Classical Multidimensional Scalings: Central Limit Theorems and Random Forests

by

Gongkai Li

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

May, 2019

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 dissertation, 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.

Now the second part is for Random Forest

Primary Reader: Carey E. Priebe

Secondary Reader: Minh Tang

Acknowledgments

The past five years in graduate school has been nothing but pure joy and excitement for me. Much of the previous statement is only true because my advisor, Carey Priebe, without whom this journey will be impossible, let along enjoyable, so I want to thank him for his unconditional support and valuable advice at all levels. To Minh Tang and Avanti Athreya, I want to express my uttermost admiration for putting up with me and explain things over and over again to me when I am confused (which is most of the time). I also want to thank my co-authors and collaborators, including Nicolas Charon, Vince Lyzinski, Youngser Park and Joshua Vogelstein, who have all offered their help with patience in discussions over the past few years.

Special thanks to Dan Naiman, Daniel Robinson for their time and suggestions in my Candidacy Exam and Raman Arora and Katia Consani for their time and suggestions in my Graduate Board Exam.

I also want express my appreciation for the faculty and staff of the Applied Mathematicss and Statistics Department at Johns Hopkins University. In particular, my thanks to John Wierman, Edward Scheinerman, Fred Torcaso, Donniell Fishkind,

ACKNOWLEDGMENTS

Tamás Budavári for their kindness advice; and to Kristin Bechtel, Sandy Kirt, and Ann Gibbins for their help. My deepest thanks to my fellow students and friends at Hopkins for coffee, soccer games, and all the time we have spent together. Finally, my warmest thanks to my parents, Weiyang Li and Huiping Xu, whose unconditional love made me who I am (well, at least the good part).

Dedication

This thesis is dedicated to my parents and the Schaufelds

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Multidimensional Scaling

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] to 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 Borg and Groenen [2005], Cox and Cox [2008], de Leeuw and Heiser [1982], Torgerson [1952]. The popular classical multidimensional scaling (CMDS) dimensionality reduction method provides an example of such an embedding scheme into Euclidean space, in which we have readily available tools to perform statistical inference. Furthermore, CMDS also forms the basis for several other more recent approaches to nonlinear dimension reduction and manifold learning Chen and Buja [2009], Schölkopf et al. [1998], such as Isomap Tenenbaum et al. [2000] and Random Forest manifold learning Criminisi and Shotton [2013] among others.

Although widely used, the behavior of CMDS 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." This paper addresses this acknowledged gap in the literature.

1.2 Random Forest for Manifold Learning

Chapter 2

Classical Multidimensional Scaling with Perturbation

2.1 Review of Classical Multidimensional Scaling

Given an $n \times n$ hollow symmetric dissimilarity matrix D, and an embedding dimension d, we seek $X \in \mathbb{R}^{n \times d}$, where the rows $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ 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 distances given by the dissimilarity matrix D. For a given matrix H, we shall denote by $H^{(2)} = H \circ H$ the element-wise squaring of the matrix H. Given D, classical multidimensional scaling involves the following steps:

1. Compute the matrix $B = -\frac{1}{2}PD^{(2)}P$ where $P = I - 1_n 1_n^\top / n$ is the double centering matrix. Here I denotes the $n \times n$ identity matrix and $1_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$.

2. Extract the d largest positive eigenvalues $s_1, \ldots, s_d$ of B and the corresponding eigenvectors $u_1, \ldots, u_d$.

3. Let $X = U_B S_B^{1/2} \in \mathbb{R}^{n \times d}$, where $U_B = (u_1, \ldots, u_d)$ and $S_B = \mathrm{diag}(s_1, \ldots, s_d)$. Each row of X represents the coordinate of a point in $\mathbb{R}^d$.

In essence, the procedure 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, resulting in an inherent issue of identifiability: X is unique only up to an orthogonal transformation. In the following presentation, we will write $X = U_B S_B^{1/2} W$ where W is some orthogonal matrix, for a suitably transformed X.

2.2 Noise Models and Embedding

2.2.1 Model 1: $\Delta^2 = D^2 + E$

In this section we propose three different but related noise models for the matrix of observed dissimilarities. Suppose that we have a latent or unobserved matrix D

of inter-point Euclidean distances between n points in $\mathbb{R}^d$, i.e. $D_{ij} = \|x_i - x_j\|$. Let $D^{(2)}$ denote the entry-wise square of D and Δ be the observed dissimilarity matrix, such as that measured via a scientific experiment.

The first noise model we consider is $\Delta^{(2)} = D^{(2)} + E$ where we think of $D^{(2)}$ as the signal matrix and E as the noise matrix; see also Zhang et al. [2016]. We shall assume that E satisfies the following conditions:

(i) $\mathbb{E}[E] = 0$, hence $\mathbb{E}[\Delta^{(2)}] = D^{(2)}$.

(ii) The matrix E is hollow and symmetric.

(iii) The entries E_{ij} are independent and $\mathrm{Var}(E_{ij}) = \sigma^2$.

(iv) There exists a finite constant C such that E_{ij} follows a sub-Gaussian distribution with variance proxy C for all i, j, i.e., $\mathbb{P}[E_{ij} \geq t] \leq 2\exp(-t^2/(2C))$ for all i, j.

2.2.2 Model 2: $\Delta = D + E$

The second error model we consider is $\Delta = D + E$. We once more require that the random matrix E satisfies conditions (i) to (iv) identical to that in the model $\Delta^{(2)} = D^{(2)} + E$ along with a constant third and fourth moment conditions, i.e., there exists finite constants γ and ξ such that (v) $\mathbb{E}[E_{ij}^3] \equiv \gamma$ and $\mathbb{E}[E_{ij}^4] \equiv \xi$ for all i, j.

2.2.3 Model 3: Matrix Completion

Finally, we consider a noise model where only a fraction of the entries of D are observed. More specifically, for a given $q \in [0, 1]$ let Δ be such that for $i < j$, with probability q we observe $\Delta_{ij} = D_{ij}$ and with probability $1 - q$, Δ_{ij} is unobserved in which case we set $\Delta_{ij} = 0$. We then have $\Delta = D + E$ where E_{ij} is distributed as $(-D_{ij}) \times \mathrm{Bernoulli}(1 - q)$. Furthermore, $\mathbb{E}[\Delta] = q \cdot D$ and $E[\Delta^{(2)}] = q \cdot D^{(2)}$. This model is motivated by the widely-studied problems of distance matrix completion and sensor localization; see e.g., Alfakih et al. [1999], Chatterjee [2015], Javanmard and Montanari [2013], Patwari et al. [2005].

For each of the above noise models, we shall apply classical multidimensional scaling to the observed Δ to obtain a configuration matrix $\widehat{X}$ whose rows are the estimate of the latent, unobserved $X = [x_1, \dots, x_n]^\top$. A natural question that arises is how the added noise affects the embedding configuration. That is, what is the relationship between the configuration X and $\widehat{X}$ obtained from classical multidimensional scaling of D and Δ ?

2.3 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

$\Delta^{(2)} = D^{(2)} + E$ along with the estimator

$$\widehat{D}^{(2)} = \arg\max_{M \in \mathcal{D}_n^{(2)}} \Big\{ \frac{1}{2} \|\Delta^{(2)} - M\|_F^2 + \lambda_n \text{trace } (-\frac{1}{2} PMP) \Big\}$$

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. Corollary 6 in Zhang et al. [2016] states that under suitable model on E, with probability approaching to one we have

$$\|\widehat{D}^2 - D^2\|_F^2 \leq 36 n \sigma^2 (r+1) \tag{2.1}$$

where σ is the variance of the noise and r is the rank of D^2. In this paper we obtain, as a corollary of ours results, a bound of the same order on $\|\widehat{D}^2 - D^2\|_F^2$. Furthermore, our central limit theorem on the configuration matrix $\widehat{X}$ provides a more refined limiting result, albeit one of a different flavor from Eq. (2.1).

On the other hand, completing a distance matrix with missing entries has been 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] . 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], where the author established an upper bound for the mean squared error on the estimator $\widetilde{M}$ for a general distance matrix M. More specifically, let (K, d) be a compact metric space and $x_1, \ldots, x_n$ be n arbitrary points in K. Let M be the $n \times n$ matrix whose ij-entry is $d(x_i, x_j)$. Let $\epsilon > 0$ be such that $q \geq n^{-1+\epsilon}$. For a given $\delta > 0$, let $N(\delta)$ be the covering number of K using balls of radius ϵ with respect to the metric d. Then there exists an estimator $\widetilde{M}$ obtained by truncating the singular value decomposition of M such that

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

where c and C are constants depending on the truncation level η for the singular values of M and $C(\epsilon)$ is a constant depending only on ϵ and η. Of particular interest is the application of this theorem to the Euclidean distance matrix, for which we obtain roughly

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

Another recent result for the configuration $\widehat{X}$ obtained from the incomplete distance matrix Δ is Theorem 1 of Taghizadeh [2014] which states that, with high probability

$$\|\widehat{X} - X\|_F \leq \mathcal{O}\Big(\frac{\sqrt{n}}{\sqrt{q}}\Big).$$

Our central limit theorem in this paper improves upon both results. It is worth mentioning that the Euclidean distance matrix completion problem can also be viewed from an optimization point of view. See Tasissa and Lai [2018] for a review of such approaches.

2.4 Main Theorems

Recall that a random variable X is sub-Gaussian if $\mathbb{P}[|X| > t] \leq 2e^{-\frac{t^2}{K^2}}$ for some constant K and for all $t \geq 0$. Associated with a sub-Gaussian random variable is a Orlicz norm defined as $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(\frac{X^2}{t^2}) \leq 2\}$. A random vector X in $\mathbb{R}^n$ is called sub-Gaussian if the one-dimensional marginals $\big\langle X, x \big\rangle$ are sub-Gaussian random variables for all $x \in \mathbb{R}^n$, and the corresponding sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup\limits_{x \in S^{n-1}} \|\big\langle X, x \big\rangle\|_{\psi_2}$.

We now present central limit theorems for the rows of the classical multidimensional scaling configuration $\widehat{X}$ for the three noise models in § 2.2.1. Intuitively speaking, the theorems established that the rows of $\widehat{X}$, after some orthogonal transformation, is approximately normally distributed around the rows of X. Furthermore, the covariance matrix 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 2.4.1 depends only on the variance σ^2 of the noise E_{ij}. This is in contrast with the covariance matrices of the model $\Delta = D + E$ and the model $\mathbb{E}[\Delta] = qD$ in Theorem 2.4.3 and Theorem 2.4.4, 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 a sketch of the proof. For ease of exposition, we denote by $(A)_i$ the i-th row of a matrix.

Theorem 2.4.1 (central limit theorem for $\Delta^2 = D^2 + E$) *Let $Z_1, \ldots, Z_n$ be independent and identically distributed according to a multivariate sub-Gaussian distribution F on $\mathbb{R}^d$. Let D be the Euclidean distance matrix generated by the Z_k's, i.e. $D_{ij} = \|Z_i - Z_j\|$. Let $\Delta^2 = D^2 + E$ where the noise matrix E satisfy the conditions (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. Denote by $\widehat{X}_n$ the classical multidimensional scaling embedding configurations of Δ into $\mathbb{R}^d$. 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*

$$\lim_{n\to\infty} \mathbb{P}\{n^{1/2}[(\widehat{X}_n W_n)_i - (Z_i - \bar{Z})] \leq \alpha\} = \Phi(\alpha, \Sigma)$$

where $\bar{Z}$ is the mean of Z_k's and $\Phi(\alpha, \Sigma)$ denotes the cumulative distribution function 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 2.4.2 *We can relax the common variance requirement (iii) in Theorem 2.4.1. Let Var$(E_{ij}) = \sigma_{ij}^2$ and suppose that, for a fixed i, the collection $(D_{ij}^2 - \Delta_{ij}^2)(Z_j - \mathbb{E}[Z_j])$ for $j \neq i$ satisfies the conditions for the Lindeberg-Feller central limit theorem. Let $\Sigma_i = n^{-1}\sum_j \sigma_{ij}^2 \mathrm{cov}(Z_k)$. We obtain the following variant of Theorem 2.4.1:*

$$n^{1/2}\Sigma_i^{-\frac{1}{2}}\big((\widehat{X_n}W_n)_i - (Z_i - \bar{Z})\big) \to \mathcal{N}(0, I).$$

Theorem 2.4.3 (Central Limit Theorem for $\Delta = D + E$) *Let $Z_1, \ldots, Z_n$ be independent and identically distributed according to a multivariate sub-Gaussian distribution F on $\mathbb{R}^d$. Let D be the Euclidean distance matrix generated by the Z_k's, i.e. $D_{ij} = \|Z_i - Z_j\|$. Let $\Delta = D + E$ and suppose that the noise matrix E satisfy, in addition to the conditions in Theorem 2.4.1, the condition (v) $\mathbb{E}[E_{ij}^3] \equiv \gamma$ and $\mathbb{E}[E_{ij}^4] \equiv \xi$. Denote by $\widehat{X}_n$ the classical multidimensional embedding of Δ into $\mathbb{R}^d$. 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,*

$$\lim_{n\to\infty} \mathbb{P}\{n^{1/2}\big((\widehat{X}_n W_n)_i - (Z_i - \bar{Z})\big) \leq \alpha\} = \int \Phi(\alpha, \Sigma(z)) dF(z)$$

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

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

is a covariance matrix depending on z.

Theorem 2.4.4 (Central Limit Theorem for $\Delta = D$ with missing entries) *Let $Z_1, \ldots, Z_n$ be independent and identically distributed according to a multivariate 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 $\widehat{X}_n$ the classical multidimensional 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,*

$$\lim_{n\to\infty} \mathbb{P}\{n^{1/2}[(\widehat{X_n}W_n)_i - q_n^{1/2}(Z_i - \bar{Z})] \leq \alpha\} = \int \Phi(\alpha, \Sigma(z))dF(\mathrm{z})$$

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 $\Sigma(z) = \Xi^{-1}\widetilde{\Sigma}(z)\Xi^{-1}$, $\Xi = \mathrm{cov}(Z_i) \in \mathbb{R}^{d\times d}$ and with $\mu = \mathbb{E}[Z_i] \in \mathbb{R}^d$,

$$\widetilde{\Sigma}(z) = \frac{1 - q_n}{4} \times \mathbb{E}_{Z_k}\Big[\|z - Z_k\|^4 (Z_k - \mu)(Z_k - \mu)^\top\Big]$$

is a covariance matrix depending on z.

Remark 2.4.5 *We emphasize that, in the statement of Theorem 2.4.4, $(\widehat{X_n}W_n)_i$ is centered around $q_n^{1/2}(Z_i - \bar{Z})$ and not around $Z_i - \bar{Z}$. Therefore, unless q_n is known or that an identifiability condition is specified, the classical multidimensional scaling configuration $\widehat{X}_n$ will only recovers an estimate of $Z - 1_n\bar{Z}$ up to an orthogonal transformation W_n and a scaling factor $q_n^{1/2}$.*

2.5 Empirical Results

2.5.1 Three Point-mass Simulated Data

As a simple illustration of our central limit theorem, we embed noisy Euclidean distances obtained from n points into $\mathbb{R}^2$. For illustrative purpose, we will focus on the error model $\Delta = D + E$ as in Theorem 2.4.3. Experimental results for the other error models are completely analogous. 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 $\pi = [0.2, 0.3, 0.5]^\top$. The resulting Euclidean inter-point distance matrix D is then subjected to uniform noise, yielding $\Delta = D + E$ where $E_{ij} \overset{i.i.d.}{\sim} \text{Uniform}(-4, +4)$ for $i < j$ and $E_{ij} = E_{ji}$. For this setting, our central limit theorem for the classical multidimensional embedding of Δ into $\mathbb{R}^2$ yields class-conditional Gaussians. For $n \in \{50, 100, 1000\}$, Figure 2.1 compares, for one realization, the theoretical vs. estimated means and covariances matrices (95% level curves). Table 2.1 shows the empirical covariance matrix for one of the point masses, $\widehat{\Sigma}^{(1)}$, behaving in accordance with Theorem 2.4.3.

Table 2.1 investigates the empirical covariance matrix for one of the point masses, and its entry-wise variance, as a function of n. The theoretical covariance matrix is

$$\Sigma^{(1)} = \begin{bmatrix} 13.56 & -3.06 \\ -3.06 & 22.65 \end{bmatrix}$$

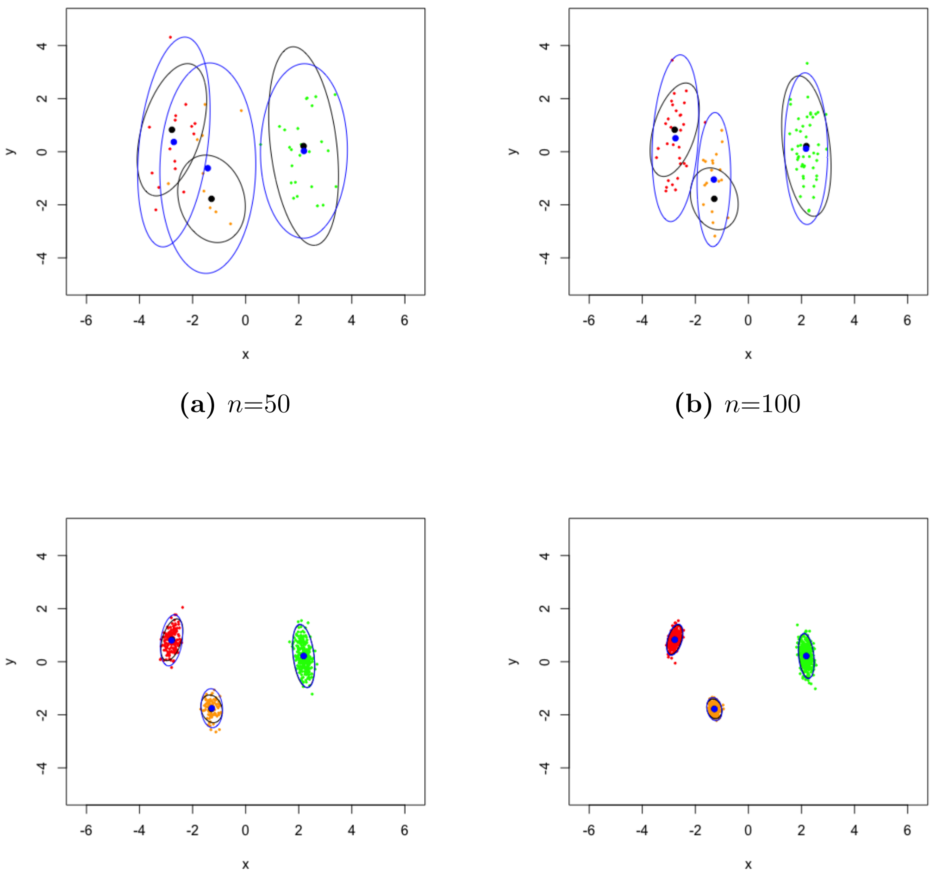
	n=50	n=100	n=500	n=1000
$\widehat{\Sigma}^{(1)}$:	$\begin{bmatrix} 14.15 & 0.25 \\ 0.25 & 79.07 \end{bmatrix}$	$\begin{bmatrix} 13.67 & -0.79 \\ -0.79 & 98.96 \end{bmatrix}$	$\begin{bmatrix} 13.65 & -2.34 \\ -2.34 & 41.02 \end{bmatrix}$	$\begin{bmatrix} 13.63 & -2.70 \\ -2.70 & 31.76 \end{bmatrix}$
$\mathrm{Var}\begin{bmatrix} \widehat{\Sigma}^{(1)}_{11} \\ \widehat{\Sigma}^{(1)}_{12} \\ \widehat{\Sigma}^{(1)}_{22} \end{bmatrix}$:	$\begin{bmatrix} 41.25 \\ 113.31 \\ 829.52 \end{bmatrix}$	$\begin{bmatrix} 19.29 \\ 68.06 \\ 984.45 \end{bmatrix}$	$\begin{bmatrix} 3.67 \\ 7.87 \\ 31.71 \end{bmatrix}$	$\begin{bmatrix} 1.71 \\ 3.25 \\ 11.08 \end{bmatrix}$

Table 2.1: Empirical average of covariance matrix $\widehat{\Sigma}^{(1)}$, and entry-wise variance (500 simulations).

Remark 2.5.1 *In this simulation we relax the requirement that the entries of Δ should be nonnegative in order to illustrate the phenomenon of decreasing covariance with increasing n.*

2.5.2 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. In this experiment, we consider boundary curves obtained from silhouettes of the Kimia shape database. Specifically, 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 2.2 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



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

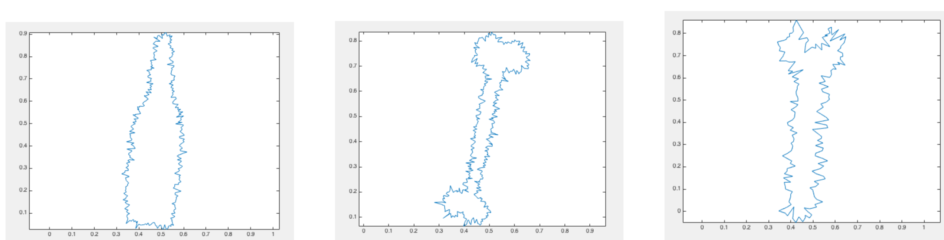
(c) n=500 (d) n=1000

Figure 2.1: Simulation results for n=50, 100, 500 and 1000 points, as described in Section 2.5.1. 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 in Theorem 2.4.3. Note that the blue and black centers and ellipses coincide for large n.

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 such as the ones displayed in Figure 2.3.

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

Figure 2.2: Examples from the Kimia Dataset.



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

Figure 2.3: 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. In this setting, we can view the resulting distance matrix as a perturbation of the ideal distances between the 9 template curves, which fits into the generic framework of our model. (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 $\widehat{d} = 3$ (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 2.4. This configuration exhibits nine fairly well-separated clusters roughly centered around the position of each of the noiseless template curves. Those, in turn, form 3 'super-clusters' consistent with the classes. Furthermore, the ellipsoidal shape of each cluster suggests that the configuration approximately follows a Gaussian distribution.

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 theorem derived here could then constitute a useful theoretical tool to evaluate the discriminating power of shape clustering methods based on CMDS.

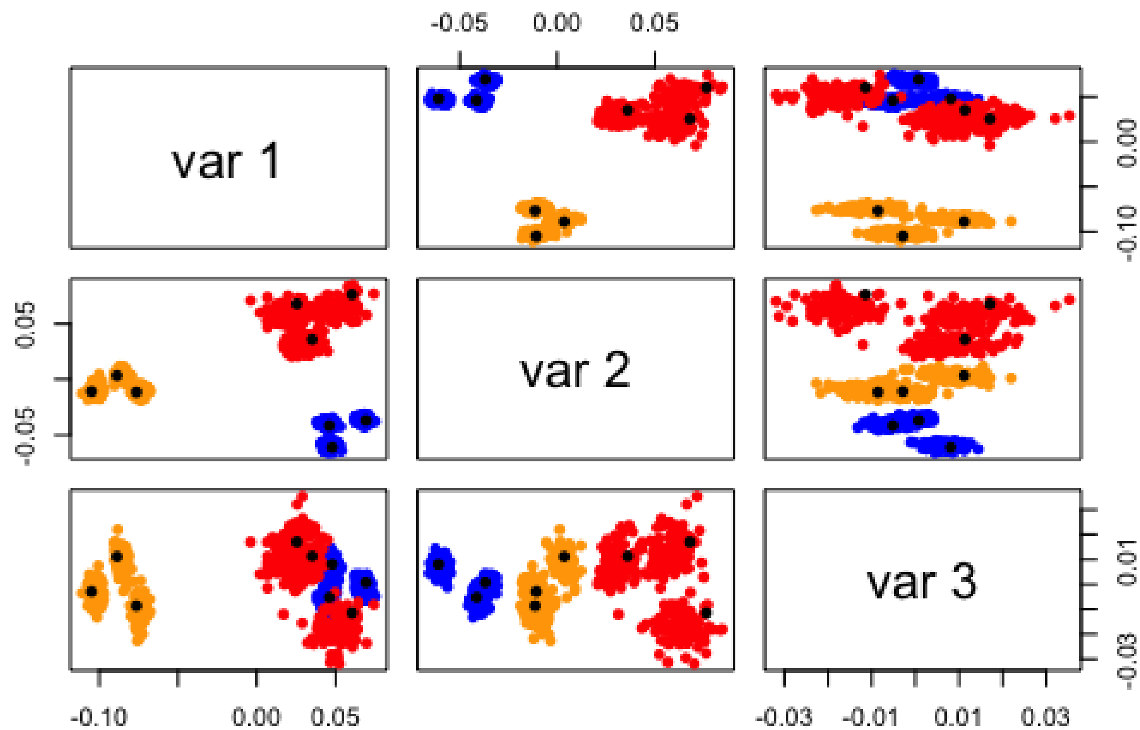
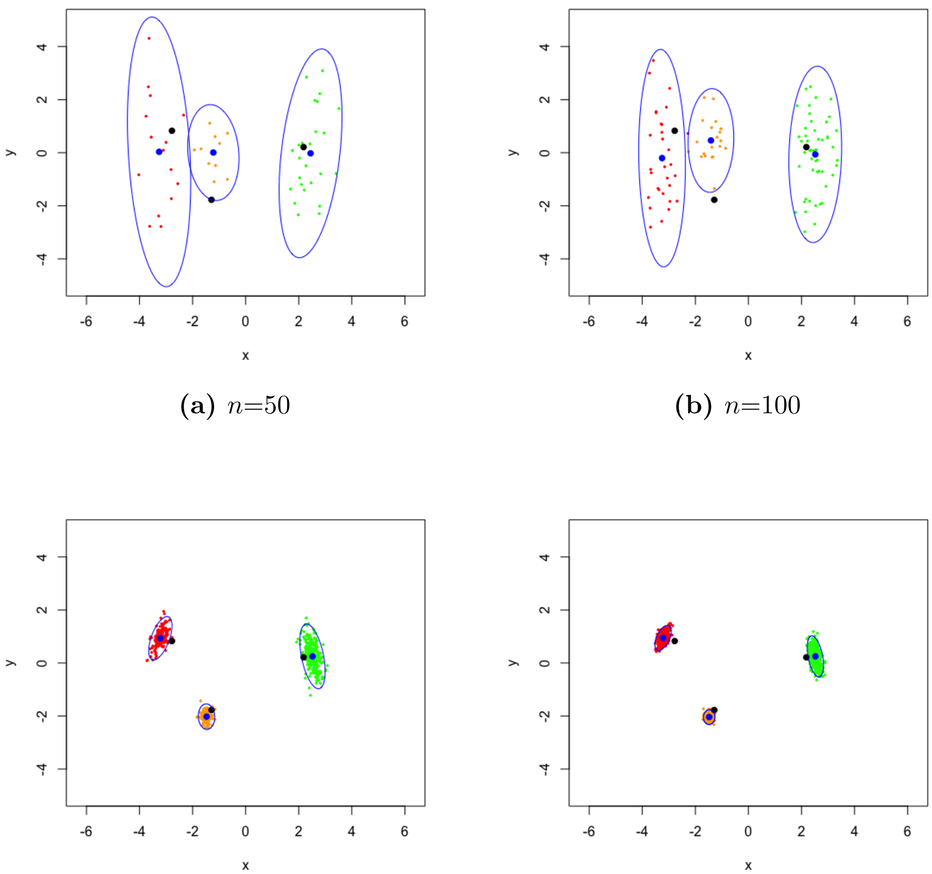


Figure 2.4: Pairs plot of CMDS into $\mathbb{R}^3$ for the noisy curves. Colors correspond to the different classes (blue for bottle, red for bone, and orange for wrench). The position of the nine template curves in the configuration are highlighted with large black dots.

2.6 Discussion

In Athreya et al. [2016] and Levin et al. [2017], the authors prove that adjacency spectral embedding of the random dot product graph gives rise to a central limit theorem for the estimated latent positions. In this work we extend these results to the previously unexplored area of perturbation analysis for CMDS, addressing a gap in the literature as acknowledged in Fan et al. [2018] and Peterfreund and Gavish [2018]. Notably, the three noise models we proposed in Section 2.2.1 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 the shape clustering data all demonstrated the validity of our results. We have avoided any discussion of the model selection problem of choosing a suitable embedding dimension $\widehat{d}$. Instead, we assume d is known – except in Section 4.2. There are many methods for choosing (spectral) embedding dimensions, see Chatterjee [2015], Jackson [1991], Zhu and Ghodsi [2006].

A practically relevant and conceptually illustrative example comes from relaxing the assumption of common variance for the entries of the noise matrix E in Section 2.2.2: the consistency result from Theorem 2.4.3 no longer holds. To illustrate this point, we return to our three-point-mass simulation presented in Section 2.5.1 and modify our noise model as follows: Let $\widetilde{E}_{ij} \overset{i.i.d.}{\sim} \text{Uniform}(-D_{ij}, +D_{ij})$ for $i < j$ and $\widetilde{E}_{ij} = \widetilde{E}_{ji}$. (The noise now depends on the entries of D, and $\Delta = D + \widetilde{E}$ no longer has



(a) n=50

(b) n=100

(c) n=500

(d) n=1000

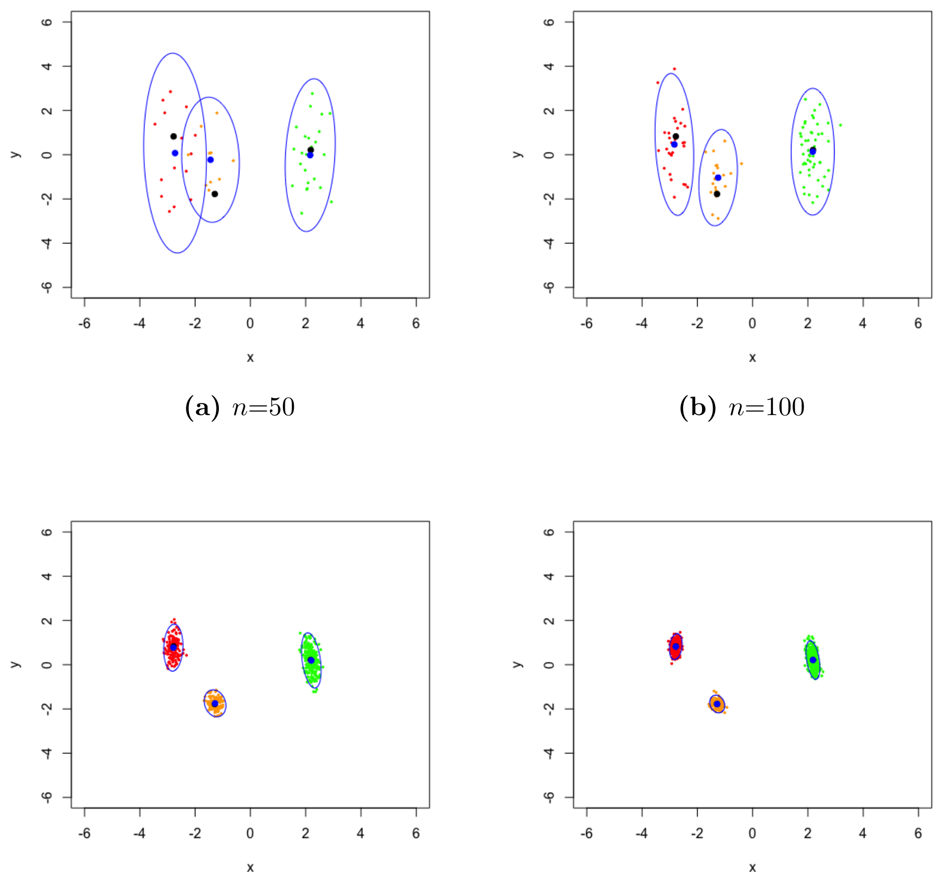
Figure 2.5: Simulation of CMDS with heteroscedastic noise $\widetilde{E}$. The black dots are the true positions for the three points. The blue dots are the empirical means and the blue ellipses are the 95% level curve of the empirical covariance matrix. Note that $\widetilde{E}$ used in this simulation is of the same order for the off-diagonal blocks as that used in Figure 2.1. NB: there is asymptotic bias.

negative entries.) The embedding of Δ into two dimensions gives class-conditional Gaussians; however, we have introduced bias into the embedding configuration. Figure 2.5 shows, for one realization, the embedding result. Note that the empirical mean and the theoretical positions do not coincide in simulation with large n, and theoretically even in the limit.

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 2.5.1, the resulting configuration is shown in Figure 2.6. 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.



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

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

Figure 2.6: 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.

2.7 Application: Omni Embedding of graphs and Hypothesis Testing

2.8 Proof of the Theorems

2.8.1 Proof of Theorem 2.4.3

We proceed to give a complete proof for Theorem 2.4.3. Theorems 2.4.1 and 2.4.4 will have different covariance matrix structures then what is given in Lemma 2.8.3 and will be dealt with later.

Given a matrix A, we denote by $\|A\|$ and $\|A\|_F$ its spectral and Frobenius norm, respectively. We will utilize the following observation repeatedly in our presentation.

Observation 2.8.1 *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 the following notations for the subsequent presentation. Recall that $B = -\frac{1}{2}PD^{(2)}P$ and $\widehat{B} = -\frac{1}{2}P\Delta^{(2)}P$ are the double centering of $D^{(2)}$ and $\Delta^{(2)}$, respectively. If $D^{(2)}$ is a Euclidean distance matrix whose elements are $D_{ij} = \|Z_i - Z_j\|$, then $B = PZZ^\top P$. In particular, $U_B S_B^{1/2} = PZ\widetilde{W}$ for some $d \times d$

orthogonal matrix $\widetilde{W}$. The ith row of $U_B S_B^{1/2}$ is then $\widetilde{W}_n^\top(Z_i - \bar{Z})$. Now let W^* be the orthogonal matrix satisfying $W^* = \arg\min_W \|U_B^\top U_{\widehat{B}} - W\|_F$. The following lemma provides a decomposition for $\widehat{X} - U_B S_B^{1/2} W^*$ into a sum of several matrices.

Lemma 2.8.2 *Let W^* be the orthogonal matrix satisfying $W^* = \arg\min_W \|U_B^\top U_{\widehat{B}} - W\|$. Then*

$$
\begin{aligned}
\widehat{X} - U_B S_B^{1/2} W^* &= (\widehat{B} - B) U_B S_B^{-1/2} W^* && (2.2)\\
&- (\widehat{B} - B) U_B (S_B^{-1/2} W^* - W^* S_{\widehat{B}}^{-1/2}) - U_B U_B^\top (\widehat{B} - B) U_B W^* S_{\widehat{B}}^{-1/2} && (2.3)\\
&+ (I - U_B U_B^\top)(\widehat{B} - B)(U_{\widehat{B}} - U_B W^*) S_{\widehat{B}}^{-1/2} && (2.4)\\
&+ U_B (U_B^\top U_{\widehat{B}} - W^*) S_{\widehat{B}}^{1/2} + U_B (W^* S_{\widehat{B}}^{1/2} - S_B^{1/2} W^*) && (2.5)
\end{aligned}
$$

Proof: We have

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

We used the facts $U_B U_B^\top B = B$ and $U_{\widehat{B}} S_{\widehat{B}}^{1/2} = \widehat{B} U_{\widehat{B}} S_{\widehat{B}}^{-1/2}$ in the above equalities. The last two terms of the above display is Eq. (2.5). Denote by R the term $(I -$

$U_B U_B^\top)(\widehat{B} - B) U_{\widehat{B}} S_{\widehat{B}}^{-1/2}$, we have

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

The four terms in the above display are identical to that in Eq. (2.2) through Eq. (2.4).

■

Lemma 2.8.2 implies

$$\widehat{X} W^{*\top} \widetilde{W}_n - U_B S_B^{1/2} \widetilde{W}_n = \widehat{X} W^{*\top} \widetilde{W}_n - PZ = (\widehat{B} - B) U_B S_B^{-1/2} \widetilde{W}_n + R_n \widetilde{W}_n$$

where R_n are the matrices in Eq. (2.3) through Eq. (2.5). The essential term is $(\widehat{B} - B) U_B S_B^{-1/2} \widetilde{W}_n$. We analyzed the rows of this matrix in Lemma 2.8.3 where we show that they converge to multivariate normals. Meanwhile, Lemma 2.8.4 shows that the rows of the matrices R_n, when scaled by $n^{1/2}$, converge to 0 in probability. Combining these results yield Theorem 2. A few minor changes to the covariance computation in the proof of Lemma 2.8.3 also yield Theorem 2.4.1 and Theorem 2.4.4.

Lemma 2.8.3 *Let $Z_1, \dots, Z_n$ be independent and identically distributed according to some multivariate sub-Gaussian distribution F. Then there exists a sequence of $d \times d$*

orthogonal matrices $\widetilde{W}_n$, such that for any fixed index i with $Z_i = z_i$, we have

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

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 z. Here $(A)_i$ or $[A]_i$ denote the ith row of a matrix A.

Proof: Recall that $PZ = U_B S_B^{1/2}\widetilde{W}_n$. We therefore have

$$\begin{aligned}
n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)U_B S_B^{-1/2}]_i &= n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)PZ\widetilde{W}_n^\top {S_B}^{-1}]_i \\
&= n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n[(\widehat{B} - B)PZ]_i \\
&= -n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n[P(D \circ E + \frac{E^2}{2})PZ]_i \\
&= -n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n\Big[P\Big(D \circ E + \frac{E^2 - \sigma^2 1_n 1_n^\top}{2}\Big)PZ\Big]_i
\end{aligned}$$

The last equality holds since $P1_n = 0$. Now $PZ = Z - 1_n\bar{Z} = Z - 1_n\mu^\top + \widetilde{R}_n$ where

$\|\widetilde{R}_n\| = O(n^{-1/2})$ with high probability. Therefore,

$$n^{1/2}\widetilde{W}_n^\top[(\widehat{B} - B)U_B S_B^{-1/2}]_i = -\, n\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n\Big[n^{-1/2}\sum_{j\neq i}^{n}\Big(D_{ij}E_{ij} + \frac{E_{ij}^2 - \sigma^2 1_n 1_n^\top}{2}\Big)(Z_j - \mu)\Big] + o(1)$$

Conditioning on $Z_i = z_i$ and ignoring the term $o(1)$ that vanishes as $n \to \infty$, the above expression is sum of $n-1$ independent mean 0 random vector. We then invoke the Lindeberg-Feller central limit theorem to show that this sum converges to a multivariate normal. We now evaluate the covariance matrix for this sum. Each summand has covariance matrix of the form

$$\mathrm{cov}[(E_{ij}D_{ij} + \frac{E_{ij}^2 - \sigma^2}{2})(Z_j - \mu)] = \mathrm{Var}\Big(E_{ij}\|z_i - Z_j\| + \frac{E_{ij}^2 - \sigma^2}{2}\Big)(Z_j - \mu)(Z_j - \mu)^\top.$$

Since $\mathbb{E}[E_{ij}] = 0$ and $\mathbb{E}[E_{ij}^2] = \sigma^2$, we also have

$$\mathrm{Var}\Big(E_{ij}\|z_i - Z_j\| + (E_{ij}^2 - \sigma^2)/2\Big) = \mathbb{E}\Big[E_{ij}^2\|z_i - Z_j\|^2 + E_{ij}\|z_i - Z_j\|(E_{ij}^2 - \sigma^2) + \frac{(E_{ij}^2 - \sigma^2)^2}{4}\Big]$$

where the expectation is taken with respect to E_{ij} and conditional on Z_j. Averaging over the indices j and then taking the limit as $n \to \infty$ yields

$$\begin{aligned}\widetilde{\Sigma}_n(z_i) &= \mathrm{Var}\Big[n^{-1/2}\sum_{j\neq i}^{n}\Big(D_{ij}E_{ij} + \frac{E_{ij}^2 - \sigma^2 1_n 1_n^\top}{2}\Big)(Z_j - \mu)\Big] \\ &\longrightarrow \mathbb{E}_{Z_k}\Big[\Big(\sigma^2\|z_i - X_k\|^2 + \mathbb{E}[E_{ij}^3]\|z_i - Z_k\| + \frac{1}{4}\mathbb{E}[E_{ij}^4] - \frac{\sigma^4}{4}\Big)(Z_k - \mu)(Z_k - \mu)^\top\Big].\end{aligned}$$

By the strong law of large numbers, we have

$$\frac{\widetilde{W}_n^\top S_B \widetilde{W}_n}{n} = \frac{1}{n} Z^\top P Z \to \Xi \in \mathbb{R}^{d\times d}$$

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

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

as desired. ∎

Finally we state the following lemma showing that any row of these matrices, when scaled by $n^{1/2}$, converges to 0 in probability.

Lemma 2.8.4 *For any fixed index i, we have, simultaneously*

$$n^{1/2}[(\widehat{B}-B)U_B(W^* S_{\widehat{B}}^{-1/2} - S_B^{-1/2}W^*)]_i \xrightarrow{P} 0 \tag{2.6}$$

$$n^{1/2}[U_B U_B^\top(\widehat{B}-B)U_B W^* S_{\widehat{B}}^{-1/2}]_i \xrightarrow{P} 0 \tag{2.7}$$

$$n^{1/2}[(I - U_B U_B^\top)(\widehat{B}-B)(\widehat{U}_B - U_B W^*)S_{\widehat{B}}^{-1/2}]_i \xrightarrow{P} 0 \tag{2.8}$$

$$n^{1/2}[U_B(U_B^\top U_{\widehat{B}} - W^*)S_{\widehat{B}}^{1/2}]_i \xrightarrow{P} 0. \tag{2.9}$$

$$n^{1/2}[U_B(W^* S_{\widehat{B}}^{1/2} - S_B^{1/2}W^*)]_i \xrightarrow{P} 0. \tag{2.10}$$

The rest of this section is devoted toward proving Lemma 2.8.4, for which we need the following technical lemmas controlling the spectral norm of $\|\widehat{B}-B\|$ and

$\|U_B^\top \widehat{U}_B - W^*\|$ (recall that W^* is the closest orthogonal matrix, in Frobenius norm, to $U_B^\top \widehat{U}_B$.) We start with a bound for the spectral norm of $B - \widehat{B}$.

Proposition 2.8.5 $\|B - \widehat{B}\| = \mathcal{O}(\sqrt{n \log n})$ *with high probability.*

Proof: We have

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

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 **?** gives the desired result.

■

Lemma 2.8.6 *Let* $X_1, \ldots, X_n, Y \overset{i.i.d}{\sim} F$ *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* $\Xi = \mathbb{E}[X_1 {X_1}^\top]$ *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 $i, j \in [d]$,

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

where $M = \max_k \|(X_{ki}X_{kj} - \mathbb{E}[Y_iY_j])\|^2_{\varphi_2}$. 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 $\|X^\top X - n\mathbb{E}[YY^\top]\|_F^2 \leq C^2d^2 n\log n$ 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

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

holds almost surely. ■

Proposition 2.8.7 *Let $W_1\Sigma {W_2}^T$ be the singular value decomposition of $U_B^\top U_{\widehat{B}}$, then with high probability, $\|U_B^\top U_{\widehat{B}} - W_1{W_2}^\top\| = \mathcal{O}(n^{-1}\log n)$.*

Proof: Let $\sigma_1, \sigma_2, \ldots, \sigma_d$ be the singular values of $U_B^\top U_{\widehat{B}}$ (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 $U_{\widehat{B}}$. The Davis-Kahan $\sin(\Theta)$ theorem [Davis and Kahan, 1970] gives

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

for sufficiently large n. Note in the last equality we used the previous two lemmas. Thus,

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

■

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 $W^*\widehat{S}_B$ and $S_B W^*$ (and similarly $W^*\widehat{S}_B^{1/2}$ and $S_B^{1/2}W^*$.) This juxtaposition is essential in showing Eq. (2.6) and Eq. (2.10) in Lemma 2.8.4.

Lemma 2.8.8 *Let $W^* = W_1 {W_2}^\top$. Then with high probability,*

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

Proof: Let $R = U_{\widehat{B}} - U_B U_B^\top U_{\widehat{B}}$. Note R is the residual after projecting $U_{\widehat{B}}$ orthogonally onto the column space of U_B, and thus $\|U_{\widehat{B}} - U_B U_B^\top U_{\widehat{B}}\|_F \leq \min_W \|U_{\widehat{B}} - U_B W\|_F$ 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

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

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

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

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

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

then we have

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

This gives

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

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

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

where $V = P U_B$. Furthermore, we have

$$V_i^\top (\Delta^2 - D^2) V_j = \sum_{k,l} V_{ik} ({\Delta_{kl}}^2 - {D_{kl}}^2) V_{jl}. \tag{2.11}$$

Recall, since X_k's are sub-Gaussian, thus equation (2.11) is a sum of mean zero sub-exponential random variables. By Bernstein's inequality [Vershynin, 2018], we have

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

where $M:=\max_{k,l}\|{\Delta_{kl}}^2-{D_{kl}}^2\|_{\psi_1}$. Since $\sum_k {V_{ik}}^2\leq 1\forall i$, we have that each entry of $U_B^\top(\widehat{B}-B)U_B\in\mathbb{R}^{d\times d}$ is $\mathcal{O}(\log n)$, and

$$\|U_B^\top(\widehat{B}-B)U_B\|_F=\mathcal{O}(\log n). \tag{2.12}$$

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

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

$$W^*{}_{ij}({\lambda_j}^{1/2}(\widehat{B})-{\lambda_i}^{1/2}(B))=W^*{}_{ij}\frac{\lambda_j(\widehat{B})-\lambda_i(B)}{{\lambda_j}^{1/2}(\widehat{B})+{\lambda_i}^{1/2}(B)}\leq W^*{}_{ij}\frac{\lambda_j(\widehat{B})-\lambda_i(B)}{\Omega(\sqrt{n})}=\mathcal{O}(n^{-\frac{1}{2}}\log n),$$

as desired (note in the last inequality, we used the first part of this Lemma. ■

We now proceed to prove Lemma 2.8.4. **Proof:** [Proof of Lemma 2.8.4] To show Eq. (2.6), we have

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

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

Let us now consider Eq. (2.7). Recall that $X = U_B S_B^{1/2} W$ for some orthogonal matrix W, and since X_i's are sub-Gaussian, $\|X_i\|$ is bounded by some constant C with high probability, i.e., $\|X_i\| = \sqrt{\sum_{j=1}^{d} \sigma_j {U_{Bij}}^2} \leq C$ with high probability, where σ_i's are the diagonal entries of $S_B^{1/2}$. Note that $\sigma_i = \Omega(n) \geq C' n$ for all i and some constant C'. We thus obtain $\sqrt{\sum_{j=1}^{d} {U_{Bij}}^2} \leq \frac{C}{\sqrt{n}}$, i.e., $||U_B||_{2\to\infty} \leq \frac{C}{\sqrt{n}}$. Hence,

$$
\begin{aligned}
\|[U_B U_B^\top (\widehat{B} - B) U_B W^* S_{\widehat{B}}^{-1/2}]_h\| &\leq \|U_B\|_{2\to\infty} \|U_B^\top (\widehat{B} - B) U_B\| \times \|S_{\widehat{B}}^{-1/2}\| \\
&\leq \frac{C}{\sqrt{n}} \mathcal{O}(\log n) \mathcal{O}(n^{-\frac{1}{2}}) \leq \frac{C \log n}{n}
\end{aligned}
$$

which also converges to 0 as $n \to \infty$ (note in the last inequality we used 2.12).

To show Eq. (2.8), we must bound $\|[(I - U_B U_B^\top)(\widehat{B} - B)(\widehat{U}_B - U_B W^*) S_{\widehat{B}}^{-1/2}]_h\|$. Define

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

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

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

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

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

Define

$$
H_1 = (I - U_B U_B^\top)(\widehat{B} - B)(I - U_B U_B^\top)U_{\widehat{B}} U_{\widehat{B}}^\top.
$$

Since the Z_i are i.i.d., the rows of H_1 are exchangeable and hence, for any fixed index

h, $n\mathbb{E}\|(H_1)_h\|^2 = \mathbb{E}[\|H_1\|_F^2]$. Markov's inequality then implies

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

Furthermore,

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

We now recall the following two observations

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

Combining the two facts above, we conclude that $\|U_{\widehat{B}} - U_B U_B^\top U_{\widehat{B}}\|_F{}^2 \leq \frac{C}{n}$ with high probability, as in Lemma 2.8.8, hence

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

with high probability. Therefore,

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

picking $t = n^{\frac{1}{4}}$, we get $\lim_{n\to\infty} Cn^{-1/2}\|\sqrt{n}(H_1)_h\| = 0$.

Finally, Eq. (2.9) and Eq. (2.10) follow from Lemma 2.8.7 and Lemma 2.8.8 and the bound $\|U_B\|_{2\to\infty} \le Cn^{-1/2}$. ■

2.8.2 Adaptation for Theorem 2.4.1 and 2.4.4

The major difference between our main theorems is the calculation of the covariance matrices. In this section, we will give those calculations.

Lemma 2.8.9 *Let $Z_1, \ldots, Z_n$ be independent and identically distributed according to some multivariate sub-Gaussian distribution F and let our model be as in Theorem 2.4.1. Then there exists a sequence of $d \times d$ orthogonal matrices $\widetilde{W}_n$, such that for any fixed index i, we have*

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

where $\Sigma = \frac{\sigma^2}{4}\Xi^{-1}$, $\Xi = \mathrm{cov}(Z_k)$. Here $(A)_i$ or $[A]_i$ denote the ith row of a matrix A.

Proof: Recall that $PZ = U_B S_B^{1/2} \widetilde{W}_n$. We therefore have

$$
\begin{aligned}
n^{1/2}\widetilde{W}_n^\top[(\widehat{B}-B)U_B S_B^{-1/2}]_i &= n^{1/2}\widetilde{W}_n^\top[(\widehat{B}-B)PZ\widetilde{W}_n^\top {S_B}^{-1}]_i \\
&= n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n[(\widehat{B}-B)PZ]_i \\
&= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n[P(D^{(2)}-\Delta^{(2)})PZ]_i \\
&= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n\Big[P(D^{(2)}-\Delta^{(2)})(I-1_n 1_n^\top/n)Z\Big]_i \\
&= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n\Big[P(D^{(2)}-\Delta^{(2)})(Z-1_n\mu^\top)\Big]_i \\
&(\text{ since } PZ = Z - 1_n\bar{Z} = Z - 1_n\mu^\top) \\
&= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n\Big[(D^{(2)}-\Delta^{(2)})(Z-1_n\mu^\top) \\
&- 1_n 1_n^\top/n(D^{(2)}-\Delta^{(2)})(Z-1_n\mu^\top)\Big]_i \\
&(\text{note that } \frac{1_n 1_n^\top}{n}(D^{(2)}-\Delta^{(2)})(Z-1_n\mu^\top) \to 0 \text{ as } n\to\infty) \\
&= \frac{1}{2}n^{1/2}\widetilde{W}_n^\top {S_B}^{-1}\widetilde{W}_n\Big[(D^{(2)}-\Delta^{(2)})(Z-1_n\mu^\top)\Big]_i \text{ as } n\to\infty
\end{aligned}
$$

Therefore,

$$
n^{1/2}\widetilde{W}_n^\top[(\widehat{B}-B)U_B S_B^{-1/2}]_i = \frac{1}{2}n\widetilde{W}_n^\top S_B^{-1}\widetilde{W}_n\Big[n^{-1/2}\sum_{j\neq i}^{n}\Big(D_{ij}^{(2)}-\Delta_{ij}^{(2)}\Big)(Z_j-\mu)\Big]
$$

Conditioning on $Z_i = z_i$, the above expression is sum of $n-1$ independent mean 0 random vectors. We then invoke the Lindeberg-Feller central limit theorem to show that this sum converges to a multivariate normal. We now evaluate the covariance matrix for this sum. Each summand has covariance matrix of the form

$$\begin{aligned}\operatorname{cov}\Big[(D_{ij}^{(2)} - \Delta_{ij}^{(2)})(Z_j - \mu)\Big] &= \operatorname{Var}\Big(D_{ij}^{(2)} - \Delta_{ij}^{(2)}\Big)(Z_j - \mu)(Z_j - \mu)^\top \\ &= \sigma^2 (Z_j - \mu)(Z_j - \mu)^\top \\ &\text{Since } \mathbb{E}[E_{ij}] = 0 \text{ and } \mathbb{E}[E_{ij}^2] = \sigma^2\end{aligned}$$

. By the strong law of large numbers, we have

$$\frac{\widetilde{W}_n^\top S_B \widetilde{W}_n}{n} = \frac{1}{n} Z^\top P Z \to \Xi \in \mathbb{R}^{d \times d}$$

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

$$n^{1/2} \widetilde{W}_n^\top [(\widehat{B} - B) U_B S_B^{-1/2}]_i \longrightarrow \mathcal{N}(0, \frac{\sigma^2}{4} \Xi^{-1})$$

as desired. ■

Lemma 2.8.10

Bibliography

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. doi: 10.1023/A:1008655427845. 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. doi: 10.1007/s13171-015-0071-x. URL https://doi.org/10.1007/s13171-015-0071-x.

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

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

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. doi: 10.1007/BF02310791. 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. doi: 10.1214/14-AOS1272.

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. doi: 10.1198/jasa.2009.0111. URL `https://doi.org/10.1198/jasa.2009.0111`.

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. doi: 10.1007/978-3-540-33037-0_14. URL `https://doi.org/10.1007/978-3-540-33037-0_14`.

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.

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.

J. Fan, Q. Sun, W. X. Zhou, and Z. Zhu. Principal component analysis for big data, January 2018. arXiv:1801.01602.

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. 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. doi: 10.1007/s10208-012-9129-5. URL `https://doi.org/10.1007/s10208-012-9129-5`.

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 random dot product graphs, 05 2017. arXiv:1705.09355.

J. A. T. Machado and M. E. Mata. Analysis of world economic variables using multidimensional scaling. *PLOS ONE*, 10(3):1–17, 03 2015. doi: 10.1371/journal.pone.0121277. URL `https://doi.org/10.1371/journal.pone.0121277`.

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. doi: 10.1109/ITWKSPS.2010.5503144.

N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal. Locating the nodes: cooperative localization in wireless sensor networks. *IEEE Signal Processing Magazine*, 22(4):54–69, July 2005. ISSN 1053-5888. doi: 10.1109/MSP.2005.1458287.

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.

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

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. doi: 10.1073/pnas.0709842104. URL http://www.pnas.org/content/105/28/9507.

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

M. J. Taghizadeh. Theoretical analysis of Euclidean distance matrix completion for ad hoc microphone array calibration. 2014.

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

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. doi: 10.1126/science.290.5500.2319. URL http://science.sciencemag.org/content/290/5500/2319.

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

R. Vershynin. *High-Dimensional Probability An Introduction with Applications in Data Science*. 2018. URL https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf.

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. doi: 10.1126/science.1250298. 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.

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. doi: 10.1111/rssb.12138. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12138`.

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.

Vita

Gongkai (Percy) Li received the B.S. degree in mathematics from the Lehigh University in 2014, the M.S.E. degree in Applied Mathematics and Statistics from the Johns Hopkins University in 2017 and enrolled in the Applied Mathematics and Statistics Ph.D. program at Johns Hopkins University in 2014. He has received the Teaching Fellow Recognition Award.