Sensitivity Analysis in Kernel Principal Component Analysis

Yoshihiro Yamanishi[1] and Yutaka Tanaka[2]

[1] Center for Computational Biology, Ecole des Mines de Paris, 35 rue Saint-Honore, 77305 Fontainebleau cedex, France yoshi@cg.ensmp.fr

[2] Department of Mathematical Sciences, Nanzan University, 27 Seto, Aichi 489-0863, Japan ytanaka@ms.nanzan-u.ac.jp

Summary. In this paper we derive empirical influence functions for features in kernel principal component analysis. Based on the derived influence functions, a sensitivity analysis procedure is proposed for detecting influential objects with respect to each feature, subspace spanned by specified eigenvectors, and configuration of the features of interest. We show the usefulness of the proposed procedure with a numerical example.

Key words:
kernel method, kernel PCA, sensitivity analysis, influence function

1 Introduction

In recent years, kernel-based statistical methods have been developed such as support vector machine (SVM), kernel regression analysis (kernel RA), kernel principal component analysis (kernel PCA), and kernel canonical correlation analysis (kernel CCA) [SS02]. However, there is a possibility that these methods are sensitive to a few influential objects. Sensitivity analysis has been well developed in various methods of ordinary multivariate data analysis such as RA, PCA, and CCA. Similarly, it is much anticipated to develop a method of detecting influential objects which have extraordinarily large effects on the results obtained by kernel methods as well, because it is undesirable that the interpretation of the result of analysis depends on a few objects. However, there has been little work on the sensitivity analysis in kernel methods so far.

In this paper, we focus on the kernel PCA and we propose a method of sensitivity analysis in the context of the kernel PCA. The kernel PCA is a useful method to investigate nonlinear structures of the data and remove noise effects. It works well in ordinary circumstances to extract a few major features from the complex data when the kernel similarity matrix is obtained. The algorithm for computing the features reduces to solving an eigenvalue problem of the kernel similarity matrix. However, based on our experiences in multivariate classical PCA, there is a possibility that the kernel PCA is sensitive to a few influential objects.

To investigate the effect of influential objects on the features in the kernel PCA, we derive empirical influence functions (EIFs) for the features by applying a variant of perturbation theory of eigenproblems. In particular, we consider influence statistics based on the EIF for each feature or principal component(PC) and the EIFs for two statistics which characterize the subspace spanned by eigenvectors of interest, and propose a sensitivity analysis method based on the EIF. We shall apply the proposed method to an artificial dataset, and show the usefulness of our approach for detecting influential objects on the result of the kernel PCA.

2 Kernel principal component analysis (KPCA)

2.1 Kernel PCA (KPCA)

Kernel PCA is a method which generalizes classical PCA [SSM98]. Its goal is to extract a feature taking into account nonlinear structures in a dataset $\{\mathbf{x}_i\}_{i=1}^N$, where each object $\mathbf{x}_i$ belongs to some set $\mathcal{X}$. To this end, the objects $\mathbf{x}_i$ are mapped to a high-dimensional space, or a Hilbert space H_x, by a mapping $\phi(.)$. Classical PCA can then be applied to the images $\{\phi(\mathbf{x}_i)\}_{i=1}^N$. The goal is to find a direction $\mathbf{w} \in H_x$ such that the feature $f(\mathbf{x}) =< \mathbf{w}, \phi(\mathbf{x}) >$ has the maximal variance, where $< \cdot, \cdot >$ indicate an inner-product in the Hilbert space. As directions orthogonal to the subspace spanned by $\phi(\mathbf{x}_1), \cdots, \phi(\mathbf{x}_N)$ do not contribute to the variance, it can be restricted that the $\mathbf{w}$ belongs to this space. The $\mathbf{w}$ can therefore be expressed as

$$\mathbf{w} = \sum_{i=1}^N \alpha_i \phi(\mathbf{x}_i). \tag{1}$$

The corresponding feature f can be calculated as

$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i < \phi(\mathbf{x}_i), \phi(\mathbf{x}) >, \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)^T$. We can consider multiple features f_j $(j = 1, 2, \cdots, p)$, choosing the corresponding directions $\mathbf{w}_j$ sequentially under the constraints $||\mathbf{w}_j||^2 = 1$ and $< \mathbf{w}_j, \mathbf{w}_k >= 0$ $(j > k)$.

The use of kernel function $k(\mathbf{x}, \mathbf{x}')$ for computing $< \phi(\mathbf{x}), \phi(\mathbf{x}') >$ enables us to avoid the explicit calculation of $\phi(\mathbf{x})$, which is called *kernel trick* [SS02]. Any kernel function $k(\cdot, \cdot)$ on $\mathcal{X}^2$ defines a Hilbert space and a mapping $\phi(.)$ such that $\forall(\mathbf{x}, \mathbf{x}') \in \mathcal{X}^2, k(\mathbf{x}, \mathbf{x}') =< \phi(\mathbf{x}), \phi(\mathbf{x}') >$. Examples of the kernel functions are linear kernel:

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}', \tag{3}$$

and Gaussian kernel with width σ:

$$k(\mathbf{x}, \mathbf{x}') = \exp\{- \parallel \mathbf{x} - \mathbf{x}' \parallel^2 /2\sigma^2\}. \tag{4}$$

The performance of the KPCA depends on the choice of the kernel functions. The cross-validation is useful for choosing the kernel function and its optimal parameters. When linear kernel is chosen as the kernel function, the result of the KPCA is equivalent to that of classical PCA [SSM98].

2.2 Algorithm of KPCA

Let now K, $(K)_{ij} := k(\mathbf{x}_i, \mathbf{x}_j)$, be a kernel matrix. Suppose that the kernel matrix K is centered in advance as follows:

$$K \quad \leftarrow \quad Q_N K Q_N, \tag{5}$$

where $Q_N = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$, I_N is an identity matrix and $\mathbf{1}_N = (1, 1, \cdots, 1)^T$. Then, the kernel matrix K is decomposed as

$$K = \sum_{a=1}^{N} \rho_a \mathbf{u}_a \mathbf{u}_a^T = UDU^T, \tag{6}$$

where $U = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_N]$ and $D = diag(\rho_1, \rho_2, \cdots, \rho_N)$. Since $||\mathbf{w}||^2 = \boldsymbol{\alpha}^T K \boldsymbol{\alpha} = 1$, $\boldsymbol{\alpha} = \mathbf{u}/\sqrt{\rho}$, the $j-$th feature $\mathbf{f}_j \in R^N$ is computed as

$$\mathbf{f}_j = K\boldsymbol{\alpha}_j = \sqrt{\rho_j}\mathbf{u}_j. \tag{7}$$

3 Sensitivity analysis in kernel PCA

3.1 Influence function

Let us introduce case-weights $Nw_i/\sum_\beta w_\beta$ for the objects as perturbation parameters, and define the first order partial derivative of estimated parameter vector or function $\hat{\theta}$ with respect to w_i as the empirical influence function (EIF) of $\hat{\theta}$ for the i-th object. Or in other words, we define the EIF for the i-th object by the first derivative with respect to ϵ after introducing the following case-weight perturbation

$$w_\beta = 1 \,\text{for all}\, \beta \;\longrightarrow\; w_\beta = N\tilde{w}_\beta / \sum_{\beta=1}^{N} \tilde{w}_\beta, \,\text{where}\, \tilde{w}_\beta = \begin{cases} 1 & (\beta \neq i) \\ 1+\epsilon & (\beta = i) \end{cases} \tag{8}$$

to the objects. Then it is easily verified that the first derivative of $\hat{\theta}$ with respect to ϵ is equal to a constant times the empirical influence curve [Ham74, Tan94].

3.2 Perturbation to the eigenproblem of KPCA

Consider a case weight perturbation to an object. The feature $\mathbf{f}$ is assumed to be expanded in a convergent power series in the neighborhood of $\epsilon = 0$ as

$$\mathbf{f}(\epsilon) = \mathbf{f} + \epsilon \mathbf{f}^{(1)} + O(\epsilon^2). \tag{9}$$

We want to evaluate the influence of each object on the feature $\mathbf{f}$ by evaluating the derivative $\mathbf{f}^{(1)}$. However, it is not possible to compute such derivative directly. Similarly, the eigenvalue ρ and eigenvector $\mathbf{u}$ of the kernel matrix K are also assumed to be expanded as

$$\rho(\epsilon) = \rho + \epsilon \rho^{(1)} + O(\epsilon^2), \quad \mathbf{u}(\epsilon) = \mathbf{u} + \epsilon \mathbf{u}^{(1)} + O(\epsilon^2). \tag{10}$$

However, we can not directly evaluate the derivatives $\rho^{(1)}$ and $\mathbf{u}^{(1)}$ as well, because it is impossible to introduce the case-weight perturbation to an $N \times N$ kernel matrix K. We therefore propose the following procedure.

First, the $N \times N$ kernel matrix K is decomposed or approximated by a set of eigenvectors associated with the largest p eigenvalues as

$$K = UDU^T = UD^{1/2}D^{1/2}U^T = BB^T, \tag{11}$$

where $D = diag(\rho_1, \cdots, \rho_p)$, $U = [\mathbf{u}_1, \cdots, \mathbf{u}_p]$, $B = UD^{1/2}$, and $p < N$ (for example, p is a rank of K). We can consider that the configuration of N rows of B provides a p-dimensional approximation to the configuration of the N objects in the feature space. Let us define a $p \times p$ matrix L as

$$L = B^T B = VDV^T, \tag{12}$$

where $D = diag(\rho_1, \cdots, \rho_p)$, $V = [\mathbf{v}_1, \cdots, \mathbf{v}_p]$, ρ and $\mathbf{v}$ are the eigenvalue and eigenvector of matrix L, respectively. Note that the eigenvalues are the same across matrices K and L. The j-th eigenvector $\mathbf{u}_j$ of $K = BB^T$ has relationships with the j-th eigenvector $\mathbf{v}_j$ of $L = B^T B$ as follows:

$$\mathbf{u}_j = \rho_j^{-1/2} B \mathbf{v}_j \quad \text{and} \quad \mathbf{v}_j = \rho_j^{-1/2} B^T \mathbf{u}_j, \quad j = 1, 2, \cdots, p. \tag{13}$$

Here we propose to introduce perturbations to the weights for N rows of the matrix B, and evaluate the perturbed ρ, $\mathbf{v}$ and $\mathbf{u}$, and then the perturbed $\mathbf{f}$.

3.3 Perturbation to the eigenproblem of L

Let $\mathbf{b}$ be a column of the matrix B^T, and $\bar{\mathbf{b}}$ be the mean vector of $\{\mathbf{b}_i\}_{i=1}^N$, where $B^T = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_N]$. Let us define the covariance matrix C as $C = \frac{1}{N}\sum_{i=1}^N (\mathbf{b}_i - \bar{\mathbf{b}})(\mathbf{b}_i - \bar{\mathbf{b}})^T$. Then, the perturbed covariance matrix C is expanded as

$$C(\epsilon) = C + \epsilon C^{(1)} + (\epsilon^2/2) C^{(2)}, \tag{14}$$

where $C^{(1)}$ and $C^{(2)}$ are given as

$$C^{(1)} = (\mathbf{b}-\bar{\mathbf{b}})(\mathbf{b}-\bar{\mathbf{b}})^T - C, \quad C^{(2)} = -2(\mathbf{b}-\bar{\mathbf{b}})(\mathbf{b}-\bar{\mathbf{b}})^T \qquad (15)$$

(see, e.g., [Cri85]). From the perturbation theory of eigenproblems, its eigenvalues and eigenvectors can be also expanded in a convergent power series in the neighborhood of $\epsilon = 0$ as

$$\rho_s(\epsilon) = \rho_s + \epsilon\rho_s^{(1)} + O(\epsilon^2), \quad \mathbf{v}_s(\epsilon) = \mathbf{v}_s + \epsilon\mathbf{v}_s^{(1)} + O(\epsilon^2), \qquad (16)$$

and we have the following formulas (see, e.g. [Sib79]):

$$\rho_s^{(1)} = a_{ss}^{(1)}, \quad \mathbf{v}_s^{(1)} = \sum_{r \neq s}(\rho_s - \rho_r)^{-1} a_{rs}^{(1)} \mathbf{v}_r. \qquad (17)$$

where $a_{rs}^{(1)} = \mathbf{v}_r^T C^{(1)} \mathbf{v}_s$.

Then, the EIFs for the j-th feature $\mathbf{f}_j$ in KPCA can be derived as

$$EIF(\mathbf{x}; \mathbf{f}_j) = B\mathbf{v}_j^{(1)}. \qquad (18)$$

Note that the expansion of eigenvectors v_s cannot be used in the case where there exist other eigenvalues which are exactly equal to or very close to ρ_s.

3.4 Influence functions related to the subspace spanned by specified eigenvectors which characterize features of interest

Suppose that we have q features $\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_q$ of interest, selected from p features ($q < p$). Here we propose to evaluate the influence on the subspace spanned by the q eigenvectors which characterize the features of interest by using the following two statistics:

$$P = V_q(V_q^T V_q)^{-1} V_q^T = V_q V_q^T, \qquad (19)$$

$$T = V_q D_q V_q^T, \qquad (20)$$

where $D_q = diag(\rho_1, \cdots, \rho_q)$, $U_q = diag(\mathbf{u}_1, \cdots, \mathbf{u}_q)$, and $V_q = [\mathbf{v}_1, \cdots, \mathbf{v}_q]$. The P indicates the orthogonal projector onto the subspace spanned by V_q, and T indicates the dominant part of eigen decomposition of $L = VDV^T$.

Then, the perturbed matrices P and T are expanded as

$$P(\epsilon) = V_q V_q^T + \epsilon(V_q V_q^T)^{(1)} + O(\epsilon^2), \qquad (21)$$

$$T(\epsilon) = V_q D_q V_q^T + \epsilon(V_q D_q V_q^T)^{(1)} + O(\epsilon^2), \qquad (22)$$

where $(V_q V_q^T)^{(1)}$ and $(V_q D_q V_q^T)^{(1)}$ can be computed, respectively, by using the following formulas in [Tan88]:

$$(V_qV_q^T)^{(1)} = \sum_{s=1}^{q}\sum_{r=q+1}^{p}(\rho_s-\rho_r)^{-1}a_{rs}^{(1)}(\mathbf{v}_s\mathbf{v}_r^T+\mathbf{v}_r\mathbf{v}_s^T), \tag{23}$$

$$\begin{aligned}(V_qD_qV_q^T)^{(1)} &= \textstyle\sum_{s=1}^{q}\sum_{r=1}^{q}(\mathbf{v}_s^TC^{(1)}\mathbf{v}_r)\mathbf{v}_s\mathbf{v}_r^T \\ &+\textstyle\sum_{s=1}^{q}\sum_{r=q+1}^{p}\rho_s(\rho_s-\rho_r)^{-1}(\mathbf{v}_s^TC^{(1)}\mathbf{v}_r)(\mathbf{v}_s\mathbf{v}_r^T+\mathbf{v}_r\mathbf{v}_s^T).\end{aligned} \tag{24}$$

Note that these formulas work well even if there exist multiple eigenvalues or eigenvalues which are very close with each other among those of interest. Therefore, we can define the EIFs for P and T as follows, respectively:

$$EIF(\mathbf{x};P) = P^{(1)} = (V_qV_q^T)^{(1)}, \tag{25}$$

$$EIF(\mathbf{x};T) = T^{(1)} = (V_qD_qV_q^T)^{(1)}. \tag{26}$$

3.5 Influence functions for the configuration of N objects

We consider the influence on the configuration of N objects. Let us define an $N \times q$ feature matrix F_q as $F_q = [\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_q]$ by selecting q features from p features, which is obtained by $F_q = BV_q$ in the q-dimensional feature space. Here we propose to use two statistics

$$\parallel F_qF_q^T \parallel, \quad \parallel \frac{F_qF_q^T}{\{tr(F_qF_q^T)^2\}^{1/2}} \parallel, \tag{27}$$

which characterize the configuration of N objects. As discussed in [RE76], the former statistic is invariant for translation and rotation, while the latter is invariant for translation, rotation and scale change.

When we introduce the perturbation, the first statistic can be expanded as

$$\parallel F_qF_q^T - F_q(\epsilon)F_q^T(\epsilon) \parallel = |\epsilon|M_1 + o(\epsilon), \tag{28}$$

where $M_1 = (tr\{(V_qV_q^T)^{(1)}B^TB\}^2)^{1/2}$, and the second statistic can be expanded as

$$\parallel \frac{F_qF_q^T}{\{tr(F_qF_q^T)^2\}^{1/2}} - \frac{F_q(\epsilon)F_q(\epsilon)^T}{\{tr(F_q(\epsilon)F_q(\epsilon)^T)^2\}^{1/2}} \parallel = |\epsilon|M_2 + o(\epsilon), \tag{29}$$

where $M_2 = (R^{(1)})^{1/2}$ with $R^{(1)}$ given as

$$R^{(1)} = \frac{tr\{(V_qV_q^T)^{(1)}B^TB\}^2}{tr(V_qV_q^TB^TB)^2} - \left[\frac{tr\{V_qV_q^TB^TB(V_qV_q^T)^{(1)}B^TB\}}{tr(V_qV_q^TB^TB)^2}\right]^2 \geq 0. \tag{30}$$

It can be verified that $R^{(1)}$ is closely related to Escoufier's RV coefficient as

$$RV(F_q, F_q(\epsilon)) = 1 - (\epsilon^2/2)R^{(1)} + o(\epsilon^2). \tag{31}$$

The coefficients of $|\epsilon|$, M_1 and M_2, can be used as influence measures. Note that both coefficients M_1 and M_2 are functions of $P^{(1)}$.

4 Numerical example

We applied the KPCA to the toy data used in [SSM98] with two kernel functions: linear kernel and Gaussian kernel with width $\sigma = 0.5$. Figure 1 shows the scatter-plot of the features (PC2 scores versus PC1 scores) for linear kernel and Gaussian kernel, respectively. It seems that the clusters can be more clearly detected by the nonlinear effect of Gaussian kernel.

Next, we applied the proposed sensitivity analysis to the result of the KPCA. We computed the EIF for the first feature (PC1 scores), for example. Figure 2 shows the index-plot of the norms of the corresponding EIFs in using linear kernel and Gaussian kernel, respectively. The objects with large effect on the PC1 direction seem to be detected as influential objects in both cases.

Finally, we studied the influence on the subspace spanned by a few eigenvectors which characterize the features of interest, and that on the configuration of N objects. We computed the EIFs for two statistics P and T with $q = 2$ and $p = 4$ in applying KPCA with Gaussian kernel. Figure 3 shows the index-plots of the Frobenius norms of the corresponding EIFs for the P and T, respectively. We computed the proposed influence measures M_1 and M_2 with $q = 2$ and $p = 4$ in applying KPCA with Gaussian kernel. Figure 4 shows the index-plots for the coefficients M_1 and M_2, respectively. The objects supporting the configuration seem to be detected as influential objects in both cases. There is a possibility that detected influential objects depend on the choice of the parameters in the kernel functions.

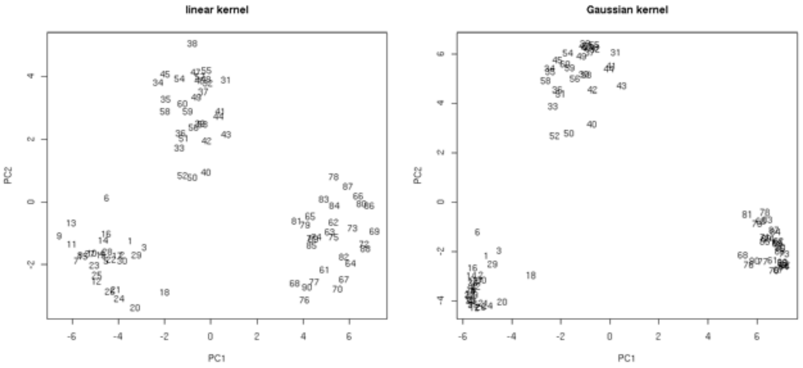

Fig. 1. Scatter-plot of features (PC2 versus PC1) in the KPCA with linear kernel (left) and Gaussian kernel (right)

5 Concluding remarks

We proposed a method of sensitivity analysis in KPCA. The key idea was to introduce case-weight perturbations to the objects in the p-dimensional

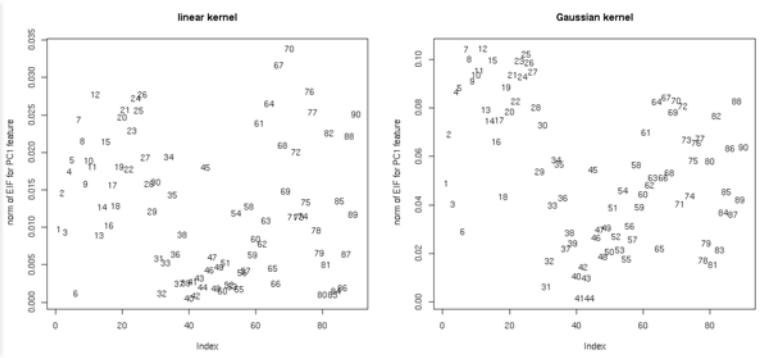

Fig. 2. Index-plot of the norms of the EIFs for PC1 feature

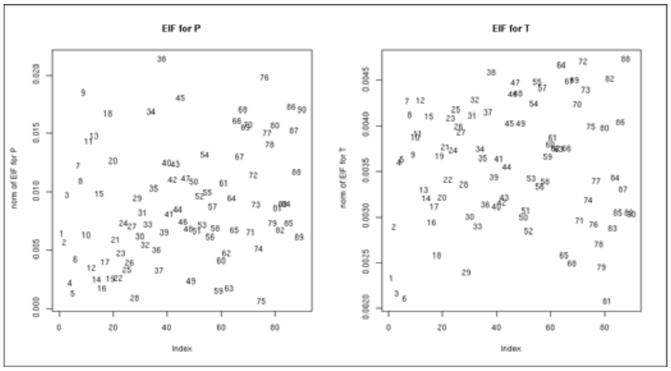

Fig. 3. Index-plot of the norms of the EIFs for P (left) and T (right)

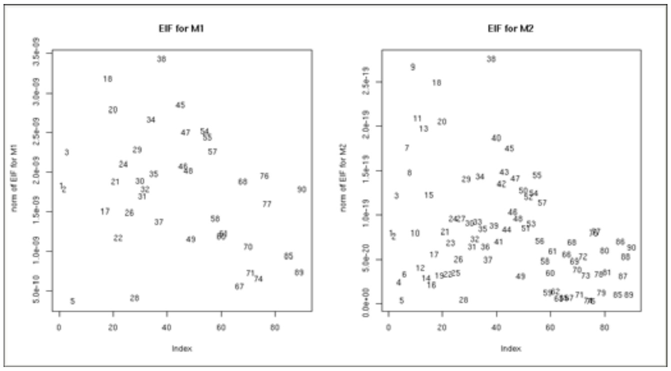

Fig. 4. Index-plot of influence measures M_1 (left) and M_2 (right)

Euclidean space derived as the eigenvalue-based approximation to the Hilbert space and then evaluate the amount of influence with influence functions. In a numerical example it is shown that the proposed procedure is useful for detecting influential objects. As influence measures we used influence functions P and T related to the subspace spanned by specified eigenvectors which characterize features of interest and coefficients M_1 and M_2 which reflect the changes of the configuration of N objects. The reason why we consider subspace and configuration in addition to individual features or individual PCs is that usually in KPCA subspaces or configurations are more important than individual PCs. For vector or matrix of influence functions we used the norm with identity metric in the present paper, while the inverse of asymptotic covariance matrix is often used as the metric to compute the Cook's D type measures in the ordinary PCA. We may define in KPCA similar measures based on internal or external estimates for the covariance matrix. Studies on metrics and on the effects of the goodness of eigenvalue-based approximation will be topics of our future study.

Acknowledgments

This research was partly supported by Japan Society of Promotion of Science (Grant-in-Aid for Scientific Research (C)(2)16500175.

References

[Cri85] Critchley, F.: Influence in principal component analysis. *Biometrika,* ***72****, 627–636.* (1985).

[Ham74] Hampel, R.R.: The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.,* **69***, 383–393.* (1974).

[RE76] Robert, P. and Escoufier, Y.: A unified tool for linear multivariate statistical methods: RV coefficient, *Applied Statistics*, 25, 257–265, (1976).

[SSM98] Schölkopf, B., Smola, A.J., and Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, *Neural Computation*, 10, 1299–1319, (1998).

[SS02] Schölkopf, B., and Smola, A.J.: Learning with Kernels, *MIT Press*, (2002).

[Sib79] Sibson, R.: Studies in the robustness of multidimensional scaling: Perturbation analysis of classical scaling, *J. R. Statist. Soc.* **B 41***, 217–229.* (1979).

[Tan88] Tanaka,Y.: Sensitivity analysis in principal component analysis: influence on the subspace spanned by principal components, *Communications in Statistics, Theory and Methods,* **17***, 3157-3175.* (1988).

[Tan89] Tanaka,Y.: Influence functions related to eigenvalue problems which appear multivariate analysis, *Communications in Statistics, Theory and Methods,* **18***, 3991-4010.* (1989).

[Tan94] Tanaka, Y.: Recent advance in sensitivity anaysis in multivariate statistical methods, *J.Japanese Soc.Comp.Statist.,7,1–25.* (1994).