S E C T I O N O F S T A T I S T I C S

DEPARTMENT OF MATHEMATICS

KATHOLIEKE UNIVERSITEIT LEUVEN

T E C H N I C A L R E P O R T

TR-07-11

DETECTING INFLUENTIAL OBSERVATIONS IN KERNEL PCA

Debruyne, M., Hubert, M. and Van Horebeek, J.

http://wis.kuleuven.be/stat/

Detecting Influential Observations in Kernel PCA

Michiel Debruyne,
Dept. of mathematics and computer science, Universiteit Antwerpen,
Middelheimlaan 1G, B-2020 Antwerpen, Belgium.

Mia Hubert,
Dept. of mathematics, K.U.Leuven - LStat,
Celestijnenlaan 200B, B-3001 Leuven, Belgium.

Johan Van Horebeek,
Center for Research in Mathematics (CIMAT),
Apartado Postal 402, Guanajuato, Gto. 36000, México

January 4, 2008

Abstract

Individual observations can be very influential when performing classical Principal Component Analysis in a Euclidean space. Robust PCA algorithms detect and neutralize such dominating data points. This paper studies robustness issues for PCA in a kernel induced feature space. The sensitivity of Kernel PCA is characterized by calculating the influence function. A robust Kernel PCA method is proposed by incorporating kernels in the Spherical PCA algorithm. Using the scores from Spherical Kernel PCA, a graphical diagnostic is proposed to detect points that are influential for ordinary Kernel PCA.

Keywords: robust statistics, spherical PCA, kernel methods, influence function.

1 Introduction

Principal Component Analysis (PCA) is a well known technique designed to reduce the dimension of a data set by projecting onto a lower dimensional subspace. Kernel PCA (Schölkopf et al., 1998) is an extension of PCA where the data are first mapped into a high dimensional feature space. Then ordinary PCA is performed in this feature space. A remarkable aspect is that the explicit feature vectors are not needed to compute the resulting scores. Only the inner products between feature vectors are required. This makes it possible to apply the kernel trick: one replaces all inner products by a kernel function that is chosen beforehand (see for example Schölkopf and Smola (2002) for an extensive overview of kernel methods). This extension from linear to Kernel PCA has found many applications in recent years. It is for instance easy to consider types of non-linear PCA, simply by defining an appropriate non-linear kernel function. Also when the data consist of objects rather than real numbers, such a kernel formulation is very attractive: it suffices to define an appropriate kernel function between any two such objects. For example in text and string analysis, kernel methods enjoy an increasing popularity (Shawe-Taylor and Cristianini, 2004).

The current paper addresses some questions about influential observations in Kernel PCA. For linear PCA this has been studied intensively. It is known that some observations are relatively less important than others. Points close to the center for example do not really help a lot determining principal components. Observations far away from the center on the other hand are much more influential. Actually, it is even possible that one or a small fraction of observations in the data set almost fully determines the principal components. Sometimes this is not desirable, since then the structure of the majority of the data is not learned anymore. Therefore many robust PCA algorithms have been proposed, for instance by Locantore et al. (1999), Hubert et al. (2002), Croux and Ruiz-Gazen (2005), Hubert et al. (2005), Maronna (2005). These methods are less affected by outliers and produce scores which do fit the majority of the data points. Additionally outliers can be detected using these methods in appropriate diagnostic tools.

The goal of this paper is to extend these robustness issues from linear PCA to general Kernel PCA. Our contribution is threefold:

- A theoretical analysis of the effect of contamination for ordinary Kernel PCA. To this extent we calculate the influence function (Hampel et al., 1986) of Kernel PCA in Section 3. We show that the influence function can be arbitrary large for unbounded kernels. This means that very small fractions of observations can completely determine the results, such that the structure of the majority of the data is not reflected. When the kernel is bounded however, the influence function is bounded as well.
- Since the influence function indicates that outliers can be malicious for a general unbounded kernel, our next step is to construct a robust Kernel PCA algorithm. To this end Spherical Kernel PCA is proposed. It is a generalization of the linear method from Locantore et al. (1999). The first step is a robust centering of the data, which is explained in detail in Section 4. The entire Spherical KPCA procedure is given in Section 5.
- In practice it is often difficult to find outliers and to know how to handle them. An important application of a robust PCA method is not only to produce robust scores, but also to detect outliers and to visualize the difference between applying the robust and the classical method. If some outliers are indeed detected, further inspection of these observations can reveal important information. Such a visual display is constructed in Section 6 applying an idea from Pison and Van Aelst (2004).

Section 7 illustrates our new methods on some specific examples. It is shown that outliers in the data can be neutralized and detected using Spherical KPCA, whereas classical KPCA fails to do so.

2 Kernel PCA

Assume that we have a sample of n observations in some non-empty set $\mathcal{X}$: $x_i \in \mathcal{X}$, $i = 1, \ldots, n$.

Definition 1 *A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a* kernel *on $\mathcal{X}$ if there exists a $\mathbb{R}$-Hilbert space $\mathcal{H}$ with inner product $\langle ., . \rangle$ and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that for all $x, x' \in \mathcal{X}$ we have*

$$K(x, x') \;=\; \langle \Phi(x), \Phi(x') \rangle \,.$$

We call Φ the feature map *and $\mathcal{H}$ the* feature space *of K.*

Then Kernel PCA basically performs linear PCA in the feature space $\mathcal{H}$ instead of the original space $\mathcal{X}$. Schölkopf et al. (1998) show that the solution of this problem can be obtained only in terms of the inner products between feature vectors. Assume that the feature vectors are mean-centered. Denote Ω the matrix containing $\langle \Phi(x_i), \Phi(x_j) \rangle$ as i, j-th entry. The first principal component equals the direction maximizing the variance of the projections onto this direction. Since a principal component is always contained in the space spanned by the observations, this

direction can be written as a linear combination of the feature vectors. Let $\alpha = (\alpha_1, \ldots, \alpha_n)^t \in \mathbb{R}^n$. Then

$$\left\| \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\|^2 = \left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \sum_{i=1}^{n} \alpha_i \Phi(x_i) \right\rangle = \alpha^t \Omega \alpha.$$

Thus the condition $\alpha^t \Omega \alpha = 1$ ensures that the norm of $\sum_{i=1}^{n} \alpha_i \Phi(x_i)$ equals 1. Moreover the projection of feature vector $\Phi(x_j)$ onto such a direction equals

$$\left\langle \sum_{i=1}^{n} \alpha_i \Phi(x_i), \Phi(x_j) \right\rangle = \sum_{i=1}^{n} \alpha_i \left\langle \Phi(x_i), \Phi(x_j) \right\rangle = (\Omega\alpha)_j.$$

Thus maximizing the variance of these projections is equivalent to

$$\max \alpha^t \Omega^2 \alpha \qquad \text{subject to} \qquad \alpha^t \Omega \alpha = 1. \tag{1}$$

The maximum is obtained for α equal to the eigenvector of Ω corresponding to the largest eigenvalue λ_1, with norm equal to $\lambda_1^{-1/2}$. Denote the unit norm eigenvector corresponding to λ_1 as $\alpha^{(1)}$. Then the score of a new feature vector $\Phi(x)$ corresponds to

$$\left\langle \sum_{i=1}^{n} \frac{\alpha_i^{(1)}}{\sqrt{\lambda_1}} \Phi(x_i), \Phi(x) \right\rangle = \sum_{i=1}^{n} \frac{\alpha_i^{(1)}}{\sqrt{\lambda_1}} \left\langle \Phi(x_i), \Phi(x) \right\rangle \tag{2}$$

It is now clear that expressions (1) and (2) depend on the feature vectors $\Phi(x_i)$ only through pairwise inner products. For a Reproducing Kernel Hilbert Space $\mathcal{H}$ these inner products can be evaluated using the underlying kernel function (Definition 1).

Note that the feature vectors were assumed to be centered around zero. However, centering around the mean can be performed explicitly. Taking this into account, Schölkopf et al. (1998) end up with the following result:

Algorithm 1 (Kernel PCA) *Given a sample $x_1, \ldots, x_n \in \mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with corresponding feature space $\mathcal{H}$. Then the kth Kernel PCA (KPCA) score function f_k at $x \in \mathcal{X}$ equals:*

$$f_k(x) = \sum_{i=1}^{n} \frac{\alpha_i^{(k)}}{\sqrt{\lambda_k}} \left(K(x_i, x) - \frac{1}{n} \sum_{l=1}^{n} K(x_l, x) \right)$$

with $\alpha^{(k)}$ the unit norm eigenvector belonging to the kth largest eigenvalue λ_k of the mean centered kernel matrix $\Omega_{c,mean}$ with entry i, j equal to

$$(\Omega_{c,mean})_{i,j} := \left(K(x_i, x_j) - \frac{1}{n} \sum_{k=1}^{n} K(x_k, x_j) - \frac{1}{n} \sum_{k=1}^{n} K(x_k, x_i) + \frac{1}{n^2} \sum_{k=1}^{n} \sum_{l=1}^{n} K(x_k, x_l) \right). \tag{3}$$

Some well known kernels when $\mathcal{X} \subset \mathbb{R}^d$ are the linear kernel

$$K(u, v) = u^t v,$$

for which the solutions are equivalent to classical linear PCA; the polynomial kernel of degree $p > 0$ with offset $\tau \in \mathbb{R}^+$

$$K(u, v) = (u^t v + \tau)^p,$$

and the RBF kernel with bandwidth $\sigma \in \mathbb{R}^+$

$$K(u, v) = e^{-||u-v||^2/\sigma^2}.$$

Many more types of kernels exist, see for instance Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004).

3 Characterizing influential observations

Suppose that we have a distribution P on the input space $\mathcal{X}$. Define the mean-centered covariance operator as follows:

$$C_P : \mathcal{H} \to \mathcal{H} : f \to C_P(f) = \mathbb{E}_P f(X)\Phi(X) - \mathbb{E}_P f(X)\mathbb{E}_P \Phi(X).$$

If $\mathbb{E}_P||\Phi(X)||^2 < \infty$, then the operator C_P is a well-defined, compact, positive and self-adjoint Hilbert-Schmidt operator (Blanchard et al., 2007). Therefore it has a countable spectrum of positive eigenvalues $\lambda_{P,1} \geq \lambda_{P,2} \geq \ldots$ with an associated orthonormal basis of eigenfunctions $\{e_{P,i}\}$. Thus for any function $f \in \mathcal{H}$ we have that

$$f = \sum_{i=1}^{\infty} \langle f, e_{P,i} \rangle e_{P,i} \qquad \text{and} \qquad C_P(f) = \sum_{i=1}^{\infty} \lambda_{P,i} \langle f, e_{P,i} \rangle e_{P,i}.$$

There is a very close connection between the operator C_P and the kernel matrix $\Omega_{c,mean}$: let P_n be the empirical distribution of a sample of size n. Then the eigenvector $\alpha^{(k)}$ obtained by Algorithm 1 converges to the eigenfunction $e_{P,k}$ if $P_n \to P$ for $n \to \infty$. Explicit learning rates were obtained by Shawe-Taylor et al. (2002), Blanchard et al. (2007). In this section we work with continuous distributions P rather than samples since we want to assess properties of the corresponding statistical functionals through the concept of the influence function. The following definitions introduce the necessary tools.

Definition 2 *Given a distribution P with $\mathbb{E}_P||\Phi(X)||^2 < \infty$, then the statistical functionals C, λ_i resp. e_i map P onto $C(P) = C_P$, $\lambda_i(P) = \lambda_{P,i}$ resp. $e_i(P) = e_{P,i}$.*

The goal of this section is to quantify the sensitivity of these statistical functionals under small contamination of the underlying distribution P. To this end the influence function (Hampel et al., 1986) is used.

Definition 3 *Given a statistical functional T mapping a distribution P onto $T(P)$. Consider the contaminated distribution*

$$P_{\epsilon,z} = (1 - \epsilon)P + \epsilon\Delta_z$$

for small enough ϵ. The distribution Δ_z is the Dirac distribution which puts all probability mass at the point z. Then the influence function of T at the distribution P is defined as

$$IF(z;T,P) = \lim_{\epsilon \downarrow 0} \frac{T(P_{\epsilon,z}) - T(P)}{\epsilon}.$$

Thus $IF(z;T,F)$ measures the effect on T under infinitesimally small contamination at the point z. For linear PCA the influence functions of the eigenvalues and the eigenvectors were derived by Critchley (1985). We prove the following theorem for KPCA.

Theorem 1 *Let P be a distribution on $\mathcal{X}$ such that $\mathbb{E}_P||\Phi(X)||^2 < \infty$. Assume that $\lambda_{P,i} \neq \lambda_{P,j}$ for all $j \neq i$. Then the influence functions of λ_i and e_i at P in $z \in \mathcal{X}$ are given by*

$$IF(z;\lambda_i,P) = \langle e_{P,i}, \Phi(z)\rangle^2 - \lambda_{P,i}.$$
$$IF(z;e_i,P) = \langle e_{P,i}, \Phi(z)\rangle \sum_{j=1,j\neq i}^{\infty} \frac{\langle e_{P,j}, \Phi(z)\rangle}{\lambda_{P,i} - \lambda_{P,j}} e_{P,j}.$$

Proof

First note that

$$\mathbb{E}_{P_{\epsilon,z}}||\Phi(X)||^2 = (1-\epsilon)\mathbb{E}_P||\Phi(X)||^2 + \epsilon||\Phi(z)||^2 = (1-\epsilon)\mathbb{E}_P||\Phi(X)||^2 + \epsilon K(z,z).$$

Thus $\mathbb{E}_P||\Phi(X)||^2 < \infty$ implies that $\mathbb{E}_{P_{\epsilon,z}}||\Phi(X)||^2 < \infty$. Thus the operator $C_{P_{\epsilon,z}}$ is a well defined, positive, compact and self-adjoint Hilbert-Schmidt operator. Hence the statistical functionals λ_i and e_i from Definition 2 exist at $P_{\epsilon,z}$ for any $\epsilon \in [0,1]$ and any $z \in \mathcal{X}$. By definition we have that

$$\langle e_i(P_{\epsilon,z}), e_i(P_{\epsilon,z})\rangle = 1.$$

Taking the derivative with respect to ϵ in $\epsilon = 0$ on both sides yields

$$\langle IF(z;e_i,P), e_{P,i}\rangle = 0.$$

Denote $\mathcal{H}^{\perp,i}$ the subspace of $\mathcal{H}$ orthogonal to the ith component. Then $IF(z;e_i) \in \mathcal{H}^{\perp,i}$. Furthermore we have that

$$\begin{aligned}\lambda_i(P_{\epsilon,z})e_i(P_{\epsilon,z}) &= C_{P_{\epsilon,z}}(e_i(P_{\epsilon,z})) \\ &= \mathbb{E}_{P_{\epsilon,z}}\langle e_i(P_{\epsilon,z}), \Phi(X)\rangle\Phi(X) - \mathbb{E}_{P_{\epsilon,z}}\langle e_i(P_{\epsilon,z}), \Phi(X)\rangle\mathbb{E}_{P_{\epsilon,z}}\Phi(X).\end{aligned}$$

Next take the derivative with respect to ϵ and simplify using

$$\mathbb{E}_P\langle e_i(P), \Phi(X)\rangle = \mathbb{E}_P\langle C_P(e_i(P))/\lambda_i, \Phi(X)\rangle = 0$$

since $\mathbb{E}_P\langle C_P(f), \Phi(X)\rangle = 0$ for all $f \in \mathcal{H}$ by definition of C_P as the mean centered covariance operator. Then

$$\begin{aligned}IF(z;\lambda_i,P)e_i(P) + \lambda_i(P)IF(z;e_i,P) &= -\mathbb{E}_P\langle e_i(P), \Phi(X)\rangle\Phi(X)) + \langle e_i(P), \Phi(z)\rangle\Phi(z) \qquad (4)\\ + \mathbb{E}_P\langle IF(z;e_i,P), \Phi(X)\rangle\Phi(X) - \mathbb{E}_P\langle IF(z;e_i,P), \Phi(X)\rangle\mathbb{E}_P\Phi(X) - \langle e_i(P), \Phi(z)\rangle\mathbb{E}_P\Phi(X).\end{aligned}$$

Now take the inner product of both sides with respect to $e_i(P)$. Then

$$IF(z; \lambda_i, P) = -\lambda_{P,i} + \langle e_{P,i}, \Phi(z) \rangle^2$$

proving the first statement. Using this result (4) can be rewritten as

$$(C_P - \lambda_i \mathrm{id}_{\mathcal{H}})(IF(z; e_i, P)) = \langle e_i(P), \Phi(z) \rangle^2 e_i(P) - \langle e_i(P), \Phi(z) \rangle \Phi(z) - \langle e_i(P), \Phi(z) \rangle \mathbb{E}_P \Phi(X).$$

The operator $(C_P - \lambda_{P,i} \mathrm{id}_{\mathcal{H}})$ does not have an eigenvalue equal to 0 in $\mathcal{H}^{\perp,i}$. Thus the Fredholm alternative (see e.g. Phelps (1986)) shows that this operator is invertible and

$$IF(z; e_i, P) = (C_P - \lambda_{P,i} \mathrm{id}_{\mathcal{H}})^{-1} \left(\langle e_i(P), \Phi(z) \rangle^2 e_i(P) - \langle e_i(P), \Phi(z) \rangle (\Phi(z) - \mathbb{E}_P \Phi(X)) \right). \quad (5)$$

Moreover we have that

$$(\lambda_{P,j} - \lambda_{P,i}) \langle IF(z; e_i, P), e_j(P) \rangle = \langle e_i(P), \Phi(z) \rangle \langle e_j(P), \Phi(z) \rangle$$

for any $j \neq i$. Thus

$$IF(z; e_i, P) = \langle e_{P,i}, \Phi(z) \rangle \sum_{j=1, j \neq i}^{\infty} \frac{\langle e_{P,j}, \Phi(z) \rangle}{\lambda_{P,i} - \lambda_{P,j}} e_{P,j}.$$

proving the second statement.

□

This theorem reveals two interesting properties. First we see that estimating an eigenfunction is very sensitive to small distributional changes when other eigenvalues are very close to its corresponding eigenvalue. This is well known for instance in linear PCA. As a limit case consider a spherical distribution. Then a first principal component is not well defined, since all directions give raise to the same projected variance. This changes of course if an arbitrary small amount of Dirac probability mass is put at any point z except for the center of the distribution. Then the direction through z and the center of the distribution will be the first principal component. In this case, an infinitesimally small amount of probability mass fully determines the first component, which is reflected in an infinitesimally large influence function of the eigenfunction.

Now suppose again that all eigenvalues are different. An important question is whether the influence function can become arbitrary large in such a case as well. Theorem 1 tells us exactly how to choose z in order to achieve this: both the score with respect to the ith component and the sum of the scores with respect to the other components should be large. Figure 1 shows a classical example of a very influential point (denoted as observation 11) having a large influence on the components of a two-dimensional Gaussian distribution, in case of a linear kernel.

For a bounded kernel however this is different. From the previous theorem upper bounds on the influence function in terms of bounds on the kernel can be derived.

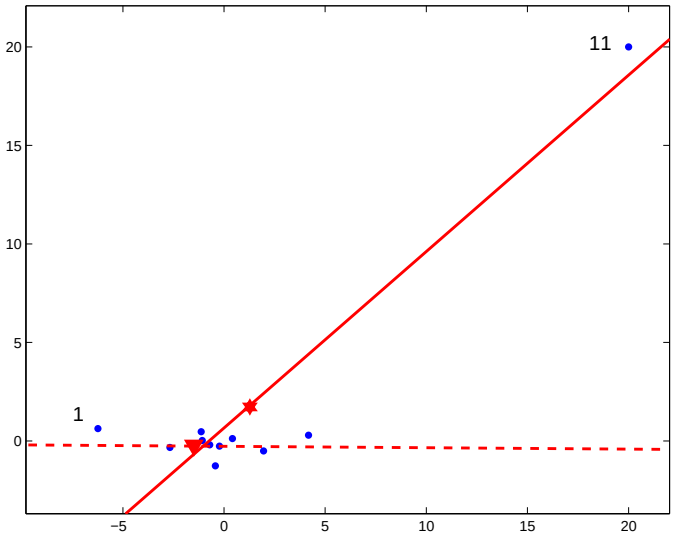


Figure 1: The influence of a single observation on the mean and on the first principal component of linear PCA can be arbitrary large. Solid line and star: classical linear PCA and mean. Dashed line and triangle: Spherical linear PCA and spatial median.

Theorem 2 *With the notation of Definition 1 and a feature map with bounded norm, i.e. there exists $M > 0$ such that $||\Phi(z)|| \leq M$, the following bounds hold:*

$$|IF(z; \lambda_i, P)| \leq M^2 + \lambda_{P,i}.$$
$$||IF(z; e_i, P)|| \leq \frac{3M^2}{\min_j |\lambda_{P,i} - \lambda_{P,j}|}$$

Proof

The Cauchy-Schwarz theorem guarantees that

$$|\langle e_{P,i}, \Phi(z)\rangle|^2 \leq ||e_{P,i}||^2 ||\Phi(z)||^2 = ||\Phi(z)||^2 < M^2$$

for any $i \in \mathbb{N}$. Together with Theorem 1 this immediately gives the upper bound for the influence function of the eigenvalues. For the eigenfunctions equation (5) shows that

$$||IF(z; e_i, P)|| \leq ||(C_P - \lambda_{P,i}\mathrm{id}_{\mathcal{H}})^{-1}|| \;\; ||\langle e_i(P), \Phi(z)\rangle^2 e_i(P) - \langle e_i(P), \Phi(z)\rangle(\Phi(z) - \mathbb{E}_P\Phi(X))||.$$

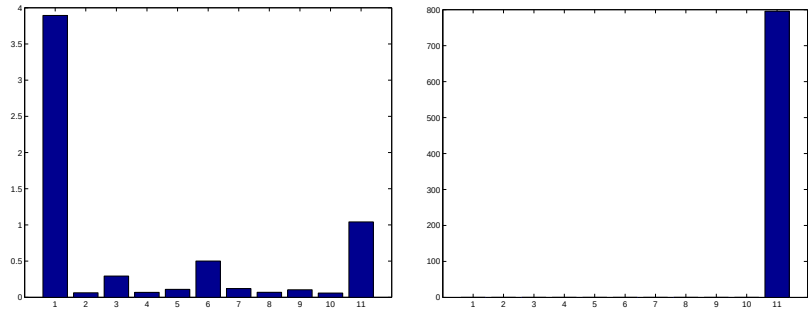
The norm of the operator in the first term is bounded by its largest eigenvalue which equals $(\min_j |\lambda_{P,j} - \lambda_{P,i}|)^{-1}$. Cauchy-Schwarz bounds the second term by $3M^2$.

□

This indicates a crucial difference between bounded and unbounded kernels in terms of robustness of kernel PCA. Similar conclusions were obtained for classification (Christmann and Steinwart, 2004) and regression (Christmann and Steinwart, 2006, Debruyne et al., 2006). For an RBF kernel for example, we can take $M = 1$ showing a bounded influence. On the other hand

the spacings between eigenvalues play an important role as well. For instance as the bandwidth σ of the RBF kernel reaches infinity, the kernel matrix converges to the matrix with all entries equal to 1. Thus all eigenvalues converge to the same value and the upper bound becomes arbitrary large as $\sigma \to \infty$. This is quite expected, since an RBF kernel with large σ approaches linear PCA for which the influence function is indeed unbounded. Therefore observations might still be rather influential even for an RBF, especially if the parameter σ is chosen in a data-driven way. Theorem 2 shows however that the worst cases appear for unbounded kernels.

The main problem with unbounded kernels is that the effect of one or few observations can be so big that their influence is not easily detected anymore. One could for instance think about plugging in the sample eigenvalues λ_i and the sample eigenvectors $\alpha^{(i)}$ obtained by algorithm 1 in the expressions for the influence function of the eigenfunction of interest (thus estimating $e_{P,i}$ by $\alpha^{(i)}$ and $\lambda_{P,i}$ by λ_i). Taking the norm of the resulting influence vector gives a data based diagnostic tool assessing the influence of each observation in the sample. In Figure 2(a) one



(a) (b)

Figure 2: Estimated influences based on (a) classical linear PCA, (b) robust linear PCA.

sees however that such an approach completely fails even for the simple example from Figure 1. If one point would be considered an outlier, it would be observation 1 for which the influence is the highest. The really influential observation 11 is only considered moderately influential. Looking back to Figure 1 one understands why. According to Theorem 1 influential points are characterized by the product of the first and second principal component scores. For observation 1 both terms are rather large leading to a large product. For observation 11 the first score is very large, but the second score is small giving only a moderately large product. In robust statistics such a phenomenon is called the *masking effect*: the influence of point 11 is so huge that it affects estimates and diagnostics so heavily that its influence is actually hidden!

One way around this problem is to use a robust method. Then the principal components are constructed in a such a way that a small fraction of the data can never demolish an entire fit. In Figure 1 the spherical PCA method of Locantore et al. (1999) was used to find the dashed line as first principal component. To visualize the difference between robust and classical PCA, Pison and Van Aelst (2004) propose to plug the robust results into the expressions for the influence function of classical PCA. This produces Figure 2(b) as resulting diagnostic plot. Now it is

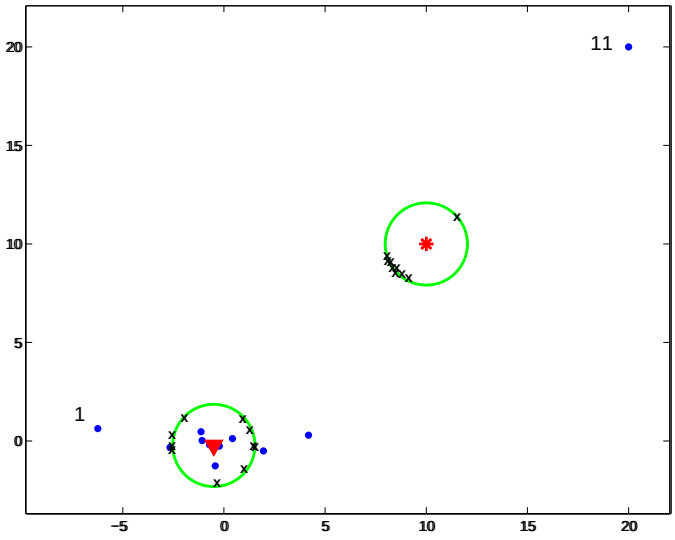


Figure 3: When projecting all data on a sphere around the star, the mean of these projections (depicted as crosses) does not equal the center of the sphere. For the triangle, it does. By definition, the triangle equals the spatial median. Note the moderate influence of observation 11.

clearly visible that observation 11 is the one causing the large difference between robust and classical PCA.

4 Robust centering

4.1 Spatial median in $\mathbb{R}^d$

The first step of PCA consists of centering the data, usually around the mean. However, the mean is not a robust measure of the center. Again one observation can have an arbitrary large influence. In Figure 1 for example the mean (pictured as a star) is clearly influenced a lot by observation 11. A first logical step in a robust PCA procedure consists of a more robust centering. In this section we propose to use the L_1 M-estimate of location, which is a multivariate extension of the univariate median and which has been around for a long time (see for instance Haldane (1948) and Huber (1981)). This location measure is also known as the spatial median. It has a nice geometrical interpretation (Small, 1990): take a point θ in $\mathbb{R}^d$ and project all observation onto a sphere around the center θ. If the mean of these projections equals θ, then θ equals the spatial median. Figure 3 shows this for the previous two-dimensional example. The mean of the data projected on the sphere (these projections are pictured as crosses) around the asterisk does not equal the asterisk at all. The asterisk is a bad estimator of location indeed. The mean of the data projected on the sphere around the triangle does equal the triangle itself. Thus the triangle indicates the position of the spatial median. Note how the sphering reduces the influence of observation 11, such that it does not affect the spatial median

more than any of the other observations.

Definition 4 *Given a sample of inputs $x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. Then the spatial median θ is defined as the solution of*

$$\sum_{i=1}^{n} \frac{x_i - \theta}{||x_i - \theta||} = 0.$$

For the computation of this center, the following simple iterative algorithm exists (Gower, 1974, Huber, 1981, Hössjer and Croux, 1995). Given an initial guess $\theta^{(0)} \in \mathbb{R}^d$, iteratively define

$$\theta^{(k)} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where

$$w_i = \frac{1}{||x_i - \theta^{(k-1)}||}.$$

Kuhn (1973) showed that this algorithm converges unless the starting point is in the domain of attraction of the data points. If the latter is the case, one can use the modification proposed by Vardi and Zhang (2000). However, in practice the simple algorithm above almost always converges.

4.2 Spatial median in feature space

Assume again that the inputs x_i are mapped into a high- (possibly infinite) dimensional feature space $\mathcal{H}$. Applying Definition 4 in feature space means that we want to find $\theta \in \mathcal{H}$ such that

$$\sum_{i=1}^{n} \frac{\Phi(x_i) - \theta}{||\Phi(x_i) - \theta||} = 0.$$

This is equivalent to demanding that

$$\left\| \sum_{i=1}^{n} \frac{\Phi(x_i) - \theta}{||\Phi(x_i) - \theta||} \right\|^2 = 0$$

or if we write out the norms as inner products

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left\langle \frac{\Phi(x_i) - \theta}{||\Phi(x_i) - \theta||}, \frac{\Phi(x_j) - \theta}{||\Phi(x_j) - \theta||} \right\rangle = 0$$

which is equivalent to

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\langle \Phi(x_i), \Phi(x_j) \rangle - \langle \theta, \Phi(x_j) \rangle - \langle \theta, \Phi(x_i) \rangle + \langle \theta, \theta \rangle}{\sqrt{\langle \Phi(x_i), \Phi(x_i) \rangle - 2\langle \Phi(x_i), \theta \rangle + \langle \theta, \theta \rangle} \sqrt{\langle \Phi(x_j), \Phi(x_j) \rangle - 2\langle \Phi(x_j), \theta \rangle + \langle \theta, \theta \rangle}} = 0. \quad (6)$$

If the mapping Φ is explicitly known, one could use this equation to find the center θ. In most kernel applications this is of course not the case. However, the spatial median naturally lies in the space spanned by the n inputs, and any point in this $\min(n, d)$-dimensional space can be parametrized as a linear combination of the inputs. Thus the spatial median can be written as

$$\theta = \sum_{k=1}^{n} \gamma_k \Phi(x_k). \quad (7)$$

Using this representation in (6) we find

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{\langle\Phi(x_i),\Phi(x_j)\rangle-\sum_{k=1}^{n}\gamma_k\langle\Phi(x_k),\Phi(x_j)\rangle}{\sqrt{A_i}\sqrt{A_j}}\right.$$
$$\left.-\frac{\sum_{k=1}^{n}\gamma_k\langle\Phi(x_k),\Phi(x_i)\rangle+\sum_{k=1}^{n}\sum_{l=1}^{n}\gamma_k\gamma_l\langle\Phi(x_k),\Phi(x_l)\rangle}{\sqrt{A_i}\sqrt{A_j}}\right\}=0 \tag{8}$$

with the notation

$$A_i=\langle\Phi(x_i),\Phi(x_i)\rangle-2\sum_{k=1}^{n}\gamma_k\langle\Phi(x_i),\Phi(x_k)\rangle+\sum_{k=1}^{n}\sum_{l=1}^{n}\gamma_k\gamma_l\langle\Phi(x_k),\Phi(x_l)\rangle.$$

Due to the parametrization of θ in (7), the spatial median can be expressed in terms of inner products only. Therefore this center can be computed in a kernel-induced feature space, using the same kernel trick as in kernel PCA replacing $\langle\Phi(u),\Phi(v)\rangle$ by $K(u,v)$.

Definition 5 *Given a sample of inputs $x_i\in\mathcal{X}$, $i=1,\ldots,n$ and a kernel function $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}:(u,v)\to K(u,v)$. Define the $n\times n$ kernel matrix as $\Omega_{i,j}=K(x_i,x_j)$. Denote $\Omega_{.,j}$ as the jth column of this matrix. Then the vector of coefficients $\gamma\in\mathbb{R}^n$ determining the spatial median in the kernel induced features space is defined by*

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\Omega_{i,j}-\gamma^t\Omega_{.,j}-\gamma^t\Omega_{.,i}+\gamma^t\Omega\gamma}{\sqrt{\Omega_{i,i}-2\gamma^t\Omega_{.,i}+\gamma^t\Omega\gamma}\sqrt{\Omega_{j,j}-2\gamma^t\Omega_{.,j}+\gamma^t\Omega\gamma}}=0.$$

To compute the vector γ the iterative algorithm in Section 4.1 can easily be modified to be computed in a kernel-induced feature space, only using the kernel inner product. Given an initial guess $\gamma^{(0)}\in\mathbb{R}^n$, iteratively define

$$\gamma^{(k)}=\frac{w}{\sum_{i=1}^{n}w_i}$$

where the components of the vector $w\in\mathbb{R}^n$ are given by

$$w_i=\frac{1}{\sqrt{\Omega_{i,i}-2(\gamma^{(k-1)})^t\Omega_{.,i}+(\gamma^{(k-1)})^t\Omega\gamma^{(k-1)}}}.$$

For the starting point we take the coefficients corresponding to the mean: $\gamma^{(0)}=(1/n,\ldots,1/n)\in\mathbb{R}^n$. In any data set we tried, the algorithm took 20 or less steps to converge to the solution giving 0 in the expression of Definition 4, indicating a similar good behavior as the original algorithm.

4.3 Centering the kernel matrix around the spatial median

The resulting center in the kernel feature space can of course not be computed. We do find the n coefficients γ_k such that the spatial median equals $\sum_{k=1}^{n}\gamma_k\Phi(x_k)$, but the feature map Φ is unknown. However, operations involving distances and inner products between feature vectors and the center often can be computed. A well known operation is for instance centering of the

data. Suppose we want to center the data in feature space around the spatial median. We define a new feature map as

$$\tilde{\Phi}(x) = \Phi(x) - \sum_{i=1}^{n} \gamma_i \Phi(x_i).$$

The corresponding centered kernel function K_c becomes

$$\begin{aligned} K_c(x,z) &= \langle \tilde{\Phi}(x), \tilde{\Phi}(z) \rangle \\ &= \langle \Phi(x) - \sum_{i=1}^{n} \gamma_i \Phi(x_i), \Phi(z) - \sum_{i=1}^{n} \gamma_i \Phi(z_i) \rangle \\ &= K(x,z) - \sum_{i=1}^{n} \gamma_i K(x,x_i) - \sum_{i=1}^{n} \gamma_i K(z,x_i) + \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \gamma_j K(x_i,x_j) \end{aligned}$$

or expressed in terms of matrix operations on the kernel matrix:

$$\Omega_{c,\text{median}} = \Omega - \gamma 1_n^t \Omega - \Omega 1_n \gamma^t + \gamma^t \Omega \gamma 1_n 1_n' \tag{9}$$

where 1_n is a vector containing 1 in its n entries.

Thus first computing γ as in Definition 4 with the algorithm from the previous paragraph and then computing (9), gives a robustly centered kernel matrix centered around the spatial median instead of the mean.

5 Spherical KPCA

5.1 Spherical PCA

Once the data is centered in an appropriate robust way, we can continue estimating the kernel principal components. We use the idea first mentioned in Locantore et al. (1999). Basically they project the data on a sphere around the L_1 median. Then the traditional components are computed for these projected data. Scores are computed by projecting the original, unsphered data on the principal directions. Figure 4 shows the algorithm in practice. Due to the sphering the influence of the outlier is obviously heavily reduced leading to principal components capturing the structure of the majority of the data much better. Marden (1999) shows that these spherical principal components are exactly equal to the original ones at population level for a rather large class of distributions.

5.2 Spherical PCA in feature space

Assume that $\gamma \in \mathbb{R}^n$ is the vector of coefficients determining the spatial median in feature space $\sum_{k=1}^{n} \gamma_k \Phi(x_k)$. In the first step we project all feature vectors onto the unit sphere around the spatial median, giving us new feature vectors

$$\Phi^*(x_i) = \frac{\Phi(x_i) - \sum_{k=1}^{n} \gamma_k \Phi(x_k)}{||\Phi(x_i) - \sum_{k=1}^{n} \gamma_k \Phi(x_k)||}. \tag{10}$$

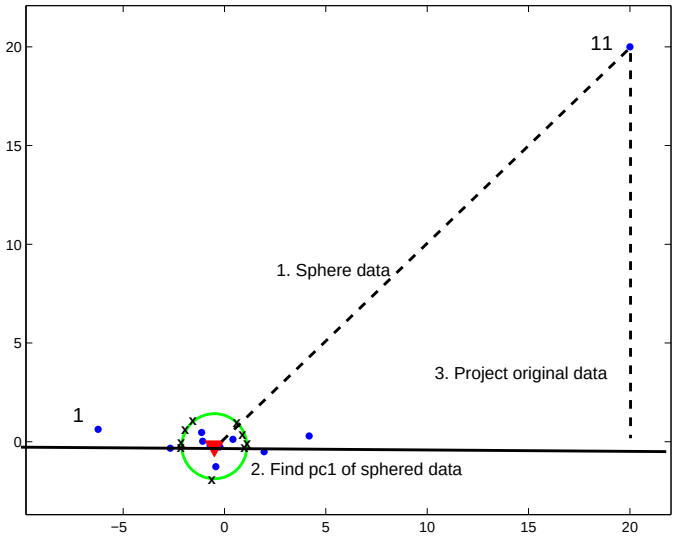


Figure 4: Spherical PCA in a simple 2-dimensional example.

This implies that

$$\langle \Phi^*(x_i), \Phi^*(x_j) \rangle = \left\langle \frac{\Phi(x_i) - \sum_{k=1}^n \gamma_k \Phi(x_k)}{||\Phi(x_i) - \sum_{k=1}^n \gamma_k \Phi(x_k)||}, \frac{\Phi(x_j) - \sum_{k=1}^n \gamma_k \Phi(x_k)}{||\Phi(x_j) - \sum_{k=1}^n \gamma_k \Phi(x_k)||} \right\rangle.$$

In terms of the original and uncentered kernel matrix Ω, this leads to a new kernel matrix Ω^* with entries

$$\begin{aligned} \Omega^*_{i,j} &:= \langle \Phi^*(x_i), \Phi^*(x_j) \rangle \\ &= \frac{\Omega_{i,j} - \gamma^t \Omega_{.,j} - \gamma^t \Omega_{.,i} + \gamma^t \Omega \gamma}{\sqrt{\Omega_{i,i} - 2\gamma^t \Omega_{.,i} + \gamma^t \Omega \gamma} \sqrt{\Omega_{j,j} - 2\gamma^t \Omega_{.,j} + \gamma^t \Omega \gamma}}. \end{aligned} \tag{11}$$

Thus once the spatial median is found, it is easy to compute the new kernel matrix Ω^* belonging to the sphered data based on the kernel matrix Ω of the original data.

In the second step, ordinary KPCA is applied to the sphered data. This means that we compute the eigenvectors and eigenvalues of Ω^* which we denote by $\alpha^{(k),*}$ resp. λ^*_k where $\lambda^*_1 \geq \lambda^*_2 \geq \ldots$ and $||\alpha^{(k),*}||^2 = 1$.

Thirdly the score $f^*_k(x)$ of any point x for the kth component is computed by

$$f^*_k(x) = \sum_{i=1}^n \frac{\alpha_i^{(k),*}}{\sqrt{\lambda_k}} \left\langle \Phi^*(x_i), \Phi(x) - \sum_{l=1}^n \gamma_l \Phi(x_l) \right\rangle.$$

Using (10) leads to the following result.

Algorithm 2 (Spherical Kernel PCA) *Given a sample $x_1, \ldots, x_n \in \mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with corresponding feature space $\mathcal{H}$. Then the kth Sperical KPCA score function f^*_k at $x \in \mathcal{X}$ equals:*

$$f^*_k(x) = \sum_{i=1}^n \frac{\alpha_i^{(k),*}}{\sqrt{\lambda^*_k}} \frac{K(x_i, x) - \sum_{l=1}^n \gamma_k K(x_l, x) - \gamma^t \Omega_{.,i} + \gamma^t \Omega \gamma}{\sqrt{\Omega_{i,i} - 2\gamma^t \Omega_{.,i} + \gamma^t \Omega \gamma}}. \tag{12}$$

with $\alpha^{(k),*}$ the eigenvector belonging to the kth largest eigenvalue λ_k^* of the median centered and sphered kernel matrix Ω^* with entry i, j equal to

$$\Omega^*_{i,j} := \frac{\Omega_{i,j} - \gamma^t\Omega_{.,j} - \gamma^t\Omega_{.,i} + \gamma^t\Omega\gamma}{\sqrt{\Omega_{i,i} - 2\gamma^t\Omega_{.,i} + \gamma^t\Omega\gamma}\sqrt{\Omega_{j,j} - 2\gamma^t\Omega_{.,j} + \gamma^t\Omega\gamma}}. \tag{13}$$

These are the Spherical KPCA scores that provide a robust alternative to the classical KPCA scores from Algorithm 1. Note that the computational complexity of both algorithms is essentially the same. The most time consuming part consists of finding eigenvectors and eigenvalues of a $n \times n$ matrix. Also notice that Spherical KPCA, just like classical KPCA, only depends on the kernel matrix. No additional tuning constants are needed. This is a nice feature of the sphering concept compared to other types of robustification. Many robust algorithms for instance use reweighting. Then an appropriate weight function has to be chosen often implying an priori assumption on the distribution of the data. Especially in a general kernel induced feature space, this would be a difficult choice to make.

6 Visualizing influential observations

The spherical KPCA scores themselves can be useful in many applications. The example in Figure 1 shows for instance that spherical KPCA with a linear kernel (dashed line) produces scores that capture the structure of the majority of the data much better than ordinary KPCA (solid line). However, in high dimensions it is more difficult to visualize the difference between both methods. In this section we propose a simple graphical display to assess the influence of observations with respect to ordinary KPCA. Our strategy is to use the spherical KPCA estimates in the expressions for the influence function in Theorem 1. For linear PCA this idea was applied by Pison and Van Aelst (2004). We use the score function $f_k^*(x)$ as a sample estimate of $e_{P,k}$. However, as explained by Marden (1999), the spherical eigenvalues λ_k^* are not always good estimates of $\lambda_{P,k}$. But since $\lambda_{P,k}$ equals the variance of the score function, we can re-estimate these eigenvalues by a measure of spread of the scores at the data points. Of course this measure of spread should not be influenced too much by individual observations either. We use the robust Median Absolute Deviation (MAD) to define

$$\lambda_k^{**} := \mathrm{MAD}(f_k^*(x_i)) = \left(\mathrm{median}(|f_k^*(x_i) - \mathrm{median}(f_k^*(x_i))|)\right)^2. \tag{14}$$

Other options, e.g. the Q_n-estimator (Rousseeuw and Croux, 1993), are possible as well of course. Now observe from Theorem 1 that

$$||IF(z; e_k, P)|| = |\langle e_{P,k}, \Phi(z)\rangle| \sqrt{\sum_{j=1, j\neq k}^{\infty} \frac{\langle e_{P,j}, \Phi(z)\rangle^2}{(\lambda_k - \lambda_j)^2}}$$

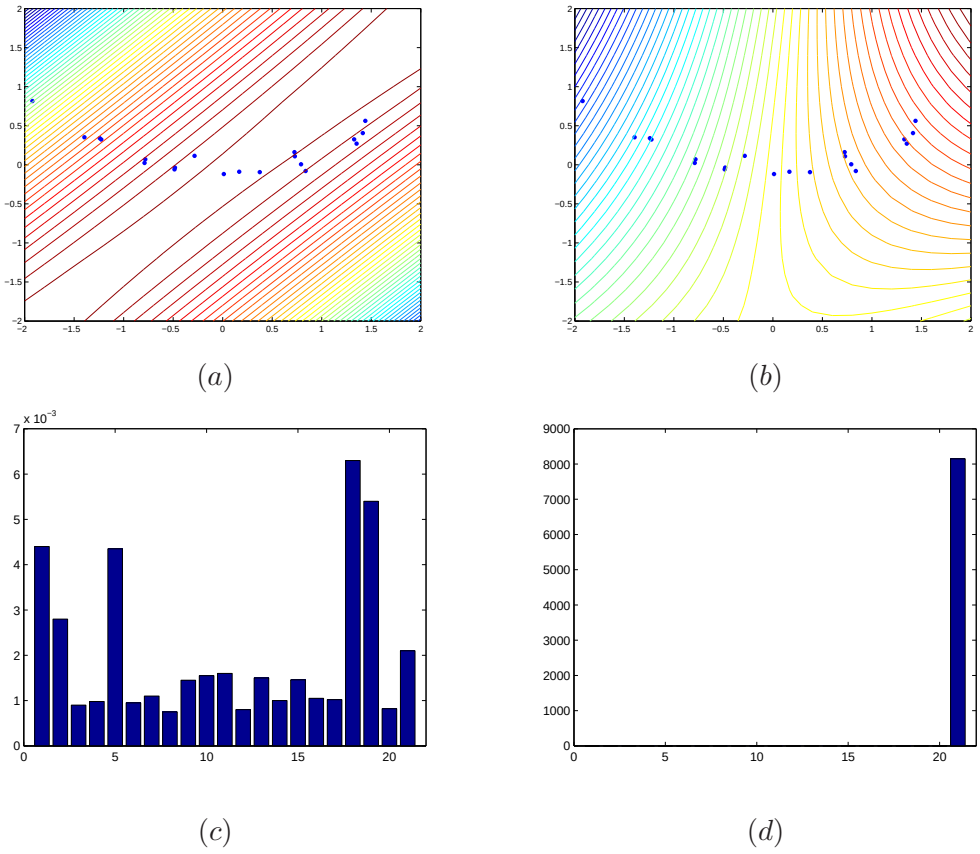
which gives us the following sample based influence diagnostic equal to the norm of the empirical influence function at z of the kth component:

$$||EIF_k(z)|| = |f_k^*(z)| \sqrt{\sum_{j=1, j\neq k}^{n} \frac{f_j^*(z)^2}{(\lambda_k^{**} - \lambda_j^{**})^2}}. \tag{15}$$

To obtain the influence of an observation x_i, just take $z = x_i$. A bar plot of these values for all 11 observations in the sample from the example in Figure 1 is shown in Figure 2(b), for a linear kernel and the first component ($k = 1$). Again note how using the classical PCA scores fails to provide an accurate description of the data (Figure 2(a)). For linear PCA other robust methods would be able to give good results as well. Our method however can deal with any type of kernel. In the next section we show some examples where spherical KPCA is able to detect influential observations in more general kernel based frameworks.

7 Examples

7.1 Toy example



(a) (b) (c) (d)

Figure 5: Score-contours and estimated influences for (a) – (c) classical KPCA, (b) – (d) spherical KPCA.

We explained the methodology on a toy example for the linear kernel (see Figure 1). Consider now a second toy example where the underlying structure of the data is not linear. In Figure 5 we generated 20 data points showing a quadratic curvature, together with 1 outlier at $(-5, 5)$ (not visible on the plot). If we construct the score-contours corresponding to the first principal component of ordinary KPCA with a polynomial kernel of degree 2, we obtain Figure 5(a). Clearly the quadratic structure is lost completely due to the single outlier. For

Spherical KPCA, Figure 5(b) depicts the corresponding score-contours. Now the quadratic structure of the majority of observations is learned, despite the outlier. Two bar plots of the empirical influence function from equation (15) for the 21 observations are shown in Figure 5 (c) using classical KPCA and (d) using spherical KPCA. A comparison of both plots visualizes the large difference between both methods. The diagnostic plot created with spherical KPCA (Figure 5(d)) correctly reveals tha observation 21 (the outlier) causes this difference and that it is highly influential for classical KPCA.

7.2 String kernel

Consider a situation where the inputs are no vectors, but strings. Then many kernels exist that can be used to identify patterns in this set of strings. Here we concentrate on one example, i.e. the all-subsequence kernel. Then the strings are represented by feature vectors of which each component represents a possible substring. For the three strings "gca", "cag" and "ggc" for instance the corresponding feature vectors are:

	$\emptyset$	a	c	g	ag	ca	cg	ga	gc	gg	gca	cag	ggc
gca	1	1	1	1	0	1	0	1	1	0	1	0	0
cag	1	1	1	1	1	1	1	0	0	0	0	1	0
ggc	1	0	1	2	0	0	0	0	1	1	0	0	1

So in this case every string can be represented as a vector with 13 components, and thus analysis could proceed in a 13-dimensional space. However, this example is extremely simple, since there are only three possible characters (a,c,g) and only strings of size three are considered. Unfortunately the dimension of the feature space increases exponentially with the size of the strings. For longer strings the explicit computation of the feature vectors thus becomes infeasible. However, when using a kernel method these explicit representations are not necessary. All we need are the inner products between any two feature vectors. For the all-subsequence kernel the kernel matrix containing these inner products can be computed with fast recursive algorithms (Shawe-Taylor and Cristianini, 2004). Since sperical KPCA does not require explicit feature vectors either, but only the kernel matrix, applying the methodology from the previous sections is straightforward.

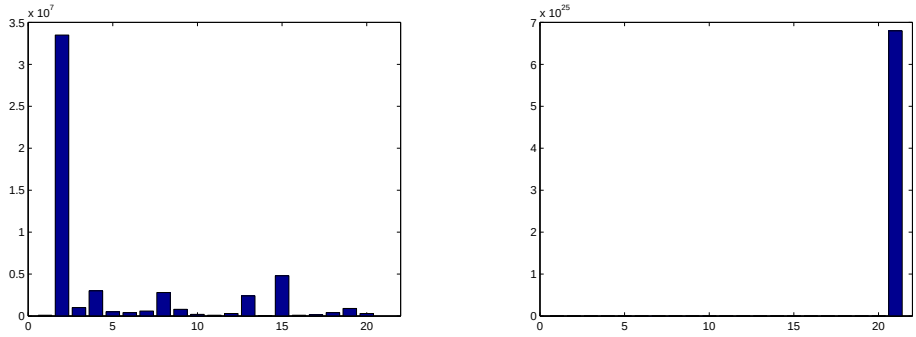
As an example take the first 20 DNA sequences in the 'Splice-junction gene sequences' database from the UCI database (http://www.ics.uci.edu/∼mlearn/MLRepository.html). This gives us 20 observations, all strings of size 60 composed out of 4 characters (A,C,G,T). The first 3 elements are shown below.

'CCAGCTGCATCACAGGAGGCCAGCGAGCAGGTCTGTTCCAAGGGCCTTCGAGCCAGTCTG',
'AGACCCGCCGGGAGGCGGAGGACCTGCAGGGTGAGCCCCACCGCCCCTCCGTGCCCCCGC',
'GAGGTGAAGGACGTCCTTCCCCAGGAGCCGGTGAGAAGCGCAGTCGGGGGCACGGGGATG'.

As an example we add one strange string to the data set, observation 21, which is the following sequence:

'CCCCCCCCCCCCCCCCAAAAAAAAAAAAAATTTTTTTTTTTTTTTGGGGGGGGGGGGGGGGG'.

Although the length of this string and the number of A's,C's,G's and T's are both similar as for the other strings, this new observation 21 is clearly different due to the specific order of the characters. Next we perform KPCA and spherical KPCA on this data set with the all-subsequence kernel. For each string we compute its influence measure as in (15) with respect to the first principal component. Figure 6(a) shows the result if we use the original KPCA scores and eigenvalues. String number 2 comes out as the most influential observation. Nevertheless it does not look extremely dominating and one would probably not suspect big problems. One would certainly not detect that observation 21 is an exceptional string, since its influence measure is very small. The results using spherical KPCA are depicted in Figure 6(b). Then it



(a) (b)

Figure 6: Estimated influences based on (a) KPCA, (b) spherical KPCA.

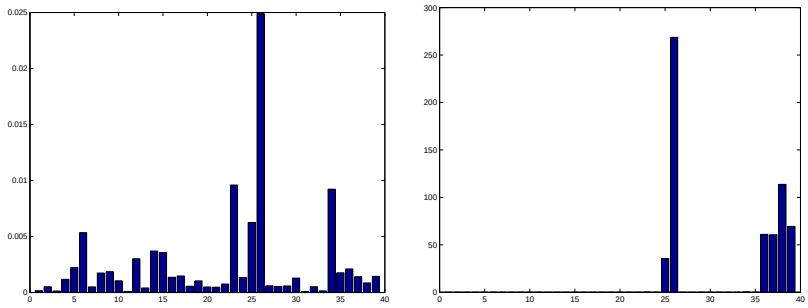
is immediately clear that we have the same effect we discussed for the simple toy example in Figures 1 and 2. Observation 21 is in reality extremely influential, dominating the estimation of the ordinary first kernel principal component completely. This first pc is completely attracted by string 21. Therefore using this component results in a misleading plot of the influences. Only by using the spherical kernel principal components a correct assessment can be made about observations deviating from the mainstream. Also note that robust linear PCA methods cannot be used. They require the explicit feature vectors corresponding to the strings. However, according to Shawe-Taylor and Cristianini (2004), the dimension of these feature vectors would be likely to exceed 4^{30} in this example, which is obviously infeasible.

7.3 Octane data

The next example is the octane data set described in Esbensen et al. (1994). It contains near-infrared (NIR) absorbance spectra over 226 wavelengths of $n = 39$ gasoline samples with certain octane numbers. It is known that six of the samples $(25, 26, 36 - 39)$ contain added alcohol. The

data set was also analyzed in Hubert et al. (2005), where it was shown that the robust linear PCA method ROBPCA was able to detect the six outlying samples in contrast to ordinary linear PCA. Now suppose that we increase the difficulty of the problem by using a polynomial kernel of degree 2. In theory the corresponding feature vectors could be computed by taking appropriately weighted squares and cross-products for all 226 variables. In practice the resulting dimension of these feature vectors will again be way too high. Explicitly calculating quadratic forms and then applying a robust method such as ROBPCA in feature space is thus infeasible.

Using a kernel method avoids this problem. All we need is the 39×39 dimensional kernel matrix, both for ordinary as spherical kernel PCA. The resulting diagnostic plots are shown in Figure 7. Part (a) of this plot depicts the results using ordinary KPCA. Of course some



(a) (b)

Figure 7: Octane data: estimated influences based on (a) KPCA, (b) spherical KPCA.

points seem more influential than others, but no dramatic effects would be detected. Part (b) shows the influence measures using spherical KPCA. Now we see what is really happening: six observations are extremely influential, dominating all others. These six observations are exactly the outlying samples that contain alcohol.

8 Conclusion

We investigated the effect of small amounts of contamination on classical Kernel PCA by calculating expressions for the influence function. We showed that the influence function can be unbounded if an unbounded kernel is used. Spherical Kernel PCA is proposed as a more robust alternative. The resulting algorithm is fast and only depends on the choice of the kernel, as in classical KPCA, so no additional tuning parameters are needed. We use this robust method to detect influential points in classical KPCA by a simple diagnostic plot. Examples illustrate that large differences can appear when outliers are present in the data. Using Spherical KPCA these outliers can be detected, whereas classical KPCA fails to do so.

References

G. Blanchard, O. Bousquet, and L. Zwald. Statistical properties of kernel principal component analysis. *Machine Learning*, 33:259–294, 2007.

A. Christmann and I. Steinwart. Consistency and robustness of kernel based regression. *Bernoulli*, 13:799–819, 2007.

A. Christmann and I. Steinwart. On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research*, 5:1007–1034, 2004.

F. Critchley. Influence in principal component analysis. *Biometrika*, 72:627–636, 1985.

C. Croux and A. Ruiz-Gazen. High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95:206–226, 2005.

M. Debruyne, A. Christmann, M. Hubert, and J. Suykens. Robustness and stability of reweighted kernel based regression. Technical report TR 06-09, Katholieke Universiteit Leuven, Department of Mathematics, Section of Statistics, 2006.

K.H. Esbensen, S. Schönkopf, and T. Midtgaard. *Multivariate Analysis in Practice*. Camo, Trondheim, 1994.

J.C. Gower. The mediancentre. *Applied Statistics*, 23:466–470, 1974.

J.B.S. Haldane. Note on the median of a multivariate distribution. *Biometrika*, 35:414–415, 1948.

F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York, 1986.

O. Hössjer and C. Croux. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. *Non-parametric statistics*, 4:293–308, 1995.

P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

M. Hubert, P.J. Rousseeuw, and S. Verboven. A fast robust method for principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 60: 101–111, 2002.

M. Hubert, P.J. Rousseeuw, and K. Vanden Branden. ROBPCA: a new approach to robust principal components analysis. *Technometrics*, 47:64–79, 2005.

H.W. Kuhn. A note on Ferniat's problem. *Mathematical Programming*, 4:98–107, 1973.

N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, and K.L. Cohen. Robust principal component analysis for functional data. *Test*, 8:1–73, 1999.

J.I. Marden. Some robust estimates of principal components. *Statistics and Probability Letters*, 43:349–359, 1999.

R.A. Maronna. Principal components and orthogonal regression based on robust scales. *Technometrics*, 47:264–273, 2005.

R. Phelps. *Convex functions, monotone operators and differentiability, volume 1364 of Lecture notes in math.* Springer, 1986.

G. Pison and S. Van Aelst. Diagnostic plots for robust multivariate methods. *Journal of Computational and Graphical Statistics*, 13:310–329, 2004.

P.J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283, 1993.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

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

J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, Cambridge, 2004.

J. Shawe-Taylor, C. Williams, N. Cristianini, and J. Kandola. Eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic learning theory: 13th international conference, ALT2002 of lecture notes in computer science*, volume 2533, pages 23–40. Springer-Verlag, 2002.

C.G. Small. A survey of multidimensional medians. *International Statistical Review*, 58:263–277, 1990.

Y. Vardi and C.-H. Zhang. The multivariate l1-median and associated data depth. *Proceedings of the National Academy of Sciences of the United States*, 97:1423–1426, 2000.