Spherical Rotation Dimension Reduction with Geometric Loss Functions

Hengrui Luo[1], Jeremy E. Purvis[2], and Didong Li[3]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA, E-mail: hrluo@lbl.gov
[2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA, E-mail: purvisj@email.unc.edu
[3]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA, E-mail: didongli@unc.edu

Abstract

Modern datasets often exhibit high dimensionality, yet the data reside in low-dimensional manifolds that can reveal underlying geometric structures critical for data analysis. A prime example of such a dataset is a collection of cell cycle measurements, where the inherently cyclical nature of the process can be represented as a circle or sphere. Motivated by the need to analyze these types of datasets, we propose a non-linear dimension reduction method, Spherical Rotation Component Analysis (SRCA), that incorporates geometric information to better approximate low-dimensional manifolds. SRCA is a versatile method designed to work in both high-dimensional and small sample size settings. By employing spheres or ellipsoids, SRCA provides a low-rank spherical representation of the data with general theoretic guarantees, effectively retaining the geometric structure of the dataset during dimensionality reduction. A comprehensive simulation study, along with a successful application to human cell cycle data, further highlights the advantages of SRCA compared to state-of-the-art alternatives, demonstrating its superior performance in approximating the manifold while preserving inherent geometric structures.

Keywords: principal component analysis, high-dimensional dataset, dimension reduction.

1 Introduction

Modern data analysis presents the challenge of high-dimensionality, where the dataset usually comes as high dimensional vectors in $\mathbb{R}^d$, with a large d. Dimension reduction (DR) methods seek low dimensional representation of high dimension data (Mukhopadhyay et al., 2020; Zhang et al., 2021) to facilitate data visualization, subsequent data exploration, and statistical modeling in machine learning (Jolliffe and Cadima, 2016). Along with the difficulty in visualizations and computation, non-linearity obstructs conventional dimension reduction methods.

1.1 Motivation: Human Cell Cycle

In traditional DR methods (e.g., Principal Component Analysis (PCA), Pearson (1901)), it has been repeatedly pointed out that normalization preprocessing, including translations (by mean) and scalings (by standard deviation), is crucial in practicing DR (Jolliffe, 1995). However, rotation as a preprocessing step is less studied in the DR context. We are motivated by preserving non-trivial geometrical structure in DR tasks, and observed that *rotations* are as important as translations and scalings if we want to design DR methods that respect the underlying structure.

A compelling example that illustrates the need for advanced dimension reduction methods respecting the underlying structure is the analysis of cell cycle data. The cell cycle is an inherently cyclical process (Schafer, 1998) that consists of four proliferative phases: G1, S, G2, and M. Fluctuations in cell cycle genes and proteins show periodic, non-linear trends, that can be represented as a circle or sphere in a lower-dimensional space. Traditional linear methods may not adequately capture these properties, leading to the loss of crucial information.

Figure 1.1 presents a 2-dimensional representation of cell cycle data proposed in Stallaert et al. (2022), which included 40 single-cell features such as the expression or localization of core cell cycle regulators and signaling proteins. These features combine to form a multivariate cell cycle signature for each cell in the entire population, collected from 8,850 individual cells. Because individual cells are naturally asynchronous during data collection, the cells are randomly sampled over the entire cyclical distribution of possible cell cycle states. The phase of each cell (G1, S, G2, or M) was assigned using its unique molecular profile. Based on the known sequence of cell cycle phases, we would expect consecutive phases, such as G2 (red) and M (green) to be neighbors in the low dimensional projection. However, ex-

isting methods such as PCA, t-distributed Stochastic Neighbor Embedding (tSNE, Van der Maaten and Hinton (2008)), and Uniform Manifold Approximation and Projection (UMAP, McInnes et al. (2018)) (selected from the best results among other methods attempted) fail to preserve this structure in their representations.

This example motivates the development of a new DR method that utilizes spheres to represent high-dimensional data in low-dimensional spaces, effectively preserving the geometric structure and inherent cyclical nature of biological processes. In contrast to other DR methods, our proposed method, provides a representation on a 2-dimensional sphere, represented by longitude and latitude in the first panel (see Section 4.4 for more details). This SRCA representation in the lower dimensional space clearly preserves the cell cycle progression: G1 $\rightarrow$ S $\rightarrow$ G2 $\rightarrow$ M $\rightarrow$ G1, where the latitude (y-axis) is in the mod 2π sense, meaning $\pi = -\pi$. This biological periodicity, or in other words sphericity, is of central importance in analyzing cell data (Stallaert et al., 2022). In general, disruption of these low-dimensional structures in a high-dimensional dataset (Luo et al., 2020; Luo and Strait, 2022) diminishes the effectiveness of the subsequent analysis procedure like clustering and classification.

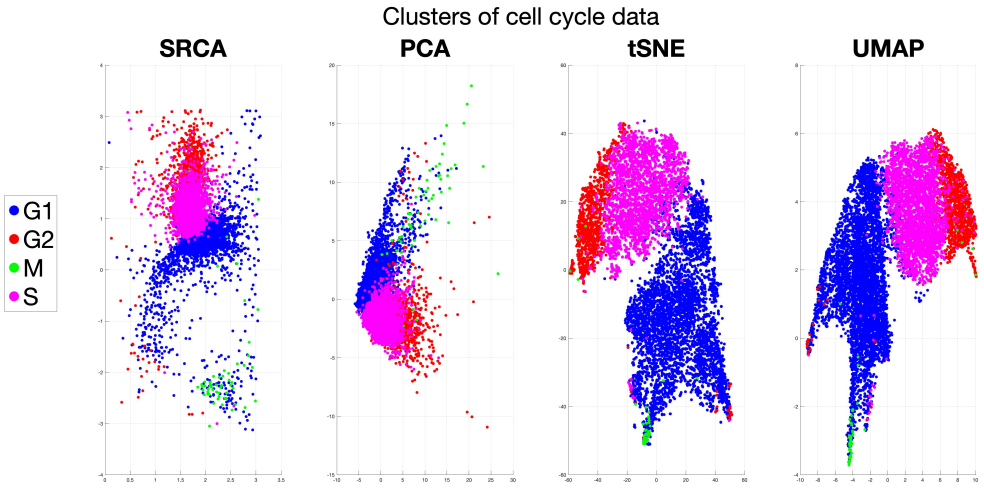

Figure 1.1: 2-dimensional representation of cell cycle data, colored by different cell phases.

	Local	Global
Manifold learning	LLE, tSNE, UMAP	MDS, Isomap, GPLVM
Manifold estimation	PCurv, GMRA, LPE, SAME, Spherelets	PCAs, SPCA, SRCA

Table 1.1: Conceptual categorization of selected dimension reduction methods

1.2 Related Literature

Sphericity induced by periodicity in the above data example requires the development of sophisticated non-linear DR methods designed to preserve certain structures in the data. The common assumption is that the observetaions $x_1, \cdots, x_n$ are near a manifold M embedded in $\mathbb{R}^d$. Table 1.1 provides a selected collection of dimension reductions methods loosely categorized in two ways. Algorithms in the first row are known as "manifold learning" (Lin and Zha, 2008), which output some low dimensional features in a new Euclidean space of dimension d' instead of an estimate of M, denoted by $\widehat{M}$. These methods include Locally Linear Embedding (LLE, Roweis and Saul (2000)), tSNE, , UMAP, Multi-Dimensional Scaling (MDS, Kruskal (1978)), Isomap (Tenenbaum et al., 2000), Gaussian Processes Latent Variable Model (GPLVM, Titsias and Lawrence (2010)), etc.

In contrast, the other type of DR methods, known as "manifold estimation", which estimates M in $\mathbb{R}^d$ directly, has been attracting researchers' attention (Genovese et al., 2012). There is an immense literature in local methods including Principal Curves (PCurv), Geometric Multi-Resolution Analysis (GMRA Allard et al. (2012)), Local Polynomial Estimator (Aamari and Levrard (2019)), Structure-Adaptive Manifold Estimation (SAME Puchkin and Spokoiny (2022)), Spherelets (Li et al., 2022), etc. The common idea behind these methods is to partition the space into local regions, and apply local, often linear, method to each small region. The intuition is that a manifold can be locally approximated by its tangent spaces. However, these local, nonparametric, complex methods are often computationally expensive and lack of interpretability.

A recent attempt to develop a spherical analogue of PCA is Li et al. (2022), which allows us to conduct dimension reduction and learn the shape of spherically distributed datasets. However, both PCA and SPCA fail when the sample size $n < d'$, the retained dimension (i.e., the dimension of the reduced dataset, the formal definition is introduced below) and are not easily applicable to high dimensional datasets. For instance, in the gene expression data, d is the number of genes, often over $20,000$ and the retained dimension d' is often chosen to be a couple of hundreds with the largest variability (Townes et al., 2019). While the sample size

could be much smaller, for example, less than 20 for certain tissues in the Genotype-Tissue Expression (GTEx) dataset (Consortium, 2020). In fact, most existing dimension reduction methods cannot handle $n < d'$ without substantial modifications.

In this paper, we focus on parametric global methods and derive an DR method called *spherical rotation dimension reduction* (SRCA), that preserves the sphericity constraints of the dataset. Unlike some competitors, this method is applicable to high-dimensional datasets regardless of retained dimensions and sample sizes. SRCA is scalable and interpretable, and will not destroy not only the geometry but also the topology of the dataset. (see Supplement A for synthetic examples).

Specifically, we focus on biological and genetic datasets, where dimension reduction is adopted by biologist directly before clustering and subsequent tasks (Johnson et al., 2022; Zhou and Sharpee, 2018). Our method also echoes and exemplifies a grander community belief that any dimension reduction should be guided by the need of its subsequent analyses and respect the structure in the original dataset.

2 Methodology

In this section, we outline the proposed procedure that aims to minimize a geometric loss function, specifically the mean squared errors between the original data points x_i and dimension-reduced data points $\widehat{x_i}$.

We denote the intrinsic dimension of the support of the reduced dataset by d' and refer to it as the *retained dimension*. It is worth noting that some literature uses the embedded dimension as d'. For example, if the reduced dataset lies in $\mathbb{S}^1$ embedded into $\mathbb{R}^2$, we would consider the dimensionality of the reduced data to be $d' = 1$ and not 2, as $\mathbb{S}^1$ is a one-dimensional manifold.

2.1 PCA and SPCA Revisited

We begin by recalling that PCA identifies a low-rank linear subspace from observations $\mathcal{X} = x_1, \cdots, x_n \subset \mathbb{R}^d$ by minimizing the sum of squared error loss function:

$$\min_{V \in \mathbb{R}^{d \times d'}} \sum_{i=1}^{n} \|x_i - \widehat{x_i}\|^2 = \sum_{i=1}^{n} \|x_i - \bar{x} - VV^T(x_i - \bar{x})\|^2, \text{s.t. } V^T V = I_{d'}.$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean calculated in $\mathbb{R}^d$. The solution to this optimization problem yields a rotation matrix V that defines a subspace, called the solution to PCA.

From an optimization perspective, PCA is a problem for a given (geometric) loss function (Journée et al., 2010), which quantifies the l_2 errors between the observation and the subspace. SPCA (Li et al., 2022) aims to find the optimal sphere S. Unlike PCA's planar solution, the solution of SPCA is a sphere with center c and radius r residing in the linear subspace V to fit the data. SPCA does not minimize the sum of squared distances between observations and the sphere; instead, it employs a two-step algorithm to minimize the sum of point-to-plane and projection-to-sphere distances.

The desired one-step algorithm was not explored in the original paper (Li et al., 2022) since the problem is theoretically more complicated and lacks a closed-form solution (See Supplement G). The SRCA can be shown to attain this one-step goal.

2.2 Geometric Loss Function

Given a sphere centered at c with radius r, we first assume that it lies in a subspace parallel to a coordinate plane in $\mathbb{R}^d$ determined by $\mathcal{I} \subset \{1, \cdots, d\}$ after a linear transformation determined by a (non-singular) matrix W. We denote such low-dimensional "sub-sphere" by $S_{\mathcal{I}}(c, r)$ and use the notation $I_{\mathcal{I}}$ to denote an identity matrix with ones in (i, i)-th entries, $i \in \mathcal{I}$ but zeros in the rest entries, then the point-to-sphere distance from a generic point x_i to this sphere can be expressed as

$$\begin{aligned} d(x_i, S_{\mathcal{I}}(c, r))^2 &= (x_i - c)^T W I_{\mathcal{I}^c}(x_i - c) + \left(\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}(x_i - c)} - r\right)^2 \\ &= (x_i - c^T) W (x_i - c) + r^2 - 2r\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}(x_i - c)}. \end{aligned}$$

Note that the assumption that the sphere lies in a coordinate plane is essential; otherwise no closed form expression can be written. With this point-to-sphere distance, the (geometric loss) function can be written as

$$\begin{aligned} \mathscr{L}(c, r, \mathcal{I} \mid \mathcal{X}, W) &= \sum_{i=1}^{n} \left((x_i - c)^T W (x_i - c) + r^2 - 2r\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}(x_i - c)} \right) \\ &=: \sum_{i=1}^{n} \rho(x_i; \theta), \end{aligned}$$

where we use the notation $\theta = (c, r)$ with a given $\mathcal{I}$, and the notation $\rho(x_i; \theta)$ is adopted to emphasize its additive form and to facilitate the later theoretic discussions. Our dimension reduction procedure can be described as solving the optimization problem below:

$$\min_{\mathcal{I}\subset\{1,\cdots,d\},c\in\mathbb{R}^d,r\in\mathbb{R}^+} \mathscr{L}(c, r, \mathcal{I} \mid \mathcal{X}, W) = \min_{\mathcal{I}\subset\{1,\cdots,d\},c\in\mathbb{R}^d,r\in\mathbb{R}^+} \sum_{i=1}^{n} \Big((x_i - c)^T W (x_i - c) + r^2 - 2r\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x_i - c)} \Big), \text{ s.t.} |\mathcal{I}| = d' + 1 \tag{2.1}$$

Using the loss function defined in (2.1), we can simultaneously estimate the center c, radius r and $\mathcal{I}$ by solving the optimization problem. Since there are at most 2^d possible choices of $\mathcal{I}$, it is straightforward to verify that this one-step optimization problem (2.1) can be equivalently solved in a two-step procedure: First, select a subset of indices $\mathcal{I} \subset \{1, \cdots, d\}$, and then estimate the center c and radius r of the dataset $\mathcal{X}$. This optimization problem can also be solved iteratively on variables $\mathcal{I}, r, c$. The binary search can be the first step, followed by esimating the center c and radius r:

1. Given c and r, perform an exhaustive binary search among all possible $\mathcal{I}$.

2. Given $\mathcal{I}$ and c, take the derivative of $\frac{\partial \mathscr{L}}{\partial r} = 0$ to obtain:

$$r = \frac{1}{n} \sum_{i=1}^{n} \sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x_i - c)}.$$

3. Given $\mathcal{I}$ and r, take the derivative of $\frac{\partial \mathscr{L}}{\partial c} = 0$ to obtain:

$$\frac{\partial \mathscr{L}}{\partial c} = \sum_{i=1}^{n} \left(-2W(x_i - c) + 2r \frac{I_{\mathcal{I}} \sqrt{W} (x_i - c)}{\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x_i - c)}} \right) = 0.$$

   Observe that if $j \in \mathcal{I}$, the j-th coordinate of the second term is zero, so we have $c_j = \frac{1}{n} \sum_{i=1}^{n} x_{i,j}$ for any $j \in \mathcal{I}^c$. For $j \in \mathcal{I}$, an analytic solution is difficult to find, but gradient descent can provide a numerical solution.

So far, we have assumed that the underlying support $S_{\mathcal{I}}$ has coordinate axes that are parallel to the coordinate axes. To make this assumption more realistic, we propose to rotate the dataset $\mathcal{X}$ so that it can be viewed in a position such that its axes are parallel to the coordinate axes. Then, we solve the optimization problem (2.1) for the rotated dataset and

rotate it back to obtain reduced dataset. The rotation can be chosen according to the types of datasets as appropriate in the procedure, as is discussed in Sec.4.5.

The exhaustive binary search over 2^d possible subsets is computationally expensive. We observe that the core of the optimization problem lies in the subset selection of index set $\{1, \cdots, d\}$. We can rephrase the optimization problem (2.1) as follows:

$$\min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \sum_{i=1}^{n} \Big((x_i - c)^T W (x_i - c) + r^2 \\ -2r\sqrt{(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c)} \Big) \quad , \text{s.t. } \|v\|_{l_0} = d' + 1, \tag{2.2}$$

Since l_0-norm is not convex, solving this problem requires a brute-force step in finding optimal v whose entries are either 0 or 1 (and hence $\mathcal{I}$ since $v^T I v = I_{\mathcal{I}}$), as detailed in Algorithm 2. Instead of using l_0 directly in the original problem, we consider the following computationally cheaper alternative:

$$\min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \sum_{i=1}^{n} \Big((x_i - c)^T W (x_i - c) + r^2 \\ -2r\sqrt{(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c)} \Big) \quad , \text{s.t. } \|v\|_{l_k} \leq d' + 1, \tag{2.3}$$

where $k \geq 1$ for a convex surrogate norm (usually l_1 is sufficient). This kind of relaxation is proposed in optimization (Boyd et al., 2004). As a practical suggestion, when there are more than 500 combinations of binary indices to search exhaustively, we recommend ℓ_1 relaxation as in (2.3), otherwise we perform an exhaustively search. For very high-dimensional dataset, the empirical performances for l_0 and l_1 penalties are similar.

2.3 SRCA Method

Algorithm 1: SRCA dimension reduction algorithm

Data: X (data matrix consisting of n samples in $\mathbb{R}^d$)

Input: d' (the dimension of the sphere), W (the covariance weight matrix, by default $W = I_d$), λ (optional, the sparse penalty parameter), rotationMethod (the method we use to construct the rotation matrix).

Result: $\hat{c}$ (The estimated center of $S_{\mathcal{I}}$ in $\mathbb{R}^d$), $\hat{r}$ (The estimated radius of $S_{\mathcal{I}}$), $\mathcal{I}_{opt}$ (The optimal index subset)

`GetRotation` (X,rotationMethod) obtains a $d \times d$ rotation matrix based on the data matrix X. The rotationMethod option specifies what method we use to construct the rotation matrix, by default, we use PCA to obtain a rotation matrix.

`ProjectToSphere` (X,c,r,k) is a projection that projects the point set X onto an l_k-sphere of center c and radius r via $X \mapsto c + \frac{X-c}{\|X-c\|_{l_k}} \cdot r$.

begin
- Standardize the dataset by subtracting its empirical mean $X = X - \bar{X}$
- Construct a rotate matrix $R =$ `GetRotation` (X,rotationMethod)
- $X_{rotated} = X * R$
- $\mathscr{L}_{opt} = \infty$
- **while** $\mathcal{I} \subset 1, \cdots, p$ **do**
  - Solve the optimization problem (2.2) with respect to c, r with a fixed $\mathcal{I}$
  - Denote the solution as $c_{cur}, r_{cur}, \mathcal{I}_{cur}$
  - **if** $\mathscr{L}(c, r, \mathcal{I} \mid \mathcal{X}) \leq \mathscr{L}_{opt}$ **then**
    - $\mathscr{L}_{opt} \leftarrow \mathscr{L}(c, r, \mathcal{I} \mid \mathcal{X})$
    - $c_{opt} \leftarrow c_{cur}$
    - $r_{opt} \leftarrow r_{cur}$
    - $\mathcal{I}_{opt} \leftarrow \mathcal{I}_{cur}$
  - **end**
- Construct the binary index vector $\eta = (\eta_i), \eta_i = 1$ iff $i \in \mathcal{I}$ and $\eta_i = 0$ otherwise.
- $\hat{c} = c_{opt} \cdot \eta * R^{-1} + \bar{X}$
- $\hat{r} = r_{opt}$
- $X_{rotated}(:, \mathcal{I}) \leftarrow 0$
- $X_{rotated} \leftarrow$ `ProjectToSphere` (X,$\hat{c}$,$\hat{r}$,k)
- $X_{output} \leftarrow X_{rotated} * R^{-1} + \bar{X}$

end

The method discussed above is refereed to as the *spherical rotation dimension reduction* (SRCA) method and presented in Algorithm 1, which employs geometric loss functions designed for spherical datasets. The key steps of our proposed SRCA dimension reduction method can be summarized into a "Rotate-Optimize-Project" scheme as follows, with l_1

algorithms detailed in Supplement C.

Rotate: Conduct the rotation. With the chosen rotation method, we construct a rotation matrix R based on the dataset $\mathcal{X}$. We translate and rotate the dataset $\mathcal{X}$ to a standard position $(\mathcal{X}-\bar{\mathcal{X}})R$, so that we can reasonably assume that the axes of the ellipsoid are parallel to the coordinate axes (Jolliffe, 1995).

Optimize: Solve the optimization for the best $d'+1$ axes. We perform dimension reduction based on the geometric loss function discussed above. As stated in (2.2), we conduct dimension reduction by minimizing the loss function based on the point-to-sphere distance to the estimated sphere $S_{\mathcal{I}}$, to obtain the optimal v_{opt} and the optimal index set $\mathcal{I}_{\text{opt}}$.

$$(v_{\text{opt}}, c_{\text{opt}}, r_{\text{opt}}) = \arg\min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \sum_{i=1}^{n} \Big((x_i-c)^T W(x_i-c) + r^2 \\ -2r\sqrt{(x_i-c)^T \sqrt{W}^T v^T I v \sqrt{W}\,(x_i-c)}\Big) \text{ s.t. } \|v\|_{l_0} = d'+1, \tag{2.4}$$

where the constraint can be relaxed by $\|v\|_{l_k} \le d'+1$.

Project: project onto the optimal sphere. Now we project the datapoints back into full space with the chosen dimension and axes, placing x_i back onto the sphere $S(c_{\text{opt}}, r_{\text{opt}})$ with the estimated center c_{opt} and radius r_{opt}, the SRCA projection is given by:

$$\widehat{x}_i = c_{\text{opt}} \cdot v_{\text{opt}} + r_{\text{opt}} \frac{(x_i - c_{\text{opt}}) v_{\text{opt}}^T I v_{\text{opt}}}{\|(x_i - c_{\text{opt}}) v_{\text{opt}}^T I v_{\text{opt}}\|}. \tag{2.5}$$

3 Theoretical Results

We have established the procedure for our propose method, SRCA, in an algorithmic way. Next, we discuss and provide some theoretical results that guarantee the performance of SRCA in applications. Proofs are deferred to supplementary materials, but we want to emphasize that the techniques of ρ-loss (Huber et al., 1967) and Γ-convergence (Braides et al., 2002) are introduced to tackle probabilistic properties for DR methods.

3.1 Convergence

Unlike SPCA, SRCA does not have a closed form solution (i.e., analytic expression of center and radius estimates in terms of dataset $\mathcal{X}$) but relies on the solution to an optimization problem. Therefore, the convergence of this optimization becomes central in our theory development. We briefly discuss the convergence guarantee for the algorithm we designed. In the binary search situation, for each fixed choice of indices, we compute the gradient of loss function. With mild assumptions, gradient descent provides linear convergence. If the optimization problem (2.4) has solutions, then the solution is clearly unique. This is because there are only finitely many v such that $\|v\|_{l_0} = d' + 1$, and the binary search in the standard algorithm would exhaustively search all possible values of v.

To these ends, we provide a basic convergence for a sub-problem in our Algorithm 1 via the gradient descent algorithm of positive constant step size. The sub-problem is defined by the following loss function:

$$\mathscr{L}_v(c, r|\mathcal{X}, W) = \mathscr{L}_v(c, r) \coloneqq \sum_{i=1}^{n} \left((x_i - c)^T W (x_i - c) + r^2 \right. \tag{3.1}$$

$$\left. -2r\sqrt{(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c)} \right) \tag{3.2}$$

The following theorem guarantees the convergence of SRCA.

Theorem 1. *For a fixed vector v (or equivalently $I_{\mathcal{I}}$), if we assume that $\|x_i - c\| \leq R_1$, $\forall i = 1 \cdots, n$, $r \leq R_2$ and $|\lambda_{\max}(W)| \leq R_3$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue, then for a positive finite constant step size independent of the iteration number k, the gradient descent algorithm (c.f., the setting in Boyd et al. (2004, 2003)) converges to the optimal value in the following sense,*

$$\lim_{k \to \infty} \mathscr{L}_{v,k} \to \mathscr{L}_v^*$$

where $\mathscr{L}_{v,k} = \mathscr{L}_v(c_k, r_k)$, (c_k, r_k) is the value in the k-th iterative step in the gradient descent algorithm, and $\mathscr{L}_v^*$ denotes the minimum of the loss function for this fixed v.

These results justify that for a fixed v (or equivalently $\mathcal{I}$) we can solve the sub-problem defined by the above function, and since we conduct an exhaustive search for the index vector v, we can find the solution to the original problem (2.4) as well.

3.2 Consistency

In this section, we assume the observed data are from a "true" but unknown sphere $S_{I_0}(c_0, r_0)$ and show that the solution of SRCA is consistent, that it, we can find the true sphere as long as we have enough samples.

Theorem 2. *Assume $x_i \in S_{\mathcal{I}_0}(c_0, r_0),\ \forall i = 1, \cdots, n$ and $n > d' + 1$. Let $\widehat{\mathcal{I}}_k, \widehat{c}_k, \widehat{r}_k$ be the solution of SRCA after k iteration in the solution of the corresponding optimization problem, then*

$$(\widehat{\mathcal{I}}_k, \widehat{c}_k, \widehat{r}_k) \xrightarrow{k\to\infty} (\mathcal{I}_0, c_0, r_0).$$

However, the assumption that are observations are exactly on a sphere is unrealistic in practice, as the data often come with measurement errors. Instead, we adopt following (common) assumption in manifold estimation: $x_i = y_i + \epsilon_i$, where the unobserved y_i's are exactly from a sphere $S(c_0, r_0)$ and ϵ_i represents the measurement error. The next theorem fills in the gap using the Γ-convergence (Braides et al., 2002), which is first applied in DR problems.

Theorem 3. *Under the following assumptions:*

(A0) The index vector $\mathcal{I}$ is fixed and the parameter $\theta = (c, r) \in \Theta \coloneqq [-C, C]^d \times [R_0, R] \subset \mathbb{R}^d \times \mathbb{R}^+$ for some $C, R_0, R > 0$.

(A1) x_i's are compactly supported.

(A2) $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \|\epsilon_i\| = 0$

the SRCA solution $\widehat{\theta}_n \to \theta_0$ as $n \to \infty$.

In other words, the SRCA estimator based on noisy samples is consistent, that is, converges to the true parameter θ_0, as low as the noise decays to zero with sample size. In fact, this assumption is even weaker than those in existing literature, see Aamari and Levrard (2019); Fefferman et al. (2018); Maggioni et al. (2016) for more details. For example, in Aamari and Levrard (2019) the amplitude of the noise is assume to be $\|\epsilon\| \sim n^{-\frac{\alpha}{d}}$ for $\alpha > 1$. In contrast, we only require $\|\epsilon\| \to 0$, so $\|\epsilon\| \sim n^{-\alpha}$ for any $\alpha > 0$ or even $\|\epsilon\| \sim \frac{1}{\log n}$ is good enough.

3.3 Asymptotics

In this section, we consider the asymptotic behavior of SRCA optimization result when the underlying dataset is assumed to be drawn from a probabilistic distribution, regardless whether it is supported on a sphere or not.

To yield the asymptotic results, we need to take the perspective of robust statistics as mentioned in the end of Section F. The asymptotic theory here is a specific case of empirical risk minimization. With a mild technical assumption that the parameter $\theta = (c, r) \in \Theta \coloneqq [-C, C]^d \times [R_0, R] \subset \mathbb{R}^d \times \mathbb{R}^+$ for some $C, R_0, R \in (0, \infty)$, our loss function and optimization problem can be expressed as

$$\min_{\mathcal{I}\subset\{1,\cdots,d\}, c\in[-C,C]^d\subset\mathbb{R}^d, r\in[R_0,R]\subset\mathbb{R}^+} \mathscr{L}(c, r, \mathcal{I} \mid \mathcal{X}, W) \tag{3.3}$$

$$= \min_{\mathcal{I}\subset\{1,\cdots,d\}, c\in[-C,C]^d\subset\mathbb{R}^d, r\in[R_0,R]\subset\mathbb{R}^+} \sum_{i=1}^{n} \Big((x_i - c)^T W (x_i - c) + r^2 \tag{3.4}$$

$$-2r\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x_i - c)} \Big) \text{ s.t. } |\mathcal{I}| = d' + 1$$

For a fixed $\mathcal{I}$, (3.4) can be written in the form of (3.1) in Huber (2004), i.e.,

$$\rho(x; \theta) = \left((x_i - c)^T W (x_i - c) + r^2 - 2r\sqrt{(x - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x - c)} \right)$$

Correspondingly, we can write Huber's ψ-type function of ρ as $\psi(\theta) = \frac{\partial \rho(\theta)}{\partial \theta}$.

Classical style asymptotic results are presented below in Theorem 4, which states that, with mild assumptions, the estimates T_n obtained by solving SRCA would estimate the center and radius of the spherical space consistently, corresponding to Huber's ρ-type estimator consistency (Huber et al., 1967); Theorem 5 states that with more stringent conditions on continuity of T_n, asymptotic normality of these estimators can also be formulated into Huber's ψ-type normality.

To apply these two asymptotic results, we need to make mild assumptions on the parameter space and assume that we already know the retained dimension $d' + 1$.

Theorem 4. *Suppose (A0) in Theorem 3 holds and the samples $x_1, \cdots, x_n \in \mathbb{X} = \mathbb{R}^d$ of size n are i.i.d. drawn from the common distribution P. P has finite second moments on the probability space $(\mathbb{X}, \mathcal{A}, \nu)$ with Borel algebra $\mathcal{A}$ and Lebesgue measure ν. Then the consistent estimator T_n for parameter $\theta = (c, r)$ defined by*

$$\frac{1}{n}\sum_{i=1}^{n} \rho(x_i; T_n) - \inf_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \rho(x_i; \theta) \xrightarrow{n\to\infty} 0, \text{ a.s. } P$$

would converge in probability and almost surely to θ_0 w.r.t. P (for the true parameter values θ_0 defined on page 53). Particularly, T_n can be realized as a solution to our optimization

problem (3.3) above.

Unlike Theorem 1, which concerns the convergence of the algorithm, we assume that the fixed i.i.d. samples $\mathcal{X}$ are drawn from a probability distribution. Similarly, we have a distributional result as follows.

Theorem 5. *In addition to the assumption in Theorem 4, we assume* $P\left(|T_n - \theta_0| \leq \eta\right) \to 1$ *as* $n \to \infty$*, then the estimator* T_n *defined by*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i; T_n) \xrightarrow{n\to\infty} 0, \, a.s. \; P$$

would satisfy $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i; T_n) + \sqrt{n}\lambda(T_n) \xrightarrow{n\to\infty} 0, a.s. \; P$ *Particularly, our loss function would satisfy differentiability at* θ_0 *and* $\sqrt{n}\left(T_n - \theta_0\right)$ *is asymptotically normal with mean zero and covariance matrix*

$$\left(\nabla_{\theta_0}\lambda^{-1}\right) \cdot \Big(\left[\psi(x_i;\theta_0) - \mathbb{E}_P\psi(x_i;\theta_0)\right]^T \left[\psi(x_i;\theta_0) - \mathbb{E}_P\psi(x_i;\theta_0)\right] \Big) \cdot \left(\nabla_{\theta_0}\lambda^{-1}\right)^T.$$

Defined by the geometric loss function $\mathscr{L}$, SRCA does not have an analytic solution, but this loss benefits from the theoretic results above and can be replaced by other types of loss functions, enabling SRCA to be applied more widely.

Note that we also assumed that the index set $\mathcal{I}$ is fixed for our statements of theorems. In the exhaustive search, these results above can be applied individually to fixed $\mathcal{I}$; but in the l_1-relaxed problem (2.4), since the optimization is a joint optimization our asymptotic results Theorem 4 and 5 in this section do not apply.

3.4 Loss Function Minimization

Next, we consider the theoretic behavior of SRCA in terms of approximating a general manifold. The following theorem compares the MSE of PCA, SRCA (when the rotation is chosen by PCA) and SPCA:

Theorem 6. *Given data* $x_1, \cdots, x_n$ *in a bounded subset of* $\mathbb{R}^d$*, let* H *be the best subspace obtained by PCA,* S_1 *be the sphere obtained by SPCA and* S_2 *be the sphere obtained by SRCA with rotation provided by PCA, then*

$$\sum_{i=1}^{n} d^2(x_i, S_2) \leq \min\left\{ \sum_{i=1}^{n} d^2(x_i, H), \sum_{i=1}^{n} d^2(x_i, S_1) \right\}.$$

That is, SRCA has the best approximation performance in terms of MSE among PCA, SPCA and SRCA, regardless of the true support of the observations.

To summarize and interpret our theoretical results briefly, Theorem 1 ensures that a gradient-descent algorithm can be used for solving the loss function minimization problem (2.1) for any finite samples with convergence guarantees; Theorem 2 and 3 show that SRCA can recover the true sphere, if it exists, when the data are clean or with measurement error. For general case where the observations are not necessarily supported by a sphere, Theorems 4 and 5 ensure that the sequence of finite-sample minimizers of our loss function asymptotically converges to minimizer θ_0; Theorem 6 points out that SRCA can better approximate the unknown support in terms of MSE than PCA and SPCA.

4 Numerical Experiments

With theoretical results above on MSE, we also wish to examine the practical performance of SRCA against the state-of-the-art dimension reduction methods on real datasets. We focus on the empirical structure-preserving and coranking measurements below, an application to the motivating dataset about cell cycle, and discuss the choice of parameters in SRCA in the end. Details of selected datasets are in Supplement D.

4.1 MSE

As a dimension reduction method, the most common and natural measurement of performance is based on the mean squared error (MSE) between the original and reduced datasets, which measures how close the manifold is to the original observations. However, most dimension reduction methods only output low dimension features, like LLE, Isomap, tSNE, UMAP, GPLVM, etc, where the MSE is not well-defined because the low dimensional features cannot be trivially embedded into original data space $\mathbb{R}^d$. Algorithms that output projected data in the original space $\mathbb{R}^d$ include SPCA, PCA, and our proposed SRCA.

Table 4.1 presents the MSEs of three competing algorithms on these datasets with $d' = \min\{d-1, 4\}$. The out-sample MSEs show a similar patterns and is postponed to the Supplement I. It is evident that SRCA has the property of MSE minimization for most datasets and most d', as predicted by the theory in Theorem 6.

Dataset	Method/$d' =$	1	2	3	4
Banknote	PCA	15.6261	6.3356	1.9479	
	SPCA	16.3717	8.1004	1.7348	
	SRCA	**13.439**	**5.5088**	**1.0743**	
Power Plant	PCA	222.2971	55.4460	23.5173	**2.9957**
	SPCA	162.8865	102.1006	45.5793	41.5251
	SRCA	**150.8041**	**52.1439**	**19.8839**	3.0373
User Knowledge	PCA	0.1921	0.1253	0.0718	0.0311
	SPCA	0.1465	0.0893	0.0477	0.0148
	SRCA	**0.1458**	**0.0887**	**0.0471**	**0.0142**
Ecoli	PCA	0.076693	0.035222	0.020522	**0.00756**
	SPCA	**0.047776**	0.032948	0.019648	0.01136
	SRCA	0.076660	**0.032799**	**0.018332**	**0.00756**
Concrete	PCA	6.7056	4.7711	3.3636	2.3218
	SPCA	5.4743	3.9934	2.9681	1.9181
	SRCA	**5.4741**	**3.9909**	**2.9552**	**1.8985**
Leaf	PCA	0.0245	0.0121	0.0059	0.0038
	SPCA	**0.0155**	0.0094	0.0072	0.0037
	SRCA	0.0230	**0.0093**	**0.0054**	**0.0033**
Climate	PCA	1.4100	1.3204	1.2323	1.1450
	SPCA	1.3563	1.2648	1.1781	1.0907
	SRCA	**1.3554**	**1.2646**	**1.1780**	**1.0905**

Table 4.1: MSE for different experiments.

4.2 Cluster Preserving

Cluster structure properties of different dimension reduction algorithms varies, however, we hope that the data points belonging to the same group in the original dataset, are close together in the dimension-reduced dataset.

For visualization purposes, we fix retained dimension to be $d' = 2$ and compare the following six algorithms: SRCA, SPCA, PCA, LLE, tSNE, UMAP. We choose these state-of-the-art competitors to visualize in 2-dimensional figures. To further quantify the how

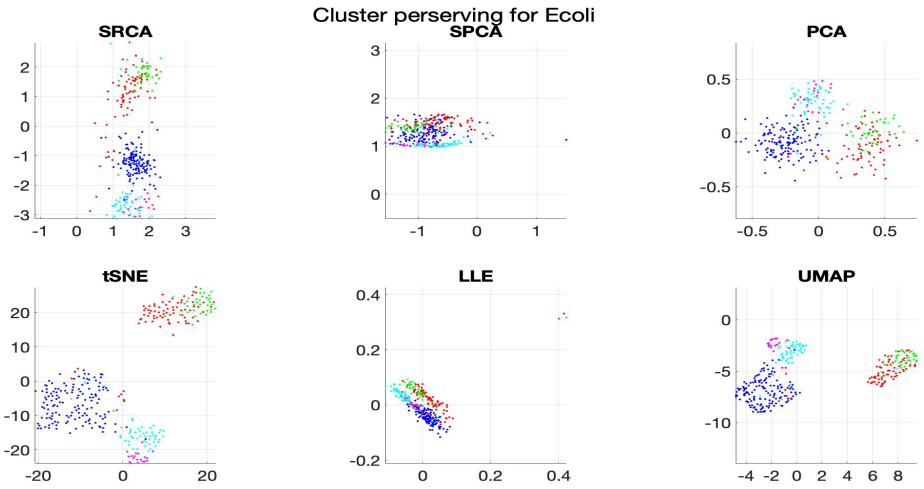

Figure 4.1: Cluster structures for Ecoli, $d' = 2$, there are five different clusters represented by different colors in the reduced dataset.

well the clustering structures are preserved, the Silhouette Score (SC, (Rousseeuw, 1987)), Calinski-Harabasz Index (CHI, (Caliński and Harabasz, 1974)) and Davis-Bouldin index (DBI, (Davies and Bouldin, 1979)) are considered. Higher SC and CHI, lower DBI imply better separation between clusters in the dataset. We provide these measures on the original labeled dataset (without any DR) as baselines.

Figure 4.1 and Table 4.2 show that SRCA outperforms SPCA, PCA and LLE in terms of all three metrics and comparable to tSNE and UMAP. SRCA also has advantage over its predecessor SPCA and simpler linear method like PCA. With these experiments and Supplement B, we conclude that if the dataset has strong spatial sphericity, we usually have good cluster preserving properties from SRCA. If the dataset is highly non-linear, tSNE and UMAP are usually better at the cost of creating fake clusters if tuning parameters are not well-chosen (Wattenberg et al., 2016; Wilkinson and Luo, 2020).

Index	Baseline	SRCA	SPCA	PCA	LLE	tSNE	UMAP
SC	0.257	0.267	0.260	0.200	0.209	0.293	0.290
CHI	133	192	190	215	46.6	376	376
DBI	1.49	1.59	1.58	2.56	2.40	1.37	1.32

Table 4.2: Clustering performance measures for Ecoli

4.3 Coranking Matrix

Another type of quantitative measures is based on the *coranking matrix* (Lee and Verleysen, 2009; Lueks et al., 2011). The coranking matrix can be viewed as the joint histogram of the ranks of original samples and the dimension-reduced samples. The coranking matrix can be used to assess results of dimension reduction methods. Entry q_{kl} in the coranking matrix is defined as $q_{kl} \coloneqq \{(i,j) \mid \rho_{ij} = k \text{ and } r_{ij} = l\}$, where $\rho_{ij} \coloneqq \{k : d(x_i, x_k) < d(x_i, x_j) \text{ or } d(x_i, x_k) = d(x_i, x_j), k < j\}$ stores the rank of the pair x_i, x_j in the original dataset; $r_{ij} \coloneqq \{k : d(\hat{x}_i, \hat{x}_k) < d(\hat{x}_i, \hat{x}_j) \text{ or } d(\hat{x}_i, \hat{x}_k) = d(\hat{x}_i, \hat{x}_j), k < j\}$ stores the rank of the pair $\hat{x}_i, \hat{x}_j$ in the dimension-reduced dataset, where the rank pair reversed in the dimension-reduced dataset. An ideal dimension reduction method should preserve all the ranks of these pairwise distances between original and reduced datasets. That is, we have identical ordering of these pairwise distances in the original space and the dimension-reduced space. Coranking matrix is a finer summary but is related to ijk rank test (See, e.g., Solomon et al. (2021)).

We provide three scores (the higher the better) related to coranking matrices of the dimension-reduced results: CC (cophenetic correlation, measuring correlation between distance matrices), AUC (area under curve for the R_{NX} score), WAUC (weighted AUC) computed from `coRanking` R-package (Kraemer et al., 2018).

To understand our subsequent analyses better, we referred our readers to the analysis of the dimension reduction result of simple examples like $\mathbb{S}^2$, $\mathbb{T}^2$ and a plane diffeomorphic to $\mathbb{R}^2$, evaluated by these coranking matrix related scores in Supplement B, where SRCA is the only dimension method that consistently behaves almost the best in plane, spheres and topologically non-trivial examples like torus when measured by coranking scores. Another advantage of SRCA over existing DR methods is that it allows $n < d'$, which happens to a variety of real datasets, specially for biomedical data where both d and d' are large. For example, in Genotype-Tissue Expression(GTEx) dataset (Consortium, 2020), some tissues are hard to collect so the sample sizes are small but the dimension is very high, like Kidney Medulla ($n = 4$), Fallopian Tube ($n = 9$) and Cervix Endocervix ($n = 10$). However, there are thousands of genes so we expect that the intrinsic dimension $d' > n$. Following the common practice of feature selection in this database, we subsetted the data to the most variable 500 genes (Townes et al., 2019). Most competitors mentioned before are not directly applicable anymore when $d' > n$, including tSNE, UMAP, Isomap, MDS, etc. For illustration purpose, we retain the first n dimensions. As a result, we present the three coranking based measurements on three tissues obtained from SRCA, SPCA, PCA and LLE for different d' in Figure 4.2.

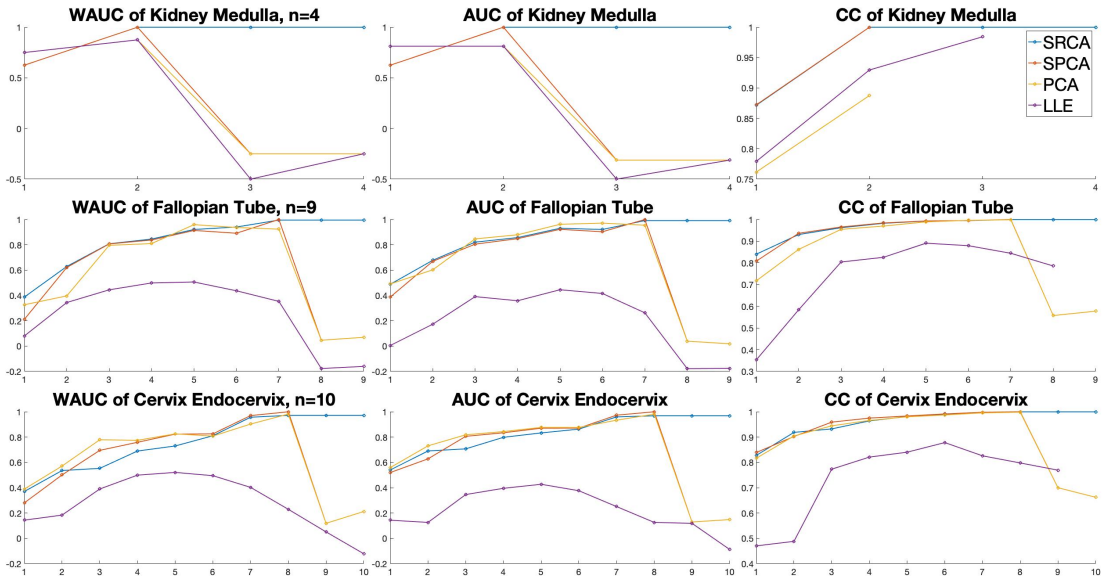

Figure 4.2: Coranking measurements of three GTEx tissues for different retained dimension, the horizontal axes are retained dimension d', the vertical axes are score values.

4.4 Application: Human Cell Cycle

The human cell cycle consists of four growth phases: G1, S, G2, and M. In a recent study, non-transformed human retinal pigmented epithelial (RPE) cells were genetically engineered to express a fluorescent cell cycle reporter that enables accurate identification of each cell's phase (i.e., G1, S, G2, or M) through time-lapse imaging (Stallaert et al., 2022). Subsequently, the cells were fixed and subjected to iterative indirect immunofluorescence imaging (4i) to measure 48 key cell cycle effectors in 8,850 individual cells. A total of 246 single-cell features were derived from this imaging dataset, including protein expression and localization (e.g., nucleus, cytosol, perinuclear region, and plasma membrane), cell morphological attributes (such as nucleus and cell size and shape), and microenvironment characteristics (like local cell density), ultimately generating a comprehensive cell cycle signature for each cell within the population.

In their study, Stallaert et al. (2022) narrowed the features to a set of 40 that most accurately predicted cell cycle phase (refer to Figure S1 panel A in Stallaert et al. (2022)). Thus, the reduced dataset has a sample size of $n = 8,850$ and an ambient dimension of $d = 40$. Our goal is to decrease the dimension to $d' = 2$ for visualization purposes, while maintaining the four clusters that correspond to the four phases (G1, S, G2, M) and the cyclic structure: G1 $\rightarrow$ S $\rightarrow$ G2 $\rightarrow$ M $\rightarrow$ G1. To account for the diverse units of the 40

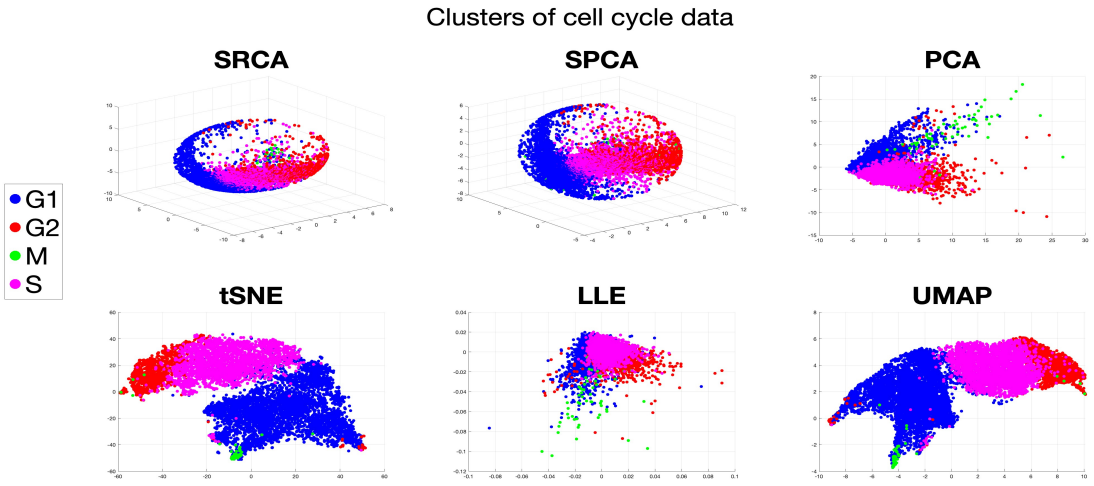

Figure 4.3: Cluster structures for human cell cycle, $d' = 2$, colored by phase.

selected features, we applied z-score normalization to the data.

Figure 4.3 displays the visualization of cell cycle data, with colors representing different cell cycle phases. While all algorithms can distinguish the four phases, PCA, tSNE, LLE, and UMAP fail to capture the cyclical structure. For example, the green points (M) should be located between the blue points (G1) and red points (G2); and magenta points (S) should be opposite to green points. In contrast, both SRCA and SPCA successfully recover the cyclical structure on 2-dimensional spheres. To compare SRCA and SPCA, we assess the MSE, as shown in the first column of Table 4.3.

	MSE	Var(G1)	Var(S)	Var(G2)	Var(M)
PCA	22.934	575	81	1297	3631
SPCA	22.252	566	110	276	763
SRCA	**22.098**	**633**	117	**434**	**475**

Table 4.3: Quantitative metrics of SRCA and SPCA for cell cycle data

Given that the true structure is cyclical, clustering metrics that depend on linear structures, such as the Silhouette score, are not suitable for this example. Instead, we use external biological information to validate our findings. Among the four phases, G1 cells are known to possess greater degrees of freedom (Chao et al., 2019; Stallaert et al., 2022), leading us to anticipate that the variance of samples within G1 will be the largest among the four phases, with variances for G2 and M being similar and variance for S being the smallest.

The variance is defined as the distance between each sample and the cluster mean, so a larger variance indicates a more dispersed distribution of points within that cluster. The final four columns in Table 4.3 demonstrate that SRCA more accurately captures the heterogeneity of cell activities across different phases.

4.5 Parameter Selection

It is a separate but important problem to select (or tune) the parameter of both classical and modern DR methods. For classical DR methods, like PCA or MDS, the parameter usually has a explicit geometric interpretation. For modern non-linear DR methods, like tSNE and UMAP, the parameters affect both reproducibility and interpretability of the resulting dimension-reduced dataset.

The first parameter that dictates the behavior of most DR methods is the retained dimension d', which can be determined by subsequent purpose (e.g., the tSNE and UMAP usually take $d' = 2, 3$ for visualizations).

The second parameter is the choice of rotation methods, which is highly data dependent and affects the clustering and visualization most. Regarding the MSE performance, Table 4.4 investigates the performance of SRCA with different kinds of rotations. We can see that the PCA rotation usually gives a reasonable result in terms of MSE performance. Both PCA and quartimax rotations used along with SRCA method outperforms dimension reduction of PCA and SPCA separately. We have similar observations for some other datasets with $n > d$ (e.g., Leaf, PowerPlant, etc., see Supplement D).

The choice of rotation methods could also be made to accommodate the type of noise in observations. In the situation where the tail behavior of the noise is close to Gaussian and the W is known (or, by default I), PCA is our default choice; but in the situation where the noise is non-Gaussian and we do not have much knowledge for W, then ICA (Hyvärinen and Oja, 2000) is a better alternative.

Based on the empirical evidence obtained from real datasets (e.g., Table 4.4), we recommend using PCA rotation as a default, but other types of rotations can be useful for specific datasets, if desired (Jolliffe, 1995).

We summarize the observations from above experiments in sections 4.1 to 4.5. Although SRCA is the slowest in terms of computational time among SRCA, SPCA and PCA, it is rather fast compared to some non–linear methods like Isomap.

When the retained dimensions $d' < \min\{d, n\}$, SRCA behaves very similarly to SPCA in terms of the MSE, both outperform PCA alone across different real datasets. The interesting

	PCA	SPCA	SRCA						
d'			PCA	varimax	orthomax	quartimax	equamax	parsimax	ICA
1	15.6	16.4	13.4	18.4	18.5	14.9	27.5	26.6	14.5
2	6.34	8.10	5.51	6.19	6.19	5.79	13.2	13.4	5.36
3	1.95	1.73	1.07	1.07	1.07	1.07	1.07	1.07	1.14

Table 4.4: MSE of Banknote (see SupplementD) for different rotation methods in the SRCA procedure. We also include two other DR methods PCA and SPCA to compare against SRCA. The first row records the DR methods (PCA, SPCA and SRCA); the second row records the optimal rotation method used by SRCA.

observation is that SRCA out-competes both of them in some simple but geometrically nontrivial examples like the ones in Supplement B, especially in clustering tasks. The choice of rotation methods can improve the SPCA but PCA rotation is usually good enough.

When $d > d' \geq n$, SRCA outperforms PCA, SPCA and other non-linear DR methods in terms of the coranking scores across different real datasets. Only SRCA yields consistently better dimension reduction results when $d' < n$ and $d' \geq n$.

5 Contribution

In this paper, we propose a novel DR method by proposing a rotation-based method with a geometric-induced loss function that minimizes the point-to-sphere distance from original to target spaces. Our motivation is to get the dimension reduction for spherical datasets (or datasets with spherical and elliptical structures) to respect the geometry in the original space. Its variant also works with a general weight matrix W and a sparsity penalty λ. The proposed method is statistically principled, and is theoretically guaranteed to perform well asymptotically.

Unlike traditional DR methods like PCA and MDS, SRCA works smoothly with a stable performance even when $d \geq d' + 1 > n$ which is extremely important in biomedical data DR, especially in gene expression data (e.g.,GTEx). Accompanying generalized algorithms for SRCA are also developed, with detailed convergence and a straightforward parallel potentiality for real-world practice. SRCA is related to PCA and SPCA but also generalizes the former into a spherical setting and the latter one into a one-step procedure. Most importantly, SRCA removes the $d' < n$ requirements in these predecessors in a unified framework using novel loss functions.

Compared to non-linear methods, SRCA has geometrical interpretation and practical con-

venience. Its unique binary search also allows parallelizations when applied to big datasets. A comprehensive experimental study of SRCA against a collection of state-of-art DR methods has been done with detailed qualitative and quantitative measures, revealing the superiority of SRCA.

Our current technique focuses on but is not limited to spherical datasets. Similar designs of loss functions can be generalized to a wider variety of spaces like symmetric spaces (Li et al., 2020) using multivariate decomposition technique like ICA, and a wider range of datasets like binary datasets (Landgraf and Lee, 2020). The geometric loss function has some relation to or is motivated by the statistical problem we are trying to solve. Then the reduced dataset should be more suitable for the subsequent statistical method to be applied.

To sum up, we provide a principled way that targets at an explicitly designed geometric loss function and provide an algorithm along with statistical guarantees and a comprehensive evaluation of this novel DR method.

There are several directions of future work that we wish to pursue. For example, it is of great interest to see how geometric or topological loss function DR methods perform in data visualization (Nigmetov and Morozov, 2022; Sigmund, 2001).

Acknowledgement

HL wants to thank Dmitriy Morozov, Leland Wilkinson for motivating discussions and comments on early manuscripts; Yin-Ting Liao for discussions in technical details; Justin D. Strait for helpful reading and suggestions. HL was supported by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231., who wants to thank the support of LBNL CRD during this research. DL wants to thank David Dunson and Tarek Zikry for motivating discussions. DL was supported by NIH/NCATS award UL1 TR002489, NIH/NHLBI award R01 HL149683 and NIH/NIEHS award P30 ES010126.

Our code for SRCA implementations and experiments are publicly available at https://github.com/hrluo/SphericalRotationDimensionReduction.

References

Aamari, E. and C. Levrard (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics 47*(1), 177–204.

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

Alon, U., N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences 96*(12), 6745–6750.

Boyd, S., S. P. Boyd, and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Boyd, S., L. Xiao, and A. Mutapcic (2003). Subgradient methods. *lecture notes of EE392, Stanford University, Autumn Quarter 2004*, 2004–2005.

Braides, A. et al. (2002). *Gamma-convergence for Beginners*, Volume 22. Clarendon Press.

Caliński, T. and J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods 3*(1), 1–27.

Chao, H. X., R. I. Fakhreddin, H. K. Shimerov, K. M. Kedziora, R. J. Kumar, J. Perez, J. C. Limas, G. D. Grant, J. G. Cook, G. P. Gupta, et al. (2019). Evidence that the human cell cycle is a series of uncoupled, memoryless phases. *Molecular systems biology 15*(3), e8604.

Chernov, N. (2010). *Circular and linear regression fitting circles and lines by least squares*. Taylor & Francis.

Consortium, G. (2020). The gtex consortium atlas of genetic regulatory effects across human tissues. *Science 369*(6509), 1318–1330.

Davies, D. L. and D. W. Bouldin (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence* (2), 224–227.

Doob, J. L. (1953). *Stochastic processes*, Volume 10. Wiley: New York.

Erichson, N. B., P. Zheng, K. Manohar, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin (2020). Sparse principal component analysis via variable projection. *SIAM Journal on Applied Mathematics 80*(2), 977–1002.

Fefferman, C., S. Ivanov, Y. Kurylev, M. Lassas, and H. Narayanan (2018). Fitting a putative manifold to noisy data. In *Conference On Learning Theory*, pp. 688–720. PMLR.

Genovese, C. R., M. Perone Pacifico, V. Isabella, and L. Wasserman (2012). Minimax manifold estimation.

Huber, P. J. (2004). *Robust statistics*, Volume 523. John Wiley & Sons.

Huber, P. J. et al. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 221–233. University of California Press.

Hyvärinen, A. and E. Oja (2000). Independent component analysis: algorithms and applications. *Neural networks 13*(4-5), 411–430.

Johnson, E. M., W. Kath, and M. Mani (2022). Embedr: Distinguishing signal from noise in single-cell omics data. *Patterns 3*(3), 100443.

Jolliffe, I. T. (1995). Rotation of principal components: choice of normalization constraints. *Journal of Applied Statistics 22*(1), 29–35.

Jolliffe, I. T. and J. Cadima (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374*(2065), 1–16.

Journée, M., Y. Nesterov, P. Richtárik, and R. Sepulchre (2010). Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research 11*(2).

Kraemer, G., M. Reichstein, and M. D. Mahecha (2018). dimred and coranking—unifying dimensionality reduction in r. *R Journal 10*(1), 342–358.

Kruskal, J. B. (1978). *Multidimensional scaling*. Number 11. Sage.

Landgraf, A. J. and Y. Lee (2020). Dimensionality reduction for binary data through the projection of natural parameters. *Journal of Multivariate Analysis 180*, 104668.

Lee, J. A. and M. Verleysen (2009). Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing 72*(7-9), 1431–1443.

Li, D., Y. Lu, E. Chevalier, and D. B. Dunson (2020). Density estimation and modeling on symmetric spaces. *arXiv preprint arXiv:2009.01983*.

Li, D., M. Mukhopadhyay, and D. B. Dunson (2022). Efficient Manifold Approximation with Spherelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lin, T. and H. Zha (2008). Riemannian manifold learning. *IEEE transactions on pattern analysis and machine intelligence 30*(5), 796–809.

Lueks, W., B. Mokbel, M. Biehl, and B. Hammer (2011). How to evaluate dimensionality reduction?–improving the co-ranking matrix. *arXiv preprint arXiv:1110.3917*.

Luo, H., S. N. MacEachern, and M. Peruggia (2020). Asymptotics of lower dimensional zero-density regions. *arXiv:2006.02568*, 1–28.

Luo, H., A. Patania, J. Kim, and M. Vejdemo-Johansson (2021). Generalized penalty for circular coordinate representation. *Foundations of Data Science*, 1–37.

Luo, H. and J. D. Strait (2022). Nonparametric multi-shape modeling with uncertainty quantification. *arXiv preprint arXiv:2206.09127*.

Maggioni, M., S. Minsker, and N. Strawn (2016). Multiscale dictionary learning: non-asymptotic bounds and robustness. *The Journal of Machine Learning Research 17*(1), 43–93.

McInnes, L., J. Healy, N. Saul, and L. Grossberger (2018). Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software 3*(29), 861.

Mukhopadhyay, M., D. Li, and D. B. Dunson (2020). Estimating densities with non-linear support by using fisher–gaussian kernels. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82*(5), 1249–1271.

Nigmetov, A. and D. Morozov (2022). Topological optimization with big steps. *arXiv preprint arXiv:2203.16748*.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572.

Puchkin, N. and V. G. Spokoiny (2022). Structure-adaptive manifold estimation. *J. Mach. Learn. Res. 23*, 40–1.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20*, 53–65.

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

Schafer, K. (1998). The cell cycle: a review. *Veterinary pathology 35*(6), 461–478.

Schölkopf, B., A. Smola, and K.-R. Müller (1997). Kernel principal component analysis. In *International conference on artificial neural networks*, pp. 583–588. Springer.

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

Schubert, E. and M. Gertz (2017). Intrinsic t-stochastic neighbor embedding for visualization and outlier detection. In *International Conference on Similarity Search and Applications*, pp. 188–203. Springer.

Sigmund, O. (2001). A 99 line topology optimization code written in matlab. *Structural and multidisciplinary optimization 21*, 120–127.

Solomon, E., A. Wagner, and P. Bendich (2021). From geometry to topology: Inverse theorems for distributed persistence. *arXiv preprint arXiv:2101.12288*.

Stallaert, W., K. M. Kedziora, C. D. Taylor, T. M. Zikry, J. S. Ranek, H. K. Sobon, S. R. Taylor, C. L. Young, J. G. Cook, and J. E. Purvis (2022). The structure of the human cell cycle. *Cell systems 13*(3), 230–240.

Stallaert, W., S. R. Taylor, K. M. Kedziora, C. D. Taylor, H. K. Sobon, C. L. Young, J. C. Limas, J. Varblow Holloway, M. S. Johnson, J. G. Cook, et al. (2022). The molecular architecture of cell cycle arrest. *Molecular Systems Biology 18*(9), e11087.

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

Titsias, M. and N. D. Lawrence (2010). Bayesian gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 844–851. JMLR Workshop and Conference Proceedings.

Townes, F. W., S. C. Hicks, M. J. Aryee, and R. A. Irizarry (2019). Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology 20*(1), 1–16.

Van der Maaten, L. and G. Hinton (2008). Visualizing data using t-sne. *Journal of machine learning research 9*(11).

Wattenberg, M., F. Viégas, and I. Johnson (2016). How to use t-sne effectively. *Distill*.

Wilkinson, L. and H. Luo (2020). A Distance-preserving Matrix Sketch. *arXiv:2009.03979*.

Zhang, L., T. Dunn, J. Marshall, B. Olveczky, and S. Linderman (2021). Animal pose estimation from video data with a hierarchical von mises-fisher-gaussian model. In *International Conference on Artificial Intelligence and Statistics*, pp. 2800–2808. PMLR.

Zhou, Y. and T. O. Sharpee (2018). Using global t-sne to preserve inter-cluster data structure. *bioRxiv*, 331611.

A Orthogonal Loop Examples

Figure A.1 shows orthogonal circles parallel to xz and xy planes respectively in $\mathbb{R}^3$. The projection of these two circles to the principal axes given by PCA is shown in Figure A.2, where only one circular structure is retained in the reduced dataset while the other circular structure is completely destroyed

In Figure A.3, we have the same but each coordinates is perturbed by a Gaussian noise with mean zero and different noise variances. As the noise variance increases, we observe that the topological structure of this example of two orthogonal loops becomes less and less obvious. We can see that SRCA is consistently achieving the lowest matched MSE defined in Section 4.1, while both SPCA and SRCA preserves the topological structure relatively well. It becomes evident that that SPCA and SRCA method better respect the topology of the original dataset under the same d'. PCA does not retain the circular structure, but SRCA puts both circles onto a larger 1-sphere congruent to $\mathbb{S}^1$.

In Table A.1, we provide MSE for more settings of noise variances to show the MSE from each different DR methods. It can be observed that PCA becomes worse quickly in terms of MSE.

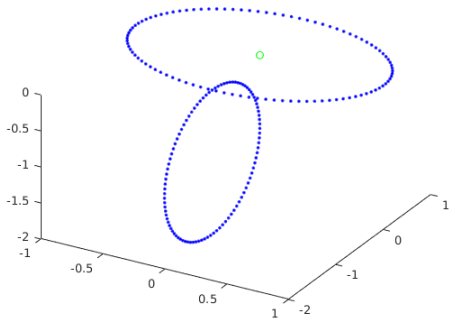

Figure A.1: An example where PCA fails, adapted from Luo et al. (2021). Note that the two orthogonal circles only intersect at one point. The DR results are summarized in Figure A.2

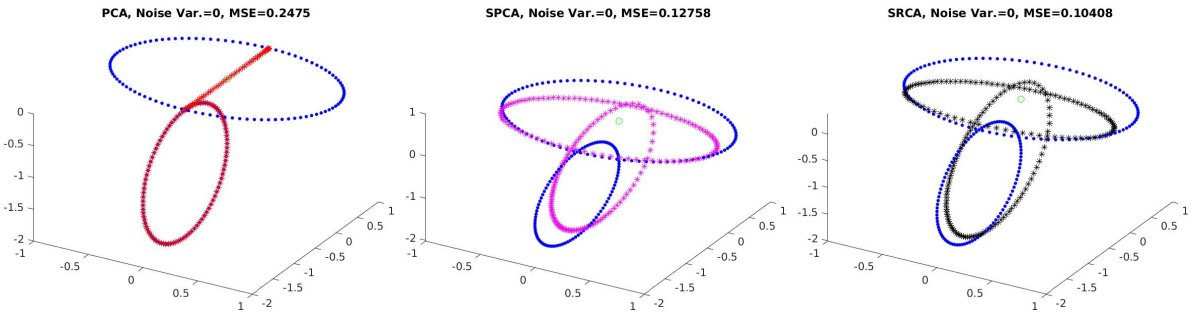

Figure A.2: The dimension-reduced dataset from the example shown in Figure A.1. The green dot represents the origin $(0, 0, 0)$. The original dataset is represented by blue points. On the left panel, the PCA dimension-reduced dataset is represented by red stars. On the middle and right panels the dimension-reduced dataset processed by SPCA and SRCA, is represented by magenta and black stars, respectively.

Noise Var.	0	0.01	0.05	0.10	0.20	0.40	1.00
PCA	0.24750	0.24889	0.25634	0.26990	0.31126	0.45045	1.2653
SRCA	0.10408	0.10421	0.10623	0.11237	0.13668	0.21834	0.64711
SPCA	0.12758	0.12764	0.12925	0.13448	0.15585	0.22861	0.65268

Table A.1: MSE for different DR methods performed on the same orthogonal loop dataset but with different noise variances in the Gaussian perturbation.

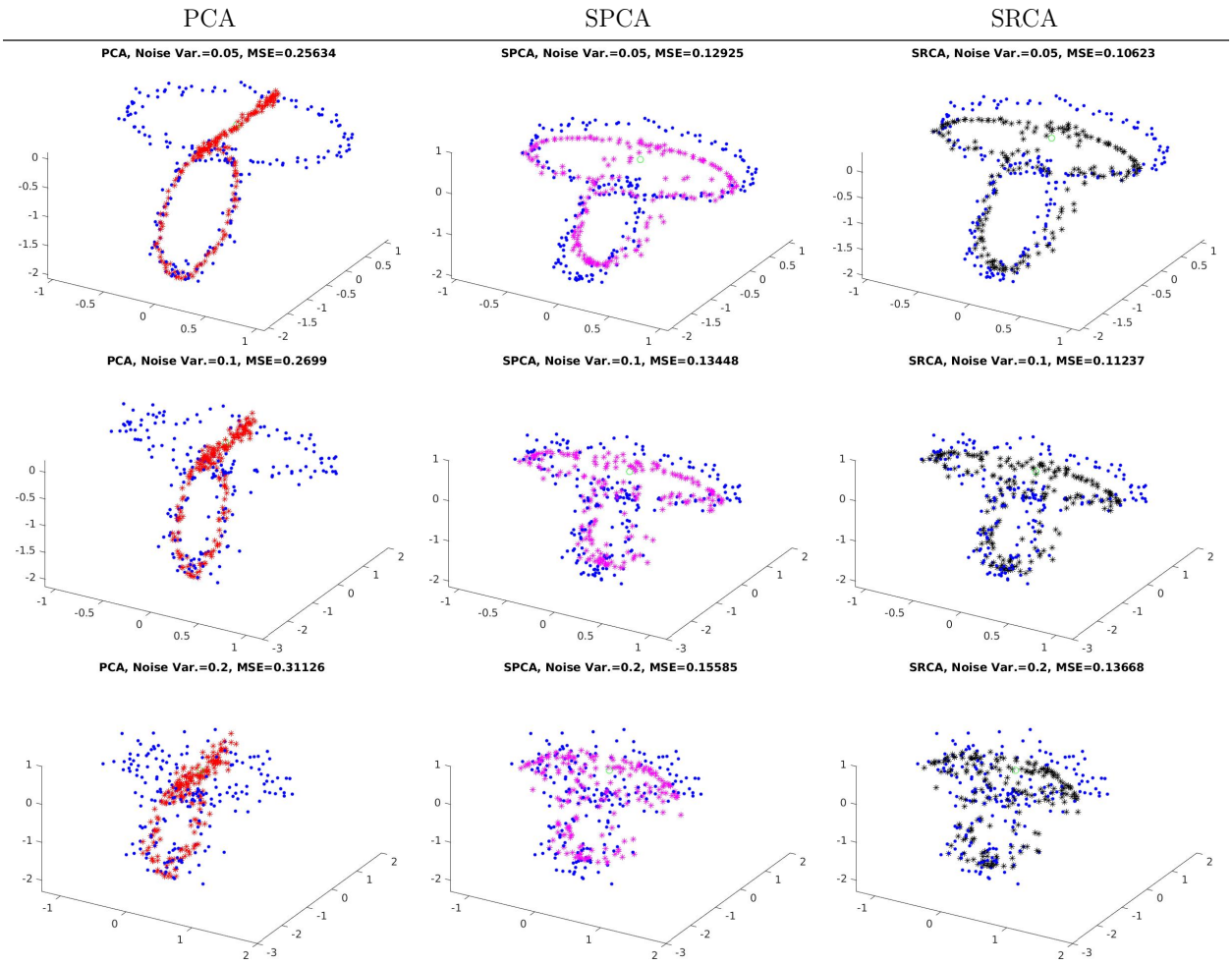

Figure A.3: On the left columns we show the original dataset in blue points, and the PCA dimension-reduced dataset in red stars. On the middle column we show the original dataset in blue points, and the SPCA dimension-reduced dataset in magenta stars. On the right column we show the original dataset in blue points, and the SRCA dimension-reduced dataset in black stars.

B Other Synthetic Examples

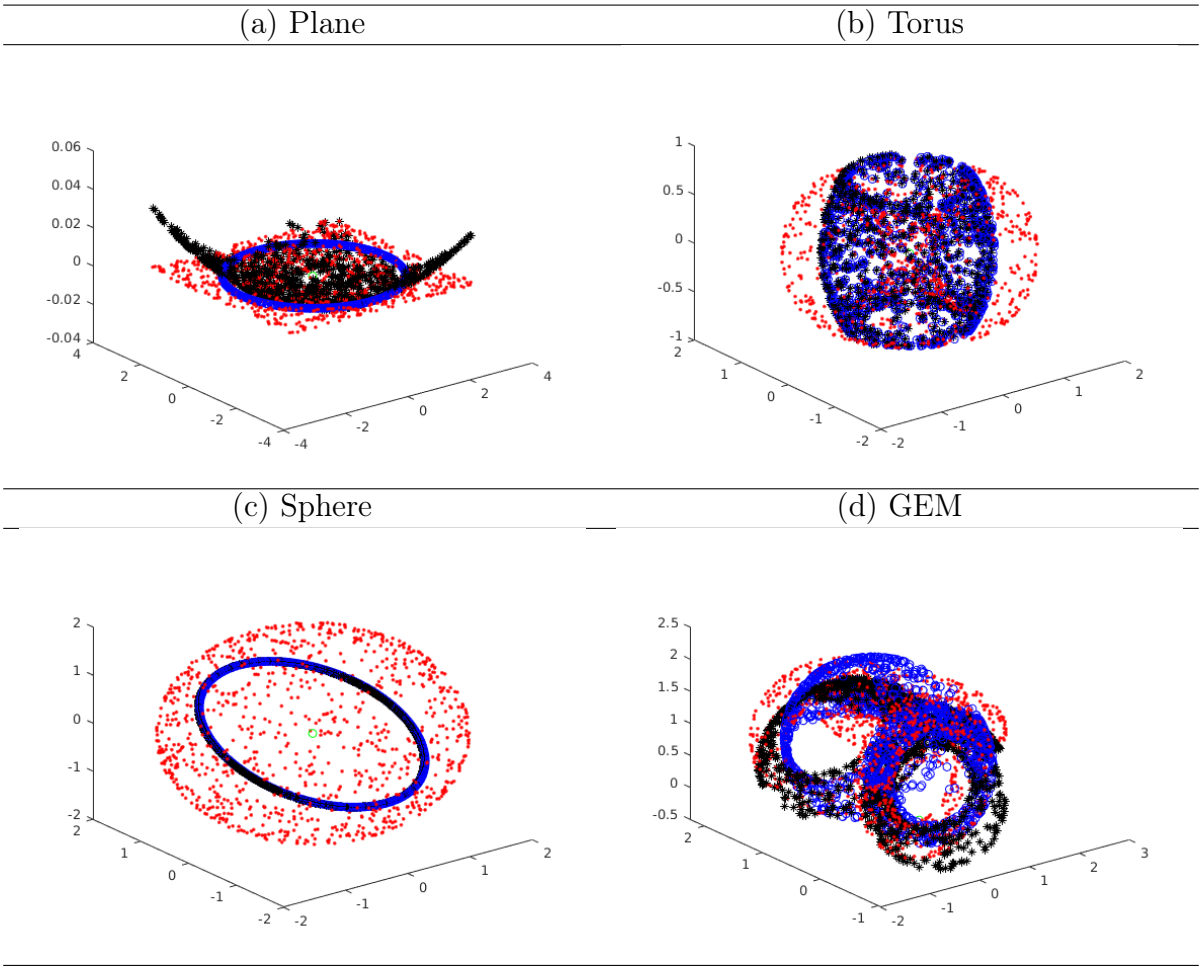

Figure B.1: The datasets for the basic example. In these examples the datasets are observed in $\mathbb{R}^3$ ($d = 3$) and we want to reduce the dataset by one dimension ($d' = 2$). In each of the above panels, the red solid dots are the original datasets sampled from a uniform distribution lying on (a) plane (b) torus (c) sphere (d) triple torus without interior intersection (GEM). The blue circles are the points in reduced dataset obtained by SPCA. The black stars are the points in reduced dataset obtained by SRCA. We do not display the result of PCA in these figures.

In this appendix, we consider several basic examples we use to illustrate the difference between PCA, SRCA, and SPCA. Below, we provide the quantitative measures for each of these examples and the coranking summaries. In (a) plane, we uniformly sample points

$(x_1, x_2, 0)$ from $[-3, 3] \times [-3, 3] \times \{0\}$. In (b) torus, we take the parameterization

$$((R_1 + R_2 \cos\theta)\cos\phi, (R_1 + R_2 \cos\theta)\sin\phi, R_2 \sin\theta)$$

where $R_1 = 1/2$ and $R_2 = 1/3$. The parameters (θ, ϕ) are uniformly sampled from $[0, 2\pi) \times [0, 2\pi)$. In (c) sphere, we take the canonical parameterization and uniform sampling on the parameter space $[0, 2\pi) \times [0, 2\pi)$, which is equivalent to von-Mises Fisher distribution with concentration $\kappa = 0$. In (d) triple torus, we independently sampled 3 batches of points with equal sample sizes. Then we multiply each of these 3 batches with the following rotation matrices and add a translation vector:

$$R_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\frac{\pi}{2} & -\sin\frac{\pi}{2} \\ 0 & \sin\frac{\pi}{2} & \cos\frac{\pi}{2} \end{pmatrix}, \tau_1 = \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix}.$$

$$R_2 = \begin{pmatrix} \cos\frac{\pi}{4} & 0 & \sin\frac{\pi}{4} \\ 0 & 1 & 0 \\ -\sin\frac{\pi}{4} & 0 & \cos\frac{\pi}{4} \end{pmatrix}, \tau_2 = \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}.$$

$$R_3 = \begin{pmatrix} \cos 0 & -\sin 0 & 0 \\ \sin 0 & \cos 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tau_3 = \begin{pmatrix} 3 \\ 3 \\ 3 \end{pmatrix}.$$

The resulting dataset contains three tori, and it is not difficult to verify that these three tori do not have interior intersections. We scale the dataset by subtracting $(1, 1, 1)^T$ from each point and multiply $1/2$ entry-wisely.

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.9999998	0.9998008	1.0000000	0.9983610	0.9774094	0.9906862
AUC	0.9995782	0.9898617	0.9998370	0.9604653	0.8670743	0.9094597
WAUC	0.9990011	0.9913487	0.9991016	0.9725947	0.8460439	0.8320849

Table B.1: Performance scores and the coranking of plane.

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.8222110	0.8208433	0.9777033	0.9751233	0.8523848	0.9072153
AUC	0.6217176	0.6194763	0.8317488	0.8251250	0.6344541	0.7257927
WAUC	0.6025524	0.6033849	0.6104916	0.6092178	0.7430941	0.6707993

Table B.2: Performance scores and the coranking of torus.

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	1.0000000	1.0000000	0.8026803	0.7262926	0.6762080	0.6810310
AUC	0.9999992	1.0000000	0.5585561	0.4876606	0.5171233	0.5696564
WAUC	0.9999998	1.0000000	0.4953096	0.4792300	0.7312668	0.7270227

Table B.3: Performance scores and the coranking of sphere.

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.9986899	0.7407685	0.9987456	0.9963411	0.5697951	0.9438497
AUC	0.9509051	0.5618742	0.9518870	0.9205247	0.4646542	0.8042776
WAUC	0.7067388	0.5204686	0.7057702	0.6973970	0.7077429	0.6735852

Table B.4: Performance scores and the coranking of GEM.

From the performance evaluation table above, we can see that:

1. In the plane example, all linear and non-linear DR methods performs well in terms of performance evaluation measures and coranking summaries.

2. In the torus example, neither SPCA nor SRCA perform as good as other DR methods in terms of scores. However, in the visualization of the reduced dataset, we can see that both SPCA and SRCA preserve the structure of torus pretty well.

3. In the sphere example, both SPCA and SRCA give almost identical results, outperforming both the simplest PCA and the more sophisticated non-linear DR methods like tSNE and UMAP. In the visualization of the reduced dataset, we can see that both SPCA and SRCA reduce the data points on the sphere ($d = 3$) to points on a geodesic circle ($d' = 2$). This preserves the spherical nature of the dataset and yield a better result.

4. In the GEM example, the SRCA is the best in terms of almost all performance scores. A visual verification also reveals that the resulting reduced dataset preserves all three holes in the tori.

C SRCA Algorithms

In this section, we present the algorithm that solves our optimization problem (2.2) and (2.3). Algorithm 1 delineates the binary search strategy which exhausts 2^d possible subsets of $\{1, \cdots, d\}$ to find an optimal subspace. Algorithm 2 delineates the l_1-relaxation strategy which formulates the original problem (2.2) into an optimization problem (2.3) to find an optimal subspace.

Detailed explanation of algorithms in this section is as follows.

1. (Step 1: Get the empirical mean for $\mathcal{X}$.) Estimate the empirical mean $\bar{\mathcal{X}}$ for the dataset $\mathcal{X}$ in $\mathbb{R}^d$, and subtract the mean $\bar{\mathcal{X}}$ to make sure that the assumption of PCA is satisfied.

$$z_i = x_i - \frac{1}{n}\sum_{i=1}^{n} x_i = x_i - \bar{x}$$

2. (Step 2: Conduct the rotation.) We choose a rotation method to construct rotation matrix R based on the datset $\mathcal{X}$. Then we rotate the dataset $\mathcal{X}$ to standard position $(\mathcal{X} - \bar{\mathcal{X}})R$. Here we use a chosen rotation matrix R (PCA, ICA or other kinds of optimal rotation) to rotate the sphere so that all its axes are parallel to the coordinate axes.

   For PCA rotation, let the covariance matrix $\text{cov}(z_i) = R\Lambda R^T$ where Λ is diagonal and the rotation matrix R is orthogonal.

$$y_i = Rz_i$$

3. (Step 3: Binary search for the best $d' + 1$ axes.) In this step we perform dimension reduction. Now that we can assume that the axes of the sphere (or ellipsoid) are parallel to the coordinate axes. We solve the binary optimization problem (2.2) with $W = I$ (or other W if extremely skewed dataset is observed) to choose the optimal directions to retain. In this step we would find the optimal v_{opt} and hence the optimal index set $\mathcal{I}_{\text{opt}}$.

   In this optimization problem, as we stated in (2.2) above, we conduct dimension reduction by minimizing the loss function based on the point-to-ellipsoid distance to the estimated sphere $S_{\mathcal{I}}$.

   The optimization problem is spelled out as (2.4).

4. (Step 4: Eliminate the un-chosen dimensions.) This simply sets drop the un-selected dimension not in $\mathcal{I}_{\text{opt}}$, equivalently, we find the v_{opt} in the notation of problems (2.2) and (2.3).

   (a) In the standard form (2.2), we may let $\eta = v_{\text{opt}}$ be the binary vector such that $\|v_{\text{opt}}\|_{l_0} = d' + 1$.

   (b) In the l_1-relaxed form (2.3), the v_{opt} would have l_1-norm less or equal than $d' + 1$ but not binary entries. We construct a binary vector η such that only the first leading $d' + 1$ entries with the largest absolute values in v_{opt} are 1; and the rest entries are 0.

5. (Step 5: Re-estimate the center and radius.) Only in the problem (2.3), we re-estimate the center c_{opt} and radius r_{opt} using the same loss function but a fixed v_{opt}. In standard problem (2.2), we use the center c_{opt} and radius r_{opt} from step 3.

6. (Step 6: Project and rotate the sphere back into full space.) After we choose the dimension and axes, we project the dataset onto a sphere with the center c_{opt} and radius r_{opt} and add back the empirical mean $\bar{\mathcal{X}}$. More specifically, we project the datapoints x_i to the sphere $S(c, r)$ as in (2.5) and then we simply we rotate the resulting dimension-reduced dataset back, using the inverse of the same rotation matrix R.

Algorithm 2: SRCA dimension reduction algorithm with l_1 relaxation

Data: X (data matrix consisting of n samples in $\mathbb{R}^d$)

Input: d' (the dimension of the sphere), W (the covariance weight matrix, by default $W = I_d$), λ (optional, the sparse penalty parameter), rotationMethod (the method we use to construct the rotation matrix).

Result: $\hat{c}$ (The estimated center of $S_{\mathcal{I}}$ in $\mathbb{R}^d$), $\hat{r}$ (The estimated radius of $S_{\mathcal{I}}$), $\mathcal{I}_{opt}$ (The optimal index subset)

`GetRotation` (X,rotationMethod) obtains a $d \times d$ rotation matrix based on the data matrix X. The rotationMethod option specifies what method we use to construct the rotation matrix, by default, we use PCA to obtain the rotation matrix.

`ProjectToSphere` (X,c,r,k) is a projection that projects the point set X onto an l_k-sphere of center c and radius r via $X \mapsto c + \frac{X-c}{\|X-c\|_{l_k}} \cdot r$.

begin

- Standardize the dataset by subtracting its empirical mean $X = X - \bar{X}$
- Construct a rotate matrix $R =$ `GetRotation` (X,rotationMethod)
- $X_{rotated} = X * R$
- Solve the optimization problem (2.3) with respect to c, r and v
- Denote the solution as c_0, r_0, v_{opt}
- Construct $\mathcal{I}$ that contains the largest/leading $d' + 1$ coordinates of v_{opt}.
- Solve the optimization problem (2.2) with respect to c, r with a fixed $\mathcal{I}$
- Denote the solution as c_{opt}, r_{opt}
- $\hat{c} = c_{opt} \cdot \eta * R^{-1} + \bar{X}$
- $\hat{r} = r_{opt}$
- $X_{rotated}(:, \mathcal{I}) \leftarrow 0$
- $X_{rotated} \leftarrow$ `ProjectToSphere` (X,$\hat{c}$,$\hat{r}$,k)
- $X_{output} \leftarrow X_{rotated} * R^{-1} + \bar{X}$

end

D Dataset Selection

We select datasets for high- and low-dimensional scenarios, covering both $n \geq d$ and $n < d$ as detailed below. Thanks to the binary search scheme we designed in our SRCA algorithm, SRCA can be parallelized when an even larger n (e.g., $n = 100,000$) presents. Neither PCA nor non-linear methods we consider below can be generalized to a quite large n in an obvious way.

The tSNE (Van der Maaten and Hinton, 2008) has known problems of not distance-preserving and creating false clusters in the dimension-reduced dataset (Schubert and Gertz, 2017). The UMAP (McInnes et al., 2018) has known problems of being sensitive to outliers and require the user to have a good understanding of their distance metrics to interpret the

dimension-reduced dataset. To highlight the advantage of our method with a geometric loss function, we also choose labeled datasets to study the structure-preserving properties and datasets that require careful normalization.

We use several public datasets for our numerical experiments, but consider only datasets with continuous variables as its features, as we recognized that dimension reduction for datasets with discrete (categorical and integer) or mixed type variables as their attributes is a different problem (Schölkopf et al., 1997, 1998).

Before the analysis of our experiments, we shall briefly introduce our datasets:

- Source
  - UCI repository (https://archive.ics.uci.edu/ml): Banknote, Climate, Concrete, Ecoli[1], Leaf, PowerPlant, UserKnowledge.
  - Microarray: Alon (Alon et al., 1999).
  - GTEx (https://gtexportal.org/home/).
- Sample size

  We understand that a valid dimension reduction method should have reasonable performance regardless of the size of the underlying datasets. We choose a wide range of datasets with sample sizes varying from 100 to 10,000. In the table below, we show the code for each dataset with its sample size and dimension ($n \times d$).

$n > d$	$n \leq d$
Banknote (1372×4), UserKnowledge (403×5),	Kidney Medulla (4×500),
Ecoli (336×7), Concrete (1030×8),	Fallopian Tube (9×500),
Climate (540×18), Leaf (340×14),	Cervix Endocervix (10×500),.
PowerPlant (9568×5),	Alon (62×2000), .

- Dimensionality

  It is of central importance to recognize that the relation between the sample size n and the original dimension d would affect dimension reduction methods. In fact, the sparse PCA (Erichson et al., 2020) were developed to take the sparsity (i.e., $n < d$) of the dataset into consideration. SRCA has a natural sparsity penalty parameter in the loss function $\mathscr{L}$ we designed. To see the performance of different methods on dense ($n > d$) and sparse ($n \leq d$) data, we also include both kinds of datasets in the above

[1] Three groups among them with sample size smaller than 5 are removed for visual convenience. The five groups left are cytoplasmic proteins (cp), inner membrane proteins without a signal sequence (im), inner brane proteins with an uncleavable signal sequence (imU), other outer membrane proteins (om) and periplasmic proteins (pp).

selection.

- Normalization

  When the attributes or features of the dataset have high mutual correlation or relationships with one another (e.g., total expenditure cannot exceed total income of an individual), normalization would introduce problems like distortion of correlation and violation of relationships. When the attributes or features of the dataset are uncorrelated or independent, normalization would convert all features to (relatively) the same scale. We include datasets that require normalization and those that do not require normalization.

  - Not normalized: Banknote, Ecoli, PowerPlant, UserKnowledge, Climate.
  - Normalized: Alon, Concrete, Leaf, GTEx.

Finally, we shall point out that we have not covered any dataset with a high d and large n. Our exploratory experiments found that the performance of existing dimension methods, including modern methods, has suffered from a very high computational cost. It is a separate problem to study how to perform dimension reduction on a large high-dimensional dataset.

Decomposition-based methods like PCA, multi-dimensional scaling (MDS) and truncated singular value decomposition (SVD) become exceedingly slow for a dataset with large n and d. Manifold learning based methods like tSNE, UMAP and IsoMap (Tenenbaum et al., 2000) have some variation in computational time due to their stochastic nature, but they are all rather slow. Therefore, we would leave this type of dataset as a separate problem that we do not experiment in the current paper.

E Coranking Performance Comparison for Section 4.3

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.987	0.925	0.988	0.833	0.640	0.635
AUC	0.869	0.774	0.860	0.598	0.469	0.459
WAUC	0.644	0.582	0.626	0.503	0.694	0.600

Table E.1: Coranking performance scores of Banknote

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.270	0.270	0.262	0.464	0.215	0.124
AUC	0.152	0.152	0.148	0.225	0.147	0.102
WAUC	0.134	0.132	0.0864	0.105	0.141	0.118

Table E.2: Coranking performance scores of Ecoli

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.987	0.815	0.987	0.928	0.620	0.847
AUC	0.886	0.605	0.886	0.731	0.446	0.651
WAUC	0.485	0.416	0.485	0.447	0.611	0.528

Table E.3: Coranking performance scores of PowerPlant

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.219	0.211	0.219	0.183	0.134	0.180
AUC	0.0780	0.0853	0.0785	0.0700	0.0707	0.0635
WAUC	0.0935	0.0952	0.0948	0.0762	0.106	0.100

Table E.4: Coranking performance scores of Leaf

	SRCA	SPCA	PCA	LLE	tSNE	UMAP
CC	0.730	0.798	0.693	0.585	0.483	0.592
AUC	0.416	0.456	0.379	0.338	0.231	0.384
WAUC	0.256	0.3233	0.251	0.212	0.214	0.328

Table E.5: Coranking performance scores of Alon.

F Sparse Penalty

λ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0
CC	0.704	0.727	0.730	0.730	0.730	0.730
AUC	0.408	0.416	0.416	0.416	0.416	0.416
WAUC	0.261	0.256	0.256	0.256	0.256	0.256

Table F.1: Coranking performance scores for different λ's on the Alon dataset, $d' = 2$.

Here, we provide a new version of SRCA with sparse penalty, which only involves an additional penalty term in the loss function we designed. Recall that the objective loss

function with a weighted matrix W in our method is

$$d(x_i, S_{\mathcal{I}}(c,r))^2 = (x_i - c)^T W (x_i - c) + r^2 - 2r\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x_i - c)}$$

which involves only the point-to-sphere distance from x_i to the estimated sphere surface $S_{\mathcal{I}}$ (based on dataset $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$). One problem we wish to address when there exists sparsity in the dataset in the procedure of dimension reduction, is that we want the sparsity being preserved.

To be more precise, if x_i's have most coordinates zeros except for a few, then we want the reduced dataset $\hat{x}_i$ to have a similar property. This can be achieved by penalizing $\|I_{\mathcal{I}}(x_i - c)\|_1$ in the optimization problem, which encourages the estimated sphere so that the data are in the affine subspace centered at c while parallel to the coordinate planes. This is different from the l_1 relaxation we propose above. The l_1 relaxation we proposed above is an approximation to the constraints we imposed on the binary optimization problem. Here, we directly penalize the reduced coordinates. The corresponding optimization problem is:

$$\min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \sum_{i=1}^{n} \Big((x_i - c)^T W (x_i - c) + r^2 \\ -2r\sqrt{(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c)} \Big) + \lambda \|I_{\mathcal{I}}(x_i - c)\|_1, \tag{F.1}$$

$$\text{s.t. } \|v\|_{l_k} \leq d' + 1, \lambda > 0, \tag{F.2}$$

with a tuning parameter $\lambda > 0$. $\|v\|_{l_k} \leq d' + 1$ can be $\|v\|_{l_k} = d' + 1$. This feature of l_1 constraint allows us to perform dimension reduction in a high-dimensional input space with SRCA. We want to consider the penalty parameter λ that controls the retained dimension d' when l_1 approximation is in place as defined in (F.1). When a strict binary search like l_0 optimization in (2.4) is used, the penalty is usually not needed. However, the caution we shall take here is that the selection of sparsity penalty parameter λ should roughly be at the same magnitude as the loss function in order to function properly.

The tuning parameter $\lambda > 0$ is part of the objective function, instead of the constraints. In some applications, the penalty term can also be replaced with $\lambda \sum_{i=1}^{n} \|I_{\mathcal{I}} x_i\|_1$. In experiments, we found that SRCA is not sensitive to λ. In very high-dimensional datasets (e.g., Alon (see Supplement D), GTEx), the choice of this parameter also affects the convergence speed in the execution of the optimization algorithm, a larger penalty parameter forces the numerical algorithm to converge slightly faster. Alternatively, we can also treat the choice

of this parameter as a apriori tuning parameter of the loss function, whose values can be selected for different datasets using cross-validation.

G Spherical Estimation

The essence of SPCA (Li et al., 2022) can be summarized as a two-step procedure:

1. First, we utilize the principal component analysis (PCA) to find a subspace $V \subset \mathbb{R}^{d'+1}$ of retained dimension based on $\mathcal{X}$ and project $\mathcal{X}$ to $\hat{\mathcal{X}}$ in V.

2. Second, we perform a circular (or d'-dimensional spherical) regression[2] with the projected image $\hat{\mathcal{X}}$ onto V.

By selecting the principal components given by PCA, we find a subspace V and determine the dimension of the S. By fitting a circular regression on a d'-dimensional sphere with the projected dataset $\hat{\mathcal{X}}$, we determine the center c and radius r of the spherical support.

Suppose the assumed sphere $S_V(c,r)$ is $d^2(x,c) = r^2$, whose dimensionality is determined by the PCA estimated linear subspace V. Two typical loss functions for the estimation of c, r are:

$$\mathscr{L}(V,c,r) = \sum_{i=1}^{n} d^2(x_i, S_V(c,r))$$

where d^2 can be chosen as geometric or algebraic loss

$$\text{geometric loss} = \sum_{i=1}^{n} \left(\sqrt{(x_i - c)^T (x_i - c)} - r \right)^2$$

$$\text{algebraic loss} = \sum_{i=1}^{n} \left((x_i - c)^T (x_i - c) - r^2 \right)^2$$

Following Li et al. (2022), we first assume that V is determined (through PCA) and attempt to estimate the center c and radius r via a two-step gradient descent with both geometric and algebraic loss functions. Through the procedure of taking derivation, we observe and explain why an analytic solution for c and r is impossible in this SPCA setup in the end and how SRCA handles this problem.

[2]Unfortunately, although methods in circular regression could be extended to spheres of intrinsic dimensions greater than 1, the term "circular regression" instead of "spherical regression" is adopted.

G.1 Geometric Loss

Let us calculate the geometric loss first, the algebraic loss is calculated at the end. This function is a quadratic polynomial of radius parameter $r > 0$, $\mathscr{L}$ has a unique global (conditional) minimum in r if $\hat{r} > 0$. When the c is assumed fixed.

We calculate its gradient

$$\begin{aligned}\frac{\partial \mathscr{L}(c,r)}{\partial r} &= \sum_{i=1}^{n} \frac{\partial}{\partial r} d^2(x_i, S(c,r)) \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial r} \left(\sqrt{(x_i - c)^T (x_i - c)} - r \right)^2 \\ &= \sum_{i=1}^{n} -2 \left(\sqrt{(x_i - c)^T (x_i - c)} - r \right)\end{aligned}$$

Setting this equation to zero, we have

$$\hat{r} = \frac{1}{n} \sum_{j=1}^{n} \sqrt{(x_j - c)^T (x_j - c)} \geq 0.$$

Plug this back into the $\mathscr{L}(c,r)$ we have

$$\begin{aligned}\mathscr{L}(c,\hat{r}) &= \sum_{i=1}^{n} d^2(x_i, S(c,\hat{r})) \\ &= \sum_{i=1}^{n} \left(\sqrt{(x_i - c)^T (x_i - c)} - \hat{r} \right)^2 \\ &= \sum_{i=1}^{n} \left(\sqrt{(x_i - c)^T (x_i - c)} - \frac{1}{n} \sum_{j=1}^{n} \sqrt{(x_j - c)^T (x_j - c)} \right)^2\end{aligned}$$

Although this cannot be simplified further (due to the fact that it is fourth power in c), we can still attempt to take its gradient

$$
\begin{aligned}
\frac{\partial \mathscr{L}(c,\hat{r})}{\partial c} &= \sum_{i=1}^{n} \frac{\partial}{\partial c} \left(\sqrt{(x_i - c)^T (x_i - c)} - \frac{1}{n} \sum_{j=1}^{n} \sqrt{(x_j - c)^T (x_j - c)} \right)^2, \\
&\text{where } \hat{r} = \frac{1}{n} \sum_{j=1}^{n} \sqrt{(x_j - c)^T (x_j - c)} \\
&= \sum_{i=1}^{n} 2 \left(\sqrt{(x_i - c)^T (x_i - c)} - \hat{r} \right) \cdot \\
&\left(\frac{\partial}{\partial c} \sqrt{(x_i - c)^T (x_i - c)} - \frac{1}{n} \sum_{j=1}^{n} \frac{\partial}{\partial c} \sqrt{(x_j - c)^T (x_j - c)} \right)
\end{aligned}
$$

The equation $\frac{\partial \mathscr{L}(c,\hat{r})}{\partial c} = 0$ would not have an analytic solution in general. However, with an appropriate gradient-based optimization method, for example, Gauss-Newton method with Levenberg-Marquardt correction (Chernov, 2010), the sequence of estimates of c, r can be proven to converge to global minimum under the regularity condition. It is also not hard to observe why the insertion of W into the $\sqrt{(x_i - c)^T W (x_i - c)}$ makes the gradient calculation even more intractable for the geometric loss function.

G.2 Algebraic Loss

However, analytic solutions for a sphere estimation can be derived for algebraic loss. It can also generalize to ellipsoid (i.e., an algebraic loss can be solved analytically for the ellipsoid $x^T W x = r$)

$$
\begin{aligned}
\frac{\partial \mathscr{L}(c,r)}{\partial r} &= \sum_{i=1}^{n} \frac{\partial}{\partial r} d^2(x_i, S(c,r)), \text{ algebraically} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial r} \left((x_i - c)^T (x_i - c) - r^2 \right)^2 \\
&= \sum_{i=1}^{n} 2 \left((x_i - c)^T (x_i - c) - r^2 \right) \cdot (-2r)
\end{aligned}
$$

which is a cubic polynomial. $\frac{\partial \mathscr{L}(c,r)}{\partial r} = 0$ is analytically solvable in r, via Cardano-Viete's formula:

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} -2\left(\left(x_i - c\right)^T \left(x_i - c\right) - r^2\right) \cdot 2r \\
0 &= \sum_{i=1}^{n} \left(\left(x_i^T x_i - 2c^T x_i + c^T c\right) - r^2\right) \cdot r \\
0 &= \sum_{i=1}^{n} \left(\left(x_i^T x_i - 2c^T x_i + c^T c\right) r - r^3\right) \\
0 &= -n \cdot r^3 + \left[\sum_{i=1}^{n} \left(x_i^T x_i - 2c^T x_i + c^T c\right)\right] \cdot r.
\end{aligned}
$$

Write it into $x^3 + px + q = 0$ form:

$$
\begin{aligned}
r^3 + \left[-\frac{1}{n}\sum_{i=1}^{n} \left(x_i^T x_i - 2c^T x_i + c^T c\right)\right] \cdot r + 0 = 0 \\
p = -\frac{1}{n}\sum_{i=1}^{n} \left(x_i^T x_i - 2c^T x_i + c^T c\right), q = 0
\end{aligned}
$$

The determinant $4p^3 + 27q^2 < 0$ obviously, the solution is

$$
\begin{aligned}
\hat{r}_k &= 2\sqrt{-\frac{p}{3}} \cdot \cos\left[\frac{1}{3}\arccos\left(\frac{3q}{2p}\sqrt{\frac{-3}{p}}\right) - \frac{2\pi k}{3}\right] \text{ for } k = 0, 1, 2. \\
&= 2\sqrt{\frac{1}{3n}\sum_{i=1}^{n} \left(x_i^T x_i - 2c^T x_i + c^T c\right)} \cdot \cos\left[\frac{1}{3} \cdot \frac{\pi}{2} - \frac{2\pi k}{3}\right]
\end{aligned}
$$

For the gradient with respect to the center c,

$$\begin{aligned}
\frac{\partial \mathscr{L}(c,\hat{r})}{\partial c} &= \sum_{i=1}^{n} \frac{\partial}{\partial c} d^2(x_i, S(c,\hat{r})), \text{ algebraically} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial c} \left((x_i - c)^T (x_i - c) - \hat{r}^2 \right)^2 \\
&= \sum_{i=1}^{n} 2 \left(\frac{\partial}{\partial c} \left[(x_i - c)^T (x_i - c) - \frac{1}{n} \sum_{j=1}^{n} (x_j - c)^T (x_j - c) \right] \right) \\
&= \sum_{i=1}^{n} 2 \left(\frac{\partial}{\partial c} \left[\left(x_i^T x_i - 2c^T x_i + c^T c \right) - \frac{1}{n} \sum_{j=1}^{n} \left(x_j^T x_j - 2c^T x_j + c^T c \right) \right] \right),
\end{aligned}$$

and the equation $\frac{\partial \mathscr{L}(c,\hat{r})}{\partial c} = 0$ solves

$$\hat{c} = \frac{1}{2} \left(\sum_{i=1}^{n} (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j)^T (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j) \right)^{-1} \sum_{i=1}^{n} \left(x_i^T x_i - \frac{1}{n} \sum_{j=1}^{n} x_j^T x_j \right) (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j).$$

Therefore, an algebraic loss would provide us a closed form solution to the estimate of both center c and radius r.

G.3 SPCA and SRCA Solution

Following the thought of the simultaneous estimation of c, r and the dimension of the sphere (or equivalently, the linear subspace $\mathbf{V} \in \mathbb{R}^{d \times (d'+1)}$ where S lives in), we can instead consider

the following geometric loss in one step

$$\begin{aligned}
\mathscr{L}(V,c,r) &= \sum_{i=1}^{n} d^2(x_i, S_V(c,r)) \\
&= \sum_{i=1}^{n} d^2(x_i, c+V) + \sum_{i=1}^{n} d^2(Pr_V(x_i), S_V(c,r)) \\
&= \sum_{i=1}^{n} \|x_i - c - VV^T(x_i - c)\|^2 + \sum_{i=1}^{n} (\|Pr_{c+V}(x_i) - c\| - r)^2 \\
&= \sum_{i=1}^{n} \|x_i - c - VV^T(x_i - c)\|^2 + \sum_{i=1}^{n} (\|c + VV^T(x_i - c) - c\| - r)^2 \\
&= \sum_{i=1}^{n} \|x_i - c - VV^T(x_i - c)\|^2 + \sum_{i=1}^{n} (\|VV^T(x_i - c)\| - r)^2
\end{aligned}$$

The second identity comes from the Pythagorean theorem and $Pr_{c+V}(x_i)$ is the linear projection of x_i to the affine subspace $c+V$.

The first sum corresponds to PCA loss function and the second term is the loss of SRCA if $V = I$. For the SRCA and the SPCA, we minimize the first sum so V is the top eigenvectors of sample covariance matrices and then plug this V to the second sum, and change the geometric sum to the algebraic loss function, since only the latter loss allows a closed form analytic solution. This minimizer from a two-step procedure obtained by SPCA is not necessarily the same as the true minimizer of the above geometric loss $\mathscr{L}(V,c,r)$. However, these two minimizers coincide when all x_i are from a sphere, otherwise SPCA solution is sub-optimal (Li et al., 2022). We adopt the two-step SPCA algorithm only because we cannot derive a closed form minimizer for $L = \mathscr{L}(V,c,r)$. Moreover, this loss function is difficult to generalize to the ellipsoid situation.

H Related Proofs

H.1 Proof of Theorem 1

For each fixed $\|v\|_{l_0} = d' + 1$, it suffices to optimize the following sub-problem of (2.4):

$$\min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \sum_{i=1}^n \left((x_i - c)^T W (x_i - c) + r^2 - 2r\sqrt{(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c)} \right) \tag{H.1}$$

$$= \min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \mathscr{L}_v(c, r; x_1, x_2, \cdots, x_n),$$

$$= \min_{c\in\mathbb{R}^d, r\in\mathbb{R}^+} \sum_{i=1}^n \mathscr{L}_v(c, r; x_i), \tag{H.2}$$

which has gradients with respect to c and r as

$$\begin{aligned}
&\frac{\partial \mathscr{L}_v}{\partial c} = \sum_{i=1}^n \frac{\partial \mathscr{L}_v}{\partial c}(c, r; x_i) \\
&= \sum_{i=1}^n \left(-2(x_i - c)^T W - 2r \cdot \frac{1}{2} \left[(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c) \right]^{-\frac{1}{2}} \right. \\
&\left. \left[-2 (x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} \right] \right) \\
&= \sum_{i=1}^n -2(x_i - c)^T \left(W + r \left[(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c) \right]^{-\frac{1}{2}} \left[\sqrt{W}^T v^T I v \sqrt{W} \right] \right),
\end{aligned}$$

and,

$$\frac{\partial \mathscr{L}_v}{\partial r} = \sum_{i=1}^n \frac{\partial \mathscr{L}_v}{\partial r}(c, r; x_i) = \sum_{i=1}^n \left(2r - 2 \left[(x_i - c)^T \sqrt{W}^T v^T I v \sqrt{W} (x_i - c) \right]^{\frac{1}{2}} \right).$$

Therefore, we can assume that the mild assumptions $\|x_i - c\| \leq R_1, r \leq R_2$ and $|\lambda_{\max}(W)| \leq R_3$. We can compute the bounds of these gradients, using Cauchy-Schwartz

inequality in the first inequality:

$$
\begin{aligned}
&\|\nabla_{(c,r)}\mathscr{L}_v(c,r)\| = \left\|\frac{\partial \mathscr{L}_v}{\partial c}(c,r)\right\| + \left\|\frac{\partial \mathscr{L}_v}{\partial r}(c,r)\right\| \\
&\leq 4\sum_{i=1}^{n}(x_i-c)^T W^T W(x_i-c) \\
&\times \sum_{i=1}^{n}\left\|\left(W + r\left[(x_i-c)^T\sqrt{W}^T v^T I v\sqrt{W}\,(x_i-c)\right]^{-\frac{1}{2}}\left[\sqrt{W}^T v^T I v\sqrt{W}\right]\right)\right\|^2 \\
&+ \sum\nolimits_{i=1}^{n}\left(2r - 2\left[(x_i-c)^T\sqrt{W}^T v^T I v\sqrt{W}\,(x_i-c)\right]^{\frac{1}{2}}\right). \\
&\leq 4\times 2nR_3^2\times nR_1^2\times n\left(R_3 + R_2\frac{\sqrt{R_3^2}}{\sqrt{R_1^2}}\right) + n\left(2R_2 + \sqrt{R_1^2R_3^2}\right) \\
&< \infty
\end{aligned}
$$

For a finite n, we can conclude that $\mathscr{L}$ is Lipschitz with a finite Lipschitz constant as bounded above. Then the gradient descent algorithm would give us a solution to the sub-problem (H.2) with linear convergence from classical results (Boyd et al., 2004). Since for fixed v, each sub-problem converges to the solution, the exhaustive search on v solves the original problem (2.4). In parallel to Boyd et al. (2003), we have proved the Theorem 1.

H.2 Proof of Theorem 2

It is clear that $\mathscr{L}(c_0, r_0, \mathcal{I}_0) = 0$ and $\widehat{\mathcal{I}}_k, \widehat{c}_k, \widehat{r}_k \to \arg\min \mathscr{L}$ by Theorem 1, it suffices to show $(c_0, r_0, \mathcal{I}_0)$ is the unique zero of L. Recall that $\mathscr{L}(c, r, \mathcal{I}) = 0$ if and only if all all x_i's are exactly on sphere $S(c, r, \mathcal{I})$, and that $d'+2$ points uniquely determine a d' dimensional sphere, then the uniqueness follows from the assumption $n > d'+1$.

H.3 Proof of Theorem 3

We consider the closed set Θ_1 on the parameter space defined by $\|x_i - c\| \leq R_1,\ \forall i = 1\cdots, n,\ r \leq R_2$ and $|\lambda_{\max}(W)| \leq R_3$ as we did in the proof of Theorem 4 and 5.

Again, let us assume $\mathcal{I}$ to be fixed index set and the

$$
f_\infty(c,r) = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\left((y_i-c)^T W(y_i-c) - r - 2r\sqrt{(y_i-c)^T\sqrt{W}^T I_{\mathcal{I}}\sqrt{W}(y_i-c)}\right)^2
$$

be the limiting form of our geometric loss function,

$$\mathcal{L}(c, r, \mathcal{I} \mid \mathcal{Y}) = f_j(c, r) = \frac{1}{j} \sum_{i=1}^{j} \left((y_i - c)^T W (y_i - c) - r - 2r \sqrt{(y_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (y_i - c)} \right)^2$$

with the dataset $\mathcal{Y} = \{y_1, \cdots, y_j\}$ and

$$\mathcal{L}(c, r, \mathcal{I} \mid \mathcal{X}) = g_j(c, r) = \frac{1}{j} \sum_{i=1}^{j} \left((x_i - c)^T W (x_i - c) - r - 2r \sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x_i - c)} \right)^2$$

with the dataset $\mathcal{X} = \{x_1, \cdots, x_j\}$. Recall that $x_i = y_i + \epsilon_i$ and $y_i \in S_W(c_0, r_0)$ lying on an ellipsoid with center c_0, radius r_0 and known covariance W.

Therefore, $\arg\min f_{\infty =} \arg\min f_j = (c_0, r_0)$ since if we plug in c_0 and r_0 the $f_\infty(c_0, r_0) = f_j(c_0, r_0) = 0$. By the definition of f_∞, f_j converges to f_∞ point-wise. In addition, by the compact assumptions, the convergence is also uniform, that is, $\sup_{\theta \in \Theta_1} |f_j(\theta) - f_\infty(\theta)| \to 0$.

The rest of our roadmap of proof is as follows. According to the Remark 1.10 of Braides et al. (2002): if a sequence of functions g_j point-wisely converges to its limit f_∞ uniformly. and f_∞ is lower semi-continous, then the same sequence of functions also converges in a Γ-convergence sense, and its Γ-limit is identical to its point-wise limit $f_\infty = \lim_{j \to \infty} f_j$. Furthermore, as we showed above, the following minimizer exists

$$\theta_* \coloneqq \arg\min_{\theta \in \Theta_1} f_\infty(\theta).$$

Then since the specific form of our geometric loss function (2.1) is coercive, by Theorem 1.21 and Remark 1.22 in Braides et al. (2002), the minimizer sequence $\{\theta_j\} = \{\arg\min_{\theta \in \Theta_1} f_j(\theta)\}$ converges to a minimum point θ_* of f_∞. Note that our assumption (A1) stating that each of $\theta_j \coloneqq \arg\min_{\theta \in \Theta_1} g_j(\theta)$ exist is essential here. Otherwise the sequence will not exist.

It's it clear that f_j uniformly converges to f_∞ on compact set Θ_1. We focus on proving g_j also converges to f_j uniformly, then through a middle-man argument, $\lim_{j \to \infty} g_j = f_\infty$

holds. The difference between two sequences f_j and g_j can be bounded as below:

$$
\begin{aligned}
&|g_j(c,r) - f_j(c,r)| && \text{(H.3)}\\
&= \left| \frac{1}{j}\sum_{i=1}^{j} (y_i - c)^T W (y_i - c) - (x_i - c)^T W (x_i - c) + \right. \\
&\left. 2r\sqrt{(x_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}(x_i - c)} - 2r\sqrt{(y_i - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}(y_i - c)} \right| && \text{(H.4)}\\
&\leq \left(4R_3^2 + R_1 R_3\right) \left| \frac{1}{j}\sum_{i=1}^{j} (\|x_i - c\| + \|y_i - c\| - 2r)(\|x_i - c\| - \|y_i - c\|) \right| && \text{(H.5)}\\
&= \left(4R_3^2 + R_1 R_3\right) \left| \frac{1}{j}\sum_{i=1}^{j} (\|x_i - c\| + \|y_i - c\| - 2r)(\|y_i - c + \epsilon_i\| - \|y_i - c\|) \right| \\
&\leq \left(4R_3^2 + R_1 R_3\right) \frac{1}{j}\sum_{i=1}^{j} |(\|x_i - c\| + \|y_i - c\| - 2r)|(\|\epsilon_i\| + 2\|\epsilon_i\|\|y_i - c\|) && \text{(H.6)}\\
&= \left(4R_3^2 + R_1 R_3\right) \frac{1}{j}\sum_{i=1}^{j} |(\|x_i - c\| + \|y_i - c\| - 2r)|(\|\epsilon_i\| + 2\|y_i - c\|) \cdot \|\epsilon_i\| \\
&\leq \left(4R_3^2 + R_1 R_3\right) \cdot (2R_1 + 2R_2) \cdot (1 + 2R_1) \cdot \frac{1}{j}\sum_{i=1}^{j} \|\epsilon_j\|, && \text{(H.7)}
\end{aligned}
$$

where we need the assumption (A2) stating that x_i, y_i and c, r are all in the closed bounded (hence compact) set $B \times \Theta_1$. We want to show that as $j \to \infty$,

$$\sup_{\theta=(c,r)\in\Theta_1} \|g_j - f_j\| \to 0,$$

to ensure uniform convergence. But this follows from (H.7) and our assumption (A3) stating that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \|\epsilon_i\| = 0$. Such an assumption is common, see, Aamari and Levrard (2019); Fefferman et al. (2018); Maggioni et al. (2016) for instances. In fact, the assumption is even weaker than those in the above references, for example, in Aamari and Levrard (2019) the amplitude of the noise is assume to be $\|\epsilon\| \sim n^{-\frac{\alpha}{d}}$ for $\alpha > 1$. In contrast, we only require $\|\epsilon\| \to 0$, so $\|\epsilon\| \sim n^{-\alpha}$ for any $\alpha > 0$ or even $\|\epsilon\| \sim \frac{1}{\log n}$ is good enough.

H.4 Proof of Theorem 4

References we mainly need for our proof below are the formulation in Huber (2004); Huber et al. (1967) and the technical separation lemma in Doob (1953).

We fix the index set $\mathcal{I}$ in the following discussions, and we assume that the parameters to be estimated can be written as a vector $\theta = (c, r) \in \Theta := [-C, C]^d \times [R_0, R] \subset \mathbb{R}^d \times \mathbb{R}^+$, which lies in a (locally) compact space with a countable base $\Theta' = \left\{[-C, C]^d \cap \mathbb{Q}^d\right\} \times \{[R_0, R] \cap \mathbb{Q}\}$, the inclusion of $r = R_0$ is needed below for compactness. We denote that estimate for θ based on n samples (by minimization of the $\mathscr{L}$) by $T_n = T_n(\mathcal{X})$.

The real-valued ρ function, based on the samples $x_1, \cdots, x_n \in \mathbb{X} = \mathbb{R}^d$ drawn from the common distribution P defined on the probability space $(\mathbb{X}, \mathcal{A}, \nu)$ with Borel algebra $\mathcal{A}$ and Lebesgue measure ν, is

$$\rho(x; \theta) = \left((x - c)^T W (x - c) + r^2 - 2r \sqrt{(x - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x - c)} \right)$$

and the $\psi(x; \theta) = \frac{\partial}{\partial \theta} \rho(x; \theta)$ is again differentiable. We show below that the assumptions in Huber et al. (1967) are satisfied, we define our estimator T_n for parameter $\theta = (c, r)$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \rho(x_i; T_n) - \inf_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \rho(x_i; \theta) \to 0, \text{a.s. } P \qquad \text{when } n \to \infty,$$

corresponding to case A in Huber et al. (1967). Since ρ is differentiable in both x, θ, this minimizer could also be expressed in form of T_n satisfying

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(x_i; T_n) \to 0, \text{a.s. } P \qquad \text{when } n \to \infty,$$

- (A-1) For $\Theta = \mathbb{R}^d \times \mathbb{R}^+$, there exists a countable basis $\Theta' = \left\{[-C, C]^d \cap \mathbb{Q}^d\right\} \times \{[R_0, R] \cap \mathbb{Q}\}$ such that for every open set $U \subset \Theta$ and every closed interval $A \subset \mathbb{R}$, two sets

$$\{x \mid \rho(x; \theta) \in A \in \mathcal{A}, \forall \theta \in U\}$$
$$\{x \mid \rho(x; \theta) \in A \in \mathcal{A}, \forall \theta \in U \cap \Theta'\}$$

  would only differ on the set of zero probability measure P. Since the measure P is fixed, by Lemma 2.1 on page 56 of Doob (1953), for each $\theta \in \Theta$ we can find $\theta' \in \Theta'$ such that

$$P\{\omega \in \mathbb{X} \mid \rho(x(\omega); \theta) \neq \rho(x(\omega); \theta'), x(\omega) \sim P\} = 0.$$

Therefore, denote the map $\tau_P : \theta \mapsto \theta'$ we can redefine our ρ by $\tilde{\rho} \coloneqq \rho \circ \tau_P$ so that it only differs from ρ on a zero measure set of the fixed P. Note that the mapping τ_P depends on the measure P and we assume P is fixed throughout our discussion. This ensures the measurability of $\inf_{\theta' \in U} \tilde{\rho}(x; \theta')$ and the measurability of its limit when an (open) neighbor hood U of θ shrinks to one-point set $\{\theta\}$. For ease of notation, we still use ρ below as assume (A-1) holds.

- (A-2) The function ρ is continuous and differentiable, therefore clearly lower semi-continuous in $\theta = (c, r)$. And this ensures that $\inf_{\theta' \in U} \rho(x; \theta') \to \rho(x; \theta)$.
- (A-3) There exists a measurable function $a(x)$ such that

$$\mathbb{E}_P \left(\rho(x; \theta) - a(x)\right)^- < \infty$$
$$\mathbb{E}_P \left(\rho(x; \theta) - a(x)\right)^+ < \infty$$

and hence $\gamma(\theta) = \mathbb{E}\left(\rho(x, \theta) - a(x)\right)$ is well-defined for all $\theta \in \Theta$. For our purpose, we choose $\theta_1 = (c_1, r_1)$ for some $\|c_1\| < \infty$ and $r_1 < \infty$. We define a function on $\mathbb{X}$

$$\begin{aligned}
a(x) = a_{\theta_1}(x) &= \left((x - c_1)^T W (x - c_1) + r_1^2 - 2 \cdot r_1 \sqrt{(x - c_1)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x - c_1)}\right) \\
\rho(x; \theta) - a(x) &= \left((x - c)^T W (x - c) - (x - c_1)^T W (x - c_1)\right) + \left(r^2 - r_1^2\right) \\
&\quad - 2r \sqrt{(x - c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x - c)} \\
&\quad + 2r_1 \sqrt{(x - c_1)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} (x - c_1)} \\
&\leq |\lambda_{\max}(W)| \left(\|x - c\|^2 - \|x - c_1\|^2\right) + \left(r^2 - r_1^2\right) \\
&\quad + 4 \max(r, r_1) \cdot \max(\|c\|, \|c_1\|) \cdot |\lambda_{\max}(W)| \cdot \|x\|
\end{aligned}$$

If we take $\mathbb{E}_P$ on both sides of inequality above and with the assumption that $|\lambda_{\max}(W)| < R_3$, then the mild assumption that P has finite second moments (hence finite first moment) ensures the finiteness. It is not hard to see that the choice of θ_1 is not essential in verifying this assumption. For simplicity, we assume $\theta_1 = (c_1, r_1) = (0, 1)$ hereafter.

$$a(x) = \left(|\lambda_{\max}(W)| \|x\|^2 + 1 - 2 \sqrt{x^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} x}\right)$$

- (A-4) There is a $\theta_0 \in \Theta$ such that $\gamma(\theta) > \gamma(\theta_0)$ for all $\theta \neq \theta_0$. To see this, we use the Fubini theorem to take differentiation inside the $\mathbb{E}_P$ (note that this is taken with

respect to x) to conclude unique minima of $\gamma(\theta)$ (notice that we assume $\mathcal{I}$ fixed and therefore the index vector v is a fixed constant vector)

$$
\begin{aligned}
&\gamma(\theta) = \mathbb{E}_P \rho(x;\theta) - a(x) \\
&= \mathbb{E}_P \left((x-c)^T W (x-c) + r^2 - 2r\sqrt{(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}\,(x-c)} \right) - a(x) \\
&\frac{\partial}{\partial \theta}\gamma(\theta) = \mathbb{E}_P \begin{pmatrix} \frac{\partial}{\partial c}\rho(x;\theta) \\ \frac{\partial}{\partial r}\rho(x;\theta) \end{pmatrix} \\
&= \begin{pmatrix} -\mathbb{E}_P 2(x-c)^T \left(W + r\left[(x-c)^T \sqrt{W}^T v^T I_p v \sqrt{W}\,(x-c) \right]^{-\frac{1}{2}} \left[\sqrt{W}^T v^T I_p v \sqrt{W} \right] \right) \\ \mathbb{E}_P 2r - 2\left[(x-c)^T \sqrt{W}^T v^T I_p v \sqrt{W}\,(x-c) \right]^{\frac{1}{2}} \end{pmatrix} \\
&= 0
\end{aligned}
$$

By letting $\frac{\partial}{\partial \theta}\gamma(\theta) = 0$ and for $x \sim P$, we derive from the second equation that

$$r_0(c) = \mathbb{E}_P \left[(x-c)^T \sqrt{W}^T v^T I_p v \sqrt{W}\,(x-c) \right]^{\frac{1}{2}} \in [0, \min(R, 2C\sqrt{|\lambda_{\max}(W)|})],$$

and from the first equation

$$\mathbb{E}_P 2\,(x-c)^T \left(W + r_0(c) \left[(x-c)^T \sqrt{W}^T v^T I_p v \sqrt{W}\,(x-c) \right]^{-\frac{1}{2}} \left[\sqrt{W}^T v^T I_p v \sqrt{W} \right] \right) = 0$$

Consider the following function

$$
\begin{aligned}
F(c) \coloneqq \mathbb{E}_P 2\,(x-c)^T \Bigg(W + r_0(c) \Big[(x-c)^T \sqrt{W}^T v^T I_p v \sqrt{W}\,(x-c) \Big]^{-\frac{1}{2}} \\
\Big[\sqrt{W}^T v^T I_p v \sqrt{W} \Big] \Bigg)
\end{aligned}
\tag{H.8}
$$

as a function of c and the above equation becomes $F(c) = 0$. Taking a sandwiching-style argument, we first note that the second term in the second bracket is always

non-negative, then we construct uniform bounding functions:

$$\begin{aligned}
&F_1(c) \coloneqq \mathbb{E}_P 2\left(x-c\right)^T W \\
&\asymp \mathbb{E}_P \left(x-c\right)^T W, \\
&F_2(c) \coloneqq \mathbb{E}_P 2\left(x-c\right)^T \Bigg(W + \min(R, 2C\sqrt{|\lambda_{\max}(W)|})\cdot \\
&|\lambda_{\max}(W)| \left[\left(x-c\right)^T \sqrt{W}^T v^T I_p v \sqrt{W} \left(x-c\right)\right]^{-\frac{1}{2}}\Bigg) W \\
&\asymp \mathbb{E}_P \left(x-c\right)^T \left(1 + \frac{K(R, C, v, |\lambda_{\max}(W)|)}{\|x-c\|_2}\right) W,
\end{aligned}$$

(where $K(R, C, v, |\lambda_{\max}(W)|)$ is a non-negative constant) such that the following bound $F_1(c) \leq F(c) \leq F_2(c)$ holds (for each component of the vector-valued F_1, F_2) uniformly in c. However, it is clear that there exists $c_1^+, c_2^- \in [-C, C]^d \subset \mathbb{R}^d$

$$\begin{aligned}
F(c_1^+) \geq F_1(c_1^+) > 0, \\
F(c_2^-) \leq F_2(c_2^-) < 0.
\end{aligned}$$

Note that F is continuous in c (we can take derivative under $\mathbb{E}_P$ since P is assumed to possess finite second moment) and $[-C, C]^d$ is connected, we apply the multivariate intermediate value theorem to assert the existence of **a** solution c_0 for $F(c) = 0$. Therefore, we can keep this solution c_0, which we know its existence but do not know its expression. Back substitution of this solution of c_0 into the expression of r_0 yields

$$r_0 = \mathbb{E}_P \left[\left(x-c_0\right)^T \sqrt{W}^T v^T I_p v \sqrt{W} \left(x-c_0\right)\right]^{\frac{1}{2}},$$

where $\theta_0 = (c_0, r_0)$ is well-defined for P with finite second moment. This verifies the assumption (A-4).

- (A-5) With the notations in (A-3), since $\Theta \coloneqq [-C, C]^d \times [R_0, R] \subset \mathbb{R}^d \times \mathbb{R}^+$ is a compact space, it suffices to verify only (i) of (A-5). There is a continuous function $b(\theta) > 0$ such that

$$b(\theta) = \left(c^T W c + r^2 - 2r\sqrt{c^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} c}\right) + 1 \in [1, 1 + C^2|\lambda_{\max}(W)| + R^2]$$

For a fixed $x \in \mathbb{X}$, the function $g_x(\theta) = 2r\sqrt{(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}\,(x-c)}$ has gradient

$$\begin{aligned}\frac{\partial}{\partial\theta} g_x(c,r) &= \begin{pmatrix} \frac{\partial}{\partial c} g_x(c,r) \\ \frac{\partial}{\partial r} g_x(c,r) \end{pmatrix} \\ &= \begin{pmatrix} r\left[(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}\,(x-c)\right]^{-\frac{1}{2}} \cdot 2\,(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} \\ 2\sqrt{(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}\,(x-c)} \end{pmatrix}\end{aligned}$$

which is bounded from above in matrix norm by $2R\sqrt{|\lambda_{\max}(W)|} \cdot 4RC\sqrt{|\lambda_{\max}(W)|} \leq 16\max(R^2,1) \cdot C|\lambda_{\max}(W)| =: L_g < \infty$. Therefore, the function $g_x(c,r)$ is a L_g-Lipschitz function. We have

$$\begin{aligned}\inf_{\theta\in\Theta} \frac{\rho(x;\theta) - a(x)}{b(\theta)} &= \inf_{\theta\in\Theta} \Big\{ \big((x-c)^T W (x-c) - x^T W x\big) + \big(r^2 - 1\big) \\ &\quad - 2r\sqrt{(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}\,(x-c)} + 2\sqrt{x^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} x} \Big\} \\ &\quad \left(c^T W c + r^2 - 2r\sqrt{c^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} c} + 1\right)^{-1} \\ &\geq \inf_{\theta=(c,r)\in\Theta} \Big\{ \big((x-c)^T W (x-c) - x^T W x\big) + \big(r^2 - 1\big) \\ &\quad - 2r\sqrt{(x-c)^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W}\,(x-c)} + 2\sqrt{x^T \sqrt{W}^T I_{\mathcal{I}} \sqrt{W} x} \Big\} \\ &\quad (1 + C^2|\lambda_{\max}(W)| + R^2)^{-1} =: h(x)\end{aligned}$$

and $\frac{\rho(x,\theta)-a(x)}{b(\theta)} \geq h(x)$ by the infimum in the definition while $h(x)$ is integrable with respect to P due to the fact that $g_x(c,r)$ is Lipschitz.

Now we verify all assumptions (A-1) to (A-5) in Huber et al. (1967), Theorem 1 in the same paper ensures that Theorem A holds. The mild assumptions that θ lies in a compact subspace of $\mathbb{R}^d \times \mathbb{R}^+$ can be relaxed by verifying a more stringent set of conditions (A-5) as pointed out by Huber et al. (1967). Since we actually verify assuming that the index set $\mathcal{I}$ is fixed, we need to point out that in the l_0 optimization for each fixed $\mathcal{I}$ the consistency result holds. But for the l_1 relaxed problem, we cannot guarantee consistency even with stronger assumptions, only algorithmic convergence is guaranteed.

H.5 Proof of Theorem 5

Now we take the second view that the estimator sequence T_n for parameter $\theta = (c, r)$ and assume a fixed index set $\mathcal{I}$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\psi(x_i; T_n) \to 0, \text{a.s. } P$$
$$n \to \infty,$$

- (N-1) For each fixed $\theta \in \Theta$, $\psi(x;\theta)$ is $\mathcal{A}$-measurable and separable. Like the construction in (A-1), we can modify the ψ into a separable version $\tilde{\psi}$ if necessary and verify this assumption. Then following functions are well-defined (with finite second moment assumption on P and the Fubini theorem)

$$\begin{aligned}\lambda(\theta) = \lambda(c, r) &\coloneqq \mathbb{E}_P\psi(x;\theta)\\ &= \mathbb{E}_P\frac{\partial}{\partial\theta}\rho(x;\theta)\\ &= \frac{\partial}{\partial\theta}\mathbb{E}_P\rho(x;\theta)\\ u(x,\theta,D) &= \sup_{\|\tau-\theta\|\le D}\left|\psi(x;\tau)-\psi(x;\theta)\right|.\end{aligned}$$

- (N-2) The same θ_0 as computed above would satisfy $\lambda(\theta_0) = 0$.
- (N-3) There are strictly positive numbers $\alpha, \beta, \gamma, \eta$ such that
  - (i) $|\lambda(\theta)| \ge \alpha|\theta - \theta_0|$ for some $\alpha > 0$ and $|\theta - \theta_0| \le \eta$ is clear since

$$\lambda(\theta) = \frac{\partial}{\partial\theta}\mathbb{E}_P\left((x-c)^T W(x-c) + r^2 - 2r\sqrt{(x-c)^T\sqrt{W}^T I_{\mathcal{I}}\sqrt{W}\,(x-c)}\right)$$

    is quadratic in both c and r, and it is bounded from below by linear part due to Taylor expansion at θ_0.
  - (ii) $\mathbb{E}_P u(x,\theta,D) = \mathbb{E}_P \sup_{\|\tau-\theta\|\le D}|\psi(x;\tau)-\psi(x;\theta)| \le \mathbb{E}_P\beta\|\tau-\theta\|$ since $\frac{\partial}{\partial r}\psi(x;\theta) = 2$ and

$$\begin{aligned}
\frac{\partial}{\partial c}\psi(x;\theta) &= \mathbb{E}_P 2c^T \left(W + r\left[(x-c)^T \sqrt{W}^T v^T I v \sqrt{W}\,(x-c)\right]^{-\frac{1}{2}} \left[\sqrt{W}^T v^T I v \sqrt{W}\right]\right) \\
&\quad - \mathbb{E}_P 2(x-c)^T (-\frac{1}{2} r \left[(x-c)^T \sqrt{W}^T v^T I v \sqrt{W}\,(x-c)\right]^{-\frac{3}{2}} \\
&\quad \cdot 2\,(x-c)^T \sqrt{W}^T v^T I v \sqrt{W} \cdot \left[\sqrt{W}^T v^T I v \sqrt{W}\right]) \\
&\quad \leq 2CR(1 + (4C^2|\lambda_{\max}(W)|^{-\frac{1}{2}})|\lambda_{\max}(W)| \\
&\quad + 4C\left(R \cdot (4C^2|\lambda_{\max}(W)|)^{-\frac{3}{2}} \cdot 2C|\lambda_{\max}(W)|^2\right) \\
&\quad \leq 16CR(4C^2|\lambda_{\max}(W)|^{-\frac{1}{2}} \max(|\lambda_{\max}(W)|, 1)^4 + |\lambda_{\max}(W)| < \infty.
\end{aligned}$$

And $\psi(x;\theta)$ is Lipschitz with coefficient

$$\beta \coloneqq 32CR(4C^2|\lambda_{\max}(W)|^{-\frac{1}{2}} \max(|\lambda_{\max}(W)|, 1)^4 + |\lambda_{\max}(W)|.$$

- (iii) $\mathbb{E}_P u(x,\theta,D)^2 = \mathbb{E}_P \left(\sup_{\|\tau-\theta\|\leq D} |\psi(x;\tau) - \psi(x;\theta)|\right)^2 \leq \max\left\{\mathbb{E}_P \left(\beta\|\tau-\theta\|\right)^2, \left(\mathbb{E}_P \beta\|\tau-\theta\|\right)^2\right\}$ and for $\gamma = \beta$ we can replace $\|\tau-\theta\| \leq \eta - D$ with $\eta - D$.

- (N-4) $\mathbb{E}_P\left[|\psi(x;\theta_0)|^2\right] < \infty$ is clear from the analytic expression of $\psi(x;\theta)$, which involves at most quadratic entries in x, and the fact that we assume P has finite second moments.

Assumptions (N-1) through (N-4) allow us to apply Theorem 3 and its corollary in Huber et al. (1967) and claim Theorem B.

The asymptotic normality result allows us to claim a Wald-type hypothesis testing for the estimated center and radius for the sphere for a fixed index set $\mathcal{I}$. that aspect in the current paper but point out that this is one of the few non-bootstrap hypothesis testing methods in manifold learning literature.

H.6 Proof of Theorem 6

First we compare SRCA with PCA. Assume $\|x_i\| \leq \alpha$ for any i, then for any $\epsilon > 0$, there exists a sphere S_ϵ such that $d(y, S_\epsilon) \leq \epsilon$ for any $y \in H$ with $\|y\| \leq \alpha$ (Li et al., 2022). Intuitively, a plane can be approximated by a sphere with infinite radius. Let $\widehat{x}_i =$

$\arg\min_{y\in H} d(x_i, y)$ be the linear projection of x_i to plane H, then by the triangle inequality,

$$d(x_i, S_\epsilon) \leq d(x_i, \widehat{x}_i) + d(\widehat{x}_i, S_\epsilon).$$

Since the linear projection of a bounded set is still bounded, $d(\widehat{x}_i, S_\epsilon) \leq \epsilon$. By the definition of SRCA,

$$\sum_{i=1}^{n} d^2(x_i, S_2) \leq \sum_{i=1}^{n} d^2(x_i, S_\epsilon) \leq \sum_{i=1}^{n} d^2(x_i, H) + 2\epsilon \sum_{i=1}^{n} d(x_i, H) + n\epsilon^2.$$

Let $\epsilon \to 0$, we conclude that

$$\sum_{i=1}^{n} d^2(x_i, S_2) \leq \sum_{i=1}^{n} d^2(x_i, H).$$

Then we compare SRCA with SPCA. Since the objective function $\mathscr{L}$, which defines SRCA, is $\min_S \sum_{i=1}^{n} d^2(x_i, S)$, it follows from $S_1 \subset H$ that

$$\sum_{i=1}^{n} d^2(x_i, S_2) \leq \sum_{i=1}^{n} d^2(x_i, S_1).$$

Note that SPCA is a restricted version of SRCA, where $\mathcal{I} = \{1, \cdots, d'+1\}$.

I Mean Square Errors for Out-of-sample Data

This section provides the out-of-sample mean square errors of PCA, SPCA and SRCA on the same datasets in Table 4.1. We provide this to show that performance evaluation measures are not really affected by the choice of testing samples.

Dataset	Method/$d' =$	1	2	3	4
Banknote	PCA	15.6094	6.2737	1.9278	
	SPCA	15.0516	7.3182	1.5511	
	SRCA	**13.2273**	**5.4120**	**1.1257**	
Power Plant	PCA	227.6291	56.0524	23.4302	**3.0244**
	SPCA	151.7555	102.3262	44.3802	41.0081
	SRCA	**151.3426**	**52.9769**	**20.0871**	4.0775
User Knowledge	PCA	0.1952	0.1281	0.0749	0.0306
	SPCA	0.1478	0.0898	0.0465	0.0145
	SRCA	**0.1479**	**0.0904**	**0.0462**	**0.0144**
Ecoli	PCA	0.0761	0.0334	0.0219	**0.0057**
	SPCA	**0.0462**	0.0351	0.0187	0.0122
	SRCA	0.0758	**0.0337**	**0.0168**	0.0058
Concrete	PCA	6.8783	4.8345	3.5462	2.5046
	SPCA	**5.5565**	**4.2051**	**3.1664**	**2.0173**
	SRCA	5.5573	4.2173	3.1842	2.0389
Leaf	PCA	0.0245	0.0126	**0.0062**	0.0040
	SPCA	**0.0163**	**0.0100**	0.0073	0.0047
	SRCA	0.0164	0.0101	**0.0062**	**0.0037**
Climate	PCA	1.4486	1.3846	1.3167	1.2447
	SPCA	1.4265	1.3637	1.2921	1.2224
	SRCA	**1.3863**	**1.3081**	**1.2278**	**1.1525**

Table I.1: Out-of-sample mean square error (MSE) table for different experiments.