Communications for Statistical Applications and Methods
2013, Vol. 20, No. 3, 175–184

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.3.175

Arrow Diagrams for Kernel Principal Component Analysis

Myung-Hoe Huh[1,a]

[a]Department of Statistics, Korea University

Abstract

Kernel principal component analysis(PCA) maps observations in nonlinear feature space to a reduced dimensional plane of principal components. We do not need to specify the feature space explicitly because the procedure uses the kernel trick. In this paper, we propose a graphical scheme to represent variables in the kernel principal component analysis. In addition, we propose an index for individual variables to measure the importance in the principal component plane.

Keywords: Principal component analysis, kernel method, radial basis function, biplot, arrow diagram.

1. Background and Aim

Consider a multivariate dataset of n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, each of which carries p numerical variables. In classical principal component analysis, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are projected onto a small number of specific ortho-normal directional vectors in $\mathbb{R}^p$, so that the projections of n observations are plotted in a reduced dimensional plane. The biplot initiated by Gabriel (1971) embeds p arrows in the graph to indicate the "flying" directions of respective variables.

In this study, we locate n observations at $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ in the q-dimensional space, where $\Phi(\cdot)$ is a nonlinear mapping from $\mathbb{R}^p$ to $\mathbb{R}^q$. To avoid the unnecessary multiplicity, we assume the center of $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ is set to the origin of $\mathbb{R}^q$: That is,

$$\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathbf{0}.$$

We project $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ onto a unit direction vector $\mathbf{v}$ of $\mathbb{R}^q$, where

$$\mathbf{v} = \sum_{i'=1}^{n} c_{i'} \Phi(\mathbf{x}_{i'}) \quad \text{and} \quad \|\mathbf{v}\|^2 = 1.$$

The projection point, called principal component score, of $\Phi(\mathbf{x}_i)$ on $\mathbf{v}$ can be written as

$$\langle \mathbf{v}, \Phi(\mathbf{x}_i) \rangle = \sum_{i'=1}^{n} c_{i'} \langle \Phi(\mathbf{x}_{i'}), \Phi(\mathbf{x}_i) \rangle ,$$

where $< \Phi(\mathbf{x}_{i'}), \Phi(\mathbf{x}_i) >$ is the inner product of $\Phi(\mathbf{x}_{i'})$ and $\Phi(\mathbf{x}_i)$ in $\mathbb{R}^q$.

[1] Professor, Department of Statistics, Korea University, Anam-Dong 5-1, Sungbuk-Gu, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr

Kernel trick assumes $< \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_{i'}) >$ is a function of $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, denoted by $K(\mathbf{x}_i, \mathbf{x}_{i'})$. Most popular form of $K(\mathbf{x}_i, \mathbf{x}_{i'})$ is the radial basis function(RBF) kernel, defined by

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp\left(-\sigma \| \mathbf{x}_i - \mathbf{x}_{i'} \|^2\right), \quad \sigma > 0.$$

Kernel principal component analysis initiated by Scholkopf *et al.* (1998) computes the principal unit vectors by eigen-decomposing $n \times n$ symmetric matrix $\tilde{K}_X$, which is obtained from

$$\tilde{K}_X = \left(I - \frac{1}{n} J\right) K_X \left(I - \frac{1}{n} J\right), \quad \text{for } K_X = (K(\mathbf{x}_i, \mathbf{x}_{i'})),$$

where I is the $n \times n$ identity matrix and J is the $n \times n$ matrix of which all elements are 1. By the spectral decomposition of $\tilde{K}_X$,

$$\tilde{K}_X = U D_\lambda U^t, \quad \text{for } U^t U = I, \ \lambda = (\lambda_1, \lambda_2, \ldots), \ \lambda_1 > \lambda_2 > \cdots,$$

the weighting vector $\mathbf{c}^{[s]}$ of the s^{th} component is obtained from $\mathbf{c}^{[s]} = \lambda_s^{-0.5} \mathbf{u}^{[s]}$, $s = 1, 2, \ldots$, where $\mathbf{u}^{[s]}$ is the s^{th} column of U.

Then, the s^{th} dimension principal component score for the i^{th} observation is given by

$$\begin{aligned}
&\sum_{i'=1}^{n} c_{i'}^{[s]} \left\langle \Phi(\mathbf{x}_{i'}) - \frac{1}{n} \sum_{i''=1}^{n} \Phi(\mathbf{x}_{i''}), \Phi(\mathbf{x}_i) - \frac{1}{n} \sum_{i'''=1}^{n} \Phi(\mathbf{x}_{i'''}) \right\rangle \\
&= \sum_{i'=1}^{n} c_{i'}^{[s]} \left(K(\mathbf{x}_{i'}, \mathbf{x}_i) - \frac{1}{n} \sum_{i''=1}^{n} K(\mathbf{x}_{i''}, \mathbf{x}_i) - \frac{1}{n} \sum_{i'''=1}^{n} K(\mathbf{x}_{i'}, \mathbf{x}_{i'''}) + \frac{1}{n^2} \sum_{i''=1}^{n} \sum_{i'''=1}^{n} K(\mathbf{x}_{i''}, \mathbf{x}_{i'''}) \right). \qquad (1.1)
\end{aligned}$$

Thus, the s^{th} dimension principal component scores for all observations are represented by

$$\left(I - \frac{1}{n} J\right) K_X \left(I - \frac{1}{n} J\right) \mathbf{c}^{[s]} \quad \text{or} \quad \tilde{K}_X \, \mathbf{c}^{[s]}.$$

Readable materials on kernel methods in general and the kernel PCA can be found at Hastie *et al.* (2009).

As a numerical example, we consider the iris data ($n = 150$, three species: Setosa, Versicolor, Virginica), of which four numerical variables (sepal length, sepal width, petal length, petal width) are standardized and put to the kernel principal component analysis via RBF with $\sigma = 0.2$. Figure 1 shows the two-dimensional principal component plane of four-variate observations in different colors by species.

In Figure 1, Setosa are well separated from the other species. Versicolor and Virginica overlap along a curved line. A natural question which may arise for further interpretation of Figure 1 is on the features that distinguish the species. For instance, what are relative characteristics of Setosa in terms of measured variables?

In this study, we build a graphical scheme for representing variables in the kernel principal component analysis. We also define an index of the importance for individual variables in the principal component plane.

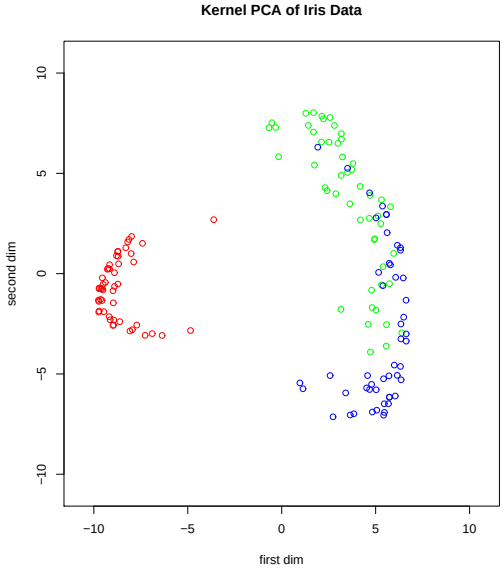

Figure 1: *Kernel PCA of the iris data*

2. Expression of Variables in the Kernel Principal Component Plane

Kernel PCA provides a nonlinear mapping of observations on a low dimensional plane of principal components. Hence it is not an easy task to grasp the directions of variables concisely on the principal component plane. We propose the following scheme for the aim.

1) For each $j\,(=1,\ldots,p)$, map $\mathbf{x}_i^{[j]} = \mathbf{x}_i + \delta\,\mathbf{e}_j$ on the plane, $i = 1,\ldots,n$, where $\delta > 0$ is a constant and $\mathbf{e}_j = (0,\ldots,1,\ldots,0)$. As an extension of Equation (1.1), the projection point of $\mathbf{x}_i^{[j]}$ is given by

$$\sum_{i'=1}^{n} c_{i'}^{[s]} \left(K\left(\mathbf{x}_{i'}, \mathbf{x}_i^{[j]}\right) - \frac{1}{n}\sum_{i''=1}^{n} K\left(\mathbf{x}_{i''}, \mathbf{x}_i^{[j]}\right) - \frac{1}{n}\sum_{i''=1}^{n} K\left(\mathbf{x}_{i'}, \mathbf{x}_{i''}\right) + \frac{1}{n^2}\sum_{i''=1}^{n}\sum_{i'''=1}^{n} K\left(\mathbf{x}_{i''}, \mathbf{x}_{i'''}\right) \right). \quad (2.1)$$

2) For each j, connect the projection points of $\mathbf{x}_i$ and $\mathbf{x}_i^{[j]}$ by arrows, $i = 1,\ldots,n$. Thus the arrows represent the hypothetical changes accompanying the move from $\mathbf{x}_i$ to $\mathbf{x}_i^{[j]}$, a perturbation of $\mathbf{x}_i$ in the direction of $\mathbf{e}_j$. The magnitude of perturbations is specified by $\delta > 0$, which can be set to 0.25, 0.5, 1 or 2 when the variables are standardized to have standard deviations 1. The choice of δ is dependent on the dataset, notably by the number of variables p, and can be finalized after trying several values.

For the kernel PCA of the iris data shown in Figure 1, we set $\delta = 0.5$ and inserted $n\,(=150)$ arrows for each variable in $p\,(=4)$ plots of Figure 2. In the top left plot for the sepal length, we see that Setosa in red color are smaller compared to the other species and that Versicolor (in green color) are a bit smaller than Virginica (in blue color).

Similar patterns can be found for the petal length (in bottom left plot) and the petal width (in the bottom right plot). The sepal width is different: Setosa tend to be larger than Versicolor and the Virginica.

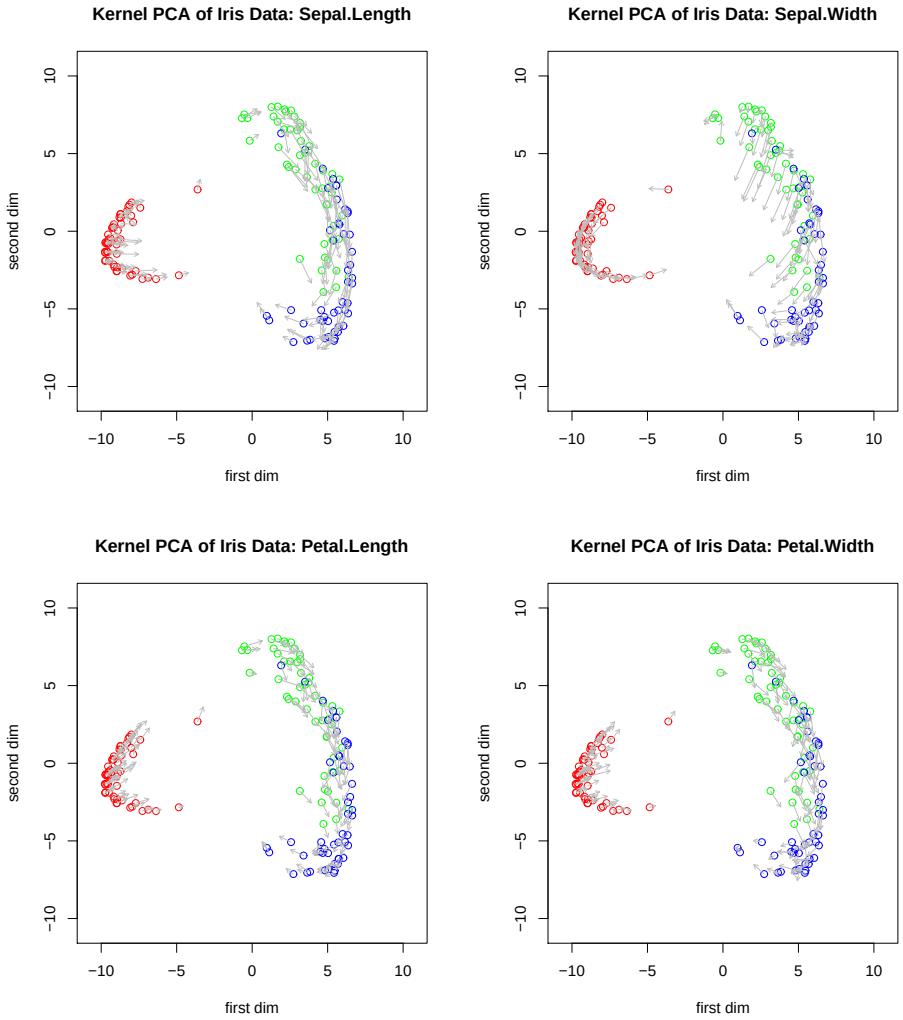

Figure 2: *Arrow diagrams for kernel PCA of the iris data*

Trial of different δ's such as 0.25 and 1 do not alter the interpretation to any meaningful degree in the iris data. All computations are carried in R System with the `kernlab` library (Karatzoglou *et al.*, 2004; Karatzoglou *et al.*, 2012).

3. Measures of Importance for Each Variable

For the case of datasets with "moderate" or "large" p, the arrow diagrams like Figure 2 may be a burden to data analysts. Since important variables may number only a few, we need an index which measures the importance for each variable in the principal component plane. We define the Importance Index for Variable $j\,(=1,\ldots,p)$ by the total of squared norms of the difference between Equation (1.1) and Equation (2.1) over a batch of observations. After evaluating p values of the index (which may depend on δ), we may select a few prominent variables to draw the arrow diagrams.

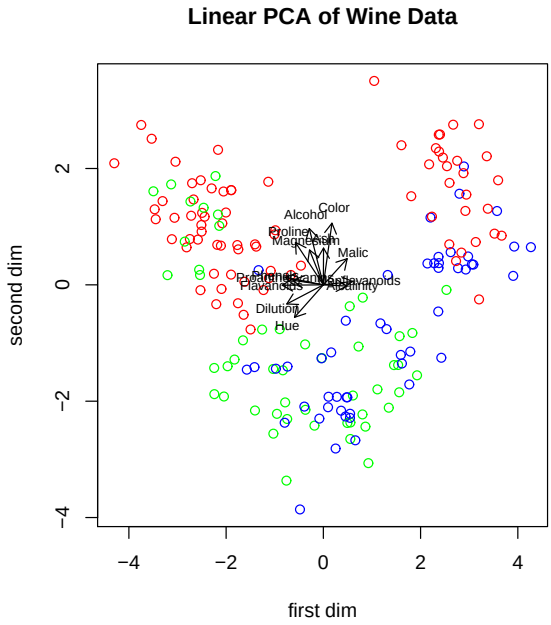

Figure 3: *Biplot for linear PCA of the wine data*

As a numerical example, we consider the wine data of the UCI Machine Learning Repository (http://www.ics.uci.edu/ mlearn/), in which 178 (= n) wine samples are measured for 13 (= p) chemical analyses. Wine samples are classified into three kinds, but we do not use the classification information in the kernel PCA. Figure 3 shows the biplot for linear PCA. We see that three types of the wine overlap with each other to a considerable degree. Hence, linear PCA is not effective for this dataset.

When RBF kernel with $\sigma = 0.04$ is applied, the importance measures in relative percentage turn out to be 14, 13, 12, 7, 7, 7, 7, 7, 7, 6, 5, 5, 4 in decreasing order for $\delta = 1$. Thus only the first three variables (Color, Alcohol, Proline) are prominent. Figure 4 shows the arrow diagrams for three prominent variables (and the least important variable Alcalinity as a contrast).

In the plots of Figure 4, we see that the second type (green) wine tends to be smaller compared to the other type (red or blue) wines in Color, Alcohol and Proline. It is remarkable that the three types of wine are well distinguishable and that the second type wine is located between the other type wines on an inverted U-shape curve.

4. The Case of Many Observations

When the dataset has a large number of observations, say more than 1,000, proposed diagrams may appear over-crowded with arrows. If it is such a case, a sensible approach is to sample a fraction of observations and to draw the diagram only for the selected observations.

As a numerical example, we consider the spam data of the UCI Machine Learning Repository in which 4,601 (= n) e-mails with 57 (= p) morphological characters or frequencies of certain words are classified into two classes (spam, non-spam).

By applying the kernel PCA with RBF parameter $\sigma = 0.01$, we obtained Figure 5, in which spam

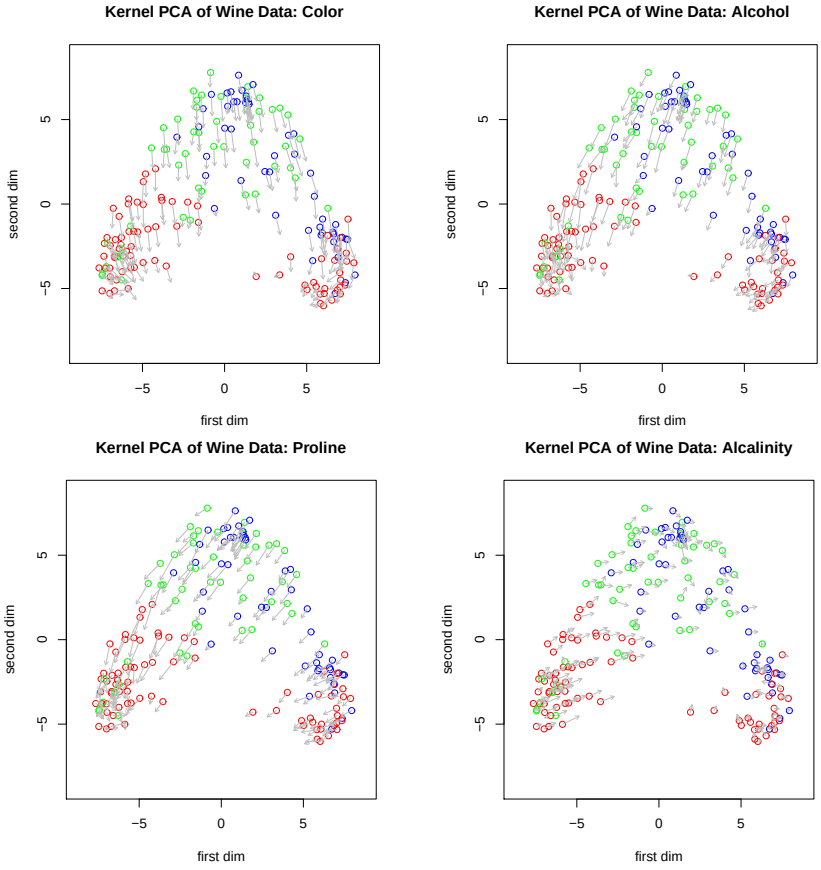

Figure 4: *Arrow diagrams for kernel PCA of the wine data for selected variables*

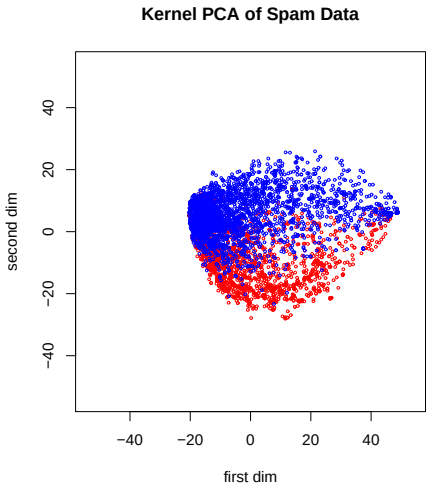

Figure 5: *Kernel PCA of the spam data*

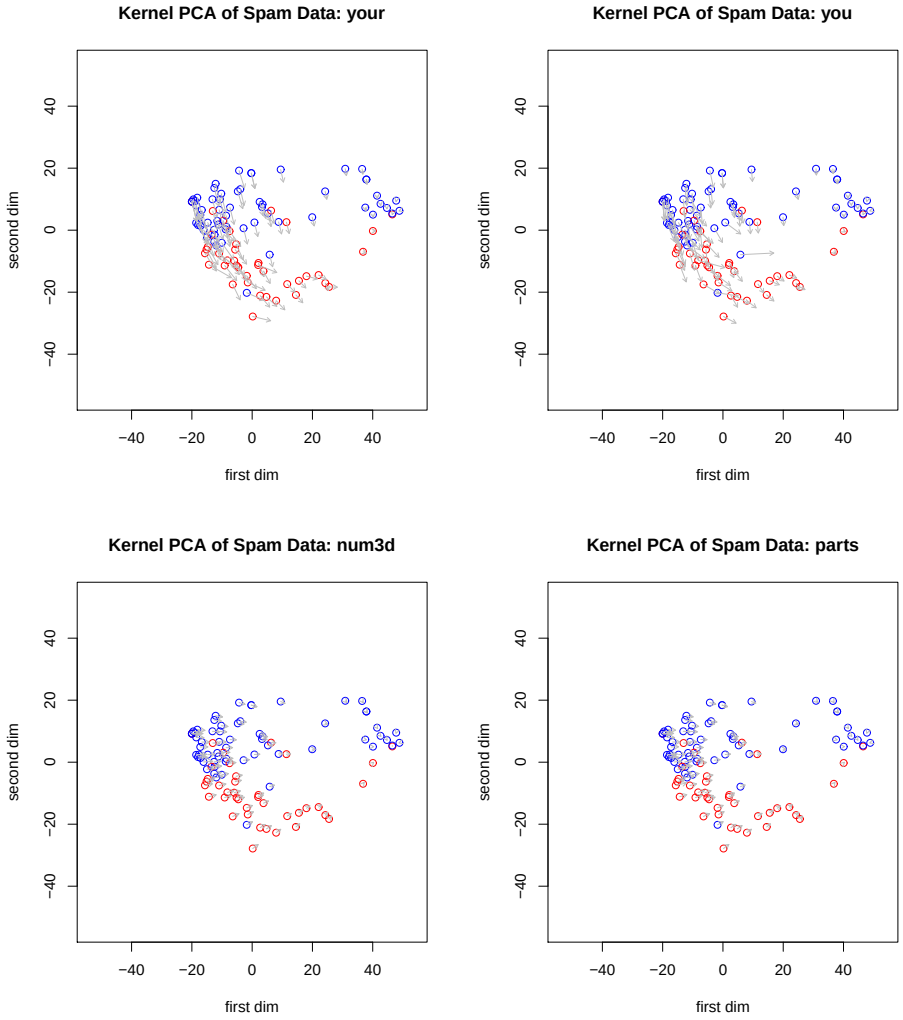

Figure 6: *Arrow diagrams for kernel PCA of the spam data for selected variables*

mails are colored in red and non-spam mails are colored in blue. Clearly, there is no room for arrows for all observations. Hence we sampled 100 ($= n_1$) observations randomly from the dataset and made the arrow diagrams with selected observations.

Figure 6 shows the diagrams ($\delta = 2$) for two most important variables, "your" and "you", and two least important variables, "num3d" and "parts", in determining principal component dimensions. We see that spam mails have larger values of "your" and "you" compared to non-spam mails.

5. Concluding Remarks

This study did not focus the optimal choice of the types of kernel functions and kernel and parameters. However, arrow diagrams proposed in this study can be valuable in assessing the appropriateness of

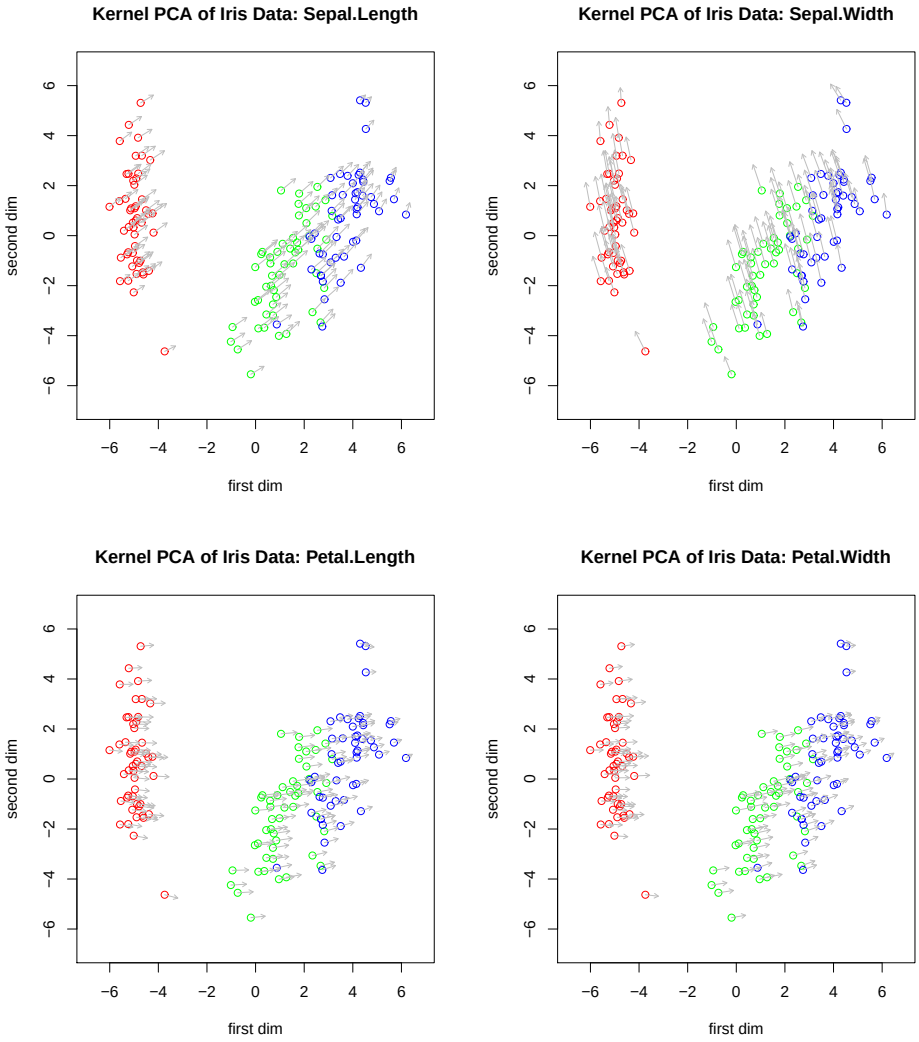

Figure 7: *Arrow diagrams for kernel PCA of the iris data with* $\sigma = 0.02$ *and* $\delta = 0.5$

kernel parameters. For instance, by setting RBF kernel parameter σ to 0.02 or 0.5 for the iris data, we have arrow diagrams Figure 7 and Figure 8, respectively. In Figure 7 ($\sigma = 0.02$), the diagrams show clearly that kernel PCA is very similar to linear PCA, since all arrows are more or less parallel and equal-sized in each plot. In Figure 8 ($\sigma = 0.5$), the diagrams are not interpretable since all arrows head for one point, implying that the sigma around 0.5 is excessive in this particular dataset. In short, the proposed graphical scheme provides visual tools to data analysts using the kernel methods.

Since the kernel PCA derives its principal weighting coefficients by eigen-decomposing $n \times n$ matrix $\tilde{K}_X$, its practical usefulness could be weakened for the dataset with large n, the number of observations, i.e. $\geq 1,000$. In such a case, a bypass is to select a fraction of the observations in building the kernel PCA. Unselected observations can be plotted over the graph using Equation (1.1).

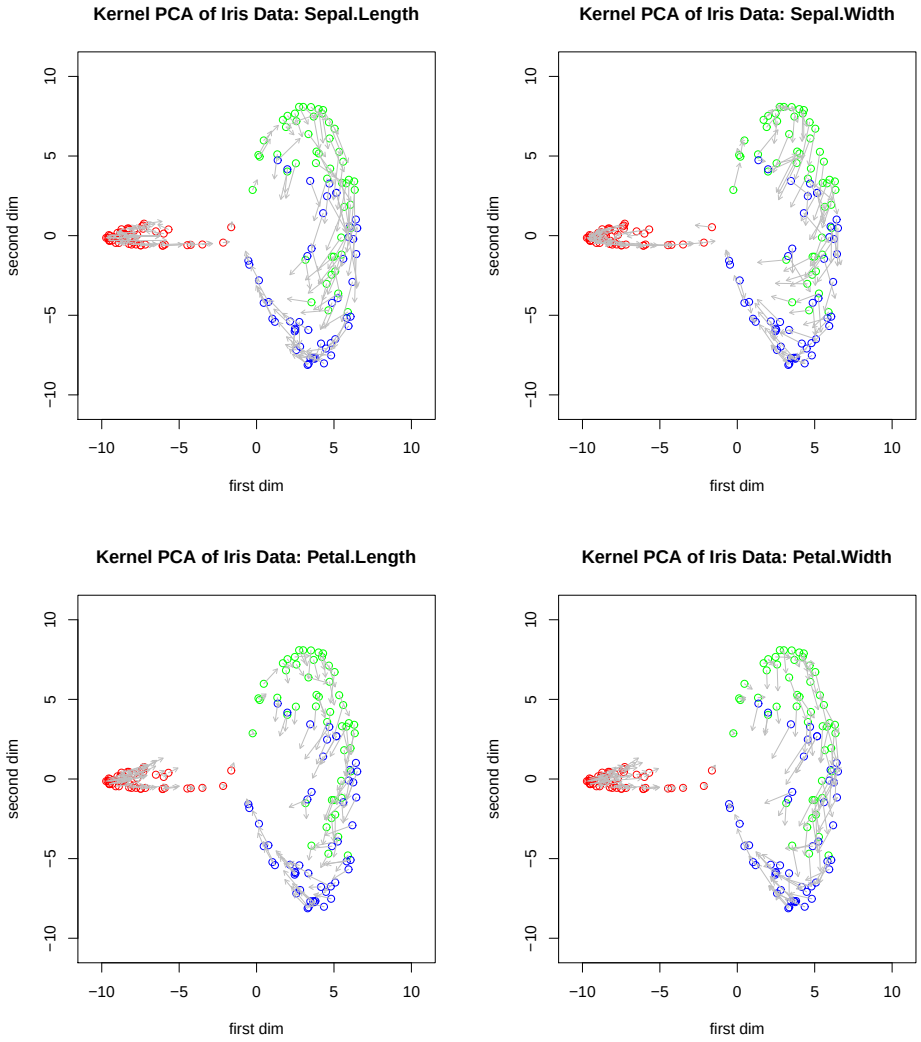

Figure 8: *Arrow diagrams for kernel PCA of the iris data with* $\sigma = 0.5$ *and* $\delta = 0.5$

References

Gabriel, K. R. (1971). The biplot display of matrices with the application to principal component analysis, *Biometrika*, **58**, 453–467.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, Second Edition, Springer, New York.

Karatzoglou, A., Smola, A., Hornik, K. and Zeileis, A. (2004). 'kernlab' - An S4 package for kernel methods in R, *Journal of Statistical Software*, **11**, 1–20.

Karatzoglou, A., Smola, A. and Hornik, K. (2012). R Package 'kernlab' (Version 0.9-15), http://cran.r-project.org/

Scholkopf, B., Smola, A. and Muller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation*, **10**, 1299–1319.

Received February 6, 2013; Revised March 11, 2013; Accepted April 8, 2013