Forward Search Algorithm for Robust Influence Analysis in Maximum Likelihood Factor Analysis

Woosung Yang,† Yutaka Tanaka††, and Jun Nakaya†

†Clinical Genome Informatics Center, Kobe University, Kobe, 650-0017 Japan
††Faculty of Mathematical Sciences and Information Engineering, Nanzan University, Seto, 489-0863 Japan

Summary
Mainly in regression analysis, numerous methods have been proposed historically for the analysis of the influence of single or multiple observations on the results of analysis. Such a sensitivity or stability problem is not special to the regression analysis, but is common to the other statistical methods including the multivariate methods. We combined the general procedure of the sensitivity analysis and the forward search method to detect the influential observations without suffering from the masking and swamping effect, and compared its performance with the other robust methods numerically. The proposed procedure can be applied to any multivariate methods with minor modification. In this paper we propose and discuss our procedure in maximum likelihood factor analysis (MLFA).

Key words:
Sensitivity analysis, Maximum likelihood factor analysis, Robust method, Forward search.

1. Introduction

We may consider that a statistical method is a system, a set of data is an input and the result of analysis is an output. We are interested in the sensitivity of this system, that is, how a small change of data affects the result of analysis. Since late 1970, mainly in regression analysis, numerous methods have been proposed for the analysis of the influence of single or multiple observations on the results of analysis, and major part of them have been summarized into several books(see, e.g., [1], [2], [3], [4], [5]).

The issue of robust estimation and/or outlier detection has been researched by many authors. Rousseeuw [6] introduced minimum volume estimator (MVE), and minimum covariance determinant (MCD) and used them for outlier detection. Other authors have used the concept of MVE or MCD in their outlier detection methods. Atkinson [7], in his outlier detection method, considered forward search from random elemental sets and chose partition of the data which had the smallest "half" sample ellipsoid volume. Rocke and Woodfuff [8] obtained a hybrid algorithm utilizing the steepest descent procedure of Hawkins [9] for obtaining the MCD, which was used as a starting point for the forward search algorithm of Atkinson [10] and Hadi [11]. Atkinson and Riani[12] proposed an idea of forward search in regression analysis to protect from the masking effect. The basic idea of forward search is to select observations forwardly in a successive manner based on their closeness to the fitted model starting from the fit to an initial subset which can be regarded as outlier free. In the later part half we discuss a robust method of sensitivity analysis in multivariate methods using the idea of the forward search method by Atkinson and Riani [12].

Covariance structure analysis (CSA) is family of multivariate methods which have a common fundamental assumption that the covariance matrix Σ of observable variables is expressed as a function of a set of parameters $\underline{\theta} = (\theta_1, \cdots, \theta_m)$, i.e., $\Sigma = \Sigma(\underline{\theta})$. This family contains confirmatory as well as exploratory factor analysis (FA), path analysis, LISREL type linear structural equation analysis among others.

Sensitivity analysis procedure have been proposed by Tanaka, Watadani and Moon [13] and Tanaka and Watadani [14] for CSA without/with equality constraints. In these papers the proposed procedures are illustrated in confirmatory and exploratory factor analysis. In this paper we deal with CSA with equality constraints and introduce a general procedure of sensitivity analysis in such kinds of CSA proposed by Tanaka and Watadani [14]. As a special case we consider maximum likelihood factor analysis (MLFA).

The main goal of this paper is combine the general procedure of influence analysis in MLFA and the forward search method to detect influential observations without suffering from the masking and swamping effect. The proposed procedure can be applied to any multivariate method with minor modification. But, here we discuss our procedure in MLFA.

2. Influence Functions in CSA

In CSA as estimate $\hat{\underline{\theta}}$ for parameter vector $\hat{\theta}$ is obtained by minimizing a function called discrepancy function

Manuscript received July 5, 2006.
Manuscript revised July 25, 2006.

$G(S,\Sigma(\underline{\theta}))$ which measures the discrepancy between the sample covariance matrix $S = n^{-1}\Sigma(\underline{x}_i - \overline{\underline{x}}_i)(\underline{x}_i - \overline{\underline{x}}_i)^T$ and the reproduced covariance matrix $\Sigma(\underline{\theta})$ using the estimated parameters. The discrepancy functions is given by

$$G_{ML}(S,\Sigma) = tr[\Sigma^{-1}S] - \log|\Sigma^{-1}S| - p \tag{1}$$

in maximum likelihood(ML) estimation(see,e.g., Joreskog, [15]).

The major objective of sensitivity analysis is to detect influential observations. To evaluate the influence of each observation, we use the idea of so-called influence function proposed by Hampel [16]. In practice, we usually focus our interest on the influence of each observations $\underline{x} = \underline{x}_i (i = 1,\cdots,n)$ on the estimate $\hat{\underline{\theta}} = \theta(\hat{F})$ of the parameter vector. Where $\hat{F}$ indicates the empirical distribution function. For this propose the empirical influence function (EIF) of $\underline{x}_i$ for $\hat{\underline{\theta}}$ is defined. Tanaka and Watadani [14] discussed the case of CSA with r equality constraints expressed as

$$\underline{h}(\underline{\theta}) = \begin{pmatrix} h_1(\underline{\theta}) \\ \cdots \\ h_r(\underline{\theta}) \end{pmatrix} = \underline{0}. \tag{2}$$

Using the Lagrangian function

$$G^*(\underline{s},\underline{\theta}^*) = G^*(\underline{s},\hat{\underline{\theta}},\hat{\underline{\lambda}}) = G(\underline{s},\underline{\theta}) + \underline{\lambda}^T \underline{h}(\underline{\theta}), \tag{3}$$

where $\underline{\lambda} = (\lambda_1,\cdots,\lambda_r)$ is an $r \times 1$ vector of Lagrangian multipliers and $\underline{\theta}^* = (\underline{\theta}^T, \underline{\lambda}^T)^T$, the EIF of $\underline{x}_i$ for $\hat{\underline{\theta}}$ is given by

$$EIF(\underline{x}_i;\hat{\underline{\theta}}) = -Q(\underline{s},\hat{\underline{\theta}},\hat{\underline{\lambda}}) \left[\frac{\partial^2 G(\underline{s},\hat{\underline{\theta}})}{\partial \underline{\theta} \partial \underline{s}^T} \right] EIF(\underline{x}_i;\underline{s}) \tag{4}$$

where $\underline{s}$ denotes a vector consists of the elements of covariance matrix S, i.e., $\underline{s} = (s_{11},\cdots,s_{p1};s_{22}\cdots s_{p2};\cdots s_{pp})^T$, the EIF of $\underline{x}_i$ for $\underline{s}$ is given by $EIF(\underline{x}_i;\underline{s}) = vech[(\underline{x}_i - \overline{\underline{x}})(\underline{x}_i - \overline{\underline{x}})^T - S]$, and $Q(\underline{s},\hat{\underline{\theta}},\hat{\underline{\lambda}})$ is the upper left $m \times m$ submatrix of the $(m+r)\times(m+r)$ matrix

$$\left[\frac{\partial^2 G(\underline{s},\hat{\underline{\theta}}^*)}{\partial \underline{\theta}^* \partial \underline{\theta}^{*T}} \right]^{-1} = \begin{bmatrix} \frac{\partial^2 G(\underline{s},\hat{\underline{\theta}})}{\partial \underline{\theta} \partial \underline{\theta}^T} + \sum_{j=1}^{r} \lambda_j \left[\frac{\partial^2 h_j(\hat{\underline{\theta}})}{\partial \underline{\theta} \partial \underline{\theta}^T} \right] & \frac{\partial \underline{h}^T(\hat{\underline{\theta}})}{\partial \underline{\theta}^T} \\ \frac{\partial \underline{h}(\hat{\underline{\theta}})}{\partial \underline{\theta}^T} & 0 \end{bmatrix}^{-1}. \tag{5}$$

To calculate the EIF for $\hat{\underline{\theta}}$ the second derivatives of discrepancy function G are required. It known that the first derivatives of G are given in the form as

$$\frac{\partial G(S,\Sigma)}{\partial \theta_i} = tr\left[\Sigma^{-1}(\Sigma - S)\Sigma^{-1} \frac{\Sigma}{\partial \theta_i} \right]. \tag{6}$$

From Eq. 6 the second derivatives of G are obtained as

$$\frac{\partial^2 G(\underline{s},\underline{\theta})}{\partial \theta_i \partial \theta_j} = tr\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right] + 2tr\left[\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} (\Sigma - S)\Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_j} \right] + tr\left[\Sigma^{-1}(\Sigma - S)\Sigma^{-1} \frac{\partial^2 \Sigma}{\partial \theta_i \partial \theta_j} \right] \tag{7}$$

and

$$\frac{\partial^2 G(\underline{s},\underline{\theta})}{\partial \theta_i \partial s_{jk}} = -tr\left[\Sigma^{-1} E_{jk}^* \Sigma^{-1} \frac{\partial \Sigma}{\partial \theta_i} \right] \tag{8}$$

where E_{jk}^* is $p \times p$ matrix with 1's in the (j,k) and (k,j) elements and 0's in the other elements. The second derivative Eq. 7 is often approximated by only the first terms of the right-hand side neglecting the remaining terms which contain the factor $(\Sigma - S)$. Influence functions of the estimate $\hat{\underline{\theta}}$ are in general vector valued. So we need to transform them into scalar measures. In this paper we use the generalized Cook's distance such as

$$D_i = EIF(\underline{x}_i;\hat{\underline{\theta}})^T \hat{a}cov(\hat{\underline{\theta}}) EIF(\underline{x}_i;\hat{\underline{\theta}}) \tag{9}$$

where $\hat{a}cov(\hat{\underline{\theta}})$ is an estimate of the asymptotic covariance matrix of $\hat{\underline{\theta}}$. It is known that $\hat{a}cov(\hat{\underline{\theta}})$ is given in the case of ML as

$$\hat{a}cov(\hat{\underline{\theta}}) = \frac{2}{n} Q(\underline{s},\hat{\underline{\theta}},\hat{\underline{\lambda}}) \tag{10}$$

in CSA with equality constraints.

Let us consider the influence of a set of k observations $A=\{\underline{x}_{i1},\cdots\underline{x}_{ik}\}$ on a parameter vector $\underline{\theta}(F)$, which is given as a functional of the cumulative distribution function(cdf). To do this we introduce a perturbation on the cdf from F to $\widetilde{F}=(1-\varepsilon)F+\varepsilon G$, where $G=k^{-1}\sum_{\underline{x}_i\in A}\delta_{\underline{x}_i}$, $\delta_{\underline{x}_i}$ being the cdf of a unit point mass at $\underline{x}_i$, and define a generalized theoretical influence function(TIF) of A as the limit $TIF(A;\underline{\theta})=\lim_{\varepsilon\to 0}[\underline{\theta}(\widetilde{F})-\underline{\theta}(F)]/\varepsilon$. Then it can be easily verified that $TIF(A;\underline{\theta})=k^{-1}\sum_{\underline{x}_i\in A}TIF(\underline{x};\underline{\theta})$, where $TIF(\underline{x};\underline{\theta})$ is the ordinary influence function of $\underline{x}_i$. The similar relation holds for the empirical influence function(EIF). Hence the parameter estimate based on the sample with a subset A omitted can be approximated as

$$\underline{\hat{\theta}}_{(A)}\cong\underline{\tilde{\theta}}_{(A)}\equiv\underline{\hat{\theta}}-(n-k)^{-1}\sum_{\underline{x}_i\in A}EIF(\underline{x}_i;\underline{\hat{\theta}}) \tag{11}$$

where symbol(~) indicates the linear approximation based on the EIF.

3. Application to Factor Analysis Models

We consider the following basic model of FA:

$$\underline{\mathbf{x}}=\mathbf{\Lambda}\underline{\mathbf{f}}+\underline{\mathbf{e}} \tag{12}$$

where $\underline{\mathbf{x}}$ a $p\times 1$ vector of observable variables, $\underline{\mathbf{f}}$ is a $q\times 1(p<q)$ vector of common factor scores, $\underline{\mathbf{e}}$ is a $p\times 1$ vector of unique factor scores and $\mathbf{\Lambda}=(\lambda_{ir})$ is a $p\times q$ factor loading matrix. Means and covariances of common and unique factors are assumed as $E(\underline{\mathbf{f}})=\underline{0},E(\underline{\mathbf{e}})=\underline{0},E(\underline{\mathbf{ff}}^T)=\Phi=(\phi_{rs})$ (every diagonal element is unity), $E(\underline{\mathbf{ee}}^T)=\Psi$ (diagonal matrix) and $E(\underline{\mathbf{fe}}^T)=0$. Then the covariance matrix Σ of observable variables $\underline{x}$ is expressed as

$$\Sigma(\Lambda,\Phi,\Psi)=\Lambda\Phi\Lambda^T+\Psi \tag{13}$$

The first and second derivatives of the covariance matrix Eq. 13 with respect to parameters are given as

$$\frac{\partial\Sigma(\Lambda,\Phi,\Psi)}{\partial\lambda_{is}}=E_{is}\Phi\Lambda^T+\Lambda\Phi E_{si},$$
$$\frac{\partial\Sigma(\Lambda,\Phi,\Psi)}{\partial\phi_{st}}=\Lambda E_{st}^*\Lambda,\qquad\frac{\partial\Sigma(\Lambda,\Phi,\Psi)}{\partial\psi_{ii}}=E_{ii}^*$$

and

$$\frac{\partial^2\Sigma(\Lambda,\Phi,\Psi)}{\partial\lambda_{is}\partial\phi_{tu}}=\delta_{st}[E_{iu}\Lambda^T+\Lambda E_{ui}]+\delta_{su}[E_{it}\Lambda^T+\Lambda E_{ti}]$$
$$\frac{\partial^2\Sigma(\Lambda,\Phi,\Psi)}{\partial\lambda_{is}\partial\lambda_{jt}}=\delta_{st}(E_{ij}+E_{ji}) \tag{14}$$

where the symbol δ represents Kronecker's delta and E_{ab} is a matrix unit with 1 in the (a,b) element and 0's in the other elements.

In exploratory FA, on the other hand, no elements are specified in advance. Instead to obtain a unique solution it is usually assumed that common factors are not correlated, i.e., $E(\underline{\mathbf{ff}}^T)=\Phi=I$, and $\Lambda^T\Phi^{-1}\Lambda$ is diagonal. Thus we may identify exploratory FA as a special case of CSA with $r=q(q-1)/2$ equality constraints

$$h_{st}=(\Lambda,\Psi)=[\Lambda^T\Phi^{-1}\Lambda]_{st}\qquad(1\le s<t\le q). \tag{15}$$

The parameters to be estimated are the elements of Λ and the diagonal elements of Ψ. The first and second derivatives of h_{st} are given as

$$\frac{\partial h_{st}(\Lambda,\Psi)}{\partial\lambda_{iu}}=\frac{\delta_{su}\lambda_{it}+\delta_{tu}\lambda_{is}}{\psi_{ii}},\frac{\partial h_{st}(\Lambda,\Psi)}{\partial\psi_{ii}}=-\frac{\lambda_{is}\lambda_{it}}{\psi_{ii}^2}$$

and

$$\frac{\partial^2 h_{st}}{\partial\lambda_{iu}\partial\lambda_{jv}}=\frac{\delta_{ij}(\delta_{su}\lambda_{it}+\delta_{tu}\lambda_{is})}{\psi_{ii}},$$
$$\frac{\partial^2 h_{st}}{\partial\lambda_{iu}\partial\psi_{jj}}=-\frac{\delta_{ij}(\delta_{su}\lambda_{it}+\delta_{tu}\lambda_{is})}{\psi_{ii}^2}, \tag{16}$$
$$\frac{\partial^2 h_{st}}{\partial\psi_{ii}\partial\psi_{jj}}=\frac{2\delta_{ij}\lambda_{is}\lambda_{it})}{\psi_{ii}^3}.$$

4. Application of the Forward Search Procedure

Atkinson and Riani [12] proposed a "forward" procedure of selecting the subsets of observations. An initial subset of the smallest possible size is selected by fitting the statistical model to a large number of subsets of the size and by evaluating the goodness of the fit. Then all observations are ordered by their closeness to this fitted model; for regression model the residuals determine closeness and for multivariate models other measures such as values of normal density function, since the data be assumed multivariate normal distribution, so that we expect outlying observations have small values, play the similar role. The subset size is increased by one based on the closeness measure and the model is refitted to the observations of the increased subset size. The process continues until all the observations are included in the subset. As the result of this forward search we have an ordering of the observations by the closeness to the assumed model. The changes of various statistics, such as Cook's D_i, are monitored in each step of this forward search.

Now we use the above forward procedure and propose the following forward algorithm for robust influential analysis in MLFA.

- Step 1. Choice of the initial subset.

If the model contains P parameters, the forward search algorithm starts with the selection of a subset of $p+1$ observations. Observations in this subset are intended to be outlier free. To choose the initial subset we make use of the MVE procedure. Using the MVE, we obtain the robust mean vector $\underline{x}_R$ and the robust covariance matrix S_R, then calculate the values of multivariate normal density function $f(\underline{x}_i;\hat{\underline{\theta}}_R)$ of all the data points, where $\hat{\underline{\theta}}_R=(\hat{\underline{\mu}}_R,\hat{\Sigma}_R)$ and $\hat{\Sigma}_R=\hat{\Lambda}_R\hat{\Lambda}_R^T+\hat{\Psi}_R$. $\hat{\Lambda}_R$ and $\hat{\Psi}_R$ indicate the estimated factor loading matrix and unique variance diagonal matrix, respectively, which are obtained from the MLFA based on the $\hat{\underline{\mu}}_R$ and the $\hat{\Sigma}_R$.

$$f(\underline{x}_i;\hat{\underline{\theta}}_R)=\frac{1}{(2\pi)^{p/2}\left|\hat{\Sigma}_R\right|^{1/2}}\exp\left[-\frac{(\underline{x}_i-\hat{\underline{\mu}}_R)^T\hat{\Sigma}_R^{-1}(\underline{x}_i-\hat{\underline{\mu}}_R)}{2}\right]$$
$$,i=1,\cdots n. \tag{17}$$

Our initial subset is selected as the set of observations which have the $p+1$ largest $f(\underline{x}_i;\hat{\underline{\theta}}_R)$ values.

- Step 2. Adding observations during the forward search.

Given a subset of size m, the forward search moves to size $m+1$ by selecting the $m+1$ observations with the largest $f(\underline{x}_i;\hat{\underline{\theta}}_J)$,

$$f(\underline{x}_i;\hat{\underline{\theta}}_R)=\frac{1}{(2\pi)^{p/2}\left|\hat{\Sigma}_J\right|^{1/2}}\exp\left[-\frac{(\underline{x}_i-\hat{\underline{\mu}}_J)^T\hat{\Sigma}_J^{-1}(\underline{x}_i-\hat{\underline{\mu}}_J)}{2}\right]$$
$$,i=1,\cdots n. \tag{18}$$

where, $\hat{\underline{\mu}}_J$ and $\hat{\Sigma}_J$ are the mean vector and covariance matrix, respectively, which are obtained from subset J. Using this subset we recompute the statistics, such as $V_J(\hat{acov}(\hat{\underline{\theta}}_J))$ and $EIF(\underline{x}_i;\hat{\underline{\theta}}_J)$. The forward search is repeated in this way until all observations are chosen in the subset, and at each step some diagnostic statistics are monitored.

- Step 3. Monitoring the search.

In monitoring the forward search we additionally calculate Cook's D_i, and study numerically how $\{D_i^J, i=1,\cdots n\}$ change in the forward search process.

$$D_i^J=\left(EIF(\underline{x}_i;\hat{\underline{\theta}}_J)\right)^T V_J^+\left(EIF(\underline{x}_i;\hat{\underline{\theta}}_J)\right) \tag{19}$$

5. Numerical Example

Let us investigate the performance of the proposed procedure in the case of MLFA. In the following analysis we study how the procedure works when there exist multiple influential observations and how large are their influences on the factor loadings and/or unique variances. To do this we generate an artificial data based on the following factor analysis model:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0.8 & 0 \\ 0.8 & 0 \\ 0.8 & 0 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} 0.6e_1 \\ 0.6e_2 \\ 0.6e_3 \\ 0.6e_4 \\ 0.6e_5 \\ 0.6e6 \end{bmatrix} \quad (20)$$

where

$$e_1, \cdots, e_6 \sim NID(0,1)$$
$$f_1, f_2 \sim NID\left(\underline{0}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right) \quad (21)$$

A set of 100 observations are generated based on the above factor analysis model. But, in generating observations #41~#45 the values of f_1, f_2 are replaced by (2.4,-2.4), and #81~#85 the values of e_1, e_2 are replaced by (2.7,-2.7).

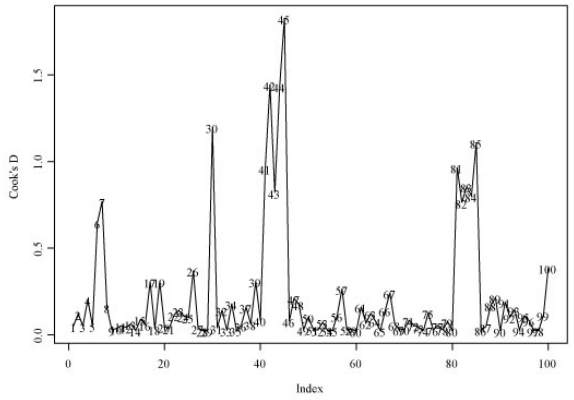

Fig. 1 Index plot of Cook's D in the ordinary procedure

Before analyzing the data set, we apply the ordinary general procedure. The result of the ordinary general procedure could not reveal the perturbed observations as most influential. Fig. 1 is the index plot of the generalized Cook's D_i. It is noted that observation #30 is more influential than most of the perturbed observations #41, #43 and #81 ~ 85.

To search for subsets of cases whose influence functions are located far and on similar directions from the origin, we applied PCA with metric $V^+(=[\hat{acov}(\hat{\underline{\theta}})]^+)$. Fig. 2 shows the scatter plot of the first two PC scores and second vs. third PC scores. In Fig. 2 the ordinary procedure was applied. The eigenvalues are 18082.6495 > 13703.8018 > 5114.3863 > 1561.8573 > …, in order of their magnitudes. This result can not reveal influential observations correctly.

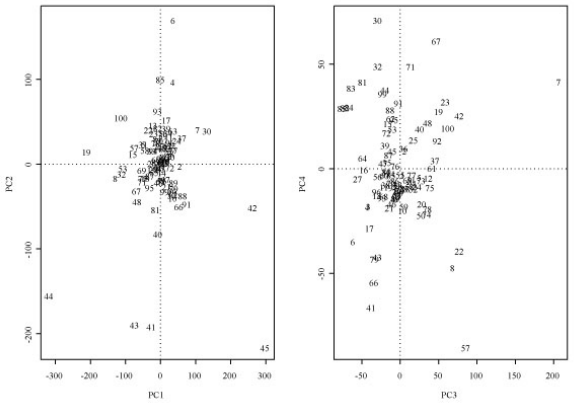

Fig. 2 Scatter plot of PC scores in the ordinary procedure

Then, the robust version of the general procedure was applied. We randomly drew 1680 subsamples of size 7, and then proceeded to the iterative process of one-step improvement. In this iterative process zero weights were assigned to 12 observations including #41~#45, #81~#85, #7 and #30. As cutoff value for these distances we used 3.80 which is the square root of the 0.975 quantile of the chi-square distribution with 6 degrees of freedom. The index plot of the Cook's D_i is shown in Fig. 3. It is obvious that all outliers are easily found to be influential. In Fig. 4 the robust version was applied and the eigenvalues obtained are 8378.791 > 4743.404 > 1605.236 > 909.820 > 438.142 > …. From Fig. 4 we can easily find the two sets of influential observation.

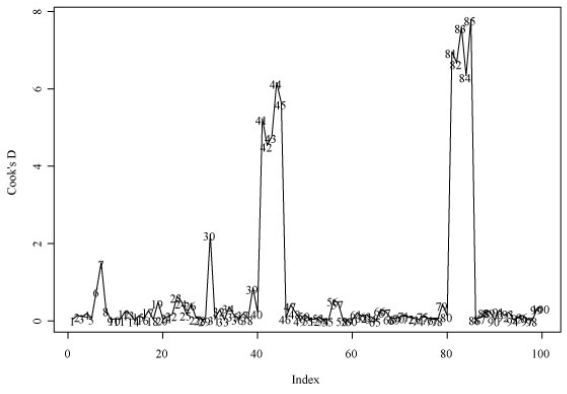

Fig. 3 Index plot of Cook's D in the robust procedure

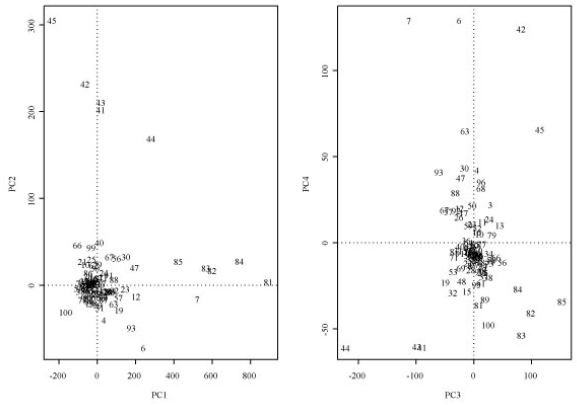

Fig. 4 Scatter plot of PC scores in the robust procedure

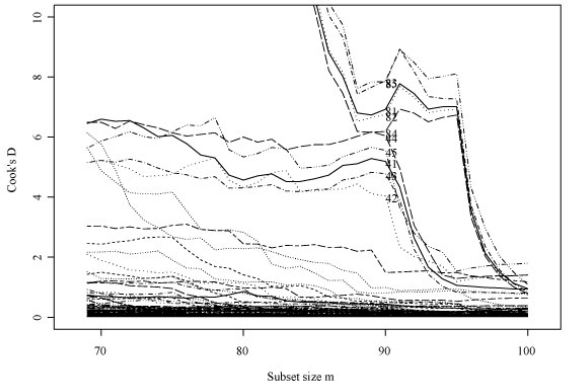

Fig. 5 Forward plot of Cook's D from the subset size m=70

Now, we apply the forward search method [17]. First, using the MVE, the robust mean vector and the robust covarianc matrix are computed and the best subset of size $P+1(=19)$ is selected in the sense that the likelihood function has the largest values. Based on the sample mean vector and covariance matrix of this subset we calculate the density function $f(\underline{x}_i;\underline{\hat{\theta}}_J)$ for all 100 observations and select 20 observations with the largest $f(\underline{x}_i;\underline{\hat{\theta}}_J)$. Similarly in each step we select successively a subset of size $m+1$, where m is the size of the previous step. In each step we recompute the statistics such as the factor loadings $\hat{\Lambda}_J$, the unique variances $\hat{\Psi}_J$, their asymptotic covariances, and their $EIFs$ which are computed using the formulas derived by Tanaka and Watadani [14]. Then, we study numerically how $\{D_i^J, i=1,\cdots n\}$ change in the forward search process. In the step after $m=91$, when one of the outlying observations enter the subset, Cook's D_i of #41~#45 decrease dramatically (Fig. 5). This phenomenon shows clearly the masking effect.

Let us look at the process more precisely. Fig. 6 is the forward plot after the step $m=70$ until the end of forward search. This shows, when the highest five influential observations #85, #83 , #84, #82, and #81 enter into the subset, Cook's D_i goes quickly down again, and at the end of the search, the groups of influential cases are mixed together and masked.

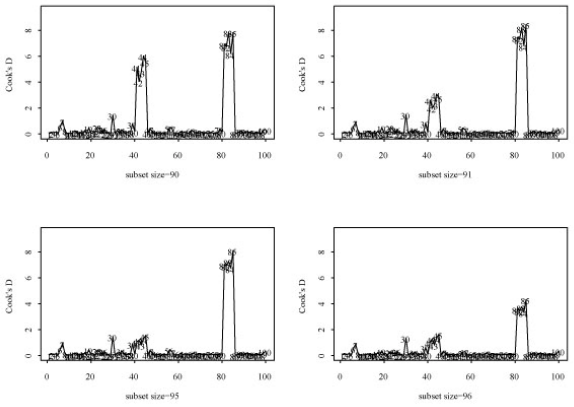

Fig. 6 Index plot of Cook's D by subset size m=90, 91, 95, 96

6. Concluding Remarks

In this paper we proposed a forward search algorithm for robust influence analysis in multivariate methods, in which the general procedure of sensitivity analysis in MLFA(Tanaka and Watadani [14]) and the forward search method(Atkinson and Riani [12]) are combined for detecting influential observations without suffering from the masking and swamping effects.

The proposed method along with the ordinary general procedure and its robust version were applied to an artificial data set which were generated in such a way that there were two groups of outlying observations. The ordinary procedure could not detect influential observations. However, the other procedures could reveal fully that there exist 10 outlying observations and that they are classified into two groups. Compared with Tanaka and Watadani [17]'s robust procedure the proposed procedure has an advantage that it can show clearly how and when the masking and/or swamping effects occur in the forward

successive process. In this sense we may say that the proposed procedure is an improved version of Tanaka and Watadani's procedure.

References

[1] Belsley, D. A., Kuh, E. and Welsch, R. E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, 1980.

[2] Cook, R. D. and Weisberg, S. Residuals and Influence in Regression. Chapman and Hall. 1982.

[3] Atkinson, A. C. Plot, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis. Oxford University Press. 1985.

[4] Chatterjee, S. and Hadi, A. S. Sensitivity Analysis in Linear Regression. John Wiley & Sons. 1988

[5] Belsley D. A. Conditioning Diagnostics: Collinearity and Week Data in Regression, John Wiley & Sons. 1991.

[6] Rousseeuw, P.J., "Multivariate Estimation with High Breakdown Point", In: Grossman, W., Pflug, G., Vincze, I., Werz, W., (Ed.), Mathematical Statistics and Applications, Vol. B, Redel, Dordrecht. 1985.

[7] Atkinson, A.C. "Fast Very Robust Methods for the Detection of Multiple Outliers", J. Amer. Statist. Assoc., 89, pp.1329-1339, 1994.

[8] Rocke, D.M. and Woodruff, D.L. "Computation of Robust Estimates of Multivariate Location and Shape", Statist. Neerlandica, 47, pp.27-42, 1993.

[9] Hawkins, D.M. "The Feasible Solution Algorithm for the Minimum Determinant Estimator in Multivariate Data", Comp. Statist. & Data Analysis, 17, pp.197-210, 1993.

[10] Atkinson, A.C. "Stalactite Plots and Robust Estimation for the Detection of Multivariate Outliers", In: Morgenthaler, S., Ronchetti, E., Stahel, W. (Ed.), Data Analysis and Robustness, Birkhauser, Basel. 1993.

[11] Hadi, A.S. "Idetifying Multiple Outliers in Multivariate Data", J. Royal Statis. Soc. Ser., B54, pp.761-771, 1992.

[12] Atkinson, A.C. and Riani, M. Robust Diagnostic Regression Analysis, Springer, 2000.

[13] Tanaka, Y., Watadani, S. and Moon, S.H. "Influence in Covariance Structure Analysis", Comm. Statist., A 20, pp.3805-3821, 1991.

[14] Tanaka, Y. and Watadani, S. "Sensitivity Analysis in Covariance Structure Analysis with Equality Constraints", Comm. Statist., A 21, pp.1501-1515, 1992.

[15] Joreskog, K.G. (1987). "Structure Analysis of Covariance and Correlation Matrices", Psychometrika, 43, pp.443-477, 1987.

[16] Hampel, F.R. "The Influence Curve and its Role in Robust Estimation", J. Amer. Statist. Assoc., 69, pp.383-393, 1974.

[17] Tanaka, Y. and Watadani, S. "Unmasking influential observations in multivariate methods", Compstat 1994, Heidelberg: Physica-Verlag, pp.292-297, 1994.