Communications for Statistical Applications and Methods
2013, Vol. 20, No. 6, 491–498

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.6.491

SVM-Guided Biplot of Observations and Variables

Myung-Hoe Huh[1,a]

[a]Department of Statistics, Korea University

Abstract

We consider support vector machines(SVM) to predict Y with p numerical variables $X_1, \ldots, X_p$. This paper aims to build a biplot of p explanatory variables, in which the first dimension indicates the direction of SVM classification and/or regression fits. We use the geometric scheme of kernel principal component analysis adapted to map n observations on the two-dimensional projection plane of which one axis is determined by a SVM model *a priori*.

Keywords: Support vector machine, kernel trick, principal component analysis, biplot.

1. Background and Aim

Suppose that we have a dataset that consists of a response variable Y and p explanatory numerical variables $X_1, \ldots, X_p$. Support vector machine(SVM) produces flexible classification and regression models in an efficient way, even in the case of a large p compared to the number of observations n, using the so-called "kernel trick". One possible criticism of the SVM relying on nonlinear kernels could be the difficulty in the interpretation of the constructed model, because multivariate observations are mapped onto a Hilbert space, that is quite different from Euclidean space. Recently, there emerged the kernel PCA that projects observations or points of the "intractable" Hilbert space on a reduced dimensional subspace (Schölkopf *et al.*, 1998).

This paper aims to build a biplot of n observations and p explanatory variables, of which the first dimension indicates the direction of SVM classification/regression fits. The second dimension is added to posit n points as widely spread as possible, using the geometric scheme of kernel PCA. Similar graph was developed by Huh and Lee (2013) for the cases of linear or logistic regression.

2. Geometry of Hilbert Space with Kernel Trick

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be p-dimensional numeric observations. We consider the transform $\Phi(\mathbf{x})$ of $\mathbf{x}$ from $\mathbb{R}^p$ to a Hilbert space H and assume that the dot product between the two images $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ of $\mathbf{x}$ and $\mathbf{x}'$ can be obtained through a kernel function $K(\mathbf{x}, \mathbf{x}')$. Then, for a given linear composite $\mathbf{w}$ of $\Phi(\mathbf{x}_1), \ldots, \Phi(\mathbf{x}_n)$ on H, *i.e.*,

$$\mathbf{w} = \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i) \tag{2.1}$$

This study was supported by Korea University Grant.

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

the projection score of $\Phi(\mathbf{x})$ on the unit-normed vector $\mathbf{w}_1$ of (2.1) is given by

$$\frac{1}{s}\sum_{i=1}^{n} c_i K(\mathbf{x}, \mathbf{x}_i), \tag{2.2}$$

where $s = (\mathbf{c}^t K \mathbf{c})^{1/2}$, $\mathbf{w}_1 = 1/s \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i)$ and $K = (k_{ii'})$, $k_{ii'} = K(\mathbf{x}_i, \mathbf{x}_{i'})$. More compactly, (2.2) can be written as

$$\frac{1}{s}\mathbf{k}^* \mathbf{c},$$

where $\mathbf{k}^*$ denotes the $1 \times n$ kernel dot product matrix between $\mathbf{x}$ and $X^t = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$. Now, we consider the projections of $\Phi(\mathbf{x}_i) - 1/n \sum_{i=1}^{n} \Phi(\mathbf{x}_i)$, $i = 1, \ldots, n$, on a linear composite of $\Phi(\mathbf{x}_1) - 1/n \sum_{i=1}^{n} \Phi(\mathbf{x}_i), \ldots, \Phi(\mathbf{x}_n) - 1/n \sum_{i=1}^{n} \Phi(\mathbf{x}_i)$, *i.e.*,

$$\mathbf{v} = \sum_{i=1}^{n} d_i \left(\Phi(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) \right). \tag{2.3}$$

Depending on the choice of $d_1, \ldots, d_n$, the total of squared norms of the projection varies. Kernel PCA answers for the maximal total of squared norms under the constraint $< \mathbf{v}, \mathbf{v} > = 1$ (Schölkopf *et al.*, 1998). The optimal $(d_1, \ldots, d_n)^t (= \mathbf{d})$ equals $\lambda_1^{-0.5}\mathbf{u}_1$, where λ_1 is the primary eigenvalue of

$$\tilde{K} = \left(I - \frac{1}{n} J \right) K \left(I - \frac{1}{n} J \right)$$

and $\mathbf{u}_1$ is the corresponding eigenvector.

For the image $\Phi(\mathbf{x})$ of arbitrary $\mathbf{x}$, the projection score on $\mathbf{v}$ of (2.3) is

$$\begin{aligned} &\sum_{i=1}^{n} d_i \left(K(\mathbf{x}, \mathbf{x}_i) - \frac{1}{n}\sum_{i'=1}^{n} K(\mathbf{x}, \mathbf{x}_{i'}) - \frac{1}{n}\sum_{i''=1}^{n} K(\mathbf{x}_i, \mathbf{x}_{i''}) + \frac{1}{n^2}\sum_{i''=1}^{n}\sum_{i'''=1}^{n} K(\mathbf{x}_{i''}, \mathbf{x}_{i'''}) \right) \\ &= \left(\mathbf{k}^* - \frac{1}{n}\mathbf{k}^* J - \frac{1}{n}\mathbf{1}^t K + \frac{1}{n^2}\mathbf{1}^t K J \right) \mathbf{d}. \end{aligned} \tag{2.4}$$

We note that (2.4) depends only on two dot product matrices, $\mathbf{k}^*$ and K, through $\mathbf{d}$.

In this study, we use the most popular kernel, radial basis function(RBF), from among the available types. RBF kernel is defined as

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

However, the essence of this study is not limited to RBF kernel. More materials on kernel principal component analysis can be found at Karatzoglou *et al.* (2004), Hastie *et al.* (2009) and Huh (2013).

3. SVM Classification and Biplot

Suppose that each of n observations is classified into one of two groups, coded as $y_i = -1$ or 1 for $i = 1, \ldots, n$. Denoting the explanatory part of the i^{th} observation by $p \times 1$ vector $\mathbf{x}_i$, SVM classifies an arbitrary case with explanatory feature $\mathbf{x}$ to one of two groups (-1 or 1) by the sign of

$$f_{SVM}(\mathbf{x}) = < \mathbf{w}, \Phi(\mathbf{x}) > + b_0, \tag{3.1}$$

where $\Phi(\mathbf{x})$ denotes the transform of $\mathbf{x}$ in $\mathbb{R}^p$ to a Hilbert space H and

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i y_i \Phi(\mathbf{x}_i), \quad \lambda_i \geq 0.$$

We assume that the dot product on H between $\Phi(\mathbf{x})$ and $\Phi(\mathbf{x}')$ for arbitrary $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbb{R}^p$ is defined through a kernel function $K(\mathbf{x}, \mathbf{x}')$. Then, (3.1) can be expressed as

$$f_{SVM}(\mathbf{x}) = \sum_{i=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}) + b_0, \tag{3.2}$$

where $c_i = \lambda_i y_i$, $i = 1, \ldots, n$.

The squared norm of $\mathbf{w}$ is equal to

$$\|\mathbf{w}\|^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} c_i K(\mathbf{x}_i, \mathbf{x}_{i'}) c_{i'} = \mathbf{c}^t K \mathbf{c},$$

where $\mathbf{c} = (c_i)$ and $K = (k_{ii'})$, $k_{ii'} = K(\mathbf{x}_i, \mathbf{x}_{i'})$. Hence, the unit vector

$$\mathbf{w}_1 = \frac{1}{s} \sum_{i=1}^{n} c_i \Phi(\mathbf{x}_i), \quad \text{for } s = (\mathbf{c}^t K \mathbf{c})^{\frac{1}{2}}$$

determines the direction of the model fits on H. Therefore, the projection scores of $\Phi(\mathbf{x}_i)$ on $\mathbf{w}_1$ are

$$\left\langle \Phi(\mathbf{x}_i), \frac{1}{s} \sum_{i'=1}^{n} c_{i'} \Phi(\mathbf{x}_{i'}) \right\rangle = \frac{1}{s} \sum_{i'=1}^{n} c_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}), \quad i = 1, \ldots, n,$$

or the n elements of

$$\frac{1}{s} K \mathbf{c}. \tag{3.3}$$

Orthogonal components $\tilde{\Phi}(\mathbf{x}_i)$ of $\Phi(\mathbf{x}_i)$ projected on $\mathbf{w}_1$ are given by

$$\tilde{\Phi}(\mathbf{x}_i) = \Phi(\mathbf{x}_i) - b_i \mathbf{w}_1, \quad \text{for } b_i = \frac{1}{s} \sum_{i'=1}^{n} c_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}).$$

Write $\mathbf{b} = (b_1, \ldots, b_n)^t$ for later use. Hence

$$\begin{aligned} \left\langle \tilde{\Phi}(\mathbf{x}_i), \tilde{\Phi}(\mathbf{x}_{i'}) \right\rangle &= < \Phi(\mathbf{x}_i) - b_i \mathbf{w}_1, \Phi(\mathbf{x}_{i'}) - b_{i'} \mathbf{w}_1 > \\ &= k_{ii'} - b_i < \Phi(\mathbf{x}_{i'}), \mathbf{w}_1 > - b_{i'} < \Phi(\mathbf{x}_i), \mathbf{w}_1 > + b_i b_{i'} < \mathbf{w}_1, \mathbf{w}_1 > \\ &= k_{ii'} - b_i b_{i'}, \quad i, i' = 1, \ldots, n. \end{aligned}$$

Thus, dot products among orthogonal components $\tilde{\Phi}(\mathbf{x}_i)$ of $\Phi(\mathbf{x}_i)$ projected on $\mathbf{w}_1$ are given by n^2 elements of

$$K - \mathbf{b}\mathbf{b}^t \, (= K'') \quad \text{or} \quad K - \frac{1}{s^2} K \mathbf{c}\mathbf{c}^t K^t.$$

Once obtained dot product matrix K'', we can derive the primary direction of $\tilde{\Phi}(\mathbf{x}_1), \ldots, \tilde{\Phi}(\mathbf{x}_n)$ in the Hilbert space H via kernel PCA algorithm. That is, the coefficient vector $\mathbf{d}''$ combining $\tilde{\Phi}(\mathbf{x}_1), \ldots, \tilde{\Phi}(\mathbf{x}_n)$ for the direction of maximal spread is determined from the eigen-decomposion of $(I - (1/n)J)K''(I - (1/n)J)\,(= \tilde{K}'')$. Thus the principal projection scores are

$$\tilde{K}''\mathbf{d}'', \tag{3.4}$$

where $\mathbf{d}'' = \lambda_1''^{-0.5}\mathbf{u}''$, where λ_1'' is the largest eigenvalue of $\tilde{K}''$ and $\mathbf{u}_1''$ is the corresponding eigenvector.

Perturbation Scheme for Arrow Diagram

Let $X^* = X + E_\delta$, $n \times p$, where E_δ denotes a perturbation of zero matrix. Specifically, E_δ could be the zero matrix with the j^{th} column replaced by $\delta\,\mathbf{1}_n$ for some $j = 1, \ldots, p$. In that case, the i^{th} row $\mathbf{x}_i^*$ of X^* is equal to $\mathbf{x}_i^* = \mathbf{x}_i + \delta\,\mathbf{e}_j$ for $\mathbf{e}_j = (0, \ldots, 1, \ldots, 0)^t$. Projection scores of $\Phi(\mathbf{x}_i^*)$ on $\mathbf{w}_1$ are given by

$$b_i^* = \frac{1}{s}\sum_{i'=1}^{n} c_{i'} K\left(\mathbf{x}_i^*, \mathbf{x}_{i'}\right), \quad i = 1, \ldots, n.$$

Thus we compute $\mathbf{b}^* = (b_1^*, \ldots, b_n^*)^t$. More compactly, the scores are obtained through

$$\frac{1}{s}K^*\mathbf{c}, \tag{3.5}$$

where $K^* = (k_{ii'}^*)$ and $k_{ii'}^* = K(\mathbf{x}_i^*, \mathbf{x}_{i'})$.

Write $\tilde{\Phi}(\mathbf{x}_i^*) = \Phi(\mathbf{x}_i^*) - b_i^*\mathbf{w}_1$, $i = 1, \ldots, n$. Then, the dot products between $\tilde{\Phi}(\mathbf{x}_i^*)$ and $\tilde{\Phi}(\mathbf{x}_{i'})$ are

$$\left\langle \tilde{\Phi}(\mathbf{x}_i^*), \tilde{\Phi}(\mathbf{x}_{i'}) \right\rangle = < \Phi(\mathbf{x}_i^*) - b_i^*\mathbf{w}_1, \Phi(\mathbf{x}_{i'}) - b_{i'}\mathbf{w}_1 > = K\left(\mathbf{x}_i^*, \mathbf{x}_{i'}\right) - b_i^* b_{i'},$$

or, the $(i, i')^{th}$ element of

$$K^* - \mathbf{b}^*\mathbf{b}^t \left(= K^{*''}\right).$$

Therefore, the primary dispersion scores for perturbed observations are given by

$$\left(K^{*''} - \frac{1}{n}K^{*''}J - \frac{1}{n}JK'' + \frac{1}{n^2}JK''J\right)\mathbf{d}''. \tag{3.6}$$

We propose "SVM-guided biplot" as the plot of (3.3) versus (3.4) on the vertical and the horizontal axes, respectively, for n observations, overlaid with the arrows oriented to the points of which the vertical and horizontal coordinates equal to (3.5) and (3.6) for each variable.

Example 1: Iris Data

As an example, we consider a subset of the iris data, restricted to the species Versicolor (-1) and Virginica $(+1)$. The number of observations is $100(= n)$ and the explanatory variates are sepal length, sepal width, petal length and petal width ($p = 4$). We use RBF kernel with $\sigma = 0.1$ and set $C = 1$.

SVM-guided biplot ($\delta = 0.5$) of the iris data is shown in Figure 1. The vertical axis represents the discrimination between the two species, Versicolor (in blue) and Virginica (in red). Thus, sepal length and sepal width are mostly irrelevant factors in distinguishing species, while petal length and petal width are key variables for species identification. It is likely that iris flowers with relatively large petal measurements are Virginica.

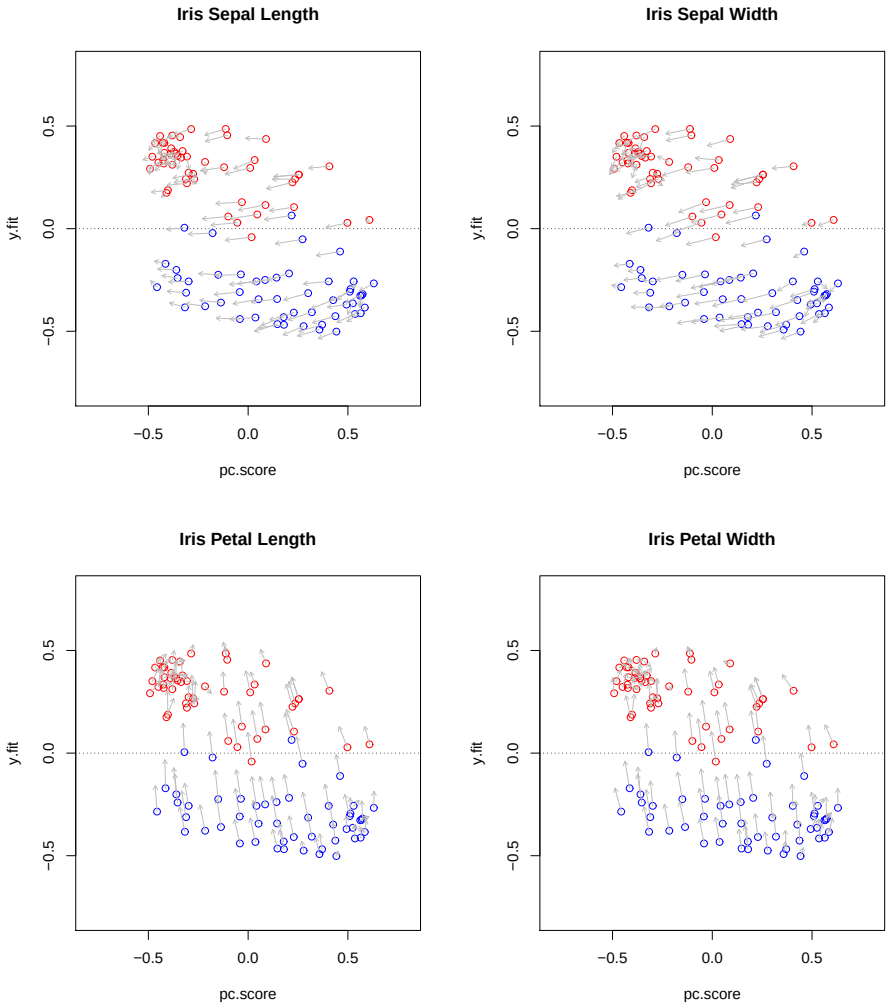

Figure 1: *SVM-guided biplot of the iris data* ($\delta = 0.5$)

Example 2: Spam Data

Spam data of the UCI Machine Learning Repository consists of 4,601 ($= n$) e-mails with 57 ($= p$) morphological characters or frequencies of certain words. Each mail is tagged as either "non-spam" (-1) or "spam" ($+1$). We use RBF kernel with $\sigma = 0.1$ and set $C = 10$, which is selected as the best tuning values by ten-fold cross-validation searched over a reasonable range of the parameters.

There are too many predictors to draw all arrow diagrams (one for each variable); therefore, we select two key variables with long arrows in either the vertical axis or horizontal axis. Formal index for importance may be defined by the total of squared arrow lengths in the direction of either axis.

Figure 2 shows two arrow diagrams ($\delta = 1$) for "free" and "charDollar" (carrying maximal information on vertical axis) and two arrow diagrams ($\delta = 1$) for "you" and "will" (carrying maximal information on horizontal axis). In the diagrams, spam mails are colored in red and non-spam mails are colored in blue.

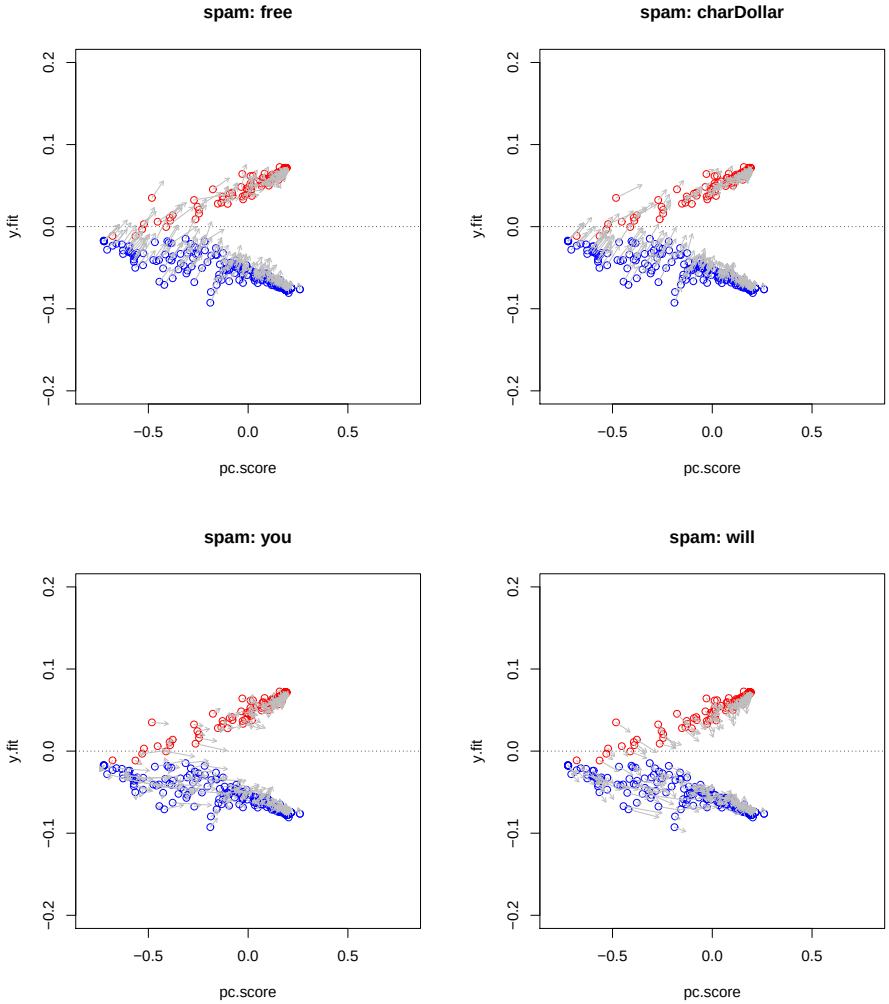

Figure 2: *SVM-guided biplot of the spam data* ($\delta = 1$)

4. SVM Regression and Biplot

Suppose that a target variable Y of numerical type is measured at each of n observations together with p explanatory variables $X_1, \ldots, X_p$. Denoting the explanatory part of the i^{th} observation by $p \times 1$ vector $\mathbf{x}_i$, SVM epsilon regression predicts the response of an arbitrary case with explanatory feature $\mathbf{x}$ to be

$$f_{SVM}(\mathbf{x}) = < \mathbf{w}, \Phi(\mathbf{x}) > + b_0$$

where $\Phi(\mathbf{x})$ denotes the transform of $\mathbf{x}$ in $\mathbb{R}^p$ to a Hilbert space H. As an optimization problem, it can be stated as

$$\min \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \right)$$

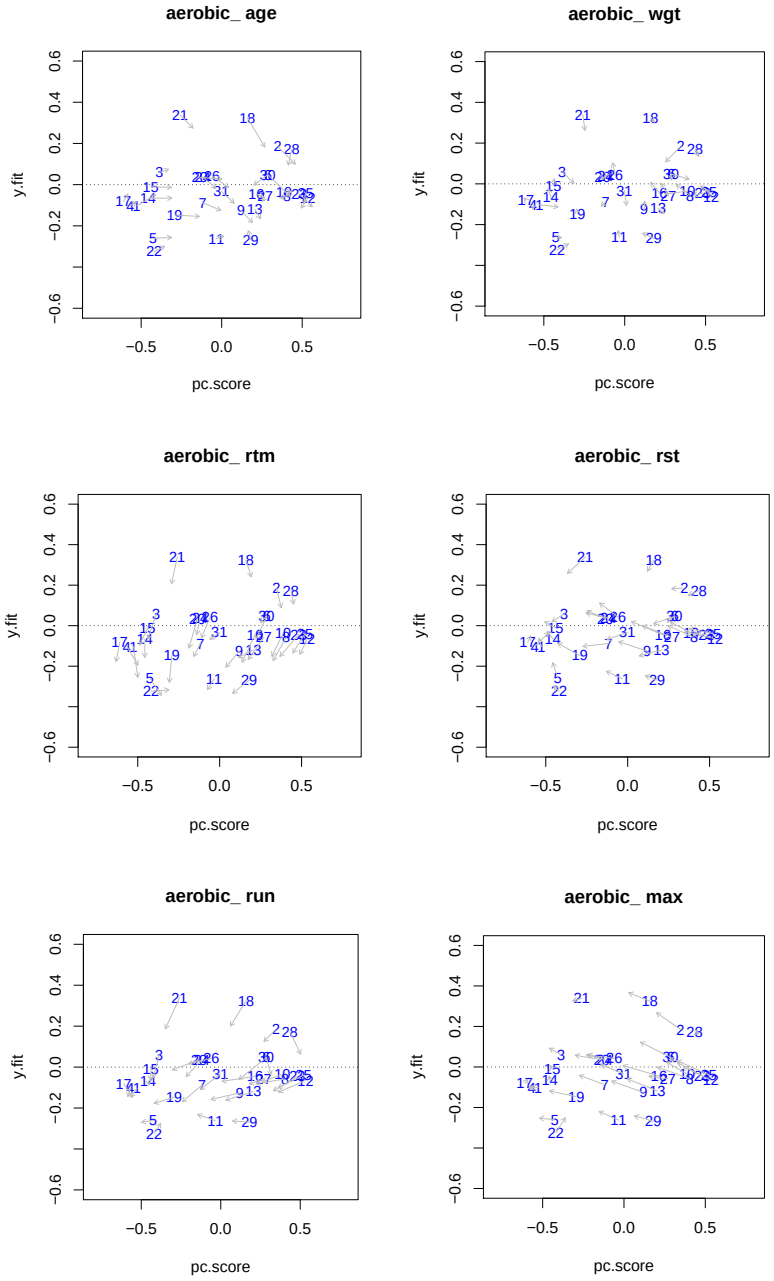

Figure 3: *SVM-guided biplot of the aerobic fitness data ($\delta = 1$)*

subject to

$$y_i - f(\mathbf{x}_i) - \xi_i \le \epsilon, \quad \text{if } y_i - f(\mathbf{x}_i) > \epsilon,$$
$$y_i - f(\mathbf{x}_i) - \xi_i \ge -\epsilon, \quad \text{if } y_i - f(\mathbf{x}_i) < -\epsilon,$$

where $\xi_i\,(i = 1, \ldots, n)$ are non-negative quantities called slack variables. The solution for $\mathbf{w}$ is given by

$$\mathbf{w} = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)\ \Phi(\mathbf{x}_i), \quad \alpha_i \geq 0,\ \alpha_i^* \geq 0.$$

Hence, once the weight vector $\mathbf{w}$ is determined, the regression case is not different from the classification case.

Example 3: Aerobic Fitness Data

In the aerobic fitness data (SAS Inc., 2009) obtained from thirty-one males, the response variable is the oxygen uptake rate (= Y) and the explanatory variables are the age (= $X1$), running time (= $X2$), run pulse (= $X3$), weight (= $X4$), max pulse (= $X5$), and rest pulse (= $X6$).

We use RBF kernel with $\sigma = 0.1$. Setting $C = 10$ and $\epsilon = 0.1$, we obtained Figure 3 ($\delta = 1$). The graphs show that the running time(rtm) is most influential for the determination of Y, the vertical axis, and that three pulses (rst, run, max) are linked to the determination of the principal spread, the horizontal axis.

5. Concluding Remark

Huh and Lee (2013) extended Gabriel (1971)'s "biplot" of observations and variables for the cases of linear and logistic regression. This study further extended the methodology to the cases in which SVM and Kernel PCA are brought in.

For the datasets with large number p of variables, it would be a messy job if one tries to examine all the diagrams for variables. In such situations, we recommend to draw arrow diagrams only for selected variables that have relatively large influence on determination of the axes.

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.

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

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.

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.

SAS Inc. (2009). *SAS/STAT V9.2 User Guide*, Second Edition, Cary, NC.

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

Received October 11, 2013; Revised November 5, 2013; Accepted November 5, 2013