2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Robust Kernel Principal Nested Spheres

Suyash P. Awate, Manik Dhar, Nilesh Kulkarni
Computer Science and Engineering Department, Indian Institute of Technology (IIT) Bombay.

***Abstract*—Kernel principal component analysis (kPCA) learns nonlinear modes of variation in the data by nonlinearly mapping the data to kernel feature space and performing (linear) PCA in the associated reproducing kernel Hilbert space (RKHS). However, several widely-used Mercer kernels map data to a Hilbert sphere in RKHS. For such directional data in RKHS, linear analyses can be unnatural or suboptimal. Hence, we propose an alternative to kPCA by extending principal nested spheres (PNS) to RKHS without needing the explicit lifting map underlying the kernel, but solely relying on the kernel trick. It generalizes the model for the residual errors by penalizing the L^p norm / quasi-norm to enable robust learning from corrupted training data. Our method, termed robust kernel PNS (rkPNS), relies on the Riemannian geometry of the Hilbert sphere in RKHS. Relying on rkPNS, we propose novel algorithms for dimensionality reduction and classification (with and without outliers in the training data). Evaluation on real-world datasets shows that rkPNS compares favorably to the state of the art.**

I. Introduction

Principal component analysis (PCA) finds modes of variation of the data, residing in a linear space, as a set of orthogonal directions that maximize the variance of the data projected onto them. The orthogonal basis defines a sequence of nested subspaces that optimally capture increasing fractions of the total variance of the data. Kernel PCA (kPCA) [1] uses the kernel trick and implicitly maps the data to a higher-dimensional kernel feature space that is associated with a reproducing kernel Hilbert space (RKHS). Then, kPCA performs PCA on the implicitly mapped data in the RKHS. Linear modes of variation in the mapped space correspond to nonlinear modes of variation in the input space, thus enabling kPCA to yield a more compact model of the data variability.

Many Mercer kernels, e.g., radial basis function kernels, map the input data to a Hilbert sphere [1] in RKHS $\mathcal{F}$ (Figure 1). The spherical structure arises because for such kernels $k(\cdot,\cdot)$, any input datum x has a self similarity $k(x,x) = 1$. The kernel defines the inner product in $\mathcal{F}$, and thus, $\langle \Phi(x), \Phi(x) \rangle_{\mathcal{F}} = 1$, implying that all mapped points $\Phi(x)$ are at a unit distance from the origin in $\mathcal{F}$. The practice of defining normalized kernels $\widetilde{k}(x,x') := k(x,x')/\sqrt{k(x,x)k(x',x')}$, e.g., pyramid match kernel [2], also leads to $\widetilde{k}(x,x) = 1$. Polynomial kernels lead to constant self similarity when the input data have constant l^2 norm, e.g., in facial image analysis [1]. When the mapped data in RKHS is *directional* [3], linear analyses, like kPCA, can be unnatural or suboptimal.

We propose an alternative to kPCA for kernels that lead to directional data in RKHS, by performing a decomposition of the Hilbert sphere. Manifold-based statistical analysis [4]–[9] explicitly models data to reside in the lower-dimensional subspace of the ambient space, representing variability in the data more efficiently (fewer degrees of freedom restricted to the manifold) and improve post-processing performance. In this way, we extend kPCA to (i) define a more meaningful sequence of nested subspaces capturing variability of the mapped data on the Hilbert sphere in RKHS and (ii) represent equivalent variability using a subspace of smaller dimension leading to a more compact model of the data. Similarly, we extend principal nested spheres (PNS) [10] to the Hilbert sphere in RKHS; we do so without needing the explicit lifting map underlying the kernel, solely using the kernel trick.

Real-world data represented in a high-dimensional space exhibits a small intrinsic dimension. kPCA attempts to capture these modes of variation via the principal eigenvectors of the implicitly mapped data in RKHS. Theoretically, PCA of directional data will typically introduce one additional / unnecessary principal mode of variation, along a radial direction and proportional to the sectional curvature. In practice, however, the unstable and erroneous behavior of PCA in high-dimensional spaces [11] interacts with the curvature of the Hilbert sphere on which the data resides. This often leads to a much poorer performance of PCA on high-dimensional directional data, contrary to our expectations on low-dimensional directional data. This behavior is evident in our empirical results, where our spherical analysis in RKHS yields larger gains than those expected in low dimensions.

Many applications involve corrupted data exhibiting weak signals, high noise, or missing values. An example is image recognition in scenarios where the visibility is low, e.g., at night or underwater, or there exist occluders [12]. In such cases, robust methods for training and recognition are vital. These typically use robust penalties (e.g., Huber loss) that penalize the L^p ($p < 2$) norm of residuals (resulting after the model fit) to reduce effect of outliers in the learning. They typically use iterative optimization that is costlier than eigen analysis in kPCA. We incorporate such a robust penalty during model learning and show that this can be achieved for PNS on the Hilbert sphere in RKHS solely using the kernel trick.

In this paper, we propose new formulations and algorithms to perform PNS on the Hilbert sphere in RKHS, without needing the explicit lifting map, but using the kernel trick. We generalize the model for the residual errors underlying PNS to enable robust learning from corrupted training data. We use our method, termed robust kernel PNS (rkPNS), for dimensionality reduction and classification. We evaluate the quality of model compactness, dimensionality reduction, and classification on real-world datasets and demonstrate advantages of rkPNS over the state of the art, including robust kPCA.

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

II. Related Work

Riemannian statistics has become an important tool for data analysis [5], [7]–[9], [13]–[15], e.g., for data lying on the manifolds of orthogonal matrices, symmetric positive definite matrices, hyperspheres, Grassmann manifold, and shape space. Some extensions of PCA to manifold-valued data rely on principal geodesic analysis (PGA) [6], [16]. Our rkPNS relies on the Riemannian geometry of the Hilbert sphere in RKHS, especially (i) the geodesic distance between two points, which is the arc cosine of their inner product, and (ii) the existence and formulation of tangent spaces [17]–[19].

Many RKHSs being infinite dimensional brings up concerns associated with statistical analysis in such spaces [20], [21]. Indeed, these same concerns arise in kPCA, and other well-known kernel methods, and thus the justification for this work is similar. First, we may assert that the covariance operator of the mapped data is of trace class or, more strongly, restricted to a finite-dimensional manifold defined by the cardinality of the input data. Second, such methods are intended mainly for data analysis instead of statistical estimation, and, thus, we intentionally work in the subspace defined by the sample size.

Modeling a probability density function (PDF) on a sphere entails fundamental trade-offs between model generality and the viability of the underlying parameter estimation. For example, although Fisher-Bingham PDFs on $\mathbb{S}^d$ can model anisotropic distributions (anisotropy around the mean) using $O(d^2)$ parameters, their parameter estimation may be intractable [3], [22]. On the other hand, parameter estimation for the $O(d)$-parameter von Mises-Fisher PDF is tractable [22], but it only models isotropic distributions. In contrast, rkPNS captures the modes of variation via a sequence of hyperspherical submanifolds with decreasing intrinsic dimension. We show that this is tractable on Hilbert spheres in RKHS, using the kernel trick. rkPNS differs from PNS by avoiding an explicit representation of the mapped points $\Phi(x)$.

Algorithms for robust kPCA appear in the recent literature [23]–[26], but *all* assume the mapped data to lie in a linear space, and all would ignore the Hilbert-sphere structure of the mapped data. Similar to our rkPNS, [24], [25] introduce a robust penalty on the residual and describe iterative optimization algorithms. Unlike our method, the spherical-kPCA, projection-pursuit, and Stahel-Donoho outlyingness based algorithms in [23] are *not* motivated as optimization problems and have algorithm components that are heuristic. The randomized algorithm in [26] can be time consuming because it repeats kPCA a number of times proportional to the sample size. Inspired by previous works on manifold-based data analyses [5], [7]–[9], [13], [15], we find that our robust kernel-based method exploiting the Hilbert-sphere structure of the data leads to advantages over linear analyses.

III. Geometry of the Hilbert Sphere in RKHS

We focus on RKHSs of infinite dimension related to popular kernels. Similar theory holds for other important kernels where the RKHS dimension is finite.

Let X be a random variable taking values x in *input space*. 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 implicitly maps x to $\Phi(x) := k(\cdot, x)$ in a RKHS $\mathcal{F}$ [1]. For vectors in RKHS represented as a linear combination of the mapped input points, i.e., $f := \sum_{i=1}^I \alpha_i \Phi(x_i)$ and $f' := \sum_{j=1}^J \beta_j \Phi(x_j)$, the inner product $\langle f, f' \rangle_{\mathcal{F}} := \sum_{i=1}^I \sum_{j=1}^J \alpha_i \beta_j k(x_i, x_j)$ and the norm $\|f\|_{\mathcal{F}} := \sqrt{\langle f, f \rangle_{\mathcal{F}}}$.

Let $Y := \Phi(X)$ be a random variable taking values y in RKHS, with $\{y_n := \Phi(x_n)\}_{n=1}^N$. Like kPCA [27], [28], we assume Y is bounded and the expectation and covariance operators of Y exist and are well defined. 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 rkPNS model applies to the Riemannian manifold of the unit Hilbert sphere [29], [30] in RKHS, centered at the origin.

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

$$\mathrm{Log}_b(a) = \frac{a - \langle a, b \rangle_{\mathcal{F}}\, b}{\|a - \langle a, b \rangle_{\mathcal{F}}\, b\|_{\mathcal{F}}} \arccos(\langle a, b \rangle_{\mathcal{F}}) \qquad (1)$$

that lies within the span of a and b and can be represented as $\sum_n \epsilon_n \Phi(x_n)$, where $\forall n : \epsilon_n \in \mathbb{R}$. $\mathrm{Log}_b(a)$ lies in the tangent space, at b, of the unit Hilbert sphere. The tangent space inherits the same structure (inner product) as the ambient space and, thus, is also a RKHS. The geodesic distance between a and b is $d_g(a,b) = \|\mathrm{Log}_b(a)\|_{\mathcal{F}} = \arccos(\langle a, b \rangle_{\mathcal{F}})$.

IV. Robust Kernel Principal Nested Spheres

We extend PNS to the Hilbert sphere in RKHS and uses a robust fitting term to deal with outliers in the data. Consider $\{y_m : \|y_m\|_{\mathcal{F}} = 1\}_{m=1}^M$ on the unit Hilbert sphere in RKHS, represented, in general, as $y_m := \sum_{n=1}^N \eta_{mn} \Phi(x_n)$.

A. Parameterizing Nested Hilbert Subspheres in RKHS

The rkPNS algorithm iteratively performs the following 2 steps: (i) fit a subsphere of one dimension lower than the dimension of the sphere on which the mapped data resides and (ii) project data on fitted subsphere. The resulting sequence of subspheres is nested, i.e., any fitted subsphere of dimension d lies completely within all fitted subspheres of dimension $> d$. The final subsphere of dimension 0 is defined as the Karcher mean of the projected data. The Karcher mean is unique if the support of the projected data is a proper subset of a half circle [31], [32], which is often observed in practice.

Although the data lie on an infinite-dimensional Hilbert sphere in RKHS, M unit-length data vectors can always be contained within a unit Hilbert subsphere isomorphic to $\mathbb{S}^{M-1}$. In rkPNS, the projected data on each fitted subsphere always lie in the span of the original mapped data. Because each projection reduces the intrinsic dimension of the data by one, the fitted subspheres are isomorphic to $\mathbb{S}^{M-2}, \mathbb{S}^{M-3}, \cdots, \mathbb{S}^1$. In this way, the number of iterations for the subsphere fitting is upper bounded by $M - 2$. This relates to modern applied

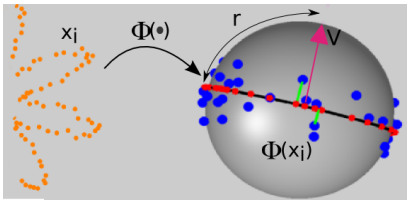

Fig. 1. **Robust Kernel Principal Nested Spheres.** Input datum x_i (**orange**) gets implicitly mapped to $\Phi(x_i)$ (**blue**) on the unit Hilbert sphere in RKHS, centered at the origin. A Hilbert subsphere $\mathcal{O}(v, r)$ (**black**) is parameterized by an axis v (**magenta**) orthogonal to itself and a geodesic distance r. In each iteration, rkPNS (i) fits a subsphere to the data and (ii) projects the data (**red**) onto the subsphere using the minimal geodesic (**green**) between them.

problems with data of high dimension (D) and low sample size ($N \ll D$) [33], where the data dimension can be reduced to $N-1$ without any loss of information.

We parameterize the Hilbert subsphere by (i) a unit-norm vector v that represents an axis orthogonal to the subsphere and (ii) the geodesic distance $r \in (0, \pi/2]$ between v and any point within the subsphere. Thus, the subsphere is $\mathcal{O}(v, r) := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} = 1, \|\mathrm{Log}_v(f)\|_{\mathcal{F}} = r\}$. Alternately, this subsphere is the intersection of the unit Hilbert sphere with the hyperplane $\{f \in \mathcal{F} : \langle f, v\rangle_{\mathcal{F}} = \cos(r)\}$, where v is orthogonal to the hyperplane. rkPNS leads to a sequence of nested subspheres $\mathcal{O}(v, r)$ parameterized by a corresponding sequence of axes that are mutually orthogonal and a sequence of corresponding radii. We represent the axis v as a linear combination of the implicitly-projected input data, i.e., $v := \sum_{m=1}^{M} \theta_m y_m$. Choosing v within the span of the data ensures that the axis v is orthogonal to the axes of every larger dimensional Hilbert subsphere that also contains the data.

B. Robust Fitting of Nested Subspheres in RKHS

Given points $\{y_m\}_{m=1}^{M}$, we formulate the problem of fitting a robust Hilbert subsphere as a constrained optimization problem in RKHS, which finds an axis $v := \sum_{m=1}^{M} \theta_m y_m$ and radius r that minimizes a robust penalty designed as a function of the geodesic distances between the subsphere $\mathcal{O}(v, r)$ and each datum $y_m := \sum_{n=1}^{N} \eta_{mn} \Phi(x_n)$; $\|y_m\|_{\mathcal{F}} = 1$. We propose the robust penalty as the p-th power of the L^p norm or quasi norm, where $0 < p \leq 2$, of the vector of residuals. Thus, we propose the best-fitting sphere to have $(r^*, \{\theta_m^*\}_{m=1}^{M}) :=$

$$\arg \min_{r, \{\theta_m\}_{m=1}^{M}} \mathcal{J}(\{y_m\}_{m=1}^{M}; r, \{\theta_m\}_{m=1}^{M})$$
$$\text{such that } r \in (0, \pi/2] \text{ and } \left\| \sum_{m=1}^{M} \theta_m y_m \right\|_{\mathcal{F}} = 1, \qquad (2)$$

where the objective function $\mathcal{J}(\{y_m\}_{m=1}^{M}; r, \{\theta_m\}_{m=1}^{M}) :=$

$$\sum_{m=1}^{M} \left(\left(\|\mathrm{Log}_v(y_m)\|_{\mathcal{F}} - r \right)^2 + \epsilon \right)^{p/2}, \qquad (3)$$

where p is a user-defined parameter (tuned via cross validation) and $\epsilon := 10^{-5}$ is used to regularize the L^p norm to make it smooth and amenable to gradient-based optimization. To solve this constrained optimization problem, we optimize the parameters v and r using projected gradient descent with step size found via line search; this guarantees convergence to a local minimum. We initialize v to a direction within the span of the projected data such that v minimizes the sum of squared distances, from the origin, of the projections of y_m onto the direction v (analogous to PCA).

Most importantly, $\|\mathrm{Log}_v(y_m)\|_{\mathcal{F}} = \arccos(\langle v, y_m\rangle_{\mathcal{F}}) = \arccos(\eta_m^\top G \eta \theta)$, where (i) η_m is the column vector with n-th element being η_{mn}, (ii) G is the Gram matrix where the element at row i and column j is $G_{ij} := \langle \Phi(x_i), \Phi(x_j)\rangle_{\mathcal{F}}$, (iii) η is the matrix with the m-th column as η_m, and (iv) θ is a column vector with the m-th element being θ_m. Thus, the objective function $\mathcal{J}(\{y_m\}_{m=1}^{M}; r, \{\theta_m\}_{m=1}^{M}) =$

$$\sum_{m=1}^{M} \left(\left(\arccos(\eta_m^\top G \eta \theta) - r \right)^2 + \epsilon \right)^{p/2}. \qquad (4)$$

This shows that the gradient of the objective function with respect to the variables r and $\{\theta_m\}_{m=1}^{M}$ solely requires the knowledge of Gram matrix G without needing the explicit mapping $\Phi(\cdot)$. Thus, we can perform the proposed subsphere fitting on the Hilbert sphere in RKHS using the kernel trick.

C. Projecting Data on the Fitted Subsphere

After a subsphere $\mathcal{O}(v, r)$ is fitted to $\{y_m\}_{m=1}^{M}$, each point y_m is projected onto $\mathcal{O}(v, r)$ by the following algorithm.

1) **Inputs:** Implicitly mapped points $\{y_m\}_{m=1}^{M}$ contained within a Hilbert subsphere isomorphic to, say, $\mathbb{S}^{D-1}$, where $y_m := \sum_{n=1}^{N} \eta_{mn} \Phi(x_n)$. Fitted subsphere $\mathcal{O}(v, r)$ with axis $v := \sum_{m=1}^{M} \theta_m y_m$.
2) We project y_m onto $\mathcal{O}(v, r)$ to give $z_m := (y_m \sin(r) + v \sin(\arccos(\langle v, y_m\rangle_{\mathcal{F}}) - r))/\sin(\arccos(\langle v, y_m\rangle_{\mathcal{F}}))$ [10] that is representable as a linear combination of $\Phi(x_n)$.
3) To put all z_m on a unit Hilbert sphere centered at the origin, translate them by $-v\cos(r)$ and rescale by $1/\sin(r)$.
4) **Outputs:** Projected points $\{z_m := \sum_{n=1}^{N} \xi_{mn} \Phi(x_n)\}_{m=1}^{M}$ on unit Hilbert subsphere $\mathcal{O}(v, r)$ isomorphic to $\mathbb{S}^{D-2}$.

This projection requires solely the knowledge of the Gram matrix G, without needing the explicit mapping $\Phi(\cdot)$.

D. Algorithm for rkPNS

First, we consider the points $\{y_m\}_{m=1}^{M}$ in RKHS to be in general position [34], i.e., the points are *not* contained in any sphere isomorphic to $\mathbb{S}^{M-2}$. Then rkPNS does the following.

1) **Inputs:** Data $\{x_n\}_{n=1}^{N}$ in input space, with or without outliers, and the associated Gram matrix G. We do *not* need the lifting map $\Phi(\cdot)$ underlying the kernel.
2) Initialize count $i := M$. Let the implicitly mapped points be $y_m^i := y_m = \Phi(x_m), \forall m$.
3) *Fit* a subsphere $\mathcal{O}(v^i, r^i)$ to $\{y_m^i\}_{m=1}^{M}$, using gradient descent optimization in Section IV-B.
4) *Project* points $\{y_m^i\}_{m=1}^{M}$ onto the fitted subsphere $\mathcal{O}(v^i, r^i)$, using the algorithm in Section IV-C, to produce the projected points $\{y_m^{i-1}\}_{m=1}^{M}$ orthogonal to v^i.
5) Reduce the count $i \leftarrow i-1$. If $i > 2$, then repeat the fitting and projection (last 2 steps); otherwise, proceed.

6) Optimize for the Karcher mean μ in RKHS on the unit Hilbert sphere (i.e., $\|\mu\|_{\mathcal{F}} = 1$) represented as a linear combination $\sum_{m=1}^{M} \rho_m y_m^2$ of the projected points $\{y_m^2\}_{m=1}^M$ that lie on a subsphere isomorphic to $\mathbb{S}^1$, using the gradient descent algorithm described in [35]. As shown by [35], finding the Karcher mean only needs the Gram matrix G, without the need for the explicit map $\Phi(\cdot)$.
7) **Outputs:** A sequence of mutually orthogonal axes $v^M, \cdots, v^3$ and distances $r^M, \cdots, r^3$ representing a sequence of nested spheres $\mathcal{O}(v^M, r^M), \cdots, \mathcal{O}(v^3, r^3)$ isomorphic to $\mathbb{S}^{M-2}, \cdots, \mathbb{S}^1$. A Karcher mean μ.

When the points $\{y_m\}_{m=1}^M$ are contained in a unit Hilbert subsphere isomorphic to $\mathbb{S}^D$ where $D \leq M-2$, the nested subsphere sequence will be shorter because fewer projections will result in the data lying on the subsphere isomorphic to $\mathbb{S}^1$. Thus, the subspheres will be isomorphic to $\mathbb{S}^{D-1}, \cdots, \mathbb{S}^1$. In such cases, which are typical in practice, we alter the stopping criterion as follows. If we fit a subsphere to points lying on a Hilbert sphere isomorphic to $\mathbb{S}^1$, then the projected points will be identical to at most two possible points. This condition can be checked at each iteration and can be used to terminate the iterations. If met, we backtrack and find the Karcher mean for points on the Hilbert sphere isomorphic to $\mathbb{S}^1$.

We see that rkPNS guarantees the axes $v^M, \cdots, v^3$ to be mutually orthogonal. Clearly, v^M is orthogonal to v^{M-1} because v^{M-1} is defined to be in the span of the projected data $\{y_m^{M-1}\}_{m=1}^M$ that is orthogonal to v^M. Similarly, v^M is also orthogonal to v^{M-2} because v^{M-2} is within the span of $\{y_m^{M-2}\}_{m=1}^M$ that is, in turn, within the span of $\{y_m^{M-1}\}_{m=1}^M$ that is orthogonal to v^M. Thus, v^M being orthogonal to v^i implies that v^M is orthogonal to all v^j for $3 \leq j \leq i-1$. Extending the argument for v^M to other v^k, each v^k is orthogonal to all v^j for $3 \leq j \leq k-1$. The Karcher mean μ must lie in the span of the projected data $\{y_m^2\}_{m=1}^M$ on a Hilbert subsphere isomorphic to $\mathbb{S}^1$ [31], [35]. Hence, the Karcher mean lies in the span of the original $\{y_m\}_{m=1}^M$.

V. rkPNS for Dimensionality Reduction

We now propose a dimensionality-reduction algorithm.

1) **Inputs:** Data $\{x_n\}_{n=1}^N$ in input space, with or without outliers. Gram matrix G. Desired embedding dimension D.
2) Perform rkPNS using the algorithm in Section IV-D.
3) Apply a sequence of projections, as per Section IV-C, to the mapped data $\{y_n\}_{n=1}^N$ so that the projected data, say $\{y_n^{D+1}\}_{n=1}^N$, lies on a Hilbert subsphere isomorphic to $\mathbb{S}^D$.
4) Map the projected data in the tangent space at μ to give vectors $\{t_n := \mathrm{Log}_\mu(y_n^{D+1})\}_{n=1}^N$.
5) Perform PCA on the tangent space vectors $\{t_n\}_{n=1}^N$. Project each vector t_n on the D eigenvectors of the sample covariance matrix, producing D coordinates $u_n \in \mathbb{R}^D$.
6) **Outputs:** The transformed data $\{u_n \in \mathbb{R}^D\}_{n=1}^N$.

VI. rkPNS for Classification

We propose algorithms for classification using rkPNS. First, we propose an algorithm for training a classifier.

1) **Inputs:** For the Q classes (denoted by $q = 1, 2, \cdots, Q$), N_q sample points $\{x_{qn}\}_{n=1}^{N_q}$ for class q. Gram matrix G, for the pooled dataset, underlying a kernel such that all diagonal elements equal 1. Parameter $D \in \mathbb{N}$.
2) Pool all the data and perform rkPNS using the algorithm in Section IV-D. For each of the principal D subspheres $\mathcal{O}(v^3, r^3), \cdots, \mathcal{O}(v^{D+2}, r^{D+2})$ that capture most of the variation in the mapped data, compute the signed residual resulting from projecting each $\Phi(x_{qn})$ onto a subsphere $\mathcal{O}(v^d, r^d)$ and scale that by $\prod_{i=d+1}^{D+2} \sin(r^i)$ (accounting for different sizes of the D subspheres [10]) to give the feature $\{u_{qn} \in \mathbb{R}^D\}_{n=1}^{N_q}$ for point n in class q.
3) Learn a classifier $\mathcal{C}$ based on features $\{u_{qn} \in \mathbb{R}^D\}_{n=1}^{N_q}$ for each class q. We train Q one-versus-all linear support vector machine (SVM) classifiers [36].
4) **Outputs:** A sequence of nested spheres $\mathcal{O}(v^M, r^M), \cdots, \mathcal{O}(v^3, r^3)$, Karcher mean μ, classifier $\mathcal{C}$.

Now, we propose an algorithm for classifying unseen data.

1) **Inputs:** The Gram matrix G for the training data. The rkPNS model represented via a sequence of nested spheres $\mathcal{O}(v^M, r^M), \cdots, \mathcal{O}(v^3, r^3)$ and the Karcher mean μ. Parameter D and classifier $\mathcal{C}$. Test image x to be classified along with the extension of the Gram matrix (one row / column) for this test image's feature vector x, giving kernel similarity of the test image's feature vector x with all training image feature vectors.
2) Get feature u for the datum x to be classified, as done during training using the sequence of D nested subspheres.
3) Use classifier $\mathcal{C}$ to classify feature u into a class, say q'.
4) **Output:** Class q'.

VII. Results and Discussion

We evaluate the proposed (rk)PNS-based statistical analyses for data having the structure of a Hilbert sphere in RKHS, comparing them to standard linear analyses that ignore the spherical structure underlying the data.

We incorporate robustness in kPCA by treating kPCA as an optimization problem for finding the Karcher mean (in RKHS) and finding a set of orthogonal directions that maximize variance of the data projected on the direction vector through the mean. Then, instead of maximizing the variance, i.e., sum of squared distances between the mean and the projected points, we maximize the sum of p-th power of the distances,

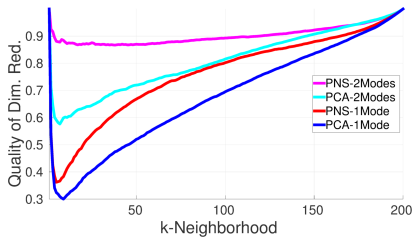

Fig. 2. **Toy Example with Simulated Data (PNS versus PCA).** For data $\{x_n\}_{n=1}^N$ drawn from a von-Mises Fisher distribution on $\mathbb{S}^2$, and using the kernel $k(x, x') = x^\top x'$, the quality of dimensionality reduction for rkPNS is far better than kPCA, for embedding dimensions 1 and 2.

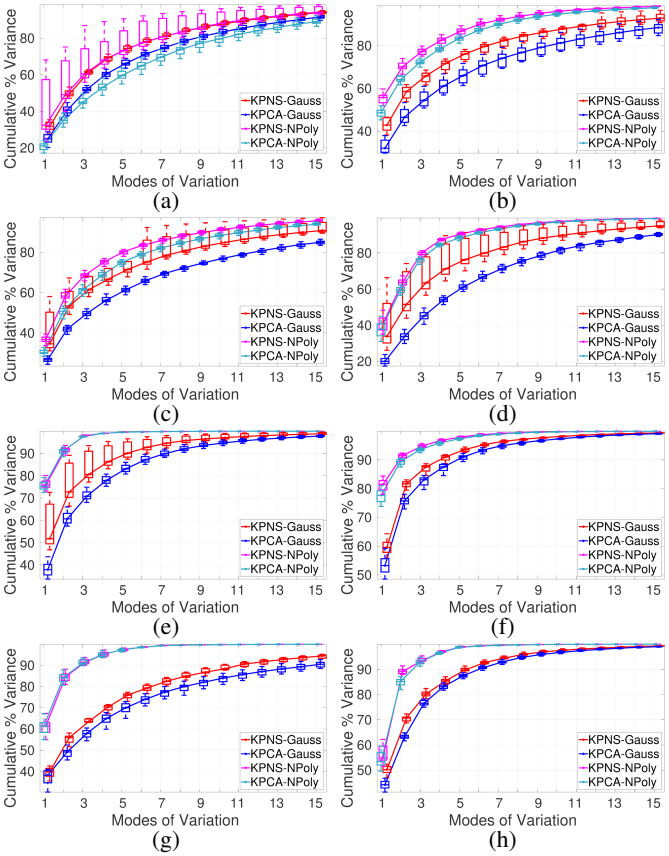

Fig. 3. **Model Compactness** captured via cumulative percent variances (= percentage of total variance explained by the chosen modes; this factors out differences in distance metrics) captured in the principal nested subspaces for UCI data [37]: **(a)** fertility. **(b)** vertebral column, **(c)** ecoli, **(d)** concrete slump test, **(e)** seeds, **(f)** iris, **(g)** glass identification, **(h)** haberman.

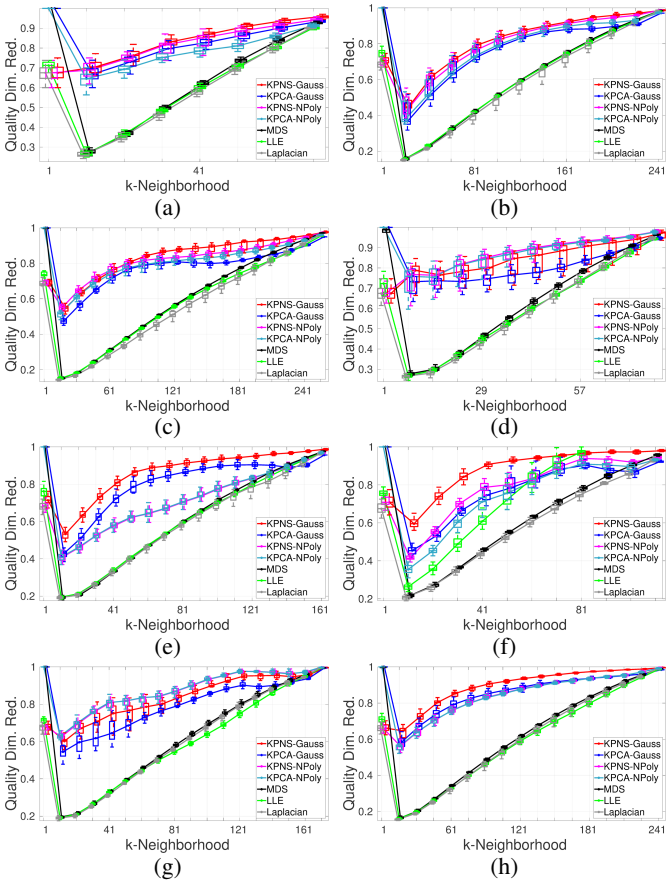

Fig. 4. **Dimensionality Reduction Quality** on UCI data [37] in the same order **(a)–(h)** as Figure 3.

where $0 < p \leq 2$ is a free parameter. We refer to this strategy as *robust kPCA* (rkPCA) in this section of the paper; this is similar to [24] with a certain influence function. If we fix $p = 2$, the proposed rkPNS and rkPCA reduce to kPNS (a subset of the proposed rkPNS) and kPCA, respectively.

To evaluate the performance of dimensionality reduction, we use the co-ranking matrix [38] 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 [38].

For each real-world dataset, we repeat the following process 25 times: we randomly select 80% data points, run all algorithms, and compute the quality metric. We evaluate rkPNS and rkPCA using (i) Gaussian kernel $k(x, x') := \exp(-0.5 \parallel x - x' \parallel^2 / \sigma^2)$, where we set σ^2 is set to the average squared distance between all pairs of points (x_i, x_j), and (ii) normalized version of the polynomial kernel $k(x, x') := (x^\top x')^q$, where we set $q := 10$. We find that these results are quite stable up to 30% perturbation in these parameter values.

Model Compactness. For rkPNS and rkPCA, we evaluate the cumulative percent variances captured in the nested subspheres / subspaces, respectively, associated with the principal modes of variation. For proof of concept of the utility of manifold-based analysis for directional data, we evaluate the methods for dimensionality reduction on simulated data, i.e., 200 points on $\mathbb{S}^2$ sampled from a von-Mises Fisher distribution, using the linear kernel $k(x, x') = x^\top x'$ that reduces kPNS to PNS and kPCA to PCA. We do *not* introduce outliers in the data and, hence, fix $p = 2$. On simulated data (Figure 2), for nested subspheres / subspaces of intrinsic dimension 1 and 2, PNS captured 85% and 100% of the total variance, respectively, while PCA captured only 37% and 70%. On UCI data [37] (Figure 3), compared to kPCA, kPNS typically captures a larger (never smaller) percentage of the variance for the same intrinsic dimension of the nested subspace.

Dimensionality Reduction. We compare rkPNS with rkPCA. On simulated data (Figure 2), PNS preserves the neighborhood structure better than PCA, using embedding dimensions of both 1 and 2. For UCI data (Figure 4), we choose the embedding dimension D to be the minimum, over all methods, of the intrinsic dimension of the nested subspace / subsphere that captures 70% of the total variance. Here, kPNS performs better (never worse) than kPCA. The results with locally linear embedding [39], multidimensional scaling [40],

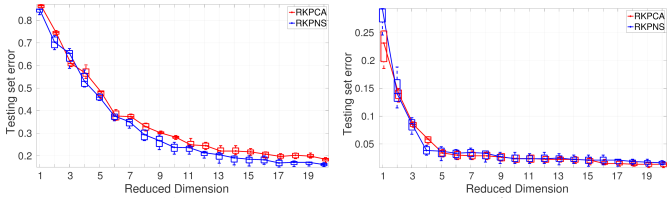

Fig. 5. **Classification.** Box plots of error rates over multiple trials for the **(a)** MNIST [42] and **(b)** Pen-Based [37] datasets.

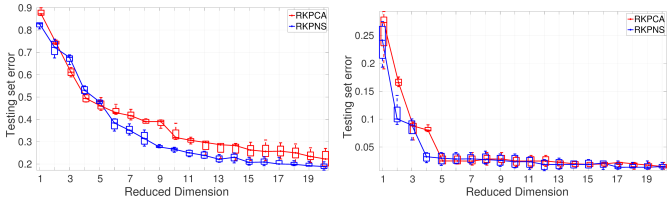

Fig. 6. **Classification (With Outliers).** Box plots of error rates over multiple trials for the **(a)** MNIST [42] and **(b)** Pen-Based [37] datasets.

and Laplacian eigenmaps [41], without using kernels, are just for context; their kernel versions are akin to kPCA [40].

Classification. We compare rkPNS and rkPCA for recognizing handwritten digits, on the MNIST dataset [42] and the Pen-Based dataset [37], over varying values of reduced-dimension parameter D (see Section VI); methods' performances will become similar for large D when *no* information is lost. Both these datasets have a small fraction of outliers inherently and, hence, we allow p to be less than 2; we tune the parameter p using 5-fold cross validation. In this case (Figure 5), we find that rkPNS often performs better then (or about as good as) rkPCA. In another experiment, we introduce outliers in both these datasets by reducing to zero the values in 20% of the randomly-chosen dimensions in the feature vector, i.e., pixel intensities in MNIST and attributes in the Pen-Based dataset. In this case (Figure 6), we find that the recognition error rates using rkPNS are almost always better than those from rkPCA when the reduced dimension is small; error rates often $5\% - 10\%$ lower for MNIST.

References

[1] B. Scholkopf and A. Smola, *Learning with Kernels*. MIT Press, 2002.

[2] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *JMLR*, vol. 8, pp. 925–60, 2007.

[3] K. Mardia and P. Jupp, *Directional Statistics*. Wiley, 2000.

[4] S. P. Awate and N. N. Koushik, "Robust dictionary learning on the Hilbert sphere in kernel feature space," in *Euro. Conf. Mach. Learn. Prac. Knowl. Disc. Data.*, vol. 1, 2016, pp. 1–18.

[5] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Jensen-Bregman logDet divergence with application to efficient similarity search for covariance matrices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2161–2174, 2012.

[6] T. Fletcher, C. Lu, S. Pizer, and S. Joshi, "Principal geodesic analysis for the study of nonlinear statistics of shape," *IEEE Trans. Med. Imag.*, vol. 23, no. 8, pp. 995–1005, 2004.

[7] M. Harandi, C. Sanderson, C. Shen, and B. Lovell, "Dictionary learning and sparse coding on Grassmann manifolds: An extrinsic solution," in *Int. Conf. Comp. Vis.*, 2013, pp. 3120–7.

[8] S. Sra, "A new metric on the manifold of kernel matrices with application to matrix geometric means," in *NIPS*, 2012, pp. 144–52.

[9] Y. Xie, J. Ho, and B. Vemuri, "On a nonlinear generalization of sparse coding and dictionary learning," *JMLR*, vol. 28, pp. 1480–1488, 2013.

[10] S. Jung, I. Dryden, and J. Marron, "Analysis of principal nested spheres," *Biometrika*, vol. 99, no. 3, pp. 551–568, 2012.

[11] J. Ahn, J. S. Marron, K. Muller, and Y.-Y. Chi, "The high-dimension, low-sample-size geometric representation holds under mild conditions," *Biometrika*, vol. 94, no. 3, pp. 760–766, 2007.

[12] J. Johnson and B. Olshausen, "The recognition of partially visible natural objects in the presence and absence of their occluders," *Vision Research*, vol. 45, pp. 3262–76, 2005.

[13] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Geometric means in a novel vector space structure on symmetric positive-definite matrices," *SIAM J. Matrix Analysis Appl.*, vol. 29, no. 1, pp. 328–47, 2007.

[14] F. Nielsen and R. Bhatia, *Matrix Information Geometry*. Springer, 2013.

[15] S. Sommer, F. Lauze, S. Hauberg, and M. Nielsen, "Manifold valued statistics, exact principal geodesic analysis and the effect of linear approximations." in *Proc. Euro. Conf. Comp. Vis.*, 2010, pp. 43–56.

[16] S. Huckemann and H. Ziezold, "Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces," *Adv. in Appl. Probab.*, vol. 38, no. 2, pp. 299–319, 2006.

[17] S. Berman, "Isotropic Gaussian processes on the Hilbert sphere," *Annals of Probability*, vol. 8, no. 6, pp. 1093–1106, 1980.

[18] W. M. Boothby, *An introduction to differentiable manifolds and Riemannian geometry*. Academic press, 1986, vol. 120.

[19] S. Kakutani *et al.*, "Topological properties of the unit sphere of a Hilbert space," *Proc. Imperial Acad.*, vol. 19, no. 6, pp. 269–271, 1943.

[20] P. Bühlmann and S. Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

[21] D. Hoyle and M. Rattray, "Limiting form of the sample covariance eigenspectrum in PCA and kernel PCA." in *NIPS*, 2003.

[22] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Mach. Learn. Res.*, vol. 6, pp. 1345–1382, 2005.

[23] M. Debruyne and T. Verdonck, "Robust kernel principal component analysis and classification," *Adv. Data Anal. Class.*, vol. 4, pp. 151–67, 2010.

[24] S. Huang, Y. Yeh, and S. Eguchi, "Robust kernel principal component analysis," *Neural Computation*, vol. 21, pp. 3179–213, 2009.

[25] M. Nguyen and F. Torre, "Robust kernel principal component analysis," in *NIPS*, 2008, pp. 1–8.

[26] H. Xu, C. Caramanis, and S. Mannor, "Outlier-robust PCA: The high dimensional case," *IEEE Trans. Info. Theory*, vol. 59, pp. 546–72, 2013.

[27] G. Blanchard, O. Bousquet, and L. Zwald, "Statistical properties of kernel principal component analysis," *Machine Learning*, vol. 66, no. 3, pp. 259–294, 2007.

[28] B. Scholkopf, A. Smola, and K.-R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[29] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford Univ. Press, 2000.

[30] M. Berger, *Panoramic View of Riemannian Geometry*. Springer, 2007.

[31] S. Buss and J. Fillmore, "Spherical averages and applications to spherical splines and interpolation," *ACM Trans. Graph.*, no. 2, pp. 95–126, 2001.

[32] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Comn. Pure Appl. Math.*, vol. 30, no. 5, pp. 509–41, 1977.

[33] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *J. R. Statist. Soc. B*, vol. 67, no. 3, pp. 427–44, 2005.

[34] A. Onishchik and R. Sulanke, *Projective and Cayley-Klein Geometries*. Springer, 2006.

[35] S. P. Awate, Y.-Y. Yu, and R. T. Whitaker, "Kernel principal geodesic analysis," in *Proc. Euro. Conf. Machine Learning and Knowledge Discovery in Databases*, vol. 8724, 2014, pp. 82–98.

[36] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[37] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[38] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomp.*, vol. 72, pp. 1432–33, 2009.

[39] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–26, 2000.

[40] J. Ham, D. Lee, S. Mika, and B. Scholkopf, "A kernel view of the dimensionality reduction of manifolds," in *ICML*, 2004, pp. 47–54.

[41] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *NIPS*, 2001, pp. 586–691.

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, 1998.