Sensitivity Analysis in Functional Principal Component Analysis

Yoshihiro Yamanishi[1] and Yutaka Tanaka[2]

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[2] Department of Environmental and Mathematical Sciences, Okayama University 3-1-1 Tsushima-Naka, Okayama, Okayama 700-8530, Japan

Summary

In the present paper empirical influence functions (EIFs) are derived for eigenvalues and eigenfunctions in functional principal component analysis in both cases where the smoothing parameter is fixed and unfixed. Based on the derived influence functions a sensitivity analysis procedure is proposed for detecting jointly as well as singly influential observations. A numerical example is given to show the usefulness of the proposed procedure. In dealing with the influence on the eigenfunctions two different kinds of influence statistics are introduced. One is based on the EIF for the coefficient vectors of the basis function expansion, and the other is based on the sampled vectors of the functional EIF. Under a certain condition it can be proved both kinds of statistics provide essentially equivalent results.

Keywords: Functional data, Principal component analysis, Statistical diagnostics, Influence function

1 Introduction

Several methods have been developed so far to analyze functional data (Besse and Ramsay, 1986; Ramsay and Dalzell, 1991; Ramsay and Silverman, 1997). Functional principal component analysis (PCA) enables us to investigate the pattern of the data over time, or over another argument. It works well in ordinary circumstances to extract a few major features from the complex functional data. However, there is a problem that the method is sensitive to a few influential observations as in the case of multivariate PCA. Sensitivity analysis has been well developed in various methods of multivariate data analysis such as regression analysis, principal component analysis, and canonical correlation analysis. Similarly, it is much anticipated to develop a method of detecting influential observations which have extraordinarily large effects on the results of functional data analysis as well, because it is undesirable that the interpretation of the result of analysis depends on a few observations. However, there has been little work in the field of the sensitivity analysis in functional data analysis so far.

In this paper, we propose a method of sensitivity analysis to investigate the effect of influential observations on the eigenvalues and eigenfunctions of functional PCA. We consider Cook's D type statistics to evaluate the influence of each observation. It is very important to apply not only single-case but also multiple-case diagnostics in order to understand the joint patterns of the influence on the parameters. So both single-case and multiple-case diagnostics in this context should be developed and some problems have to be pointed out in actual data analysis. In this study, PCA with an appropriate metric is used to carry out the multiple-case diagnostics by making use of the idea of general procedure by Tanaka (1994) in the case of multivariate methods. In particular, we consider two kinds of influence statistics to deal with the influence on the eigenfunctions. One is based on the EIF for the coefficient vectors of the basis function expansion, and the other is based on the sampled vectors of the functional EIF.

The remainder of this paper is organized as follows. In section 2, we make a brief review of functional PCA. In section 3, we derive influence functions for eigenvalues and eigenfunctions in functional PCA using perturbation theory of eigenproblem. In sections 4 and 5 we describe the procedures for single-case and multiple-case diagnostics, respectively. In section 6, we give some discussions, mainly on the relationship between the two kinds of influence to deal with the influence on the eigenfunctions. It was proved mathematically and numerically that two kinds of statistics provide essentially equivalent results. Finally we give some concluding remarks in section 7.

2 Functional principal component analysis

2.1 Ordinary functional principal component analysis

Suppose we have a set of functional data $\{x_i(s)\}_{i=1}^N$. Weight function $\xi(s)$ is chosen in such a way that it maximizes the variance

$$PCASV = \int \int \xi(s) v(s,t) \xi(t) ds dt, \tag{1}$$

where $v(s,t)$ indicates a variance-covariance function defined by $v(s,t) = N^{-1} \sum_{i=1}^N \{x_i(s) - \bar{x}(s)\}\{x_i(t) - \bar{x}(t)\}$ where $\bar{x}(t) = N^{-1} \sum_{i=1}^N x_i(t)$. The maximization of $PCASV$ under the constraints

$$\int \xi_l(t)^2 dt = 1 \ , \ \int \xi_l(t) \xi_m(t) dt = 0 \, (l < m) \tag{2}$$

leads to an integral eigenproblem as follows:

$$\int v(s,t) \xi(t) dt = \rho \xi(s). \tag{3}$$

2.2 Penalized functional principal component analysis

2.2.1 Penalized functional principal component analysis

Here a penalty function is introduced to incorporate smoothing into the principal components (PCs). The most popular form of the penalty for ξ is given by the integral squared second derivative of ξ , i.e., $PEN_2(\xi) =< D^2\xi, D^2\xi >=\| D^2\xi \|^2$, where $< \cdot, \cdot >$ and $\| \cdot \|^2$ indicate the inner product and the squared norm, respectively. In this case the penalized variance can be expressed by

$$PCAPSV = \frac{PCASV}{\| \xi \|^2 + \lambda \times PEN_2(\xi)}, \tag{4}$$

where λ is a smoothing parameter. This expression means that the trade-off between maximizing the sample variance and keeping the smoothness of ξ is controlled by a smoothing parameter λ. The solution ξ is obtained as the eigenfunction associated with the largest eigenvalue of the penalized eigenproblem expressed by

$$\int v(s,t) \xi(t) dt = \rho (I + \lambda D^4) \xi(s), \tag{5}$$

using the relation $\| D^2\xi \|^2 =< \xi, D^4\xi >$, derived from natural or periodic boundary conditions. For details refer to Ramsay and Silverman (1997).

2.2.2 Choice of the smoothing parameter λ by cross-validation

Consider how much principal components (PCs) of interest explain the original data. Let M be the number of PCs of interest, and define the cross-validation score as

$$CV(\lambda) = \sum_{i=1}^{N} \| \, x_i(s) - \sum_{m=1}^{M} \sum_{l=1}^{M} [\mathbf{G}^{-1}]_{ml}^{[-i]} \left(\int \xi_m^{[-i]}(t) x_i(t) dt \right) \xi_l^{[-i]}(s) \, \|^2, \quad (6)$$

where $\mathbf{G}$ is the $M \times M$ matrix whose (m, l) element is the inner product $\int \xi_m(t)\xi_l(t)dt$ and the superscript $[-i]$ means the omission of the $i-$th observation. Choose λ which minimizes $CV(\lambda)$.

2.2.3 Algorithm

Suppose that a data function $x_i(s)$ and a weight function $\xi(s)$ can be expanded as

$$x_i(s) = \sum_{k=1}^{K} c_{ik}\phi_k(s) = \mathbf{c}_i^T \phi(s), \quad \xi(s) = \sum_{k=1}^{K} y_k \phi_k(s) = \mathbf{y}^T \phi(s), \quad (7)$$

by using basis functions $\phi(s) = (\phi_1(s), \cdots, \phi_K(s))^T$, where K is the number of basis functions. Define $\mathbf{V}$ as the covariance matrix of coefficient $\mathbf{c}_i$ and let $\mathbf{J}_\phi = \int \phi(s)\phi(s)^T ds$ and $\mathbf{K}_\phi = \int (D^2\phi(s))(D^2\phi(s))^T ds$. Then the functional eigenproblem (5) is transformed to the following matrix generalized eigenproblem

$$(\mathbf{J}_\phi \mathbf{V} \mathbf{J}_\phi)\mathbf{y} = \rho(\mathbf{J}_\phi + \lambda \mathbf{K}_\phi)\mathbf{y}. \quad (8)$$

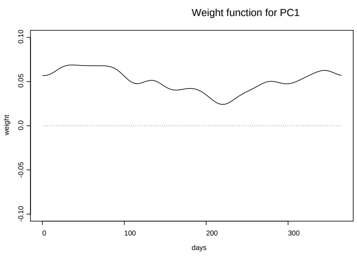
By applying Cholesky factorization $\mathbf{L}\mathbf{L}^T = \mathbf{J}_\phi + \lambda \mathbf{K}_\phi$, the above generalized eigenproblem leads to an eigenproblem of a symmetrical matrix as

$$(\mathbf{S}\mathbf{J}_\phi \mathbf{V} \mathbf{J}_\phi \mathbf{S}^T)(\mathbf{S}^{-T}\mathbf{y}) = \rho(\mathbf{S}^{-T}\mathbf{y}), \quad (9)$$

where $\mathbf{S} = \mathbf{L}^{-1}$ and $\mathbf{S}^{-T} = (\mathbf{S}^{-1})^T$.

2.2.4 Numerical example

We shall apply the penalized functional PCA to the mean daily temperature data of 50 weather stations in Japan (Japan Meteorological Agency, 1999). In this study we used Fourier series as the basis functions and fixed the number of basis functions as 20. Figure 1 and Figure 2 shows the weight functions for PC1 and PC2 using the optimum λ value determined by cross-validation, where $\lambda = 170$. Looking at these weight functions, we can interpret the PCs as follows: The PC1 is a measure of overall temperatures throughout the year, while the PC2 represents the contrast between the temperatures in summer and in winter.



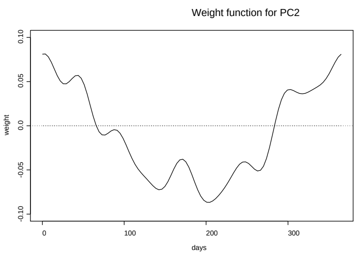


Figure 1: Weight functions for PC1.

Figure 2: Weight functions for PC2.

3 Sensitivity analysis in functional PCA

3.1 Influence function

To evaluate the influence of each individual we make use of the idea of influence function, in particular, empirical influence function (EIF). Here we define the EIF by the partial derivative of the estimated parameter function or vector with respect to a perturbation parameter. Let us introduce weights $Nw_\alpha / \sum_\beta w_\beta$ for the α-th observation as perturbation parameters, and define the first order partial derivative of parameter vector or function $\hat{\theta}$ with respect to w_α as the influence function of $\hat{\theta}$ for the α-th observation. Or in other words, we define the influence function for the α-th observation by the first derivative with respect to ϵ after introducing the following case-weight perturbation

$$w_\alpha = 1 \,\text{for all}\, \alpha \longrightarrow w_\alpha = N\tilde{w}_\alpha / \sum_\beta \tilde{w}_\beta, \,\text{where}\, \tilde{w}_\beta = \begin{cases} 1 & (\beta \neq \alpha) \\ 1+\epsilon & (\beta = \alpha) \end{cases} \tag{10}$$

to the α-th observation. Also, as in the case of multivariate data analysis we define the sample influence function (SIF) by

$$SIF(x_i; \hat{\theta}) = -(N-1)(\hat{\theta}^{[-i]} - \hat{\theta}). \tag{11}$$

3.2 Perturbation of eigenproblem

Recall that penalized functional eigenproblem (5) is eventually transformed to the corresponding eigenproblem (9) by applying basis function expansion algorithm. Consider a case weight perturbation to an individual functional observation. It is obvious that when matrix $\mathbf{SJVJS}^T$ in equation (9) can be expressed as a convergent power series of ϵ. 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

$$\rho_s(\epsilon) = \rho_s + \epsilon\rho_s^{(1)} + (\epsilon^2/2)\rho_s^{(2)} + O(\epsilon^3), \tag{12}$$

$$\mathbf{u}_s(\epsilon) = \mathbf{u}_s + \epsilon\mathbf{u}_s^{(1)} + (\epsilon^2/2)\mathbf{u}_s^{(2)} + O(\epsilon^3), \tag{13}$$

where $\mathbf{u}_s$ is defined as $\mathbf{u}_s = (\mathbf{S}^{-T}\mathbf{y})_s$, and we have the following formulas (see, e.g. Sibson, 1979, Tanaka, 1984):

$$\begin{cases} \rho_s^{(1)} = a_{ss}^{(1)}, \\ \mathbf{u}_s^{(1)} = \sum_{r\neq s}(\rho_s - \rho_r)^{-1} a_{rs}^{(1)}\mathbf{u}_r. \\ \rho_s^{(2)} = a_{ss}^{(2)} + 2\sum_{r\neq s}(\rho_s - \rho_r)^{-1}(a_{rs}^{(1)})^2, \\ \mathbf{u}_s^{(2)} = \sum_{r\neq s}(\rho_s - \rho_r)^{-1}\left\{a_{rs}^{(2)} + 2\sum_{t\neq s}(\rho_s - \rho_t)^{-1} a_{st}^{(1)}(a_{rt}^{(1)} - a_{ss}^{(1)}\delta_{rt})\right\}\mathbf{u}_r \\ \qquad - \parallel \mathbf{u}_s^{(1)} \parallel^2 \mathbf{u}_s. \end{cases} \tag{14}$$

where $a_{rs}^{(k)} = (\mathbf{S}^{-T}\mathbf{y})_r^T \left(\mathbf{SJVJS}^T\right)^{(k)} (\mathbf{S}^{-T}\mathbf{y})_s$ $(k = 1, 2,)$, superscript (k) indicates the k-th derivative in ϵ evaluated at $\epsilon = 0$ and δ_{rt} is the Kronecker delta. When the number of the basis functions is fixed, the influence functions for eigenvalues and eigenfunctions can be expressed as follows:

$$EIF(x; \rho_s) = \rho_s^{(1)}, \tag{15}$$

$$EIF(x; \xi_s) = \xi_s^{(1)} = (\mathbf{y}_s^{(1)})^T\phi. \tag{16}$$

3.3 Influence functions with fixed λ

Let us introduce the perturbation given by (10). Then the perturbed matrix $\mathbf{SJVJS}^T$ in the matrix eigenproblem (9) can be expanded as

$$\left(\mathbf{SJVJS}^T\right)(\epsilon) = \mathbf{SJVJS}^T + \epsilon\left(\mathbf{SJVJS}^T\right)^{(1)} + (\epsilon^2/2)\left(\mathbf{SJVJS}^T\right)^{(2)} + O(\epsilon^3). \tag{17}$$

When the smoothing parameter λ is fixed, $\left(\mathbf{SJVJS}^T\right)^{(k)}$ is given by

$$\left(\mathbf{SJVJS}^T\right)^{(k)} = \mathbf{SJV}^{(k)}\mathbf{JS}^T \quad (k = 1, 2), \tag{18}$$

and it is known that the perturbed covariance matrix $\mathbf{V}$ is expanded as

$$\mathbf{V}(\epsilon) = \mathbf{V} + \epsilon \mathbf{V}^{(1)} + (\epsilon^2/2)\mathbf{V}^{(2)} + O(\epsilon^3). \tag{19}$$

Let $\mathbf{c}$ be a vector of coefficients of the data function, and $\bar{\mathbf{c}}$ be the mean vector of $\{\mathbf{c}_i\}_{i=1}^N$. The $\mathbf{V}^{(1)}$ and $\mathbf{V}^{(2)}$ can be computed as

$$\mathbf{V}^{(1)} = (\mathbf{c}-\bar{\mathbf{c}})(\mathbf{c}-\bar{\mathbf{c}})^T - \mathbf{V}, \quad \mathbf{V}^{(2)} = -2(\mathbf{c}-\bar{\mathbf{c}})(\mathbf{c}-\bar{\mathbf{c}})^T \tag{20}$$

(see, e.g., Critchley, 1985). Therefore, we can define the EIFs for eigenvalues and eigenfucntions respectively as follows:

$$EIF(x;\rho_s) = \rho_s^{(1)} = (\mathbf{S}^{-T}\mathbf{y})_r^T \left(\mathbf{SJVJS}^T\right)^{(1)} (\mathbf{S}^{-T}\mathbf{y})_s, \tag{21}$$

$$EIF(x;\xi_s(t)) = (\mathbf{y}_s^{(1)})^T \phi(t) = \left[\mathbf{S}^T(\mathbf{S}^{-T}\mathbf{y})_s^{(1)}\right]^T \phi(t). \tag{22}$$

3.4 Influence functions with variable λ

If we take into consideration the possibility that the smoothing parameter λ, which is decided by cross-validation, varies due to a small perturbation to the case weights, the perturbed matrix $\mathbf{SJVJS}^T$ is given by

$$\begin{aligned}
\left(\mathbf{SJVJS}^T\right)^{(1)} &= \mathbf{S}^{(1)}\mathbf{JVJS}^T + \mathbf{SJV}^{(1)}\mathbf{JS}^T + \mathbf{SJVJ}(\mathbf{S}^{(1)})^T, \\
\left(\mathbf{SJVJS}^T\right)^{(2)} &= \mathbf{SJV}^{(2)}\mathbf{JS}^T + \mathbf{S}^{(2)}\mathbf{JVJS}^T + \mathbf{SJVJ}(\mathbf{S}^{(2)})^T \\
&\quad 2\left(\mathbf{S}^{(1)}\mathbf{JV}^{(1)}\mathbf{JS}^T + \mathbf{SJV}^{(1)}\mathbf{JS}^{(1)} + \mathbf{S}^{(1)}\mathbf{JVJ}(\mathbf{S}^{(1)})^T\right),
\end{aligned} \tag{23}$$

where $\mathbf{S}^{(1)} = \frac{\partial\lambda}{\partial\epsilon}\left(\frac{\partial\mathbf{S}}{\partial\lambda}\right)$ and $\mathbf{S}^{(2)} = \frac{\partial\lambda}{\partial\epsilon}\left(\frac{\partial\mathbf{S}^{(1)}}{\partial\lambda}\right)$. The derivatives $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ can be calculated as $\mathbf{S}^{(1)} = (\mathbf{L}^{-1})^{(1)} = -\mathbf{L}^{-1}\mathbf{L}^{(1)}\mathbf{L}^{-1}$, $\mathbf{S}^{(2)} = (\mathbf{L}^{-1})^{(2)} = -\mathbf{L}^{-1}\mathbf{L}^{(2)}\mathbf{L}^{-1} + 2\mathbf{L}^{-1}\mathbf{L}^{(1)}\mathbf{L}^{-1}\mathbf{L}^{(1)}\mathbf{L}^{-1}$, respectively (see, Tanaka and Tarumi, 1989, p.18-19). Therefore, we can define the EIFs for eigenvalues and eigenfucntions respectively as follows:

$$EIF(x;\rho_s) = \rho_s^{(1)} = (\mathbf{S}^{-T}\mathbf{y})_r^T \left(\mathbf{SJVJS}^T\right)^{(1)} (\mathbf{S}^{-T}\mathbf{y})_s, \tag{24}$$

$$EIF(x;\xi_s(t)) = (\mathbf{y}_s^{(1)})^T \phi(t) = \left[\mathbf{S}^T(\mathbf{S}^{-T}\mathbf{y})_s^{(1)} + (\mathbf{S}^{(1)})^T(\mathbf{S}^{-T}\mathbf{y})_s\right]^T \phi(t). \tag{25}$$

It is not easy to derive analytically the EIF for λ determined by cross-validation. So we use the difference of λ's for a small ϵ in calculating the EIF for λ .

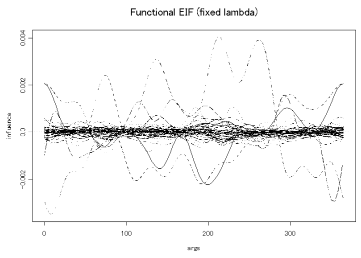


Figure 3: Functional EIFs for PC1 weight function with fixed λ.

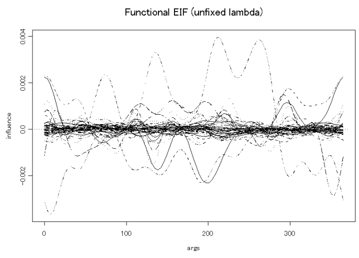


Figure 4: Functional EIFs for PC1 weight function with unfixed λ.

3.5 Numerical example

We shall apply sensitivity analysis to the penalized functional PCA of the temperature data. For example, we evaluate the influence of each individual for the PC1 weight function. Figure 3 shows the EIF for PC1 weight function when λ is fixed. Figure 4 shows the EIF for PC1 weight function when the effect through λ is taken into account. Our final aim is to approximate the SIF by using the EIF. To confirm the validity of the derived EIF, in Figure 5 and Figure 6, the scatterplots are drawn between the EIFs with fixed λ and the SIFs, where the latter are calculated by actually omitting each observation. On the other hand, in Figure 7 and Figure 8, the scatterplots are drawn between the EIFs with variable λ and SIFs. Figure 5 and Figure 7 correspond to linear approximation, while Figure 6 and Figure 8 correspond to quadratic approximation. From these figures, it is found that almost all points are located on or near the straight line through the origin. Therefore we can conclude that the EIF can be used instead of the SIF in evaluating the influence of each individual. Comparing the Figure 5 and Figure 7, we can find some improvement of approximation by taking the influence through λ into consideration. However, it seems that the effect of the smoothing parameter is quite small compared to the other effects. We may conclude that the influence function based on the fixed λ is enough in actual applications, because the calculation of the influence of λ requires considerable computational burden and mathematical difficulty.

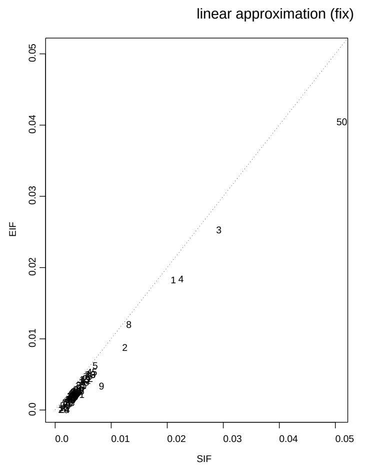


Figure 5: EIFs with fixed λ versus SIFs (linear approximation).

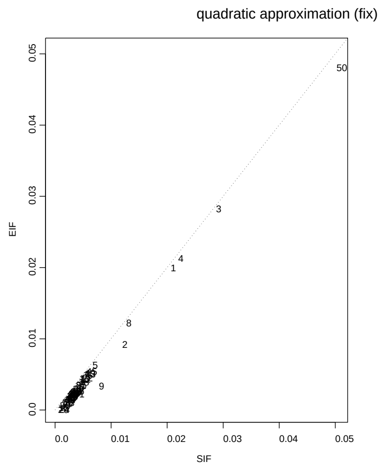


Figure 6: EIFs with fixed λ versus SIFs (quadratic approximation).

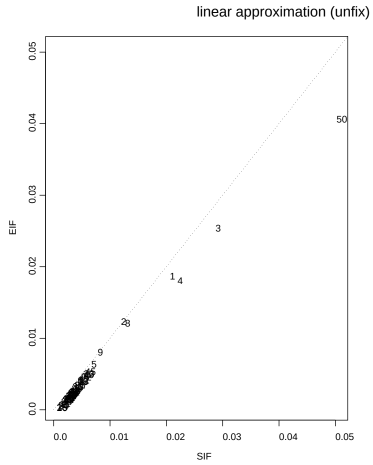


Figure 7: EIFs with unfixed λ versus SIFs (linear approximation).

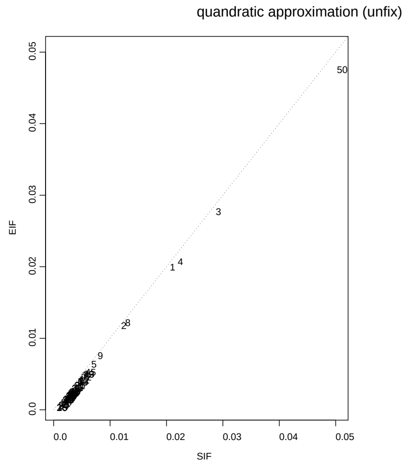


Figure 8: EIFs with unfixed λ versus SIFs (quadratic approximation).

4 Single-case diagnostics

4.1 What is single-case diagnostics?

Among various measures to evaluate the influence of each individual on a parameter vector, so-called generalized Cook's D may be the most popular measure. It is defined by

$$D_i = [EIF_i]^T [\widehat{acov}]^{-1} [EIF_i], \tag{26}$$

where $\widehat{acov}$ is the asymptotic covariance matrix for the parameters of interest. So far, the theory of probability in functional context has not been developed yet, so we can not apply the above Cook's D directly to a functional parameter. In this section, we define the generalized Cook's D in functional data analysis in two different ways.

4.2 Cook's D based on the coefficients

When the number of the basis functions is fixed, each functional observation $x_i(s)$ can be determined by the corresponding coefficient vector $\mathbf{c}_i$. Then, by regarding $\mathbf{c}_i$ as the original multivariate observation, we can define Cook's D in the similar manner as in the case of ordinary multivariate analysis. In other words, we define $EIF_i = \mathbf{y}_i^{(1)}$ as the EIF for the eigenvector, where $\mathbf{y}$ is the coefficient vector of ξ. To evaluate $\widehat{acov}$ we use a nonparametric method, for example, the Jackknife estimate defined by

$$\widehat{acov}_{\mathbf{JK}} := \frac{1}{N(N-1)} \sum_{i=1}^{N} JIF_i JIF_i^T \tag{27}$$

where $JIF(x_i; \hat{\theta}) = (N-1)(\hat{\theta}^{[-i]} - \hat{\theta})$. In this case, $\widehat{acov}$ is computed as

$$\widehat{acov}_{\mathbf{JK}}(\mathbf{y}) = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\mathbf{y}^{[-i]} - \mathbf{y})(\mathbf{y}^{[-i]} - \mathbf{y})^T. \tag{28}$$

4.3 Cook's D based on the sampled functions

We may also define Cook's D based on the functional EIF $\mathbf{y}_i^{(1)T}\phi(t)$. To do this we first take samples from $\xi_i^{(1)} = \mathbf{y}_i^{(1)T}\phi(t)$ at appropriate grid points, i.e., we define $EIF_i = (\xi_i^{(1)}(t_1), \cdots, \xi_i^{(1)}(t_H))^T = \Phi^T \mathbf{y}_i^{(1)}$, where Φ is a $K \times H$ matrix whose (k,h)th element is given by $\phi_k(t_h)$. Here in defining Φ we take H grid points so that the rank of Φ is equal to K. If we use the Jackknife

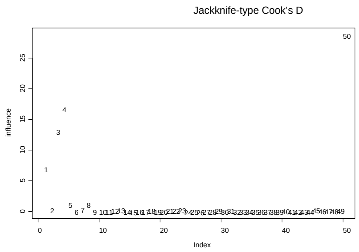


Figure 9: Cook's D based on coefficient.

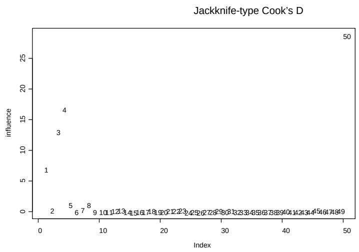


Figure 10: Cook's D based on sampled function.

estimates as in the previous section, the asymptotic covariance matrix of $\Phi^T \mathbf{y}_i$ can be obtained by

$$\widehat{acov}_{\mathbf{JK}}(\Phi^T \mathbf{y}) = \Phi^T \widehat{acov}_{\mathbf{JK}}(\mathbf{y})\Phi, \tag{29}$$

where $\widehat{acov}_{\mathbf{JK}}(\mathbf{y})$ and $\widehat{acov}_{\mathbf{JK}}(\Phi^T \mathbf{y})$ indicate the Jackknife estimates of the covariance matrices of $\mathbf{y}$ and $\Phi^T \mathbf{y}$, respectively.

4.4 Numerical example

Here we apply both methods to the derived functional EIFs defined by eq.(22) and use the Jackknife estimates of the covariance matrices of $\mathbf{y}$ and $\Phi^T \mathbf{y}$. Figure 9 shows the Cook's Ds based on the coefficient of the functional EIF, while Figure 10 shows the Cook's Ds based on the sampled function of the functional EIF. From the Cook's D, we can detect observations No.1, No.3, No.4 and No.50 as candidates for highly influential observations. This method makes it easier for us to evaluate the influence of each individual. Comparing these two figures we can find that the behavior of the Cook's D based on sampled functions is almost the same as that based on the coefficients. In section 6, we will mathematically discuss the relationship of Cook's Ds based on the coefficient and sampled function.

5 Multiple-case diagnostics

5.1 What is multiple-case diagnostics?

So far we have discussed methods for detecting singly influential observations. However, it is important to detect jointly influential observations as well. Tanaka has proposed a general procedure of applying principal component analysis (PCA) with an appropriate metric to the influence functions for detecting influential subsets of observations in multivariate methods (see e.g., Tanaka, 1994). In short the basic idea is described as follows. Since the influence of multiple observations can be evaluated with the sum of their influence functions, influential subsets of observations should be composed of the observations with influence functions which have large norms and similar directions. It works well in ordinary multivariate data analysis, but it can not be applied in functional data analysis directly. In this section, we conduct multiple-case diagnostics in functional data analysis.

5.2 PCA of influence functions

We propose to use PCA with an appropriate metric to a set of EIFs. The PCA is formulated by a generalized eigenvalue problem as

$$\left(\frac{1}{N} EIF_i EIF_i^T\right) a = \rho \mathbf{M} a, \tag{30}$$

where $\mathbf{M}$ is a nonnegative definite matrix. In the similar manner as in single-case diagnostics, we consider applying in two ways. One is to apply this type of PCA to the multivariate data of the coefficients of the basis function expansion. The other is to apply the PCA to the multivariate data generated by sampling from the functional EIF. The inverse or g-inverse of the corresponding asymptotic covariance matrix $\widehat{acov}$ is used as metric $\mathbf{M}$, where $\widehat{acov}$ is estimated by the Jackknife method as well.

5.3 Numerical example

Now we apply the PCA to the functional EIFs for PC1 weight function. Figure 11 and Figure 12 show the scatter plots between PC1 scores and PC2 scores of the EIFs of the coefficients and of the sampled EIFs, respectively. Figure 11 corresponds to the PCA of the EIF of the coefficients, while Figure 12 corresponds to the PCA of the sampled functions. Figure 13 and Figure 14 show the result of varimax rotation to all the PC scores of the influence function, respectively. It seems that varimax rotation works well

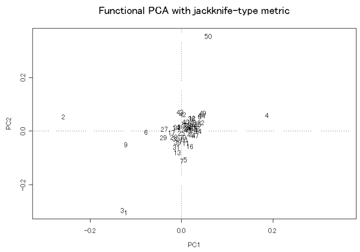


Figure 11: PCA based on coefficient.

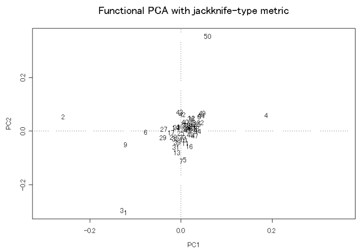


Figure 12: PCA based on sampled function.

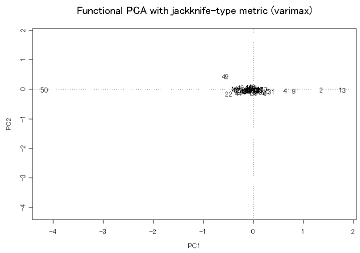


Figure 13: PCA based on coefficient. (applying varimax rotation to all the PC scores)

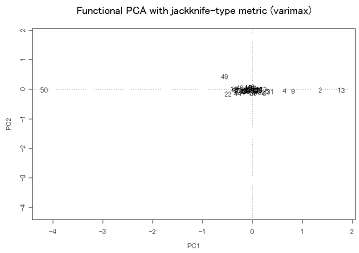


Figure 14: PCA based on sampled function. (applying varimax rotation to all the PC scores)

for getting the result easier to interpret. Looking at these figures we can detect two influential subsets. One is the subset which consists of observations such as No.1, No.2, No.3, No.4 and No.9, while the other is the subset which consists of observation No.50. We can guess that the above two influential subsets have exact opposite patterns of the influence. Recall that the PC1 is a total measure of overall temperatures throughout the year, and the weight on the winter is larger than that of any other season. No.1, No.2, No.3 and No.4 are located in Hokkaido. Hokkaido is the area famous for cold climate and cold winter in Japan, while No.50 Okinawa is the weather station located on the southernmost part of Japan, which has a warm climate throughout the year. So we can conclude that these results are valid in view of scientific account.

6 Mathematical relationship

In single-case diagnostics, we defined Cook's D in two different manners. One is based on the EIF for the coefficient vector of the basis function expansion, i.e.,

$$D_{1i} = \mathbf{y}_i^{(1)T} \mathbf{V}_y^{-1} \mathbf{y}_i^{(1)}, \tag{31}$$

where $\mathbf{V}_y$ is an estimate for the asymptotic covariance matrix of $\mathbf{y}_i$. The other is based on the sampled vector of the functional EIF $\mathbf{y}_i^{(1)T}\phi(t)$, i.e.,

$$D_{2i} = \mathbf{y}_i^{(1)T} \Phi(\Phi^T \mathbf{V}_y \Phi)^- \Phi^T \mathbf{y}_i^{(1)}. \tag{32}$$

If we choose a $K \times H$ matrix Φ so that rank $\Phi = K$, then we can prove that the relation $\mathbf{V}_y^{-1} = \Phi(\Phi^T \mathbf{V}_y \Phi)^- \Phi^T$ holds (see, Searle, 1982, p.224), and therefore $D_{1i} \equiv D_{2i}$ for any i. It implies that we do not need to take samples more than K grid points in the second formulation and that at least from the aspect of single-case diagnostics it is enough to develop sensitivity analysis based on the EIF for the coefficients of the basis function expansion.

In multiple-case diagnostics we applied PCA with metric $[\widehat{acov}]^-$ to the EIF to detect influential subsets of observations (see, Tanaka, 1994; Tanaka and Zhang, 1999). Here we can use both of the EIF for the coefficient vectors and the EIF for the sampled functional EIF. We obtain eigenvalue problems

$$\left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i^{(1)} \mathbf{y}_i^{(1)T} - \nu \mathbf{V}_y \right) \mathbf{a} = 0, \tag{33}$$

in the former formulation, and

$$\left(\frac{1}{N} \sum_{i=1}^{N} \Phi^T \mathbf{y}_i^{(1)} \mathbf{y}_i^{(1)T} \Phi - \nu (\Phi^T \mathbf{V}_y \Phi) \right) \mathbf{b} = 0, \tag{34}$$

in the latter formulation. Assume that rank $\Phi = K$. Then, multiplying $(\Phi\Phi^T)^{-1}\Phi$ to equation (34) from the left we obtain an eigenproblem of $\Phi\mathbf{b}$ which is just equivalent to equation (33). Therefore, we may conclude that, though we derived two kinds of EIF, we need not develop sensitivity analysis based on the functional EIF in addition to sensitivity analysis based on the EIF for the coefficients of the basis function expansion.

7 Concluding remarks

In the present paper we derived empirical influence functions for eigenvalues and eigenfunctions of penalized PCA and proposed multiple-case as well as single-case diagnostics procedures based on them.

In deriving the EIFs for eigenvalues and eigenfunctions we studied two cases depending on whether or not to consider the effects through the change of the smoothing parameter λ. The result of numerical example shows that the effects are so small that we need not take the effects through λ into consideration.

In the formulation of functional PCA we utilized the so-called basis function expansion, and derived the EIFs for eigenvalues and eigenfunctions by applying the perturbation theory of (matrix) eigenvalue problems. Then, for single-case and multiple-case diagnostics we proposed to define influence statistics based on two approaches. One is to regard the coefficient vectors of basis function expansion as the multivariate observations and define influence statistics using the EIFs for the coefficient vectors just as in the case of multivariate analysis. The other is to transform the functions, i.e., original functions and also their functional EIFs, into sets of vectors by sampling at appropriate grid points and define influence statistics in the similar manner as in sensitivity analysis in multivariate methods. It was found that, if we select enough number of grid points in such a way that the rank of the sampled functions is equal to the number of the basis functions, both approaches give the same results, and therefore we can construct influence statistics on the basis of EIFs for the coefficient vectors of basis function expansion.

Acknowledgements

We are grateful to the editor and reviewers for their careful reading. Their suggestions are helpful for the improvement of presentation. This research was partly supported by Japan Society of the Promotion of Science (Grant-in-Aid for Scientific Research (C) 13680374).

References

[1] Besse, P. and Ramsay, J. O. (1986). Principal components analysis of sampled functions, *Psychometrika, 51, 285-311.*

[2] Caussinus, H. and Ruiz, A. (1990). Interesting projections of multidimensional data by means of generalized principal component analysis. *Compstat, Heidelberg: Physica-Verlag, 121-126.*

[3] Critchley, F. (1985). Influence in principal component analysis. *Biometrika,* **72***, 627-636.*

[4] Japan Meteorological Agency (1999). Annual report of Automated Meteorological Data Acquisition System, *Japan Meteorological Business Support Center (JMBSC).*

[5] Ramsay, J. O. and Dalzell, C. (1991). Some tools for functional data analysis (with discussion), *J. Royal Statist. Soc.*, B 53, 539-572.

[6] Ramsay, J. O. and Silverman, B.W. (1997). Functional Data Analysis, *Springer.*

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

[8] Searle, S. R. (1982). Matrix Algebra Useful for Statistics, *Wiley.*

[9] Tanaka, Y. (1984). Sensitivity analysis in Hayashi's third method of quantification, *Behaviormetrika, 16, 31-44.*

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

[11] Tanaka,Y. and Tarumi,T.(1989). Sensitivity analysis in canonical factor analysis. *J. Japanese Soc. Comp. Statist.2, 9-20.*

[12] Tanaka, Y. and Zhang, F. (1999). R-mode and Q-mode influence analyses in statistical modelling: Relationship between influence function approach and local influence approach, *Computational Statistics and Data Analysis, 32, 197-218.*