$\max_j |\nu_j(B) - \nu_j(A)| \leq \|B - A\|, \tag{50}$$

we have

$$\|B^\ddagger\|^{-1} = \nu_k(B) \geq (\nu_k(A) - \|B - A\|)_+ \geq (\|A^\ddagger\|^{-1} - \|B - A\|)_+. \tag{51}$$

By combining Equations (49) and (51), we get

$$\|B^\ddagger - A^\ddagger\|_p \leq \sqrt{2}\left(\|A^\ddagger\| \vee \frac{1}{(\|A^\ddagger\|^{-1} - \|B - A\|)_+}\right)^2 \|B - A\|_p, \tag{52}$$

from which the result follows. ■

3. For $p = \infty$, the factor $\sqrt{2}$ in (49) can be removed, giving a tighter bound in this case.

Some elementary matrix inequalities The following lemmas are elementary inequalities involving Schatten norms.

Lemma 3 *For any two matrices A and B of same size such that $A^\top B = 0$ or $AB^\top = 0$,*

$$\|A + B\|_p \geq \|A\|_p \vee \|B\|_p. \tag{53}$$

Proof Assume without loss of generality that $A^\top B = 0$. In that case, $(A+B)^\top(A+B) = A^\top A + B^\top B$, which is not smaller than $A^\top A$ or $B^\top B$ in the Loewner order. Therefore,

$$\|A\|_p = \|A^\top A\|_{p/2}^{1/2} \leq \|A^\top A + B^\top B\|_{p/2}^{1/2} \tag{54}$$

$$= \|(A+B)^\top (A+B)\|_{p/2}^{1/2} = \|A+B\|_p, \tag{55}$$

applying several of the properties listed above for Schatten (quasi)norms. ■

Lemma 4 *For any matrix A and any positive semidefinite matrix B, we have*

$$\|A\|_p \leq \|A(B+I)\|_p, \tag{56}$$

where I denotes the identity matrix, with the same dimension as B .

Proof We write

$$A(B+I)(B+I)^\top A^\top = A(B^2 + 2B + I)A^\top = AA^\top + A(B^2+2B)A^\top,$$

with $A(B^2+2B)A^\top \succeq 0$. Therefore, for all k,

$$\nu_k(A(B+I)(B+I)^\top A^\top) \geq \nu_k(AA^\top),$$

which then implies that $\nu_k(A(B+I)) \geq \nu_k(A)$ for all k, which finally yields the result from the mere definition of the p-Schatten norm. ■

8.2. Proof of Theorem 1

Suppose $X, Y \in \mathbb{R}^{n\times d}$ and let $P \in \mathbb{R}^{n\times n}$ be the orthogonal projection onto the column space of X, which can be expressed as $P = XX^\ddagger$. Define $Y_1 = PY$ and $Y_2 = (I-P)Y$, and note that $Y = Y_1 + Y_2$ with $Y_2^\top Y_1 = 0$, and also $Y_2^\top X = 0$.

Define $M = X^\ddagger Y \in \mathbb{R}^{d\times d}$, and apply a singular value decomposition to obtain $M = UDV^\top$, where U and V are orthogonal matrices of size d, and D is diagonal with nonnegative entries. Indeed columns of U span the row space of X and columns of V span the row space of Y. Then define $Q = UV^\top$, which is orthogonal. We show that the bound (5) holds for this orthogonal matrix.

We start with the triangle inequality,

$$\|Y - XQ\|_p = \|Y_1 - XQ + Y_2\|_p \leq \|Y_1 - XQ\|_p + \|Y_2\|_p. \tag{57}$$

Noting that $Y_1 = XX^\ddagger Y = XM$, we have

$$\begin{aligned}\|Y_1 - XQ\|_p = \|XM - XQ\|_p &= \|XUDV^\top - XUV^\top\|_p \\ &= \|XU(D-I)V^\top\|_p \le \|XU(D-I)\|_p. \end{aligned} \tag{58}$$

Now by Lemma 4, we have

$$\|XU(D-I)\|_p \le \|XU(D-I)(D+I)\|_p = \|XU(D^2-I)\|_p\,. \tag{59}$$

Now by unitary invariance, we have

$$\|XU(D^2-I)\|_p = \|XU(D^2-I)U^\top\|_p = \|XUD^2U^\top - XUU^\top\|_p = \|XUD^2U^\top - X\|_p\,, \tag{60}$$

where in the last step we used the fact that columns of U span the row space of X and hence $UU^\top X^\top = X^\top$. Combining (58), (59) and (60), we obtain

$$\begin{aligned}\|Y_1 - XQ\|_p &\le \|XUD^2U^\top - X\|_p && (61)\\ &= \|(XMM^\top - X)(X^\ddagger X)^\top\|_p && (62)\\ &\le \|X^\ddagger\|\|XMM^\top X^\top - XX^\top\|_p && (63)\\ &= \|X^\ddagger\|\|Y_1Y_1^\top - XX^\top\|_p, && (64)\end{aligned}$$

where the first equality holds since $X^\ddagger X = I$, given that X has full column rank.

Coming from the other end, so to speak, we have

$$\begin{aligned}\varepsilon^2 = \|YY^\top - XX^\top\|_p &= \|Y_1Y_1^\top - XX^\top + Y_1Y_2^\top + Y_2Y_1^\top + Y_2Y_2^\top\|_p && (65)\\ &\ge \|Y_1Y_1^\top - XX^\top + Y_1Y_2^\top\| \vee \|Y_2Y_1^\top + Y_2Y_2^\top\|_p && (66)\\ &\ge \|Y_1Y_1^\top - XX^\top\|_p \vee \|Y_1Y_2^\top\|_p \vee \|Y_2Y_1^\top\|_p \vee \|Y_2Y_2^\top\|_p, && (67)\end{aligned}$$

using Lemma 3 thrice, once based on the fact that

$$(Y_1Y_1^\top - XX^\top + Y_1Y_2^\top)^\top(Y_2Y_1^\top + Y_2Y_2^\top) = \underbrace{(Y_1Y_1^\top - XX^\top + Y_2Y_1^\top)Y_2}_{=0}(Y_1^\top + Y_2^\top) = 0,$$

and then based on the fact that

$$(Y_1Y_1^\top - XX^\top)(Y_1Y_2^\top)^\top = \underbrace{(Y_1Y_1^\top - XX^\top)Y_2}_{=0}Y_1^\top = 0,$$

and

$$(Y_2Y_1^\top)(Y_2Y_2^\top)^\top = Y_2\underbrace{Y_1^\top Y_2}_{=0}Y_2^\top.$$

From (67), we extract the bound $\|Y_1Y_1^\top - XX^\top\|_p \le \varepsilon^2$, from which we get (based on the derivations above)

$$\|Y_1 - XQ\|_p \le \|X^\ddagger\|\varepsilon^2. \tag{68}$$

Recalling the inequality (57), we proceed to bound $\|Y_2\|_p$. From (67), we extract the bound $\|Y_2 Y_2^\top\|_p \le \varepsilon^2$, and combine it with

$$\|Y_2 Y_2^\top\|_p = \|Y_2\|_{2p}^2 \ge d^{-1/p}\|Y_2\|_p^2\,,$$

where d is the number of columns and the inequality is Cauchy-Schwarz's, to get

$$\|Y_2\|_p \le d^{1/2p}\varepsilon\,.$$

We next derive another upper bound for $\|Y_2\|_p$, for the case that $\|X^\ddagger\|\varepsilon < 1$. Denote by $\lambda_1 \ge \ldots \ge \lambda_d$ be the singular values of X and by $\nu_1 \ge \ldots \ge \nu_d$ the singular values of Y_1. Given that X has full column rank we have $\lambda_d > 0$ and so $\|X^\ddagger\| = 1/\lambda_d$. Further, by an application of Mirsky's inequality (Stewart and Sun, 1990, Thm IV.4.11), we have

$$\max_i |\nu_i^2 - \lambda_i^2| \le \|Y_1 Y_1^\top - XX^\top\| \le \|Y_1 Y_1^\top - XX^\top\|_p \le \varepsilon^2,$$

using Equation (67). Therefore $\nu_d^2 > \lambda_d^2 - \varepsilon^2 > 0$ by our assumption that $\|X^\ddagger\|\varepsilon^2 < 1$, which implies that Y_1 has full column rank. Now, by an application of Lemma 1, we obtain

$$\|Y_2\|_p = \|Y_2^\top\|_p \le \|Y_1^\ddagger\|\|Y_1 Y_2^\top\|_p \le \varepsilon^2\|Y_1^\ddagger\|\,, \tag{69}$$

where we used (67) in the last step. Also,

$$\|Y_1^\ddagger\| = \frac{1}{\nu_d} \le \frac{1}{(\lambda_d^2 - \varepsilon^2)^{1/2}} = \frac{\lambda_d^{-1}}{(1 - \varepsilon^2\lambda_d^{-2})^{1/2}} = \|X^\ddagger\|(1 - \varepsilon^2\|X^\ddagger\|^2)^{-1/2}\,. \tag{70}$$

Combining (70) and (69) we obtain

$$\|Y_2\|_p \le \varepsilon^2\|X^\ddagger\|(1 - \varepsilon^2\|X^\ddagger\|^2)^{-1/2}\,, \qquad \text{if} \quad \|X^\ddagger\|\varepsilon < 1\,. \tag{71}$$

Combining the the bounds (69) with (71) and (68) in the inequality (57), we get (4). The bound (5) follows readily from (4).

8.3. Proof of Theorem 2

Let $\bar{a}$ denote the average dissimilarity vector defined in Algorithm 3 based on Y, and define $\bar{b}$ similarly based on Z. Let Θ denote the matrix of dissimilarities between $\tilde{Y}$ and Z, and let $\hat{Y}$ denote the result of Algorithm 3 with inputs Z and Θ. From Algorithm 3, we have

$$\tilde{Y}^\top = \frac{1}{2}Y^\ddagger(\bar{a}1^\top - \tilde{\Delta}^\top), \quad \hat{Y}^\top = \frac{1}{2}Z^\ddagger(\bar{b}1^\top - \Theta^\top), \quad \tilde{Z}^\top = \frac{1}{2}Z^\ddagger(\bar{b}1^\top - \tilde{\Lambda}^\top), \tag{72}$$

due to the fact that the algorithm is exact.

We have

$$\|\tilde{Z} - \tilde{Y}\|_2 \le \|\tilde{Z} - \hat{Y}\|_2 + \|\hat{Y} - \tilde{Y}\|_2. \tag{73}$$

On the one hand,

$$2\|\tilde{Z} - \hat{Y}\|_2 \le \|Z^\ddagger\|\|\tilde{\Lambda} - \Theta\|_2 \le \|Z^\ddagger\|(\|\tilde{\Lambda} - \tilde{\Delta}\|_2 + \|\tilde{\Delta} - \Theta\|_2). \tag{74}$$

On the other hand, starting with the triangle inequality,

$$\begin{aligned}2\|\hat{Y}-\tilde{Y}\|_2 &= \|Z^\ddagger(\bar{b}1^\top-\Theta^\top)-Y^\ddagger(\bar{a}1^\top-\tilde{\Delta}^\top)\|_2\\ &\le \|Z^\ddagger(\bar{b}1^\top-\Theta^\top)-Z^\ddagger(\bar{a}1^\top-\tilde{\Delta}^\top)\|_2+\|Z^\ddagger(\bar{a}1^\top-\tilde{\Delta}^\top)-Y^\ddagger(\bar{a}1^\top-\tilde{\Delta}^\top)\|_2\\ &\le \|Z^\ddagger\|(\|\bar{b}1^\top-\bar{a}1^\top\|_2+\|\Theta-\tilde{\Delta}\|_2)+\|\bar{a}1^\top-\tilde{\Delta}^\top\|\|Z^\ddagger-Y^\ddagger\|_2.\end{aligned}$$

Together, we find that

$$2\|\tilde{Z}-\tilde{Y}\|_2 \le \|Z^\ddagger\|(\|\tilde{\Lambda}-\tilde{\Delta}\|_2+2\|\Theta-\tilde{\Delta}\|_2+\sqrt{m}\|\bar{b}-\bar{a}\|)+\|\bar{a}1^\top-\tilde{\Delta}^\top\|\|Z^\ddagger-Y^\ddagger\|_2. \quad (75)$$

In the following, we bound the terms $\|\bar{a}1^\top-\tilde{\Delta}^\top\|$, $\|\Theta-\tilde{\Delta}\|_2$ and $\|\bar{b}-\bar{a}\|$, separately. First, using Lemma 1 and the fact that $(Y^\ddagger)^\ddagger = Y$ has full rank,

$$\|\tilde{Y}\| = \frac{1}{2}\|Y^\ddagger(\bar{a}1^\top-\tilde{\Delta}^\top)\| \ge \frac{1}{2}\|Y\|^{-1}\|\bar{a}1^\top-\tilde{\Delta}^\top\|. \quad (76)$$

Therefore,

$$\|\bar{a}1^\top-\tilde{\Delta}^\top\| \le 2\|Y\|\|\tilde{Y}\|. \quad (77)$$

Next, set $Y=[y_1,\cdots,y_m]^\top$ and $Z=[z_1,\cdots,z_m]^\top$, as well as $\tilde{Y}=[\tilde{y}_1,\cdots,\tilde{y}_n]^\top$. Since

$$(\Theta-\tilde{\Delta})_{ij} = 2\tilde{y}_i^\top(y_j-z_j)+\|z_j\|^2-\|y_j\|^2, \quad (78)$$

we have

$$\|\Theta-\tilde{\Delta}\|_2 = \|2\tilde{Y}(Y^\top-Z^\top)+1c^\top\|_2 \le 2\|\tilde{Y}\|\|Y-Z\|_2+\sqrt{m}\|c\|, \quad (79)$$

with $c=(c_1,\ldots,c_m)$ and $c_j=\|z_j\|^2-\|y_j\|^2$. Note that

$$\begin{aligned}\|c\|^2 &= \sum_{j\in[m]}(\|z_j\|^2-\|y_j\|^2)^2\\ &\le \sum_{j\in[m]}\|z_j-y_j\|^2(\|z_j\|+\|y_j\|)^2\\ &\le (\rho_\infty(Y)+\rho_\infty(Z))^2\|Z-Y\|_2^2,\end{aligned}$$

so that

$$\|\Theta-\tilde{\Delta}\|_2 \le 2\|\tilde{Y}\|\|Y-Z\|_2+\sqrt{m}(\rho_\infty(Y)+\rho_\infty(Z))\|Z-Y\|_2. \quad (80)$$

Finally, recall that $\bar{a}$ and $\bar{b}$ are respectively the average of the columns of the dissimilarity matrix for the landmark Y and the landmark Z. Using the fact that the y's are centered and that the z's are also centered, we get

$$\bar{b}-\bar{a} = c + c_{\text{avg}}1, \quad (81)$$

where $c_{\text{avg}} = \frac{1}{m}\sum_{j\in[m]}c_j$, and therefore

$$\|\bar{b}-\bar{a}\|^2 \le \sum_{j\in[m]}(c_j+c_{\text{avg}})^2 = \|c\|^2+3mc_{\text{avg}}^2 \le 4\|c\|^2, \quad (82)$$

using the Cauchy-Schawrz inequality at the last step.

Combining all these bounds, we obtain the bound stated in (25). The last part comes from the triangle inequality and an application of Lemma 2.

8.4. Proof of Corollary 2

If the half-width $\omega = 0$, the claim becomes trivial. Hence, we assume $\omega > 0$, which implies that $Y = [y_1 \dots y_m]^\top \in \mathbb{R}^{m\times d}$ is of rank d. Recall that $\nu_d(Y)$ denotes the d-th largest singular value of Y. By characterization (20) and since Y has full column rank, we have $\nu_d(Y) = \sqrt{m}\omega$.

We denote by $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{m\times m}$ and $\Delta = (\delta_{ij}) \in \mathbb{R}^{m\times m}$ and represent the centering matrix of size m, by H. Using (43) and the fact that $\|H\|_\infty = 1$ (since H is an orthogonal projection), we have

$$\varepsilon_0^2 \equiv \tfrac{1}{2}\|H(\Lambda - \Delta)H\| \le \|\Lambda - \Delta\| \le \|\Lambda - \Delta\|_2 = m\eta^2\,. \tag{83}$$

By our assumption $\frac{\eta}{\omega} \le \frac{1}{\sqrt{2}} < 1$, which along with (83) yields

$$\varepsilon_0^2 < m\omega^2 = \nu_d^2(Y)\,. \tag{84}$$

In addition, by (13) and since $Y1 = 0$ (data points are centered), we have $YY^\top = HYY^\top H = -\frac{1}{2}H\Delta H$, and as a result $\nu_d(-\frac{1}{2}H\Delta H) = \nu_d^2(Y)$. By using the Weyl's inequality, we have

$$\nu_d(-\frac{1}{2}H\Lambda H) \ge \nu_d(-\frac{1}{2}H\Delta H) - \varepsilon_0^2 = \nu_d^2(Y) - \varepsilon_0^2 > 0\,,$$

where the last step holds by (84). In words, the first top d eigenvalues of $(-1/2)H\Lambda H$ are positive. Therefore, if $Z = [z_1, \dots, z_m]^\top \in \mathbb{R}^{m\times d}$ is the output of the classical scaling with input Λ, we have that $ZZ^\top$ is indeed the best rank d- approximation of $(-1/2)H\Lambda H$. Given that $(-1/2)H\Delta H$ is of rank d, this implies that

$$\|ZZ^\top + \tfrac{1}{2}H\Lambda H\|_2 \le \|\tfrac{1}{2}H(\Lambda - \Delta)H\|_2\,. \tag{85}$$

Thus, by triangle inequality

$$\begin{aligned}
\varepsilon^2 \equiv \|ZZ^\top - YY^\top\| &\le \|ZZ^\top + \tfrac{1}{2}H\Lambda H\| + \|\tfrac{1}{2}H(\Lambda - \Delta)H\| \\
&\le \|ZZ^\top + \tfrac{1}{2}H\Lambda H\|_2 + \|\tfrac{1}{2}H(\Lambda - \Delta)H\|_2 \\
&\le \|H(\Lambda - \Delta)H\|_2 \\
&\le \|\Lambda - \Delta\|_2 \\
&\le m\eta^2 \le m\omega^2/2\,.
\end{aligned} \tag{86}$$

where in the penultimate line we used (43) and the fact that $\|H\|_\infty = 1$. The last line follows from the definition of η and our assumption on η, given in the theorem statement.

We next apply Theorem 1 with $p = \infty$. Note that by invoking Equations (19) and (20), we get

$$\|Y^\ddagger\|\varepsilon = \frac{\varepsilon}{\sqrt{m}\omega} \le \frac{1}{\sqrt{2}}\,, \tag{87}$$

Hence, by using Theorem 1 we have

$$\begin{aligned}\min_{Q\in\mathcal{O}}\left(\frac{1}{m}\sum_{i\in[m]}\|z_i-Qy_i\|^2\right)^{1/2} &\le \sqrt{\frac{d}{m}}\min_{Q\in\mathcal{O}}\|Z-YQ\| \\ &\le \sqrt{\frac{d}{m}}(\rho/\omega+2)\frac{\varepsilon^2}{\sqrt{m}\omega} \\ &\le \sqrt{d}(\rho/\omega+2)\frac{\eta^2}{\omega} \le \frac{3\sqrt{d}\rho\eta^2}{\omega^2}\,, \end{aligned} \tag{88}$$

where the last line follows from (86) and the fact that $\omega \le \rho$.

8.5. Proof of Corollary 3

We apply Theorem 2 to $\tilde{Y} = [\tilde{y}_1, \cdots, \tilde{y}_n]^\top$, $Y = [y_1, \cdots, y_m]^\top$, and $Z = [z_1, \cdots, z_m]^\top$. To be in the same setting, we need Z to have full rank. As we point out in Remark 5, this is the case as soon as $\|Y^\ddagger\|\|Z-Y\| < 1$. Since $\|Y^\ddagger\| = (\sqrt{m}\omega)^{-1}$ and $\|Z-Y\| \le \sqrt{m}\max_{i\in[m]}\|z_i-y_i\| \le \sqrt{m}\varepsilon$, the condition is equivalent to $\varepsilon < \omega$, which is fulfilled by assumption. Continuing, we have

$$\|Z-Y\|_2 \le \sqrt{m}\max_{i\in[m]}\|z_i-y_i\| \le \sqrt{m}\epsilon. \tag{89}$$

Hence, by (27),

$$\|Z^\ddagger - Y^\ddagger\|_2 \le \frac{\frac{2}{m\omega^2}\sqrt{m}\epsilon}{(1-\frac{1}{\sqrt{m}\omega}\sqrt{m}\epsilon)_+^2} \le \frac{8\varepsilon}{\sqrt{m}\omega^2} \le \frac{4}{\sqrt{m}\omega}\,, \tag{90}$$

using the fact that $\varepsilon/\omega \le 1/2$. Hence,

$$\|Z^\ddagger\| \le \|Y^\ddagger\| + \|Z^\ddagger - Y^\ddagger\| \le \frac{5}{\sqrt{m}\omega}\,, \tag{91}$$

Further, $\|\tilde{\Delta}-\tilde{\Lambda}\|_2 = \sqrt{mn}\eta^2$. In addition, $\|\tilde{Y}\| \le \sqrt{n}\zeta$. Likewise, $\|Y\| \le \sqrt{m}\rho$. Therefore, by applying Theorem 2, we get

$$\begin{aligned}\|\tilde{Z}-\tilde{Y}\|_2 &\le \frac{5}{\sqrt{m}\omega}\Big[\frac{1}{2}\sqrt{nm}\eta^2 + 2\sqrt{nm}\zeta\epsilon + 2\sqrt{m}(2\rho_\infty+\epsilon)\sqrt{m}\epsilon\Big] + (\sqrt{m}\rho)(\sqrt{n}\zeta)\frac{8\varepsilon}{\sqrt{m}\omega^2} \\ &\le 20\left(\frac{\sqrt{n}\eta^2}{\omega} + \frac{\sqrt{n}\zeta\epsilon}{\omega} + \frac{\rho_\infty+\epsilon}{\omega}\sqrt{m}\epsilon + \frac{\sqrt{n}\rho\zeta\varepsilon}{\omega^2}\right)\,, \end{aligned} \tag{92}$$

from which we get the stated bound, using the fact that $\varepsilon \le \omega \le \rho \le \rho_\infty$.

8.6. Proof of Corollary 5

Without loss of generality, suppose the chosen landmark points are $x_1, \ldots, x_\ell$. Using $\{\gamma_{ij} : i, j \in [\ell]\}$, we embed them using classical scaling, obtaining a centered point set $z_1, \ldots, z_\ell \in \mathbb{R}^d$. Note that by our assumption on the number of landmarks $\ell \ge 1$, we have

$$\xi < (72\sqrt{d})^{-1}(\rho/\omega_*)^{-3} < \frac{1}{24}(\rho/\omega_*)^{-2}\,,$$

since $\omega_* \leq \rho$. Hence the assumption on ξ in Corollary 4 holds and by applying this corollary, we have

$$\min_{Q\in\mathcal{O}} \left(\frac{1}{\ell}\sum_{i\in[\ell]} \|z_i - Qy_i\|^2\right)^{1/2} \leq \frac{36\sqrt{d}\rho_*^3}{\omega_*^2}\xi, \tag{93}$$

where ρ_* and ω_* are the max-radius and half-width of $\{y_1, \ldots, y_\ell\}$. We may assume that the minimum above is attained at $Q = I$ without loss of generality, in which case we have

$$\varepsilon \equiv \max_{i\in[\ell]} \|z_i - y_i\| \leq \frac{36\sqrt{d\ell}\rho^3}{\omega_*^2}\xi, \tag{94}$$

using the fact that $\rho_* \leq \rho$.

The next step consists in trilaterizing the remaining points based on the embedded landmarks. With η as in (35), and noting that $\varepsilon/\omega_* \leq 1/2$ by our assumption on ξ, we may apply Corollary 3 (with the constant C_0 defined there) to obtain

$$\begin{aligned}
\frac{1}{C_0}\left(\frac{1}{n-\ell}\sum_{i=\ell+1}^{n} \|\tilde{z}_i - \tilde{y}_i\|^2\right)^{1/2} &\leq \frac{\eta^2}{\omega_*} + \left[\frac{\rho_*\rho}{\omega_*^2} + \frac{\sqrt{\ell}\rho_*}{\sqrt{n-\ell}\,\omega_*}\right]\varepsilon && (95)\\
&\leq \frac{\eta^2}{\omega_*} + \frac{2\rho^2\varepsilon}{\omega_*^2} && (96)\\
&\asymp \frac{\rho^2\xi}{\omega_*} + \frac{\rho^2}{\omega_*^2}\frac{\sqrt{d\ell}\rho^3\xi}{\omega_*^2} && (97)\\
&\asymp \frac{\sqrt{d}\rho^5}{\omega_*^4}\sqrt{\ell}\,\xi\,, && (98)
\end{aligned}$$

using the fact that $\omega_* \leq \rho_*$.

With this and the fact that

$$\left(\frac{1}{\ell}\sum_{i=1}^{\ell} \|z_i - y_i\|^2\right)^{1/2} \leq \varepsilon, \tag{99}$$

along with the bound on ε, we have

$$\min_{Q\in\mathcal{O}} \left(\frac{1}{n}\sum_{i\in[\ell]} \|z_i - Qy_i\|^2 + \frac{1}{n}\sum_{i\in[n-\ell]} \|\tilde{z}_i - Q\tilde{y}_i\|^2\right)^{1/2} \lesssim \frac{\sqrt{d}\rho^5}{\omega_*^4}\sqrt{\ell}\,\xi \asymp \frac{\rho^2}{\omega_*}\,,$$

using our assumption on the number of landmarks.

Appendix

A.1. A succinct proof that Algorithm 3 is correct

To prove that Algorithm 3 is exact, it suffices to do so for the case where we want to position one point, i.e., when $n = 1$, and we denote that point by $\tilde{y}$. In that case, $\tilde{\Delta}$ is in fact a (row) vector, which we denote by $\tilde{\delta}^\top$. We have $\|\tilde{y} - y_i\|^2 = \|\tilde{y}\|^2 + \|y_i\|^2 - 2y_i^\top\tilde{y}$, so that $\delta = \|\tilde{y}\|^2 1 + \zeta - 2Y\tilde{y}$, where $\zeta = (\|y_1\|^2, \ldots, \|y_m\|^2)^\top$. We also have $\|y_j - y_i\|^2 =$

$\|y_j\|^2 + \|y_i\|^2 - 2y_j^\top y_i$, so that $\bar{a} = b1 + \zeta$, where $b = \frac{1}{m}(\|y_1\|^2 + \cdots + \|y_m\|^2)$, using the fact that $\frac{1}{m}\sum_{i=1}^m y_i = 0$. Hence, $\bar{a} - \tilde{\delta} = (b - \|\tilde{y}\|^2)1 + 2Y\tilde{y}$, and therefore,

$$\frac{1}{2}Y^\ddagger(\bar{a} - \tilde{\delta}) = \frac{1}{2}(b - \|\tilde{y}\|^2)Y^\ddagger 1 + Y^\ddagger Y\tilde{y}. \tag{100}$$

We now use the fact that $Y^\ddagger = (Y^\top Y)^{-1}Y^\top$. On the one hand, $Y^\ddagger 1 = (Y^\top Y)^{-1}Y^\top 1 = 0$ since $Y^\top 1 = 0$ (because the point set is centered). On the other hand, $Y^\ddagger Y = (Y^\top Y)^{-1}Y^\top Y = I$. We conclude that $\frac{1}{2}Y^\ddagger(\bar{a} - \tilde{\delta}) = \tilde{y}$, which is what we needed to prove.

A.2. Proof of Proposition 1

The data points are denoted $x_1, \ldots, x_n \in \mathcal{M}$, and by assumption we assume that $x_i = \varphi(y_i)$, where $\varphi : \mathcal{D} \to \mathcal{M}$ is a one-to-one isometry, with $\mathcal{D}$ being a convex subset of $\mathbb{R}^d$. Fix $i, j \in [n]$, and note that $g_{ij} = g_{\mathcal{M}}(x_i, x_j) = \|y_i - y_j\|$.

If $g_{ij} \leq r$, then $\|x_i - x_j\| \leq g_{ij} \leq r$, so that i and j are neighbors in the graph, and in particular $\gamma_{ij} = \|x_i - x_j\|$. We may thus conclude that, in this situation, $\gamma_{ij} \leq g_{ij}$, which implies the stated bound.

Henceforth, we assume that $g_{ij} > r$. Consider $z_k = y_i + (k/m)(y_j - y_i)$, where $m = \lceil 2g_{ij}/r \rceil \geq 2$. Note that $z_0 = y_i$ and $z_m = y_j$. Let y_{i_k} be the closest point to z_k among $\{y_1, \ldots, y_n\}$, with $i_0 = i$ and $i_m = j$. By the triangle inequality, we have

$$\|y_{i_{k+1}} - y_{i_k}\| \leq \|z_{k+1} - z_k\| + \|y_{i_{k+1}} - z_{k+1}\| + \|y_{i_k} - z_k\| \tag{101}$$

$$\leq \frac{1}{m}g_{ij} + 2a \leq r/2 + 2a \leq r, \tag{102}$$

if $a/r \leq 1/4$. Therefore,

$$\|x_{i_{k+1}} - x_{i_k}\| \leq g_{\mathcal{M}}(x_{i_{k+1}}, x_{i_k}) = \|y_{i_{k+1}} - y_{i_k}\| \leq r, \tag{103}$$

implying that $(i_k : k = 0, \ldots, m)$ forms a path in the graph.

So far, the arguments are the same as in the proof of (Bernstein et al., 2000, Thm 2). What makes our arguments sharper is the use of the Pythagoras theorem below. To make use of that theorem, we need to construct a different sequence of points on the line segment. Let $\tilde{z}_k$ denote the orthogonal projection of y_{i_k} onto the line (denoted $\mathcal{L}$) defined by y_i and y_j. See Figure 6 for an illustration.

In particular the vector $\tilde{z}_k - y_{i_k}$ is orthogonal to $\mathcal{L}$, and

$$\|\tilde{z}_k - y_{i_k}\| = \min_{z \in \mathcal{L}} \|z - y_{i_k}\| \leq \|z_k - y_{i_k}\| \leq a. \tag{104}$$

It is not hard to see that $\tilde{z}_k$ is in fact on the line segment defined by y_i and y_j. Moreover, they are located sequentially on that segment. Indeed, using the triangle inequality,

$$\|\tilde{z}_k - y_i\| \leq \|z_k - y_i\| + \|z_k - \tilde{z}_k\| \tag{105}$$

$$\leq \|z_k - y_i\| + \|z_k - y_{i_k}\| + \|y_{i_k} - \tilde{z}_k\| \tag{106}$$

$$\leq \|z_k - y_i\| + 2a \tag{107}$$

$$= \frac{k}{m}g_{ij} + 2a, \tag{108}$$

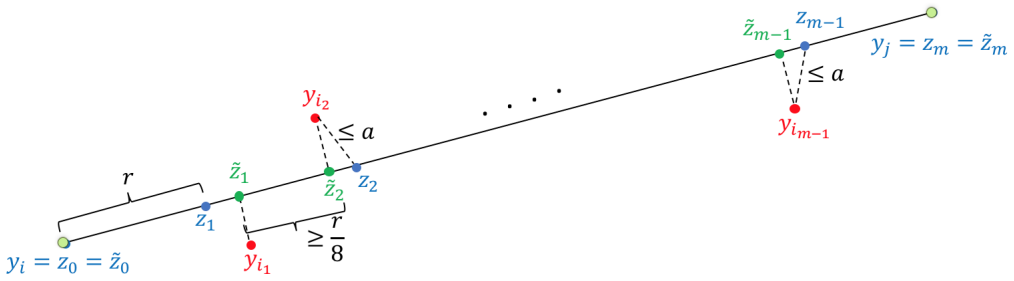


Figure 6: illustration for the proof of Proposition 1

while, similarly,

$$\|\tilde{z}_{k+1} - y_i\| \geq \|z_{k+1} - y_i\| - 2a = \frac{k+1}{m} g_{ij} - 2a, \tag{109}$$

so that $\|\tilde{z}_k - y_i\| < \|\tilde{z}_{k+1} - y_i\|$ as soon as $g_{ij}/m > 4a$. Noting that $g_{ij} > (m-1)r/2$, this condition is met when $a/r \leq (m-1)/8m$. Recalling that $m \geq 2$, it is enough that $a/r \leq 1/16$. From the same derivations, we also get

$$\|\tilde{z}_{k+1} - \tilde{z}_k\| \geq \frac{1}{m} g_{ij} - 4a \geq \frac{(m-1)r}{2m} - 4a \geq r/8, \tag{110}$$

if $a/r \leq 1/32$.

Since $(i_k : k = 0, \ldots, m)$ forms a path in the graph, we have

$$\gamma_{ij} \leq \sum_{k=0}^{m-1} \|x_{i_{k+1}} - x_{i_k}\| \leq \sum_{k=0}^{m-1} \|y_{i_{k+1}} - y_{i_k}\|. \tag{111}$$

By the Pythagoras theorem, we then have

$$\|y_{i_{k+1}} - y_{i_k}\|^2 = \|\tilde{z}_{k+1} - \tilde{z}_k\|^2 + \|y_{i_{k+1}} - \tilde{z}_{k+1} + \tilde{z}_k - y_{i_k}\|^2 \tag{112}$$

$$\leq \|\tilde{z}_{k+1} - \tilde{z}_k\|^2 + (2a)^2, \tag{113}$$

so that, using (110),

$$\|y_{i_{k+1}} - y_{i_k}\| \leq (1 + (2a)^2/(r/8)^2)^{1/2} \|\tilde{z}_{k+1} - \tilde{z}_k\| = (1 + C(a/r)^2)\|\tilde{z}_{k+1} - \tilde{z}_k\|, \tag{114}$$

where $C \leq 128$, yielding

$$\gamma_{ij} \leq (1 + C(a/r)^2) \sum_{k=0}^{m-1} \|\tilde{z}_{k+1} - \tilde{z}_k\| = (1 + C(a/r)^2) g_{ij}. \tag{115}$$

A.3. Proof of Proposition 2

We use concentration bounds for random matrices developed by Tropp (2012). Consider a point set $\mathcal{Y} = \{y_1, \ldots, y_n\}$, assumed centered without loss of generality. We apply Random to

select a subset of ℓ points chosen uniformly at random with replacement from $\mathcal{Y}$. We denote the resulting (random) point set by $\mathcal{Z} = \{z_1, \dots, z_\ell\}$. Let $Y = [y_1 \cdots y_n]$ and $Z = [z_1 \cdots z_\ell]$. We have that $\mathcal{Y}$ has squared half-width equal to $\omega^2 \equiv \nu_d(Y^\top Y)/n$, and similarly, $\mathcal{Z}$ has squared half-width equal to $\omega_Z^2 = \nu_d(Z^\top Z - \ell\, \bar{z}\bar{z}^\top)/\ell$, where $\bar{z} = (z_1 + \cdots + z_\ell)/\ell$. Note that, by (50),

$$\omega_Z^2 \geq \nu_d(Z^\top Z)/\ell - \nu_1(\bar{z}\bar{z}^\top) = \nu_d(Z^\top Z)/\ell - \nu_1(\bar{z})^2 = \nu_d(Z^\top Z)/\ell - \|\bar{z}\|^2. \tag{116}$$

We bound the two terms on the right-hand side separately.

First, we note that $Z^\top Z = \sum_j z_j z_j^\top$, with $z_1 z_1^\top, \dots, z_\ell z_\ell^\top$ sampled independently and uniformly from $\{y_1 y_1^\top, \dots, y_n y_n^\top\}$. These matrices are positive semidefinite, with expectation $Y^\top Y/n$, and have operator norm bounded by $\max_i \|y_i y_i^\top\| = \max_i \|y_i\|^2 = \rho_\infty^2$. We are thus in a position to apply (Tropp, 2012, Thm 1.1, Rem 5.3), which gives that

$$\mathbb{P}\left(\nu_d(Z^\top Z)/\ell \leq \tfrac{1}{2}\omega^2\right) \leq d \exp\big[-\tfrac{1}{8}\ell\omega^2/\rho_\infty^2\big]. \tag{117}$$

Next, we note that $\ell\bar{z} = \sum_j z_j$, with $z_1, \dots, z_n$ being iid uniform in $\{y_1, \dots, y_n\}$. These are here seen as rectangular $d \times 1$ matrices, with expectation 0 (since the y's are centered), and operator norm bounded by $\max_i \|y_i\| = \rho_\infty$. We are thus in a position to apply (Tropp, 2012, Thm 1.6), which gives that, for all $t \geq 0$,

$$\mathbb{P}\left(\|\bar{z}\| \geq t/\ell\right) \leq (d+1) \exp\big[-t^2/(2\sigma^2 + \tfrac{1}{3}\rho_\infty t)\big], \tag{118}$$

where

$$\sigma^2 = (\ell/n)\big(\|Y^\top Y\| \vee \textstyle\sum_i \|y_i\|^2\big) = (\ell/n)\textstyle\sum_i \|y_i\|^2 \leq \ell\rho_\infty^2. \tag{119}$$

In particular,

$$\mathbb{P}\left(\|\bar{z}\| \geq \tfrac{1}{4}\omega^2\right) \leq (d+1) \exp\big[-\tfrac{1}{4}\omega^2\ell^2/(2\rho_\infty\ell + \tfrac{1}{3}\rho_\infty \tfrac{1}{2}\omega\ell)\big] \tag{120}$$

$$\leq (d+1) \exp\big[-\tfrac{1}{9}\ell\omega^2/\rho_\infty^2\big], \tag{121}$$

using in the last line the fact that $\omega \leq \rho_\infty$.

Combining these inequalities using the union bound, we conclude that

$$\mathbb{P}\left(\omega_Z \leq \tfrac{1}{2}\omega\right) \leq d \exp\big[-\tfrac{1}{8}\ell\omega^2/\rho_\infty^2\big] + (d+1) \exp\big[-\tfrac{1}{9}\ell\omega^2/\rho_\infty^2\big], \tag{122}$$

from which the stated result follows.

Acknowledgements

We are grateful to Vin de Silva, Luis Rademacher, and Ilse Ipsen for helpful discussions and pointers to the literature. Part of this work was performed while the first and second authors were visiting the Simons Institute[4] on the campus of the University of California, Berkeley. EAC was partially supported by the National Science Foundation (DMS 0915160, 1513465, 1916071). AJ was partially supported by an Outlier Research in Business (iORB) grant from the USC Marshall School of Business, a Google Faculty Research Award and the NSF CAREER Award DMS-1844481. BP was partially supported by a grant from the French National Research Agency (ANR 09-BLAN-0051-01).

4. The Simons Institute for the Theory of Computing (`https://simons.berkeley.edu`)

References

Ery Arias-Castro and Thibaut Le Gouic. Unconstrained and curvature-constrained shortest-path distances and their approximation. *arXiv preprint arXiv:1706.09441*, 2017.

Ery Arias-Castro and Bruno Pelletier. On the convergence of maximum variance unfolding. *The Journal of Machine Learning Research*, 14(1):1747–1770, 2013.

Mikhail Belkin and Partha Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

M. Bernstein, V. De Silva, J.C. Langford, and J.B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Technical report, Department of Psychology, Stanford University, 2000.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Pratik Biswas, Tzu-Chen Liang, Kim-Chuan Toh, Yinyu Ye, and Ta-Chung Wang. Semidefinite programming approaches for sensor network localization with noisy distance measurements. *Transactions on Automation Science and Engineering*, 3(4):360–371, 2006.

R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

V. de Silva and J.B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in Neural Information Processing Systems (NIPS)*, 15:705–712, 2003.

Vin de Silva and Joshua B Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Technical report, Stanford University, 2004.

David L Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100 (10):5591–5596, 2003.

Christos Faloutsos and King-Ip Lin. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *ACM SIGMOD International Conference on Management of Data*, volume 24, pages 163–174, 1995.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93:418–491, 1959.

Christopher R Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under hausdorff loss. *The Annals of Statistics*, 40(2):941–963, 2012.

Evarist Giné and Vladimir Koltchinskii. Empirical graph Laplacian approximation of laplace–beltrami operators: Large sample results. In *High dimensional probability*, pages 238–259. Institute of Mathematical Statistics, 2006.

Yair Goldberg, Alon Zakai, Dan Kushnir, and Ya'acov Ritov. Manifold learning: The price of normalization. *Journal of Machine Learning Research*, 9(Aug):1909–1939, 2008.

John C Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.

Matthias Hein, Jean-Yves Audibert, and Ulrike Von Luxburg. From graphs to manifolds: Weak and strong pointwise consistency of graph Laplacians. In *Conference on Computational Learning Theory (COLT)*, pages 470–485. Springer, 2005.

John T Holodnak and Ilse CF Ipsen. Randomized approximation of the gram matrix: Exact computation and probabilistic bounds. *SIAM Journal on Matrix Analysis and Applications*, 36(1):110–137, 2015.

Ilse CF Ipsen and Thomas Wentworth. The effect of coherence on sampling from matrices with orthonormal columns, and preconditioned least squares problems. *SIAM Journal on Matrix Analysis and Applications*, 35(4):1490–1520, 2014.

Adel Javanmard and Andrea Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, 2013.

Arlene KH Kim and Harrison H Zhou. Tight minimax rates for manifold estimation under hausdorff loss. *Electronic Journal of Statistics*, 9(1):1562–1582, 2015.

Joseph B Kruskal and Judith B Seery. Designing network diagrams. In *General Conference on Social Graphics*, pages 22–50, 1980.

Dragoş Niculescu and Badri Nath. DV based positioning in ad hoc networks. *Telecommunication Systems*, 22(1-4):267–280, 2003.

Alexander Paprotny and Jochen Garcke. On a connection between maximum variance unfolding, shortest path problems and isomap. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 859–867, 2012.

John Platt. Fastmap, MetricMap, and Landmark MDS are all Nystrom algorithms. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2005.

George AF Seber. *Multivariate observations*. John Wiley & Sons, 2004.

Yi Shang, Wheeler Ruml, Ying Zhang, and Markus PJ Fromherz. Localization from mere connectivity. In *Symposium on Mobile Ad Hoc Networking & Computing*, pages 201–212, 2003.

Robin Sibson. Studies in the robustness of multidimensional scaling: Perturbational analysis of classical scaling. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 217–229, 1979.

A. Singer. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

A.K. Smith, X. Huo, and H. Zha. Convergence and rate of convergence of a manifold-based dimension reduction algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1529–1536, 2008.

Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for sensor network localization. *Mathematical Programming*, 109(2-3):367–384, 2007.

Inge Söderkvist. Perturbation analysis of the orthogonal procrustes problem. *BIT Numerical Mathematics*, 33(4):687–694, 1993.

G. W. Stewart and Ji Guang Sun. *Matrix perturbation theory*. Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1990. ISBN 0-12-670230-6.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Warren S Torgerson. *Theory and methods of scaling*. Wiley, 1958.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

U. von Luxburg, M. Belkin, and O. Bousquet. Consistency of spectral clustering. *The Annals of Statistics*, 36(2):555–586, 2008.

Jason Tsong-Li Wang, Xiong Wang, King-Ip Lin, Dennis Shasha, Bruce A Shapiro, and Kaizhong Zhang. Evaluating a class of distance-mapping algorithms for data mining and clustering. In *Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 307–311, 1999.

GA Watson. The solution of orthogonal procrustes problems for a family of orthogonally invariant norms. *Advances in Computational Mathematics*, 2(4):393–405, 1994.

Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006a.

Kilian Q Weinberger, Fei Sha, Qihui Zhu, and Lawrence K Saul. Graph Laplacian regularization for large-scale semidefinite programming. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1489–1496, 2006.

Killan Q. Weinberger and Lawrence K. Saul. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. In *Conference on Artificial Intelligence*, volume 2, pages 1683–1686. AAAI, 2006b.

Qiang Ye and Weifeng Zhi. Discrete hessian eigenmaps method for dimensionality reduction. *Journal of Computational and Applied Mathematics*, 278:197–212, 2015.

Forrest W Young. *Multidimensional scaling: History, theory, and applications*. Psychology Press, 2013.

H. Zha and Z. Zhang. Continuum isomap for manifold learnings. *Computational Statistics & Data Analysis*, 52(1):184–200, 2007.

H. Zha and Z. Zhang. Spectral properties of the alignment matrices in manifold learning. *SIAM Review*, 51(3):545, 2009.