Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 00 (2010) 1–12
Neurocomputing

An Iterative Algorithm for Robust Kernel Principal Component Analysis

Hsin-Hsiung Huang[a], Yi-Ren Yeh[b]

[a]*The H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, US*
[b]*Research Center for Information Technology Innovation, Academia Sinica , Taipei, Taiwan*

Abstract

We introduce a technique to improve iterative kernel principal component analysis (KPCA) robust to outliers due to undesirable artifacts such as noises, alignment errors, or occlusion. The proposed iterative robust KPCA (rKPCA) links the iterative updating and robust estimation of principal directions. It inherits good properties from these two ideas for reducing the time complexity, space complexity, and the influence of these outliers on estimating the directions of principal components. In the asymptotic stability analysis, we also show that our iterative rKPCA converges to the weighted kernel principal kernel components from the batch rKPCA. Experimental results are presented to confirm that our iterative rKPCA achieves the robustness as well as time saving better than batch KPCA.

Keywords:
Kernel principal component analysis, iterative update, outliers, robust estimation.

1. Introduction

Principal component analysis (PCA) is a classical dimension reduction method which has been applied in many applications, such as data visualization, image reconstruction, biomedical study, etc. It is a linear transformation that searches for an orthonormal basis of a low-dimensional subspace, so called the principal components subspace, which explains the variability of the data as much as possible [1]. PCA's utility and success stem from the simplicity of the method that calculates the eigenvectors and eigenvalues of the sample covariance matrix of the data set.

While in some cases, a linear transformation is not suitable for capturing the nonlinear structures of the data, in order to represent the nonlinear structure, the kernel PCA (KPCA) has been formulated [2] in a reproducing kernel Hilbert space framework. In KPCA, the computational cost depends on the sample size. When the sample size is very large, it is impractical to compute the principal components via a direct eigenvalue decomposition. Given a matrix $X = [x_1, x_2, \cdots, x_m] \in \mathbb{R}^{n\times m}$ of m observations in an n-dimensional space, the batch KPCA through direct eigenvalue decomposition on the covariance of the kernel matrix $K \in \mathbb{R}^{m\times m}$ needs $O(m^2)$ space and $O(m^3)$ computation time, where $K_{ij} = K(x_i, x_j)$ is a kernel function $K(x, u)$. In contrast, given l principal components, the iterative KPCA costs $O(m(n + l^2))$ time complexity and $O(m(n + l))$ space in each updating iteration [3].

Furthermore, the conventional PCA is quite sensitive to outliers. Several approaches to achieve robustness have been proposed in the literature [4]. It is also recognized that KPCA is sensitive to outliers, and the analysis based on the contaminated principal components can be misleading [5]. A robust mechanism to reduce the influence of outliers is introduced [5]. In taking the advantages of iterative updating and outliers resistance, in this article we aim for an iterative robust KPCA (rKPCA) algorithm, which updates the principal directions via each kernel datum and automatically assigns a weight to this datum for the robust estimation at the same time. We will also study its theoretical properties as well as plan for real data applications.

Email addresses: hhhuang@gatech.edu (Hsin-Hsiung Huang), yryeh@citi.sinica.edu.tw (Yi-Ren Yeh)

The following is the organization of this paper. In Section 2, we review the rKPCA based on a robust psi-weight function and the iterative KPCA which is the same as the kernel Hebbian algorithm (KHA) [6]. Then in Section 3, how to adapt the estimation of rKPCA using gradient descent is presented. We show that the mean and the principal components due to the iterative rKPCA are asymptotically stable. The asymptotic behavior of the eigenvectors of sample covariance matrices based on the gradient descent update is summarized in the same section as well. In Section 4, natural and artificial experiments are designed to confirm that the iterative rKPCA can reduce the influence of outliers of data. Conclusive remarks are in Section 5. In the appendix, mathematical proofs confirm the sufficient conditions of the asymptotic properties for the mean and kernel principal components obtained by the iterative rKPCA.

2. Related Previous Work

2.1. Robust Kernel Principal Component Analysis

Given a matrix $X \in \mathbb{R}^{n\times m}$ of observational data and a positive definite kernel function $K(x,u)$, the associated separable Hilbert space consists of the closure of all finite kernel mixtures $\sum_{i=1}^{\tilde{n}} a_i K(x,u_i)$, where $\tilde{n} \in \mathbb{N}, u_i \in \mathbb{R}^n, a_i \in \mathbb{R}$. The norm $\|v\|_{\mathcal{H}} = (\langle v,v\rangle_{\mathcal{H}})^{1/2}$ is induced by the inner product: $\langle K(x,\cdot), K(u,\cdot)\rangle_{\mathcal{H}} = K(x,u)$. This Hilbert space is known as an RKHS denoted by $\mathcal{H}$.

Assumed that $\mathcal{H}$ is separable and hence has a countable orthonormal (o.n.) basis set, denoted by $\{e_i\}_{i=1}^{\infty}$. Particularly, one may take $e_i = \sqrt{\pi_i}\,\phi_i$, where π_i's and ϕ_i's are the eigenvalues and the corresponding eigenfunctions of $K(x,u)$, and then the spectrum decomposition is $K(x,u) = \sum_i \pi_i\phi_i(x)\phi_i(u)$ with $\sum_i \pi_i < \infty$. Each input element x is mapped into an element $h_x = K(x,\cdot)$. Besides, each element $h \in \mathcal{H}$ can be expanded in terms of this o.n. system as

$$h = \sum_i a_i e_i, \text{ where } a_i = \langle h, e_i\rangle_{\mathcal{H}}. \tag{1}$$

Let O_l be the collection of all rank-l linear operators of the form $\Gamma = \sum_{i=1}^{l} \gamma_i \otimes e_i$ with $\{\gamma_i\}_{i=1}^{l}$ being orthonormal in $\mathcal{H}$ where $\otimes$ denotes tensor product. Its transpose can be expressed as $\Gamma^\top = \sum_{i=1}^{l} e_i \otimes \gamma_i$. For $\Gamma \in O_l$, let $\mathcal{P}_\Gamma$ and $\mathcal{P}_{\Gamma^\perp}$ denote, respectively, the orthogonal projections onto the subspace $\mathrm{Im}(\Gamma) = \mathrm{span}\{\gamma_1,\ldots,\gamma_l\}$ and its orthogonal complement $\mathrm{Im}(\Gamma^\perp)$. For $\mu \in \mathcal{H}$ and $\Gamma \in O_l$, define

$$\begin{aligned} z(h_x,\mu,\Gamma) &= \frac{1}{2}\left\{\|h_x-\mu\|_{\mathcal{H}}^2 - \|\Gamma^\top(h_x-\mu)\|_{\mathcal{H}}^2\right\} \\ &= \|\mathcal{P}_{\Gamma^\perp}(h_x-\mu)\|_{\mathcal{H}}^2. \end{aligned} \tag{2}$$

Note that $z(h_x,\mu,\Gamma)$ is the squared residual norm by fitting $h_x-\mu$ to the subspace $\mathrm{Im}(\Gamma)$. The KPCA of leading l principal components looks for $\Gamma \in O_l$ that satisfies the smallest expected residual norm. Assumed the data belong to distribution G, the KPCA solves the following minimization problem:

$$\underset{\mu\in\mathcal{H},\Gamma\in O_l,\Gamma^\top\Gamma=I}{\operatorname{argmin}} \mathrm{E}_G z(h_x,\mu,\Gamma), \tag{3}$$

where $h_x = K(x,\cdot)$ and $x \sim G$. Therefore the minimization problem for samples from a population belonging to distribution G is of the following form

$$\underset{\mu\in\mathcal{H},\Gamma\in O_l,\Gamma^\top\Gamma=I}{\operatorname{argmin}} \sum_{t=1}^{m} z(h_t,\mu,\Gamma) \tag{4}$$

where $h_t = K(x_t,\cdot)$ and x_t is iid from G. It can be shown that problems (3) and (4) lead to the solutions (μ,Γ), respectively, as

$$\mu = \mathrm{E}_G h_x \text{ and } \Gamma = \mathrm{eigen}_l(\Sigma), \tag{5}$$

where $\Sigma = \mathrm{E}_G\{(h_x-\mu)\otimes(h_x-\mu)\}$;

$$\hat{\mu} = \frac{1}{m}\sum_{t=1}^{m} h_t \text{ and } \hat{\Gamma} = \mathrm{eigen}_l(\hat{\Sigma}), \tag{6}$$

where $\hat{\Sigma} = \frac{1}{m}\sum_{t=1}^{m}(h_t-\hat{\mu})\otimes(h_t-\hat{\mu})$ and $\mathrm{eigen}_l(M)$ stands for the leading l eigenfunctions of M, denoted by γ_i, collected together to form the rank-l operator $\sum_{i=1}^{l}\gamma_i \otimes e_i \in O_l$. In other words, the KPCA solves for the leading eigenfunctions for a covariance operator Σ or $\hat{\Sigma}$.

However, the conventional KCPA formulation (4) is not robust [5]. For example, if data are polluted with intra-sample outliers, the estimated principal directions are affected by these outliers. As a result, principal components may not lead to greatest variances of unpolluted data. For mitigation of the influence of outliers, Huang et al. [5] proposed a robust version of iterative KPCA based on the the principle of robust loss function that enables the incorporation of robustness and is taken as a constrained minimization problem

$$\underset{\mu,\Gamma,\Gamma^\top\Gamma=I}{\operatorname{argmin}} \sum_{t=1}^{m} \Psi(z(h_t,\mu,\Gamma)), \tag{7}$$

where $\Psi(z)$ is a monotonic increasing function of z and it will result in assigning smaller weights to the outliers so that the influence of these outliers reduces. Here are a few choices of $\Psi(z)$ and corresponding unnormalized weight function $\dot{\Psi}(z) = \frac{d\Psi(z)}{dz}$:

- $\Psi_0(z) = z$ with $\dot{\Psi}_0(z) = 1$, which leads to the conventional KPCA.

- $\Psi_1(z) = (1 - e^{-\beta z})/\beta$ with $\dot{\Psi}_1(z) = e^{-\beta z}$.
- $\Psi_2(z) = -\frac{1}{\beta}\log\left\{\frac{1+e^{-\beta(z-\xi)}}{2}\right\}$ with $\dot{\Psi}_2(z) = 1 - \frac{1}{1+e^{-\beta(z-\xi)}}$.

2.2. Iterative Kernel Principal Component Analysis

We apply a gradient descent method to solve the minimization problem (4) iteratively which implements in the same way as the kernel Hebbian algorithm (KHA) [6]. The KHA applies the generalized Hebbian algorithm [7] to the kernel expansion of data and considers the following minimization problem to update the current μ_t and Γ_t via each arriving kernel datum h_t:

$$\underset{\mu_t,\Gamma_t,\Gamma_t^\top\Gamma_t=I}{\operatorname{argmin}} \; z(h_t,\mu_t,\Gamma_t). \tag{8}$$

By introducing the Lagrange multipliers Λ, the Lagrangian function of (8) can be expressed as follows:

$$\begin{aligned}\mathcal{L}(\mu_t,\Gamma_t,\Lambda_t) = \quad & z(h_t,\mu_t,\Gamma_t) + \sum_{i=1}^{l}\frac{1}{2}\lambda_{t_{ii}}(\|\gamma_{t_i}\|^2 - 1) \\ & + \sum_{i>j}^{l}\lambda_{t_{ij}}\langle\gamma_{t_i},\gamma_{t_j}\rangle_{\mathcal{H}}.\end{aligned} \tag{9}$$

where $\gamma_i(t)$'s are columns of Γ and $\Lambda_{t_{ij}} = \lambda_{t_{ij}}$. Taking the derivatives of (9) with respect to μ_t and Γ_t, we have the gradients of $\mathcal{L}(\mu_t,\Gamma_t,\Lambda_t)$ with respect to μ_t and Γ_t as follows:

$$\begin{aligned} g_t^{(\mu)} &= -(h_t - \mu_t), \qquad (10)\\ g_t^{(\Gamma)} &= -((h_t-\mu_t)\otimes(h_t-\mu_t))\Gamma_t - \Gamma_t\mathcal{UT}(\Lambda_t), \end{aligned}$$

where $\mathcal{UT}(\cdot)$ is the operator that picks the upper triangular entries of a matrix and Lagrange multipliers $\Lambda_t \in \mathbb{R}^{l\times l}$ are updated as follows

$$\begin{cases} \lambda_{t_{ii}} = \gamma_{t_i}^\top(h_t-\mu_t)(h_t-\mu_t)^\top\gamma_{t_i}, \\ \lambda_{t_{ij}} = \gamma_{t_i}^\top(h_t-\mu_t)(h_t-\mu_t)^\top\gamma_{t_j}. \end{cases} \tag{11}$$

These lead to iterative updating for $\mu \in \mathbb{R}^{m\times 1}$

$$\mu_{t+1} = \mu_t - \eta_t g_t^{(\mu)}, \tag{12}$$

and for $\Gamma \in \mathbb{R}^{m\times l}$

$$\Gamma_{t+1} = \Gamma_t - \eta_t g_t^{(\Gamma)}, \tag{13}$$

where $\eta_t = \frac{1}{t}$ is a learning rate at the t^{th} iteration.

3. Iterative Robust KPCA and Its Asymptotic Stability

When data come in a sequential manner, iterative algorithms are appropriate to be considered. We propose an iterative method for performing a robust kernel principal component analysis. For the sequential data, Higuchi et al. [8] propose an approach which applies a weight function to PCA when sequential data have outliers. Here, we extend this method into KPCA for sequential high-dimensional data. The dimension of kernel in KPCA is increasing as each observation is added. We assumes that there are only finite main features of these observations. Therefore, we can choose an arbitrary fixed basis of the kernel so that the size of the kernel is fixed. Thus the main issue is to find a way to choose the number of principal components to compress the data size, and to prove convergence of our proposed method.

3.1. Formulation of Iterative Robust KPCA

For linking the iterative updating and robust estimation of kernel principal directions, we consider the following single observation version of (7):

$$\underset{\mu_t,\Gamma_t,\Gamma_t^\top\Gamma_t=I}{\operatorname{argmin}} \; \Psi(z(h_t,\mu_t,\Gamma_t)). \tag{14}$$

Now we introduce Lagrange multipliers for the minimization problem subject to the orthonormal basis constraints. Then we have the Lagrangian function of (14):

$$\begin{aligned}\mathcal{L}(\mu_t,\Gamma_t,\Lambda_t) = \quad & \Psi(z(h_t,\mu_t,\Gamma_t)) + \sum_{i=1}^{l}\frac{1}{2}\lambda_{t_{ii}}(\|\gamma_{t_i}\|^2 - 1) \\ & + \sum_{i>j}^{l}\lambda_{t_{ij}}\langle\gamma_{t_i},\gamma_{t_j}\rangle_{\mathcal{H}}.\end{aligned} \tag{15}$$

The gradients of $\mathcal{L}(\mu_t,\Gamma_t,\Lambda_t)$ with respect to μ and Γ of (15) while receiving a kernel observation h_t are

$$\begin{aligned} g_t^{(\mu)} &= -\dot{\Psi}(z(h_t;\mu_t,\Gamma_t))(h_t-\mu_t), \qquad (16)\\ g_t^{(\Gamma)} &= -\dot{\Psi}(z(h_t;\mu_t,\Gamma_t))\{[(h_t-\mu_t)\otimes(h_t-\mu_t)]\Gamma_t \\ &\quad -\Gamma_t\mathcal{UT}(\Lambda_t)\}, \qquad (17)\end{aligned}$$

where $\dot{\Psi}(z)$ gives a smaller weight to reduce the effect of misleading updating from an outlier. Here the Lagrange multiplier $\lambda_{t_{ij}}$ in Λ_t is calculated by using the following Property 1.

Property 1. *In the solutions, $\lambda_{t_{ij}}$'s take the following values:*

$$\lambda_{t_{ij}} = \begin{cases} \dot{\Psi}(z(h_t,\mu_t,\Gamma_t))\gamma_{t_i}^\top(h_t-\mu_t)(h_t-\mu_t)^\top\gamma_{t_i}, & i = j\,, \\ \dot{\Psi}(z(h_t,\mu_t,\Gamma_t))\gamma_{t_i}^\top(h_t-\mu_t)(h_t-\mu_t)^\top\gamma_{t_j} & i < j\,. \end{cases} \tag{18}$$

With the gradient $g_t^{(\mu)}$ and $g_t^{(\Gamma)}$, the updating for $\mu \in \mathbb{R}^{\tilde{m}\times 1}$ in iterative rKPCA

$$\mu_{t+1} = \mu_t - \eta_t g_t^{(\mu)}, \tag{19}$$

and for $\Gamma \in \mathbb{R}^{\tilde{m}\times l}$

$$\Gamma_{t+1} = \Gamma_t - \eta_t g_t^{(\Gamma)}, \tag{20}$$

where the details of the entire updating procedure are described in Algorithm 1.

The initial mean μ_1 and principal directions Γ_1 in Algorithm 1 can be set either simply by $\mathbf{0}$ and δI with a small value δ or using an appropriate initial ones which can accelerate the convergence via prior-knowledge. For generating the kernel basis, the usual way is using the whole original data X to generate the kernel basis, and obtain a full kernel. However, calculating a full kernel might cause the curse of high dimensionality and increase the computational cost. A reduced kernel can be used to replace the full kernel to avoid these problems [9, 10]. The idea of reduced kernel is extracting A from a small portion of X for fast approximation. Also note that A can be selected independently from X. This method is useful for large data problems, when n goes to infinity.

3.2. Asymptotic Stability

When we gather a small number of observations, we can use batch KPCA to analyze them so that we can ensure that there are no outliers. Then, we can choose basis of kernel based on this small group. we show that in the iterative rKPCA μ_t converges to

$$\frac{\mathrm{E}_G[\dot{\Psi}(z(h_x,\mu,\Gamma))h_x]}{\mathrm{E}_G[\dot{\Psi}(z(h_x,\mu,\Gamma))]}, \tag{21}$$

and Γ_t converges to Γ whose columns are the first l eigenfunctions of

$$\frac{\mathrm{E}_G[\dot{\Psi}(z(h_x,\mu,\Gamma))(h_x-\mu)\otimes(h_x-\mu)]}{\mathrm{E}_G[\dot{\Psi}(z(h_x,\mu,\Gamma))]} \tag{22}$$

in descending eigenvalue order, respectively. Given the kernel basis at observations, if the initial values of μ and Γ are close to stable points, μ_t and Γ_t are asymptotically stable [11, 7]. If outliers are the first data in our algorithm, then our method may lead to principal components which are very different from batch KPCA. To ensure the asymptotically stability. Asymptotic stability of μ is proved by the solution of $\lim_{t\to\infty} g_t^{(\mu)}$. Now we state the convergence properties of the online rKPCA in Theorems 1 and 2 (we put the proofs in Appendix). The theorem in [7] is applied to prove the asymptotic stability of Γ.

Theorem 1 (Asymptotic Stability of μ_t). *In our proposed iterative rKPCA, μ_t converges to $\frac{E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))h_x]}{E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))]}$ almost surely.*

Theorem 2 (Asymptotic Stability of Γ_t). *Assume that the leading l eigenvalues of $\frac{E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))(h_x-\mu)\otimes(h_x-\mu)]}{E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))]}$ have unit multiplicity. Given a finite kernel basis set $\{K(\cdot,u_{t'})\}_{t'=1}^{\tilde{m}}$, when the initial Γ_0 is randomly assigned and the learning rate $\eta_t = \frac{1}{t}$, Γ_t in the iterative rKPCA converges to the matrix with columns of these l eigenvectors almost surely. Moreover, the updating system is locally asymptotically stable.*

We use Ljung's theorem [11] to prove that the updating systems for the mean and principal components are asymptotically stable. We should notice that the asymptotically stable values of μ and Γ are not equal to corresponding values resulted from batch KPCA of unpolluted data, but they are functions of the Ψ function that contains outliers. Though the mean and principal components resulted from the iterative rKPCA do not converge to those from the KPCA of the unpolluted data, since the Ψ assigns smaller weights on outliers, the mean and principal components are more closer to those from KPCA of the unpolluted data. Therefore they may lead to better data reconstruction than those resulted iterative KPCA directly. Another difficulty for the iterative rKPCA is that we cannot identify what are outliers when we only gather a few observations. Hence, if we collect outliers in the beginning of the iterative updating, the weight function will give lower weight on those really important data and thus lead to a biased results. Therefor. we propose a remedy approach is to obtain initial mean and principal components from batch KPCA when gathering a group of data in the beginning, then we start to do the iterative rKPCA. Moreover, we are supposed to know types of observations, so we can determine whether we can import data by random sampling or sequential neighbors.

4. Experimental Results

The behavior of iterative rKPCA is illustrated in four natural and synthetic experiments. The first example is a 2-dimensional synthetic example with polynomial kernel. The second one is the phoneme data, a functional data, that has three principal components as a result of the previous work [5]. In the third and fourth examples, we use wavelet kernels for multiresolution analysis [12, 13]. Level-1 Haar wavelet kernel and Order-4 Symlet wavelet kernel [14] are applied to the Lena picture and a series of face images respectively in order to show how the iterative rKPCA reduces the influence of outliers in image reconstruction. In our experiments, we insert artificial contaminated data

Input: Data matrix $X \in \mathbb{R}^{n\times m}$, a kernel function $K(x,u) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and data points $\{u_{t'} \in \mathbb{R}^{n\times 1}\}_{t'=1}^{\tilde{m}}$ for generating kernel basis, initial mean $\mu_1 \in \mathbb{R}^{\tilde{m}\times 1}$, and initial principal directions $\Gamma_1 \in \mathbb{R}^{\tilde{m}\times l}$.
Output: The resulting principal directions $\Gamma \in \mathbb{R}^{\tilde{m}\times l}$.
begin
  for $t = 1, 2, \ldots$ **do**
    $h_t \leftarrow [K(x_t, u_1) \cdots K(x_t, u_{\tilde{m}})]^\top$; // Calculate the kernel datum $h_t \in \mathbb{R}^{\tilde{m}\times 1}$
    $w_t \leftarrow \Psi(z(h_t, \mu_t, \Gamma_t))$; // update weights
    $\Lambda_t \leftarrow {\Gamma_t}^\top (h_t - \mu_t) \otimes (h_t - \mu_t)\Gamma_t$; // update the Lagrange multipliers
    $g_t^{(\mu)} \leftarrow -w_t(h_t - \mu_t)$; // update the gradient of μ
    $g_t^{(\Gamma)} \leftarrow -w_t \{(h_t - \mu_t) \otimes (h_t - \mu_t)\Gamma_t - \Gamma_t \mathcal{UT}(\Lambda_t)\}$; // update the gradient of Γ
    $\eta_t \leftarrow \frac{1}{t}$; // update the learning rate
    $\mu_{t+1} \leftarrow \mu_t - \eta_t \cdot g_t^{(\mu)}$; // update μ with learning rate η
    $\Gamma_{t+1} \leftarrow \Gamma_t - \eta_t \cdot g_t^{(\Gamma)}$; // update Γ with learning rate η
end

Algorithm 1: Iterative rKPCA through gradient descent

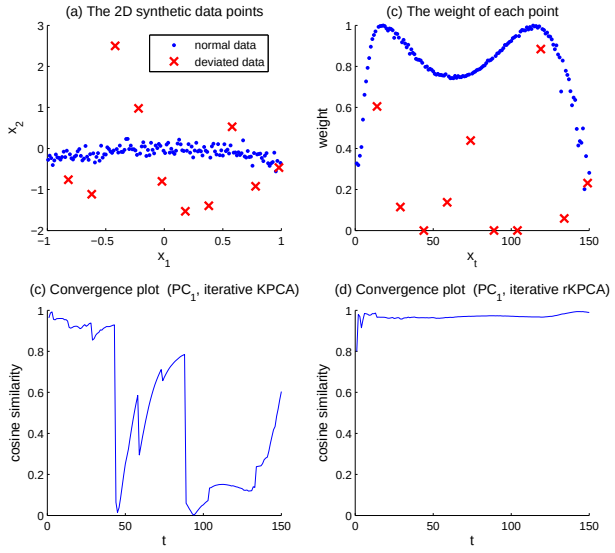

Figure 1: The convergence plots and the psi-principle weights of the iterative KPCA and the rKPCA.

into the data as outliers. These results show that our proposed iterative rKPCA can extract the principal components close to those resulted from the data without outliers.

Example 1 (2-dimensional synthetic data) We generate this 2-dimensional data points $\{(x_{11}, x_{12}), \ldots, (x_{n1}, x_{n2})\}$ with $n = 150$, where x_{i1}'s are equally spaced over (-1,1) and $x_{i2} = -0.3x_{i1}^2 + 0.1\,\mathsf{e}_i$, where e_i's are iid from standard normal. For generating contaminated data, 10 points from the original data are replaced by points via $N(\pm 0.5, \sigma^2)$ in x_2 with $\sigma = 1.5$ (five points for each). All these data points are plotted in Figure 1(a). Kernel data are made by using polynomial kernel $K(x,u) = (x^\top u)^2$, $x, u \in \mathbb{R}^2$.

To check the convergence of the principal directions, the absolute value of the cosine similarity is used to measure the goodness of the convergence. Here we also placed these 10 contaminated data points (replaced ten original data points) equally over the online updating procedure to observe the influence of the contaminated data more clearly. From Figure 1(c), the iterative KPCA can not resist the influence from the contaminated data. The estimated principal direction is affected by these 10 contaminated data and can not achieve to estimate the true principal direction. The absolute value of the cosine similarity by using the iterative KPCA is 0.6038. On the other hand, our iterative rKPCA can reduce the influence of the contaminated data and reach good estimation of the true principal direction with the 0.9892 absolute value of the cosine similarity (see Figure 1(d)). Figure 1(b) also shows that our iterative rKPCA decreases the weights of the contaminated data via the Ψ_1 function. Here we use $\beta = 0.07$ in this 2-dimensional synthetic data.

Example 2 (Phoneme functional data) In this example, we evaluate our method with real world functional data. The functional phoneme data are extracted from the TIMIT database (TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS, US Dept of Com-

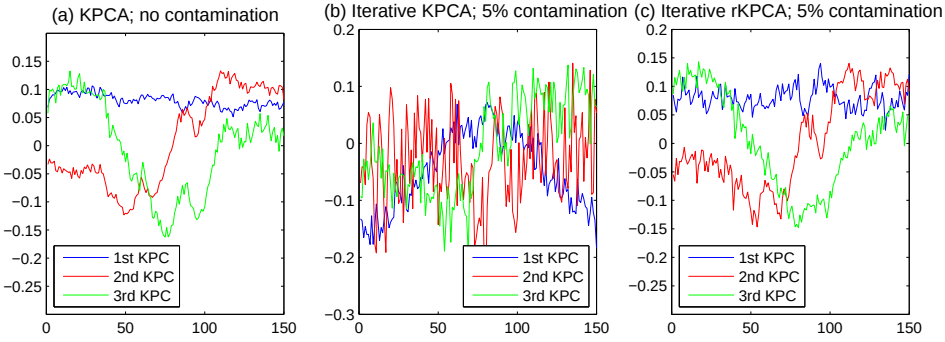

Figure 4: The three leading principal curves estimated by batch KPCA (no contamination), iterative KPCA (5% contamination), and iterative rKPCA (5% contamination.)

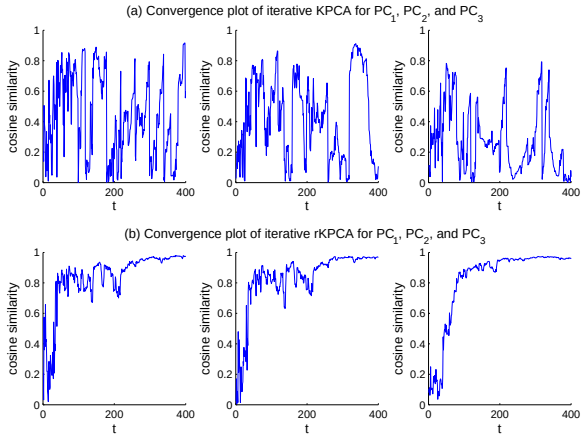

Figure 2: The convergence plots of the three leading principal components by the iterative KPCA and the iterative rKPCA.

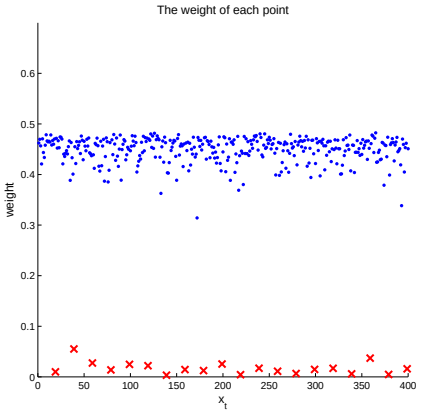

Figure 3: The weight of each functional datum.

merce)[1]. Five phonemes for classification based on digitized speech are: "sh", "iy", "dcl", "aa", and "ao". This functional data consist of 2000 pairs of $\{(f_i, y_i)\}_{i=1}^{2000}$, where y_i represents the class (phoneme) membership and f_i is the discretized log-periodogram of length 150 frequencies. Our experiments focus on the iterative rKPCA, so only one group, "sh" sound with 400 curves are applied in our analysis. 5% of data (20 curves) are randomly drawn to be contaminated with additive perturbation as $\tilde{f}_i(t) = f_i(t) + 5\cos(2\pi t) + 3\mathsf{e}_i(t)$ with $t = 0.5/150, 1.5/150, 2.5/150, \ldots 149.5/150$, where $\mathsf{e}_i(t)$ are iid from uniform$(0, 1)$. In this experiment, three leading functional principal directions are extracted from the contaminated data from the iterative KPCA and our iterative rKPCA. Note that we use the Ψ_2-based iterative rKPCA in this functional data example. The pair of parameters (β, η) in Ψ_2 function is $(0.002, 2)$.

For comparison of the performance of the estimated principal directions, we apply KPCA and iterative KPCA on the contaminated data as the comparison. Using the setting as Example 1, we placed these 5% contaminated data points equally over the online updating procedure. In Figure 2, 3, and 4, we can find that our iterative rKPCA still can reach the true principal directions closely while the iterative KPCA is affected by the deviated data $\tilde{f}(t)$ chooses wrong directions of principal components. We also report the absolute values of the cosine as a similarity indication between the true first three principal directions generated by iterative KPCA and the iterative rKPCA (See Table 1). On the other hand, we can observe that the psi-principle weights of the normal data are closer to each other than those generated by the Ψ_1 function.

Example 3 (Lena image) In this example, we test our

[1]This dataset is available in http://www.lsp.ups-tlse.fr/staph/npfda/.

	PC_1	PC_2	PC_3
iterative rKPCA	0.9726	0.9648	0.9588
iterative KPCA	0.5988	0.1068	0.0378

Table 1: The absolute values of the cosine similarity between the true first three principal directions and the estimated principal directions by the iterative KPCA and the iterative rKPCA.

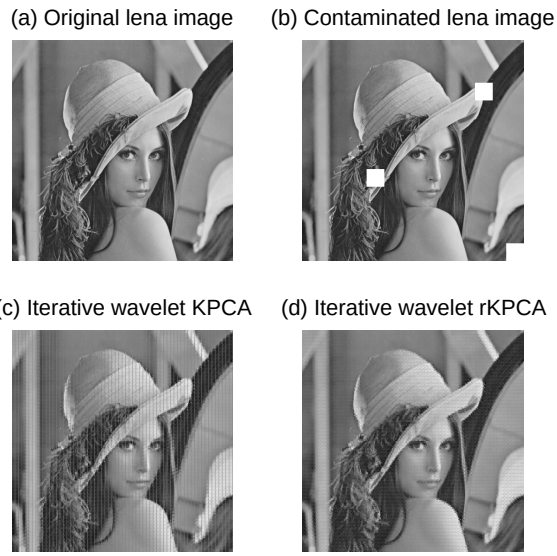

Figure 5: (a) original Lena image (b) image contaminated by three 40×40 pixels white windows (c) image recovered by the iterative conventional KPCA (d) image recovered by the iterative robust KPCA

method on the "Lena" image (see Figure 5(a)). Under compression ratio $1/8$, the iterative robust kernel principal component algorithm is evaluated by examining its use for image compression with level-1 Haar wavelet kernel. This image is contaminated by setting several 40×40 white blocks (see Figure 5(b)). It is a 256-gray-level image of size 512×512. The image was coded 8×8 disjoint blocks. Each block is arranged into a 64×1 column vector, so the image is a form of 64×4096. To reconstruct the image from eight principal components, the uncontaminated images are projected onto the subspace spanned by two basis sets from the iterative KPCA and iterative rKPCA respectively.

For the iterative KPCA, the estimated basis set is disarranged by the white block and could not converge to the correct basis set in this one-pass updating procedure (see Figure 5(c)). This unstability is caused by the single datum updating procedure. The influence of outliers is enlarged while only using local information (single datum) to adjust the principal directions. On the other hand, Figure 5(d) shows that our iterative rKPCA with Ψ_1 function can resist the contamination of the white blocks and gives a better result of extracting the principal directions. In this image example, we use the parameter $\beta = 3 \times 10^{-6}$ in Ψ_1 function and an initial matrix of 16×8 dimension where each entry is generated from $N(250, 1)$.

Example 4 (Series images) In this series images example, there are 90 gray-level series human faces images with size 28×20 [2]. From 10th image to 90th image, every 10th image is replaced by the "fortune cat" image. The original images of the training set shown in Figure 7(a) are replaced by the Fortune Cats shown in Figure 7(b) . Here the iterative KPCA and the iterative rKPCA uses the 81 human faces and 9 fortune cats as the training set and Order-4 symlet wavelet kernel. We apply the sym4 wavelet coefficients to extract the main resolution of each image, and use PCA to capture 9 principal components of these coefficients. The reconstructed images made by the iterative KPCA shown in Figure 7(c). The shadow of Fortune cats appear in those images reconstructed by the iterative KPCA. Alternatively, the iterative rKCA preserves the rotation of faces and leads to a cleared reconstructed images (see Figure 7(d)).

5. Conclusion

We have presented a method for online robust kernel principal component analysis that can be used for nonlinear models from data that may be contaminated by outliers. The approach extends previous work in the image reconstruction by modeling outliers occur sparsely. The method has been applied on natural and synthetic data and shown improved robustness to outliers when compared with other techniques.

It is known that the KPCA is sensitive to outliers. There are various robustness techniques in a variety of ways. We work on applications of robustness-principle for rKPCA through automatic and iterative data weighting scheme. The mean and the principal components generated by the iterative rKPCA are asymptotically stable. Furthermore, we demonstrate the characteristics by experiments of synthetic and real data that the proposed method uses less time to calculate and have equivalent robustness as the batch KPCA.

Yet, how to tune the parameter of the psi-function still needs to be surveyed. Since we obtain the data sequentially, it is not reasonable to use the cross validation in

[2] These series images data is available in http://www.cs.toronto.edu/~ roweis/data.html

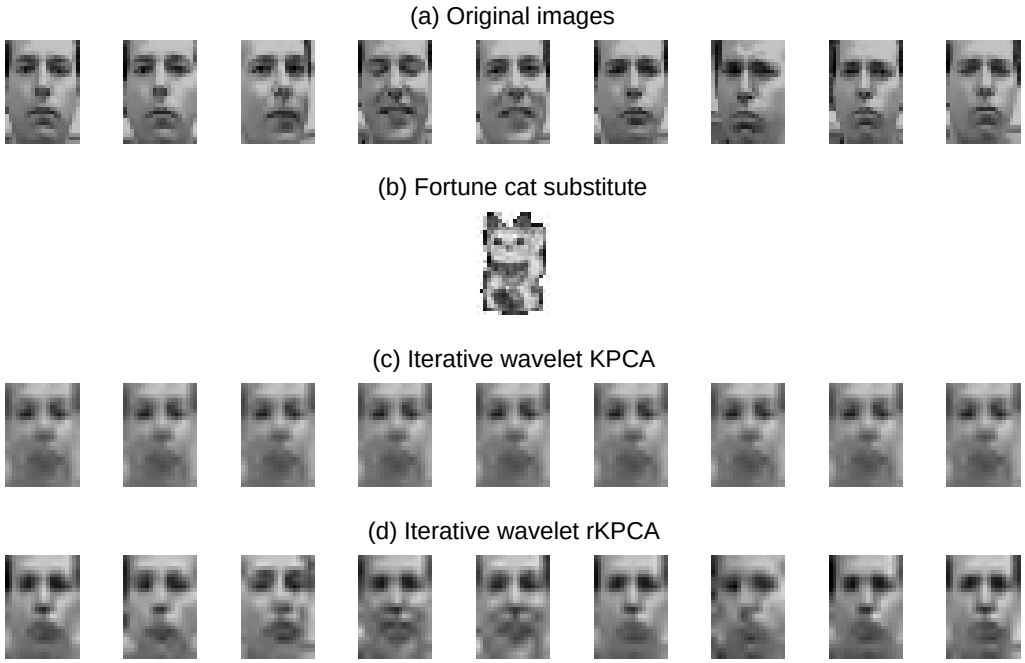

Figure 6: The original images are in the first row. The fortune cats which replace the images are in the second row. The reconstructed images resulted from the iterative KPCA are in the third row. The reconstructed images resulted from the iterative rKPCA are in the fourth row.

tuning the parameters. Besides, the definition of outliers is challenging if we have only small number of initial data. Practically, it is better to provide appropriate initial principal components based on a small batch KPCA. Though these natural drawbacks exist in iterative KPCA. The robust weight function improves the robustness in high-dimensional data.

6. Appendix: Proofs of Theorems 1 and 2

6.1. Proof for Theorem 1

Owning to the fact that the noises wash out [11], as t is sufficiently large, we have an ordinary differential equation (ODE)

$$\frac{d}{dt}\mu_t := m(\Gamma_t, \mu_t) - \kappa(\Gamma_t, \mu_t)\mu_t, \tag{23}$$

where $m(t) = m(\Gamma_t, \mu_t) = \mathrm{E}_G[\dot{\Psi}(z(h_x, \mu_t, \Gamma_t))h_x]$ and $\kappa(t) = \kappa(\Gamma_t, \mu_t) = \mathrm{E}_G[\dot{\Psi}(z(h_x, \mu, \Gamma))]$. The ODE can be solved by using an integrating factor

$$f(t) = \exp\left\{\int_0^t \kappa(s)ds\right\} \tag{24}$$

on both sides, then we have

$$\mu_t = \frac{\int_0^t m(s)f(s)ds + a}{f(t)}, \tag{25}$$

where a is an arbitrary constant. Because $f'(t) = \kappa(t)f(t) > 0$, this implies that $f(t) \to \infty$ as $t \to \infty$. By l'Hôpital's rule, μ_t converges to

$$\frac{m(\infty)}{\kappa(\infty)} = \frac{\mathrm{E}_G[\dot{\Psi}(z(h_x, \mu, \Gamma))h_x]}{\mathrm{E}_G[\dot{\Psi}(z(h_x, \mu, \Gamma))]} \tag{26}$$

as $t \to \infty$.

6.2. Proof for Theorem 2

There are six sufficient conditions for Theorem 2 in order that the principal components are asymptotically stable:

1. η_t is a sequence of positive real numbers such that

$$\eta_t \to 0, \ \sum_{t=1}^{\infty} \eta_t^p < \infty,$$

   for $p > 1$, and $\sum_{t=1}^{\infty} \eta_t = \infty$.
2. Γ_t is bounded almost surely.
3. $g_t^{(\Gamma)}(\Gamma_t, h_t - \mu_t)$ is continuously differentiable in Γ_t and $h_t - \mu_t$, and its derivative is bounded in time.
4. $\lim_{t\to\infty} E_G[g_t^{(\Gamma)}(\Gamma_t, h_t - \mu_t)]$ exists for Γ_t in attraction $D(S)$.

5. There is a locally asymptotically stable set S for the differential equation
$$\frac{d}{dt}\Gamma_t = g_t^{(\Gamma)}$$
with domain of attraction $D(S)$.
6. Γ_t enters some compact subset $A \subset D(S)$ infinitely often with probability 1.

Checking Conditions 1 to 4 Here we choose $\eta_t = \frac{1}{t}$ so that η_t satisfies Condition 1. For Condition 2, h_t is bounded by given a bounded basis set $\{K(\cdot, u_{t'})\}_{t'=1}^{\tilde{m}}$ and a bounded input data x_t . For Condition 3, since $g^{(\Gamma)}$ is composed of differentiable functions $\dot{\Psi}$, $\Sigma\Gamma$ and $\Gamma\mathcal{UT}(\Gamma^\top\Sigma\Gamma)$, thus $g^{(\Gamma)}$ is continuously differentiable with respect to Γ. For the bounded derivative requirement, we have to show
$$\sup_{h_t-\mu_t,\Gamma_t}\left|\frac{d}{dt}g_t^{(\Gamma)}\right| < \text{a constant}. \tag{27}$$
Under a fixed kernel basis set, $|\dot{\Psi}_t| \leq 1$, bounded Γ_t and Condition 2, $\|h_t - \mu_t\|$ is bounded. Since $\sup_{h_t-\mu_t,\Gamma_t}\left|\frac{d}{dt}g_t^{(\Gamma)}\right|$ is a tensor product of $|\dot{\Psi}_t|$, $\|h_t-\mu_t\|$ and Γ_t, thus $\sup_{h_t-\mu_t,\Gamma_t}\left|\frac{d}{dt}g_t^{(\Gamma)}\right|$ is bounded so that Condition 3 is satisfied. If $\lim_{t\to\infty}\mu_t$ and $\lim_{t\to\infty}\Gamma_t$ exist, then Condition 4 holds with $\lim_{t\to\infty}E_G[g_t^{(\Gamma)}(\Gamma_t, h_t-\mu)] = E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))\{(h_x-\mu)\otimes(h_x-\mu)\Gamma - \Gamma\mathcal{UT}(\Lambda)\}]$. Since Theorem 1 confirms the existence of the limit of μ_t, the rest part that we have to show is the existence of Γ_t. We show it in the procedure of checking Condition 5.

Checking Condition 5 In the following analysis, we use ODE method to check Condition 5. The ODE derived from the expectation of $g_t^{(\Gamma)}(\Gamma_t, h_t-\mu)$ is as
$$\frac{d\Gamma_t}{dt} = \Sigma'\Gamma_t - \Gamma_t\mathcal{UT}\left(\Gamma_t^\top\Sigma'\Gamma_t\right), \tag{28}$$
where $\frac{E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))(h_x-\mu)\otimes(h_x-\mu)]}{E_G[\dot{\Psi}(z(h_x,\mu,\Gamma))]} = \Sigma'$. Here we assume that all the eigenvalues of Σ' are not zero. This local asymptotic stability can be proved by Lyapunov's second method ([15]) and mathematical induction. We put the basis (base case) in Step 1 and the inductive step in Steps 2.

Step 1: Here we show that the solution of the ODE of the first eigenvector $\gamma_1(t)$ is locally asymptotically stable. We use the ODE method to show the convergence of $\gamma_1(t)$ and asymptotic stability by the differential equation
$$\frac{d\gamma_1(t)}{dt} = \Sigma'\gamma_1(t) - \gamma_1(t)\left(\gamma_1(t)^\top\Sigma'\gamma_1(t)\right). \tag{29}$$
Assume we choose l principal components which are denoted as $q_1(t), q_2(t), \cdots, q_l(t)$ corresponding to eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_l$, then we can expand $\gamma_1(t)$ in terms of the entire orthonormal set of eigenvectors as
$$\gamma_1(t) = \sum_{k=1}^{l}\theta_k(t)q_k, \quad \text{where} \quad \theta_k(t) = \gamma_i(t)^\top q_k. \tag{30}$$
Plugging (30) together with $\Sigma' q_k = \lambda_k(t)q_k$ into (29) gives
$$\begin{aligned}
&\frac{d\gamma_1(t)}{dt} = \sum_{k=1}^{l}\frac{d\theta_k(t)}{dt}q_k \\
= &\sum_{k\geq i}\lambda_k(t)\theta_k(t)q_k - \left(\sum_{v=1}^{l}\lambda_v(t)\theta_v^2(t)\right)\left(\sum_{k=1}^{l}\theta_k(t)q_k\right) \\
= &\sum_{k\geq 1}\left(\lambda_k(t) - \sum_{l=1}^{l}\lambda_l(t)\theta_l^2(t)\right)\theta_k(t)q_k := \sum_{k=1}^{l}\frac{d\theta_k(t)}{dt}q_k,
\end{aligned} \tag{31}$$
where the last equality results from the orthonormality of eigenvectors. Since $\gamma_1(t) = \sum_{v=1}^{l}\theta_v(t)q_v$ and $\Sigma'\left(\sum_{v=1}^{l}\theta_v(t)q_v\right) = \sum_{v=1}^{l}\lambda_v\theta_v q_v$, hence the second term of (29) can be written as
$$\begin{aligned}
&\left(\gamma_1(t)^\top\Sigma'\gamma_1(t)\right)\gamma_1(t) \\
= &\left[\left(\sum_{u=1}^{l}\theta_u q_u\right)^\top\sum_{v=1}^{l}\lambda_v\theta_v q_v\right]\left[\sum_{k=1}^{l}\theta_k q_k\right] \\
= &\left(\sum_{v=1}^{l}\lambda_v\theta_v^2\right)\left(\sum_{k=1}^{l}\theta_k q_k\right),
\end{aligned} \tag{32}$$
where the last equality results from the orthonormality of eigenvectors. Then we can multiply $q_k^\top$ to both sides of (31), and therefore by the orthonormality of $\{q_k\}_{k=1}^{l}$
$$\frac{d\theta_k(t)}{dt} = \theta_k(t)\left(\lambda_k(t) - \sum_{i=1}^{l}\lambda_i(t)\theta_i^2(t)\right). \tag{33}$$
As $t\to\infty$, $\{\lambda_v\}_{v=1}^{l}$ are independent of t. Thus we have the asymptotic ODE
$$\frac{d\theta_k(t)}{dt} = \theta_k(t)\left[\lambda_k - \lambda_1\theta_1^2(t) - \sum_{v=2}^{l}\lambda_v^2\theta_v^2(t)\right]. \tag{34}$$
Then the solution of the ODE depends on the number of the basis of Σ', l, which is either greater than one or equal to one.

Case $l > 1$: When $l \geq 2$, as $t\to\infty$ $\{\theta_k(t)\}_{k\geq 2}$ is not empty. Then we can define a variable $\omega_k(t)$ by $\theta_k(t)/\theta_i(t)$ (here $\theta_i \neq 0$), and (34) implies the following ODE
$$\frac{d\omega_k(t)}{dt} = \frac{1}{\theta_i(t)}\left(\frac{d\theta_k(t)}{dt} - \omega_k\frac{d\theta_i(t)}{dt}\right) \quad \text{for all } k = 1,\ldots,l. \tag{35}$$

After replacing $\frac{d\theta_k(t)}{dt}$ and $\frac{\theta_i(t)}{dt}$ by the right hand side of (34), we have

$$\begin{aligned}\frac{d\omega_k(t)}{dt} = & \frac{1}{\theta_i(t)}[\theta_k(t)(\lambda_k - \sum_{k=1}^{l}\lambda_k\theta_k^2(t)) \\ & -\omega_k\theta_i(t)(\lambda_i - \sum_{k=1}^{l}\lambda_k\theta_k^2(t))]\end{aligned} \tag{36}$$

which can be simplified to

$$\frac{d\omega_k(t)}{dt} = \omega_k(t)(\lambda_k - \lambda_i). \tag{37}$$

The solution to the above differential equation is

$$\omega_k(t) = \omega_k(0)\exp[(\lambda_k - \lambda_i)t]. \tag{38}$$

Since the eigenvalues are indexed in decreasing order, λ_i is the largest eigenvalue in $\{\lambda_i, \ldots, \lambda_l\}$, and $\lambda_k - \lambda_i < 0$ for $k > i$. Therefore, $\omega_k(t)$ exponentially converges to zero with any $\omega_k(0)$ for $k > 1$.

Case $l = 1$: By the equation (34),

$$\begin{aligned}\frac{d\theta_1(t)}{dt} = & \theta_1(t)\left[\lambda_1 - \sum_{v=1}^{l}\lambda_v\theta_1^2(t)\right] \\ = & \theta_1(t)\left[\lambda_1 - \lambda_1\theta_1^2(t) - \sum_{v=2}^{l}\lambda_v\theta_v^2(t)\right].\end{aligned} \tag{39}$$

Since there is only one principal component, thus $\theta_v(t) \to 0$ for $v > 1$ so that the last term above approaches zero. Therefore, we drop the last term, and (34) becomes

$$\frac{d\theta_1(t)}{dt} = \lambda_1\theta_1(t)\left[1 - \theta_1^2(t)\right]. \tag{40}$$

To show that the solution of the above ODE is locally asymptotically stable, we need a Lyapunov function, $V(t)$, satisfying the following requirement:

1. $V(t) > (<)0$,
2. $\frac{d}{dt}V(t) < (>)0$
3. $V(t)$ has a minimum (maximum).

Define $V(t) = \left[\theta_1^2 - 1\right]^2$. Then we can use the Lyapunov theorem to show the asymptotic stability of γ_t. If there is a Lyapunov function for a $\frac{d}{dt}\theta_1(t)$, then the following differential solution is asymptotically stable:

$$\frac{dV(t)}{dt} = \frac{d}{dt}[\theta_1^2 - 1]^2 = 4\theta_1[\theta_i^2 - 1]\frac{d\theta_1}{dt} = -4\lambda_1\theta_1^2[\theta_1^2 - 1]^2. \tag{41}$$

Accordingly, $\frac{V(t)}{dt} \le 0$. Thus $V(t)$ is a Lyapunov function and has a minimum at $|\theta_1| = 1$ and $\theta_1 = 0$, which we have excluded by the assumption. Therefore, $\frac{V(t)}{dt} < 0$, and then the system is locally asymptotically stable.

Step 2: By the updating of γ_k, we have the ODE

$$\frac{d\gamma_k(t)}{dt} = \Sigma'\gamma_k(t) - \gamma_k(t)\mathcal{UT}\left(\gamma_k(t)^\top\Sigma'\gamma_k(t)\right). \tag{42}$$

Expanding of each column of Σ' in terms of the eigenvectors of Σ' is

$$\gamma_k(t) = \sum_{v=1}^{l}\theta_v(t)q_v, \tag{43}$$

and combining (43) and (42), we have

$$\begin{aligned}\frac{d\gamma_k(t)}{dt} = & \Sigma'\gamma_k(t) - \left((\gamma_k(t))^\top\Sigma'\gamma_k(t)\right)\gamma_k(t) \\ & - \sum_{v<k}\left((\gamma_k(t))^\top\Sigma'\theta_k q_k\right)\theta_v q_v.\end{aligned} \tag{44}$$

Hence, the third term of (44) can be written as

$$\begin{aligned}& \sum_{v<k}\left((\gamma_k(t))^\top\Sigma'\theta_v q_v\right)\theta_v q_v \\ = & \sum_{v<k}\left((\gamma_k(t))^\top\theta_v\Sigma' q_v\right)q_v \\ = & \sum_{v<k}\lambda_v\left((\gamma_k(t))^T q_v\right)q_v = \sum_{v<k}\lambda_v\theta_v(t)q_v\end{aligned} \tag{45}$$

such that

$$\begin{aligned}& \Sigma'\gamma_k(t) - \sum_{v<k}\left((\gamma_k(t))^\top\Sigma'\theta_v q_v\right)\theta_v q_v \\ = & \sum_{v=1}^{l}\lambda_v\theta_v(t)q_v - \sum_{v<k}\lambda_v\theta_v(t)q_v \\ = & \sum_{v\ge k}\lambda_v\theta_v(t)q_v.\end{aligned} \tag{46}$$

As t is large enough, $\{\lambda_v\}_{v=1}^{l}$ are independent of t. After plugging the above expansion together with $\Sigma' q_k = \lambda_k q_k$ into (44), we have

$$\begin{aligned}\frac{d\gamma_k(t)}{dt} = & \sum_{u\ge k}\lambda_v\theta_v(t)q_v - \left(\sum_{u=1}^{l}\lambda_u\theta_u^2(t)\right)\left(\sum_{v=1}^{l}\theta_v q_v\right) \\ = & \sum_{u\ge k}\left(\lambda_u - \sum_{v=1}^{l}\lambda_v\theta_v^2(t)\right)\theta_u(t)q_u \\ & - \sum_{u<k}\left(\sum_{v=1}^{l}\lambda_v\theta_v^2(t)\right)\theta_u(t)q_u \\ := & \sum_{u=1}^{l}\frac{d\theta_u(t)}{dt}q_u,\end{aligned} \tag{47}$$

where

$$\frac{d\theta_u(t)}{dt} \equiv \begin{cases}\theta_u(t)\left(\lambda_u - \sum_{v=1}^{l}\lambda_v\theta_v^2(t)\right), & \text{for } u \ge k; \\ -\theta_u(t)\sum_{v=1}^{l}\lambda_v\theta_v^2(t), & \text{for } u < k.\end{cases} \tag{48}$$

For $u < k$, the solution to the differential equation is

$$\theta_u(t) = \theta_u(0) \exp\left[-\left(\sum_{v=1}^{l} \lambda_v \theta_v^2\right) t\right]. \quad (49)$$

Moreover, since Σ' is positive definite, $\lambda_v > 0$. Thus $-\left(\sum_{v=1}^{l} \lambda_v \theta_v^2\right) < 0$, and then $\theta_u(t)$ converges to zero exponentially with any $\theta_u(0)$ for $k < j$.

When $u > k$, we can define a variable ω_u by θ_u/θ_k (here $\theta_k \neq 0$), and then prove local asymptotic stability of solutions of the following ODE

$$\frac{d\omega_u(t)}{dt} = \frac{1}{\theta_k(t)}\left(\frac{d\theta_u(t)}{dt} - \omega_u \frac{d\theta_k(t)}{dt}\right). \quad (50)$$

After replacing $\frac{d\theta_u(t)}{dt}$ and $\frac{d\theta_k(t)}{dt}$ by the right hand side of (48), we have

$$\begin{aligned} \frac{d\omega_u(t)}{dt} = \; & \frac{1}{\theta_k}[\theta_u(t)(\lambda_u - \sum_{v=1}^{l} \theta_v^2(t)\lambda_v) \\ & -\omega_u(t)\theta_k(t)(\lambda_k - \sum_{v=1}^{l} \theta_v^2 \lambda_v)], \end{aligned} \quad (51)$$

which is simplified to

$$\frac{d\omega_u(t)}{dt} = \omega_u(t)(\lambda_u - \lambda_k). \quad (52)$$

The solution to the above differential equation is

$$\omega_u(t) = \omega_u(0) \exp\left[(\lambda_u - \lambda_k)\, t\right]. \quad (53)$$

Since $\lambda_u < \lambda_k$, thus $\omega_u(t)$ converges to zero exponentially with an arbitrary $\omega_u(0)$ for all $u > k$.

When $u = k$, (48) implies that

$$\frac{d\theta_k(t)}{dt} = \theta_k(t)\left(\lambda_k - \theta_k(t) \sum_{v>k} \lambda_v \theta_v^2(t) - \theta_k(t) \sum_{v<k} \lambda_v \theta_v^2(t)\right). \quad (54)$$

It has been shown in the previous cases that $\theta_v(t) \to 0$ for $v < k$ and $\theta_v(t) \to 0$ for $v > k$. Hence we can drop the last two terms, and the equation becomes

$$\frac{d\theta_k(t)}{dt} = \theta_k(t)\lambda_k(1 - \theta_k^2(t)). \quad (55)$$

To show that $\theta_j(t)$ converges, we define another function

$$P = (\theta_k^2(t) - 1)^2. \quad (56)$$

After plugging (55) into $\frac{dP(t)}{dt}$, we have

$$\frac{dP(t)}{dt} = -4\lambda_k \theta_k^2(t)(\theta_k^2(t) - 1)^2, \quad (57)$$

so we observe that $\frac{dP(t)}{dt} \leq 0$. Since $\theta_k^2(t) \neq 1$, thus $\frac{dP(t)}{dt} < 0$, and hence the solution is locally asymptotically stable. Moreover, $\{\gamma_k(t)\}_{k=1}^{l}$ converge almost surely to $\pm q_1$, the normalized principal eigenvector of Σ'.

Checking Condition 6 For Condition 6, the existence of the basin of attraction is needed to be checked. That is, there is a compact set A of the set of all matrices such that $\Gamma_t \in A$ as $t \to \infty$. Here we show that there is a compact subset $D(S)$ of the set of all matrices such that $\Gamma_t \in D(S)$ as $t \to \infty$. Let A be the compact subset of $\mathbb{R}^{\tilde{m} \times l}$ given by the set of matrices with norm less than or equal to some constant b. By the properties of an operator norm which is defined as $\|L\| := \max_{x \neq 0} \frac{\|Lx\|}{\|x\|}$,

$$\begin{aligned} \|\Gamma_{t+1}\| \leq \; & \|\eta_t \dot{\Psi}(h_t - \mu_t) y(t)\| \\ & + \|I - \dot{\Psi}_t \eta_t \mathcal{U}\mathcal{T}(\Lambda)\|\|\Gamma_t\| \\ \leq \; & |\eta_t| \|(h_t - \mu_t)^\top y(t)\| \\ & + \|I - \dot{\Psi}_t \eta_t \mathcal{U}\mathcal{T}(y(t)^\top y(t))\|\|\Gamma_t\|, \end{aligned} \quad (58)$$

where $y(t) = (h_t - \mu_t)^\top \Gamma_t$. Since $\|y(t)\|$, $\|h_t - \mu_t\|$ and $|\dot{\Psi}_t|$ are all bounded, and $\eta_t \to 0$ as $t \to \infty$, thus $|\dot{\Psi}_t \eta_t| \to 0$ so that $\|I - \dot{\Psi}_t \eta_t \mathcal{U}\mathcal{T}(y(t)^\top y(t))\| \to 1$. This implies that $\|\Gamma_{t+1}\| \leq \|\Gamma_t\|$ when $t \to \infty$. Thus Γ_t will be eventually in A as $t \to \infty$. Since the basin of attraction $D(S)$ includes all matrices with bounded norm, thus A belongs to $D(S)$. By the virtue of Ljung's theorem [11], the above checking ensures that Γ_t converges to the matrix whose columns are the first l eigenvectors of Σ' in descending eigenvalues order.

References

[1] I. T. Jolliffe, Principal Component Analysis, Springer, 2002.

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

[3] S. Günter, N. N. Schraudolph, S. V. N. Vishwanathan, Fast iterative kernel principal component analysis, Journal of Machine Learning Research 8 (2007) 1893–1918.

[4] J. E. Jackson, A User's Guide to Principal Components, Wiley and Sons, 1991.

[5] S.-Y. Huang, Y.-R. Yeh, S. Eguchi, Robust kernel principal component analysis, Neural Computation 21 (2009) 3179–3213.

[6] K. I. Kim, M. O. Franz, B. Schölkopf, Iterative kernel principal component analysis for image modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (9) (2005) 1351–1366.

[7] T. D. Sanger, Optimal unsupervised learning in a single-layer linear feedforward neural network, Neural Networks 2 (6) (1989) 459–473.

[8] I. Higuchi, S. Eguchi, Robust principal component analysis with adaptive selection for tuning parameters, Journal of Machine Learning Research 5 (2004) 453–471.

[9] Y.-J. Lee, S.-Y. Huang, Reduced support vector machines: A statistical theory, IEEE Transactions on Neural Networks 18 (1) (2007) 1–13.

[10] Y.-J. Lee, O. L. Mangasarian, Rsvm: Reduced support vector machines, in: Proceedings of the First SIAM International Conference on Data Mining, 2001.
[11] L. Ljung, Analysis of recursive stochastic algorithms, IEEE Transactions on Automatic Control 22 (4) (1977) 551–575.
[12] S. Canu, X. Mary, A. Rakotomamonjy, Functional Learning Through Kernel, IOS Press, Amsterdam, 2003.
[13] B. Vidakovic, Statistical Modeling by Wavelets, Wiley Interscience, 1999.
[14] I. Daubechies, Ten Lectures on Wavelets, SIAM, Philadelphia, 1992.
[15] E. Oja, A simplified neuron model as a principal component analyzer, Journal of Mathematical Biology 15 (3) (1982) 267–273.