

# Outlier Detection Using the Smallest Kernel Principal Components

Alan J. IZENMAN and Yan SHEN

---

The smallest principal components have not attracted much attention in the statistics literature. This apparent lack of interest is due to the fact that, compared with the largest principal components that contain most of the total variance in the data, the smallest principal components only contain the noise of the data and, therefore, appear to contribute minimal information. However, because outliers are a common source of noise, the smallest principal components should be useful for outlier detection. This article proposes a new and novel method for outlier detection using the smallest kernel principal components in a feature space induced by the radial basis function kernel. We show that the eigenvectors corresponding to the smallest kernel principal components can be viewed as those for which the residual sum of squares is minimized, and we use those components to identify outliers based upon a simple graphical technique. A threshold between “large” and “small” kernel principal components is proposed in this paper, and a nonparametric method for locating the smallest kernel principal component is suggested. Simulation studies show that under a univariate outlier situation, the proposed method is as good as the best method available, and real-data examples also suggest that this method is often better.

KEY WORDS: Kernel methods; Kernel principal component analysis; Mahalanobis distance; Multivariate outlier detection; Principal component analysis.

---

## Authors' Footnote

Alan J. Izenman is Professor of Statistics in the Department of Statistics, and Director of the Center for Statistical and Information Science, Office of the Vice-President for Research and Graduate Studies, Temple University, Philadelphia, PA 19122 (E-mail: *alan@temple.edu*). Yan Shen is biostatistician in the Department of Biometrics and Clinical Informatics at Johnson and Johnson Pharmaceutical Research & Development, L.L.C., New Jersey, NJ 08869 (E-mail: *yshen10@prdus.jnj.com*).

## 1. INTRODUCTION

Principal component analysis (PCA), first introduced by Hotelling (1933), is a well-established dimension-reduction method. It replaces a set of correlated variables by a smaller set of uncorrelated linear combinations of those variables, such that these linear combinations explain most of the total variance. It is also a way of identifying inherent patterns, relations, regularities, or structure in the data. Because such patterns are difficult to detect in high-dimensional data, PCA can be a powerful tool.

As a linear statistical technique, PCA cannot accurately describe all types of structures in a given dataset, specially nonlinear structures. Kernel principal component analysis (KPCA) has recently been proposed as a nonlinear extension of PCA (Schölkopf, Smola, and Müller, 1998). See also Schölkopf and Smola (2002). Kernel methods were introduced into the computer science literature specifically for pattern analysis (see, e.g., Shawe-Taylor and Cristianini, 2004). These are powerful techniques, which have been applied to many types of statistical methods, including support vector machines, canonical correlation analysis, and independent component analysis.

KPCA maps the data from the original space into a feature space via a nonlinear transformation, and then performs linear PCA on the mapped data. The principal components (PCs) found by this process are called kernel principal components (KPCs). In pattern analysis, a feature space is an abstract space where each transformed sample is represented as a point in  $n$ -dimensional space. Because the feature space could be of very high dimension, possibly even infinite dimensional, kernel methods employ inner-products instead of carrying out the mapping explicitly. Mercer's theorem of functional analysis (Mercer, 1909) states that if  $k$  is a continuous kernel of a positive integral operator, we can construct a mapping into

a space where  $k$  acts as an inner-product.

Much of the linear PCA literature (and also that of KPCA) has been focused primarily on the largest few PCs because they account for most of the variation in the data, as opposed to the smallest PCs, which account for only the noise in the data. Yet, outliers can be viewed as a common source of noise. In this article, we study a new approach to multivariate outlier detection using the smallest KPCs. We show that the eigenvectors corresponding to the smallest KPCs can be viewed as those for which the residual sum of squares is minimized, and we propose a threshold (or cutoff) value, which distinguishes “large” from “small” KPCs. A nonparametric method is suggested to determine the location of the smallest KPC. We can then detect outliers easily by plotting the smallest KPC against the second smallest KPC. We show that this approach is competitive with (and often an improvement over) existing methods.

The remainder of this article is organized as follows. In Section 2, we introduce the necessary background on linear PCA and KPCA. We then present in Section 3 the proposed outlier detection method in detail. Numerical results are given in Section 4, using both simulated data and real-data examples, and conclusions are drawn in Section 5.

## 2. PCA AND KPCA

### 2.1 Principal Component Analysis

Given a  $p$ -vector,  $\mathbf{x} = (x_1, \dots, x_p)^\tau$ , in input space  $\mathcal{X}$ , PCA finds uncorrelated linear combinations of the  $p$  variables such that fewer than  $p$  of these combinations contain most of the variation in the data. Assume that  $\mathbf{x}$  has zero mean and  $(p \times p)$  covariance matrix  $\Sigma$ , whose eigenvalue-eigenvector pairs are given by

$(\mu_1, \mathbf{u}_1), (\mu_2, \mathbf{u}_2), \dots, (\mu_p, \mathbf{u}_p)$ , where  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_p \geq 0$ . Then,

$$\xi_1 = \mathbf{u}_1^T \mathbf{x}, \xi_2 = \mathbf{u}_2^T \mathbf{x}, \dots, \xi_p = \mathbf{u}_p^T \mathbf{x}, \quad (1)$$

are the  $p$  PCs, where

$$\text{var}\{\xi_i\} = \mathbf{u}_i^T \mathbf{\Sigma} \mathbf{u}_i = \mu_i \text{ and } \text{cov}\{\xi_i, \xi_j\} = \mathbf{u}_i^T \mathbf{\Sigma} \mathbf{u}_j = 0, \quad j < i. \quad (2)$$

Note that in (1) and (2),  $\mathbf{u}^T$  is the transpose of the column-vector  $\mathbf{u}$ . Principal component analysis orthogonalizes the covariance matrix  $\mathbf{\Sigma}$  into a maximum of  $p$  PCs. The first PC is the linear combination of the variables that explains the greatest amount of the total variation in  $\mathbf{x}$ . The second PC is the linear combination of the variables that explains the next largest amount of variation and is uncorrelated with the first PC, and so on. If the first  $l$  (say, three) components contain most of the total variation (say, 90%), then the original variables can be replaced by these components without too much loss of variance information. The coefficients of the PCs (i.e., the elements of the  $\xi$ s) can be derived through least-squares optimization or by the use of Lagrangian multipliers. For details, see Izenman (2008, Section 7.2).

## 2.2 Kernel PCA

Instead of reducing the dimensionality directly in the original space, KPCA works in a higher-dimensional feature space by forming inner-products of a transformation function  $\Phi$ . A mapping is performed through  $\Phi : \mathcal{R}^p \rightarrow \mathcal{H}$ , where the original data lie in  $\mathcal{R}^p$  and the features lie in a Hilbert space  $\mathcal{H}$ . The transformation  $\Phi$  is usually nonlinear, but the remarkable thing is that it does not have to be specified explicitly. By Mercer's theorem, under suitable conditions of nonnegative-definiteness, the feature space  $\mathcal{H}$  induced by the kernel function,

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle, \quad (3)$$

exists and can be constructed from eigenfunctions and positive eigenvalues.

Assume that the data in feature space are centered; that is,  $\sum_{i=1}^n \Phi(\mathbf{x}_i) = 0$ . Then, the estimated covariance matrix is  $\mathbf{\Omega} = \frac{1}{n} \sum_{i=1}^n \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_i)^\tau$ . To find KPCs in this feature space, we solve the eigenequations,  $\lambda \mathbf{v} = \mathbf{\Omega} \mathbf{v}$ , for nonzero eigenvalues  $\lambda$  and corresponding eigenvectors  $\mathbf{v}$ . Note that all solutions  $\mathbf{v}$  with  $\lambda \neq 0$  lie in the span of  $\Phi(\mathbf{x}_1), \dots, \Phi(\mathbf{x}_n)$ . This implies that (1) the eigenequations can be written as  $\lambda \langle \Phi(\mathbf{x}_i), \mathbf{v} \rangle = \langle \Phi(\mathbf{x}_i), \mathbf{\Omega} \mathbf{v} \rangle, i = 1, 2, \dots, n$ , and (2) there exist coefficients  $\alpha_i, i = 1, 2, \dots, n$ , such that  $\mathbf{v} = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i)$ . This enables us to rewrite the eigenequations as

$$\lambda \sum_{i=1}^n \alpha_i \langle \Phi(\mathbf{x}_k), \Phi(\mathbf{x}_i) \rangle = \frac{1}{n} \sum_{i=1}^n \alpha_i \left\langle \Phi(\mathbf{x}_k), \sum_{j=1}^n \Phi(\mathbf{x}_j) \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}_i) \rangle \right\rangle, \quad (4)$$

which, in matrix form, becomes  $n\lambda \mathbf{K} \boldsymbol{\alpha} = \mathbf{K}^2 \boldsymbol{\alpha}$ , where  $\mathbf{K} = (K_{ij}), K_{ij} = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ , is an  $(n \times n)$  Gram matrix, and  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\tau$ .

Because  $\mathbf{K}$  is symmetric, it has a set of eigenvectors that spans the whole space. Thus,  $n\lambda \boldsymbol{\alpha} = \mathbf{K} \boldsymbol{\alpha}$  gives all solutions  $\boldsymbol{\alpha}$  of the eigenvectors and  $n\lambda$  of the eigenvalues. For the sake of simplicity, let  $\lambda_i$  represent the eigenvalues of  $\mathbf{K}$  equivalent to  $n\lambda_i$ , where  $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_l$ , with  $\lambda_l$  being the last nonzero eigenvalue, and  $\boldsymbol{\alpha}^1, \dots, \boldsymbol{\alpha}^l$  the corresponding eigenvectors. We normalize the  $\boldsymbol{\alpha}^1, \dots, \boldsymbol{\alpha}^l$  by requiring that the corresponding vectors in  $\mathcal{H}$  be normalized, i.e.,  $\langle \mathbf{v}^k, \mathbf{v}^k \rangle = 1, k = 1, 2, \dots, l$ . This can be translated into a normalization condition for  $\boldsymbol{\alpha}^1, \dots, \boldsymbol{\alpha}^l$ :

$$1 = \sum_{i,j=1}^n \alpha_i^k \alpha_j^k \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = \sum_{i,j=1}^n \alpha_i^k \alpha_j^k K_{ij} = \langle \boldsymbol{\alpha}^k, \mathbf{K} \boldsymbol{\alpha}^k \rangle = \lambda_k \langle \boldsymbol{\alpha}^k, \boldsymbol{\alpha}^k \rangle. \quad (5)$$

For a point  $\mathbf{x}$  in the original space  $\mathcal{R}^p$  with an image  $\Phi(\mathbf{x})$  in the feature space  $\mathcal{H}$ , the projection

$$\langle \mathbf{v}^k, \Phi(\mathbf{x}) \rangle = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^n \alpha_i^k \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle \quad (6)$$

can be called its  $k$ th KPC corresponding to  $\Phi, k = 1, 2, \dots, l$ .

In practice, we cannot assume that the points in feature space have mean zero. Therefore, we subtract  $n^{-1} \sum_{i=1}^n \Phi(\mathbf{x}_i)$  from all points. This leads to a minor change of the Gram matrix  $\mathbf{K}$ , namely  $\mathbf{K}^* = \mathbf{H}\mathbf{K}\mathbf{H}$ , where  $\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{J}_n$  is the centering matrix,  $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^\tau$ , and  $\mathbf{1}_n$  is an  $n$ -vector of all ones.

### 3. OUTLIER DETECTION USING THE SMALLEST PRINCIPAL COMPONENTS

The earliest work that incorporates the smallest PCs occurs in Gnanadesikan and Wilk’s (1966, 1968) generalization of PCA to the nonlinear case. Building upon these ideas, Gnanadesikan and Kettenring (1972) state that “with  $p$ -dimensional data, the projection onto the ‘smallest’ principal component would be relevant for studying the deviation of an observation from a hyperplane of closest fit, while the projections on the ‘smallest’  $q$  principal component coordinates would be relevant for studying the deviation of an observation from a fitted linear subspace of dimensionality  $p-q$ .” See also Gnanadesikan (1977, Section 2.4.2).

More recently, Donnell, Buja and Stuetzle (1994) introduced the concept of “additive principal components” (APCs) to help identify additive dependencies and concavities among the predictor variables in a regression model. The smallest additive principal component is defined as an additive function of the data,  $\sum_i \phi_i(x_i)$ , with smallest variance. The function  $\phi_i$  is usually nonlinear and may be different for each  $x_i$ . If the  $\phi_i$ s are linear and identical, then the problem reduces to standard PCA. They outlined some analytical methods for theoretical calculations of APCs, and showed that by plotting the last few smallest APCs against the raw variables, they could discover any existing nonlinear dependencies between those variables.

Our approach is different from these authors in that we use the kernel function  $k$  to transform the data from input space to feature space, and then apply standard

PCA to find the smallest KPCs in that space. From Section 2, we know that if the kernel functions satisfy Mercer’s conditions, we are doing a standard PCA in feature space. Consequently, all mathematical and statistical properties of PCA carry over to KPCA (Schölkopf and Smola, 2002).

### 3.1 Definitions

We characterize the smallest KPC by extending a definition of the smallest PC. For the smallest KPC,  $\mathbf{v}^n$  is the eigenvector corresponding to the smallest eigenvalue of the eigenequation  $\lambda \mathbf{v} = \mathbf{\Omega} \mathbf{v}$ . Because  $\Phi$  is usually unknown, we are unable to calculate  $\mathbf{\Omega}$  directly and solve for  $\mathbf{v}^n$ . Instead, we replace  $\mathbf{v}$  by  $\sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i)$  in an equivalent eigenequation  $\lambda \langle \Phi(\mathbf{x}_i), \mathbf{v} \rangle = \langle \Phi(\mathbf{x}_i), \mathbf{\Omega} \mathbf{v} \rangle$ . Hence, in this characterization, we need to find  $\boldsymbol{\alpha}^n$ , the eigenvector corresponding to the smallest eigenvalue  $\lambda_n$  of the eigenequation  $n\lambda \boldsymbol{\alpha} = \mathbf{K} \boldsymbol{\alpha}$ .

KPCA could find up to  $n$  non-zero distinct KPCs. However, in practice, when  $n$  is large, eigenvalues equal to or approximately equal to 0 are likely to occur, so that the corresponding KPCs contain almost no information, even about the noise. Therefore, finding the smallest KPC is not a trivial task. Hence, we will use the notation  $\mathbf{v}^l(\boldsymbol{\alpha}^l)$  instead of  $\mathbf{v}^n(\boldsymbol{\alpha}^n)$  to represent the eigenvector corresponding to the smallest eigenvalue of the eigenequation, where  $l$  needs to be determined. Now we are ready to define the smallest KPC.

*Definition 1.* In a feature space properly defined through a kernel function  $k$ , the smallest kernel principal component is a random vector  $\boldsymbol{\phi}^l = \sum_{i,j=1}^n \alpha_i^l \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ , which has smallest variance subject to  $\lambda_l \langle \boldsymbol{\alpha}^l, \boldsymbol{\alpha}^l \rangle = 1$ , where  $\boldsymbol{\alpha}^l$  is the eigenvector corresponding to the smallest non-zero eigenvalue  $\lambda_l$  of the eigenequation  $n\lambda \boldsymbol{\alpha} = \mathbf{K} \boldsymbol{\alpha}$ ,  $\mathbf{K} = (K_{ij})$ , and  $K_{ij} = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$ .

Analogously, we can define the smallest KPCs with the additional constraint

that they are mutually uncorrelated.

*Definition 2.* In a feature space properly defined through a kernel function  $k$ , the  $m$ th-smallest kernel principal component is a random vector

$$\boldsymbol{\phi}^{(m)} = \sum_{i,j=1}^n \alpha_i^{(m)} \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \quad (7)$$

with smallest variance, subject to the constraints

$$\lambda_{(m)} \langle \boldsymbol{\alpha}^{(m)}, \boldsymbol{\alpha}^{(m)} \rangle = 1, \quad \text{Cov}(\boldsymbol{\phi}^{(m)}, \boldsymbol{\phi}^{(t)}) = 0, \quad t = m - 1, m - 2, \dots, 1. \quad (8)$$

Here,  $\boldsymbol{\alpha}^{(m)}$  is the eigenvector corresponding to the  $m$ th-smallest non-zero eigenvalue  $\lambda_{(m)}$  of the eigenequation  $n\lambda\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}$ , and  $K_{ij} = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = k(\mathbf{x}_i, \mathbf{x}_j)$ .

Both PCA and KPCA boil down to eigenproblems. Therefore, to understand the use of the smallest KPCs, we study their eigenproperties.

### 3.2 Eigenvectors Corresponding to the Smallest KPCs

The Gram matrix  $\mathbf{K}$  is the matrix whose elements are the inner-products of the mapped data;  $\mathbf{K}$  is symmetric and nonnegative-definite if the kernel function  $k$  satisfies Mercer's conditions. For the sake of simplicity, let  $\mathbf{K} = \mathbf{X}\mathbf{X}^\tau$ , where  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$  is the  $(p \times n)$  data matrix whose  $i$ th column  $\mathbf{x}_i \in \mathcal{X}$ ,  $i = 1, 2, \dots, n$ .

We start with the singular value decomposition of  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\tau$ , where  $\mathbf{U}$  and  $\mathbf{V}$  are orthonormal matrices and  $\mathbf{D}$  is a diagonal matrix containing singular values in descending order of magnitude. We represent  $\mathbf{K}$  as  $\mathbf{K} = \mathbf{X}\mathbf{X}^\tau = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\tau$ , where  $\boldsymbol{\Lambda} = \mathbf{D}^2$ . Alternatively, we can construct a matrix  $\mathbf{K}^* = \mathbf{X}^\tau\mathbf{X} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\tau$ , with the same eigenvalues as  $\mathbf{K}$ . By the Courant–Fischer Min-Max Theorem (McDiarmid, 1989) and the definition of a projection operator, the first (or largest) eigenvalue of

$\mathbf{K}$  is given by

$$\begin{aligned}
\lambda_1(\mathbf{K}) &= \max_{0 \neq \alpha \in \mathcal{R}^n} \frac{\alpha^\top \mathbf{K}^* \alpha}{\alpha^\top \alpha} \\
&= \max_{0 \neq \alpha \in \mathcal{R}^n} \frac{\alpha^\top \mathbf{X}^\top \mathbf{X} \alpha}{\alpha^\top \alpha} \\
&= \max_{0 \neq \alpha \in \mathcal{R}^n} \frac{\|\mathbf{X}\alpha\|^2}{\alpha^\top \alpha} \\
&= \max_{0 \neq \alpha \in \mathcal{R}^n} \sum_{i=1}^n \|\mathbf{P}_\alpha(\mathbf{x}_i)\|^2 \\
&= \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \min_{0 \neq \alpha \in \mathcal{R}^n} \sum_{i=1}^n \|\bar{\mathbf{P}}_\alpha(\mathbf{x}_i)\|^2, \tag{9}
\end{aligned}$$

where  $\mathbf{P}_\alpha(\mathbf{x})$  is the projection of  $\mathbf{x}$  onto the space spanned by  $\alpha$ , and  $\bar{\mathbf{P}}_\alpha(\mathbf{x})$  is the projection of  $\mathbf{x}$  onto the space perpendicular to  $\alpha$ . Equation (9) suggests that the first eigenvector can be characterized as the direction for which the residual sum of squares is minimized. Applying the same line of reasoning to the general form of the Courant-Fischer Min-Max Theorem, the  $m$ th eigenvalue of  $\mathbf{K}$  can be expressed as

$$\lambda_m(\mathbf{K}) = \max_{\dim(\mathcal{T})=m} \min_{0 \neq \alpha \in \mathcal{T}} \sum_{i=1}^n \|\mathbf{P}_\alpha(\mathbf{x}_i)\|^2, \tag{10}$$

which implies that if  $\alpha^m$  is the  $m$ th eigenvector of  $\mathbf{K}^*$ , then

$$\lambda_m(\mathbf{K}) = \sum_{i=1}^n \|\mathbf{P}_{\alpha^m}(\mathbf{x}_i)\|^2. \tag{11}$$

Consequently, if  $\mathcal{T}_m$  is the space spanned by the first  $m$  eigenvectors, then

$$\sum_{j=1}^m \lambda_j(\mathbf{K}) = \sum_{i=1}^n \|\mathbf{P}_{\mathcal{T}_m}(\mathbf{x}_i)\|^2 = \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \sum_{i=1}^n \|\bar{\mathbf{P}}_{\mathcal{T}_m}(\mathbf{x}_i)\|^2. \tag{12}$$

By induction over the dimension of  $\mathcal{T}$ , it readily follows that we can characterize the sum of the first  $m$  and the sum of the last  $(n - m)$  eigenvalues by

$$\sum_{j=1}^m \lambda_j(\mathbf{K}) = \max_{\dim(\mathcal{T})=m} \sum_{i=1}^n \|\mathbf{P}_{\mathcal{T}}(\mathbf{x}_i)\|^2 = \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \min_{\dim(\mathcal{T})=m} \sum_{i=1}^n \|\bar{\mathbf{P}}_{\mathcal{T}}(\mathbf{x}_i)\|^2, \tag{13}$$

$$\sum_{j=m+1}^n \lambda_j(\mathbf{K}) = \sum_{i=1}^n \|\mathbf{x}_i\|^2 - \sum_{j=1}^m \lambda_j(K) = \min_{\dim(\mathcal{T})=m} \sum_{i=1}^n \|\bar{\mathbf{P}}_{\mathcal{T}}(\mathbf{x}_i)\|^2, \quad (14)$$

respectively. Hence, when  $m = n - 1$ , it implies that the subspace spanned by the last eigenvector is characterized as that for which the residual sum of squares is minimized.

We can now generalize all of the above into a kernel-defined feature space by replacing every vector  $\mathbf{x}$  by  $\Phi(\mathbf{x})$ , where  $\Phi$  is the corresponding feature map, and  $\mathbf{K} = (K_{ij})$ , where  $K_{ij} = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ , is the Gram matrix. Because the solution of KPCA is achieved by the eigendecomposition of  $\mathbf{K}$ , the interpretation of the smallest KPC has much in common with the analysis of residuals.

Consider the kernel operator  $\mathcal{K}(f)$  and its eigenspace. The kernel operator is defined as

$$\mathcal{K}(f)(\mathbf{x}) = \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{z}) f(\mathbf{z}) p(\mathbf{z}) d\mathbf{z}, \quad (15)$$

where  $p(\mathbf{x})$  is the underlying probability density function in input space and  $\mathcal{K}(f)$  is a linear operator for the function  $f$ . Assuming the operator  $\mathcal{K}$  is nonnegative-definite, by Mercer's Theorem, we can decompose  $k(\mathbf{x}, \mathbf{z})$  as a sum of eigenfunctions,

$$k(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{z}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle, \quad (16)$$

where  $\lambda_i = \lambda_i(\mathcal{K}(f))$  are the eigenvalues of  $\mathcal{K}(f)$ . The functions

$$\boldsymbol{\psi}(\mathbf{x}) = (\psi_1(\mathbf{x}), \dots, \psi_i(\mathbf{x}), \dots) \quad (17)$$

form a complete orthonormal basis with respect to the inner-product,  $\langle f, g \rangle = \int_{\mathcal{X}} f(\mathbf{x}) g(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ , and  $\Phi(\mathbf{x})$  is the feature space mapping,

$$\Phi : \mathbf{x} \rightarrow \phi_i(\mathbf{x}) = \sqrt{\lambda_i} \psi_i(\mathbf{x}) \in \mathcal{H}, \quad i = 1, 2, \dots \quad (18)$$

Note that  $\psi_i(\mathbf{x})$  has norm 1; that is,  $\int_{\mathcal{X}} \psi_i(\mathbf{x}) \psi_j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \delta_{ij}$ , where  $\delta_{ij} = 1$  if  $i = j$ , and 0 otherwise. Also,  $\psi_i(\mathbf{x})$  satisfies  $\psi_i(\mathbf{x}) = \frac{1}{\lambda_i} \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{z}) \psi_i(\mathbf{z}) p(\mathbf{z}) d\mathbf{z}$ , so

that

$$\int_{\mathcal{X} \times \mathcal{X}} k(\mathbf{x}, \mathbf{z}) \psi_i(\mathbf{x}) \psi_i(\mathbf{z}) p(\mathbf{x}) p(\mathbf{z}) d\mathbf{x} d\mathbf{z} = \lambda_i. \quad (19)$$

For a general function  $g(\mathbf{x})$ , we define the vector  $\mathbf{g} = \int_{\mathcal{X}} g(\mathbf{x}) \boldsymbol{\psi}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$ . Then, the expected squared-norm of the projection of  $\Phi(\mathbf{x})$  onto the vector  $\mathbf{g}$  is given by

$$\begin{aligned} & \mathbb{E}[\|P_{\mathbf{g}}(\Phi(\mathbf{x}))\|^2] \\ &= \int_{\mathcal{X}} \|P_{\mathbf{g}}(\Phi(\mathbf{x}))\|^2 p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X}} (\mathbf{g}^\top \Phi(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \int_{\mathcal{X}} [g(\mathbf{y}) \boldsymbol{\psi}(\mathbf{y})^\top \Phi(\mathbf{x}) p(\mathbf{y}) d\mathbf{y}] [g(\mathbf{z}) \boldsymbol{\psi}(\mathbf{z})^\top \Phi(\mathbf{x}) p(\mathbf{z}) d\mathbf{z}] p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X} \times \mathcal{X} \times \mathcal{X}} g(\mathbf{y}) g(\mathbf{z}) \times \\ & \quad \left[ \sum_{i=1}^{\infty} \sqrt{\lambda_i} \psi_i(\mathbf{y}) \psi_i(\mathbf{x}) p(\mathbf{y}) d\mathbf{y} \right] \left[ \sum_{j=1}^{\infty} \sqrt{\lambda_j} \psi_j(\mathbf{z}) \psi_j(\mathbf{x}) p(\mathbf{z}) d\mathbf{z} \right] p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X} \times \mathcal{X}} g(\mathbf{y}) g(\mathbf{z}) \times \\ & \quad \left[ \sum_{i,j=1}^{\infty} \sqrt{\lambda_i} \psi_i(\mathbf{y}) p(\mathbf{y}) d\mathbf{y} \sqrt{\lambda_j} \psi_j(\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \right] \int_{\mathcal{X}} \psi_i(\mathbf{x}) \psi_j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathcal{X} \times \mathcal{X}} g(\mathbf{y}) g(\mathbf{z}) \left[ \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{y}) \psi_i(\mathbf{z}) \right] p(\mathbf{y}) d\mathbf{y} p(\mathbf{z}) d\mathbf{z} \\ &= \int_{\mathcal{X} \times \mathcal{X}} g(\mathbf{y}) g(\mathbf{z}) k(\mathbf{y}, \mathbf{z}) p(\mathbf{y}) d\mathbf{y} p(\mathbf{z}) d\mathbf{z}. \end{aligned} \quad (20)$$

The sum of the finite-case characterization of eigenvalues and eigenvectors (19) can, therefore, be replaced by the expectation,

$$\lambda_m(\mathcal{K}(f)) = \max_{\dim(\mathcal{T})=m} \min_{0 \neq \alpha \in \mathcal{T}} \mathbb{E}[\|P_{\alpha}(\Phi(\mathbf{x}))\|^2]. \quad (21)$$

Similarly, the sum of the first  $m$  and sum of the last  $(n - m)$  eigenvalues can be expressed as

$$\sum_{j=1}^m \lambda_j(\mathcal{K}(f)) = \max_{\dim(\mathcal{T})=m} \mathbb{E}[\|P_{\mathcal{T}}(\Phi(\mathbf{x}))\|^2]$$

$$= \mathbb{E}[\|\Phi(\mathbf{x})\|^2] - \min_{\dim(\mathcal{T})=m} \mathbb{E}[\|\bar{P}_{\mathcal{T}}(\Phi(\mathbf{x}))\|^2], \quad (22)$$

$$\begin{aligned} \sum_{j=m+1}^n \lambda_j(\mathcal{K}(f)) &= \mathbb{E}[\|\Phi(\mathbf{x})\|^2] - \sum_{j=1}^m \lambda_j(\mathcal{K}(f)) \\ &= \min_{\dim(\mathcal{T})=m} \mathbb{E}[\|\bar{P}_{\mathcal{T}}(\Phi(\mathbf{x}))\|^2], \end{aligned} \quad (23)$$

where  $P_{\mathcal{T}}(\Phi(\mathbf{x}))$  is the projection of  $\Phi(\mathbf{x})$  onto the subspace  $\mathcal{T}$ , and  $\bar{P}_{\mathcal{T}}(\Phi(\mathbf{x}_i))$  is the projection of  $\Phi(\mathbf{x})$  onto the space orthogonal to  $\mathcal{T}$ . The above results again imply that, in the feature space induced by the map  $\Phi$ , the smallest KPC can be interpreted in a similar way as that of residuals.

### 3.3 Threshold between Small and Large Kernel Principal Components

The eigenvalues are bounded below by 0 because  $\mathbf{K}$  is nonnegative-definite. The existence of an eigenvalue equal to 0 indicates degeneracy or nonlinear dependency. In practice, exact 0 rarely happens due to computational accuracy, but near-zero or duplicates of near-zero eigenvalues usually exist, especially when  $n$  is large. Therefore, we have to determine the magnitude of the smallest eigenvalues so that they do contain important information about the noise. Meanwhile, because our attention is focused on the smallest KPCs, we need a clear separation between “large” and “small” KPCs.

Intuitively, we could do this by separating the eigenvalues corresponding to the “large” and “small” KPCs. However, for a given dataset, applying different kernels leads to different kernel matrices that need to be eigendecomposed and, hence, generate different eigenstructures. For example, when applying a polynomial kernel,  $k(\mathbf{x}, \mathbf{z}) = (\langle \mathbf{x}, \mathbf{z} \rangle + c)^d$ , the magnitudes of the eigenvalues increase along with the increase of the power  $d$ . Hence, a universal threshold (with respect to all kernel functions) to separate “large” from “small” KPCs is impossible. Therefore, we have

to enforce some restrictions on the kernel functions being chosen.

We consider only kernel functions that represent probability measures or empirical distributions, for then the eigenvalues derived from those kernels are bounded above. In particular, we use a Gaussian radial basis function (RBF) kernel,

$$K_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right), \quad \sigma > 0, \quad (24)$$

where the matrix  $\mathbf{K} = (K_{ij})$  induced by the RBF kernel (24) has full rank. Define a vector space by taking a linear combination of the form,

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, \mathbf{x}_i). \quad (25)$$

Next, define an inner-product between  $f$  and another function  $g$  as

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(\mathbf{x}_i, \mathbf{y}_j), \quad (26)$$

where

$$g(\cdot) = \sum_{j=1}^m \beta_j k(\cdot, \mathbf{y}_j). \quad (27)$$

Here,  $n$  and  $m$  are integers,  $\alpha_i, \beta_j \in \mathcal{R}$ , and  $\mathbf{x}_i, \mathbf{y}_j \in \mathcal{X}$ ,  $i = 1, 2, \dots, n$ ,  $j = 1, 2, \dots, m$ . Expression (26) explicitly contains the expansion coefficients, which need not be unique. To see that the inner-product is nevertheless well-defined, note that  $\langle f, g \rangle = \sum_{j=1}^m \beta_j f(\mathbf{y}_j)$ , using  $k(\mathbf{y}_j, \mathbf{x}_i) = k(\mathbf{x}_i, \mathbf{y}_j)$ . The sum in (26), however, does not depend on the particular expansion of  $f$ . Similarly, for  $g$ ,  $\langle f, g \rangle = \sum_{i=1}^n \alpha_i g(\mathbf{x}_i)$ . Thus,  $\langle \cdot, \cdot \rangle$  is symmetric and nonnegative-definite,  $\langle f, g \rangle = \langle g, f \rangle$ ,  $\langle f, f \rangle = \sum_{i,j=1}^n \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ , and, hence, is bilinear. Thus, the space  $\mathcal{H}$  is a vector space and is endowed with an inner-product. We turn the space  $\mathcal{H}$  into a Hilbert space by completing it in the norm corresponding to the inner-product,  $\|f\| = \sqrt{\langle f, f \rangle}$ , where ‘‘completeness’’ means that every Cauchy sequence of elements in the space converges to an element in the space, in the sense that the norm of differences approaches zero. The constraint of the smallest KPC optimization problem

can now be cast as a restriction to the unit ball in  $\mathcal{H}$ :  $\text{var}\{\Phi(\mathbf{x})\} = \|\Phi(\mathbf{x})\|^2 = 1$ . It follows from an elementary theorem in  $L_2$  theory that there exists a bounded, symmetric, linear operator  $\mathbf{P}$  on  $\mathcal{H}$  such that the optimization problem for the smallest KPC can be written as:

$$\min_{\mathcal{H}} \langle \Phi(\mathbf{x}), \mathbf{P}\Phi(\mathbf{x}) \rangle, \text{ subject to } \|\Phi(\mathbf{x})\|^2 = 1. \quad (28)$$

We need to identify the operator  $\mathbf{P}$  that satisfies these conditions.

*Lemma 1.* Define the operator  $\mathbf{P} : \mathcal{H} \rightarrow \mathcal{H}$  by the component mappings

$$(\mathbf{P}\Phi(\mathbf{x}))_i = \mathbf{P}_i\Phi(\mathbf{x}_j), \quad i, j = 1, 2, \dots, n, \quad (29)$$

where  $\mathbf{P}_i$  is the orthogonal projection onto the subspace  $\mathcal{H}_i$  of the feature space  $\mathcal{H}$ . Then,  $\mathbf{P}$  is symmetric, nonnegative-definite, and bounded above by  $n$ .

*Proof.* The operator  $\mathbf{P}$  is nonnegative:

$$\begin{aligned} \langle \Phi(\mathbf{x}), \mathbf{P}\Phi(\mathbf{x}) \rangle &= \sum_{i,j} \langle \Phi(\mathbf{x}_i), \mathbf{P}_i\Phi(\mathbf{x}_j) \rangle \\ &= \sum_{i,j} \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \\ &= \left\langle \sum_i \Phi(\mathbf{x}_i), \sum_j \Phi(\mathbf{x}_j) \right\rangle \\ &= \left\| \sum_i \Phi(\mathbf{x}_i) \right\|^2 \\ &\geq 0, \end{aligned}$$

and  $\mathbf{P}$  is bounded above by  $n$ :

$$\|\mathbf{P}\Phi(\mathbf{x})\|^2 = \sum_i \|\mathbf{P}_i\Phi(\mathbf{x}_j)\|^2 \leq \sum_i \|\Phi(\mathbf{x}_j)\|^2 = n\|\Phi(\mathbf{x}_j)\|^2 = n.$$

The eigencharacterization of the smallest KPC now follows from standard results about symmetric operators. Because we are working in a possibly-infinite Hilbert

space, the existence of the smallest eigenvalue of  $\mathbf{P}$  is not guaranteed. Therefore, we use the phrase “if it exists” whenever necessary.

*Proposition 1.* The eigenvector corresponding to the smallest eigenvalue of the operator  $\mathbf{P}$ , if it exists, is the vector for the smallest KPC.

*Proposition 2.* The eigenvector corresponding to the  $m$ th smallest eigenvalue of the operator  $\mathbf{P}$ , if it exists, is the vector for the  $m$ th smallest KPC.

*Corollary 1.* The variance of the  $m$ th smallest KPC is  $\lambda_{(m)}$ .

This result follows because

$$\text{var}(\phi^{(m)}) = \langle \Phi(\mathbf{x}), \mathbf{P}_{(m)}\Phi(\mathbf{x}) \rangle = \langle \Phi(\mathbf{x}), \lambda_{(m)}\Phi(\mathbf{x}) \rangle = \lambda_{(m)}\|\Phi(\mathbf{x})\|^2 = \lambda_{(m)}. \quad (30)$$

Note that the dimension of  $\mathcal{H}$  could be infinite in feature space. Although the spectrum of  $\mathbf{P}$  is bounded above, the existence of eigenvalues is complicated in that  $\mathbf{P}$  may have spectral values that are not eigenvalues (Gohberg, Goldberg, and Kaashoek, 2003). In other words, the existence of the smallest eigenvalue of  $\mathbf{P}$  is not granted à priori. These undesirable possibilities can be ruled out by adopting the usual compactness assumption. An operator  $\mathbf{K}$  is said to be *compact* if it maps the closed unit ball onto a relatively compact set. This implies that the image of the unit ball, or any norm-bounded set, is a relatively compact set in the norm topology. It is known that the compact operator  $\mathbf{K}$  on an infinite-dimensional Banach space (i.e., a complete normed vector space) has a spectrum that is either a finite subset of  $\mathcal{C}$  which includes 0, or a countably infinite subset of  $\mathcal{C}$  which has 0 as its only limit point. Moreover, in either case, the nonzero elements of the spectrum are eigenvalues of  $\mathbf{K}$  with finite multiplicities (Jorgens, 1982).

We now assume that the restricted projections  $\mathbf{P}_{i|\mathcal{H}_k} = \mathbf{P}_{i|k} : \mathcal{H}_k \rightarrow \mathcal{H}_i$  are compact operators. Under this assumption,  $\mathbf{P}$  is not compact; but we know from

Donnell, Buja and Stuetzle (1994) that the operator  $\mathbf{P} - \mathbf{I} : \mathcal{H} \rightarrow \mathcal{H}$  is compact, where  $\mathbf{I}$  represents the identity operator. If this compactness assumption holds, then the characteristics of the spectrum of  $\mathbf{P} - \mathbf{I}$  are similar to those of a finite-dimensional symmetric matrix, except for the limiting behavior that is vacuous in finite dimensions:

1. There exists a sequence of eigenvalues  $\{l_k, k = 1, 2, \dots\}$  of  $\mathbf{P} - \mathbf{I}$  for which  $|l_1| \geq |l_2| \geq \dots \geq |l_k| \geq \dots$ .
2. The limit of the sequence is 0, that is:  $\lim_{k \rightarrow \infty} l_k = 0$ .
3. The eigenspaces for distinct eigenvalues are orthogonal.
4. The nonzero eigenvalues have finite multiplicity.

The spectrum of  $\mathbf{P} - \mathbf{I}$  is thus a countable, bounded set with 0 as the only possible accumulation point. The eigenvalues,  $\{l_k\}$ , of  $\mathbf{P} - \mathbf{I}$  are related to the eigenvalues,  $\{\lambda_k\}$ , of  $\mathbf{P}$  through  $\lambda_k = l_k + 1$ . Hence, the eigenvalues and eigenspaces of  $\mathbf{P}$  inherit all the above properties, with  $l_k$  replaced by  $\lambda_k - 1$ . In particular, we have the following result.

*Corollary 2.* The only accumulation point of the eigenvalues of  $\mathbf{P}$  is +1.

This explains why +1 is the natural threshold between “large” and “small” KPCs.

### 3.4 Location of the Smallest KPC

KPCA can find up to  $n$  KPCs. However, it is not true that the  $n$ th KPC is the “smallest” KPC, except that the dimension of the feature space is exactly equal to  $n$ . In other words, the  $n$  eigenvalues are all nonzero and distinct. If the feature space has dimension smaller than  $n$ , then the smallest KPC corresponds to the last

non-zero eigenvalue (say the  $l$ th) and the remaining eigenvalues should each equal 0. If the dimension of the feature space is greater than  $n$ , then all  $n$  eigenvalues should be nonzero and distinct, but the  $n$ th eigenvalue is not the true smallest eigenvalue.

In practice, when  $n$  is large, near-zero eigenvalues are likely to occur. Because we usually do not know the dimension of the feature space, we need to determine a magnitude for the smallest eigenvalue. Consequently, the corresponding smallest KPC preserves enough information about the noise of the distribution in order to detect any outliers. This task is made more difficult by the fact that no distribution is being assumed in feature space. Hence, the distribution of the eigenvalues is hard to obtain. The choice of kernel function is subjective as are the values of the parameters in the selected kernel function; so, the solution should be interpreted carefully.

Because  $+1$  is the threshold value that separates “large” from “small” eigenvalues, we know that the smallest eigenvalue is less than 1, but is not small enough to be trivial. We now propose a nonparametric method that is simple, but practical, to locate the position of the smallest KPC. Start with all  $n$  eigenvalues in decreasing order of magnitude and locate the pair of successive eigenvalues (say  $\lambda_k$  and  $\lambda_{k+1}$ ) that surround the value 1. Then, construct a new sequence using the last  $n - k$  eigenvalues,  $\lambda_{k+1}, \lambda_{k+2}, \dots, \lambda_n$ , as follows:

$$S = \left\{ \frac{\lambda_{k+1}}{\sum_{i=1}^n \lambda_i}, \frac{\lambda_{k+2}}{\sum_{i=1}^n \lambda_i}, \dots, \frac{\lambda_n}{\sum_{i=1}^n \lambda_i} \right\}, \quad (31)$$

where each element represents the proportion of variance explained by the  $i$ th KPC ( $i = k + 1, k + 2, \dots, n$ ). We only keep those elements that are greater than 0.01%; that is, we ignore the small KPCs that explain less than 0.01% of the total variance, so that the new sequence is further truncated at  $n'$  ( $\leq n$ ),

$$S' = \left\{ \frac{\lambda_{k+1}}{\sum_{i=1}^n \lambda_i}, \frac{\lambda_{k+2}}{\sum_{i=1}^n \lambda_i}, \dots, \frac{\lambda_{n'}}{\sum_{i=1}^n \lambda_i} \right\}. \quad (32)$$

Because the mean of the eigenvalues  $\lambda_{k+1}, \lambda_{k+2}, \dots, \lambda_{n'}$ ,

$$\bar{\lambda}^* = \frac{\sum_{i=k+1}^{n'} \lambda_i}{n' - k}, \quad (33)$$

approximates the average of the “small” eigenvalues, we can compare each element in the truncated sequence  $S'$  with the threshold

$$C = \frac{\bar{\lambda}^*}{\sum_{i=1}^n \lambda_i}, \quad (34)$$

which approximates the average proportion of variance explained by the small KPCs. Find the element of  $S'$  that is just above the threshold. Then, the numerator of that element is the eigenvalue corresponding to the selected smallest KPC.

After locating the smallest KPC, we draw the scatterplot of the smallest KPC against the second smallest KPC. From Section 3.2, we know that the interpretation of the smallest KPCs can be viewed as the analysis of residuals. Therefore, the last two smallest KPCs should have a strong relationship. Hence, in the scatterplot, any points that deviate significantly from the majority of the data are identified as outliers.

## 4. NUMERICAL EXAMPLES

### 4.1 Simulations

To illustrate how the proposed method works in practice, we set up several small simulation studies. These simulation studies only consider the problem of identifying outliers in a univariate outlier situation. The details and S-PLUS code can be found in Shen (2007). The RBF kernel (24) is applied with fixed  $\sigma = 8$ .

We first generate a dataset containing 100 observations from a multivariate standard normal distribution in five dimensions. The standard deviation of the outliers in one randomly chosen dimension is forced to be 10 times of that of the regular

Table 1: Summary of simulation studies with one outlier generated.

KPCA	MD		RMD	
	0	1	0	1
0	29	0	29	0
1	1	20	0	21

Table 2: Summary of simulation studies with three outliers generated.

KPCA	MD			RMD				
	0	1	2	3	0	1	2	3
0	7	0	0	0	6	1	3	0
1	1	14	0	0	0	12	0	0
2	4	17	1	0	0	3	19	0
3	5	1	0	0	0	0	2	4

points in order to ensure the inclusion of a given number of outliers in the data. The outliers are generated so that they occupy the last few positions of the dataset; this makes the outliers easy to label without confusing them with the regular points. For example, if three outliers are generated, their positions in the dataset are the 98th, 99th and 100th observations, respectively. We calculate the ordinary and robust Mahalanobis distance of the data in the original space as references, and the ordinary Mahalanobis distance of the last two smallest KPCs in feature space. This gives us three sets of distance measurements. For each set of distance measurements, “outliers” are identified if the observations are below  $F_L - 3(F_U - F_L)$  or above  $F_U + 3(F_U - F_L)$ , where  $F_L$  and  $F_U$  are the lower- and upper-fourths, respectively, of the data (see Tukey, 1977). Three scenarios are considered in which different numbers of outliers (1, 3 and 5, respectively) are generated. In each scenario, the experiment is run 50 times. The results are given in Tables 1, 2, and 3, corresponding to the numbers of outliers generated.

Table 3: Summary of Simulation Studies with five outliers generated.

KPCA	MD					RMD						
	0	1	2	3	4	5	0	1	2	3	4	5
0	4	0	0	0	0	0	2	0	0	1	1	0
1	5	4	0	0	0	0	0	5	4	0	0	0
2	8	13	3	0	0	0	0	2	21	1	0	0
3	5	4	0	0	0	0	0	2	0	5	2	0
4	1	2	0	0	0	0	0	1	1	0	1	0
5	1	0	0	0	0	0	0	0	0	0	1	0

In each of the three Tables, the column titles represent the three sets of distance measurement: ordinary Mahalanobis distance (MD) of the data in the original space, robust Mahalanobis distance (RMD) of the data in the original space, and the proposed method using ordinary Mahalanobis distance of the two smallest KPC in feature space (KPCA). The entries are the counts of combinations of the number of outliers detected by the three methods. For example, consider Table 2, in which three outliers were generated. The number 4 in the last row, last column of the table indicates that, out of the 50 runs, there are four times that both the proposed method (KPCA) and the robust Mahalanobis distance in the original space (RMD) could detect all 3 outliers. If the two methods do about as well, then most of the entries should be on the diagonal. If the proposed method is superior, then most of the entries should be on the lower triangle. The results suggest that, when there is only a single outlier (Table 1), no distance measurement can claim superiority over the others. But when there are multiple outliers (Tables 2 and 3), the ordinary Mahalanobis distance in the original space (MD) clearly shows its inability to compete with the other two. This is not surprising because it is well known that ordinary Mahalanobis distance is not sensitive to multiple outliers. There are mixed results when comparing the proposed method (KPCA) with the robust Mahalanobis dis-

tance in the original space (RMD) where each method wins over the other in some cases.

Figures 1 and 2 show scatterplots of the last two smallest KPCs in examples where the robust Mahalanobis distance in the original space detects more of the outliers than the proposed method. For example, the lower-right panel of Figure 1 is a scatterplot from run 44 where three outliers were generated. By applying the methods described above, all three outliers are identified by using the robust Mahalanobis distance in the original space (RMD), but only two out of the three outliers are identified by using the ordinary Mahalanobis distance calculated on the last two smallest KPCs (KPCA). However, the plot clearly suggests that all three outliers stand out from the majority of the data points. If we calculate the robust Mahalanobis distance on the last two smallest KPCs, then the remaining outlier can be identified as well. This is true for all instances where the robust Mahalanobis distance wins over the proposed method. From these simulations, we are confident that the proposed method performs as well as the robust Mahalanobis distance in the original space, which is the best method currently available for detecting univariate outliers. Although the simulation model assumes the outliers have larger variance than the regular points, all of them share the same mean from a multivariate normal distribution. Therefore, the conclusion made is only applicable to this specific model. Further research will explore models where difference in location or difference in distributions will be considered.

## 4.2 Hawkins–Bradu–Kass Data

This example uses a dataset specially constructed by Hawkins, Bradu, and Kass (1984), which consists of 75 points in 3 dimensions plus a response. In our application, we only study the three predictor variables. We refer to this dataset as HBK. The first 14 points were designed to be outliers. Hawkins et al. show that by using

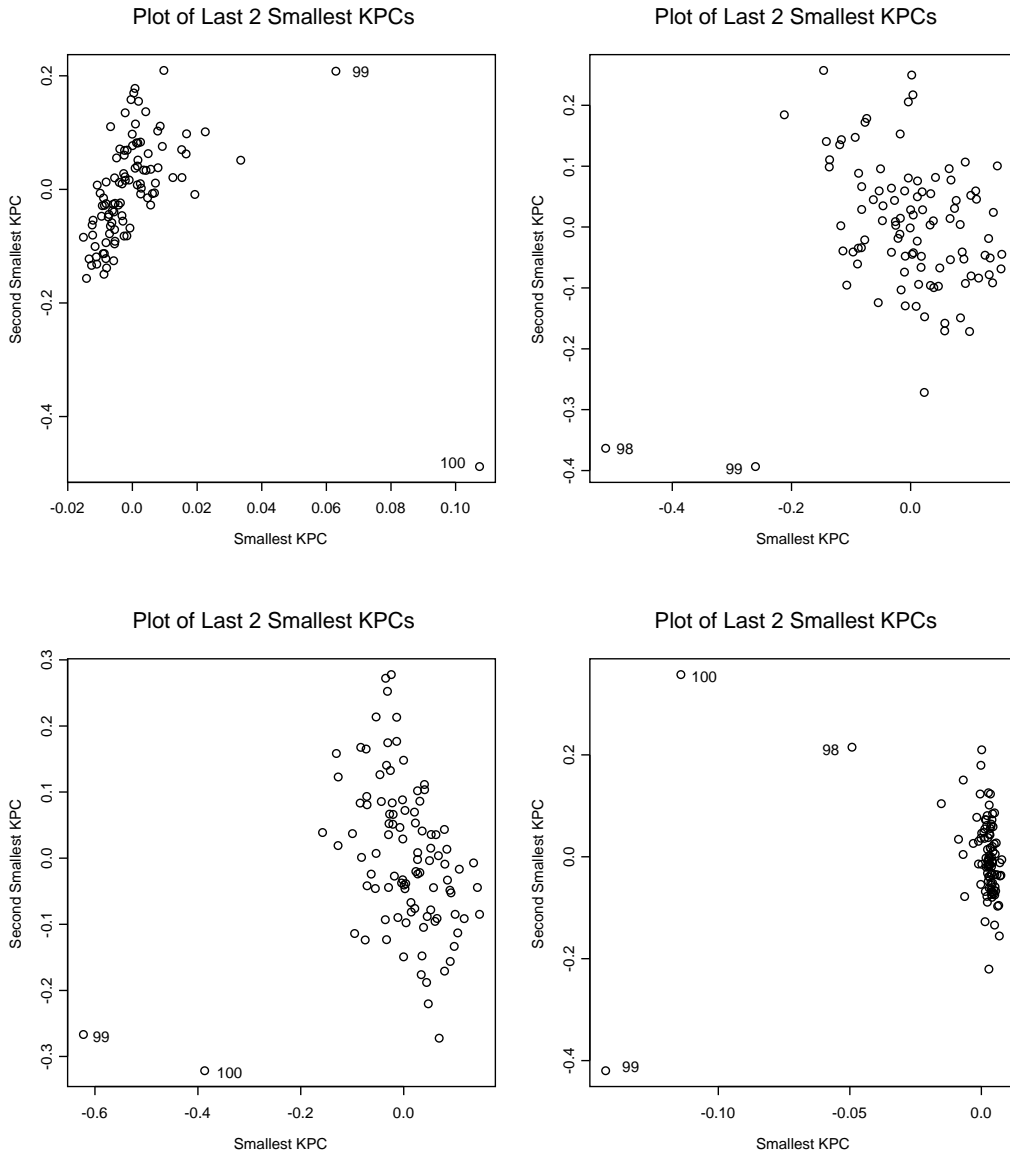


Figure 1: Simulation study for the detection of three univariate outliers from five-dimensional data. Selected scatterplots of the last two smallest KPCs from 50 different runs. Upper-left panel: run 8; upper-right panel: run 15; lower-left panel: run 19; lower-right panel: run 44.

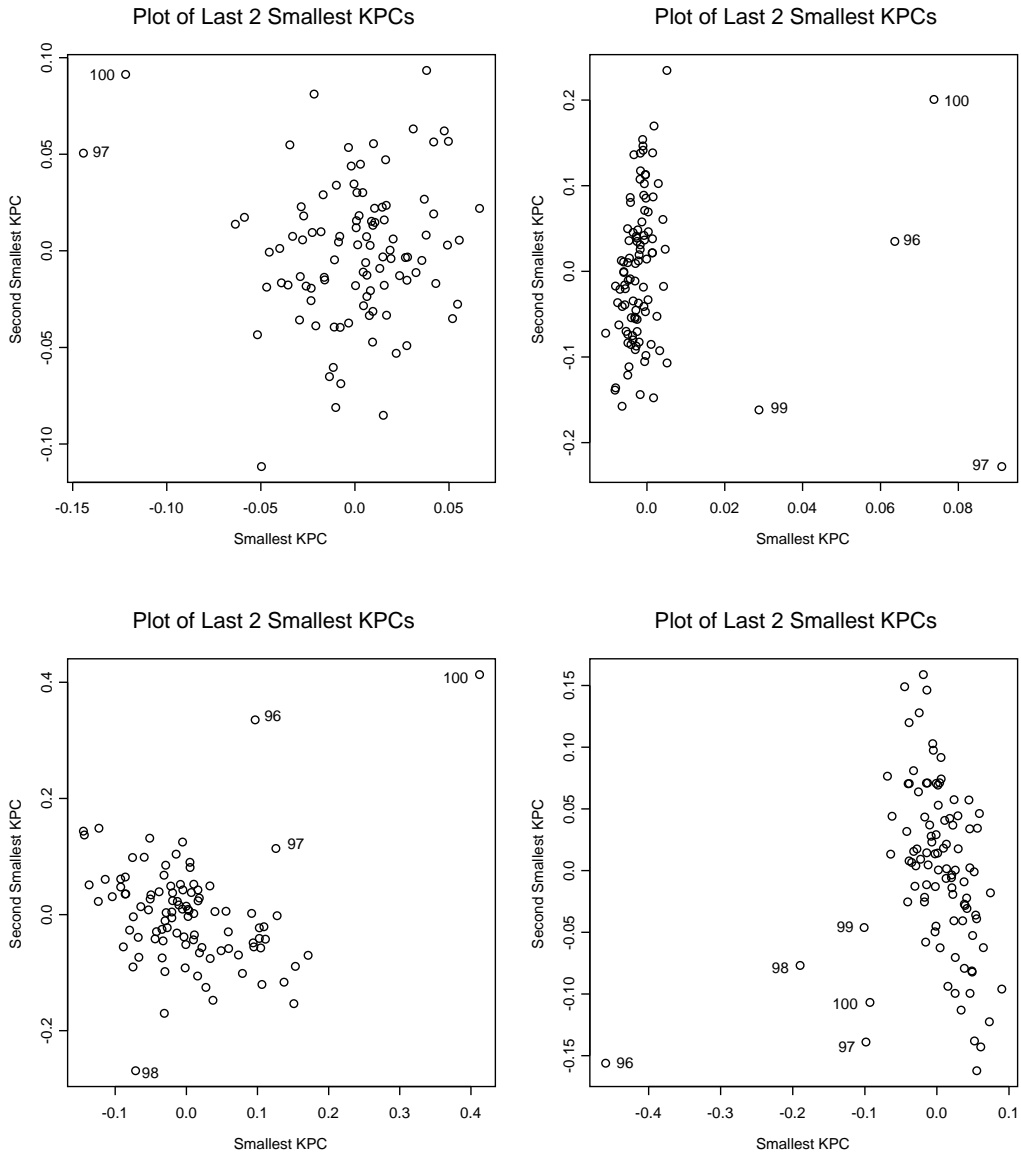


Figure 2: Simulation study for the detection of five univariate outliers from five-dimensional data. Selected scatterplots of the last two smallest KPCs from 50 different runs. Upper-left panel: run 8; upper-right panel: run 15; lower-left panel: run 22; lower-right panel: run 29.

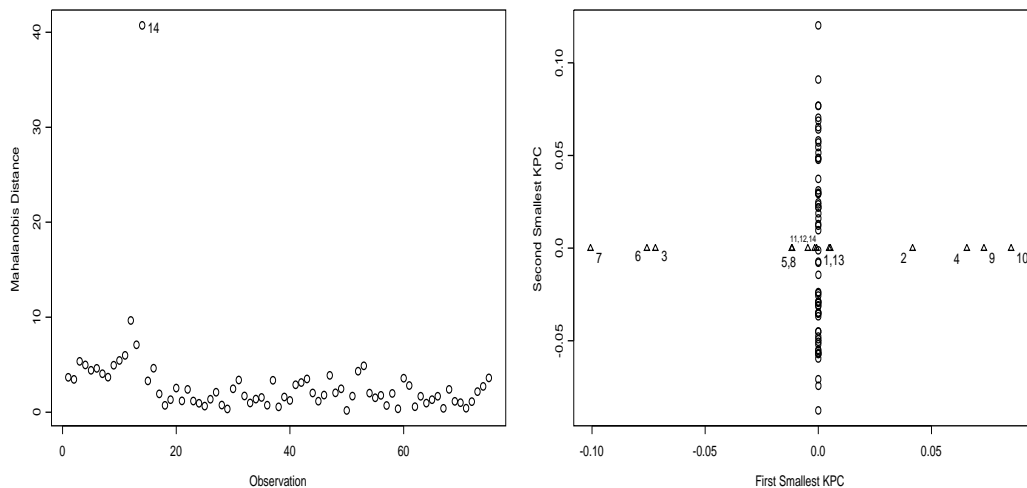


Figure 3: HBK Data. Left panel: Scatterplot of Mahalanobis distance suggests only one outlier could be detected; Right panel: Scatterplot of the last two smallest KPCs suggests all outliers could be detected

an unweighted median and a weighted median, they can separate the 14 outliers into 10 high-leverage outliers and four high-leverage inliers. Rocke and Woodruff (1996) also identify these 14 particular outliers. If we use ordinary Mahalanobis distance, only point 14 is detected, as seen in the top panel of Figure 3. In this application, the other 13 points are said to be “masked” by the 14th point (Hawkins, 2006).

Performing KPCA using RBF kernel with  $\sigma = 1$ , we find six “large” eigenvalues and 69 “small” eigenvalues. The threshold  $C$  is estimated to be 0.0026, so that we ignore all the KPCs that explain less than 0.26% of the total variance. Therefore, the 15th KPC is chosen as the smallest KPC. When we plot the smallest KPC against the second smallest KPC, as shown in the bottom panel of Figure 3, we see that the first 14 points are perpendicular to the remaining points. The KPCA scores of the smallest KPC, given in Table 4, show that the magnitudes of the first 14 points are, on average, 1000 times that of the remaining points. Therefore, we

Table 4: HBK data. Scores of the smallest KPCs.

(1)	$5 \times 10^{-3}$	$4 \times 10^{-2}$	$-7 \times 10^{-2}$	$7 \times 10^{-2}$	$-1 \times 10^{-2}$	$-8 \times 10^{-2}$	$-1 \times 10^{-1}$
(8)	$-1 \times 10^{-2}$	$7 \times 10^{-2}$	$9 \times 10^{-2}$	$-1 \times 10^{-3}$	$-5 \times 10^{-3}$	$5 \times 10^{-3}$	$-2 \times 10^{-3}$
(15)	$3 \times 10^{-6}$	$8 \times 10^{-6}$	$1 \times 10^{-5}$	$-5 \times 10^{-6}$	$-5 \times 10^{-6}$	$1 \times 10^{-5}$	$6 \times 10^{-6}$
(22)	$-5 \times 10^{-6}$	$-3 \times 10^{-6}$	$-5 \times 10^{-6}$	$7 \times 10^{-6}$	$-4 \times 10^{-6}$	$-3 \times 10^{-6}$	$-6 \times 10^{-6}$
(29)	$-2 \times 10^{-6}$	$-6 \times 10^{-6}$	$1 \times 10^{-6}$	$6 \times 10^{-6}$	$-5 \times 10^{-6}$	$2 \times 10^{-5}$	$8 \times 10^{-6}$
(36)	$-3 \times 10^{-6}$	$2 \times 10^{-6}$	$-3 \times 10^{-6}$	$-3 \times 10^{-7}$	$5 \times 10^{-6}$	$3 \times 10^{-6}$	$7 \times 10^{-6}$
(43)	$-3 \times 10^{-6}$	$-6 \times 10^{-6}$	$1 \times 10^{-6}$	$-7 \times 10^{-6}$	$-3 \times 10^{-7}$	$2 \times 10^{-6}$	$2 \times 10^{-5}$
(50)	$-8 \times 10^{-6}$	$-3 \times 10^{-7}$	$-1 \times 10^{-7}$	$-5 \times 10^{-6}$	$-5 \times 10^{-6}$	$5 \times 10^{-7}$	$-5 \times 10^{-6}$
(57)	$-3 \times 10^{-6}$	$-4 \times 10^{-6}$	$-9 \times 10^{-6}$	$5 \times 10^{-6}$	$-5 \times 10^{-6}$	$1 \times 10^{-6}$	$-2 \times 10^{-7}$
(64)	$-5 \times 10^{-6}$	$-6 \times 10^{-6}$	$1 \times 10^{-5}$	$-2 \times 10^{-6}$	$5 \times 10^{-6}$	$-1 \times 10^{-6}$	$-9 \times 10^{-6}$
(71)	$-5 \times 10^{-6}$	$4 \times 10^{-7}$	$-3 \times 10^{-6}$	$1 \times 10^{-5}$	$4 \times 10^{-7}$		

conclude that the these points are outliers.

### 4.3 Real-Data Examples

This section illustrates the methodology of this paper — by which we use the smallest KPCs method for finding outliers in feature space — to two well-known datasets. Only the RBF kernel is applied with different values of the scale parameter  $\sigma$  in order to reach the best solution.

#### 4.3.1 Bushfire Data

These data were used by Campbell (1989) to locate bushfire scars. The data consist of satellite measurements at  $p = 5$  frequency bands, corresponding to each of  $n = 38$  pixels. Maronna and Yohai (1995) analyzed the data and concluded that there were two groups of outliers: points 7–11 and points 32–38. The data were also studied by Rocke and Woodruff (1996), who found points 8, 9, 32–38 to be extreme outliers and points 7, 10, 11, and 31 to be less extreme outliers. Performing KPCA using a RBF kernel with  $\sigma = 8$ , we found two “large” eigenvalues and 36 “small” eigenvalues.

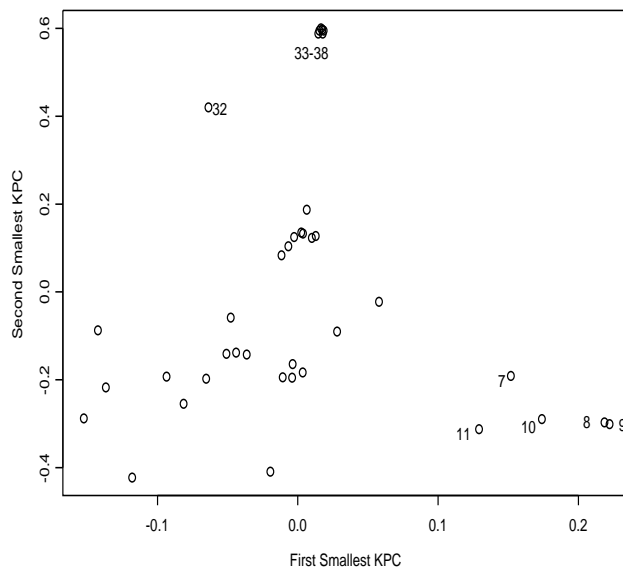


Figure 4: The Bushfire data. Scatterplot of the smallest KPC vs. the second smallest KPC show that there are two groups of outliers, points 32–38 (top) and 7–11 (lower right).

Figure 4 displays a scatterplot of the smallest and the second smallest KPCs, where we confirm the division into two groups of outliers, namely, 32–38 (top of Figure 4) and 7–11 (lower right in Figure 4).

### 4.3.2 Education Expenditure Data

These data were used by Chatterjee, Hadi, and Price (2000) as an example of heteroscedasticity. The data give the education expenditures for the 50 U.S. States as projected in 1975. The data were also studied by Rousseeuw and Leroy (1987, pp. 109–112). There are three explanatory variables,  $X_1$  (number of residents per thousand residing in urban areas in 1970),  $X_2$  (per capita personal income in 1973),  $X_3$  (number of residents per thousand under 18 years of age in 1974), and one re-

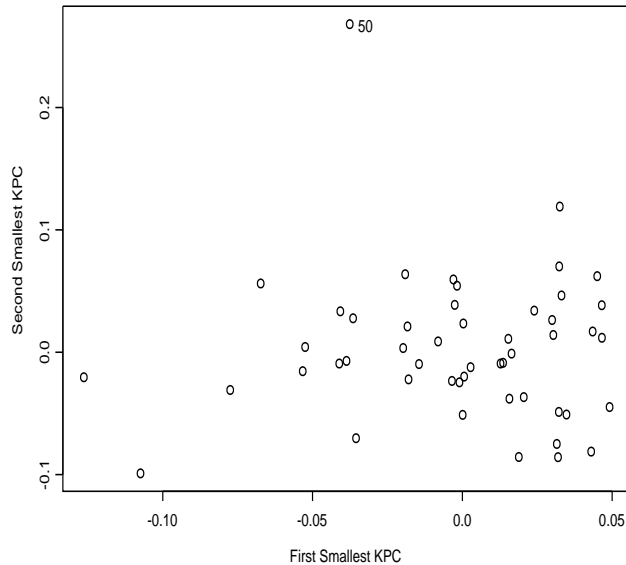


Figure 5: The Education Expenditure data. Scatterplot of smallest KPC vs. the second smallest KPC shows that the fiftieth data point is an outlier.

sponse variable  $Y$  (per capita expenditure on public education in 1975). Chatterjee and Price analyzed these data by using weighted least-squares regression. They considered the fiftieth case (Hawaii) as an outlier and decided to omit it. Rousseeuw and Leroy also identified Hawaii as an outlier by analyzing the residual plot from a least median-of-squares regression. Performing KPCA using a RBF kernel with  $\sigma = 4$ , we found four “large” eigenvalues and 46 “small” eigenvalues. A scatterplot of the smallest and the second smallest KPC is displayed in Figure 5; we see that the fiftieth case is clearly identified as an outlier.

## 5. CONCLUDING REMARKS

In this article, we investigate a new method for outlier detection using the smallest kernel principal components (KPCs). We show that the eigenvectors correspond-

ing to the smallest KPCs can be viewed as those for which the residual sum of squares is minimized, so that we could use those components to detect outliers with simple graphical techniques. The threshold between “large” and “small” eigenvalues is determined, and a method to determine the smallest KPC is suggested. Simulation studies show that in the univariate outlier situation, the proposed method performs as well as the best method available. The given examples show that this method is at least as useful as other methods, and sometimes is better. Possible directions for future research include fine-tuning the method we propose here, especially the way the smallest KPC is determined.

## References

- [1] Campbell, N.A. (1989), “Bushfire Mapping Using NOAA AVHRR Data”, *Technical Report*, CSIRO.
- [2] Chatterjee, S., Hadi, A.S., and Price, B. (2000), *Regression Analysis by Example, Third Edition*, New York: John Wiley.
- [3] Donnell, D.J., Buja, A. and Stuetzle, W. (1994) “Analysis of Additive Dependencies and Concurvities Using Smallest Additive Principal Components”, *The Annals of Statistics*, 22, 1635–1673.
- [4] Gnanadesikan, R. (1977), *Methods for Statistical Analysis of Multivariate Observations*, New York: John Wiley.
- [5] Gnanadesikan, R. and Kettenring, J.R. (1972), “Robust Estimates, Residuals, and Outlier Detection With Multivariate Data,” *Biometrics*, 28, 81–124.
- [6] Gnanadesikan, R. and Wilk, M.B. (1966), “Data Analytic Methods in Multivariate Statistical Analysis.” *General Methodology Lecture on Multivariate*

- Analysis*, 126th Annual Meeting, American Statistical Association, Los Angeles.
- [7] Gnanadesikan, R. and Wilk, M.B. (1969), “Data Analytic Methods in Multivariate Statistical Analysis,” In: *Multivariate Analysis II* (P.R. Krishnaiah, ed.), New York: Academic Press, pp. 593–638.
  - [8] Gohberg, I., Goldberg, S. and Kaashoek, M.A. (2003), “Basic Classes of Linear Operators,” *Birkhäuser*.
  - [9] Hawkins, D.M. (2006), “Masking and Swamping,” *Encyclopedia of Statistical Sciences*, New York: Wiley.
  - [10] Hawkins, D.W., Bradu, D., and Kass, G.V. (1984), “Location of Several Outliers in Multiple Regression Data Using Elemental Sets,” *Technometrics*, 26, 197–208.
  - [11] Hotelling, H. (1933), “Analysis of a Complex of Statistical Variables with Principal Components,” *Journal of Educational Psychology*, 24, 498–520.
  - [12] Izenman, A.J. (2008), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, New York: Springer.
  - [13] Jorgens, K. (1982), “Linear Integral Operators,” *Pitman Advanced Publ. Program, Boston–London–Melbourne*.
  - [14] Maronna, R.A. and Yohai, V.J. (1995), “The Behavior of the Stahel–Donoho Robust Multivariate Estimator,” *Journal of the American Statistical Association*, 90, 330–341.
  - [15] McDiarmid, C. (1989), “On the Method of Bounded Differences,” *Surveys in Combinatorics*, Cambridge University Press, 148–188.

- [16] Mercer, J. (1909), “Functions of Positive and Negative Type and Their Connection with the Theory of Integral Equations,” *Philosophical Transactions of the Royal Society, London*.
- [17] Rocke, D.M. and Woodruff, D.L. (1996), “Identification of Outliers in Multivariate Data,” *Journal of the American Statistical Association*, **91**, 1047–1061.
- [18] Rousseeuw, P.J. and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.
- [19] Schölkopf, B. and Smola, A.J. (2002), *Learning with Kernels*, Cambridge, MA: MIT Press.
- [20] Schölkopf, B., Smola, A.J., and Müller, K.-R. (1998), “Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” *Neural Computation*, **10**, 1299–1319.
- [21] Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K., Rätsch, G., and Smola, A. (1999), “Input Space Versus Feature Space in Kernel-Based Methods,” *IEEE Transactions on Neural Networks*, **10**(5), 1000–1017.
- [22] Shawe-Taylor, J. and Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge, U.K.: Cambridge University Press.
- [23] Shen, Y. (2007), *Outlier Detection Using the Smallest Kernel Principal Components*, Ph.D. dissertation, Department of Statistics, Temple University.
- [24] Tukey, J.W. (1977), *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.