Electronic Journal of Statistics
Vol. 14 (2020) 2362–2394
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1720

Central limit theorems for classical multidimensional scaling

Gongkai Li

Department of Applied Mathematics and Statistics
Johns Hopkins University
e-mail: ligkpercy@gmail.com

Minh Tang

Department of Statistics
North Carolina State University
e-mail: mtang8@ncsu.edu

Nicolas Charon and Carey Priebe

Department of Applied Mathematics and Statistics
Johns Hopkins University
e-mail: charon@cis.jhu.edu; cep@jhu.edu

Abstract: Classical multidimensional scaling is a widely used method in dimensionality reduction and manifold learning. The method takes in a dissimilarity matrix and outputs a low-dimensional configuration matrix based on a spectral decomposition. In this paper, we present three noise models and analyze the resulting configuration matrices, or embeddings. In particular, we show that under each of the three noise models the resulting embedding gives rise to a central limit theorem. We also provide compelling simulations and real data illustrations of these central limit theorems. This perturbation analysis represents a significant advancement over previous results regarding classical multidimensional scaling behavior under randomness.

AMS 2000 subject classifications: Primary 62H12; secondary 62H30, 62B10.
Keywords and phrases: Classical multidimensional scaling, dissimilarity matrix, perturbation analysis, central limit theorem.

Received March 2020.

1. Background and overview

Inference based on dissimilarities is of fundamental importance in statistics, data mining and machine learning (Pekalska and Duin, 2005), with applications ranging from neuroscience (Vogelstein et al., 2014) to psychology (Carroll and Chang, 1970) and economics (Machado and Mata, 2015). In each of these fields, rather than directly observing the feature values of the objects, often we observe only the dissimilarities or "distances" between pairs of objects (inter-point distances). A common approach to dimensionality reduction and subsequent inference problems involving dissimilarities is to embed the observed distances into some (usually Euclidean) space to recover a configuration that faithfully preserves observed distances, and then proceed to perform inference based on the resulting configuration (de Leeuw and Heiser, 1982; Borg and Groenen, 2005; Torgerson, 1952; Cox and Cox, 2008). The popular classical multidimensional scaling (CMDS) method provides an example of such an embedding scheme into Euclidean space, in which we have readily available tools to perform statistical inference. CMDS can also be regarded as a powerful dimension reduction technique for high dimensional data. Indeed, the ubiquitous PCA (a linear dimension reduction technique), is equivalent to CMDS on a matrix of pairwise Euclidean distance between feature vectors. CMDS also forms the basis for several recent and popular approaches to nonlinear dimension reduction and manifold learning (Schölkopf et al., 1998; Chen and Buja, 2009), such as Isomap (Tenenbaum et al., 2000), Laplacian eigenmaps (Belkin and Niyogi, 2003), diffusion maps (Coifman and Lafon, 2006), locally linear embedding Roweis and Saul (2000), and random forest manifold learning (Criminisi and Shotton, 2013). These procedures can be formulated as kernelized variants of classical PCA (see e.g., Ham et al. (2004)) and thus correspond to CMDS embedding of Euclidean distances in some high or infinite dimensional vector spaces.

To summarize, classical multidimensional scaling is the problem of, given an $n \times n$ hollow symmetric dissimilarity matrix D and an embedding dimension d, find a $n \times d$ matrix $X \in \mathbb{R}^{n \times d}$ where the rows $X_1, X_2, \dots, X_n \in \mathbb{R}^d$ of X represent coordinates of points in $\mathbb{R}^d$ such that the overall inter-point distances between X_i and X_j are "as close as possible" to the entries of D. The specific steps are as follows.

1. Compute the matrix $B = -\frac{1}{2}(I - \frac{11^\top}{n})D^2(I - \frac{11^\top}{n})$, where D^2 is obtained by *element-wise squaring* the entries of D. The matrix B is termed the *double centering* of D^2. Here I denotes the $n \times n$ identity matrix and $11^\top$ denote the matrix of all ones.
2. Extract the d largest *positive* eigenvalues $s_1, \dots, s_d$ of B and the corresponding eigenvectors $u_1, \dots, u_d$.
3. Let $X = U_B S_B^{1/2} \in \mathbb{R}^{n \times d}$, where $U_B = (u_1, \dots, u_d)$ is a $n \times d$ matrix and $S_B = \operatorname{diag}(s_1, \dots, s_d)$ is a diagonal $d \times d$ matrix. Each row of X represents the coordinate of a point in $\mathbb{R}^d$ so that $\|X_i - X_j\| \approx D_{ij}$ where D_{ij} is the ij-th entry of D. We shall refer to X as the configuration matrix or the embedding configuration of D into $\mathbb{R}^d$.

In essence, CMDS minimizes the Strain loss function defined as $L(X) := \|XX^\top - B\|_F$ where $\| \cdot \|_F$ denote the Frobenius norm of a matrix. Furthermore, the resulting configuration X centers all points around the origin and is unique only up to an orthogonal transformation, i.e., for any configuration X, the configuration XW with W a $d \times d$ orthogonal matrix yields the same interpoint distances as X.

CMDS can be applied whenever the notion of dissimilarity come into play, be it the difference between two time series (Vogelstein et al., 2014) or, in psychometric applications, the difference between two people's perception of the same object (Jaworska and Chupetlovska-Anastasova, 2009). See also the books Pekalska and Duin (2005); Cox and Cox (2010); Borg and Groenen (2005) and the references therein. Note that as the entries of D represents dissimilarities, they do not need to satisfy the triangle inequality, i.e., we can have $D_{ij} > D_{ik} + D_{kj}$. As we allude to earlier, by using CMDS one obtains a Euclidean representation of the data even when the true representation is unknown or possibly ill-defined and/or complex and high-dimensional. The Euclidean representation then allows for the use of a large and extremely diverse suite of classical, robust, and efficient inference methodologies. However, in most of the real world applications, the measurement of distance or dissimilarity is greatly affected by a wide range of factors, such as instrument precision or faulty sensors, which in turn introduce randomness (noise) into the "observed" distance matrix. While CMDS is widely used in these settings, its behaviour under randomness remains largely unexplored. Several recent papers have highlighted this omission. Zhang et al. (2016) write "Despite the popularity of multi-dimensional scaling, very little is known about to what extent the distances between the embedded points could faithfully reflect the true pairwise distances when observed with noise."; Fan et al. (2018) write "[W]e are not aware of any statistical results measuring the performance of MDS under randomness, such as perturbation analysis when the objects are sampled from a probabilistic model." and Peterfreund and Gavish (2018) write "To the best of our knowledge, the literature does not offer a systematic treatment on the influence of ambient noise on MDS embedding quality." The current paper addresses this acknowledged gap in the literature.

2. Noise model and embedding

In this section, we propose three different but related noise models for the matrix of observed dissimilarities. Suppose we have inter-point distances of n points in $\mathbb{R}^d$, and the resulting distance matrix is given by $D \in \mathbb{R}^{n\times n}$, i.e. $D_{ij} = \|X_i - X_j\|$. Let D^2 denote the element-wise squaring of the entries of D and Δ or Δ^2 (the element-wise squaring of the entries of Δ) be the *observed* dissimilarity matrix (such as measured via a scientific experiment). We consider three error models for Δ or Δ^2.

2.1. Model 1: $\boldsymbol{\Delta^2 = D^2 + E}$

An error model proposed in R.Sibson (1979) and Zhang et al. (2016) is $\Delta^2 = D^2 + E$, where we think of D^2 as the *true but unobserved* "signal" matrix and E as the "noise". We shall assume that E satisfies the following conditions:

(i) $\mathbb{E}[E] = 0$, hence $\mathbb{E}[\Delta^2] = D^2$.
(ii) E is hollow and symmetric.
(iii) The entries E_{ij} for $i < j$ are independent and $\mathrm{Var}(E_{ij}) = \sigma^2$ for all i, j.
(iv) Each E_{ij}, for $i < j$, follows a sub-Gaussian distribution.

One possible criticism of this noise model is that some of the entries of Δ^2 could be negative. Since the entries of Δ^2 have interpretations as squared dissimilarities, it is more desirable that they are all non-negative. However, as we will see later, the CMDS embedding of Δ^2 is well-defined and valid even when Δ^2 have negative entries as it is based on the truncated eigendecomposition of Δ^2, a symmetric matrix.

2.2. Model 2: $\boldsymbol{\Delta = |D + E|}$

Another realistic error model is $\Delta = |D + E|$ where the absolute value is taken *element wise*. In this model the noise is added directly onto the distance $D_{ij} = \|X_i - X_j\|$ as opposed to Model 1 in which the noise is added onto the squared distance. For this model we will require that the random matrix E satisfies conditions (i) to (iv) in Section 2.1 along with the following constant third and fourth moment conditions, i.e.,

(v) $\mathbb{E}[E_{ij}^3] \equiv \gamma$ for all i, j.
(vi) $\mathbb{E}[E_{ij}^4] \equiv \xi$ for all i, j.

We emphasize that, under this noise model, the ij-th entry of Δ is $\Delta_{ij} = |D_{ij} + E_{ij}|$ and are guaranteed to be non-negative. Thus, in contrast to Model 1, all of the entries of Δ are proper dissimilarities. Meanwhile the ij-th entry of Δ^2 is $\Delta_{ij}^2 = D_{ij}^2 + 2E_{ij}D_{ij} + E_{ij}^2$ and hence, taking $\tilde{E}_{ij} = 2E_{ij}D_{ij} + E_{ij}$, the noise perturbations $\tilde{E}_{ij}$ on the entries of Δ^2 satisfy

1. $\mathbb{E}[\tilde{E}_{ij}] = \sigma^2 > 0$
2. $\mathrm{Var}[\tilde{E}_{ij}] = \sigma^2 D_{ij}^2 + 2\gamma D_{ij} + \xi$ and thus the $\tilde{E}_{ij}$ does not have constant variances.
3. The $\tilde{E}_{ij} = E_{ij}^2$ are sub-exponential, as opposed to sub-Gaussian, random variables.

Since CMDS is computed via the eigendecomposition of $B = -\frac{1}{2}(I - \frac{11^\top}{n})\Delta^2(I - \frac{11^\top}{n})$, the non-constant variance and the sub-exponential of the noise $\tilde{E}_{ij}$ for Δ^2 lead to quite different finite-sample and limit results for Model 2, when compared to Model 1. We illustrate these differences in Section 3 and Section 4.1.

2.3. Model 3: Matrix completion

The third noise model is related to the problem of recovering a true distance matrix from a noisy and *partially observed subset* of its entries, see e.g., Javanmard and Montanari (2013); Chatterjee (2015). Restricting our attention to the Euclidean distance, we propose the following matrix completion model:

- With probability q we observe $\Delta_{ij} = \Delta_{ji} = D_{ij} = D_{ji}$. Here Δ_{ij} is the ij-th entry of Δ.
- With probability $1 - q$, both Δ_{ij} and Δ_{ji} is missing (in which case we can set $\Delta_{ij} = 0$).

The above model can be rewritten in the form $\Delta = D + E$ where E_{ij} is a Bernoulli random variable which takes value $-D_{ij}$ with probability $1 - q$ and takes value 0 with probability q. It is easy to see that $\mathbb{E}[\Delta] = q \cdot D$ and $\mathbb{E}[\Delta^2] = q \cdot D^2$.

For each of the above noise models, we apply CMDS to the $n \times n$ matrix Δ to get the resulting configuration matrix $\hat{X}$, and use the following notations for this procedure:

1. Let $\hat{B} = -\frac{1}{2}(I - 11^\top/n)\Delta^2(I - 11^\top/n)$.
2. Let $S_{\hat{B}} \in \mathbb{R}^{d\times d}$ be the diagonal matrix of d largest eigenvalues of $\hat{B}$ and $U_{\hat{B}} \in \mathbb{R}^{n\times d}$ be the matrix whose orthogonal columns are the corresponding eigenvectors.
3. The matrix $\hat{X} = U_{\hat{B}} S_{\hat{B}}^{1/2} \in \mathbb{R}^{n\times d}$ is the "embedding of Δ" into $\mathbb{R}^d$, i.e., the ith row of the $n \times d$ matrix $\hat{X}$ yield the coordinates of the point $\hat{X}_i$ such that $D_{ij} \approx \|\hat{X}_i - \hat{X}_j\|$.

A natural question arises regarding how the added noise affects the embedding configuration. That is, what is the relationship between the embedding X from D as in Section 1 and the embedding $\hat{X}$ from Δ? We emphasize that the goal is to recover the $\{X_i\}$ and thus it is generally not the case that $\|\hat{X}_i - \hat{X}_j\| \approx \Delta_{ij}$ but rather that $\|\hat{X}_i - \hat{X}_j\| \approx D_{ij}$.

Finally we remark that the assumption that the missing entries of Δ are set to 0 is an arbitrary choice that is made for ease of exposition. This choice is firstly quite common in the literature, see e.g., Chatterjee (2015, Section 2.3) and Javanmard and Montanari (2013), and secondly, the limit results in Section 3 still hold when we set the missing entries to any other *fixed* but finite value C. Indeed, suppose we set the missing entries to some fixed value $C > 0$. Then $\mathbb{E}[\Delta] = qD + (1-q)C \times 11^\top$ and $\mathbb{E}[\Delta^2] = qD^2 + (1-q)C^2 \times 11^\top$ where $11^\top$ is the $n \times n$ matrix of all ones. The double centering of Δ^2 satisfies

$$\mathbb{E}[-0.5 \times (I - 11^\top/n)\Delta^2(I - 11^\top/n)] = -\frac{q}{2}(I - 11^\top/n)D^2(I - 11^\top/n)$$

which does not depend on C, i.e., the choice of C does not matter in the subsequent theoretical analysis. Finally, even if we set the missing entries to NA, most matrix completion algorithm will first initialize/replace these NA's with some arbitrary value.

2.4. Related works

The problem of recovering an Euclidean distance matrix from noisy or imperfect observations of pairwise dissimilarity scores arises naturally in many different contexts. For example, in Zhang et al. (2016), the authors considered the model $\Delta^2 = D^2 + E$ and studied the behaviour of the estimator

$$\hat{D}^2 := \underset{M\in\mathcal{D}_n^{(2)}}{\arg\min}\Big\{\|\Delta^2 - M\|_F^2 - \lambda_n \operatorname{trace}\Big(I - \frac{11^\top}{n}\Big)M\Big(I - \frac{11^\top}{n}\Big)\Big\}$$

for D^2. Here $\mathcal{D}_n^2$ is the set of $n \times n$ *squared* Euclidean distance matrix and λ_n is a tuning parameter. In particular, Corollary 6 in Zhang et al. (2016) states that if the noise perturbation E satisfy the assumptions in Section 2.1 then, with probability approaching to one, we have

$$\|\hat{D}^2 - D^2\|_F^2 \leq 36n\sigma^2(r+1) \tag{1}$$

where $\sigma^2 = \mathbb{E}[E_{ij}^2]$ is the variance of the noise and r is the rank of D^2. In this paper we obtain, as a corollary of our results, a more refined bound for $\hat{D}^2 - D^2$ in terms of the uniform error $\max_{ij} |D_{ij}^2 - \hat{D}_{ij}^2|$ (see Remark 3 below). Furthermore, our central limit theorem on the configuration matrix is also a more refined limiting result, albeit of a slightly different flavor, when compared to Eq. (1). More specifically, for the noise model of Zhang et al. (2016), our Theorem 3.1 indicates that the *marginal* distribution of the i-th row of the CMDS embedding is, up to some orthogonal transformation, normally distributed around the true but unknown latent positions, and that statistical inference using the rows of the CMDS embedding can proceed as if they were independent multivariate normal random vectors centered around the *true but unknown* latent positions.

The problem of completing a distance matrix with missing entries is also a popular problem in the engineering and social sciences; see, for example, Alfakih et al. (1999); Bakonyi and Johnson (1995); Singer (2008); Spence and Domoney (1974) and distance matrix completion is closely related to multidimensional scaling (Borg and Groenen, 2005; Chatterjee, 2015; Javanmard and Montanari, 2013; Oh et al., 2010). Especially noteworthy is Theorem 2.5 of Chatterjee (2015) which gives an upper bound for the mean squared error for recovering a general, not necessarily Euclidean, distance matrix M. More specifically, let (K, d) be a compact metric space and $x_1, x_2, \dots, x_n$ be n arbitrary points in K. Let M be the $n \times n$ matrix whose ij-th entry is $d(x_i, x_j)$. Let $\epsilon > 0$ be such that $q \geq n^{-1+\epsilon}$. Recall that q is the *proportion* of observed entries of M. For a given $\delta > 0$, let $N(\delta)$ be the covering number of K using balls of radius δ with respect to the metric d. Then there exists an estimator $\tilde{M}$ obtained by truncating the singular value decomposition of M such that

$$\mathrm{MSE}(\tilde{M}) \leq C \inf_{\delta > 0} \min\Bigl\{\frac{\delta + \sqrt{N(\delta/4)/n}}{\sqrt{q}}, 1\Bigr\} + C(\epsilon)e^{-ncq}$$

where c and C are constants depending on the truncation of the singular values of M and $C(\epsilon)$ is a constant depending only on ϵ. When M is a Euclidean distance matrix the above bound yields

$$\mathrm{MSE}(\tilde{M}) \leq \frac{Cn^{-1/3}}{\sqrt{q}}.$$

This result can be improved, e.g., Theorem 1 in Taghizadeh et al. (2015) states that, with high probability, $\mathrm{MSE}(\tilde{M}) = \mathcal{O}((nq)^{-1})$. Theorem 3.3 in the current paper implies the same bound and furthermore, also yields more refined limit results for the rows of the embedding configuration $\hat{X}$ as well as bounds for the maximum entry-wise different for the recovered Euclidean distance matrix (see Remark 3).

Finally, we remark that the Euclidean distance matrix completion problem can also be formulated as a problem of minimizing the nuclear norm of a matrix M subject to the constraint that the non-missing entries of Δ and the corresponding entries of M are equal. More specifically, let S be the set of non-missing entries in Δ. Then one can consider solving

$$\min \|M\|_* \qquad \text{subject to } \Delta_{ij} = M_{ij} \text{ for all } i, j \in S.$$

Here $\|\cdot\|_*$ denote the nuclear norm. This is a convex optimization problem and, under certain regularity conditions on the *coherence* of the true distance matrix D, the solution is unique and, with high probability, equal to D, provided that the number of non-missing entries is of order $\Omega(nd\nu \log n)$ where ν is the coherence of the matrix D (see Theorem 1 in Tasissa and Lai (2018) for a more detailed statement). While these type of results are certainly powerful, they nevertheless depends the coherence of the matrix D which could be as large as $\Theta(n)$ in the worst case. Our results for the distance matrix completion setting do not depend on the coherence of D, and thus do not guarantee that the recovered distances are *exactly equal* to D.

3. Main results

Recall that a random variable ζ is sub-Gaussian if there exists a constant $K > 0$ such that $\mathbb{P}[|\zeta| > t] \leq 2e^{-\frac{t^2}{K^2}}$ for all $t \geq 0$. The Orlicz ψ_2 norm of ζ is defined by $\|\zeta\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(\frac{\zeta^2}{t^2}) \leq 2\}$. A random vector Y in $\mathbb{R}^n$ is called sub-Gaussian if the one-dimensional marginals $\langle Y, v\rangle$ are sub-Gaussian for all $v \in \mathbb{R}^n$; the corresponding Orlicz ψ_2 norm of Y is defined as $\|Y\|_{\psi_2} = \sup_{y \in S^{n-1}} \|\langle Y, v\rangle\|_{\psi_2}$.

3.1. Main theorems

We now present central limit theorems for the CMDS embedding $\hat{X}$ for the three noise models in Section 2. Recall that, given a $n \times n$ dissimilarity matrix Δ, CMDS outputs a $n \times d$ matrix $\hat{X}$ with $d \ll n$ such that the rows of $\hat{X}$ represent the embedding coordinates in $\mathbb{R}^d$ of the rows of Δ. Intuitively speaking, the following theorems established that the rows of $\hat{X}$, after some suitable orthogonal transformation, is approximately normally distributed around the rows of the true X. The covariance matrices of the rows of $\hat{X}$ will depend on the noise model and the true distribution of the points in the underlying space and are substantially different between the three noise models considered. In particular, the covariance matrix for the noise model $\Delta^2 = D^2 + E$ in Theorem 3.1 depends only on the variance σ^2 of the noise E_{ij}. This is in contrast with the covariance matrices of the model $\Delta = |D + E|$ in Theorem 3.2 and the model $\mathbb{E}[\Delta] = qD$

in Theorem 3.3, both of which depend also on the underlying true distances D_{ij}. The machinery involved in proving these results are by and large the same and we refer the reader to the Appendix for detailed proofs. Finally, for ease of exposition, we denote by $(A)_i$ the i-th row of a matrix, and for vectors $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$, $\alpha \leq \beta$ denote that each entry of α is less than the corresponding entry of β.

Theorem 3.1 (Central Limit Theorem for CMDS of $\Delta^2 = D^2 + E$)**.** *Let $Z_1, Z_2, \dots, Z_n \overset{i.i.d.}{\sim} F$ for some sub-Gaussian distribution F on $\mathbb{R}^d$. Let D be the Euclidean distance matrix generated by the Z_k's, i.e. $D_{ij} = \|Z_i - Z_j\|$, and suppose that $\max_i \sum_j D_{ij}^2 \gg \log^4 n$. Let $\Delta^2 = D^2 + E$ where the noise matrix E satisfy the conditions in Section 2.1, i.e, (i) $\mathbb{E}[E] = \mathbf{0}$, (ii) E is hollow and symmetric, (iii) the entries E_{ij} are independent for $i \leq j$ with $\mathrm{Var}[E_{ij}] \equiv \sigma^2$, and (iv) each E_{ij} follows a sub-Gaussian distribution. Note that the E_{ij} need not be identically distributed. Denote by $\hat{X}^{(n)}$ the $n \times d$ matrix representing the CMDS embedding of Δ into $\mathbb{R}^d$. Then there exists a sequence of $d \times d$ orthogonal matrices $\{W^{(n)}\}_{n=1}^{\infty}$ such that for any $\alpha \in \mathbb{R}^d$ and any fixed row index i, we have*

$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}[(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})] \leq \alpha\} = \Phi(\alpha, \Sigma)$$

where $\bar{Z} = n^{-1} \sum_k Z_k$ and $\Phi(\alpha, \Sigma)$ denotes the CDF of a multivariate Gaussian with mean 0 and covariance matrix Σ, evaluated at α. Here $\Sigma = \frac{\sigma^2}{4} \Xi^{-1}$ where $\Xi = \mathrm{Cov}(Z_k) \in \mathbb{R}^{d\times d}$.

Remark 1. We can relax the common variance requirement (iii) in Theorem 3.1. Let $\mathrm{Var}(E_{ij}) = \sigma_{ij}^2$ and suppose that, for each fixed i, the collection $\{(D_{ij}^2 - \Delta_{ij}^2)(Z_j - \mu_z)\}_{j=1}^n$ satisfy the multivariate Lindeberg-Feller condition. Define, for each fixed i, $\Sigma_i = \frac{1}{n} \sum_j \sigma_{ij}^2 \mathrm{Cov}(Z_k)$. We then obtain the following variant of Theorem 3.1:

$$\lim_{n\to\infty} n^{1/2} \Sigma_i^{-\frac{1}{2}} [(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})] \longrightarrow \mathcal{N}(0, I).$$

Theorem 3.2 (Central Limit Theorem for CMDS of $\Delta = |D + E|$)**.** *Let $Z_1, Z_2, \dots, Z_n \overset{i.i.d.}{\sim} F$ for some sub-Gaussian distribution F on $\mathbb{R}^d$. Let D be the Euclidean distance matrix generated by the Z_k's, i.e. $D_{ij} = \|Z_i - Z_j\|$ and suppose that $\max_i \sum_j D_{ij}^2 \gg \log^4 n$. Let $\Delta = |D + E|$ and suppose that the noise matrix E satisfy, in addition to the conditions in Theorem 3.1, the condition (v) $\mathbb{E}[E_{ij}^3] \equiv \gamma$ and $\mathbb{E}[E_{ij}^4] \equiv \xi$. Denote by $\hat{X}^{(n)}$ the $n \times d$ matrix representing the CMDS embedding of Δ into $\mathbb{R}^d$. Then there exists a sequence of $d \times d$ orthogonal matrices $\{W^{(n)}\}_{n=1}^{\infty}$ such that for any $\alpha \in \mathbb{R}^d$ and any fixed row index i,*

$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}[(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})] \leq \alpha\} = \int_{\mathrm{supp}(F)} \Phi(\alpha, \Sigma(\mathrm{z}))\, dF(\mathrm{z})$$

where $\bar{Z}$ is the mean of Z_k's and $\Phi(\alpha, \Sigma)$ denotes the CDF of a multivariate Gaussian with mean 0 and covariance matrix Σ, evaluated at α. Here $\Sigma(\mathrm{z}) =$

$\Xi^{-1}\widetilde{\Sigma}(\mathrm{z})\Xi^{-1}$ *where* $\Xi := \mathrm{Cov}(Z_i) \in \mathbb{R}^{d\times d}$ *and, with* $\mu_z = \mathbb{E}[Z_i] \in \mathbb{R}^d$,

$$\widetilde{\Sigma}(z) := \mathbb{E}_{Z_k}\Big[\big(\sigma^2\|z - Z_k\|^2 + \gamma\|z - Z_k\| + \frac{1}{4}(\xi - \sigma^4)\big)(Z_k - \mu_z)(Z_k - \mu_z)^\top\Big] \quad (2)$$

is a covariance matrix depending on z.

Theorem 3.3 (Central Limit Theorem for CMDS of $\Delta = D$ with missing entries)**.** *Let* $Z_1, Z_2, \dots, Z_n \overset{i.i.d.}{\sim} F$ *for some sub-Gaussian distribution* F *on* $\mathbb{R}^d$. *Let* D *be the Euclidean distance matrix generated by the* Z_i*'s, i.e.* $D_{ij} = \|Z_i - Z_j\|$. *Suppose that with probability* $q_n \in [0, 1]$ *we observe the distance* D_{ij} *and with probability* $1 - q_n$ *it is missing, i.e.,* $\Delta = D + E$ *where* $E_{ij} = (-D_{ij}) \times \mathrm{Bernoulli}(1 - q_n)$. *Denote by* $\hat{X}^{(n)}$ *the* $n \times d$ *CMDS embedding of* Δ *into* $\mathbb{R}^d$. *Then there exists a sequence of* $d \times d$ *orthogonal matrices* $\{W^{(n)}\}_{n=1}^{\infty}$ *such that if* $nq_n = \omega(\log^4 n)$, *then for any* $\alpha \in \mathbb{R}^d$ *and any fixed row index* i,

$$\lim_{n\to\infty} \mathbb{P}\{\sqrt{n}[(\hat{X}^{(n)}W^{(n)})_i - \sqrt{q_n}(Z_i - \bar{Z})] \le \alpha\} = \int_{\mathrm{supp}(F)} \Phi(\alpha, \Sigma(\mathrm{z}))dF(\mathrm{z})$$

where $\bar{Z}$ *is the mean of* Z_i*'s and* $\Phi(\alpha, \Sigma)$ *denotes the CDF of a multivariate Gaussian with mean* 0 *and covariance matrix* Σ, *evaluated at* α. *Here* $\Sigma(\mathrm{z}) = \Xi^{-1}\widetilde{\Sigma}(\mathrm{z})\Xi^{-1}$, $\Xi := \mathrm{Cov}(Z_i) \in \mathbb{R}^{d\times d}$ *and with* $\mu_z = \mathbb{E}[Z_i] \in \mathbb{R}^d$,

$$\widetilde{\Sigma}(z) := \frac{1}{4}(1 - q_n)\mathbb{E}\Big[\|z - Z_k\|^4(Z_k - \mu_z)(Z_k - \mu_z)^\top\Big] \quad (3)$$

is a covariance matrix depending on z.

Remark 2. As we alluded to earlier, the covariance matrix in Theorem 3.1 depends on the (common) variance of the noise E_{ij} and the covariance matrix of the latent positions Z_k. In contrasts, the covariance matrix in Theorem 3.2 depends on the third and fourth moment of the noise E_{ij} as well as $\|z - Z_k\|$ and $\|z - Z_k\|^2$. We note that, in the case when the E_{ij} are symmetric, $\gamma = \mathbb{E}[E_{ij}^3] = 0$ and the covariance matrix in Eq. (2) simplifies to

$$\widetilde{\Sigma}(z) := \mathbb{E}_{Z_k}\Big[\big(\sigma^2\|z - Z_k\|^2 + \frac{1}{4}(\xi - \sigma^4)\big)(Z_k - \mu_z)(Z_k - \mu_z)^\top\Big].$$

In general, as the covariance matrix in Theorem 3.2 depends on $\|z - Z_k\|$, any point z that is considered an “outlier” will be associated with a covariance matrix $\tilde{\Sigma}(z)$ that is substantially larger (in the positive semidefinite ordering), then points z that are “close” to the center μ_z. The rows of the embedding $\hat{X}$ associated with the outliers will thus be much more noisy than the non-outliers points. This phenomenon is even more pronounced in the noisy distance completion setting of Theorem 3.3; indeed, the covariance matrix in Theorem 3.3 now depends on $\|z - Z_k\|^4$.

Theorem 3.1 through Theorem 3.3 indicate that for all three noise models, the *marginal* distribution of the i-th row of the CMDS embedding $\hat{X}^{(n)}$ is, up to some orthogonal transformation $W^{(n)}$, normally distributed around the true but unknown latent position Z_i. The proofs of these theorems also imply, by the Cramer-Wold device, that for any *finite* collection of indices $S = \{i_1, i_2, \dots, i_K\}$, the random vectors

$$\{\sqrt{n}[(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})]\}_{i \in S}$$

are *jointly independent*. For example, in the setting of Theorem 3.1,

$$\lim_{n \to \infty} \mathbb{P}\{\bigcap_{i \in S} \sqrt{n}[(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})] \leq \alpha_i\} = \prod_{i \in S} \Phi(\alpha_i, \Sigma)$$

These results suggest that we can perform statistical inference using the rows of $\hat{X}_n$ as if they were independent multivariate normal random vectors centered around the *true but unknown* latent positions Z_i. Note that as long as the inference procedure is invariant with respect to orthogonal transformation, the fact that $W^{(n)}$ is unknown is immaterial; some examples include K-means and hierarchical clustering using Euclidean distances, classification using k-NN or linear discriminant analysis or support vector machines with radial basis kernels.

Remark 3. Our presentation emphasizes the central limit theorem mainly because it is a succinct limit result. Nevertheless the proof techniques used to establish Theorem 3.1 through Theorem 3.3 also yield uniform or global error bounds for $\|(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})\|$. More specifically, for a fixed index i, the central limit theorems (as presented) are consequences of the Lindeberg-Feller central limit theorem applied to a sum of independent mean 0 random variables (see Lemam A.3 and Eq. (15) in the appendix). However, if instead of the Lindeberg-Feller central limit theorem we apply a concentration inequality a la Hoeffding/Bernstein inequality then, in the setting of Theorem 3.1 and Theorem 3.2, for any index i, $\|(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})\| \leq C n^{-1/2}$ with high probability. A union bound over the n rows of $X^{(n)}$ then implies

$$\sup_{i \in [n]} \|(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})\| \leq C \sqrt{\frac{\log n}{n}}, \tag{4}$$

$$n^{-1} \sum_i \|(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})\| \leq C \sqrt{\frac{\log n}{n}}. \tag{5}$$

Eq. (4) furthermore implies that the pairwise distances between the rows of $(\hat{X}^{(n)})_i$ is a good approximation to D element-wise, i.e., letting $D(\hat{X})$ be the $n \times n$ matrix whose ij-th element is $\|(\hat{X}^{(n)})_i - (\hat{X}^{(n)})_j\|$, we have

$$\begin{aligned} \|D(\hat{X}) - D\|_{\max} &= \max_{ij} \Big| \|(\hat{X}^{(n)})_i - (\hat{X}_j^{(n)})\| - \|Z_i - Z_j\| \Big| \\ &= \Big| \|(\hat{X}^{(n)} W^{(n)})_i - (\hat{X}^{(n)} W^{(n)})_j\| - \|(Z_i - \bar{Z}) - (Z_j - \bar{Z})\| \Big| \\ &\leq 2 \max_i \|(\hat{X}^{(n)} W^{(n)})_i - (Z_i - \bar{Z})\| \leq C \sqrt{\frac{\log n}{n}}. \end{aligned} \tag{6}$$

We note that when the Z_i belongs to a *compact* subset of $\mathbb{R}^d$, Eq. (6) also implies a uniform error bound for $\hat{D}^2 - D^2$ of the same order (recall here that D^2 denote the matrix whose entries are the elementwise squares of D). This result is a considerable refinement of the Frobenius norm error bound from Zhang et al. (2016) (which we reproduced earlier in Eq. (1)). Analogous results can also be derived for the setting in Theorem 3.3.

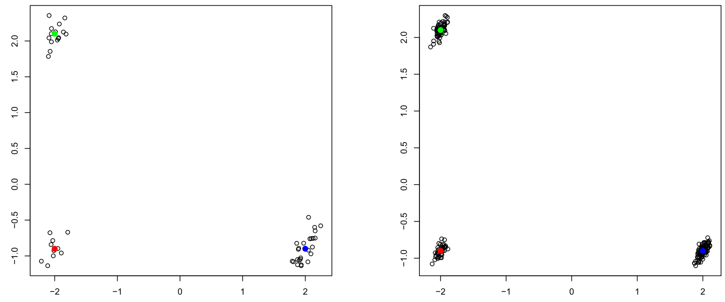
4. Empirical results

4.1. Three point-mass simulated data

As a simple illustration of our central limit theorem, we embed noisy Euclidean distances obtained from n points into $\mathbb{R}^2$. We consider three points $x_1, x_2, x_3 \in \mathbb{R}^2$ for which the inter-point distances are 3,4 and 5 (these three points form a right triangle) and generate $n_k = \pi_k n$ points equal to x_k, $k = 1, 2, 3$, where $\pi = [0.2, 0.3, 0.5]^\top$. We first consider the noise model $\Delta^2 = D^2 + E$ where the noise entries E_{ij} are i.i.d. Uniform$(-4, +4)$ for $i < j$ with $E_{ij} = E_{ji}$. For this setting, Theorem 3.1 indicates that the CMDS embedding of the dissimilarity matrix Δ into $\mathbb{R}^2$ result in a mixture of three multivariate Gaussians with different means but the same covariance matrix Σ. Figure 1 compares the embedding of one realization of Δ with $n = 50$ against the embedding of one realization of Δ with $n = 200$. Theoretically, for sufficiently large n, the rows of the embedding configuration $\hat{X}$ is centered around the three point masses located at the centroids

$$\mu_{\text{red}} = (-2, -1); \quad \mu_{\text{blue}} = (2, -1); \quad \mu_{\text{green}} = (-2, 2). \tag{7}$$

Note that these centroids are only unique up to translation and/or orthogonal transformations. The empirical covariance matrices for $n = 50$ and $n = 200$



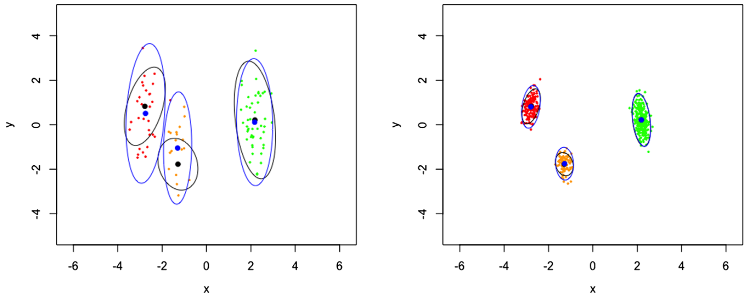
(a) n=50 (b) n=200

FIG 1. *Simulation results for the noise model $\Delta^2 = D^2 + E$ where Δ is the dissmilarity matrix on $n = 50$ and $n = 200$ points as described in Section 4.1. For this noise model, $n = 200$ is already large enough for the CMDS embedding $\hat{X}$ to exhibits the pattern of a mixture of multivariate Gaussians as specified in Theorem 3.1*

points and the theoretical covariance matrix are given below; the empirical covariance matrices are estimated using 1000 Monte Carlo replicates, i.e., for each n and each Monte Carlo replicate we generate a noisy dissimilarity matrix Δ, compute the embedding $\hat{X}$ and the sample covariance matrices for each centroid, and then average these estimates over the 1000 replicates to get the empirical covariance matrices presented below.

$$\hat{\Sigma}_n = \begin{bmatrix} 0.45 & 0.42 \\ 0.42 & 0.99 \end{bmatrix} \text{ for } n = 50; \quad \hat{\Sigma}_n = \begin{bmatrix} 0.55 & 0.51 \\ 0.51 & 1.17 \end{bmatrix} \text{ for } n = 200;$$
$$\Sigma = \begin{bmatrix} 0.54 & 0.48 \\ 0.48 & 1.23 \end{bmatrix}. \tag{8}$$

We next consider the noise model $\Delta = |D+E|$ where the noise entries E are once again of the form $E_{ij} \overset{i.i.d.}{\sim} \text{Uniform}(-4,+4)$ for $i<j$ and $E_{ij}=E_{ji}$. Theorem 3.2 indicates that the CMDS embedding of Δ into $\mathbb{R}^2$ still results in a mixture of three multivariate Gaussians but the Gaussian components now have different means (given in Eq. (7)) and possibly different covariance matrices. Figure 2 compares the embedding of one realization of Δ with $n = 100$ against the embedding of one realization of Δ with $n = 500$. Table 1 then presents the empirical covariance matrices for the three mixture components as n changes; these empirical covariance matrices converge to the true theoretical covariance matrices given in Theorem 3.2.



(a) n=100 (b) n=500

FIG 2. *Simulation results for the noise model $\Delta = |D+E|$ where Δ is the dissmilarity matrix on $n = 100$ and $n = 500$ points as described in Section 4.1. The blue ellipses are the 95% level curves of the empirical covariance matrix, and the blue dots are the empirical centers for three classes. The black dots are the true positions of x_1, x_2 and x_3, and the black ellipses are the 95% level curve for the theoretical covariance matrices as given in Theorem 3.2. For this noise model $n = 500$ is large enough for the CMDS embedding $\hat{X}$ to exhibits the pattern of a mixture of three multivariate Gaussians as specified in Theorem 3.2*

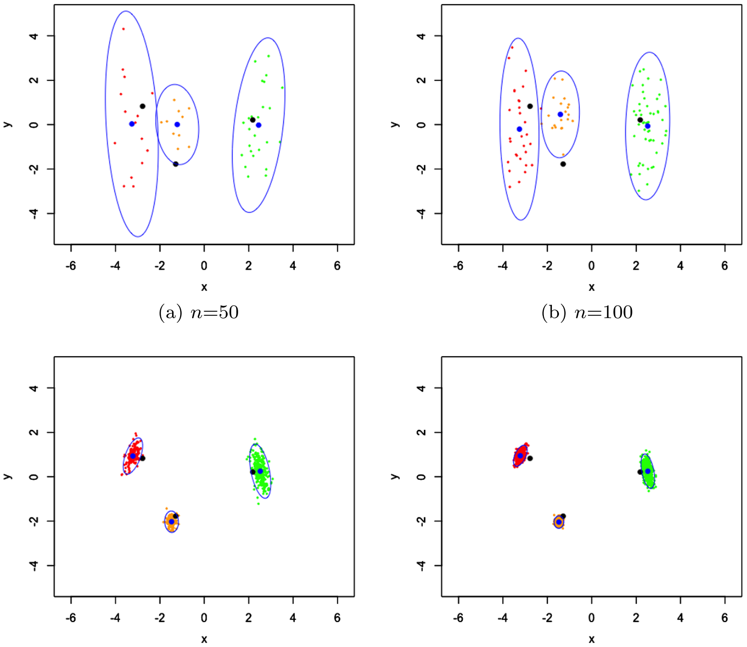
Remark 4. We note that while the noise E is identical for the two model $\Delta^2 = D^2 + E$ and $\Delta = |D+E|$ in the previous examples, its effects on the embedding configurations are quite different. In particular, for the model $\Delta^2 = D^2 + E$, the Gaussian mixture components all have the same covariance matrix; these components have different covariance matrices in the model $\Delta = |D+E|$. Furthermore, as evidenced by the magnitude of the entries of the covariance

Table 1. *Empirical estimates of the covariance matrices $\hat{\Sigma}^{(1)}$, $\hat{\Sigma}^{(2)}$ and $\hat{\Sigma}^{(3)}$ and the corresponding standard errors (in parenthesis) for the noise model $\Delta = |D + E|$. The estimates, for each value of n, are obtained from 500 Monte Carlo replicates. As n increases the empirical averages will converge to the true theoretical covariance matrices given in the last column. The empirical estimates $\hat{\Sigma}^{(i)}$ and the theoretical covariance matrices differ between the three blocks/latent positions. This is in contrasts to the model $\Delta^2 = D^2 + E$ where the covariance matrix is independent of the latent positions (see Eq. (8)).*

	$n = 500$	$n = 1000$	$n = 2000$	**Theoretical**
$\hat{\Sigma}^{(1)}$	$\begin{bmatrix} 17.71(2.45) & 10.14(2.82) \\ 10.14(2.82) & 36.03(4.95) \end{bmatrix}$	$\begin{bmatrix} 15.44(1.55) & 6.26(1.57) \\ 6.26(1.57) & 29.35(2.94) \end{bmatrix}$	$\begin{bmatrix} 14.29(1.01) & 4.35(1) \\ 4.35(1) & 26.3(1.79) \end{bmatrix}$	$\begin{bmatrix} 13.15 & 2.37 \\ 2.37 & 23.04 \end{bmatrix}$
$\hat{\Sigma}^{(2)}$	$\begin{bmatrix} 41.41(4.64) & 34.81(4.86) \\ 34.81(4.86) & 54.08(6.41) \end{bmatrix}$	$\begin{bmatrix} 37.41(3.11) & 28.29(2.81) \\ 28.29(2.81) & 42.65(3.35) \end{bmatrix}$	$\begin{bmatrix} 35.87(2.08) & 25.27(1.74) \\ 25.27(1.74) & 37.13(2) \end{bmatrix}$	$\begin{bmatrix} 34.15 & 22.37 \\ 22.37 & 31.93 \end{bmatrix}$
$\hat{\Sigma}^{(3)}$	$\begin{bmatrix} 30.96(2.57) & 40.39(3.78) \\ 40.39(3.78) & 105.22(7.93) \end{bmatrix}$	$\begin{bmatrix} 30.05(1.72) & 39.23(2.63) \\ 39.23(2.63) & 103.85(5.74) \end{bmatrix}$	$\begin{bmatrix} 29.59(1.21) & 38.45(2.03) \\ 38.5(2.03) & 102.61(4.31) \end{bmatrix}$	$\begin{bmatrix} 29.16 & 37.93 \\ 37.93 & 102.06 \end{bmatrix}$

matrices, the variability of the embedding configuration $\hat{X}$ is generally larger in the model $\Delta = |D + E|$ when compared to the model $\Delta^2 = D^2 + E$.

We recall that the statement of Theorem 3.2 assumes that the variance of the noise terms E_{ij} are the same. A practically relevant and conceptually illustrative example comes from relaxing this assumption; now the consistency result from Theorem 3.2 no longer holds. To illustrate this point we modify our noise model so that $\widetilde{E}_{ij} \overset{i.i.d.}{\sim} \text{Uniform}(-D_{ij}, D_{ij})$ for $i < j$ and $\widetilde{E}_{ij} = \widetilde{E}_{ji}$, i.e., the noise now depend on the entries of D. Figure 3 shows, for this non-constant variance setting, the embedding of one realization of Δ for different values of n. These embedding of Δ into $\mathbb{R}^2$ still appears as a mixture of class-conditional Gaussians; however, we have introduced bias into the embedding configuration in that the empirical centroids differ quite a bit from the theoretical centroids even for sufficiently large values of n.



(a) n=50

(b) n=100

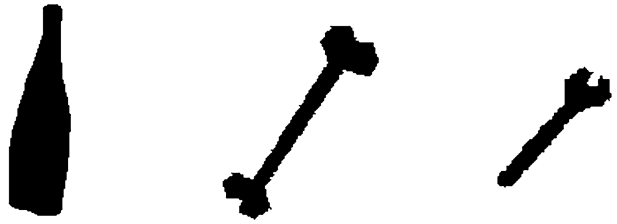
(c) n=500

(d) n=1000

FIG 3. *Simulation of CMDS with heteroscedastic noise* $\widetilde{E}$. *The black dots are the true positions for the three points. The blue dots are the empirical means and the blue ellipses are the 95% level curve of the empirical covariance matrix. NB: there is asymptotic bias.*

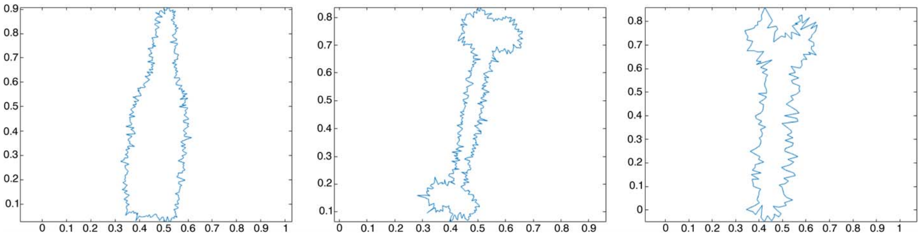
4.2. Shape clustering

As a second illustration of the effect of noise on CMDS we examine a more involved clustering experiment in the (non-Euclidean) shape space of closed curves. We consider here boundary curves obtained from silhouettes of the Kimia shape database Sharvit et al. (1998); we restrict attention to three predefined classes of objects (bottle, bone, and wrench) and take from each class three different examples of shapes all given by planar closed polygonal curves representing the objects' outline. Figure 4 shows one instance for each of the bottle, bone, and wrench class. A database of noisy curves is then created as follows: for each of the nine template shapes, we generate 100 noisy realizations in which vertices of the curve are moved along the curve's normal vectors with random distances drawn from independent Gaussian distributions at each vertex. This results in a total of 900 noisy versions of the initial curves. See Figure 5 for some examples of these noisy curves.



(a) Bottle (b) Bone (c) Wrench

Fig 4. *Examples from the Kimia Dataset.*



(a) Bottle (b) Bone (c) Wrench

Fig 5. *Noisy versions of examples from the Kimia Dataset.*

We then compute the pairwise distance matrix between all the curves (including the noiseless templates) based on a shape distance which was introduced in Glaunès et al. (2008) and later extended in the work of Kaltenmark et al. (2017). This type of metric is based on the representation of shapes in a particular distribution space called currents; see Kaltenmark et al. (2017) for details. In our context, this metric offers several advantages: (i) the distance is completely geometrical in the sense that it is independent of the sampling of the curves and does not rely on predefined pointwise correspondences between vertices; (ii) it has an intrinsic smoothing effect that provides robustness to noise to a certain degree; (iii) it can be computed in closed form with minimal computational time which is critical given the large number of pairwise distances to evaluate. We can thus view the resulting distance matrix as a perturbation of the ideal distances between the 9 template curves, i.e., we assume that we are given a dissimilarity matrix Δ arising from the noise model $\Delta = |D + E|$ where D is some true but unknown distance matrix and the noise E arises due to the noisy realizations of the templates and the smoothing effect inherent in the metric Kaltenmark et al. (2017). Note that we leave aside the issue of checking the technical assumptions on the matrix E which may be quite involved for this noise model and distance.

We proceed to perform CMDS on this distance matrix. A scree plot investigation shows that an appropriate embedding dimension here is $\hat{d} = 3$ (the top three eigenvalues are 2.20, 0.68, 0.06 with the fourth $\ll 0.01$). The resulting embedding configuration is shown in Figure 6. This configuration exhibits nine fairly well-separated clusters which are roughly centered around the position of each of the noiseless template curves. Those, in turn, form 3 'super-clusters' consistent with the classes. The ellipsoidal shape of each cluster furthermore suggests that the configuration approximately follows a mixture of multivariate Gaussians. For a somewhat more quantitative assessment of this approximation we perform, for each clusters, a goodness-of-fit test for multivariate normality. We used three different test procedures; one is based on multivariate skewness and kurtosis (Mardia, 1970) and the other two are based on the weighted L_2 distance between the empirical and theoretical characteristic functions (Henze and Zirkler, 1990; Szekely and Rizzo, 2013). We fail to reject, for all three test procedures and all nine clusters, the null hypothesis that the points in the cluster are multivariate normal. For all three test procedures, the reported p-values for the nine clusters ranges from 0.07 to 0.95; these reported p-values had not been corrected for multiple comparisons. While these preliminary shape clustering results are obtained with a specific and simple distance on the space of curves, future work will investigate whether similar properties hold with different, more elaborate metrics and/or geometric noise models. The central limit theorems derived here could then constitute a useful theoretical tool to evaluate the discriminating power of shape clustering methods based on CMDS.

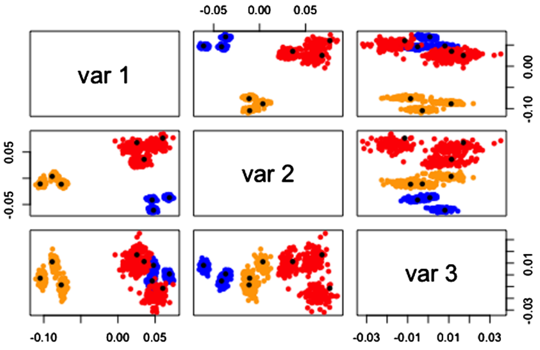
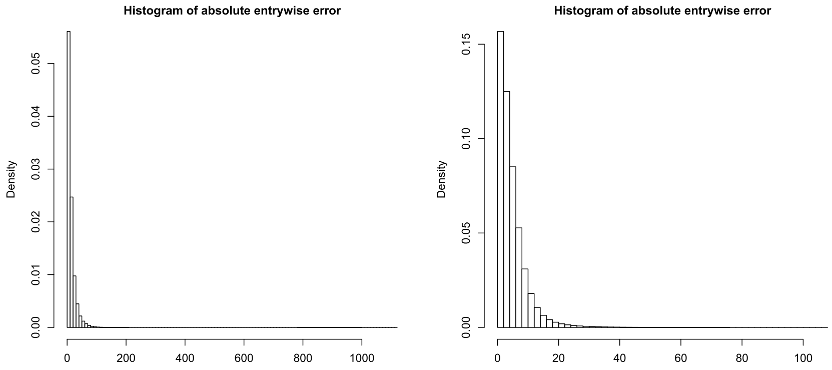


Fig 6. *Pairs plot of CMDS into $\mathbb{R}^3$ for the noisy curves. Colors correspond to the different classes (blue for bottle, red for bone, and orange for wrench). The position of the nine template curves in the configuration are highlighted with large black dots.*



(a) $q = 0.2$ (b) $q = 0.5$

Fig 7. *Histogram plot of the reconstruction error for the big cities distances matrix with* $q =$ 20% *observed entries (left plot) and* $q = 50\%$ *observed entries (right plot). The accompanying Table 2 indicates that the reconstruction error is uniformly small for most of the missing entries.*

4.3. Distance matrix completion

Our last example on the effect of noise on CMDS concerns the problem of recovering the missing entries in a partially observed distance matrix D. We use a dataset on the locations (longitude and latitude coordinates) of the 4000 most populous cities in the world; this dataset is part of the `mdsr` package (Baumer et al., 2019) in `R`. We first construct a Euclidean distance matrix D between the cities using the latitude and longitude coordinates. We then keep a subsample of $q \times 100\%$ of the upper triangular entries of D; the remaining $(1-q) \times 100\%$ upper triangular entries are set to `NA` to denote missing entries. Note that the lower triangular entries of D are kept, or denote missing, in exact correspondence with the upper triangular entries of D. Let Δ_q denote the matrix obtained from D after this subsampling process. We then replace the `NA`'s with the average of the non-missing entries of Δ. We then do CMDS of Δ_q into $\mathbb{R}^2$, yielding a configuration $\hat{\mathbf{X}}$ as a 4000×2 matrix. We estimate the original distance matrix D by computing the pairwise Euclidean distances between the rows of $q^{-1/2}\hat{\mathbf{X}}$. The resulting estimates and their errors for two values of $q = 0.2$ and $q = 0.5$ are given in Figure 7 and Table 2, i.e., the plots in Figure 7 are histogram plots for $\{|\hat{D}_{ij} - D_{ij}| \colon (i,j) \in S_q\}$ where S_q denote the indices of the missing entries in Δ_q while Table 2 displays the quantiles of $\{|\hat{D}_{ij} - D_{ij}| \colon (i,j) \in S_q\}$. Figure 7 and Table 2 indicate that the entry-wise absolute error $|\hat{D}_{ij} - D_{ij}|$ are generally small but with a somewhat heavy-tailed and right-skewness; these phenomena corroborate with the theory in Theorem 3.3. Indeed, the covariance matrix in Theorem 3.3 depends on $\|z - Z_k\|^4$ and thus outlier points and/or outlier distances will, in general, have large residuals.

TABLE 2. *Reconstruction error for the big cities distances matrix with $q = 20\%$ (first row) and $q = 50\%$ (second row) observed entries. The columns are the quantiles level of the reconstruction error versus the true distances, e.g., if we observe 50% of the true D_{ij} then the reconstruction error for the remaining entries have* median *absolute error of* 3.43. *The median value for the unobserved Euclidean distances is* 83.2.

	$\alpha = 0.01$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.975$	$\alpha = 0.99$	$\alpha = 1$
$\lvert\hat{D}_{ij} - D_{ij}\rvert,\ q = 0.2$	0.14	3.83	8.54	16.63	29.18	39.97	52.23	68.58	1113.41
$\lvert\hat{D}_{ij} - D_{ij}\rvert,\ q = 0.5$	0.61	1.57	3.43	6.253	9.961	13.118	15.843	18.313	106.352
D_{ij}	3.8	43.6	83.2	127.1	184.4	207.5	219.8	227.9	340.7

5. Discussion

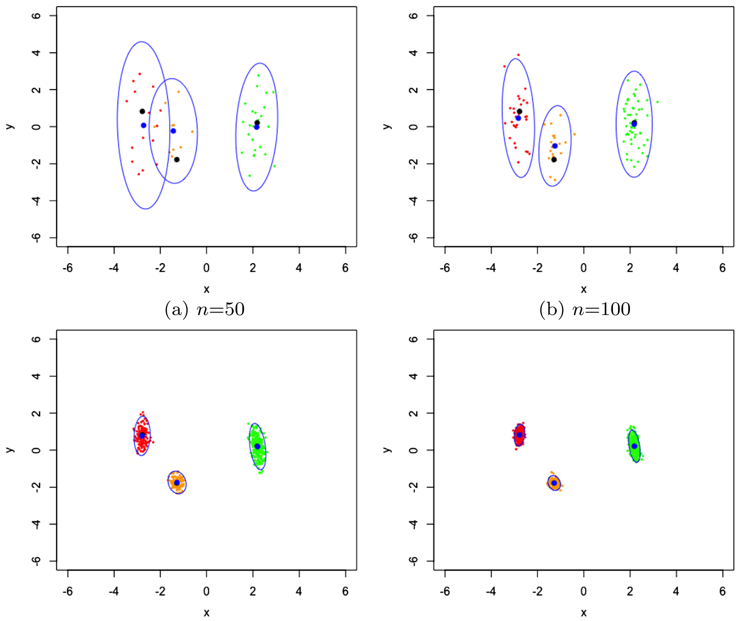
The authors of Athreya et al. (2016) and Levin et al. (2017) prove that adjacency spectral embedding of random dot product graphs result in central limit theorems for the estimated latent positions. In this work we extend these results to the previously unexplored area of perturbation analysis for CMDS, thereby addressing a gap in the literature as acknowledged in e.g., Fan et al. (2018) and Peterfreund and Gavish (2018). Notably, the three noise models we proposed in Section 2 each give rise to a central limit theorem; that is, for Euclidean distance matrix, the rows of the configuration matrix given by CMDS under noise will center around the corresponding rows of the true configuration matrix. Furthermore our simulations on the synthetic data together with experiments on the shape clustering data and the distance matrix recovery all demonstrated the validity of our results. We have avoided any discussion of the model selection problem of choosing a suitable embedding dimension $\hat{d}$. Instead, we assume d is known – except in Section 4.2. There are many methods for choosing (spectral) embedding dimensions, see Zhu and Ghodsi (2006); Jackson (1991); Chatterjee (2015).

Another natural, and important, practical question is how to estimate the parameter σ in the noise model of interests. We note, however, that consistent estimation of σ is not necessary for our embedding method and the corresponding theoretical results. Indeed, the classical multidimensional scaling algorithm does not require estimation of σ, but rather the dimension d of the original data points (see the description of classical multidimensional scaling in Section 1). Under all of our noise model, $\|\mathbf{E}\| \leq \sigma\sqrt{n}$ and provided that we choose d such that $\lambda_d > n^{1/2+\epsilon}$ for any $\epsilon > 0$, then our theoretical limit results apply. For concreteness, we can choose $\epsilon = 1/3$ and thus as long as we choose the embedding dimension $\hat{d}$ satisfying $\lambda_{\hat{d}}(B) \geq n^{2/3}$, then $\hat{d} \to d$ almost surely and our central limit theorem applies.

Throughout this paper, we assume that d is fixed as $n \to \infty$. Therefore, given a central limit theorem for the embedding into d dimension, one can derive a central limit theorem for the embedding into $d' < d$ dimension in a straightforward manner. More specifically, given a dissimilarity matrix Δ and positive integers $d' \leq d$, the classical multidimensional scaling of Δ into $\mathbb{R}^{d'}$ is equivalent to the classical multidimensional scaling of Δ into $\mathbb{R}^{d}$ and keeping the first $d' < d$ columns (see the description of classical multidimensional scaling in Section 1). Thus, our limit results can be rephrased to say that, letting $\hat{X}_n^{(d')}$ denote the classical multidimensional scaling of Δ into $\mathbb{R}^{d'}$ for $d' < d$, that there exists a sequence of $d' \times d'$ orthogonal matrix $W_n^{(d')}$ and a sequence of $d \times d'$ matrices with orthonormal columns T_n such that

$$\sqrt{n}\Big((\hat{X}_n^{(d')} W_n^{(d')})_i - T_n(Z_n - \bar{Z}_n)_i\Big)$$

converges to a mixture of multivariate normal. For a given n, T_n corresponds to the principal component projection of the $n \times d$ matrix $[Z_1 \mid Z_2 \mid \dots, Z_n]^\top$ into



(a) n=50 (b) n=100

(c) n=500 (d) n=1000

FIG 8. *Simulation of MDS using raw stress criterion for $n = 50$, 100, 500 and 1000 points. The black dots are the true positions of x_1, x_2 and x_3, the blue dots are the empirical mean of the simulation and the blue ellipses are the 95% level curve of the empirical covariance matrix.*

$\mathbb{R}^{d'}$. We emphasize that T_n is not necessarily unique (indeed, the eigenvalues of the covariance matrix for Z_n are not necessarily distinct).

We further note that the dependency on d in our limit results is implicit in the covariance matrices. Naively speaking, we can say that the estimation accuracy is inversely proportional to d. This is most visible in the statement of Eq. (1) (which is also a corollary of our results), since as d increases r also increases, note that $r \leq d + 2$. A more precise description is that the accuracy of our limit results depends on the covariance matrix Σ, which is a $d \times d$ matrix. Since the squared norm of a mean 0 multivariate Gaussian is the trace of its covariance matrix, we see that as d increases, the trace of Σ generally increases but the rate at which it increases need not depends on d. Indeed, the trace of Σ depends purely on the distribution F of the underlying data points; in the case where the data points are sampled from a multivariate normal with mean 0 and identity matrix in $\mathbb{R}^d$, then as d increases, the trace of Σ also increases linearly.

Finally we note that CMDS is just one of a wide variety of multidimensional scaling techniques. Minimizing the raw stress criterion is another commonly used MDS technique (de Leeuw and Heiser, 1982), i.e., given a $n \times n$ observed dissimilarity matrix Δ and an embedding dimension d, one seeks to minimize the objective function

$$\sigma_r = \sigma_r(X) = \sum_{i<j}(\Delta_{ij} - \|X_i - X_j\|)^2.$$

The minimization of $\sigma_r(X)$ is with respect to all configurations $X \in \mathbb{R}^{n\times d}$ and usually proceeds via an iterative algorithm which updates the configuration matrix X until a stopping criterion is met. Keeping the simulation settings as in Section 4.1, the resulting configuration is shown in Figure 8. This suggests that the CLT may hold for raw stress just as well as for CMDS. However, this claim is at best a conjecture at present as perturbation analysis of stress minimization algorithms is significantly more involved.

References

A. Y. Alfakih, A. Khandani, and H. Wolkowicz. Solving Euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12(1):13–30, Jan 1999. ISSN 1573-2894. URL https://doi.org/10.1023/A:1008655427845.

A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman. A limit theorem for scaled eigenvectors of random dot product graphs. *Sankhya A*, 78(1):1–18, Feb 2016. ISSN 0976-8378. URL https://doi.org/10.1007/s13171-015-0071-x. MR3494576

M. Bakonyi and C. Johnson. The Euclidian distance matrix completion problem. *SIAM Journal on Matrix Analysis and Applications*, 16(2):646–654, 1995. URL https://doi.org/10.1137/S0895479893249757.

Ben Baumer, Nicholas Horton, and Daniel Kaplan. *mdsr: Complement to 'Modern Data Science with R'*, 2019. URL https://CRAN.R-project.org/package=mdsr. R package version 0.1.7.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003. ISSN 0899-7667. URL http://dx.doi.org/10.1162/089976603321780317.

I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, 2005. MR2158691

J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319, Sep. 1970. ISSN 1860-0980. URL https://doi.org/10.1007/BF02310791.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015. MR3285604

L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104(485):209–219, 2009. URL https://doi.org/10.1198/jasa.2009.0111. MR2504374

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. ISSN 1063-5203. URL http://www.sciencedirect.com/science/article/pii/S1063520306000546. Special Issue: Diffusion Maps and Wavelets. MR2238665

M. A. A. Cox and T. F. Cox. *Multidimensional Scaling*, pages 315–347. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-33037-0. URL https://doi.org/10.1007/978-3-540-33037-0_14.

T. F. Cox and M. A. A. Cox. *Multidimensional scaling*. CRC Press, 2010.

A. Criminisi and J. Shotton. Manifold forests. In A. Criminisi and J. Shotton, editors, *Decision Forests for Computer Vision and Medical Image Analysis*, chapter 7, pages 79–94. Springer, London, 2013.

C. Davis and M. Kahan, W. The rotation of eigenvectors by a perturbation III. *SIAM Journal of Numerical Analysis*, 7:1–46, 1970. MR0264450

J. de Leeuw and W. Heiser. Theory of multidimensional scaling. In P. R. Krishnaiah and L. Kanal, editors, *Handbook of Statistics II*, pages 285–316. North Holland Publishing Company, Amsterdam, The Netherlands, 1982. MR0716709

Jianqing Fan, Qiang Sun, Wen-Xin Zhou, and Ziwei Zhu. *Principal Component Analysis for Big Data*, pages 1–13. American Cancer Society, 2018. ISBN 9781118445112. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat08122.

J. Glaunès, A. Qiu, M. I. Miller, and L. Younes. Large deformation diffeomorphic metric curve mapping. *International Journal of Computer Vision*, 80(3):317–336, 2008. ISSN 0920-5691. URL http://dx.doi.org/10.1007/s11263-008-0141-9.

J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A kernel view of the dimensionality reduction of manifolds. In *Proceedings of the 21st International Conference on Machine Learning*, 2004.

N. Henze and B. Zirkler. A class of invariant consistent tests for multivariate normality. *Communications in Statistics: Theory and Methods*, 19:3595–3617, 1990. MR1089501

J. E. Jackson. *A User's Guide to Principal Components*. Wiley & Sons, New York, 1991.

A. Javanmard and A. Montanari. Localization from incomplete noisy distance measurements. *Foundations of Computational Mathematics*, 13(3):297–345, Jun 2013. ISSN 1615-3383. URL https://doi.org/10.1007/s10208-012-9129-5.

Natalia Jaworska and Angelina Chupetlovska-Anastasova. A review of multidimensional scaling (mds) and its utility in various psychological domains. *Tutorials in quantitative methods for psychology*, 5(1):1–10, 2009.

I. Kaltenmark, B. Charlier, and N. Charon. A general framework for curve and surface comparison and registration with oriented varifolds. *Computer Vision and Pattern Recognition (CVPR)*, 2017.

K. Levin, A. Athreya, M. Tang, V. Lyzinski, and C. E. Priebe. A central limit theorem for an omnibus embedding of multiple random dot product graphs. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 964–967, 2017.

J. A. T. Machado and M. E. Mata. Analysis of world economic variables using multidimensional scaling. *PLOS ONE*, 10(3):1–17, 03 2015. URL https://doi.

org/10.1371/journal.pone.0121277.

K. V. Mardia. Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57:519–530, 1970. MR0397994

S. Oh, A. Montanari, and A. Karbasi. Sensor network localization from local connectivity: Performance analysis for the mds-map algorithm. In *2010 IEEE Information Theory Workshop on Information Theory (ITW 2010, Cairo)*, pages 1–5, Jan 2010. MR2973771

E. Pekalska and R. P. W. Duin. *The Dissimilarity Representation for Pattern Recognition: Foundations and Applications*. World Scientific Publishing Company Inc, Singapore, 2005.

E. Peterfreund and M. Gavish. Multidimensional Scaling of Noisy High Dimensional Data, January 2018. arXiv:1801.10229.

S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

R. Sibson. Studies in the robustness of multidimensional scaling: Perturbation analysis of classical scaling. *Journal of the Royal Statistical Society*, 41:217–229, 1979. MR0547248

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. URL https://doi.org/10.1162/089976698300017467.

D. Sharvit, J. Chan, H. Tek, and B. B. Kimia. Symmetry-based indexing of image databases. *Journal of Visual Communication and Image Representations*, 9:366–380, 1998.

A. Singer. A remark on global positioning from local distances. *Proceedings of the National Academy of Sciences*, 105(28):9507–9511, 2008. ISSN 0027-8424. URL http://www.pnas.org/content/105/28/9507. MR2430205

I. Spence and D. W. Domoney. Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika*, 39(4):469–490, Dec 1974. ISSN 1860-0980. . URL https://doi.org/10.1007/BF02291669.

G. J. Szekely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143:1249–1272, 2013. MR3055745

Mohammad J. Taghizadeh, Reza Parhizkar, Philip N. Garner, Hervé Bourlard, and Afsaneh Asaei. Ad hoc microphone array calibration: Euclidean distance matrix completion algorithm and theoretical guarantees. *Signal Processing*, 107:123–140, 2015. ISSN 0165-1684. URL http://www.sciencedirect.com/science/article/pii/S0165168414003508. Special Issue on ad hoc microphone arrays and wireless acoustic sensor networks Special Issue on Fractional Signal Processing and Applications.

A. Tasissa and R. Lai. Exact reconstruction of Euclidean distance geometry problem using low-rank matrix completion. *CoRR*, abs/1804.04310, 2018. MR3951386

J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. ISSN 0036-8075. URL http://science.sciencemag.org/content/290/5500/2319.

W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952. MR0054219

J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2012. ISSN 1615-3383. URL https://doi.org/10.1007/s10208-011-9099-z. MR2946459

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. MR3837109

J. T. Vogelstein, Y. Park, T. Ohyama, R. A. Kerr, J. W. Truman, C. E. Priebe, and M. Zlatic. Discovery of brainwide neural-behavioral maps via multiscale unsupervised structure learning. *Science*, 344(6182):386–392, 2014. ISSN 0036-8075. URL http://science.sciencemag.org/content/344/6182/386.

Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102:351–323, 2015. MR3371006

L. Zhang, G. Wahba, and M. Yuan. Distance shrinkage and Euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):849–867, 2016. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12138. MR3534353

M. Zhu and A. Ghodsi. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis*, 51(2):918–930, 2006. MR2297497

Appendix: Proofs of stated results

We now present detailed arguments for the proof of Theorem 3.2. We note that the machinery involved in proving Theorem 3.1 and Theorem 3.3 are by and large the same as that used in proving Theorem 3.2 and will thus be omitted. Given a matrix A, we denote by $\|A\|$ and $\|A\|_F$ the spectral and Frobenius norm of A, respectively. We will utilize the following observation repeatedly in our presentation.

Observation A.1. *Let A and B be matrices of appropriate dimensions. Then*

$$\|AB\|_F = \|B^\top A^\top\|_F \leq \min\{\|A\| \times \|B\|_F, \|B\| \times \|A\|_F\}.$$

We remind our readers of the following notations that are used in the subsequent presentation. Let $P = (I - 11^\top/n)$. Recall that $B = -\frac{1}{2}PD^2P$ and $\hat{B} = -\frac{1}{2}P\Delta^2P$ are the double centering of D^2 and Δ^2, respectively. If D^2 is a (squared) Euclidean distance matrix whose elements are $D_{ij}^2 = \|Z_i - Z_j\|^2$, then $B = PZZ^\top P$ and $PZ = U_B S_B^{1/2} \tilde{W}_n$ for some orthogonal matrix $\tilde{W}_n$. Now let W^* be the orthogonal matrix satisfying $W^* = \arg\min_W \|U_B^\top \hat{U}_B - W\|$. Our main goal is to investigate the quantity $\hat{X} - U_B S_B^{1/2} W^*$. The following lemma provides a decomposition for $\hat{X} - U_B S_B^{1/2} W^*$ into a sum of several matrices.

Lemma A.2. *Let W^* be the orthogonal matrix satisfying $W^* = \arg\min_W \|U_B^\top \hat{U}_B - W\|$, then*

$$\begin{aligned}
\hat{X} - U_B S_B^{1/2} W^* = {} & (\hat{B} - B) U_B S_B^{-1/2} W^* && (9)\\
& - (\hat{B} - B) U_B (S_B^{-1/2} W^* - W^* S_{\hat{B}}^{-1/2}) && (10)\\
& - U_B U_B^\top (\hat{B} - B) U_B W^* S_{\hat{B}}^{-1/2} && (11)\\
& + (I - U_B U_B^\top)(\hat{B} - B)(U_{\hat{B}} - U_B W^*) S_{\hat{B}}^{-1/2} && (12)\\
& + U_B (U_B^\top U_{\hat{B}} - W^*) S_{\hat{B}}^{1/2} && (13)\\
& + U_B (W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*). && (14)
\end{aligned}$$

Proof. We have

$$\begin{aligned}
\hat{X} - U_B S_B^{1/2} W^* &= U_{\hat{B}} S_{\hat{B}}^{1/2} - U_B W^* S_{\hat{B}}^{1/2} + U_B (W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*)\\
&= U_{\hat{B}} S_{\hat{B}}^{1/2} - U_B U_B^\top U_{\hat{B}} S_{\hat{B}}^{1/2} + U_B U_B^\top U_{\hat{B}} S_{\hat{B}}^{1/2} - U_B W^* S_{\hat{B}}^{1/2}\\
&+ U_B (W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*)\\
&= (I - U_B U_B^\top) \hat{B} U_{\hat{B}} S_{\hat{B}}^{-1/2} + U_B (U_B^\top U_{\hat{B}} - W^*) S_{\hat{B}}^{1/2}\\
&+ U_B (W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*)\\
&= (I - U_B U_B^\top)(\hat{B} - B) U_{\hat{B}} S_{\hat{B}}^{-1/2} + U_B (U_B^\top U_{\hat{B}} - W^*) S_{\hat{B}}^{1/2}\\
&+ U_B (W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*).
\end{aligned}$$

Note that we used the facts $U_B U_B^\top B = B$ and $U_{\hat{B}} S_{\hat{B}}^{1/2} = \hat{B} U_{\hat{B}} S_{\hat{B}}^{-1/2}$ in the above equalities. The last two terms of the above display is Eq. (13) and Eq. (14) in the statement of the Lemma. We now consider the term $(I - U_B U_B^\top)(\hat{B} - B) U_{\hat{B}} S_{\hat{B}}^{-1/2}$. We have

$$\begin{aligned}
& (I - U_B U_B^\top)(\hat{B} - B) U_{\hat{B}} S_{\hat{B}}^{-1/2}\\
&= (I - U_B U_B^\top)(\hat{B} - B)(U_B W^* + \hat{U}_B - U_B W^*) S_{\hat{B}}^{-1/2}\\
&= (\hat{B} - B) U_B W^* S_{\hat{B}}^{-1/2} - U_B U_B^\top (\hat{B} - B) U_B W^* S_{\hat{B}}^{-1/2}\\
&+ (I - U_B U_B^\top)(\hat{B} - B)(\hat{U}_B - U_B W^*) S_{\hat{B}}^{-1/2}\\
&= (\hat{B} - B) U_B S_B^{-1/2} W^* - (\hat{B} - B) U_B (S_B^{-1/2} W^* - W^* S_{\hat{B}}^{-1/2})\\
&- U_B U_B^\top (\hat{B} - B) U_B W^* S_{\hat{B}}^{-1/2}\\
&+ (I - U_B U_B^\top)(\hat{B} - B)(\hat{U}_B - U_B W^*) S_{\hat{B}}^{-1/2}.
\end{aligned}$$

The four terms in the above display correspond to the terms in Eq. (9) through Eq. (12). □

Note that from Lemma A.2, we have $\hat{X}W^{*\top}\tilde{W}_n - U_B S_B^{1/2}\tilde{W}_n = (\hat{B} - B)U_B S_B^{-1/2}\tilde{W}_n$ + remainder terms given in Eq. (10) through Eq. (14). The essential term is $(\hat{B} - B)U_B S_B^{-1/2}\tilde{W}_n$ and we showed, in Lemma A.3 below, that the rows of this matrix converge to multivariate normals. As for the remainder terms, Lemma A.4 implies that the rows of the matrices in Eq. (10) through Eq. (14), when scaled by $\sqrt{n}$, converge to 0 in probability. Combining these results yield the proof of Theorem 3.2. Indeed, the term $\hat{X}W^{*\top}\tilde{W}_n$ can be written as $\hat{X}W_n$ for some orthogonal matrix $W_n = W^{*\top}\tilde{W}_n$ that appeared in the statements of Theorem 3.1 through Theorem 3.3 while the rows of $U_B S_B^{1/2}\tilde{W}_n$ is, as we observed earlier, simply $(Z_i - \bar{Z})$.

Lemma A.3. *Let $Z_1, Z_2, \ldots, Z_n$ be the rows of Z and that $Z_1, \ldots, Z_n \overset{i.i.d}{\sim} F$ for some sub-Gaussian distribution F. Then there exists a sequence of $d \times d$ orthogonal matrices $\tilde{W}_n$, such that for any fixed index i, we have*

$$\sqrt{n}\tilde{W}_n^\top[(\hat{B} - B)U_B S_B^{-1/2}]_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma(z_i))$$

where $\Sigma(z_i) = \Xi^{-1}\widetilde{\Sigma}(z_i)\Xi^{-1}$, $\Xi = \mathbb{E}[Z_k {Z_k}^\top] \in \mathbb{R}^{d\times d}$, $\mu = \mathbb{E}[Z_k] \in \mathbb{R}^d$. and

$$\widetilde{\Sigma}(z_i) = \mathbb{E}_{Z_k}[(\sigma^2||z_i - Z_k||^2 + \mathbb{E}[E_{ij}^3]\|z_i - Z_k\| + \frac{1}{4}\mathbb{E}[E_{ij}^4] - \frac{\sigma^4}{4})(Z_k - \mu)(Z_k - \mu)^\top] \\ \in \mathbb{R}^{d\times d}$$

is a covariance matrix depending on x_i. Here, for ease of notation, we denote by $(A)_i$ or $[A]_i$ the i-th row of matrix A.

Proof. Recall that $PZ = U_B S_B^{1/2}\tilde{W}_n$, i.e., $U_B S_B^{1/2} = PZ\tilde{W}_n^\top$. We therefore have

$$\begin{aligned}
&\sqrt{n}\tilde{W}_n^\top[(\hat{B} - B)U_B S_B^{-1/2}]_i \\
&= \sqrt{n}\tilde{W}_n^\top[(\hat{B} - B)PZ\tilde{W}_n^\top {S_B}^{-1}]_i \\
&= \sqrt{n}\tilde{W}_n^\top {S_B}^{-1}\tilde{W}_n[(\hat{B} - B)PZ]_i \\
&= -\sqrt{n}\tilde{W}_n^\top {S_B}^{-1}\tilde{W}_n\Big[P\Big(D \circ E + \frac{E^2}{2}\Big)PZ\Big]_i \\
&= -\sqrt{n}\tilde{W}_n^\top {S_B}^{-1}\tilde{W}_n\Big[\Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\Big)\Big(D \circ E + \frac{E^2}{2}\Big)\Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\Big)Z\Big]_i \\
&= -\sqrt{n}\tilde{W}_n^\top {S_B}^{-1}\tilde{W}_n\Big[\Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\Big)\Big(D \circ E + \frac{E^2}{2}\Big)(Z - \mathbf{1}\bar{Z}^\top)\Big]_i \\
&= -\sqrt{n}\tilde{W}_n^\top {S_B}^{-1}\tilde{W}_n\Big[\Big(I - \frac{\mathbf{1}\mathbf{1}^\top}{n}\Big)\Big(D \circ E + \frac{E^2 - \sigma^2\mathbf{1}\mathbf{1}^\top}{2}\Big)(Z - \mathbf{1}\bar{Z}^\top)\Big]_i.
\end{aligned}$$

The last equality in the above display holds since $(I - \frac{\mathbf{1}\mathbf{1}^\top}{n})\frac{\sigma^2\mathbf{1}\mathbf{1}^\top}{2}(Z - \mathbf{1}\bar{Z}^\top) = 0$. Now by the strong law of large numbers, we have

$$\frac{\mathbf{1}\mathbf{1}^\top}{n}(D \circ E + \frac{E^2 - \sigma^2\mathbf{1}\mathbf{1}^\top}{2})(Z - \mathbf{1}\mu^\top + \mathbf{1}\mu^\top - \mathbf{1}\bar{Z}^\top) \longrightarrow 0$$

as $n \to \infty$. We therefore have, for sufficiently large n, that

$$\begin{aligned}
&\sqrt{n}\tilde{W}_n^\top[(\hat{B}-B)U_B S_B^{-1/2}]_i \\
&= -\sqrt{n}\tilde{W}_n^\top S_B{}^{-1}\tilde{W}_n\Big[\Big(D\circ E + \frac{E^2-\sigma^2\mathbf{1}\mathbf{1}^\top}{2}\Big)(Z-\mathbf{1}\bar{Z}^\top)\Big]_i + o_P(1) \\
&= -n\tilde{W}_n^\top S_B{}^{-1}\tilde{W}_n\Big[\frac{1}{\sqrt{n}}\Big(\sum_{j=1}^{n}\Big[\Big(D\circ E + \frac{E^2-\sigma^2\mathbf{1}\mathbf{1}^\top}{2}\Big)_{ij}(Z-\mathbf{1}\mu^\top)_j\Big]\Big)\Big] + o_P(1)
\end{aligned}$$

Ignoring the $o_P(1)$ term (which converges to 0 as $n \to \infty$), the above display simplifies to

$$\begin{aligned}
&\sqrt{n}\tilde{W}_n^\top[(\hat{B}-B)U_B S_B^{-1/2}]_i \\
&= -n\tilde{W}_n^\top S_B{}^{-1}\tilde{W}_n[\frac{1}{\sqrt{n}}(\sum_{j=1}^{n}[(D_{ij}\cdot E_{ij} + \frac{E_{ij}^2-\sigma^2\mathbf{1}\mathbf{1}^\top}{2})(Z_j-\mu)])]. \qquad (15)
\end{aligned}$$

Condition on $Z_i = z_i$, (15) is then the sum of $n-1$ independent mean 0 random variables (since $D_{ii} = 0$), each with the same covariance matrix (for $j \neq i$)

$$\tilde{\Sigma}(z_i) = \mathrm{Cov}[(E_{ij}\|z_i - Z_j\| + \frac{E_{ij}^2-\sigma^2}{2})(Z_j-\mu)].$$

Now by the law of total variance, since $\mathbb{E}[E_{ij} \mid Z_j] = 0$ and $\mathbb{E}[E_{ij}^2 - \sigma^2 \mid Z_j = 0]$, we have

$$\begin{aligned}
\Sigma(\tilde{z}_i) &= \mathbb{E}\Big[\mathbb{E}\Big[E_{ij}^2\|z_i-Z_j\| + E_{ij}\|z_i-Z_j\|(E_{ij}^2-\sigma^2) \\
&\quad + \frac{(E_{ij}^2-\sigma^2)^2}{4}\Big|Z_j\Big](Z_j-\mu)(X_j-\mu)^\top\Big] \\
&= \mathbb{E}_{Z_j}\Big[(\sigma^2\|z_i-Z_j\|^2 + \gamma\|z_i-Z_j\| + \frac{\xi}{4} - \tfrac{\sigma^4}{4})(Z_j-\mu)(Z_j-\mu)^\top\Big].
\end{aligned}$$

Finally, by the strong law of large numbers, we have

$$\frac{\tilde{W}_n^\top S_B\tilde{W}_n}{n} = \frac{1}{n}(PZ)^\top PZ \longrightarrow \Xi := \mathrm{Cov}(Z_j) \in \mathbb{R}^{d\times d}$$

almost surely. Hence $(n\tilde{W}_n^\top S_B^{-1}\tilde{W}_n) \to \Xi^{-1}$ almost surely. Slutsky's theorem then yields

$$\sqrt{n}\tilde{W}_n^\top[(\hat{B}-B)U_B S_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \Xi^{-1}\widetilde{\Sigma}(x_i)\Xi^{-1})$$

in distribution as $n \to \infty$. □

We now look at the matrices in Eq. (10) through Eq. (14). The following lemma show that any row of these matrices, when scaled by $\sqrt{n}$, will converge to 0 in probability.

Lemma A.4. *We have, simultaneously*

$$\sqrt{n}[(\hat{B}-B)U_B(W^*S_{\hat{B}}^{-1/2}-S_B^{-1/2}W^*)]_i \xrightarrow{P} 0, \tag{16}$$

$$\sqrt{n}[U_BU_B^\top(\hat{B}-B)U_BW^*S_{\hat{B}}^{-1/2}]_i \xrightarrow{P} 0, \tag{17}$$

$$\sqrt{n}[(I-U_BU_B^\top)(\hat{B}-B)(\hat{U}_B-U_BW^*)S_{\hat{B}}^{-1/2}]_i \xrightarrow{P} 0, \tag{18}$$

$$\sqrt{n}[U_B(U_B^\top U_{\hat{B}}-W^*)S_{\hat{B}}^{1/2}]_i \xrightarrow{P} 0, \tag{19}$$

$$\sqrt{n}[U_B(W^*S_{\hat{B}}^{1/2}-S_B^{1/2}W^*)]_i \xrightarrow{P} 0. \tag{20}$$

The rest of this Appendix is devoted toward proving Lemma A.4, for which we need the following technical lemmas controlling the spectral norm of $\|\hat{B}-B\|$ and $\|U_B^\top\hat{U}_B-W^*\|$ (recall that W^* is the closest orthogonal matrix, in Frobenius norm, to $U_B^\top\hat{U}_B$.) We start with a bound for the spectral norm of $B-\hat{B}$.

Proposition A.5. $\|B-\hat{B}\| = \mathcal{O}(\sqrt{n\log n})$ *with high probability.*

Proof. We have

$$\begin{aligned}\|B-\hat{B}\| &= \|-\frac{1}{2}PD^2P+\frac{1}{2}P(D+E)^2P\| \\ &= \|PD\circ EP+\frac{1}{2}PE^2P\| \text{ (where } \circ \text{ is the Hadamard product)} \\ &\le \|D\circ E\|+\frac{1}{2}\|E^2-\mathbb{E}[E^2]\| \text{ (since } \|P\|=1.) \\ &= \mathcal{O}(\sqrt{n})+\mathcal{O}(\sqrt{n\log n}).\end{aligned}$$

Note that here we used $\mathbb{E}[D\circ E]=0$ and $\mathbb{E}[\frac{1}{2}PE^2P]=0$. Each entries of $D\circ E$ is of sub-Gaussian distribution with mean 0 and each entries of $E^2-\mathbb{E}[E^2]$ is of sub-exponential distribution with mean 0. An application of Theorem 4.4.5 in Vershynin (2018) and Matrix Bernstein for the sub-exponential case in Tropp (2012) gives the desired result. □

Lemma A.6. *Let* $X_1,\dots,X_n,Y \overset{i.i.d}{\sim} F$ *for some sub-Gaussian distribution* F*, where* X_i *is the ith row of the configuration matrix* X *of* B *viewed as a column vector. Let* $\Xi=\mathbb{E}[X_1X_1^\top]$ *be of rank* d*, then* $\lambda_i(B)=\Omega(n)$ *almost surely.*

Proof. For any matrix H, the nonzero eigenvalues of $H^\top H$ are the same as those $HH^\top$, so $\lambda_i(XX^\top)=\lambda_i(X^\top X)$. In what follows, we remind the reader that X is a matrix whose rows are the transposes of the column vectors X_i, and Y is a d-dimensional vector that is independent from and has the same distribution as that of the X_i. We observe that $(X^\top X-n\mathbb{E}[YY^\top])_{ij}=\sum_{k=1}^{n}(X_{ki}X_{kj}-\mathbb{E}[Y_iY_j])$ is a sum of n independent mean-zero sub-Gaussian random variables. By a general Hoeffding's inequality for sub-gaussian random variables (Vershynin, 2018), for all $i,j\in[d]$,

$$\mathbb{P}[|(X^\top X-n\mathbb{E}[YY^\top])_{ij}|\ge t] \le 2\exp\Big\{\frac{-ct^2}{nM}\Big\},$$

where $M = \max_k \|(X_{ki}X_{kj} - \mathbb{E}[Y_iY_j])\|_{\varphi_2}^2$. Therefore,

$$\mathbb{P}[|(X^\top X - n\mathbb{E}[YY^\top])_{ij}| \geq C\sqrt{n \log n}] \leq 2n^{\frac{-2C^2}{M^2}}.$$

A union bound over all $i, j \in [d]$ implies that $\|X^\top X - n\mathbb{E}[YY^\top]\|_F^2 \leq C^2 d^2 n \log n$ with probability at least $1 - 2n^{-2C^2/M^2}$, i.e. $\|X^\top X - n\mathbb{E}[YY^\top]\|_F \leq Cd\sqrt{n \log n}$ with high probability for any $C > \frac{M}{\sqrt{2}}$. By the Hoffman-Wielandt inequality, $|\lambda_i(XX^\top) - n\lambda_i(\mathbb{E}[YY^\top])| \leq Cd\sqrt{n \log n}$, and by reverse triangle inequality, we obtain that

$$\lambda_i(XX^\top) \geq \lambda_d(XX^\top) \geq |n\lambda_d(\Xi)| - Cd\sqrt{n \log n} = \Omega(n)$$

holds almost surely. □

Proposition A.7. *Let $W_1 \Sigma W_2{}^T$ be the singular value decomposition of $U_B^\top U_{\hat{B}}$, then with high probability, $\|U_B^\top U_{\hat{B}} - W_1 W_2{}^\top\| = \mathcal{O}(n^{-1} \log n)$.*

Proof. Let $\sigma_1, \sigma_2, \ldots, \sigma_d$ be the singular values of $U_B^\top U_{\hat{B}}$ (the diagonal entries of Σ). Then $\sigma_i = \cos(\theta_i)$ where θ_i's are the principal angles between the subspace spanned by U_B and $U_{\hat{B}}$. The Davis-Kahan $\sin(\Theta)$ theorem (Davis and Kahan, 1970) gives

$$\|U_{\hat{B}} U_{\hat{B}}^\top - U_B U_B^\top\| = \max_i |\sin(\theta_i)| \leq \frac{C\|B - \hat{B}\|}{\lambda_d(B)} = \mathcal{O}(\sqrt{\frac{\log n}{n}})$$

for sufficiently large n. Note that we have used Proposition A.5 and Lemma A.6 to bound $\|\hat{B} - \hat{B}\|$ and $\lambda_d(B)$ in the above expression, respectively. We thus have

$$\begin{aligned} ||U_B^\top U_{\hat{B}} - W_1 W_2{}^\top||_F = ||\Sigma - I||_F = \sqrt{\sum_{i=1}^d (1 - \sigma_i)^2} &\leq \sum_{i=1}^d (1 - \sigma_i) \\ &\leq \sum_{i=1}^d (1 - \sigma_i{}^2) \\ &= \sum_{i=1}^d \sin(\theta_i)^2 \leq d||U_{\hat{B}} U_{\hat{B}}^\top - U_B U_B^\top||^2 = \mathcal{O}(\frac{\log n}{n}). \end{aligned}$$ □

Recall that a random vector X is sub-exponential if $\mathbb{P}[|X| > t] \leq 2e^{-\frac{t}{K}}$ for some constant K and for all $t \geq 0$. Associated with a sub-exponential random variable there is a Orlicz norm defined as $\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}\exp(\frac{|X|}{t}) \leq 2\}$. Furthermore, a random variable X is sub-Gaussian if and only if X^2 is sub-exponential, and $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$. We now have the following lemma which allows us to juxtapose the ordering in the matrix product $W^*\hat{S}_B$ and $S_B W^*$ (and similarly $W^*\hat{S}_B^{1/2}$ and $S_B^{1/2}W^*$.) This juxtaposition is essential in showing Eq. (16) and Eq. (20) in Lemma A.4.

Lemma A.8. *Let $W^* = W_1 {W_2}^\top$. Then with high probability,*

$$\|W^* S_{\hat{B}} - S_B W^*\|_F = \mathcal{O}(\log n); \quad \textit{and} \quad \|W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*\|_F = \mathcal{O}(n^{-\frac{1}{2}} \log n).$$

Proof. Let $R = U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}$. Note R is the residual after projecting $U_{\hat{B}}$ orthogonally onto the column space of U_B, and thus $\|U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}\|_F \leq \min_W \|U_{\hat{B}} - U_B W\|_F$ where the minimization is over all orthogonal matrices W. By a variant of the Davis-Kahan $\sin\Theta$ theorem (Yu et al., 2015), we have

$$\min_W \|U_B W - U_{\hat{B}}\|_F \leq \frac{C\sqrt{d}\|B - \hat{B}\|}{\lambda_d(B)},$$

and hence $\|R\|_F \leq \mathcal{O}(\sqrt{\frac{\log n}{n}})$. Now consider

$$\begin{aligned} W^* S_{\hat{B}} &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top U_{\hat{B}} S_{\hat{B}} \\ &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top \hat{B} U_{\hat{B}} \\ &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top (\hat{B} - B) U_{\hat{B}} + U_B^\top B U_{\hat{B}} \\ &= (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top (\hat{B} - B) R + U_B^\top (\hat{B} - B) U_B U_B^\top U_{\hat{B}} + S_B U_B^\top U_{\hat{B}}. \end{aligned}$$

Note here we use the fact $U_{\hat{B}} S_{\hat{B}} = \hat{B} U_{\hat{B}}$. Now write

$$S_B U_B^\top U_{\hat{B}} = S_B (U_B^\top U_{\hat{B}} - W^*) + S_B W^*.$$

We then have

$$\begin{aligned} W^* S_{\hat{B}} - S_B W^* = (W^* - U_B^\top U_{\hat{B}}) S_{\hat{B}} + U_B^\top (\hat{B} - B) R + U_B^\top (\hat{B} - B) U_B U_B^\top U_{\hat{B}} \\ + S_B (U_B^\top U_{\hat{B}} - W^*). \end{aligned}$$

Let $\zeta = \|W^* S_{\hat{B}} - S_B W^*\|_F$. Then

$$\begin{aligned} \zeta &\leq \|U_B^\top (\hat{B} - B) R\|_F + \|U_B^\top (\hat{B} - B) U_B U_B^\top U_{\hat{B}}\|_F \\ &\leq \|(U_B^\top U_{\hat{B}} - W^*)\|_F (\|S_{\hat{B}}\| + \|S_B\|) + \|U_B^\top (\hat{B} - B) R\|_F \\ &+ \|U_B^\top (\hat{B} - B) U_B U_B^\top U_{\hat{B}}\|_F \\ &\leq \|W_1 W_2^\top - U_B^\top U_{\hat{B}}\|_F (\mathcal{O}(n) + \mathcal{O}(n)) + \|U_B^\top (\hat{B} - B) R\|_F \\ &+ \|U_B^\top (\hat{B} - B) U_B\|_F \\ &\leq \mathcal{O}(n^{-1})(\mathcal{O}(n) + \mathcal{O}(n)) + \mathcal{O}(\log n) + \|U_B^\top (\hat{B} - B) U_B\|_F \\ &= \mathcal{O}(\log n) + \|U_B^\top (\hat{B} - B) U_B\|_F. \end{aligned}$$

Now consider the term $U_B^\top (\hat{B} - B) U_B \in \mathbb{R}^{d \times d}$. If we denote U_i be the ith column of U_B, then for each i, jth entry, we have

$$(U_B^\top (\hat{B} - B) U_B)_{ij} = U_i^\top (\hat{B} - B) U_j = \frac{1}{2} V_i^\top (\Delta^2 - D^2) V_j,$$

where $V = PU_B$. Furthermore, we have

$$V_i^\top(\Delta^2 - D^2)V_j = \sum_{k,l} V_{ik}({\Delta_{kl}}^2 - {D_{kl}}^2)V_{jl}. \tag{21}$$

We recall that the X_k's are sub-Gaussian. Eq. (21) is thus sum of mean zero sub-exponential random variables and hence, by Bernstein's inequality (Vershynin, 2018), we have

$$\begin{aligned}
&\mathbb{P}[|\sum_{k,l}({\Delta_{kl}}^2 - {D_{kl}}^2)V_{ik}V_{jl}| > t] \\
&\le 2\exp\Big\{-C\min\Big(\frac{t^2}{M^2\sum_{k,l}{V_{ik}}^2{V_{kl}}^2}, \frac{t}{M\max_{k,l}(V_{ik}V_{jl})}\Big)\Big\},
\end{aligned}$$

where $M := \max_{k,l}\|{\Delta_{kl}}^2 - {D_{kl}}^2\|_{\psi_1}$. Since $\sum_k {V_{ik}}^2 \le 1$ for all i, the entries of the $d \times d$ matrix $U_B^\top(\hat{B} - B)U_B \in \mathbb{R}^{d\times d}$ are uniformly bounded by $\mathcal{O}(\log n)$, and

$$\|U_B^\top(\hat{B} - B)U_B\|_F = \mathcal{O}(\log n). \tag{22}$$

This gives $\|W^* S_{\hat{B}} - S_B W^*\|_F = \mathcal{O}(\log n)$, with high probability.

Finally, consider $\|W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*\|_F$. The i, jth entry of $W^* S_{\hat{B}}^{1/2} - S_B^{1/2} W^*$ is

$$\begin{aligned}
{W^*}_{ij}({\lambda_j}^{1/2}(\hat{B}) - {\lambda_i}^{1/2}(B)) &= {W^*}_{ij}\frac{\lambda_j(\hat{B}) - \lambda_i(B)}{{\lambda_j}^{1/2}(\hat{B}) + {\lambda_i}^{1/2}(B)} \\
&\le {W^*}_{ij}\frac{\lambda_j(\hat{B}) - \lambda_i(B)}{\Omega(\sqrt{n})} = \mathcal{O}(n^{-\frac{1}{2}}\log n),
\end{aligned}$$

as desired. Note that we had used the first part of this lemma to derive the bound for the last inequality above. □

We now proceed to prove Lemma A.4.

Proof of Lemma A.4. We now show Eq. (16). We have

$$\begin{aligned}
&\sqrt{n}\|(\hat{B} - B)U_B(W^* S_{\hat{B}}^{-1/2} - S_B^{-1/2}W^*)\|_F \\
&\le \sqrt{n}\|(\hat{B} - B)U_B\| \times \|W^* S_{\hat{B}}^{-1/2} - S_B^{-1/2}W^*\|_F \\
&\le \sqrt{n}\|(\hat{B} - B)\| \times \|W^* S_{\hat{B}}^{-1/2} - S_B^{-1/2}W^*\|_F \\
&= \sqrt{n}\mathcal{O}(\sqrt{n\log n})\mathcal{O}(n^{-\frac{3}{2}}\log n) = \frac{C\log n\sqrt{\log n}}{\sqrt{n}},
\end{aligned}$$

which converges to 0 as $n \to \infty$.

Let us now consider Eq. (17). Recall that $PZ = U_B S_B^{1/2} W$ for some orthogonal matrix W, and since the Z_i's are sub-Gaussian, $\|Z_i\|$ is bounded by some constant C with high probability, i.e., $\|Z_i\| = \sqrt{\sum_{j=1}^{d}\sigma_j {U_{Bij}}^2} \le C$ with high probability, where σ_i's are the diagonal entries of $S_B^{1/2}$. Note that $\sigma_i = \Omega(n) \ge C'n$

for all i and some constant C'. We thus obtain $\sqrt{\sum_{j=1}^{d} {U_B}_{ij}^2} \leq \frac{C}{\sqrt{n}}$, i.e., $\max_i \|(U_B)_i\|_{\ell_2} \leq \frac{C\sqrt{\log n}}{\sqrt{n}}$ with high probability. Hence,

$$\begin{aligned}
&\max_i \|[U_B U_B^\top(\hat{B}-B)U_B W^* S_{\hat{B}}^{-1/2}]_i\|_{\ell_2} \\
&\leq \max_i \|(U_B)_i\|_{\ell_2} \times \|U_B^\top(\hat{B}-B)U_B\| \times \|S_{\hat{B}}^{-1/2}\| \\
&\leq \frac{C\sqrt{\log n}}{\sqrt{n}} \mathcal{O}(\log n)\mathcal{O}(n^{-\frac{1}{2}}) \leq \frac{C\log^{3/2} n}{n},
\end{aligned}$$

which also converges to 0 as $n \to \infty$ (note that we used Eq. (22) in bounding the last inequality).

To show Eq. (18), we must bound $\|[(I-U_B U_B^\top)(\hat{B}-B)(\hat{U}_B - U_B W^*)S_{\hat{B}}^{-1/2}]_i\|$. Define

$$\begin{aligned}
G_1 &= (I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} S_{\hat{B}}^{-1/2}, \\
G_2 &= (I - U_B U_B^\top)(\hat{B}-B)U_B(U_B^\top U_{\hat{B}} - W^*)S_{\hat{B}}^{-1/2}.
\end{aligned}$$

Note that $(I - U_B U_B^\top)(\hat{B}-B)(\hat{U}_B - U_B W^*){S_{\hat{B}}}^{-1/2} = G_1 + G_2$. We now only need to bound the i-th row of G_1 and G_2. We have

$$\begin{aligned}
\|G_2\|_F &\leq \|(I - U_B U_B^\top)(\hat{B}-B)U_B\| \times \|U_B^\top U_{\hat{B}} - W^*\|_F \times \|{S_{\hat{B}}}^{-\frac{1}{2}}\| \\
&\leq \|(I - U_B U_B^\top)\| \times \|\hat{B}-B\| \times \|U_B^\top U_{\hat{B}} - W^*\|_F \times \|{S_{\hat{B}}}^{-\frac{1}{2}}\| \\
&= \mathcal{O}(1)\mathcal{O}(\sqrt{n\log n})\mathcal{O}(n^{-1})\mathcal{O}(n^{-\frac{1}{2}}) = \mathcal{O}\Big(\frac{\sqrt{\log n}}{n}\Big).
\end{aligned}$$

Thus $\|\sqrt{n}G_2\|_F$ converges to 0 as $n \to \infty$. We now consider the rows of G_1. Note that $U_{\hat{B}}^\top U_{\hat{B}} = I$ and hence

$$\begin{aligned}
\|(G_1)_h\| &= \|[(I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} S_{\hat{B}}^{-1/2}]_i\| \\
&= \|[(I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top U_{\hat{B}} S_{\hat{B}}^{-1/2}]_i\| \\
&= \|U_{\hat{B}} S_{\hat{B}}^{-1/2}\| \times \|[(I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top]_i\| \\
&\leq \frac{C}{\sqrt{n}}\|[(I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top]_i\|.
\end{aligned}$$

Let us define $H_1 = (I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top$. Since the Z_i are i.i.d., the rows of H_1 are exchangeable and hence, for any fixed index i, $n\mathbb{E}\|(H_1)_i\|^2 = \mathbb{E}[\|H_1\|_F^2]$. Markov's inequality then implies

$$\begin{aligned}
\mathbb{P}[\|\sqrt{n}(H_1)_i\| > t] &\leq \frac{n\mathbb{E}\|[(I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top]_i\|^2}{t^2} \\
&= \frac{\mathbb{E}\big(\|(I - U_B U_B^\top)(\hat{B}-B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top\|_F^2\big)}{t^2}.
\end{aligned}$$

Furthermore,

$$\|(I - U_B U_B^\top)(\hat{B} - B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top\|_F \leq \|\hat{B} - B\| \times \|U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}\|_F.$$

We now recall the following two observations

- The optimization problem $\min_{T \in \mathbb{R}^{d\times d}} \|U_{\hat{B}} - U_B T\|_F{}^2$ is solved by $T = U_B^\top U_{\hat{B}}$.
- By theorem 2 of Yu et al. (2015), there exists $W \in \mathbb{R}^{d\times d}$ orthogonal, such that
$$\|U_{\hat{B}} - U_B W\|_F \leq C\|U_{\hat{B}} U_{\hat{B}}^\top - U_B U_B^\top\|_F.$$

Combining the two facts above, we conclude that $\|U_{\hat{B}} - U_B U_B^\top U_{\hat{B}}\|_F{}^2 \leq \frac{C}{n}$ with high probability, as in Lemma A.8, hence

$$\|(I - U_B U_B^\top)(\hat{B} - B)(I - U_B U_B^\top)U_{\hat{B}} U_{\hat{B}}^\top\|_F \leq \mathcal{O}(\sqrt{n \log n})\frac{C}{\sqrt{n}} = \mathcal{O}(\sqrt{\log n}),$$

with high probability. Therefore,

$$\mathbb{P}(\|\sqrt{n}(H_1)_i\| > t) \leq \frac{\sqrt{\log n}}{t^2}.$$

Letting $t = n^{\frac{1}{4}}$, we get $\lim_{n\to\infty} Cn^{-1/2}\|\sqrt{n}(H_1)_i\| = 0$. Finally, Eq. (19) and Eq. (20) follow from Lemma A.7 and Lemma A.8 and the bound $\max_i \|(U_B)_i\|_{\ell_2} \leq C(\log n)^{1/2} n^{-1/2}$. □