Efficient Manifold and Subspace Approximations with Spherelets

Didong Li[1], Minerva Mukhopadhyay[2] and David B Dunson[1,2]
Department of Mathematics[1] and Statistical Science[2], Duke University

Data lying in a high dimensional ambient space are commonly thought to have a much lower intrinsic dimension. In particular, the data may be concentrated near a lower-dimensional subspace or manifold. There is an immense literature focused on approximating the unknown subspace, and in exploiting such approximations in clustering, data compression, and building of predictive models. Most of the literature relies on approximating subspaces using a locally linear, and potentially multiscale, dictionary. In this article, we propose a simple and general alternative, which instead uses pieces of spheres, or spherelets, to locally approximate the unknown subspace. Theory is developed showing that spherelets can produce lower covering numbers and MSEs for many manifolds. We develop spherical principal components analysis (SPCA). Results relative to state-of-the-art competitors show gains in ability to accurately approximate the subspace with fewer components. In addition, unlike most competitors, our approach can be used for data denoising and can efficiently embed new data without retraining. The methods are illustrated with standard toy manifold learning examples, and applications to multiple real data sets.

Key Words: Curvature; Data visualization; Denoising; Dimensionality reduction; Manifold learning; Spherical principal component analysis; Subspace learning.

1 Introduction

Given a data set $X = \{X_i\}_{i=1}^n$, where $X_i \in \mathbb{R}^D$ are independent and identically distributed (i.i.d.) from a probability measure ρ, it is common to assume that ρ is supported on or near a lower dimensional subspace or manifold M with dimension $d \ll D$. The lower-dimensional subspace is commonly assumed to be linear; for example, principal components analysis (PCA) and latent factor models assume a linear subspace.

PCA is an extremely useful and versatile approach, including for dimension reduction, data visualization, and as a key component of many unsupervised and supervised learning algorithms. However, the linearity assumption of PCA is restrictive. A more flexible assumption is to consider the subspace to be a potentially nonlinear d-dimensional manifold embedded in D-dimensional ambient space. "Manifold learning" is the problem of estimating a mapping from the D-dimensional ambient space of the observed data to the intrinsic d-dimensional space.

A rich variety of algorithms are available for manifold learning and nonlinear dimension reduction, including both global and local methods. Popular global approaches in-

clude Curvilinear-CA (Demartines and Hérault, 1997), Kernel PCA (Schölkopf et al., 1998), Isomap (Tenenbaum et al., 2000), Laplacian eigenmap (Belkin and Niyogi, 2002), Nonlinear PCA (Scholz et al., 2005), Probabilistic PCA (Lawrence, 2005) and Diffusion Map (Coifman and Lafon, 2006). Examples of local methods are local PCA (Kambhatla and Leen, 1997), local linear embedding (Roweis and Saul, 2000) and their variants like Donoho and Grimes (2003), Zhang and Wang (2007). See Lee and Verleysen (2007) for an overview. As for PCA, these methods provide a mapping from the high-dimensional space to the low-dimensional embedding, and can be used for data visualization for $d \leq 3$. Ironically, almost all available algorithms only produce lower dimensional coordinates and do not in general produce an estimate of the manifold. In addition, one cannot embed new data points $i = n+1, \ldots, N$ without completely reimplementing the algorithm as data are added, eliminating the possibility of using cross validation (CV) for out-of-sample assessments and tuning parameter choice. Finally, it is not in general possible to obtain *denoised* estimates $\hat{X}_i$, for $i = 1, \ldots, n$, of the original data exploiting the manifold structure. Real data are often corrupted by measurement errors or noise so that the original data points do not lie exactly on the manifold; in practice, it is also typically not reasonable to assume that the noise is bounded. It is often of substantial utility to exploit an estimate of the manifold structure of the data to adjust for measurement errors and denoise the data; for example, in medical imaging the raw data may consist of error prone measurements of a tumor or organ that is known to have a smooth boundary, a 2-dimensional manifold embedded in $\mathbb{R}^3$, and the goal is to denoise the data and estimate the boundary.

Relative to competitors ranging from local linear embeddings (LLE) to diffusion maps, our spherelets approach has the major advantages of producing an estimate of the manifold, allowing new data points to be embedded quickly and efficiently without reimplementing the algorithm, and easily providing denoised estimates of the data adjusting for measurement errors.

Most other methods that have similar advantages rely on variations of locally projecting data in a neighborhood to the best approximating hyperplane (see Walder and Schölkopf (2009), Chen and Maggioni (2011), Allard et al. (2012), Maggioni et al. (2016), Liao and Maggioni (2016), Liao et al. (2016) and Xianxi et al. (2017)). Key disadvantages of such approaches are the lack of statistical efficiency and parsimony of the representation, particularly when the true manifold has large curvature. For example, if the manifold is a circle with small radius (large curvature), a huge number of tangent lines will be needed to fit the circle well and achieve a small approximation error. Such disadvantages also hold when M is not a single manifold but is a more complex collection of manifolds.

We consider a simple and efficient alternative to PCA, which uses spheres instead of hyperplanes to locally approximate the unknown subspace. One can replace PCA with spherical PCA (SPCA) in any setting in which PCA is used, potentially improving performance. For example, a more accurate approximation to a curved subspace can be obtained by using SPCA within local neighborhoods instead of PCA. Taking curvature into consideration, it is reasonable to use locally quadratic instead of locally linear approximations. However, locally

quadratic approximation increases the number of parameters per local piece from $O(D)$ to $O(D^2)$, while adding considerably to computational complexity. As a practical alternative that generalizes locally linear approximations, we use a collection of spheres to build up a local approximation to a manifold, or to a collection of disconnected manifolds. Spheres with large radius approximate hyperplanes, but when the subspace is not approximately flat, our new sphere-based basis will potentially perform much better. The improvement will increase with the curvature of the unknown subspace being approximated.

We formalize this gain through providing theory bounding the covering number, showing that dramatic gains are possible for manifolds having sufficiently large curvature but not too many subregions having locations with large change in directional curvature. We also prove the consistency of empirical SPCA and provide an upper bound on estimation error of SPCA.

Our simulation results show significant reductions in the number of components needed to provide an approximation within a given error tolerance. Comparing to local PCA, our *spherelets* approach has much lower mean square error using dramatically fewer components. Substantial practical gains are shown in multiple examples.

Subspace approximation is not the only application of spherelets. We apply spherelets to denoise data on manifolds. There are several competitors designed for denoising including Gaussian Blurring Mean Shift (GBMS) (Zemel and Carreira-Perpiñán (2005), Carreira-Perpiñán (2006), Carreira-Perpinán (2008), Hein and Maier (2007b), Hein and Maier (2007a)), Manifold Blurring Mean Shift (MBMS) (Wang and Carreira-Perpinán (2010), Wang et al. (2011)), and its special case Local Tangent Projection (LTP). Spherelets outperforms these methods in terms of denoising performance in our experiments.

Visualization algorithms project data to a lower dimensional space, while preserving geometric structures hidden in the original data set. Popular algorithms include Isomap, Diffusion Map, and t-distributed Stochastic Neighborhood Embedding (tSNE) (Maaten and Hinton, 2008). We show that all these algorithms can be modified easily to produce spherical versions, potentially improving performance.

The paper is organized as follows. In section 2, we prove our main theorem on the covering numbers of hyperplanes and spherelets. In section 3, we develop an algorithm for estimating an optimal approximating sphere, and propose spherical principal component analysis (SPCA). We prove upper bounds on approximation errors along with some convergence properties of empirical SPCA. In section 4, we propose a simple approximation algorithm along with some numerical experiments. We also consider manifold denoising and data visualization, providing algorithms and numerical results. In section 5, we discuss some open problems and future work. Proofs are mainly in the Appendix.

2 Covering Number

We define the covering number as the minimum number of local bases needed to approximate the manifold within ϵ error. Our main theorem shows the covering number of spherelets is smaller than that of hyperplanes.

Definition 1. *Let M denote a $d-$dimensional compact C^3 Riemannian manifold, and $\mathcal{B}$ a dictionary of basis functions. Then the $\epsilon > 0$ covering number $N_{\mathcal{B}}(\epsilon, M)$ is defined as*

$$N_{\mathcal{B}}(\epsilon, M) := \inf_{N} \left\{ N : \exists \{C_k, \text{Proj}_k, B_k\}_{k=1}^{K} \ s.t. \ \left\| x - \sum_{k=1}^{K} \mathbf{1}_{\{x \in C_k\}} \text{Proj}_k(x) \right\| \leq \epsilon, \ \forall x \in M \right\},$$

where $\{C_k\}_{k=1}^{K}$ is a partition of $\mathbb{R}^D$, $B_k \in \mathcal{B}$ is a local basis, $\text{Proj}_k : C_k \to B_k$, $x \mapsto \underset{y \in B_k}{\text{argmin}} \|x - y\|^2$ is the corresponding local projection.

We focus on two choices of $\mathcal{B}$: all d-dimensional hyperplanes in $\mathbb{R}^D$, denoted by $\mathcal{H}$, and the class of all $d-$dimensional spheres in $\mathbb{R}^D$, denoted by $\mathcal{S}$. Including spheres with infinite radius in $\mathcal{S}$, we have $\mathcal{H} \subset \mathcal{S}$ implying the following Proposition.

Proposition 1. *For any compact C^3 Riemannian manifold M, and $\epsilon > 0$,*

$$N_{\mathcal{S}}(\epsilon, M) \leq N_{\mathcal{H}}(\epsilon, M).$$

Before proving a non-trivial upper bound, we define some important features of M.

Definition 2. *Let M be a d-dimensional C^3 Riemannian manifold. The volume of M is denoted by V. Let $K : UTM \to \mathbb{R}$ denote the curvature of M at point p in direction v:*

$$K(p, v) := \left\| \frac{d^2 \exp_p(tv)}{dt^2} \bigg|_{t=0} \right\|,$$

where UTM is the unit sphere bundle over M and $\exp_p(\cdot)$ is the exponential map from the tangent plane at p to M. Then $K := \sup_{(p,v) \in UTM} K(p, v) < \infty$ is the maximum curvature. Similarly,

$$T := \sup_{(p,v) \in UTM} \left\| \frac{d^3 \exp_p(tv)}{dt^3}(0) \right\|$$

is the maximum of the absolute rate of change of the curvature.

Definition 3. *Given any $\epsilon > 0$ and letting T_pM denote the tangent plane to M at $p \in M$,*

$$F_\epsilon := \left\{ p \in M : \sup_{v \in UT_pM} K(p, v) - \inf_{v \in UT_pM} K(p, v) \leq \left(\frac{2\epsilon}{K} \right)^{\frac{1}{2}} \right\}$$

is called the set of ϵ-spherical points on M, where UT_pM is the unit ball in T_pM. Let $B(p,\epsilon)$ be the geodesic ball centered at p with radius ϵ, then $M_\epsilon := \bigcup_{p\in F_\epsilon} B\Big(p, (\frac{6\epsilon}{3+T})^{\frac{1}{3}}\Big)$ is called the spherical submanifold of M, and the volume is $V_\epsilon \coloneqq \mathrm{Vol}(M_\epsilon)$. M is called an ϵ sphere if $V_\epsilon = V$.

Example 1. *A space form, a complete, simply connected Riemmanian manifold of constant sectional curvature, is an ϵ sphere for any manifold dimension d and $\epsilon > 0$.*

Example 2. *A one dimensional manifold (a curve) is an ϵ sphere for any $\epsilon > 0$.*

Example 3. *Let $M = \{(u - \frac{1}{3}u^3 + uv^2, -v - u^2v + \frac{1}{3}v^3, u^2 - v^2) \in \mathbb{R}^3 | u^2 + v^2 \le R^2\}$ be the compact truncation of the Enneper surface, which is an interesting surface in differential geometry that has varying curvature. By definition, M is a compact smooth surface. We calculate the spherical points, spherical submanifold as well as its volume. We just present the results here and the proof is in the appendix.*

Proposition 2. *For the Enneper surface,*

$$F_\epsilon = \begin{cases} \{(u - \frac{1}{3}u^3 + uv^2, -v - u^2v + \frac{1}{3}v^3, u^2 - v^2) | \frac{2}{\sqrt{\epsilon}} - 1 \le u^2 + v^2 \le R^2\} & \epsilon \ge \frac{4}{(R^2+1)^2}, \\ \emptyset & \epsilon < \frac{4}{(R^2+1)^2}. \end{cases}$$

When $\epsilon < \frac{4}{(R^2+1)^2}$, there are no spherical point so $F_\epsilon = M_\epsilon = \emptyset$ and $V_\epsilon = 0$. For $\epsilon \ge \frac{4}{(R^2+1)^2}$, $T = \frac{1}{1024\sqrt{7}}$ for $\frac{1}{\sqrt{7}} \le R$ and $\frac{4R}{(R^2+1)^4}$ for $\frac{1}{\sqrt{7}} > R$. Assuming $R \ge \frac{1}{\sqrt{7}}$,

$$M_\epsilon = \Big\{(u - \frac{1}{3}u^3 + uv^2, -v - u^2v + \frac{1}{3}v^3, u^2 - v^2) | u^2 + v^2 \ge \alpha^2\Big\},$$

where $\alpha = \max\left\{0, \sqrt{\frac{2}{\sqrt{\epsilon}} - 1} - \left(\frac{6\epsilon}{3+\frac{1}{1024\sqrt{7}}}\right)^{\frac{1}{3}}\right\}$, then

$$V_\epsilon = \pi\Big(R^2 + \frac{1}{2}R^4 - \alpha^2 - \frac{1}{2}\alpha^4\Big), \quad \frac{V_\epsilon}{V} = \frac{2R^2 + R^4 - 2\alpha^2 - \alpha^4}{2R^2 + R^4}.$$

An extreme case is $\epsilon \ge \frac{4}{(R^2+1)^2}$ and $\frac{2}{\sqrt{\epsilon}} - 1 - \Big(\frac{6\epsilon}{3+\frac{1}{1024\sqrt{7}}}\Big)^{\frac{1}{3}} \le 0$. In this case, although $F_\epsilon \ne M$, the union of small geodesic balls centered on the spherical points is M, that is, $M_\epsilon = M$, so $\alpha = 0$ and $\frac{V_\epsilon}{V} = 1$, which means M is an ϵ sphere.

Theorem 1 (Main Theorem)**.** *For any $\epsilon > 0$ and compact C^3 d-dimensional Riemannian manifold M,*

$$N_{\mathcal{H}}(\epsilon, M) \lesssim V\Big(\frac{2\epsilon}{K}\Big)^{-\frac{d}{2}}, \tag{1}$$

$$N_{\mathcal{S}}(\epsilon, M) \lesssim V_\epsilon\Big(\frac{6\epsilon}{3+T}\Big)^{-\frac{d}{3}} + (V - V_\epsilon)\Big(\frac{2\epsilon}{K}\Big)^{-\frac{d}{2}}. \tag{2}$$

When $d = 1$, $M = \gamma$ is a curve and we have the following Corollary.

Corollary 1. *For any $\epsilon > 0$ and compact C^3 curve γ, $N_{\mathcal{S}}(\epsilon, \gamma) \lesssim V(\frac{6\epsilon}{T})^{-\frac{d}{3}}$.*

Remark 1. *The curse of dimensionality comes in through the term $\epsilon^{-\frac{d}{2}}$, but we can decrease its impact from $\epsilon^{-\frac{d}{2}}$ to $\epsilon^{-\frac{d}{3}}$.*

Proposition 3. *The upper bounds of covering number $N_{\mathcal{H}}(\epsilon, M)$ and $N_{\mathcal{S}}(\epsilon, M)$ are both tight.*

The bounds in our main theorem are *tight*, implying that spherelets often require many fewer pieces than locally linear dictionaries to approximate M to any fixed accuracy level ϵ; particularly large gains occur when a non-negligible subset of M is covered by the closure of points having not too large change in directional curvature. As each piece involves $O(D)$ unknown parameters, these gains in covering numbers should lead to real practical gains in statistical performance; indeed this is what we have observed in applications. In the next section, we develop algorithms for fitting local data by a sphere, providing a spherical alternative to local PCA.

3 Spherical PCA

3.1 Estimation

Let $S_V(c, r) \coloneqq \{x : \|x - c\| = r, x - c \in V, \dim(V) = d + 1\}$, where c is the center, r is the radius, and V determines an affine subspace the sphere lies in. Our goal in this section is to estimate (V, c, r) to obtain the best approximating sphere through data $X_1, \ldots, X_n$ consisting of samples in $\mathbb{R}^D$.

Lemma 1. *For any $x \in \mathbb{R}^D$, the closest point $y \in S_V(c, r)$ is*

$$\underset{y \in S_V(c,r)}{\operatorname{argmin}} \; d^2(x, y) = c + \frac{r}{\|VV^\top(x - c)\|} VV^\top(x - c),$$

where $VV^\top(x - c)$ is the projection of x onto the affine subspace $c + V$.

In fact, x is projected to an affine subspace first and then to the sphere (see the proof of Lemma 1). To find the optimal affine subspace, we use $d + 1$-dimensional PCA, obtaining

$$\widehat{V} = (v_1, \cdots, v_{d+1}), \quad v_i = \operatorname{evec}_i\{(X - 1\bar{X}^\top)^\top (X - 1\bar{X}^\top)\}, \tag{3}$$

where $\operatorname{evec}_i(S)$ is the ith eigenvector of S in decreasing order. Letting $Y_i = \bar{X} + \widehat{V}\widehat{V}^\top(X_i - \bar{X})$, we then find the optimal sphere through points $\{Y_i\}_{i=1}^n$. A sphere can be expressed as the set of zeros of a quadratic function $y^\top y + f^\top y + b$, where $c = -\frac{1}{2}f$ is the center and $r^2 = \frac{1}{4}f^\top f - b$.

When this quadratic function has positive value, y is outside the sphere, and y is inside the sphere if the function has negative value. Hence, we define the loss function

$$g(f,b) \coloneqq \sum_{i=1}^{n}(Y_i^\top Y_i + f^\top Y_i + b)^2.$$

By simple calculation, we can show that

$$\widehat{b} \coloneqq \underset{b}{\operatorname{argmin}}\ g(f,b) = -\frac{1}{n}\sum_{i=1}^{n}(Y_i^\top Y_i + f^\top Y_i).$$

Hence, if we define

$$g(f) \coloneqq \sum_{i=1}^{n}(Y_i^\top Y_i + f^\top Y_i + \widehat{b})^2 = \sum_{i=1}^{n}\left\{Y_i^\top Y_i + f^\top Y_i - \frac{1}{n}\sum_{j=1}^{n}(Y_j^\top Y_j + f^\top Y_j)\right\}^2,$$

we can then minimize $g(f)$ to obtain $\widehat{f}$ and let $\widehat{c} = -\frac{1}{2}\widehat{f}$. The resulting $\widehat{f}$ is shown in the following Theorem.

Theorem 2. *The minimizer of $g(f)$ is given by*

$$\widehat{f} = -H^{-1}\xi,$$

where

$$H = \sum_{i=1}^{n}\big(Y_i - \bar{Y}\big)(Y_i - \bar{Y})\big)^\top, \quad \xi = \sum_{i=1}^{n}\left(\|Y_i^\top Y_i\| - \frac{1}{n}\sum_{j=1}^{n}\|Y_j^\top Y_j\|\right)\big(Y_i - \bar{Y}\big).$$

We refer to the resulting estimates $(\widehat{V}, \widehat{c}, \widehat{r})$ as *Spherical PCA* (SPCA), where

$$\widehat{V} = (v_1, \cdots, v_{d+1}),\ \widehat{c} = -\frac{1}{2}\widehat{f},\ \widehat{r} = \frac{1}{n}\sum_{i=1}^{n}\|Y_i - \widehat{c}\|. \tag{4}$$

Remark 2. *Alternatively, we could have minimized $\sum_{i=1}^{n} d^2(X_i, S_V(c,r))$, corresponding to the sum of squared residuals. However, the resulting optimization problem is non convex, lacks an analytic solution, and iterative algorithms may be slow to converge, while only producing local minima.*

Corollary 2. *If $X_i \in S_V(c,r)$ for all i, SPCA will find the same minimizer as the loss function in Remark 2, corresponding to exactly (V,c,r).*

Definition 4. *Supposing data $X_i \sim \rho$, we let (V^*, c^*, r^*) denote the values of $(\widehat{V}, \widehat{c}, \widehat{r})$ obtained plugging in exact moments of the population distribution in place of sample values.*

Next we focus on the consistency of empirical SPCA, characterized by $(\widehat{V}, \widehat{c}, \widehat{r})$ and $\widehat{\text{Proj}}(x) := \widehat{c} + \frac{\widehat{r}}{\|\widehat{V}\widehat{V}^\top(x-\widehat{c})\|}\widehat{V}\widehat{V}^\top(x-\widehat{c})$. The corresponding population version is characterized by (V^*, c^*, r^*) and $\text{Proj}^*(x) := c^* + \frac{r^*}{\|V^*V^{*\top}(x-c^*)\|}V^*V^{*\top}(x-c^*)$.

We denote the population variance-covariance matrix and sample variance-covariance matrix by Σ and $\hat{\Sigma}$, respectively. Furthermore, we rely on the following two assumptions:

(A) *Distributional Assumption:* The $n \times D$ data matrix $X = V^*\Lambda^{*1/2}Z$ where $Z = ((z_{i,j}))$ is a $n \times D$ matrix whose elements $z_{i,j}$'s are independent and identically distributed (i.i.d.) non-degenerate random variables with $E(z_{i,j}) = 0$, $E(z_{i,j}^2) = 1$ and $E(z_{i,j}^6) < \infty$.

(B) *Spike Population Model:* If $\lambda_1, \lambda_2, \ldots, \lambda_D$ are the ordered eigenvalues of Λ^* then there exists $m > d$ such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m > \lambda_{m+1} = \ldots = \lambda_D = 1$.

An assumption similar to (A) is also considered in Lee et al. (2010). The spike population model is defined in Johnstone (2001).

Theorem 3. *Under the assumptions A and B, for any x, we have*

$$\widehat{\text{Proj}}(x) \xrightarrow{p} \text{Proj}^*(x), \quad n \to \infty,$$

where $\xrightarrow{p}$ denotes convergence in probability.

3.2 Local SPCA

A single sphere will typically not be sufficient to approximate the entire manifold M, but instead we partition $\mathbb{R}^D$ into local neighborhoods and implement SPCA separately in each neighborhood. This follows similar practice to popular implementations of local PCA, but we apply SPCA locally instead of PCA. We divide $\mathbb{R}^D$ into non-overlapping subsets $C_1, \ldots, C_K$. For the kth subset, we let $X_{[k]} = \{X_i : X_i \in C_k\}$, $(\widehat{V}_k, \widehat{c}_k, \widehat{r}_k)$ denote the results of applying SPCA to data $X_{[k]}$, $\widehat{P}_k$ denote the projection map from $x \in C_k$ to $y \in S_{\widehat{V}_k}(\widehat{c}_k, \widehat{r}_k)$ obtained by Lemma 1, and $\widehat{M}_k = S_{\widehat{V}_k}(\widehat{c}_k, \widehat{r}_k) \cap C_k$. Then, we approximate M by $\widehat{M} = \bigcup_{k=1}^K \widehat{M}_k$.

In general, $\widehat{M}$ will not be continuous or a manifold but instead is made up of a collection of pieces of spheres chosen to approximate the manifold M. There are many ways in which one can choose the subsets $\{C_k\}_{k=1}^K$, but in general the number of subsets K will be chosen to be increasing with the sample size with a constraint so that the number of data points in each subset cannot be too small, as then $\widehat{M}_k$ cannot be reliably estimated. Below we provide theory on mean square error properties of the estimator $\widehat{M}$ under some conditions on how the subsets are chosen but without focusing on a particular algorithm for choosing the subsets.

There are a wide variety of algorithms for multiscale partitioning of the sample space, ranging from cover trees (Beygelzimer et al., 2006) to METIS (Karypis and Kumar, 1998)

to iterated PCA (Szlam, 2009). As the scale becomes finer, the number of partition sets increases exponentially and the size of each set decreases exponentially. Assume $U \subset M$ is an arbitrary submanifold of M, $\rho_U = \rho|_U$ is the density of data X_i conditionally on $X_i \in U$ and

$$\text{diam}(U) = \sup_{x,y \in U} d(x,y) = \alpha.$$

For example, if we bisect the unit cube in $\mathbb{R}^D$ j times, then the diameter of each piece will be $\alpha \sim 2^{-j}$. The approximation error depends on α: as $\alpha \to 0$, each local neighborhood is getting smaller so tangent plane or spherical approximations perform better.

Definition 5. *Given fixed (M, ρ), assume $U \subset M$ is a submanifold of M and ρ_U is the conditional density. Let (V_U, c_U, r_U) be the solution of SPCA (population version, see Definition 4), which depends on U only, then the sphere $S_{V_U}(c_U, r_U)$ is called the spherelet of U.*

Theorem 4. *Assume there exists $\delta > 0$ such that for any submanifold $U \subset M$, $r_U \geq \delta$ where $S_{V_U}(c_U, r_U)$ is the spherelet of U. Then there exists $\theta > 0$ that depends only on (M, ρ) such that for any submanifold $U \subset M$ with spherelet $S_{V_U}(c_U, r_U)$,*

$$E_{\rho_U} d^2(x, S_{V_U}(c_U, r_U)) \leq \theta \alpha^4.$$

Remark 3. *The assumption $r_U \geq \delta$ for all $U \subset M$ is weak and reasonable when M is compact. Recall that $K := \sup_{(p,v) \in UTM} K(p,v) < \infty$ (2) so the curvature at any point $p \in M$ in any direction $v \in T_pM$ is bounded by K. Since SPCA is aimed at finding the best sphere to approximate the manifold locally, the curvature of each spherelet is approximately upper bounded by K. Since the curvature of a circle is the reciprocal of its radius, we have the relation $\frac{1}{r_U} \lesssim K, r_U \gtrsim \frac{1}{K}$. As a result, there exists $\delta \approx \frac{1}{K} > 0$ such that $r_U \geq \delta$ for any $U \subset M$.*

Combining Theorem 3 and Theorem 4, we have the following Corollary:

Corollary 3. *Assume $\widehat{\text{Proj}}$ is the empirical spherical projection obtained from SPCA on $U \subset M$ with $\text{diam}(U) = \alpha$, then under assumptions A, B, there exists $\theta \in \mathbb{R}_+$ that depends only on (M, ρ) such that for any x and $\epsilon > 0$,*

$$\lim_{n \to \infty} P(\|x - \widehat{\text{Proj}}(x)\|^2 > \theta \alpha^4 + \epsilon) = 0.$$

Remark 4. *Corollary 3 does not imply that applying SPCA in local neighborhoods leads to better approximation error than applying PCA. However, this is not surprising since we are not putting restrictions on the curvature. In the special case in which the curvature is zero, we expect SPCA and PCA to have similar performance. However, outside of such an extreme case, when curvature is not very close to zero, SPCA is expected to yield notably improved performance except for very small local regions. This remark is consistent with the empirical results in the following sections.*

4 Applications, Algorithms and Numerical Experiments

Spherelets can be applied to various problems including subspace approximation, manifold denoising and data visualization. In this section we demonstrate these three main applications with algorithms and some benchmark examples. The examples presented are a representative subset of a broader class of examples we have considered, but do not have space to present.

4.1 Subspace approximation

In subspace approximation we attempt to find an estimator of the unknown manifold M, say $\widehat{M}$. As local SPCA provides an estimator of a submanifold $U \subset M$ in a neighborhood, we split $\mathbb{R}^D$ into pieces and apply local SPCA to estimate the manifold in each piece.

There are many existing partitioning algorithms for subdividing the sample space into local neighborhoods. Popular algorithms, such as cover trees, iterated PCA and METIS, have a multi-scale structure, repeatedly bisecting $\mathbb{R}^D$ until a stopping condition is achieved. Given any such algorithm, let $C_1, \ldots, C_K$ be the partitions of the ambient space, and $M_k = C_k \cap M$ be the sub-manifold of M restricted to C_k. We find an estimate $\widehat{M}_k$ of M_k using local SPCA (see Section 3.2 for details), and set $\widehat{M} = \bigcup_{k=1}^{K} \widehat{M}_k$. The map which projects a data point x to the estimated manifold $\widehat{M}$ is denoted by $\text{Proj} : \mathbb{R}^D \to \widehat{M}$. Algorithm 1 explains the calculation of Proj and $\widehat{M}$ given a partition of $\mathbb{R}^D$.

For simplicity, we consider a multi-scale partitioning scheme which iteratively splits the sample space based on the first principal component score until a predefined bound ϵ on mean square error (MSE) is met for each of partition sets or the sample size n_k no longer exceeds a minimal value n_0. The MSE within set C_k is estimated as

$$\text{MSE}_k = \frac{1}{n_k} \sum_{X_i \in X_{[k]}} \|X_i - \text{Proj}\,(X_i)\|^2\,, \tag{5}$$

where $X_{[k]}$ and n_k are as defined in Algorithm 1. If $\text{MSE}_k > \epsilon$ and $n_k > n_0$, we calculate $\text{PC}_1 = (X_{[k]} - \mu_k)v_{1,k}$, where $\mu_k = \bar{X}_{[k]}$ and $v_{1,k}$ is the first eigenvector of the covariance matrix of $X_{[k]}$. Next we split C_k into two sub-partitions $C_{k,1}$ and $C_{k,2}$ based on the sign of PC_1, i.e, i^{th} sample of $X_{[k]}$ is assigned to $C_{k,1}$ if $\text{PC}_{1,i} > 0$ and to $C_{k,2}$ otherwise.

Out-of-sample Projection The projection operator $\widehat{\text{Proj}}$ can be estimated using a training dataset X, and then applied to test dataset Y without retraining to obtain predictive values $\widehat{Y}_i = \widehat{\text{Proj}}(Y_i)$ and a predictive MSE= $\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \|Y_i - \widehat{Y}_i\|^2$. This out of sample MSE provides a useful measure of performance not subject to over-fitting.

Below we validate the performance of local SPCA using a synthetic and a real data example. In each case, we compare our performance with that of local PCA. For PCA we use the same partitioning scheme. We compare PCA and SPCA with respect to the rate of

Algorithm 1: Spherelets

input : Data X; intrinsic dimension d; Partition $\{C_k\}_{k=1}^K$

output: Estimated manifold $\widehat{M}_k$ and projection map $\widehat{\text{Proj}}_k$ for each $k = 1, \ldots, K$; Estimated manifold $\widehat{M}$ of M and the projection map $\widehat{\text{Proj}}$

for $k = 1 : K$ **do**

  Define $X_{[k]} = X \cap C_k$;

  Calculate $\widehat{V}_k, \widehat{c}_k, \widehat{r}_k$, *by 4*;

  Calculate $\widehat{\text{Proj}}_k(x) = \widehat{c}_k + \frac{\widehat{r}_k}{\|\widehat{V}_k \widehat{V}_k^\top (x - \widehat{c}_k)\|}(x - \widehat{c}_k)$;

  Calculate $\widehat{M}_k = S_{\widehat{V}_k}(\widehat{c}_k, \widehat{r}_k) \cap C_k$;

end

Calculate $\widehat{\text{Proj}}(x) = \sum_{k=1}^K \mathbf{1}_{\{x \in C_k\}} \widehat{\text{Proj}}_k(x)$, *and* $\widehat{M} = \bigcup_{k=1}^K \widehat{M}_k$.

decrease in MSE with number of partitions. For the first example where $D = 2,\ d = 1$, we depict the plots of estimated manifolds by representing the different partitions in different colors.

Example 1: Euler Spiral We generate 2500 training and 2500 test samples from manifold:

$$\gamma(s) = \left[\int_0^s \cos(t^2) \mathrm{dt}, \int_0^s \sin(t^2) \mathrm{dt} \right], s \in [0, 2].$$

The Euler spiral is a well studied smooth curve in differential geometry with curvature linear with respect to the arc length parameter, that is, $\kappa(s) = s$ for all s. Figure 1(c) shows the comparative performance of PCA and SPCA with respect to the number of partitions vs MSE. Clearly SPCA has much better performance than PCA, as it requires only 14 partitions to achieve an MSE of about 10^{-4}, while PCA requires 120 partitions to achieve a similar error. Figure 1(a) and (b) show the projected test dataset with different partitions described in different colors. It is clear that there are fewer pieces of circles than tangent lines and the estimated $\widehat{M}$ is smoother in the second panel, reflecting better approximation by SPCA.

Example 2: Seals dataset Seals dataset is available in R-package ggplot2. This dataset has $D = 4$ attributes, and 1155 samples. We choose 867 samples randomly as the training set and the remaining 288 samples as the test set and set $d = 1$. Figure 1(d) shows the comparative performance of PCA and SPCA. From the figure it is clear that for the same number of partitions, the MSE of SPCA is much lower than PCA.

The above examples show that local SPCA has smaller MSE than local PCA given the same number of partitions. Equivalently, given a fixed error ϵ, the number of spheres needed to approximate the manifold is smaller than that of hyperplanes. This coincides with the

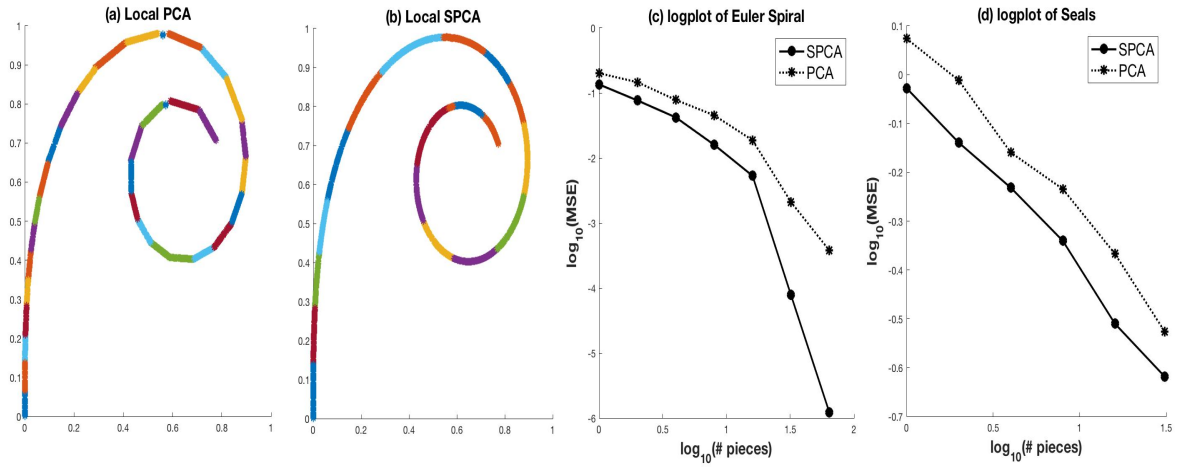

Figure 1: (a) local linear projection of Euler spiral data by local PCA; (b) locally spherical projection of Euler spiral data by local SPCA; (c) logplot of MSE vs. number of pieces for Euler spiral data; (d) logplot of MSE vs. number of pieces for Seals dataset.

statement of Theorem 1. When training sample size is small to moderate, there will be limited data available per piece and local PCA will have high error regardless of the number of pieces.

4.2 Manifold denoising

Another important application of SPCA is manifold denoising, where based on a set of noisy samples one tries to approximate the underlying manifold. Manifold Blurring Mean Shift (MBMS) is designed for denoising data which are assumed to lie in some lower dimensional manifold. MBMS first shifts the original data toward their local mean. Then the shifted data are projected to the tangent space, approximating the manifold by standard PCA. There are several variants of MBMS but the basic idea is very similar. Removing the linear projection step, we obtain Gaussian Blurring Mean Shift (GBMS) which only averages the data using Gaussian weights. If we remove the averaging step, that is, do local linear projection only, the method is called Local Tangent Projection (LTP). Hence, MBMS is a combination of GBMS and LTP so we can expect that MBMS performs the best among the three methods. We modify MBMS by applying spherelets to provide a locally spherical projection instead of linear by replacing local PCA by local SPCA. We call the new algorithm Spherical Manifold Blurring Mean Shift (SMBMS) or Local Spherical Projection (LSP). The SMBMS algorithm is described below.

As before we consider a synthetic and a real data example. We depict the original and denoised data in each case.

Algorithm 2: Manifold denoising (SMBMS)

input : Data $X \in \mathbb{R}^{n\times D}$; intrinsic dimension d; number of neighbors k; Gaussian bandwidth σ

output: Denoised data $\hat{X}$

for $i = 1 : n$ **do**

  Find $X_{[i]}$*, set of k-nearest neighbors of* X_i;

  Shift X_i *toward the Gaussian mean:*

$$Y_i = \sum_{X_j \in X_{[i]}} \frac{\exp\{-\|X_i - X_j\|^2/2\sigma^2\}}{\sum_{X_l \in X_{[i]}} \exp\{-\|X_i - X_l\|^2/2\sigma^2\}} X_j;$$

  For $Y_{[i]}$*, find spherelet* (V_i, c_i, r_i) *by 4*;

  Calculate spherical projection $\hat{X}_i = c_i + \dfrac{r_i}{\|V_i V_i^\top (Y_i - c_i)\|}(Y_i - c_i)$

end

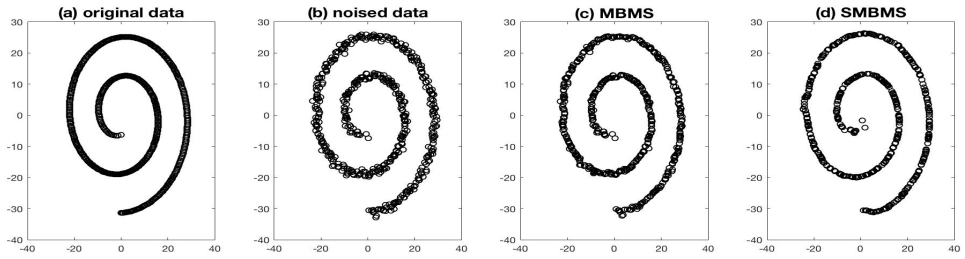

Figure 2: Noisy spiral. (a) clean data; (b) noisy data; (c) denoised data by MBMS; (d) denoised data by SMBMS.

Example 3: Noisy Spiral We generate 500 samples from the spiral with equation:

$$\gamma(t) = [2t\cos(t), 2t\sin(t)], t \in [\pi, 4\pi].$$

Then we add white noise to get the noisy spiral. We run both MBMS and SMBMS with the same tuning parameters. We tuned the parameters manually to find the best combination. Here the neighborhood size k is 36 and the Gaussian bandwidth is $\sigma = 1$.

Figure 2(a) is the data without noise, a clean spiral, while (b) is the noisy spiral. Panel (c) is the denoised spiral by MBMS, which is much cleaner than panel 2. However, the denoised spiral by SMBMS, shown in (d), is even closer to the true one.

Example 4: USPS Digit Dataset This is one of the standard datasets for handwritten digit recognition. Each point in the data set is an image containing a handwritten digit

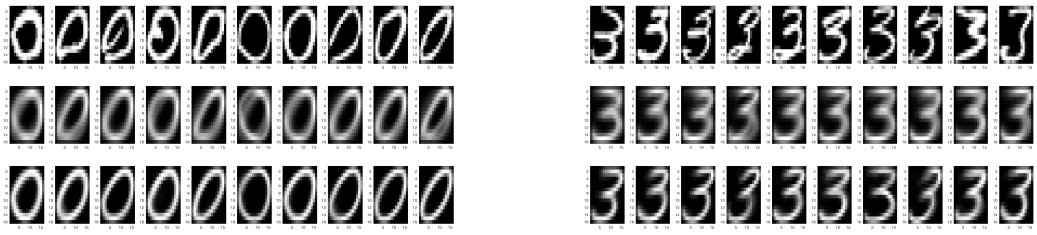

Figure 3: USPS Hand written digits images. The first row: original data; the second row: denoised data by MBMS; the third row: denoised data by SMBMS.

in $16 * 16 = 256$ pixels. The data are noisy and contain many images that are hard to read. We run both MBMS and SMBMS on this dataset and present the result in Figure 3. The first row in Figure 3 contains original noisy images. The second and third rows are the denoised images obtained by MBMS and SMBMS, respectively. Obviously SMBMS outperforms MBMS, as the noise level is much lower in the third row.

Another advantage of spherical MBMS is that the algorithm is more stable than MBMS when the Gaussian bandwidth varies, according to our observation on many numerical experiments. The reason might be the fact that spherelets can approximate the manifold locally with smaller error than PCA. Thus the spherical projection is more stable than linear projection when the data are not clean.

4.3 Data Visualization

Data visualization is an increasingly popular area due to the need for tools for understanding structure in multivariate data. When the dimension D is greater than 3 it's hard to visualize the data. The algorithm having the best performance in our experiments is tSNE (Maaten and Hinton, 2008). tSNE is focused on minimizing the Kullback-Leibler divergence between a discrete density determined by the original data through a Gaussian kernel and another density determined by the projected data through a t-distributed kernel. The first step is to obtain a pairwise distance matrix between data points relying on Euclidean distance. We assume that the manifold can be approximated locally by a sphere, so the spherical distance is better than the Euclidean distance as an estimator of the geodesic distance. In general, for a sphere centered at c with radius r, the distance between two points $x, y \in S(c, r)$ has an analytic form:

$$d_S(x, y) = r \arccos\left(\frac{1}{r^2}(x - c)^\top (y - c)\right).$$

As a result, we can replace the Euclidean distance in the pairwise distance matrix by this spherical distance to obtain a better estimation of the geodesic distance locally. One can then

use this modified pairwise distance matrix in any algorithm. As an example, we compare the original tSNE with spherical tSNE (S-tSNE).

Algorithm 3: S-tSNE

input : Data $X \in \mathbb{R}^{n\times D}$; intrinsic dimension d; dimension of the projected space $m \leq 3$; number of neighbors k; Gaussian bandwidth σ

output: Projected data $Y \in \mathbb{R}^{n\times m}$

for $i = 1 : n$ **do**
  Find $X_{[i]}$, *set of* k*-nearest neighbors of* X_i;
  Do SPCA for $X_{[i]}$ *to find* (V_i, c_i, r_i) *by 4*;
  for $j \in [i]$ **do**
    $D_{ij} = r_i \arccos\left(\frac{1}{r_i^2}(X_i - c_i)^\top (X_j - c_i)\right)$;
  end
end

$p_{j|j} = \dfrac{\exp\{-D_{ij}/\sigma^2\}}{\sum_{l\neq i}\exp\{-D_{il}/\sigma^2\}}$;

$p_{ij} = \dfrac{p_{j|i} + p_{i|j}}{2}$;

$\min_Y KL(P||Q) = \sum_{i\neq j} p_{ij}\log\dfrac{p_{ij}}{q_{ij}}$, *where* $q_{ij} = \dfrac{(1+\|Y_i - Y_j\|)^{-1}}{\sum_{l\neq i}(1+\|Y_i - Y_l\|)^{-1}}$, $Y_i \in \mathbb{R}^m$;

Example 5: Iris dataset This is a well known data set in the UCI Machine Learning Repository. This dataset has $D = 4$ attributes, and 150 labeled samples classified into three groups. Figure 4 shows the 2 dimensional projection given by tSNE and S-tSNE with two different parameters $\sigma = 60, 80$. The first advantage of S-tSNE is that there are fewer overlapping points across classes than tSNE, for both $\sigma = 60$ and $\sigma = 80$. The second advantage is that S-tSNE is more stable as the tuning parameter varies. This relative stability can be seen not only in data visualization but also in manifold denoising, as discussed in the previous subsection.

5 Discussion

There are several natural next directions building on the spherelets approach proposed in this article. The current version of spherelets is not constrained to be discontinuous, so that the estimate $\widehat{M}$ of the manifold M will in general have gaps and hence $\widehat{M}$ will not correspond to a single manifold. We view this as an advantage in many applications, because it avoids restricting consideration to subspaces that are only one manifold and instead

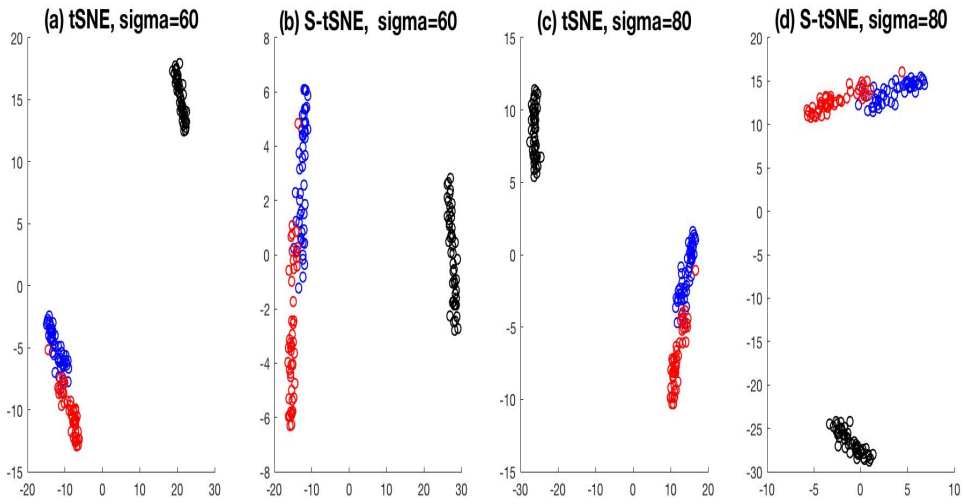

Figure 4: Iris dataset. (a) tSNE, $\sigma = 60$; (b) S-tSNE, $\sigma = 60$; (c) tSNE, $\sigma = 80$; (d) S-tSNE, $\sigma = 80$.

accommodates true discontinuities. Nevertheless, in certain applications it is very useful to obtain a continuous estimate; for example, when we have prior knowledge that the true manifold is continuous and want to use this knowledge to improve statistical efficiency and produce a more visually appealing and realistic estimate. A typical case is in imaging when D is 2 or 3 and d is 1 or 2 and we are trying to estimate a known object from noisy data. In addition, by restricting the subspace to be a single manifold and producing a continuous estimation, we have the possibility of using $\widehat{M}$ to estimate geodesic distances between data points; a problem of substantial interest in the literature (Bernstein et al. (2000), Yang (2004), Meng et al. (2007), Meng et al. (2008)). Possibilities in terms of incorporating continuity constraints include (a) producing an initial $\widehat{M}$ using spherelets and then closing the gaps through linear interpolation; and (b) incorporating a continuity constraint directly into the objective function, to obtain essentially a type of higher dimensional analogue of splines.

Another important direction is to extend the approach to accommodate observed data that do not consist of D-dimensional Euclidean vectors but have a more complex form. For example, the original data may themselves be functions or surfaces or networks or may have a discrete form; e.g, consisting of high-dimensional binary vectors. There are two natural possibilities to account for more elaborate data. The first is to modify the spherelets algorithm to take as input a pairwise distance matrix in place of the original data, as already addressed in the data visualization case. The second is to develop a model-based version of spherelets in which we define a likelihood function for the data, which can be fitted from a frequentist or Bayesian perspective. We are currently considering the latter direction,

through a mixture of spherelets model that includes Fisher von-Mises kernels on different spheres with Gaussian noise. Once we have such a generative probability model for the data, it becomes natural to include modifications to account for more elaborate data structures; e.g. exploiting the flexibility of the Bayesian hierarchical modeling framework. Such an approach also has the advantage of include uncertainty quantification in manifold learning and associated tasks.

An additional direction is improving the flexibility of the basis by further broadening the dictionary beyond simply pieces of spheres. Although one of the main advantages of spherelets is that we maintain much of the simplicity and computational tractability of locally linear bases, it is nonetheless intriguing to include additional flexibility in an attempt to obtain more concise representations of the data with fewer pieces. Possibilities we are starting to consider include the use of quadratic forms to obtain a higher order local approximation to the manifold and extending spheres to ellipses. In considering such extensions, there are major statistical and computational hurdles; the statistical challenge is maintaining parsimony, while the computational one is to obtain a simple and scalable algorithm. As a good compromise to clear both of these hurdles, one possibility is to start with spherelets and then perturb initial sphere estimates (e.g., to produce an ellipse) to better fit the data.

Finally, there is substantial interest in scaling up to very large D cases; the current algorithm will face problems in this regard similar to issues faced in applying usual PCA to high-dimensional data. To scale up spherelets, one can potentially leverage on scalable extensions of PCA, such as sparse PCA (Zou et al. (2006), Johnstone and Lu (2009)). The availability of a very simple closed form solution to spherical PCA makes such extensions conceptually straightforward, but it remains to implement such approaches in practice and carefully consider appropriate asymptotic theory. In terms of theory, it is interesting to consider optimal rates of simultaneously estimating M and the density of the data on (or close) to M, including in cases in which D is large and potentially increasing with sample size.

Acknowledgments

The authors acknowledge support for this research from the Office of Naval Research grant N000141712844 and the National Institutes of Health grants R01 CA 193655.

References

Allard, W. K., Chen, G., and Maggioni, M. (2012). Multi-scale geometric methods for data sets ii: Geometric multi-resolution analysis. *Applied and Computational Harmonic Analysis*, 32(3):435–462.

Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for em-

bedding and clustering. In *Advances in Neural Information Processing Systems*, pages 585–591.

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

Beygelzimer, A., Kakade, S., and Langford, J. (2006). Cover trees for nearest neighbor. In *Proceedings of the 23rd International Conference on Machine learning*, pages 97–104. ACM.

Carreira-Perpiñán, M. Á. (2006). Fast nonparametric clustering with gaussian blurring mean-shift. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 153–160. ACM.

Carreira-Perpinán, M. A. (2008). Generalised blurring mean-shift algorithms for nonparametric clustering. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Chen, G. and Maggioni, M. (2011). Multiscale geometric and spectral analysis of plane arrangements. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2825–2832. IEEE.

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

Demartines, P. and Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154.

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

Hein, M. and Maier, M. (2007a). Manifold denoising. In *Advances in Neural Information Processing Systems*, pages 561–568.

Hein, M. and Maier, M. (2007b). Manifold denoising as preprocessing for finding natural representations of data. In *AAAI*, pages 1646–1649.

Hu, T.-C., Rosalsky, A., and Volodin, A. (2008). On convergence properties of sums of dependent random variables under second moment and covariance restrictions. *Statistics and Probability Letters*, 78(14):1999–2005.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, pages 295–327.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693.

Kambhatla, N. and Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural Computation*, 9(7):1493–1516.

Karypis, G. and Kumar, V. (1998). A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20(1):359–392.

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6(Nov):1783–1816.

Lee, J. A. and Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Lee, S., Zou, F., and Wright, F. A. (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Annals of Statistics*, 38(6):3605.

Liao, W. and Maggioni, M. (2016). Adaptive geometric multiscale approximations for intrinsically low-dimensional data. *arXiv preprint arXiv:1611.01179*.

Liao, W., Maggioni, M., and Vigogna, S. (2016). Learning adaptive multiscale approximations to data and functions near low-dimensional sets. In *Information Theory Workshop (ITW), 2016 IEEE*, pages 226–230. IEEE.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Maggioni, M., Minsker, S., and Strawn, N. (2016). Multiscale dictionary learning: non-asymptotic bounds and robustness. *Journal of Machine Learning Research*, 17.

Meng, D., Leung, Y., Xu, Z., Fung, T., and Zhang, Q. (2008). Improving geodesic distance estimation based on locally linear assumption. *Pattern Recognition Letters*, 29(7):862–870.

Meng, D., Xu, Z., Gu, N., and Dai, M. (2007). Estimating geodesic distances on locally linear patches. In *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, pages 851–854. IEEE.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

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

Scholz, M., Kaplan, F., Guy, C. L., Kopka, J., and Selbig, J. (2005). Non-linear pca: a missing data approach. *Bioinformatics*, 21(20):3887–3895.

Szlam, A. (2009). Asymptotic regularity of subdivisions of euclidean domains by iterated pca and iterated 2-means. *Applied and Computational Harmonic Analysis*, 27(3):342–350.

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

Walder, C. and Schölkopf, B. (2009). Diffeomorphic dimensionality reduction. In *Advances in Neural Information Processing Systems*, pages 1713–1720.

Wang, W. and Carreira-Perpinán, M. A. (2010). Manifold blurring mean shift algorithms for manifold denoising. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1759–1766. IEEE.

Wang, W., Carreira-Perpinán, M. A., and Lu, Z. (2011). A denoising view of matrix completion. In *Advances in Neural Information Processing Systems*, pages 334–342.

Xianxi, L., Li, S., Guoquan, L., Menghua, X., and Wei, W. (2017). Nonlinear system monitoring with piecewise performed principal component analysis. In *Control Conference (CCC), 2017 36th Chinese*, pages 9665–9669. IEEE.

Yang, L. (2004). K-edge connected neighborhood graph for geodesic distance estimation and nonlinear data projection. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 196–199. IEEE.

Zemel, R. S. and Carreira-Perpiñán, M. Á. (2005). Proximity graphs for clustering and manifold learning. In *Advances in Neural Information Processing Systems*, pages 225–232.

Zhang, Z. and Wang, J. (2007). Mlle: Modified locally linear embedding using multiple weights. In *Advances in Neural Information Processing Systems*, pages 1593–1600.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286.

6 Appendix

6.1 Covering Number

6.1.1 Curves (d=1)

Throughout this section, $\gamma : [0, L] \to \mathbb{R}^2$ is a C^3 curve with input s, the arc length parameter. Let $s_0 \in [0, L]$ be fixed. Let $\kappa(s)$ be the curvature at point $\gamma(s)$.

Proposition 4. *Let $L(s)$ be the tangent line of γ at point $\gamma(s_0)$, then L is the unique line such that $d(\gamma(s), L) \leq \frac{K}{2}|s - s_0|^2$, so*

$$\lim_{s \to s_0} \frac{\|L(s) - \gamma(s)\|}{|s - s_0|} = 0,$$

where $K = \sup_s |\kappa(s)|$.

Proof. This is a standard result so we will not prove it in this paper. □

Corollary 4. *Let $\epsilon > 0$, then there exists $N \leq L(\frac{2\epsilon}{K})^{-\frac{1}{2}}$ tangent lines $L_1, \cdots, L_N$ at points $\gamma(s_1), \cdots, \gamma(s_n)$ such that $\forall p \in \gamma$, $\exists i$ s.t. $d(p, L_i) \leq \epsilon$.*

Proof. Let $s_1 = 0$, then by Proposition 4, $\|\gamma(s) - L_1(s)\| < \frac{K}{2}|s|^2$, so if $s < (\frac{2\epsilon}{K})^{1/2}$, $\|\gamma(s) - L_1(s)\| < \epsilon$. Then start from $(\frac{2\epsilon}{K})^{1/2}$, repeat this process to find N tangent lines, so $N \leq L/(\frac{2\epsilon}{K})^{\frac{1}{2}} = L(\frac{2\epsilon}{K})^{-\frac{1}{2}}$. □

Proposition 5. *Let $C(s)$ be the osculating circle of γ at point $\gamma(s_0)$. Then if $\kappa(s_0) \neq 0$, C is the unique circle such that $d(\gamma(s), C) \leq \frac{T}{6}|s - s_0|^3$, so*

$$\lim_{s \to s_0} \frac{\|C(s) - \gamma(s)\|}{|s - s_0|^2} = 0,$$

where $T = \sup_s |\gamma^{(3)}(s)|$.

Note 1. *Proposition 5 holds only when the osculating circle is non degenerate. If the curvature of γ at $\gamma(s_0)$ is $\kappa(s_0) = 0$, the osculating circle C degenerates to tangent line L. In this case, Proposition 4 applies.*

Proof. The osculating circle C has radius $r = \frac{1}{|\kappa(s_0)|}$ and center $\gamma(s_0) + \frac{1}{\kappa(s_0)}\mathbf{n}$, where $\mathbf{n} = \frac{\gamma''(s_0)}{\|\gamma''(s_0)\|}$ is the unit normal vector. Let $\{-\mathbf{n}, t\}$ be the Frenet frame, where $t = \gamma'(s_0)$. Without loss of generality, assume $s_0 = 0$ and $\gamma(s_0) = 0$. Under the Frenet frame, we can rewrite the osculating circle as

$$C(s) = \begin{bmatrix} -\frac{1}{\kappa} \\ 0 \end{bmatrix} + \frac{1}{\kappa} \begin{bmatrix} \cos(\kappa s) \\ \sin(\kappa s) \end{bmatrix} = \begin{bmatrix} r(-1 + \cos(\kappa s)) \\ r \sin(\kappa s) \end{bmatrix}.$$

The Taylor expansion for γ can be written as

$$\begin{aligned} \gamma(s) &= \gamma(0) + \gamma'(0)s + \frac{1}{2}\gamma''(0)s^2 + R_2(s) \\ &= 0 + s \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \frac{s^2}{2} \begin{bmatrix} -\kappa \\ 0 \end{bmatrix} + R_2(s) \\ &= \begin{bmatrix} -\frac{\kappa s^2}{2} \\ s \end{bmatrix} + R_2(s), \end{aligned}$$

where $|R_2(s)| \leq \frac{T}{6}|s|^3$, so $\lim_{s\to s_0} \frac{R_2(s)}{s^2} = 0$. As a result,

$$\begin{aligned} C(s) - \gamma(s) &= \begin{bmatrix} \frac{1}{\kappa}(-1+\cos(\kappa s)) + \frac{\kappa s^2}{2} \\ \frac{1}{\kappa}\sin(\kappa s) - s \end{bmatrix} - R_2(s) \\ &= \begin{bmatrix} \frac{1}{\kappa}(-1+1-\frac{\kappa^2 s^2}{2} + o(s^3)) + \frac{\kappa s^2}{2} \\ \frac{1}{\kappa}(\kappa s - o(s^3)) - s \end{bmatrix} - R_2(s) \\ &= \begin{bmatrix} o(s^3) \\ o(s^3) \end{bmatrix} - R_2(s) \end{aligned}$$

As a result,

$$\|C(s) - \gamma(s)\| \leq |R_2(s)| \leq \frac{T}{6}|s|^3.$$

Now we prove the uniqueness. Observe the first entry of $C(s) - \gamma(s)$:

$$\begin{aligned} &\frac{1}{\kappa'}\left(-1 + 1 - \frac{\kappa^2 s^2}{2} + o(s^3)\right) + \frac{\kappa s^2}{2} = o(s^3) \\ &\iff \frac{k^2 s^2}{\kappa'} = k s^2 \\ &\iff \kappa' = \kappa, \end{aligned}$$

which means $C(s)$ is the osculating circle. □

Corollary 5. *Let $\epsilon > 0$, then there exists $N \leq L(\frac{6\epsilon}{T})^{-\frac{1}{3}}$ osculating circles $C_1, \cdots, C_N$ at points $\gamma(s_1), \cdots, \gamma(s_n)$ such that $\forall p \in \gamma$, $\exists i$ s.t. $d(p, C_i) \leq \epsilon$.*

Proof. Let $s_1 = 0$, then by Proposition 5, $\|\gamma(s) - C_1(s)\| < \frac{T}{6}|s|^3$, so if $s < (\frac{6\epsilon}{T})^{1/3}$, $\|\gamma(s) - C_1(s)\| < \epsilon$. Then start from $(\frac{6\epsilon}{T})^{1/3}$, repeat this process to find N osculating circles, so $N \leq L/(\frac{6\epsilon}{T})^{\frac{1}{3}} = L(\frac{6\epsilon}{T})^{-\frac{1}{3}}$. □

6.1.2 Surfaces (d=2)

Throughout this section, $M : U \to \mathbb{R}^3$ is a regular C^3 surface parametrized by $x = x(u,v), y = y(u,v), z = z(u,v)$ where U is a compact subset of $\mathbb{R}^2$. Let $K : UTM \to \mathbb{R}$ be the directional curvature, that is, $K(p,v) = \kappa(\exp_p(tv))|_{t=0}$ is the curvature in direction v, UTM is the unit sphere bundle of M. Without loss of generality, assume $X_0 = (x(0,0), y(0,0), z(0,0)) = (0,0,0) \in M$ is fixed.

Proposition 6. *Letting $H(u,v)$ be the tangent plane of S at point X_0, H is the unique plane such that $\|H(u,v) - M(u,v)\|^2 \leq \frac{K}{2}\|(u,v)\|^2$, so*

$$\lim_{(u,v)\to(0,0)} \frac{\|H(u,v) - M(u,v)\|}{\|(u,v)\|} = 0,$$

where $K = \sup_{(p,v)\in UTM} |K(p,v)|$.

This is a higher dimensional analogue of Proposition 4, but a similar analogue of Proposition 5 does not exist, which is a direct result from the following lemma.

Lemma 2. *There exists a sphere* $S(u,v)$ *such that*

$$\lim_{(u,v)\to(0,0)} \frac{\|S(u,v) - M(u,v)\|}{\|(u,v)\|^2} = 0$$

if and only if X_0 *is an umbilical point.*

Proof. X_0 is an umbilical point $\Leftrightarrow \frac{L}{E} = \frac{M}{F} = \frac{N}{G} = \alpha$, where I$= \begin{bmatrix} E & F \\ F & G \end{bmatrix}$ is the first fundamental form and II$= \begin{bmatrix} L & M \\ M & N \end{bmatrix}$ is the second fundamental form. Without loss of generality, assume (u,v) are locally orthonormal parameters, which means $\langle M_u, M_v\rangle = 0$ and $\|M_u\| = \|M_v\| = 1$. In this case, $E = G = 1, F = 0$, so $L = N = \alpha, M = 0$. As a result, the quadratic terms in the Taylor expansion

$$dX = X_0 + M_u du + M_v dv + \alpha du^2 + \alpha dv^2 + o\left(\sqrt{u^2+v^2}^2\right)$$

is nothing but a sphere.

□

Although we can't find a sphere so that the error is third order in general, we can still find a sphere with this property in some direction, as shown in the following proposition.

Proposition 7. *Let* $k_1 \geq k_2$ *be the principal curvatures of* M *at* X_0*,* e_1, e_2 *be the corresponding principal directions and* $\mathbf{n}$ *be the normal vector at* X_0*. Then for any* $k \in [k_2, k_1]$*, let* S_k *be the sphere centered at* $c = X_0 - \frac{1}{k}\mathbf{n}$ *with radius* $\frac{1}{|k|}$*, there exists a curve* γ *on* M *and a constant* T *such that*

$$d(\gamma(s), S_k) \leq \frac{T}{6}|s|^3.$$

Proof. Since $k \in [k_2, k_1]$, there exists a direction, represented by unit vector ξ at $(0,0) \in U$ so that k is the normal curvature in direction ξ. To be more specific, let γ_ξ be the curve on M in direction ξ; that is, $\gamma_\xi(s) = M(s\xi), s \in [L_1, L_2]$, where $L_1 = \inf\{s | s\xi \in U\}$, $L_2 = \inf\{s|s\xi \in U\}$. By the above construction, the curvature of γ_ξ at X_0 is just k, so from Proposition 5, we know that

$$d(\gamma_\xi(s), C_\xi) \leq \frac{T}{6}|s|^3,$$

where $T = \sup_s |\gamma_\xi^{(3)}(s)|$ and C_ξ is a circle centered at $c = X_0 - \frac{1}{k}\mathbf{n}$ with radius $\frac{1}{k}$. Since C_ξ is just a great circle of S_k, we have the desired inequality:

$$d(\gamma_\xi(s), S_k) \leq d(\gamma_\xi(s), C_\xi) \leq \frac{T}{6}|s|^3.$$

□

6.1.3 General Cases

Throughout this section, M is a d-dimensional C^3 compact manifold embedded in $\mathbb{R}^D$. Let $p \in M$ be a fixed point and we can assume $p = 0$ without loss of generality. Then we have the following proposition that is similar to Proposition 4 and Proposition 6.

Proposition 8. *Let T_pM be the tangent space of M at p, then $d(x, T_pM) \leq K\|x\|^2$.*

Proof of Theorem 1. Firstly, we prove inequality 1 in main theorem. First we focus on one local neighborhood of $p \in M$. Proposition 8 shows that there exists a hyperplane H such that when $\|x\| < \left(\frac{2\epsilon}{K}\right)^{\frac{1}{2}}$, $d(x, H) \leq \epsilon$. Locally, $d|_U \approx d|_{\log U}$ where U is a local neighborhood on M and log is the inverse of the exponential mapping on U. So the inequality $d(q, H) \leq \epsilon$ holds when $q \in B\left(p, \left(\frac{2\epsilon}{K}\right)^{\frac{1}{2}}\right)$. When ϵ is small, the volume of the geodesic ball is

$$\mathrm{Vol}\left\{B\left(p, \left(\frac{2\epsilon}{K}\right)^{\frac{1}{2}}\right)\right\} \approx \left(\frac{2\epsilon}{K}\right)^{\frac{d}{2}}.$$

Let $V = \mathrm{Vol}(M)$ and we cover the manifold by geodesic balls with radius $(\frac{2\epsilon}{K})^{\frac{1}{2}}$, the number of balls $N \lesssim \frac{V}{(\frac{2\epsilon}{K})^{\frac{d}{2}}} = V(\frac{2\epsilon}{K})^{-\frac{d}{2}}$. Then we prove inequality 2 in main theorem by considering two cases.

1. M_ϵ^{c}. Firstly we consider the worst part: the complement of M_ϵ. We cover this subset of M by smaller geodesic balls $B(p, (\frac{2\epsilon}{K})^{\frac{1}{2}})$ with square distance less than or equal to ϵ. The first part of the proof shows that this covering exists, and the number of balls is less than or equal to $\frac{(V - V_\epsilon)}{(\frac{2\epsilon}{K})^{\frac{d}{2}}} = (V - V_\epsilon)(\frac{2\epsilon}{K})^{-\frac{d}{2}}$.

2. M_ϵ. We cover this part by bigger geodesic balls. For any point $p \in F_\epsilon$, we have $|k_1(p) - k_d(p)| \leq (\frac{2\epsilon}{K})^{\frac{1}{2}}$. Let k^* be the curvature of the sphere obtained by SPCA on $U \coloneqq B(p, (\frac{6\epsilon}{3+T})^{\frac{1}{3}})$. Then for any $q \in U$, if $q \in B(p, (\frac{2\epsilon}{K})^{\frac{1}{2}})$, when case 1 shows that the error is less than or equal to ϵ, so we only need to consider $q \in U - B(p, (\frac{2\epsilon}{K})^{\frac{1}{2}})$, that is, $(\frac{2\epsilon}{K})^{\frac{1}{2}} \leq d(p, q) \leq (\frac{6\epsilon}{3+T})^{\frac{1}{3}}$. Let $\gamma_q(s) = \exp_p(s \log q)$ be the geodesic connecting

p and q, assume γ_q at p is k_q. Since both k_q and k^* are in $[k_d(p), k_1(p)]$, we have the following relation:

$$|k_q - k^*| \leq |k_1(p) - k_d(p)| \leq \left(\frac{2\epsilon}{K}\right)^{\frac{1}{2}}.$$

Recall in the proof of Proposition 5, only the first three terms in the Taylor expansion of γ_q matter; that is, $\gamma_q(0)$, $\gamma_q^{'}(s)$ and $\gamma_q^{''}(s)$, so we only need to consider the first two coordinates of $\gamma_q(s)$ and $C(s)$ while other coordinates are all $o(s^3)$. Similar to the proof of Proposition 5, the first two coordinates of $C(s) - \gamma_q(s)$ are

$$\begin{aligned}
&\begin{bmatrix} \frac{1}{k_q}(-1 + \cos(k_q s)) + \frac{k^* s^2}{2} \\ \frac{1}{k_q}\sin(k_q s) - s \end{bmatrix} - R_2(s) \\
&= \begin{bmatrix} \frac{1}{k_q}(-1 + (1 - \frac{k_q^2 s^2}{2} + o(s^3)) + \frac{k^* s^2}{2} \\ \frac{1}{k_q}(k_q s - o(s^3)) - s \end{bmatrix} - R_2(s) \\
&= \begin{bmatrix} \frac{s^2}{2}(k_q - k^*) + o(s^3) \\ o(s^3) \end{bmatrix} - R_2(s),
\end{aligned}$$

where $|R_2(s)| \leq \frac{T}{6}|s|^3$. We claim that $\|C(s_q) - q\| \leq \epsilon$ where $s_q = d(p, q)$. Since

$$|k_q - k^*| \leq \left(\frac{2\epsilon}{K}\right)^{\frac{1}{2}} \leq |s| \leq \left(\frac{6\epsilon}{3+T}\right)^{\frac{1}{3}},$$

$$\begin{aligned}
\|C(s_q) - q\| &\leq \frac{s^2}{2}|k_q - k^*| + \frac{T}{6}|s|^3 \\
&\leq \left(\frac{1}{2} + \frac{T}{6}\right)|s|^3 \\
&\leq \frac{3+T}{6}\frac{6\epsilon}{3+T} = \epsilon.
\end{aligned}$$

As a result, we can cover M_ϵ by geodesic balls with radius $(\frac{6\epsilon}{3+T})^{\frac{1}{3}}$ so that the error is less than or equal to ϵ. So the number of balls needed to cover M_ϵ is less than or equal to $V_\epsilon(\frac{6\epsilon}{3+T})^{-\frac{d}{3}}$.

Based on the above two cases, the total number of balls $N_S(\epsilon, M) \leq V_\epsilon(\frac{6\epsilon}{3+T})^{-\frac{d}{3}} + (V - V_\epsilon)(\frac{2\epsilon}{K})^{-\frac{d}{2}}$. □

Note 2. *1. Corollary 4 is a special case of Theorem 1.*
2. When d increases, the performance will be worse and worse, which is another representation of the curse of dimensionality. But fortunately, d is the intrinsic dimension of M, which is assumed to be small in most cases.

Proof of Proposition 3. It suffices to provide two manifolds with covering numbers achieving the upper bounds in Theorem 1.

1. Let $\gamma(t) = (t, t^2),\ t \in (0,1)$ so $\gamma^{''}(t) = (0,2)$ is constant. The covering number $N_{\mathcal{H}}(\epsilon, \gamma)$ follows.

2. Let $\gamma(t) = (t, t^3),\ t \in (0,1)$ so $\gamma^{(3)}(t) = (0,6)$ is constant. The covering number $N_{\mathcal{S}}(\epsilon, \gamma)$ follows.

□

6.2 SPCA

6.2.1 Proof of Lemma 1

Let $\Phi_{V,c}$ be the orthogonal projection to the affine subspace $c + V$; that is, $\Phi_{V,c}(x) = c + VV^\top(x - c)$. Then observe that $x - \Phi_{V,c}(x) \perp \Phi_{V,c} - y,\ \forall y \in S_V(c,r)$, so

$$\|x - y\|^2 = \|x - \Phi_{V,c} + \Phi_{V,c} - y\|^2 = \|x - \Phi_{V,c}(x)\|^2 + \|\Phi_{V,c}(x) - y\|^2.$$

That is, the optimization problem $\underset{y \in S_V(c,r)}{\operatorname{argmin}} \|x - y\|^2$ is equivalent to $\underset{y \in S_V(c,r)}{\operatorname{argmin}} \|\Phi_{V,c}(x) - y\|^2$. Since the second problem only involves the affine subspace $c + V$, we can translate it to the following problem:

$$\min_{y \in S(c,r) \subset \mathbb{R}^{d+1}} \|x - y\|^2,$$

where x is any point in $\mathbb{R}^{d+1}$ and $S(c,r) = \{y \in \mathbb{R}^{d+1} :\ \|y - c\| = r\}$. So we only need to prove

$$\Psi_{V,c}(x) \coloneqq \underset{y \in S(c,r)}{\operatorname{argmin}} \|x - y\|^2 = c + \frac{r}{\|x - c\|}(x - c).$$

On one hand,

$$\begin{aligned}
\|x - \Psi_{V,c}(x)\|^2 &= \left\| x - c - \frac{r}{\|x - c\|}(x - c) \right\|^2 \\
&= \left\| \left(1 - \frac{r}{\|x - c\|}\right)(x - c) \right\|^2 \\
&= \left(1 - \frac{r}{\|x - c\|}\right)^2 \|x - c\|^2 \\
&= (\|x - c\| - r)^2
\end{aligned}$$

On the other hand, for any $y \in S(c,r)$,

$$\begin{aligned}
\|x-y\|^2 &= \|x-c+c-y\|^2 \\
&= \|x-c\|^2 + \|c-y\|^2 - 2(x-c)^\top (y-c) \\
&= \|x-c\|^2 + r^2 - 2(x-c)^\top (y-c) \\
&\geq \|x-c\|^2 + r^2 - 2\|x-c\|\|y-c\| \\
&= \|x-c\|^2 + r^2 - 2r\|x-c\| \\
&= (\|x-c\| - r)^2.
\end{aligned}$$

□

6.3 Proof of Theorem 2

The objective function is

$$g(f) = \sum_{i=1}^n \Big\{ Y_i^\top Y_i + f^\top Y_i - \frac{1}{n}\sum_{j=1}^n (Y_j^\top Y_j + f^\top Y_j) \Big\}^2.$$

Let $l_i = Y_i^\top Y_i$, $\bar{l} = \dfrac{1}{n}\sum_{i=1}^n l_i$ and $\bar{Y} = \dfrac{1}{n}\sum_{i=1}^n Y_i$, then

$$\begin{aligned}
g(f) &= \sum_{i=1}^n \Big\{ Y_i^\top Y_i + f^\top Y_i - \frac{1}{n}\sum_{j=1}^n (Y_j^\top Y_j + f^\top Y_j) \Big\}^2 \\
&= \sum_{i=1}^n \Big(l_i + f^\top Y_i - \bar{l} - f^\top \bar{Y} \Big)^2 \\
&= \sum_{i=1}^n \Big((l_i - \bar{l}) + f^\top (Y_i - \bar{Y}) \Big)^2 \\
&= \sum_{i=1}^n \Big\{ f^\top (Y_i - \bar{Y})(Y_i - \bar{Y})^\top f + 2(l_i - \bar{l}) f^\top (Y_i - \bar{Y}) + (l_i - \bar{l})^2 \Big\} \\
&= f^\top \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})^\top f + 2 f^\top \sum_{i=1}^n (l_i - \bar{l})(Y_i - \bar{Y}) + \sum_{i=1}^n (l_i - \bar{l})^2
\end{aligned}$$

is a quadratic function. So $\widehat{f} = -H^{-1}\xi$ where

$$H = \sum_{i=1}^n \big(Y_i - \bar{Y}\big)\big(Y_i - \bar{Y}\big)^\top, \quad \xi = \sum_{i=1}^n \Big(\|Y_i^\top Y_i\| - \frac{1}{n}\sum_{j=1}^n \|Y_j^\top Y_j\| \Big) \big(Y_i - \bar{Y}\big).$$

As a result,

$$\widehat{c} = -\frac{1}{2}\widehat{f} = \frac{1}{2}H^{-1}\xi.$$

□

Then we prove Corollary 2, that is, when the data are sampled from some sphere $S_{\widehat{V}}(\widehat{c},\widehat{r})$, then SPCA and the sum of squared residuals have the same minimizer, which is exactly $\widehat{c}$:

$$\widehat{c} = \operatorname*{argmin}_c g(c) = \operatorname*{argmin}_c \sum_{i=1}^n \left\| Y_i - \left(c + \frac{\sum_{j=1}^n \|Y_j - c\|}{n\|Y_i - c\|}(Y_i - c)\right)\right\|^2 =: \operatorname*{argmin}_c \tilde{g}(c).$$

Proof of Corollary 2. We split the proof into three parts, where the first two parts are the proofs of the following two equations.

$$g(c) = n\,\mathrm{var}(\|Y_i - c\|^2),\ \tilde{g}(c) = n\,\mathrm{var}(\|Y_i - c\|).$$

1. We prove the first equation.

$$\begin{aligned} n\,\mathrm{var}(\|Y_i - c\|^2) &= \sum_{i=1}^n \left(\|Y_i - c\|^2 - \frac{1}{n}\sum_{j=1}^n \|Y_j - c\|^2\right)^2 \\ &= \sum_{i=1}^n \left(Y_i^\top Y_i - 2c^\top Y_i + c^\top c - \frac{1}{n}\sum_{j=1}^n (Y_j^\top Y_j - 2c^\top Y_j + c^\top c)\right)^2 \\ &= \sum_{i=1}^n \left(Y_i^\top Y_i - 2c^\top Y_i - \frac{1}{n}\sum_{j=1}^n (Y_j^\top Y_j - 2c^\top Y_j)\right)^2 = g(c). \end{aligned}$$

2. We prove the second equation.

$$\begin{aligned} \tilde{g}(c) &= \sum_{i=1}^n \left\| Y_i - \left(c + \frac{\sum_{j=1}^n \|Y_j - c\|}{n\|Y_i - c\|}(Y_i - c)\right)\right\|^2 \\ &= \sum_{i=1}^n \left\{ \left(1 - \frac{\sum_{j=1}^n \|Y_j - c\|}{n\|Y_i - c\|}\right)\|Y_i - c\| \right\}^2 \\ &= \sum_{i=1}^n \left(\|Y_i - c\| - \frac{\sum_{j=1}^n \|Y_j - c\|}{n}\right)^2 = n\,\mathrm{var}(\|Y_i - c\|) \end{aligned}$$

3. Since $X_i \in S_{\widehat{V}}(\widehat{c},\widehat{r})$, $Y_i = \widehat{\mu} + \widehat{V}\widehat{V}^\top(X_i - \widehat{\mu}) = X_i$ so $\|Y_i - \widehat{c}\| = \widehat{r}$ for all $i = 1, \cdots, n$, that is

$$g(\widehat{c}) = n\,\mathrm{var}(\|Y_i - \widehat{c}\|^2) = 0 = n\,\mathrm{var}(\|Y_i - \widehat{c}\|) = \tilde{g}(\widehat{c}).$$

□

6.4 Proof of Theorem 3

Notations For notational convenience we drop the superscript $*$ from the population versions V^*, c^*, r^*, H^*, f^*, Proj* and Λ^*. Consider the following notations:

a. v_i: The eigenvector corresponding to the i^{th} largest eigenvalue of Σ. $\widehat{v}_i$: The eigenvector corresponding to the i^{th} largest eigenvalue of sample variance-covariance matrix $\widehat{\Sigma}$.

b. $V^{D\times(d+1)} = [v_1, v_2, \ldots, v_{d+1}]$, and $\widehat{V}^{D\times(d+1)} = [\widehat{v}_1, \widehat{v}_2, \ldots, \widehat{v}_{d+1}]$.

c. $H = VV^T\Sigma VV^T$, and $\widehat{H} = \widehat{V}\widehat{V}^T\widehat{\Sigma}\widehat{V}\widehat{V}^T$.

d. $\xi = E\left[\left\{\|V^TX\|^2 - E\left(\|V^TX\|^2\right)\right\}VV^T\left\{X - E(X)\right\}\right]$, where $X \sim \rho$, and

$$\widehat{\xi} = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\|\widehat{V}^TX_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}\|\widehat{V}^TX_i\|^2\right)\widehat{V}\widehat{V}^T\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X_i\right)\right\}.$$

e. The center of the Spherelet $c = -H^{-1}\xi/2$. The estimator of c, $\widehat{c} = -\widehat{H}^{-1}\widehat{\xi}/2$.

f. Radius of the Spherelet $r = E\|X - c\|$, and $\widehat{r} = \frac{1}{n}\sum_{i=1}^{n}\|X_i - \widehat{c}\|$.

g. Non-linear embedding of any sample $X \sim \rho$,

$$\text{Proj}(X) = c + r\frac{VV^T(X-c)}{\|VV^T(X-c)\|}, \quad \text{and} \quad \widehat{\text{Proj}}(X) = \widehat{c} + \widehat{r}\frac{\widehat{V}\widehat{V}^T(X-\widehat{c})}{\|\widehat{V}\widehat{V}^T(X-\widehat{c})\|}$$

Lemma 3. *Suppose assumptions (A)-(B) hold. If* $\frac{D}{n} \to 0$ *as* $n \to \infty$, *then*

i. $l_k \xrightarrow{a.s.} \lambda_k$ *for* $k = 1, 2, \ldots, D$, *where* l_k *and* λ_k *are the* k^{th} *largest eigenvalue of* $\widehat{\Sigma}$ *and* Σ, *respectively.*

ii. $|\widehat{v}_k^Tv_k| \xrightarrow{p} 1$ *for* $k = 1, 2, \ldots, m$, *and*
$|\widehat{v}_k^Tv_k| \xrightarrow{p} 0$ *for* $k = m+1, m+2, \ldots, D$.

The proof can be found in Lee et al. (2010).

Proof of Theorem 3. The proof is split into three main parts.

I. *We first show that* $\widehat{c} \xrightarrow{p} c$. The proof is split into three subparts.

i. $\|\widehat{H}^{-1} - H^{-1}\| \xrightarrow{p} 0$, $\|\cdot\|$ *is the spectral norm.* First note that $H = VV^T\Sigma VV^T$. Let the spectral value decomposition of $\Sigma = U\Lambda_DU^T$, and $U = (V\ W)$, $\Lambda_D = \text{diag}(\Lambda, L)$. Then $H = VV^T\Sigma VV^T = V\Lambda V^T$, where Λ is the diagonal matrix of the first $(d+1)$ largest eigenvalues of Σ. Similarly, $\widehat{H} = \widehat{V}\widehat{\Sigma}_{d+1}\widehat{V}^T$, where $\widehat{\Sigma}_{d+1}$

is the vector of first $(d+1)$ largest eigenvalues of $\widehat{\Sigma}$. Further, from the properties of Moore-Penrose inverse, we have $H^{-1} = V\Lambda^{-1}V^T$, and $\widehat{H}^{-1} = \widehat{V}\widehat{\Sigma}_{d+1}^{-1}\widehat{V}^T$.
Let $\|\cdot\|_F$ denote the Frobenius norm, then consider the following:

$$\begin{aligned}\|\widehat{V}\widehat{\Sigma}_{d+1}^{-1}\widehat{V}^T - V\Lambda^{-1}V^T &\leq \|\widehat{V}\widehat{\Sigma}_{d+1}^{-1}\widehat{V}^T - V\Lambda^{-1}V^T\|_F \\ &\leq \|(\widehat{V}-V)\widehat{\Sigma}_{d+1}^{-1}\widehat{V}^T\|_F + \|V(\widehat{\Sigma}_{d+1}^{-1} - \Lambda^{-1})\widehat{V}^T\|_F + \|V\Lambda^{-1}(\widehat{V}-V)^T\|_F. \quad (6)\end{aligned}$$

Consider the first term in (6), $\|(\widehat{V}-V)\widehat{\Sigma}_{d+1}^{-1}\widehat{V}^T\|_F^2 = \|(\widehat{V}-V)\widehat{\Sigma}_{d+1}^{-1}\|_F^2 \leq l_{n,d+1}^{-1}\|(\widehat{V}-V)\|_F$ where $l_{n,d+1}$ is the $(d+1)^{th}$ largest eigenvalue of $\widehat{\Sigma}$, as $\widehat{V}^T\widehat{V} = I_{d+1}$.
Further, $\|(\widehat{V}-V)\|_F^2 = 2(d+1) - \text{tr}(V^T\widehat{V}) - \text{tr}(\widehat{V}^TV) = 2(d+1) - 2(v_1^T\widehat{v}_1 + v_2^T\widehat{v}_2 + \ldots + v_{d+1}^T\widehat{v}_{d+1})$. As $(d+1) < m$, from Lemma 3 we have $\|(\widehat{V}-V)\|_F^2 \xrightarrow{p} 0$ by Stutsky's lemma. Again, as $(d+1) < m$, $\lambda_{d+1} > 1$, by Lemma 3 (i) $l_{n,d+1} \geq 1$ for sufficiently large n. Thus, the first part of (6) converges to zero in probability.
Consider the second term in (6). Note that

$$\|V(\widehat{\Sigma}_{d+1}^{-1} - \Lambda^{-1})\widehat{V}^T\|_F^2 = \|\widehat{\Sigma}_{d+1}^{-1} - \Lambda^{-1}\|_F^2 = (l_{n,1}^{-1} - \lambda_1^{-1})^2 + \ldots + (l_{n,d+1}^{-1} - \lambda_{d+1}^{-1})^2,$$

as $V^TV = \widehat{V}^T\widehat{V} = I_{d+1}$. Using part (ii) of Lemma 3 and the Continuous Mapping Theorem, it can be shown that the second term of (6) converges in probability to zero.
Finally, following similar argument as before, it can be shown that the third part of (6) converges to zero in probability.

We next show that $\|\widehat{\xi} - \xi\| \xrightarrow{p} 0$.

ii. Note that $E(X_i) = 0$. So,

$$\xi = E\left(\|V^TX\|^2VV^TX\right) - E\left(\|V^TX\|^2\right)E\left(VV^TX\right)$$

First we will show the following three convergences in order:

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{V}^TX_i\|^2 \xrightarrow{p} E\left(\|V^TX\|^2\right) \quad (7)$$

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{V}\widehat{V}^TX_i \xrightarrow{p} E\left(VV^TX\right) \quad (8)$$

$$\frac{1}{n}\sum_{i=1}^{n}\|\widehat{V}^TX_i\|^2\widehat{V}\widehat{V}^TX_i \xrightarrow{p} E\left(\|V^TX\|^2VV^TX\right) \quad (9)$$

The convergence in (7) is proved in two steps. The steps are:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|\widehat{V}^TX_i\|^2 - \frac{1}{n}\sum_{i=1}^{n}\|VX_i\|^2\right| \xrightarrow{p} 0 \quad (10)$$

and

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|V^TX_i\|^2 - E\left(\|V^TX\|^2\right)\right| \xrightarrow{p} 0 \tag{11}$$

To see (10), observe that the left hand side (LHS) of (10) is equal to

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\left\{\left(\frac{X_i}{\|X_i\|}\right)^T\left(\widehat{V}\widehat{V}^T - VV^T\right)\frac{X_i}{\|X_i\|}\right\}\right|$$
$$\leq \left|\lambda_{\max}\left(\widehat{V}\widehat{V}^T - VV^T\right)\right|\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2.$$

Note that $\|X_i\|^2 = \Lambda_D\mathbf{z}_i$, $i = 1, 2, \ldots, n$, are independent random variables having mean $\mathrm{tr}(\Lambda_D)$ and finite variance by assumption (A). Therefore, by the Strong Law of Large Numbers (SLLN) (see, e.g., Hu et al. (2008)) the following holds:

$$\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \xrightarrow{a.s.} \mathrm{tr}(\Lambda_D).$$

Further,

$$\lambda_{\max}\left(\widehat{V}\widehat{V}^T - VV^T\right) \leq \lambda_{\max}\left(\widehat{V}\widehat{V}^T - V\widehat{V}^T\right) + \lambda_{\max}\left(V\widehat{V}^T - VV^T\right).$$

Now,

$$\lambda_{\max}\left(\widehat{V}\widehat{V}^T - V\widehat{V}^T\right) \leq \left\|\left(\widehat{V}^T - V\right)\widehat{V}^T\right\|_F = \left\|\widehat{V}^T - V\right\|_F \xrightarrow{p} 0,$$

as shown in part **i**.
To see (11) note that $X_i^TVV^TX_i = \sum_{k=1}^{d+1}\lambda_k z_{i,k}^2$, for $i = 1, 2, \ldots, n$, are independent random variables having finite variance and zero covariance by assumption (A). Therefore, by SLLN the following holds:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|V^TX_i\|^2 - E\left(\|V^TX\|^2\right)\right| \xrightarrow{a.s.} 0$$

Thus (7) is proved.

Next consider the convergence in (8). As before we split the proof in two parts:

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{V}\widehat{V}^T - VV^T\right)X_i\right\|^2 \xrightarrow{p} 0 \tag{12}$$

$$\left\|\frac{1}{n}\sum_{i=1}^{n}VV^TX_i - E\left(VV^TX\right)\right\|^2 \xrightarrow{p} 0 \tag{13}$$

To see (12) observe that the LHS of (12) is less than or equal to

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\left(\widehat{V}\widehat{V}^T - VV^T\right)X_i\right\|^2 \leq \left\|\widehat{V}\widehat{V}^T - VV^T\right\|_F^2\left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2\right)$$

As before, we write

$$\left\|\widehat{V}\widehat{V}^T - VV^T\right\|_F^2 \leq \left\|\widehat{V}\widehat{V}^T - \widehat{V}V^T\right\|_F^2 + \left\|\widehat{V}V^T - VV^T\right\|_F^2.$$

As before it's easy to show that the two terms in right hand side (RHS) of the last equation converge to zero in probability. Further, we have seen that $\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2$ converges almost surely to $\mathrm{tr}(\Lambda_D)$, and therefore is bounded in probability. Combining these facts we conclude that (12) holds.

Next consider (13). Let $u_i = VV^TX_i$. To show $\|\frac{1}{n}\sum_{i=1}^{n}(u_i - E(u_i))\| \xrightarrow{p} 0$, note that $E(u_i) = 0$ and $\mathrm{var}(u_i) = V\Lambda V^T$. Therefore, each component of u_i has expectation zero and $\mathrm{var}(u_{i,k}) \leq \mathrm{tr}(\Lambda)$, for $k = 1, \ldots, D$. Further, $u_{i,k}$ and $u_{j,k}$ are independent for all $i \neq j$ and $k = 1, \ldots, D$. Thus, $\frac{1}{n}\sum_{i=1}^{n}u_{i,k} - E(u_{i,k}) \xrightarrow{a.s.} 0$ for all $k = 1, \ldots, D$, which implies (13).

Finally, we show (9). Again, we split the proof in two main steps:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\|\widehat{V}^TX_i\|^2\widehat{V}\widehat{V}^TX_i - \|V^TX_i\|^2VV^TX_i\right) \xrightarrow{p} 0 \tag{14}$$

$$\frac{1}{n}\sum_{i=1}^{n}\|V^TX_i\|^2VV^TX_i - E\left(\|V^TX\|^2VV^TX\right) \xrightarrow{p} 0 \tag{15}$$

To show (14), we split the problem into two parts

$$\frac{1}{n}\sum_{i=1}^{n}\left(\|\widehat{V}^TX_i\|^2\widehat{V}\widehat{V}^TX_i - \|\widehat{V}^TX_i\|^2VV^TX_i\right) \xrightarrow{p} 0 \tag{16}$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\|\widehat{V}^TX_i\|^2VV^TX_i - \|V^TX_i\|^2VV^TX_i\right) \xrightarrow{p} 0 \tag{17}$$

The proof of (16) is similar to that of (12), except here we have to show that $\frac{1}{n}\sum_{i=1}^{n}\|\widehat{V}^TX_i\|^2\|X_i\|$ is bounded in probability. To show the boundedness, it is enough to show that $\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^3$ converges (as $\widehat{V}\widehat{V}^T \leq I_D$). Again, the random variables $\|X_i\|^3 = \left(z_i^T\Lambda_D z_i\right)^{3/2}$, $i = 1, 2, \ldots, n$, are independent, have finite second order moments by assumption (A). Therefore $\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^3 \xrightarrow{a.s.} E\left\{\left(z^T\Lambda_D z\right)^{3/2}\right\}$, by SLLN. Finally, (17) follows by straightforward application of SLLN on each component of $\|V^TX_i\|^2VV^TX_i$.

iii. *Remaining steps towards showing* $\widehat{c} \xrightarrow{p} c$. Observe that,

$$\widehat{c} - c = \frac{1}{2}\left\{\left(\widehat{H}^{-1} - H^{-1}\right)\widehat{\xi} + H^{-1}\left(\widehat{\xi} - \xi\right)\right\}.$$

Therefore,

$$2\|\widehat{c} - c\| \leq \|\widehat{H}^{-1} - H^{-1}\|\|\widehat{\xi}\| + \|H^{-1}\|\|\widehat{\xi} - \xi\|$$

Thus, to show that $\widehat{c} \xrightarrow{p} c$, it is enough to show that $\|\widehat{\xi}\|$ is bounded in probability, and $\|H^{-1}\|$ is bounded. From (ii) we observe that each component of $\widehat{\xi}$ converges in probability to ξ. Consider a number $N = N_\epsilon$ such that $P(\|\widehat{\xi} - \xi\| > \frac{N}{2}) \leq \epsilon$, and $\|\xi\| < \frac{N}{2}$, then

$$\begin{aligned} P\left(\|\widehat{\xi}\| > N\right) &\leq P\left(\|\widehat{\xi} - \xi\| + \|\xi\| > N\right) \\ &\leq P\left(\|\widehat{\xi} - \xi\| > \frac{N}{2}\right) + P\left(\|\xi\| > \frac{N}{2}\right) \leq \epsilon. \end{aligned}$$

Lastly, we have already seen that $H^{-1} = V\Lambda^{-1}V^T$. Thus,

$$\|H^{-1}\| \leq \|H^{-1}\|_F = \sqrt{\mathrm{tr}(\Lambda^{-2})} \leq \lambda_{d+1}^{-1}\sqrt{d+1} \leq \sqrt{d+1},$$

as $d + 1 \leq m$ (see assumption (B)). Hence, the proof follows.

II. The next step is to show $\widehat{r} \xrightarrow{p} r$. Recall that $r = E\|X - c\|$, and $\widehat{r} = \frac{1}{n}\sum_{i=1}^n \|X_i - \widehat{c}\|$. We will prove this in the following two steps:

$$\left|\frac{1}{n}\sum_{i=1}^n \left(\|X_i - \widehat{c}\| - \|X_i - c\|\right)\right| \xrightarrow{p} 0 \tag{18}$$

$$\left|\frac{1}{n}\sum_{i=1}^n \|X_i - c\| - E\|X - c\|\right| \xrightarrow{p} 0. \tag{19}$$

We first prove (18). Observe that,

$$\left|\frac{1}{n}\sum_{i=1}^n \|X_i - \widehat{c}\| - \|X_i - c\|\right| = \left|\frac{1}{n}\sum_{i=1}^n \frac{(c - \widehat{c})^T\tilde{c}_i}{\|\tilde{c}_i\|}\right| \leq \|c - \widehat{c}\|$$

by multivariate mean value theorem, where $\tilde{c}_i = X_i - \widehat{c} + s_i(\widehat{c} - c)$, where $s_i \in (0, 1)$, $i = 1, 2, \ldots, n$. Therefore, (18) is proved.

Equation (19) can be shown by SLLN applied on $\|X_i - c\|$, $i = 1, 2, \ldots, n$, as they are independent and identically distributed with finite second order moments.

III. *Showing* $\|\text{Proj}(X_i) - \widehat{\text{Proj}}(X_i)\| \xrightarrow{p} 0$, *for any* $i = 1, 2, \ldots, n$. For any i such that $i = 1, 2, \ldots, n$, observe that

$$\begin{aligned}\|\text{Proj}(X_i) - \widehat{\text{Proj}}(X_i)\| \quad &\leq \quad \|\widehat{c} - c\| + \left\| (\widehat{r} - r) \frac{\widehat{V}\widehat{V}^T(X_i - \widehat{c})}{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|} \right\| \\ &\quad + r \left\| \frac{\widehat{V}\widehat{V}^T(X_i - \widehat{c})}{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|} - \frac{VV^T(X_i - c)}{\|VV^T(X_i - c)\|} \right\|. \end{aligned} \tag{20}$$

The proof is complete if we can show that the last part of (20) converges to zero in probability. Again, observe that

$$\begin{aligned}&\left\| \frac{\widehat{V}\widehat{V}^T(X_i - \widehat{c})}{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|} - \frac{VV^T(X_i - c)}{\|VV^T(X_i - c)\|} \right\| \\ &\quad \leq \left\| \frac{\widehat{V}\widehat{V}^T(X_i - \widehat{c})}{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|} - \frac{\widehat{V}\widehat{V}^T(X_i - \widehat{c})}{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|} \cdot \frac{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|}{\|VV^T(X_i - c)\|} \right\| \\ &\qquad + \left\| \frac{\widehat{V}\widehat{V}^T(X_i - \widehat{c})}{\|VV^T(X_i - c)\|} - \frac{VV^T(X_i - c)}{\|VV^T(X_i - c)\|} \right\| \end{aligned} \tag{21}$$

We first show that $\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\| / \|VV^T(X_i - c)\| \xrightarrow{p} 1$. To see this observe that

$$\begin{aligned}1 - \frac{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c}) - VV^T(X_i - c)\|}{\|VV^T(X_i - c)\|} &\leq \frac{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c})\|}{\|VV^T(X_i - c)\|} \\ &\leq 1 + \frac{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c}) - VV^T(X_i - c)\|}{\|VV^T(X_i - c)\|}. \end{aligned} \tag{22}$$

Again

$$\begin{aligned}\frac{\|\widehat{V}\widehat{V}^T(X_i - \widehat{c}) - VV^T(X_i - c)\|}{\|VV^T(X_i - c)\|} &\leq \frac{\left\|\widehat{V}\widehat{V}^T - VV^T\right\| \|X_i - c\|}{\|VV^T(X_i - c)\|} \\ &\quad + \frac{\|\widehat{V}^T\| \|\widehat{c} - c\|}{\|VV^T(X_i - c)\|} \xrightarrow{p} 0. \end{aligned} \tag{23}$$

The above convergence is due to the following facts:

i. For any vector z and any real matrix A with lowest singular value σ, $\frac{\|Az\|}{\|z\|} \leq \sigma$.

ii. As X is a non-degenerate random variable, $P(\|X_i - c\| > \epsilon) > 1 - \delta$ for suitably chosen $\epsilon > 0$ and $\delta \in (0, 1)$.

iii. $\|VV^T - \widehat{V}\widehat{V}^T\| \xrightarrow{p} 0$ and $\|\widehat{c} - c\| \xrightarrow{p} 0$.

These facts imply that $\left[\|\widehat{V}\widehat{V}^T(X_i-\widehat{c})\| \,/\|VV^T(X_i-c)\|\right] \xrightarrow{p} 1$

Also note that the last part of the RHS of (21) is the same as the last part of the RHS of equation (22). Therefore, the proof of this part results from the convergence (23).

□

6.5 Proof of Theorem 4

Recall that a hyperplane can be viewed as a sphere with infinite radius, which is rigorously stated in the following lemma.

Lemma 4. *Let V be a d-dimensional subspace of $\mathbb{R}^{d+1}$ and $\alpha > 0$ is a fixed positive real number. Then for any ϵ, there exists a sphere $S(c,r)$ such that $\|x - \mathrm{Proj}(x)\| < \epsilon$ for any $x \in V$ with $\|x\| \le \alpha$, where $\mathrm{Proj}(x) = c + \frac{r}{\|x-c\|}(x-c)$ is the spherical projection to $S(c,r)$.*

Proof. This is a direct corollary of Taylor expansion. In fact we can let $c = r\mathbf{n}$ where $\mathbf{n}$ is the unit normal vector of V and $r = \frac{\alpha^2}{\epsilon}$. □

Lemma 5. *Let $H^* = (\mu^*, V^*) = \underset{H\in\mathcal{H}}{\operatorname{argmin}} \mathbb{E}_{\rho_U} d^2(x,H)$, then there exists $\theta > 0$ such that*

$$\mathbb{E}_{\rho_U} d^2(x, H^*) \le \theta\alpha^4,$$

where $\alpha = \mathrm{diam}(U)$.

Proof. This is a another direct corollary of the proof in 6.2.2. □

Now we prove theorem 4. Let (c^*, r^*) be the solution of SPCA, then

$$(\|x-c^*\| - r^*)^2 = \frac{(\|x-c^*\|^2 - r^{*2})^2}{(\|x-c^*\| + r^*)^2} \le \frac{(\|x-c^*\|^2 - r^{*2})^2}{r^{*2}} \le \frac{(\|x-c^*\|^2 - r^{*2})^2}{\delta^2}.$$

As a result, it suffices to find the upper bound of $(\|x-c^*\|^2 - r^{*2})^2$ which is the loss function of SPCA. Lemma 5 implies that there exists an affine subspace $\mu + V$ and $\theta > 0$ such that $d^2(x, \mu+V) \le \theta\alpha^4$ for any $x \in U$, then set $\epsilon = \theta\alpha^2$ in Lemma 4 so there exists c, r such that $d\big(y, c + \frac{r}{\|y-c\|}(y-c)\big) \le \epsilon = \theta\alpha^2$ for any $y = \mu + VV^\top x$ where $x \in U$. In fact, $r = \frac{\alpha^2}{\epsilon} = \frac{1}{\theta}$. For convenience, when x is the original point in U, let y be the linear projection of x onto the affine subspace $\mu + V$ and z be the spherical projection to sphere $S(c,r)$. By the triangular inequality, we have

$$\|x-z\| \le \|x-y\| + \|y-z\| \le \theta\alpha^2 + \theta\alpha^2 = 2\theta\alpha^2 = \theta'\alpha^2$$

by Lemma 4 and 5. Then we evaluate the loss function at such (c, r):

$$\begin{aligned}
(\|x-c\|^2 - r^2)^2 &= (x^\top x - 2c^\top x + c^\top c - r^2)^2 \\
&= ((z+x-z)^\top (z+x-z) - 2c^\top (z+x-z) + c^\top c - r^2)^2 \\
&= (z^\top z - 2c^\top z + c^\top c - r^2 + (x-z)^\top (x-z) + 2z^\top (x-z) - 2c^\top (x-z))^2 \\
&= (0 + \|x-z\|^2 + 2(z-c)^\top (x-z))^2 \\
&\leq (\|x-z\|^2 + 2|(z-c)^\top (x-z)|)^2 \\
&\leq (\theta^2\alpha^4 + 2\|z-c\|\|x-z\|)^2 \\
&\leq (\theta^2\alpha^4 + 2r\theta\alpha^2)^2 \\
&\sim 4r^2\theta^2\alpha^4 = \theta'\alpha^4
\end{aligned}$$

when α is sufficiently small. To conclude,

$$\mathbb{E}_{\rho_U} d^2(x, S_{V^*}(c^*, r^*)) \leq \frac{\mathbb{E}_{\rho_U}(\|x-c^*\|^2 - r^{*2})^2}{\delta^2} \leq \frac{\mathbb{E}_{\rho_U}(\|x-c\|^2 - r^2)^2}{\delta^2} \leq \theta\alpha^4.$$

□