Communications for Statistical Applications and Methods
2014, Vol. 21, No. 4, 317–325

DOI: http://dx.doi.org/10.5351/CSAM.2014.21.4.317
Print ISSN 2287-7843 / Online ISSN 2383-4757

Global and Local Views of the Hilbert Space Associated to Gaussian Kernel

Myung-Hoe Huh[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

Consider a nonlinear transform $\Phi(\mathbf{x})$ of $\mathbf{x}$ in $\mathbb{R}^p$ to Hilbert space H and assume that the dot product between $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ in H is given by $< \Phi(\mathbf{x}), \Phi(\mathbf{x}') >= K(\mathbf{x}, \mathbf{x}')$. The aim of this paper is to propose a mathematical technique to take screen shots of the multivariate dataset mapped to Hilbert space H, particularly suited to Gaussian kernel $K(\cdot\,,\cdot)$, which is defined by $K(\mathbf{x}, \mathbf{x}') = \exp(-\sigma \| \mathbf{x} - \mathbf{x}' \|^2)$, $\sigma > 0$. Several numerical examples are given.

Keywords: Data visualization, Hilbert space, Gaussian kernel, principal component analysis.

1. Background and Aim

Suppose that one has a dataset containing p-variate n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and that all variables being considered are numerical. Data analysts have always wanted effective dimensional reduction techniques for exploratory visualization.

Gabriel's (1971) biplot is certainly a handy tool for this purpose. The method produces a scatterplot of observations overlaid with arrows for the directional flow of respective variables. However, the biplot is effective only for linearly structured datasets, because its mechanism consists of linear projections in $\mathbb{R}^p$. Recent trends in data science push the analysts to work on the datasets of large p and, thus, not to rely on the linearity assumptions.

In this paper, I propose a mathematical technique to take screen shots of the multivariate dataset mapped to Hilbert space H, particularly suited to Gaussian kernel K:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\sigma \left\| \mathbf{x} - \mathbf{x}' \right\|^2\right), \quad \sigma > 0. \tag{1.1}$$

The proposed method utilizes the kernel projection of nonlinear transforms $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ in H of $\mathbf{x}_1, \ldots, \mathbf{x}_n$. In that respect, it is a kind of kernel principal component analysis, originally developed by Schölkopf *et al.* (1998). However, the proposed technique differs from the standard method in dealing with central tendency of transformed observations in the Hilbert space. In addition, the proposed method produces several (two or more) pictures of transformed observations in H, to help the analysts by providing diverse views of a big dataset which could be complex in nature.

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

Published online 31 July 2014 / journal homepage: http://csam.or.kr

2. Gaussian Kernel for Hilbert Space and Principal Components

Suppose that the dot product between two image vectors $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ in H is given by Gaussian kernel of (1.1). Since

$$\|\Phi(\mathbf{x})\|_H^2 = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle = 1, \quad \text{for all } \mathbf{x},$$

the norms of all images of $\mathbf{x}$ in $\mathbb{R}^p$ are the same and equal to 1. Hence the nonlinear transform $\Phi(\mathbf{x})$ takes $\mathbf{x}$ to the surface of S, a unit-radius sphere embedded in H. This fact is a unique characteristic property of the Gaussian kernel.

Note that the squared distance between two image vectors $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ in H is given by

$$\begin{aligned}\|\Phi(\mathbf{x}) - \Phi(\mathbf{x}')\|_H^2 &= \langle \Phi(\mathbf{x}) - \Phi(\mathbf{x}'), \Phi(\mathbf{x}) - \Phi(\mathbf{x}') \rangle \\ &= 2 - 2K(\mathbf{x}, \mathbf{x}') \\ &= 2 - 2\exp\left(-\sigma \|\mathbf{x} - \mathbf{x}'\|^2\right).\end{aligned}$$

Thus the order of all pairwise distances is invariant of the transformation. However, the magnitude of distances is altered via a monotonic concave function, *i.e.* $2 - 2e^{-\sigma D}$, where D denotes the squared distance in $\mathbb{R}^p$. An implication of this property is that the longer distance between two observation vectors in $\mathbb{R}^p$ is relatively shortened after the transformation.

Furthermore, note that $< \Phi(\mathbf{x}), \Phi(\mathbf{x}') >$ is positive for all $\mathbf{x}$ and $\mathbf{x}'$, since the inner product between $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ is equal to $K(\mathbf{x}, \mathbf{x}')$ of (1.1). Hence, the angle between two image vectors $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ is less than $\pi/2$, and one may search for the primary orientation of n image vectors in the following way.

Consider the projections of $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ on the unit-length vector $\mathbf{v}$ in H. Naturally, one determines $\mathbf{v}$ in a way that

$$\max_{\mathbf{v}} \sum_{i=1}^{n} < \Phi(\mathbf{x}_i), \mathbf{v} >^2 \quad \text{subject to } \|\mathbf{v}\|_H^2 = 1.$$

As argued in Schölkopf *et al.* (1998), it suffices to consider

$$\mathbf{v} = \sum_{i=1}^{n} \Phi(\mathbf{x}_i)\, d_i, \tag{2.1}$$

where $d_1, \ldots, d_n$ are constants to be determined. Therefore, the optimization is formulated as

$$\max_{\mathbf{d}} \mathbf{d}^t K^2 \mathbf{d} \quad \text{subject to } \mathbf{d}^t K \mathbf{d} = 1,$$

where K is the $n \times n$ symmetric positive definite matrix consisting of the dot products between $\Phi(\mathbf{x}_i)$ and $\Phi(\mathbf{x}_{i'})$, that is given by $K(\mathbf{x}_i, \mathbf{x}_{i'})$.

To get the solution, eigen-decompose K:

$$K = U D_\lambda U^t,$$

where U is an orthogonal matrix of eigenvectors and λ is the vector of eigenvalues ≥ 0. Optimal $\mathbf{d}$ is $\lambda_1^{-0.5}\mathbf{u}_1$ ($:= \mathbf{d}_1$), where λ_1 is the largest eigenvalue and $\mathbf{u}_1$ is the corresponding eigenvector. Kernel principal component scores are given by elements of $K\,\mathbf{d}_1$, and the sum of all eigenvalues is n since it equals to the trace of K.

To explore the Hilbert space further, one needs to examine other dimensions beyond the primary dimension. For the purpose, consider (2.1) again with the constraint

$$\left\langle \sum_{i=1}^{n} \Phi(\mathbf{x}_i)\, d_i^1, \sum_{i=1}^{n} \Phi(\mathbf{x}_i)\, d_i \right\rangle = 0,$$

where $\mathbf{d}_1 = (d_1^1, \ldots, d_n^1)$. Mathematical optimization leads to $\lambda_2^{-0.5}\mathbf{u}_2(:= \mathbf{d}_2)$ and, in a similar way, the third dimension is given by $\lambda_3^{-0.5}\mathbf{u}_3(:= \mathbf{d}_3)$, where λ_2, λ_3 are the second and the third largest eigenvalues of K and $\mathbf{u}_2$, $\mathbf{u}_3$ are corresponding eigenvectors. The 3D kernel principal component image is then obtained by $Z_{[3]} := (K\,\mathbf{d}_1, K\,\mathbf{d}_2, K\,\mathbf{d}_3)$.

The above approach is similar, but not identical, to the steps taken by Schölkopf *et al.* (1998). They subtracted the mean of transformed observations from all individual observations and set

$$\tilde{\mathbf{v}} = \sum_{i=1}^{n} \left(\Phi(\mathbf{x}_i) - \frac{1}{n} \sum_{i'=1}^{n} \Phi(\mathbf{x}_{i'}) \right) d_i \tag{2.2}$$

as principal directional vector. Subsequently, they derived the solution via eigen-decomposition of

$$\tilde{K} = \left(I_n - \frac{1}{n} J_n \right) K \left(I_n - \frac{1}{n} J_n \right),$$

where I_n denotes the $n \times n$ identity matrix and J_n denotes the $n \times n$ matrix of elements all equal to 1.

Conceptually, the first and the second dimensions of Schölkopf *et al.*'s method correspond to the second and third dimensions of the proposed method, provided that the direction of $\mathbf{d}_1$ is close to that of $\mathbf{1}_n$, possibly up to the sign. In the iris data, the absolute cosine of the angle between $\mathbf{d}_1$ and $\mathbf{1}_n$ is 0.98, indicating that the two vectors are oriented toward the very close directions. One possible criticism of Schölkopf *et al.*'s method is that the mean of $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ does not lie on the spherical surface when the Gaussian kernel is in action. In contrast, $\sum_{i=1}^{n} \Phi(\mathbf{x}_1) d_i^1$ lies on the spherical surface.

As a numerical illustration, consider the iris data. Four numerical measurements are put to the kernel principal component analysis proposed above. Gaussian kernel with $\sigma = 0.1$ is used. Figure 1 shows two screen shots of the 3D images of 150 observations mapped to the Hilbert space, colored according to the Species. By rotating the 3D box, one feels that the image points are roughly on the spherical surface.

The choice of kernel parameter σ affects the kernel principal component images. So far, there exists no systematic way for finding suitable values of σ. In general, Gaussian kernel with too small σ lenders a trivial deformation of the data cloud, whereas Gaussian kernel with too large σ yields a grotesque deformation. Hence, with several different σ values, produce a multitude of graphs and select the best view of the data. Some degree of subjectivity seems inevitable.

As for the goodness of 3D view, a natural measure is the proportion of first three eigenvalues over the sum of all eigenvalues:

$$G_1 = \frac{1}{n}(\lambda_1 + \lambda_2 + \lambda_3)$$

since $\sum_{i=1}^{n} \lambda_i$ is equal to n. For the iris data, $G_1 = 0.893$.

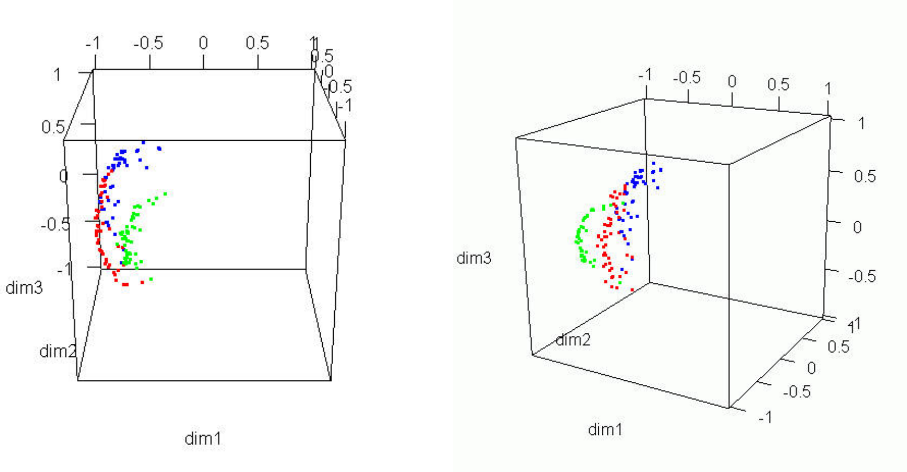

Figure 1: *Kernel principal component images of the iris data*

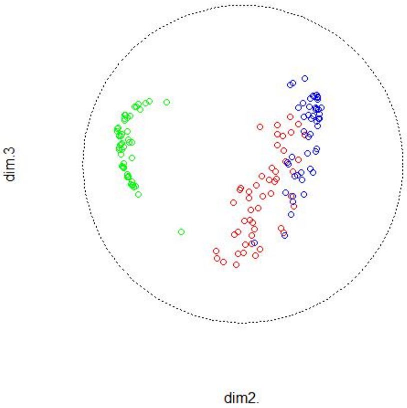

Figure 2: *Global view of the iris data transformed to the Hilbert space*

For the view of spreadness from the center, one needs to plot the second versus the third principal component dimension by decomposing K. Figure 2 is a plot for the iris data. In such a graph, the goodness measure can be defined as

$$G_2 = \frac{(\lambda_2 + \lambda_3)}{(\lambda_2 + \lambda_3 + \cdots + \lambda_n)}.$$

In the case of center-adjusted decomposition of the kernel matrix (Schölkopf *et al.*, 1998), the goodness measure for their 2D view is

$$\tilde{G}_2 = \frac{\left(\tilde{\lambda}_1 + \tilde{\lambda}_2\right)}{\left(\tilde{\lambda}_1 + \tilde{\lambda}_2 + \cdots + \tilde{\lambda}_{n-1}\right)}.$$

where $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_{n-1}$ are eigenvalues of $\tilde{K}$. For the iris data, $G_2 = 0.749$ and $\tilde{G}_2 = 0.738$. Thus, they are comparable each other.

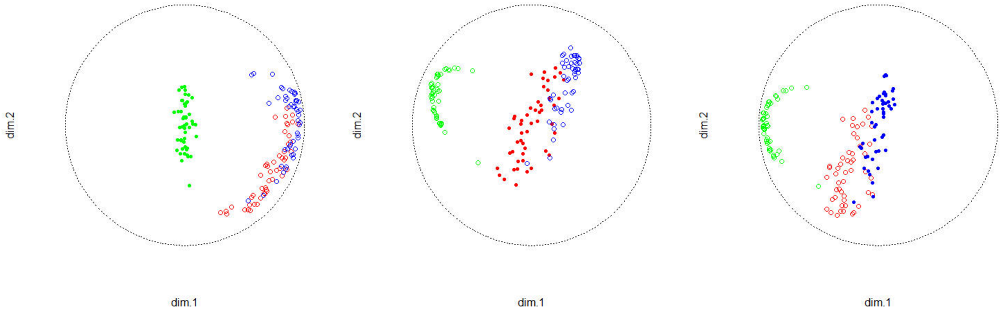

Figure 3: *Local views of the iris data. Center groups are defined by the Species*

3. Local 3D View of the Hilbert Space

In 2D views of the 3D sphere, the boundary of the sphere is a circle, so that clusters of image points deviated from the center appear curved concave toward the center even though it is linear on the spherical surface.

To see the data clouds more correctly, one needs to rotate 3D box like the one in Figure 1 and save several screen shots. To formalize things, I propose the following procedure:

- First, partition the dataset into a number of groups or clusters. Denote the number of groups by g.

- Second, compute the centers of observation groups on 3D space with the data points at $Z_{[3]}$. Denote them by $\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_g$ (of which the squared norms are scaled to 1).

- Third, for the group $j\,(= 1, \ldots, g)$, set up the orthogonal base for the 3D sphere. The base triplet for the group j should include m_j. The second and third base vectors will be denoted by $\mathbf{s}_j$ and $\mathbf{t}_j$. Specifically,

$$\mathbf{s}_j = \frac{\mathbf{e}_2 - \left(\mathbf{m}_j^t\mathbf{e}_2\right)\mathbf{m}_j}{\left\|\mathbf{e}_2 - \left(\mathbf{m}_j^t\mathbf{e}_2\right)\mathbf{m}_j\right\|}, \qquad \mathbf{f}_j = \frac{\mathbf{e}_3 - \left(\mathbf{m}_j^t\mathbf{e}_3\right)\mathbf{m}_j}{\left\|\mathbf{e}_3 - \left(\mathbf{m}_j^t\mathbf{e}_3\right)\mathbf{m}_j\right\|}$$

  and

$$\mathbf{t}_j = \frac{\mathbf{f}_j - \left(\mathbf{s}_j^t\mathbf{f}_j\right)\mathbf{s}_j}{\left\|\mathbf{f}_j - \left(\mathbf{s}_j^t\mathbf{f}_j\right)\mathbf{s}_j\right\|},$$

  where $\mathbf{e}_2 = (0, 1, 0)^t$ and $\mathbf{e}_3 = (0, 0, 1)^t$. Then, it is guaranteed that $\mathbf{m}_j$, $\mathbf{s}_j$ and $\mathbf{t}_j$ are orthonormal.

- Fourth, for each $j\,(= 1, \ldots, g)$, plot n image points at $(Z_{[3]}\mathbf{s}_j, Z_{[3]}\mathbf{t}_j)$.

Figure 3 shows the output graphs for the iris data with the Species as group indicator. In the left panel, one can see the Species Setosa observations at the center and two other Species on the sides. In the middle and right panels, one can see the Species Versicolor and Virginica overlap a little. More realistic workout case studies are given in the next two sections.

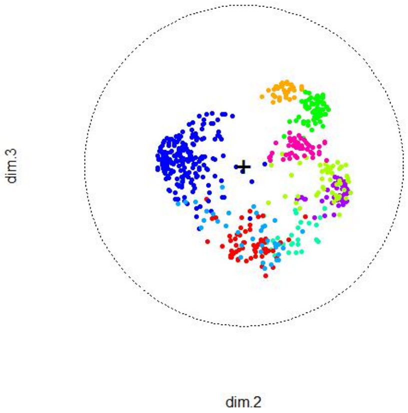

Figure 4: *Global views of Olives data transformed to the Hilbert space*

4. Olives Data

Olives data of R package `classifly` consists of the percentage composition of eight fatty acids (palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, eicosenoic) for 572 Italian olive oils, collected in nine areas. Kernel parameter σ is set to 0.1.

Figure 4 shows one global view of olives data, colored differently according to collection areas. In the plot, one sees that several groups overlap to a considerable degree. The goodness measure G_2 is 0.365. In comparison, $\tilde{G}_2 = 0.369$.

Figure 5 shows nine local views of olives data, one for each collection area. In the pictures, one can see that Area 1, Area 5 and Area 6 overlap and that Area 2, Area 4 and Area 9 are close to each other. Certainly, several local pictures show more details than a single global picture.

5. Spam Data

Spam data of R package `kernlab` consists of 57 features of 4,601 e-mails, each of which is classified into either non-spam or spam. Kernel parameter σ is set to 0.01.

Figure 6 shows one global view of the spam data, derived from a 25% subset data. The number of variables p is 57, larger than previous examples. The goodness measure G_2 is only 0.10. Apparently, even in the Hilbert space, representing high-dimensional data on a low dimensional plane is a difficult task. However, one sees that the two groups form the shape of bird wings suggesting the separation is plausible economically.

By applying K-means clustering to the subset data not using the classification marker, five groups are obtained. Figure 7 shows a number of local views: Cluster 1 is a small group of outliers. largely green-colored (non-spam), Clusters 2 is almost entirely spam mail (colored in red), Cluster 3 is largely non-spam mails (colored in green). Cluster 4 is entirely non-spam mail. Cluster 5 is largely spam mail. With such perceptual information, one may fit the support vector machine for classifying non-spam vs spam. SVM with the kernel parameter $\sigma = 0.01$ and the cost parameter 10 produces 7.8% classification error by 10-fold cross-validation.

Graphical representation of clusters in multiple plots is not new. Huh and Lee (2013) proposed a similar scheme that uses the results of principal component analysis of individual cluster.

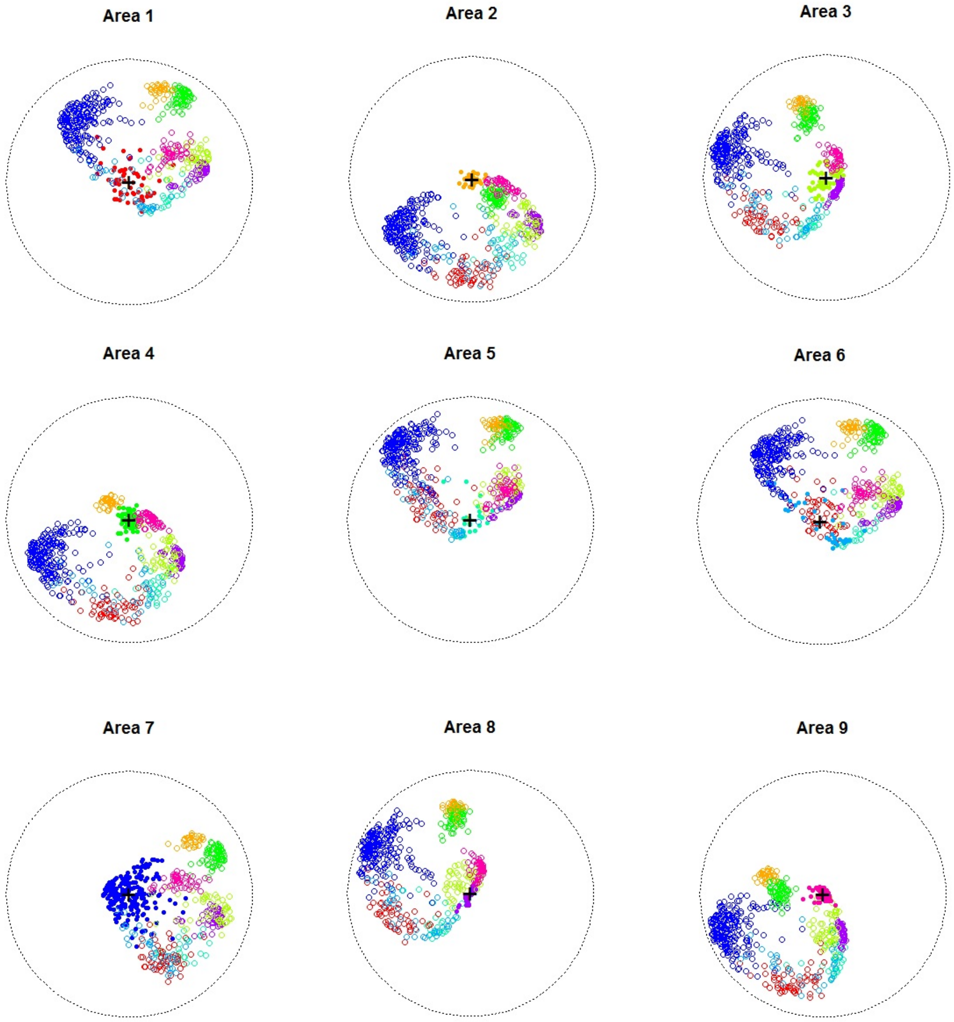

Figure 5: *Nine local views of Olives data*

6. Concluding Remarks

This study proposes graphing methods for the dataset mapped to Hilbert space, where the dot products are computed via Gaussian kernel. Mathematically, it is nothing else but a kernel principal component analysis, which was set up by Schölkopf *et al.* (1998). However, the proposed method differs from the original method in two ways. First, the new method is based on the decomposition of K, while the existing method is based on the decomposition of $\tilde{K}$. Second, the new method produces multiple plots, while the existing method produces a single plot.

All graphs appearing in this paper are plots of observations. The graphs will be more valuable if the information of variables is reflected on. For that purpose, one may draw arrow diagrams, which was proposed and studied by Huh (2013a).

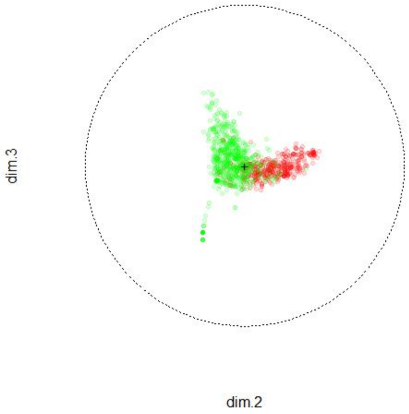

Figure 6: *Global view of the spam data transformed to the Hilbert space*

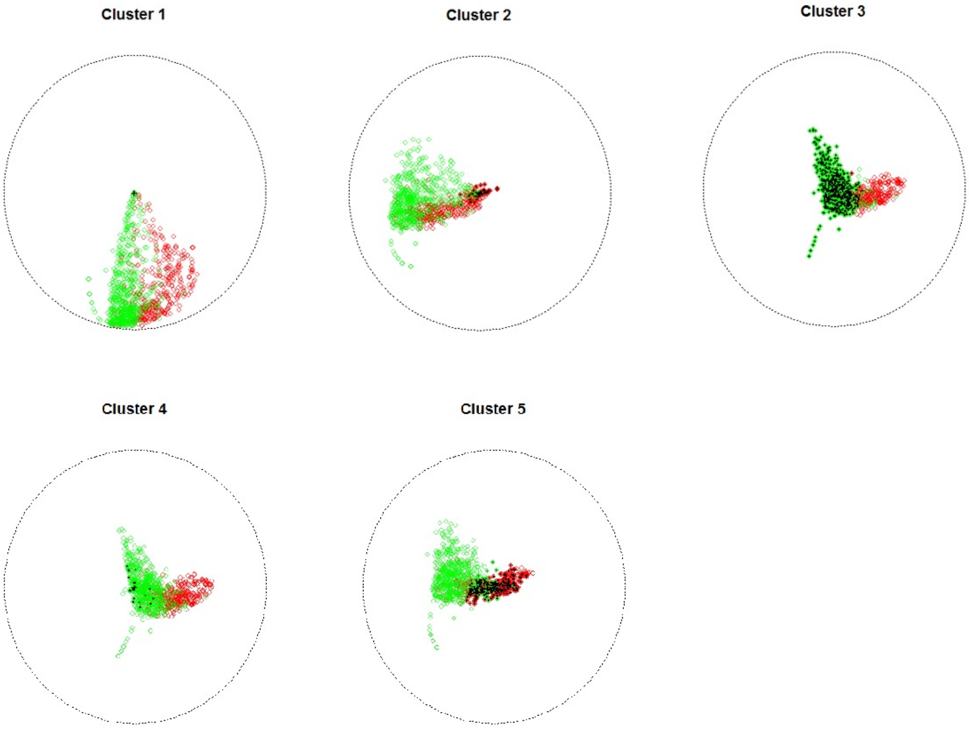

Figure 7: *Five local views of the spam data*

For the dataset with a dependent variable of either binary or continuous type, Huh (2013b) proposed a visualization method, called the SVM-guided biplot, that combines support vector machine classification (or regression) and kernel principal component analysis.

References

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

Huh, M. H. (2013a). Arrow diagrams for kernel principal component analysis, *Communications for Statistical Applications and Methods*, **20**, 175–184.

Huh, M. H. (2013b). SVM-guided biplot of observations and variables, *Communications for Statistical Applications and Methods*, **20**, 491–498.

Huh, M. H. and Lee, Y. G. (2013). Biplots of multivariate data guided by linear and/or logistic regression, *Communications for Statistical Applications and Methods*, **20**, 129–136.

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

Received May 13, 2014; Revised June 14, 2014; Accepted June 14, 2014