Kernel Principal Geodesic Analysis

Suyash P. Awate[1,2,⋆], Yen-Yun Yu[1], and Ross T. Whitaker[1]

[1] Scientific Computing and Imaging (SCI) Institute, School of Computing, University of Utah
[2] Computer Science and Engineering Department, Indian Institute of Technology (IIT) Bombay

Abstract. Kernel principal component analysis (kPCA) has been proposed as a dimensionality-reduction technique that achieves nonlinear, low-dimensional representations of data via the mapping to kernel feature space. Conventionally, kPCA relies on Euclidean statistics in kernel feature space. However, Euclidean analysis can make kPCA inefficient or incorrect for many popular kernels that map input points to a *hypersphere* in kernel feature space. To address this problem, this paper proposes a novel adaptation of kPCA, namely *kernel principal geodesic analysis* (kPGA), for hyperspherical statistical analysis in kernel feature space. This paper proposes tools for statistical analyses on the Riemannian manifold of the Hilbert sphere in the reproducing kernel Hilbert space, including algorithms for computing the sample weighted Karcher mean and eigen analysis of the sample weighted Karcher covariance. It then applies these tools to propose novel methods for (i) dimensionality reduction and (ii) clustering using mixture-model fitting. The results, on simulated and real-world data, show that kPGA-based methods perform favorably relative to their kPCA-based analogs.

1 Introduction

Kernel principal component analysis (kPCA) [47] maps points in *input space* to a (high-dimensional) *kernel feature space* where it estimates a best-fitting linear subspace via PCA. This mapping to the kernel feature space is typically denoted by $\Phi(\cdot)$. For many of the most useful and widely used kernels (e.g., Gaussian, exponential, Matern, spherical, circular, wave, power, log, rational quadratic), the input data x gets mapped to a *hypersphere*, or a *Hilbert sphere*, in the kernel feature space. Such a mapping also occurs when using (i) kernel normalization, which is common, e.g., in pyramid match kernel [28], and (ii) polynomial and sigmoid kernels when the input points have constant l^2 norm, which is common in digit image analysis [46]. This special structure arises because for these kernels $k(\cdot,\cdot)$, the self similarity of any data point x equals unity (or some constant), i.e., $k(x,x) = 1$. The kernel defines the inner product in the kernel feature space $\mathcal{F}$, and thus, $\langle \Phi(x), \Phi(x) \rangle_{\mathcal{F}} = 1$, which, in turn, equals the distance of the mapped point $\Phi(x)$ from the origin in $\mathcal{F}$. Thus, all of the mapped points $\Phi(x)$ lie on a Hilbert sphere in kernel feature space. Figure 1(a) illustrates this behavior.

The literature shows that for many high-dimensional real-world datasets, where the data representation uses a large number of dimensions, the intrinsic dimension is often quite small, e.g., between 5–20 in [18,43,24,29,42]. The utility of kPCA lies in capturing the intrinsic dimension of the data through the few principal (linear) modes of variation in kernel feature space. This paper proposes a novel extension of kPCA to model

⋆ We thank NIH support via NCRR CIBC P41-RR12553 and NCBC NAMIC U54-EB005149.

T. Calders et al. (Eds.): ECML PKDD 2014, Part I, LNCS 8724, pp. 82–98, 2014.

distributions on the Hilbert sphere manifold in kernel feature space. Manifold-based statistical analysis explicitly models data to reside in a lower dimensional subspace of the ambient space, representing variability in the data more efficiently (fewer degrees of freedom). In this way, the proposed method extends kPCA to (i) define more meaningful modes of variation in kernel feature space by explicitly modeling the data on the Hilbert sphere in kernel feature space, (ii) represent variability using fewer modes, and (iii) reduce curvature of distributions by modeling them explicitly on the Hilbert sphere, instead of modeling them in the ambient space, to avoid artificially large measurements of variability observed in the ambient space. Figure 1(b) illustrates this idea.

Typically, Euclidean PCA of spherical data introduces one additional (unnecessary) component, aligned orthogonally to the sphere and proportional to the sectional curvature. In practice, however, PCA in high-dimensional spaces (e.g., kernel feature space) is known to be unstable and prone to error [4], which interacts with the curvature of the Hilbert sphere on which the data resides. Thus, our empirical results demonstrate that the actual gains in our hyperspherical analysis in kernel feature space surpass what we would expect for the low-dimensional case.

While several works in the literature [3,21,23,27,46] address the properties and uses of kernel feature spaces, these works do *not* systematically explore this special structure of kernel feature space and its implications for PCA in kernel feature space; that is the focus of this paper. Recently, [21] have, in an independent development, examined the use of the Karcher mean in kernel feature spaces, but they propose a different estimation strategy and they do *not* formulate, estimate, or demonstrate the use of principle components on the sphere, which is the main purpose of this work.

This paper makes several contributions. It proposes new formulations and algorithms for computing the sample Karcher mean on a Hilbert sphere in reproducing kernel Hilbert space (RKHS). To analyze sample Karcher covariance, this paper proposes a kernel-based PCA on the Hilbert sphere in RKHS, namely, *kernel principal geodesic analysis* (kPGA). It shows that just as kPCA leads to a standard eigen-analysis problem, kPGA leads to a generalized eigen-analysis problem. This paper evaluates the utility of kPGA for (i) nonlinear dimensionality reduction and (ii) clustering with a Gaussian mixture model (GMM) and an associated expectation maximization (EM) algorithm on the Hilbert sphere in RKHS. Results on simulated and real-world data show that kPGA-based methods perform favorably with their kPCA-based analogs.

2 Related Work

There are several areas of related work that inform the results in this paper. The Karcher mean and associated covariance have recently become important tools for statistical analysis [39]. The algorithm for the Karcher mean proposed in [17] is restricted to analyzing the intrinsic mean and does *not* address how to capture covariance for data lying on spheres, even in finite-dimensional spaces. Other algorithms for the Karcher mean exist and may be more efficient numerically [34]. To capture covariance structure on Riemannian manifolds, Fletcher et al. [25] propose PGA and an associated set of algorithms. Likewise, a small body of work relies on the local geometric structure of Riemannian spaces of covariance matrices for subsequent statistical analysis [7,20,50].

Because many RKHSs are infinite dimensional, we must acknowledge the problem of modeling distributions in such spaces [30] and the corresponding theoretical problems [16]. Of course, these same theoretical concerns arise in kPCA, and other well-known kernel methods, and thus the justification for this work is similar. First, we may assume or assert that the covariance operator of the mapped data is of trace class or, even more strongly, restricted to a finite-dimensional manifold defined by the cardinality of the input data. Second, the proposed methods are intended primarily for data analysis rather than statistical estimation, and, thus, we intentionally work in the subspace defined by the data (which is limited by the data sample size).

In addition to the dimensionality structure, the Hilbert sphere imposes its own structure and has an associated geometry with underlying theoretical implications. The proposed approach in this paper extends PGA [25] to the Hilbert sphere in RKHS. The important geometrical properties of the sphere for the proposed extension concern (i) the geodesic distance between two points, which depends on the arc cosine of their dot product, and (ii) the existence and formulation of tangent spaces [11,15,31].

The work in [21] is more directly related to the proposed method, because it uses logarithmic and exponential maps on the Hilbert sphere in RKHS for data analysis. However, [21] does *not* define a mean or a covariance on the Hilbert sphere in RKHS; it also requires the solution of the ill-posed preimage problem. Unlike [21], we define covariance and its low-dimensional approximations on the Hilbert sphere, represented in terms of the Gram matrix of the data, and incorporate this formulation directly into novel algorithms for dimensionality reduction and clustering via EM [22], including geodesic Mahalanobis distance on the Hilbert sphere in RKHS.

We apply the proposed method for (i) dimensionality reduction for machine-learning applications and (ii) mixture modeling. This builds on the work in kPCA [47], and therefore represents an alternative to other nonlinear mapping methods, such as Sammon's nonlinear mapping [45], Isomap [51] and other kernel-based methods [35,52]. For applications to clustering, the proposed approach generalizes kernel k-means [47] and kernel GMMs [53], where we use formulations of means and/or covariances that respect the hyperspherical geometry of the mapped points in RKHS.

3 Geometry of the Hilbert Sphere in Kernel Feature Space

Many popular kernels are associated with a RKHS that is infinite dimensional. Thus, the analysis in this paper focuses on such spaces. Nevertheless, analogous theory holds for other important kernels (e.g., normalized polynomial) where the RKHS is finite dimensional.

Let X be a random variable taking values x in *input space* $\mathcal{X}$. Let $\{x_n\}_{n=1}^N$ be a set of observations in input space. Let $k(\cdot,\cdot)$ be a real-valued Mercer kernel with an associated map $\Phi(\cdot)$ that maps x to $\Phi(x) := k(\cdot, x)$ in a RKHS $\mathcal{F}$ [6,46]. Consider two points in RKHS: $f := \sum_{i=1}^I \alpha_i \Phi(x_i)$ and $f' := \sum_{j=1}^J \beta_j \Phi(x_j)$. The inner product $\langle f, f' \rangle_{\mathcal{F}} := \sum_{i=1}^I \sum_{j=1}^J \alpha_i \beta_j k(x_i, x_j)$. The norm $\|f\|_{\mathcal{F}} := \sqrt{\langle f, f \rangle_{\mathcal{F}}}$. When $f, f' \in \mathcal{F} \backslash \{0\}$, let $f \otimes f'$ be the rank-one operator defined as $f \otimes f'(h) := \langle f', h \rangle_{\mathcal{F}} f$. Let $Y := \Phi(X)$ be the random variable taking values y in RKHS.

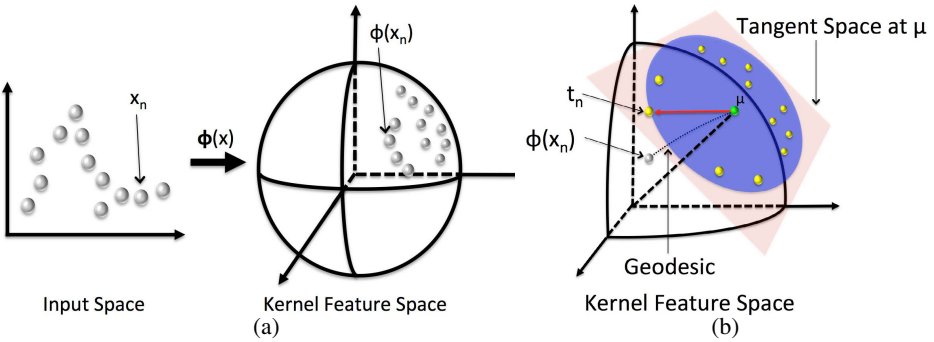


Fig. 1. Kernel Principal Geodesic Analysis (kPGA). (a) Points in input space get mapped, via several popular Mercer kernels, to a hypersphere or a Hilbert sphere in kernel feature space. **(b)** Principal geodesic analysis on the Hilbert sphere in kernel feature space.

Assuming Y is bounded and assuming the expectation and covariance operators of Y exist and are well defined, kPCA uses observations $\{y_n := \Phi(x_n)\}_{n=1}^N$ to estimate the eigenvalues, and associated eigenfunctions, of the covariance operator of Y [14,47]. The analysis in this paper applies to kernels that map points in input space to a Hilbert sphere in RKHS, i.e., $\forall x : k(x,x) = \kappa$, a constant (without loss of generality, we assume $\kappa = 1$). For such kernels, the proposed kPGA modifies kPCA using statistical modeling on the Riemannian manifold of the unit Hilbert sphere [5,10] in RKHS.

Consider a and b on the unit Hilbert sphere in RKHS represented, in general, as $a := \sum_n \gamma_n \Phi(x_n)$ and $b := \sum_n \delta_n \Phi(x_n)$. The *logarithmic map*, or Log map, of a with respect to b is the vector

$$\mathrm{Log}_b(a) = \frac{a - \langle a,b\rangle_{\mathcal{F}} b}{\|a - \langle a,b\rangle_{\mathcal{F}} b\|_{\mathcal{F}}} \arccos(\langle a,b\rangle_{\mathcal{F}}) = \sum_n \zeta_n \Phi(x_n), \text{ where } \forall n : \zeta_n \in \mathbb{R}. \tag{1}$$

Clearly, $\mathrm{Log}_b(a)$ can always be written as a weighted sum of the vectors $\{\Phi(x_n)\}_{n=1}^N$. The *tangent vector* $\mathrm{Log}_b(a)$ lies in the *tangent space*, at b, of the unit Hilbert sphere. The tangent space to the Hilbert sphere in RKHS inherits the same structure (inner product) as the ambient space and, thus, is also a RKHS. The geodesic distance between a and b is $d_g(a,b) = \|\mathrm{Log}_b(a)\|_{\mathcal{F}} = \|\mathrm{Log}_a(b)\|_{\mathcal{F}}$.

Now, consider a tangent vector $t := \sum_n \beta_n \Phi(x_n)$ lying in the tangent space at b. The *exponential map*, or Exp map, of t with respect to b is

$$\mathrm{Exp}_b(t) = \cos(\|t\|_{\mathcal{F}})b + \sin(\|t\|_{\mathcal{F}})\frac{t}{\|t\|_{\mathcal{F}}} = \sum_n \omega_n \Phi(x_n), \text{ where } \forall n : \omega_n \in \mathbb{R}. \tag{2}$$

Clearly, $\mathrm{Exp}_b(t)$ can always be written as a weighted sum of the vectors $\{\Phi(x_n)\}_{n=1}^N$. $\mathrm{Exp}_b(t)$ maps a tangent vector t to the unit Hilbert sphere, i.e., $\|\mathrm{Exp}_b(t)\|_{\mathcal{F}} = 1$.

4 PCA on the Hilbert Sphere in Kernel Feature Space

This section proposes the kPGA algorithm for PCA on the unit Hilbert sphere in RKHS.

4.1 Sample Karcher Mean

The sample Karcher mean on Riemannian manifolds is a consistent estimator of the theoretical Karcher mean of the underlying random variable [12,13]. The sample weighted Karcher mean of set of observations $\{y_m\}_{m=1}^M$, on the unit Hilbert sphere in RKHS, with associated weights $\{p_m \in \mathbb{R}^+\}_{m=1}^M$ is defined as

$$\mu := \arg\min_{\nu} \sum_m p_m d_g^2(\nu, y_m). \tag{3}$$

The existence and uniqueness properties of the Karcher mean on the Riemannian manifold of the unit Hilbert sphere are well studied [1,32,33]; a study on finite-dimensional Hilbert spheres appears in [17]. The sample Karcher mean on a Hilbert sphere exists and is unique if the pointset is contained within (i) an open convex Riemannian ball of radius $\pi/2$ [33], i.e., an open hemisphere, or (ii) a similar closed ball if one of the points lies in its interior [17]. Thus, the sample Karcher mean exists and is unique for all kernels that map points within a single orthant of the Hilbert sphere in RKHS; this is true for all positive-valued kernels, e.g., the Gaussian kernel.

Clearly, a Karcher mean μ must lie within the space spanned by $\{y_m\}_{m=1}^M$; if not, we could project the assumed "mean" ν' onto the span of $\{y_m\}_{m=1}^M$ and reduce all distances $d_g(y_m, \nu')$ on the Hilbert sphere because of the spherical Pythagoras theorem, thereby resulting in a more-optimal mean ν'' with $d_g(y_m, \nu'') < d_g(y_m, \nu'), \forall m$ and a contradiction to the initial assumption. Therefore, if the points y_m are represented using another set of points $\{\Phi(x_n)\}_{n=1}^N$, i.e., $\forall m, y_m := \sum_n w_{mn}\Phi(x_n)$, then the mean μ can be represented as $\mu = \sum_n \xi_n \Phi(x_n)$, where $\forall n : \xi_n \in \mathbb{R}$.

We propose the following gradient-descent algorithm to compute the mean μ.

1. **Input:** A set of points $\{y_m\}_{m=1}^M$ on the unit Hilbert sphere in RKHS. Weights $\{p_m\}_{m=1}^M$. As described previously, we assume that, in general, each y_m is represented using another set of points $\{\Phi(x_n)\}_{n=1}^N$ and weights w_{mn} on the unit Hilbert sphere in RKHS, i.e., $y_m := \sum_n w_{mn}\Phi(x_n)$.
2. Initialize iteration count: $i = 0$. Initialize the mean estimate to
$$\mu^0 = \frac{\sum_m p_m y_m}{\|\sum_m p_m y_m\|_{\mathcal{F}}} = \sum_n \xi_n \Phi(x_n), \text{ where } \xi_n = \frac{\sum_m p_m w_{mn}}{\|\sum_m p_m y_m\|_{\mathcal{F}}}. \tag{4}$$
3. Iteratively update the mean estimate, until convergence, by (i) taking the Log maps of all points with respect to the current mean estimate, (ii) performing a weighted average of the resulting tangent vectors, and (iii) taking the Exp map of the weighted average scaled by a step size τ^i, i.e.,
$$\mu^{i+1} = \text{Exp}_{\mu^i}\left(\frac{\tau^i}{M}\sum_m p_m \text{Log}_{\mu^i}(y_m)\right), \text{ where } \tau^i \in (0, 1). \tag{5}$$
4. **Output:** Mean μ lying on the unit Hilbert sphere in RKHS.

In practice, we use a gradient-descent algorithm with an adaptive step size τ^i such that the algorithm (i) guarantees that the objective-function value is non increasing every iteration and (ii) increases/decreases the step size each iteration to aid faster convergence.

We detect convergence as the point when the objective function cannot be reduced using any non-zero step size. Typically, in practice, a few iterations suffice for convergence.

The convergence of gradient descent for finding Karcher means has been studied [2,17]. In certain conditions, such as those described earlier when the sample Karcher mean on a Hilbert sphere is unique, the objective function becomes convex [19], which leads the gradient descent to the global minimum.

4.2 Sample Karcher Covariance and Eigen Analysis

Given the sample weighted Karcher mean μ, consider a random variable $Z := \mathrm{Log}_\mu(Y)$ taking values in the tangent space at μ. Assuming that both the expectation and covariance operators of Z exist and are well defined (this follows from the similar assumption on Y), the sample weighted Karcher covariance operator, in the tangent space at μ, is

$$C := (1/M) \sum_m p_m z_m \otimes z_m, \text{ where } z_m := \mathrm{Log}_\mu(y_m). \tag{6}$$

Because the tangent space is a RKHS, the theoretical analysis of covariance in RKHS in standard kPCA [14,48] applies to C as well (note that the set $\{z_m\}_{m=1}^M$ is empirically centered by construction; i.e., $\sum_m z_m = 0$). Thus, as the sample size $M \to \infty$, the partial sums of the empirically-computed eigenvalues converge to the partial sums of the eigenvalues of the theoretical covariance operator of Z.

Using the Log map representation in Section 3, $z_m = \sum_{n'} \beta_{n'm} \Phi(x_{n'})$ leading to

$$C = \sum_{n'} \sum_{n''} E_{n'n''} \Phi(x_{n'}) \otimes \Phi(x_{n''}), \text{ where } E_{n'n''} = \frac{1}{M} \sum_m p_m \beta_{n'm} \beta_{n''m}. \tag{7}$$

If λ is a positive eigenvalue of C and v is the corresponding eigenfunction, then

$$v = \frac{Cv}{\lambda} = \frac{1}{\lambda} \sum_{n'} \sum_{n''} E_{n'n''} \Phi(x_{n'}) \otimes \Phi(x_{n''}) v = \sum_{n'} \alpha_{n'} \Phi(x_{n'}),$$

$$\text{where } \alpha_{n'} = \sum_{n''} \frac{E_{n'n''}}{\lambda} \langle \Phi(x_{n''}), v \rangle_{\mathcal{F}}. \tag{8}$$

Thus, any eigenfunction v of C lies within the span of the set of points $\{\Phi(x_n)\}_{n=1}^N$ used to represent $\{y_m\}_{m=1}^M$. For any $\Phi(x_\eta) \in \{\Phi(x_n)\}_{n=1}^N$ and the eigenfunction v,

$$\langle \Phi(x_\eta), Cv \rangle_{\mathcal{F}} = \lambda \langle \Phi(x_\eta), v \rangle_{\mathcal{F}}. \tag{9}$$

$$\text{Thus, } \langle \Phi(x_\eta), \sum_{n'} \sum_{n''} E_{n'n''} \Phi(x_{n'}) \otimes \Phi(x_{n''}) \sum_{n'''} \alpha_{n'''} \Phi(x_{n'''}) \rangle_{\mathcal{F}} =$$

$$\lambda \langle \Phi(x_\eta), \sum_{n'''} \alpha_{n'''} \Phi(x_{n'''}) \rangle_{\mathcal{F}}. \tag{10}$$

$$\text{Thus, } \sum_{n'''} \left(\sum_{n'} K_{\eta n'} \sum_{n''} E_{n'n''} K_{n''n'''} \right) \alpha_{n'''} = \lambda \sum_{n'''} K_{\eta n'''} \alpha_{n'''}, \tag{11}$$

$$\text{where } K_{ij} := \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{F}}$$

is the element in row i and column j of the Gram matrix K. Considering E and K as $N \times N$ real matrices and defining $F := EK$ and $G := KF$ leads to

$$\sum_{n''} E_{n'n''} K_{n''n'''} = F_{n'n'''} \text{ and } \sum_{n'} K_{\eta n'} \sum_{n''} E_{n'n''} K_{n''n'''} = G_{\eta n'''}. \tag{12}$$

Therefore, the left hand side of Equation 9 equals $G_{\eta\bullet}\alpha$, where (i) $G_{\eta\bullet}$ is the η^{th} row of the $N \times N$ matrix G and (ii) α is the $N \times 1$ column vector with the n^{th} component as α_n. Similarly, the right hand side of Equation 9 equals $K_{\eta\bullet}\alpha$, where $K_{\eta\bullet}$ is the η^{th} row of the $N \times N$ matrix K. Using Equation 9 to form one equation for all $\eta = 1, \cdots, N$, gives the following generalized eigen-analysis problem

$$G\alpha = \lambda K\alpha. \tag{13}$$

If $k(\cdot,\cdot)$ is a symmetric positive-*definite* (SPD) Mercer kernel and the points $\{\Phi(x_n)\}_{n=1}^N$ are *distinct*, then K is SPD (hence, invertible) and the generalized eigen-analysis problem reduces to the standard eigen-analysis problem

$$EK\alpha = \lambda\alpha. \tag{14}$$

Thus, (i) the eigenvalues $\{\lambda_n\}_{n=1}^N$ are same as the eigenvalues of the sample covariance operator C and (ii) each eigenvector α gives one eigenfunction of C through Equation 8. Note that standard kPCA requires eigen decomposition of the (centralized) matrix K.

The definition of the sample covariance operator C implies that the rank of C is upper bounded by the sample size M. Because the eigenvalues of C are the same as those for EK or for the pair (G, K), if $M < N$, then the rank of the $N \times N$ matrices EK and G are also upper bounded by M. While K is an $N \times N$ symmetric positive (semi) definite matrix of rank at-most N, E is an $N \times N$ symmetric positive (semi) definite matrix of rank at-most M because $E = BPB^T$ where (i) B is a $N \times M$ matrix where $B_{nm} = \beta_{nm}$ and (ii) P is an $M \times M$ diagonal matrix where $P_{mm} = p_m/M$.

4.3 Kernel Principal Geodesic Analysis (kPGA) Algorithm

We summarize the proposed **kPGA** algorithm below.

1. **Input:** (i) A set of points $\{y_m\}_{m=1}^M$ on the unit Hilbert sphere in RKHS. (ii) Weights $\{p_m\}_{m=1}^M$. As described previously, we assume that, in general, each y_m is represented using another set of points $\{\Phi(x_n)\}_{n=1}^N$ and weights w_{mn} on the unit Hilbert sphere in RKHS, i.e., $y_m := \sum_n w_{mn}\Phi(x_n)$.
2. Compute the Gram matrix K.
3. Compute the Karcher mean μ using the algorithm in Section 4.1.
4. Compute the matrix E or $G = KEK$ as described in Section 4.2.
5. To analyze the Karcher covariance, perform eigen analysis for the linear system $G\alpha = \lambda K\alpha$ or $EK\alpha = \lambda\alpha$ to give eigenvalues $\{\lambda_\eta\}_{\eta=1}^N$ (sorted in non-increasing order) and eigenvectors $\{\alpha_\eta\}_{\eta=1}^N$.
6. **Output:** (i) Mean μ lying on the unit Hilbert sphere in RKHS. (ii) Principal components or eigenfunctions $\{v_n = \sum_{n'} \alpha_{\eta n'}\Phi(x_{n'})\}_{n=1}^N$ in the tangent space at μ. (iii) Eigenvalues $\{\lambda_n = \lambda_\eta\}_{n=1}^N$ capturing variance along principal components.

5 Applications

This section proposes kPGA-based algorithms for (i) nonlinear dimensionality reduction and (ii) clustering using a mixture model fitted using EM.

5.1 Nonlinear Dimensionality Reduction

First, we propose the following algorithm for dimensionality reduction using kPGA.

1. **Input:** A set of points $\{x_n\}_{n=1}^N$ along with their maps $\{\Phi(x_n)\}_{n=1}^N$ on the unit Hilbert sphere in RKHS. Weights $\{p_n = 1\}_{n=1}^N$.
2. Apply the kPGA algorithm in Section 4.2 to the observed sample $\{\Phi(x_n)\}_{n=1}^N$ to compute mean μ, eigenvalues $\{\lambda_n\}_{n=1}^N$ (sorted in non-increasing order), and corresponding eigenfunctions $\{v_n\}_{n=1}^N$.
3. Select the largest $Q < N$ eigenvalues $\{\lambda_q\}_{q=1}^Q$ that capture a certain fraction of energy in the eigenspectrum. Select the corresponding subspace $\mathbb{G}_Q =< v_1, \cdots, v_Q >$.
4. Project the Log map of each point $\Phi(x_n)$ on the subspace $\mathbb{G}_Q$ to give the embedding coordinates $e_{nq} := \langle \mathrm{Log}_\mu \Phi(x_n), v_q \rangle_{\mathcal{F}}$ and projected tangent vectors $t_n = \sum_q e_{nq} v_q$ in the tangent space at the mean μ.
5. Take the Exp map of projections $\{t_n\}_{n=1}^N$ to produce $\{y_n = \mathrm{Exp}_\mu(t_n)\}_{n=1}^N$ lying within a Q-dimensional subsphere on the unit Hilbert sphere in RKHS.
6. **Output**: Embedding subspace (lower dimensional) $\mathbb{G}_Q$, embedding coordinates $\{(e_{n1}, \cdots, e_{nQ})\}_{n=1}^N$, and (re)mapped points on the Hilbert subsphere $\{y_n\}_{n=1}^N$.

5.2 Clustering Using Mixture Modeling and Expectation Maximization

We now propose an algorithm for clustering a set of points $\{x_n\}_{n=1}^N$, into a fixed number of clusters, by fitting a mixture model on the unit Hilbert sphere in RKHS.

The proposed approach entails mixture modeling in a finite-dimensional subsphere of the unit Hilbert sphere in RKHS, after the dimensionality reduction of the points $\{\Phi(x_n)\}$ to a new set of points $\{y_n\}$ (as in Section 5.1). Modeling PDFs on Hilbert spheres entails fundamental trade-offs between model generality and the viability of the underlying parameter estimation. For instance, although Fisher-Bingham PDFs on $\mathbb{S}^d$ are able to model generic anisotropic distributions (anisotropy around the mean) using $O(d^2)$ parameters, their parameter estimation may be intractable [9,37,40]. On the other hand, parameter estimation for the $O(d)$-parameter von Mises-Fisher PDF is tractable [9], but that PDF can only model isotropic distributions. We take another approach that uses a tractable approximation of a normal law on a Riemannian manifold [41], allowing modeling of anisotropic distributions through its covariance parameter in the tangent space at the mean. Thus, the proposed PDF evaluated at $\Phi(x)$ is $P(\Phi(x)|\mu, C) \doteq \exp\left(-0.5 d_g^2(\mu, \Phi(x); C)\right) / ((2\pi)^{Q/2}|C|^{1/2})$, where $|C| = \Pi_{q=1}^Q \lambda_q$ and $d_g(\mu, \nu; C)$ is the *geodesic Mahalanobis distance* between the point $\Phi(x)$ and mean μ, given covariance C.

The geodesic Mahalanobis distance relies on a regularized sample inverse-covariance operator [38] $C^{-1} := \sum_{q=1}^Q (1/\lambda_q) v_q \otimes v_q$, where λ_q is the q^{th} sorted eigenvalue of

C, v_q is the corresponding eigenfunction, and $Q \leq \min(M, N)$ is a regularization parameter. Then, the corresponding square-root inverse-covariance operator is $C^{-1/2} := \sum_q (1/\sqrt{\lambda_q}) v_q \otimes v_q$ and the geodesic Mahalanobis distance of the point ν from mean μ is $d_g(\nu, \mu; C) := (\langle C^{-1/2} t, C^{-1/2} t \rangle_{\mathcal{F}})^{0.5}$ where $t := \text{Log}_\mu(\nu)$.

Let Y be a random variable that generates the N independent and identically-distributed data points $\{y_n\}_{n=1}^N$ as follows. For each n, we first draw a cluster number $l \in \{1, 2, \cdots, L\}$ with probability w_l (where $\forall l : w_l > 0$ and $\sum_l w_l = 1$) and then draw y_n from $P(Y|\mu_l, C_l)$. Thus, the probability of observing y_n is $P(y_n) = \sum_l w_l P(y_n|\mu_l, C_l)$.

The *parameters* for $P(Y)$ are $\theta = \{w_l, \mu_l, C_l\}_{l=1}^L$. We solve for the maximum-likelihood estimate of θ via EM. Let $\{S_n\}_{n=1}^N$ be *hidden* random variables that give, for each n, the cluster number $s_n \in \{1, \cdots, L\}$ that generated data point y_n.

EM performs iterative optimization. Each EM iteration involves an E step and an M step. At iteration i, given parameter estimates θ^i, the E step defines a function $\mathcal{Q}(\theta|\theta^i) := E_{P(\{S_n\}_{n=1}^N | \{y_n\}_{n=1}^N, \theta^i)}[\log P(\{S_n, y_n\}_{n=1}^N | \theta)]$. For our mixture model,

$$\mathcal{Q}(\theta|\theta^i) = \sum_n \sum_l P(s_n = l|y_n, \theta^i) \left(\log w_l - 0.5 \log |C_l| - 0.5 d_g^2(\mu_l, y_n; C_l) \right) + \text{ constant, where} \tag{15}$$

$$P(s_n = l|y_n, \theta^i) = \frac{P(s_n = l|\theta^i) P(y_n|s_n = l, \theta^i)}{P(y_n|\theta^i)} = \frac{w_l^i P(y_n|\mu_l^i, C_l^i)}{\sum_l w_l^i P(y_n|\mu_l^i, C_l^i)}. \tag{16}$$

We denote $P(s_n = l|y_n, \theta^i)$ in shorthand by the class membership P_{nl}^i. We denote $\sum_n P_{nl}^i$ in shorthand by P_l^i. Simplifying gives

$$\mathcal{Q}(\theta|\theta^i) = \sum_l P_l^i \left(\log w_l - 0.5 \log |C_l| \right) - 0.5 \sum_n \sum_l P_{nl}^i d_g^2(\mu_l, y_n; C_l) + \text{ constant.} \tag{17}$$

The M step maximizes $\mathcal{Q}(\theta)$, under the constraints on w_l, using the method of Lagrange multipliers, to give the optimal values and, hence, the updates, for parameters θ.

Thus, the proposed clustering algorithm is as follows.

1. **Input:** A set of points $\{\Phi(x_n)\}_{n=1}^N$ on the unit Hilbert sphere in RKHS with all associated weights p_n set to unity.
2. Reduce the dimensionality of the input using the algorithm in Section 5.1 to give points $\{y_n\}_{n=1}^N$ on a lower-dimensional subsphere of the Hilbert sphere in RKHS.
3. Initialize iteration count $i := 0$. Initialize parameters $\theta^0 = \{w_l^0, \mu_l^0, C_l^0\}_{l=1}^L$ as follows: run farthest-point clustering [26] (with kernel-based distances; with randomly-selected first point) to initialize kernel k means [47] that, in turn, initializes μ_l^0 and C_l^0 to be the mean and covariances of cluster l, respectively, and w_l^0 equal to the number of points in cluster l divided by N.
4. Iteratively update the parameter estimates, until convergence, as follows.
5. Evaluate probabilities $\{P_{nl}^i\}$ using current parameter estimates θ^i.

6. Update means $\mu_l^{i+1} = \arg\min_\mu \sum_n P_{nl}^i d_g^2(\mu_l, y_n; C_l)$ using a gradient-descent algorithm similar to that used in Section 4.1 for the sample weighted Karcher mean.
7. Update covariances $C_l^{i+1} = \sum_n (P_{nl}^i / P_l^i) \mathrm{Log}_{\mu_l^{i+1}}(y_n) \otimes \mathrm{Log}_{\mu_l^{i+1}}(y_n)$.
8. Update probabilities $w_l^{i+1} = P_l^i / (\sum_l P_l^i)$.
9. **Output:** Parameters: $\theta = \{w_l, \mu_l, C_l\}_{l=1}^L$. Labeling: Assign $\Phi(x_n)$ to the cluster l that maximizes $P(y_n|\mu_l, C_l)$.

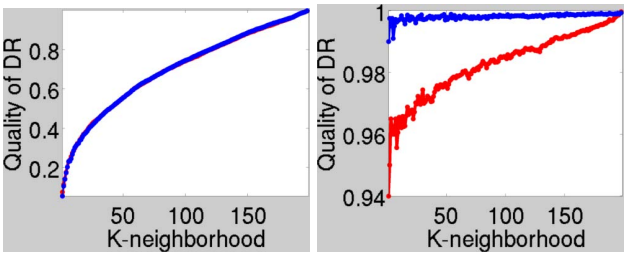
6 Results and Discussion

This section shows results on simulated data, real-world face images from the Olivetti Research Laboratory (ORL) [44], and real-world data from the University of California Irvine (UCI) machine learning repository [8].

6.1 Nonlinear Dimensionality Reduction

We employ kPCA and the proposed kPGA for nonlinear dimensionality reduction on simulated and real-world databases. To evaluate the quality of dimensionality reduction, we use the co-ranking matrix [36] to compare rankings of pairwise distances between (i) data points in the original high-dimensional space (i.e., without any dimensionality reduction) and (ii) the projected data points in the lower-dimensional embedding found by the algorithm. Based on this motivation, a standard measure to evaluate the quality of dimensionality-reduction algorithms is to average, over all data points, the fraction of other data points that remain inside a κ neighborhood defined based on the original distances [36]. For a fixed number of reduced dimensions, an ideal dimensionality-reduction algorithm would lead to this quality measure being 1 for every value of $\kappa \in \{1, 2, \cdots, N-1\}$, where N is the total number of points in the dataset.

Simulated Data – Points on a High-Dimensional Unit Hilbert Sphere. We generate $N = 200$ data points lying on the unit Hilbert sphere in $\mathbb{R}^{100}$. We ensure the intrinsic



Reduced Dimension $Q = 1$ Reduced Dimension $Q = 2$

Fig. 2. Nonlinear Dimensionality Reduction on Simulated Data. The performance for the proposed kPGA is in blue and that for the standard kPCA is in red. The horizontal axis shows values of κ in the κ neighborhood [36]. The quality measure on the vertical axis indicates the preservation of κ-sized neighborhoods based on distances in the original space (see text). For a fixed number of reduced dimensions Q, the ideal performance is a quality measure of 1 for all κ.

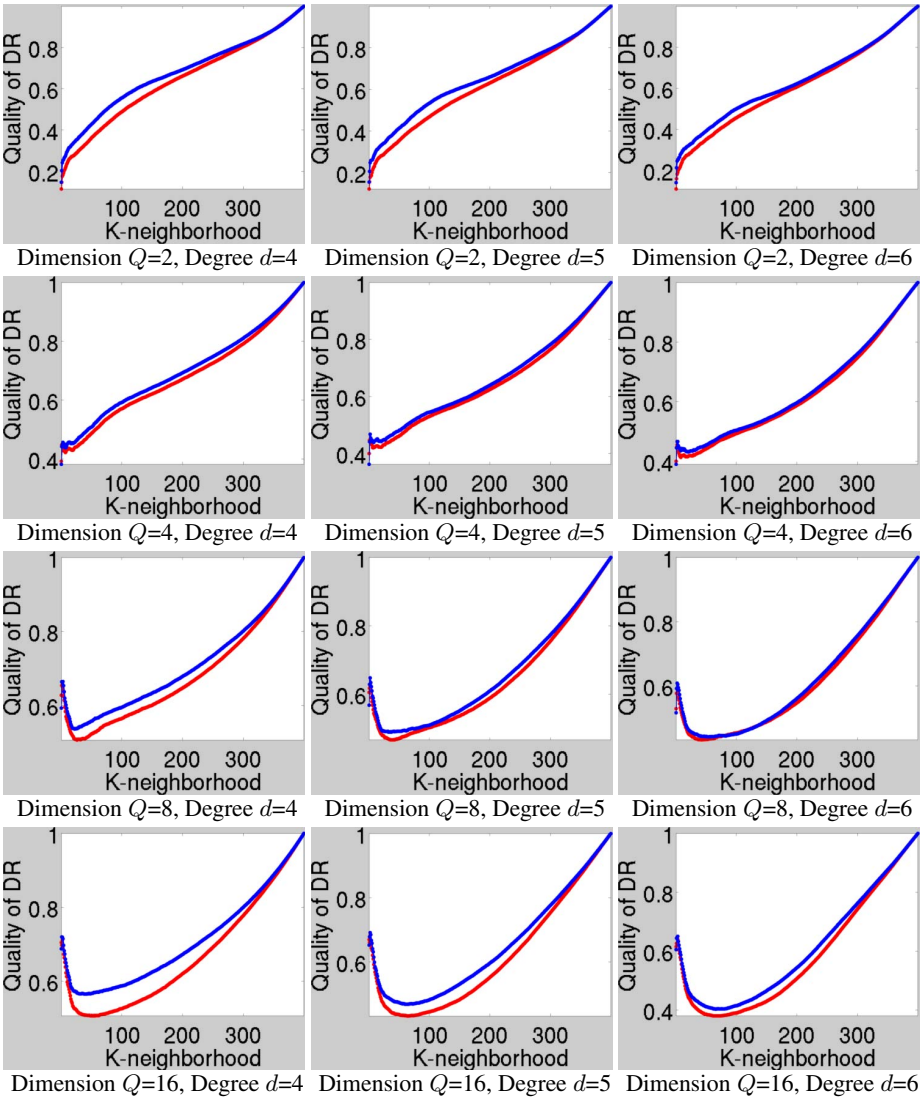


Fig. 3. Nonlinear Dimensionality Reduction on ORL Face Images. The **blue** curves represent the **proposed kPGA** and the **red** curves represent **standard kPCA**. Each subfigure plots quality measures (on vertical axis) for reduced-dimension values $Q = 2, 4, 8, 16$ and polynomial-kernel-parameter values $d = 4, 5, 6$. Within each subfigure (on horizontal axis), $\kappa = 1, \cdots, 399$. See Figure 4 for additional results with reduced-dimension values $Q = 32, 64, 128, 256$.

dimensionality of the dataset to be 2 by considering a subsphere $\mathbb{S}^2$ of dimension 2 and sampling points from a von Mises-Fisher distribution on $\mathbb{S}^2$ [37]. We set the kernel as $k(x, y) := \langle x, y \rangle$ that reduces the map $\Phi(\cdot)$ to identity (i.e., $\Phi(x) := x$) and, thereby, performs the analysis on the original data that lies on a Hilbert sphere in input space. Figure 2 shows the results of the dimensionality reduction using kPCA and kPGA.

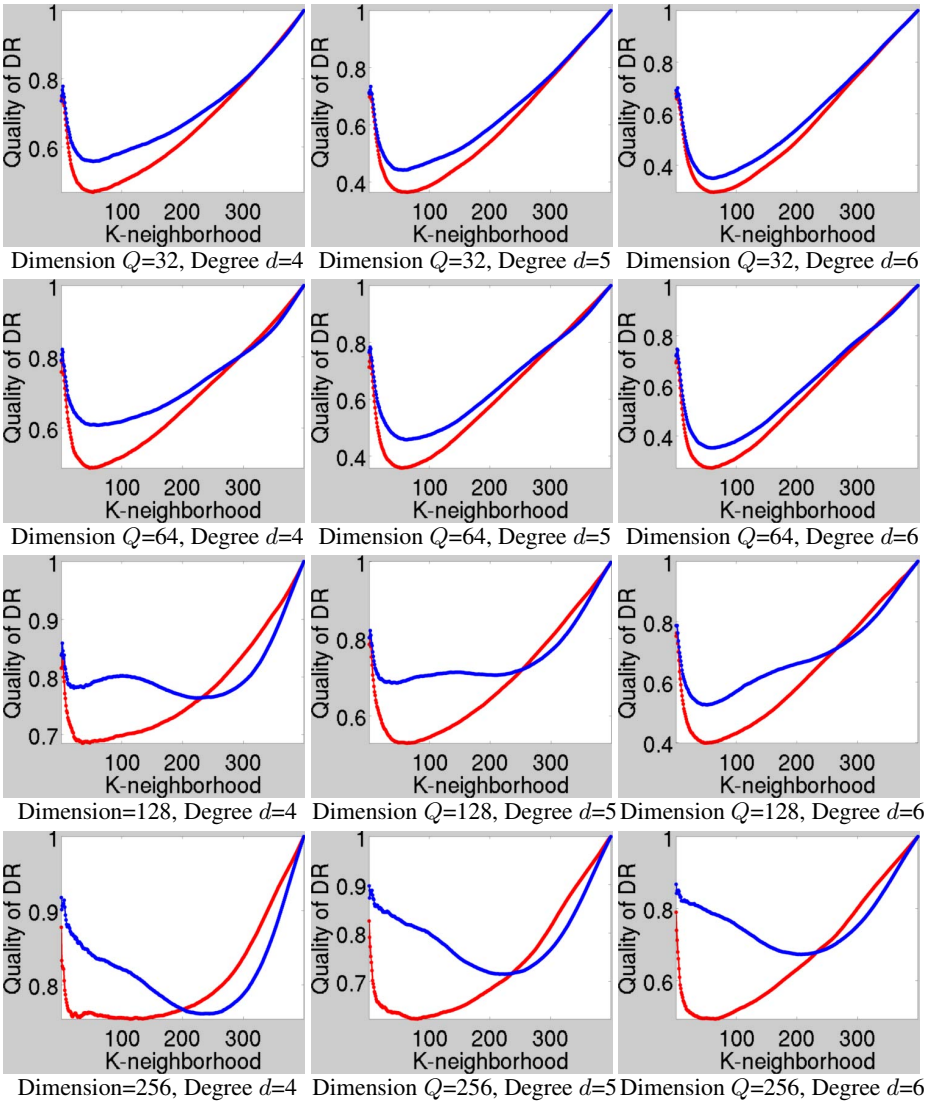


Fig. 4. Nonlinear Dimensionality Reduction on ORL Face Images. Continued from Figure 3.

When the reduced dimensionality is forced to be 1, which we know is suboptimal, both kPCA and kPGA perform comparably. However, when the reduced dimensionality is forced to 2 (which equals the intrinsic dimension of the data), then kPGA clearly outperforms kPCA; kPGA preserves the distance-based κ neighborhoods for almost every value of $\kappa \in \{1, \cdots, 199\}$. The result in Figure 2 is also consistent with the covariance eigenspectra produced by kPCA and kPGA. Standard kPCA, undesirably, gives 3 non-zero eigenvalues $(0.106, 0.0961, 0.0113)$ that reflect the dimensionality of the data representation for points on $\mathbb{S}^2$. On the other hand, the proposed kPGA gives

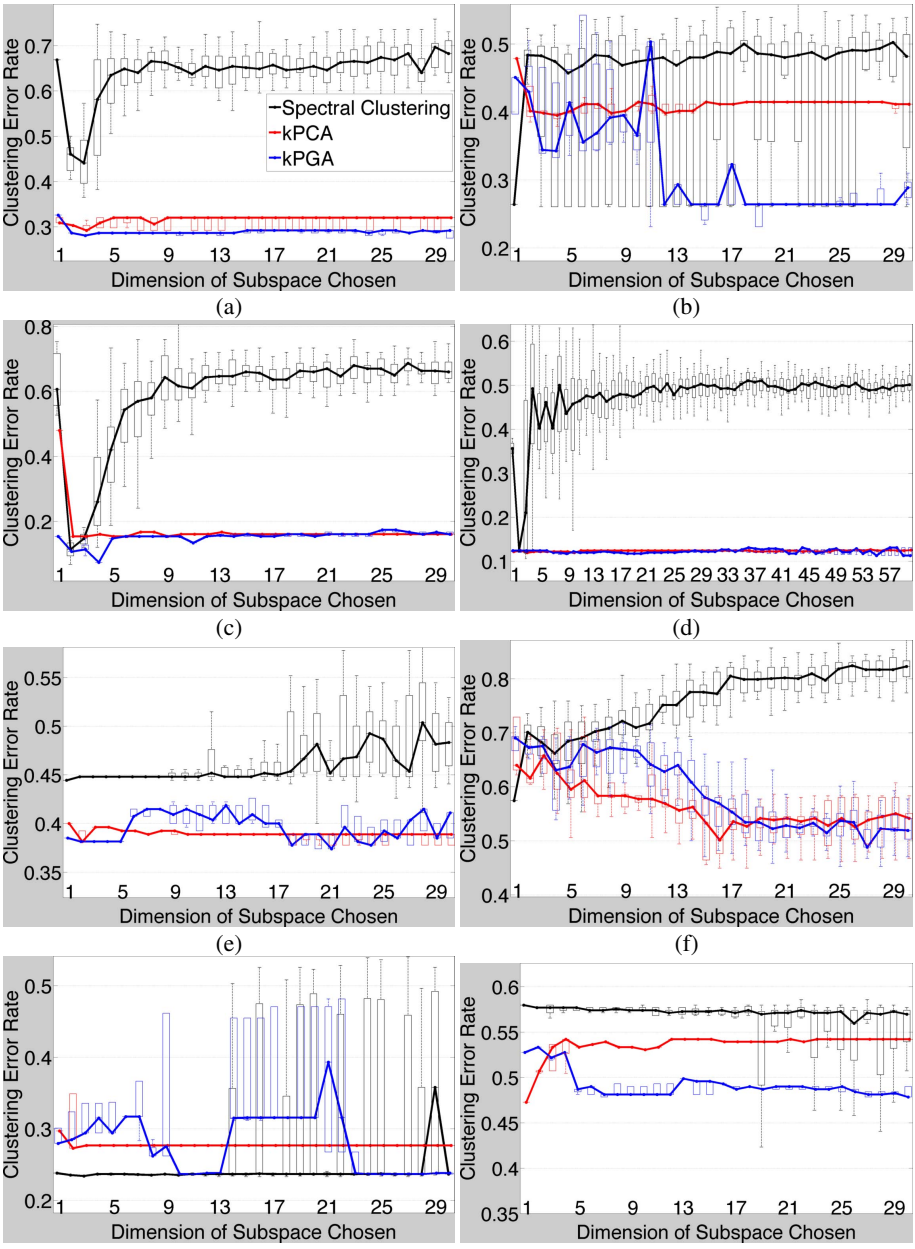


Fig. 5. Clustering on UCI Datasets. Box plots of error rates from clustering random subsets of the dataset. We use a Gaussian kernel. **(a)–(h)** show results on Wine, Haberman, Iris, Vote, Heart, Ecoli, Blood, and Liver datasets, respectively.

only 2 non-zero eigenvalues $(0.1246, 0.1211)$ that reflect the intrinsic dimension of the data. Thus, kPGA needs fewer components/dimensions to represent the data.

Real-World Data – ORL Face Image Database. The ORL database [44] comprises $N = 400$ face images of size 112×92 pixels. To measure image similarity, a justifiable kernel is the polynomial kernel $k(x, y) := (\langle x, y \rangle)^d$ after normalizing the intensities in each image x (i.e., subtract mean and divide by standard deviation) so that $\langle x, x \rangle = 1 = k(x, x)$ [46]. Figure 3 and Figure 4 show the results of nonlinear dimensionality reduction using standard kPCA and the proposed kPGA. For a range of values of the reduced dimension (i.e., $2, 4, 8, 16, 32, 64, 128, 256$) and a range of values of the polynomial kernel degree d (i.e., $d = 4, 5, 6$), the proposed kPGA outperforms standard kPCA with respect to the κ-neighborhood based quality measure.

6.2 Clustering Using Mixture Modeling and Expectation Maximization

We use the UCI repository to evaluate clustering in RKHS. Interestingly, for all but 2 of the UCI datasets used in this paper, the number of modes in kPCA (using the Gaussian kernel) capturing 90% of the spectrum energy ranges from 3–15 (mean 8.5, standard deviation 4.5). For only 2 datasets is the corresponding number of modes more than 20. This number is usually close to the intrinsic dimension of the data.

Real-World Data – UCI Machine Learning Repository. We evaluate clustering algorithms by measuring the error rate in the assignments of data points to clusters; we define error rate as the fraction of the total number of points in the dataset assigned to the incorrect cluster. We evaluate clustering error rates on a wide range of subspace dimensions $Q \in \{1, \cdots, 30\}$. For each Q, we repeat the following process 50 times: we randomly select 70% points from each cluster, run the clustering algorithm, and compute the error rate. We use the Gaussian kernel $k(x_i, x_j) = \exp(-0.5\|x_i - x_j\|_2^2/\sigma^2)$ and set σ^2, as per convention, to the average squared distance between all pairs (x_i, x_j).

Figures 5 compares the performance of spectral clustering [49], standard kPCA, and the proposed kPGA. In Figures 5(a)–(f), kPGA gives the lowest error rates (over all Q) and outperforms spectral clustering. In Figures 5(a)–(d), kPGA performs better or as well for almost all choices of Q. In Figure 5(g), kPGA performs as well as spectral clustering (over all Q). In Figure 5(h), kPGA performs slightly worse than kPCA (over all Q), but kPGA performs the best whenever $Q > 2$.

7 Conclusion

This paper addresses the hyperspherical geometry of points in kernel feature space, which naturally arises from many popular kernels and kernel normalization. This paper proposes kPGA to perform PGA on the Hilbert sphere manifold in RKHS, through algorithms for computing the sample weighted Karcher mean and the eigenvalues and eigenfunctions of the sample weighted Karcher covariance. It leverages kPGA to propose methods for (i) nonlinear dimensionality reduction and (ii) clustering using mixture-model fitting on the Hilbert sphere in RKHS. The results, on simulated and real-world data, show that kPGA-based methods perform favorably with their kPCA-based analogs.

References

1. Afsari, B.: Riemannian L^p center of mass: Existence, uniqueness, and convexity. Proc. Am. Math. Soc. 139(2), 655–673 (2011)
2. Afsari, B., Tron, R., Vidal, R.: On the convergence of gradient descent for finding the Riemannian center of mass. SIAM J. Control and Optimization 51(3), 2230–2260 (2013)
3. Ah-Pine, J.: Normalized kernels as similarity indices. In: Proc. Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining, vol. 2, pp. 362–373 (2010)
4. Ahn, J., Marron, J.S., Muller, K., Chi, Y.Y.: The high-dimension, low-sample-size geometric representation holds under mild conditions. Biometrika 94(3), 760–766 (2007)
5. Amari, S., Nagaoka, H.: Methods of Information Geometry. Oxford Univ. Press (2000)
6. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. 68(3), 337–404 (1950)
7. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Log-Euclidean metrics for fast and simple calculus on diffusion tensors. Mgn. Reson. Med. 56(2), 411–421 (2006)
8. Bache, K., Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
9. Banerjee, A., Dhillon, I., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von Mises-Fisher distributions. J. Mach. Learn. Res. 6, 1345–1382 (2005)
10. Berger, M.: A Panoramic View of Riemannian Geometry. Springer (2007)
11. Berman, S.: Isotropic Gaussian processes on the Hilbert sphere. Annals of Probability 8(6), 1093–1106 (1980)
12. Bhattacharya, R., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds. I. Annals Stats. 31(1), 1–29 (2005)
13. Bhattacharya, R., Patrangenaru, V.: Large sample theory of intrinsic and extrinsic sample means on manifolds. II. Annals Stats. 33(3), 1225–1259 (2005)
14. Blanchard, G., Bousquet, O., Zwald, L.: Statistical properties of kernel principal component analysis. Machine Learning 66(3), 259–294 (2007)
15. Boothby, W.M.: An introduction to differentiable manifolds and Riemannian geometry, vol. 120. Academic Press (1986)
16. Bühlmann, P., Van De Geer, S.: Statistics for high-dimensional data: methods, theory and applications. Springer (2011)
17. Buss, S., Fillmore, J.: Spherical averages and applications to spherical splines and interpolation. ACM Trans. Graph. 20(2), 95–126 (2001)
18. Carter, K., Raich, R., Hero, A.: On local intrinsic dimension estimation and its applications. IEEE Trans. Signal Proc. 58(2), 650–663 (2010)
19. Charlier, B.: Necessary and sufficient condition for the existence of a Frechet mean on the circle. ESAIM: Probability and Statistics 17, 635–649 (2013)
20. Cherian, A., Sra, S., Banerjee, A., Papanikolopoulos, N.: Jensen-Bregman logDet divergence with application to efficient similarity search for covariance matrices. IEEE Trans. Pattern Anal. Mach. Intell. 35(9), 2161–2174 (2012)
21. Courty, N., Burger, T., Marteau, P.-F.: Geodesic analysis on the Gaussian RKHS hypersphere. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part I. LNCS, vol. 7523, pp. 299–313. Springer, Heidelberg (2012)
22. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. Royal Statistical Society B(39), 1–38 (1977)
23. Eigensatz, M.: Insights into the geometry of the Gaussian kernel and an application in geometric modeling. Master thesis. Swiss Federal Institute of Technology (2006)
24. Felsberg, M., Kalkan, S., Krueger, N.: Continuous dimensionality characterization of image structures. Image and Vision Computing 27(6), 628–636 (2009)

25. Fletcher, T., Lu, C., Pizer, S., Joshi, S.: Principal geodesic analysis for the study of nonlinear statistics of shape. IEEE Trans. Med. Imag. 23(8), 995–1005 (2004)
26. Gonzalez, T.: Clustering to minimize the maximum intercluster distance. Theor. Comp. Sci. 38, 293–306 (1985)
27. Graf, A., Smola, A., Borer, S.: Classification in a normalized feature space using support vector machines. IEEE Trans. Neural Networks 14(3), 597–605 (2003)
28. Grauman, K., Darrell, T.: The pyramid match kernel: Efficient learning with sets of features. Journal of Machine Learning Research 8, 725–760 (2007)
29. Hein, M., Audibert, J.Y.: Intrinsic dimensionality estimation of submanifolds in R^d. In: Int. Conf. Mach. Learn., pp. 289–296 (2005)
30. Hoyle, D.C., Rattray, M.: Limiting form of the sample covariance eigenspectrum in PCA and kernel PCA. In: Int. Conf. Neural Info. Proc. Sys. (2003)
31. Kakutani, S., et al.: Topological properties of the unit sphere of a Hilbert space. Proceedings of the Imperial Academy 19(6), 269–271 (1943)
32. Karcher, H.: Riemannian center of mass and mollifier smoothing. Comn. Pure Appl. Math. 30(5), 509–541 (1977)
33. Kendall, W.S.: Probability, convexity and harmonic maps with small image I: uniqueness and fine existence. Proc. Lond. Math. Soc. 61, 371–406 (1990)
34. Krakowski, K., Huper, K., Manton, J.: On the computation of the Karcher mean on spheres and special orthogonal groups. In: Proc. Workshop Robotics Mathematics, pp. 1–6 (2007)
35. Lawrence, N.: Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J. Mach. Learn. Res. 6, 1783–1816 (2005)
36. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 72, 1432–1433 (2009)
37. Mardia, K., Jupp, P.: Directional Statistics. Wiley (2000)
38. Mas, A.: Weak convergence in the function autoregressive model. J. Multiv. Anal. 98, 1231–1261 (2007)
39. Nielsen, F., Bhatia, R.: Matrix Information Geometry. Springer (2013)
40. Peel, D., Whiten, W., McLachlan, G.: Fitting mixtures of Kent distributions to aid in joint set identification. J. Amer. Stat. Assoc. 96, 56–63 (2001)
41. Pennec, X.: Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. J. Mathematical Imaging and Vision 25(1), 127–154 (2006)
42. Raginsky, M., Lazebnik, S.: Estimation of intrinsic dimensionality using high-rate vector quantization. In: Proc. Adv. Neural Information Processing Systems, pp. 1–8 (2005)
43. de Ridder, D., Kuoropteva, O., Okun, O., Pietikainen, M., Duin, R.: Supervised locally linear embedding. In: Kaynak, O., Alpaydın, E., Oja, E., Xu, L. (eds.) ICANN/ICONIP 2003. LNCS, vol. 2714, pp. 333–341. Springer, Heidelberg (2003)
44. Samaria, F., Harter, A.: Parameterisation of a stochastic model for human face identification. In: Proc. IEEE Workshop on Applications of Computer Vision, pp. 138–142 (1994)
45. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Trans. Computers 18(5), 401–409 (1969)
46. Scholkopf, B., Smola, A.: Learning with Kernels. MIT Press (2002)
47. Scholkopf, B., Smola, A., Muller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation 10, 1299–1319 (1998)
48. Shawe-Taylor, J., Williams, C., Cristianini, N., Kandola, J.: On the eigenspectrum of the Gram matrix and the generalisation error of kernel PCA. IEEE Trans. Info. Th. 51(7), 2510–2522 (2005)

49. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22(8), 888–905 (2000)
50. Sommer, S., Lauze, F., Hauberg, S., Nielsen, M.: Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part VI. LNCS, vol. 6316, pp. 43–56. Springer, Heidelberg (2010)
51. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
52. Walder, C., Schölkopf, B.: Diffeomorphic dimensionality reduction. In: Int. Conf. Neural Info. Prof. Sys., pp. 1713–1720 (2008)
53. Wang, J., Lee, J., Zhang, C.: Kernel trick embedded Gaussian mixture model. In: Gavaldá, R., Jantke, K.P., Takimoto, E. (eds.) ALT 2003. LNCS (LNAI), vol. 2842, pp. 159–174. Springer, Heidelberg (2003)