ASSESSMENT OF INFLUENCE IN FACTOR ANALYSIS REGRESSION*

ZENAIDA F. MATEO
Department of Mathematical Sciences and Physics
Central Luzon State University, Muñoz, 3120 Nueva Ecija

YUTAKA TANAKA
Department of Environmental and Mathematical Sciences
Okayama University, Tsushima, Okayama, Japan

ABSTRACT

Influence functions are derived for the parameters in the factor analysis (FA) regression and some measures are proposed to evaluate the influence on the estimated regression weights $\hat{\beta}_F$ and on the squared multiple correlation coefficient $r_2(\hat{\beta}_F)$. As an application, a numerical study is conducted to illustrate the present procedure and to show fundamental properties of influence in FA regression from the sensitivity perspective.

INTRODUCTION

Sensitivity analysis in factor analysis (FA) has been studied by several authors to evaluate the amount of influence of small changes of data. Among them are works written by Tanaka and Odaka (1989 a,b), Tanaka, Castaño-Tostado and Odaka (1990) and their related papers. For all these various versions of FA model, some measures of influence on the estimates, their precision and the goodness of fit have been used for detecting single or multiple influential observations. In the present paper, we try to develop a method of sensitivity analysis in FA regression based on the influence function derived by Tanaka and Odaka (1989b).

FA regressions is a statistical method which can deal with the problem of linear prediction where the independent variables are subject to errors. There have been some authors who devoted to study on FA regression. Among them include Horst (1941), Scott (1968), Lawley and Maxwell (1973), Isogawa and Okamoto (1980) and Browne (1988) who discussed this topic. FA regression was first proposed by

*Paper presented at the 19th Annual Statistic Meeting of the National Academy of Science and Technology held at the Philippine International Convention, Manila on July 9-10, 1997.

Horst (1941) as a method for improving the predictive validity when the given number of prediction variables is large compared to the number of cases. It was suggested by Scott (1966) as a method of linear prediction which is an alternative estimation procedure to the classical least squares regression. This works for situations in which the independent variates are subject to measurement errors and when there are high intercorrelations that cause the least squares procedures to break down. Furthermore, Scott (1966) points out that the method used in FA regression explicitly accounts for the errors as the "unique factors" in the FA model. More directly, the factor model has been employed by Isogawa and Okamoto (1980) as a basis for a prediction model in an effort to elucidate the effects of measurement errors in prediction. Lawley and Maxwell (1973) formulated an optimality problem for the factor analysis model and obtained the optimal biased and unbiased predictors as the solutions of the unrestricted and restricted problems, respectively.

Our main concern in this study is to develop a method of sensitivity analysis in FA regression for detecting influential observations. Here we consider situations where the problem is to predict an observable dependent variate from observable independent variates with measurement errors and where a factor analysis model holds for both of independent and dependent variates. The theoretical formulation given below is closely patterned after the work of Browne (1988). Detailed properties of the underlying model and the maximum likelihood estimates are studied by Lawley and Maxwell (1973) and Browne (1988). First, we derive the empirical influence functions $\widehat{\beta}_F^{(1)}$ and $[r_2(\widehat{\beta}_F)]^{(1)}$ for the FA regression weights $\widehat{\beta}_F$ and the squared multiple correlation $r_2(\widehat{\beta}_F)$. As an application of the proposed procedure, we try to conduct a numerical study to illustrate some fundamental properties of the influence in FA regression.

Linear Prediction

Let us consider a sample of n observations $x_i = (x_{1i}, x_{2i}^T)^T$, $i = 1, \ldots, n$, on a $p \times 1$ vector variable $\bar{x}_i = (\bar{x}_1, \bar{x}_2^T)^T$. Our concern is to predict the first variable x_1 by means of a linear function of the remaining variables represented by the $(p-1) \times 1$ vector x_2. Let $x = (x_1, x_2^T)^T$ and

$$S = \begin{pmatrix} s_{11} & s_{22}^T \\ s_{21} & S_{22} \end{pmatrix}$$

be the sample mean vector and covariance matrix. If $\widehat{\beta}$ is a $(p-1) \times 1$ vector of weights, the predicted value of x_{1i} obtained from x_{2i}, $i = 1, \ldots, n$, is given by

$$x_{1i} = (x_1 + \widehat{\beta}^T (x_{2i} - x_2). \tag{1}$$

The usual sample multiple regression weights derived by the ordinary least squares method (OLS) are obtained as

$$\hat{\beta} = S_{22}^{-1} s_{21}. \tag{2}$$

These weights yield the minimum

$$s_e^2(\hat{\beta}_F) = s_{11} - s_{22}^T S_{22}^{-1} s_{21} \tag{3}$$

of the residual variance and the maximum

$$r^2(\hat{\beta}_F) = 1 - s_e^2(\hat{\beta})/s_{11} \tag{4}$$

of the squared multiple correlation coefficient $r^2(\hat{\beta})$.

Factor Analysis Model

Now consider a factor analysis (FA) model for a $p \times 1$ observation vector x given by

$$\mathrm{x} = \mu + \Lambda f + e, \tag{5}$$

where μ is the mean vector, Λ is a $p \times k (k < p)$ factor loading matrix, f is a $k \times 1$ common factor score vector, and e is a $p \times 1$ unique factor score vector. And also, we assume

$$E(f) = 0, \quad E(e) = 0,$$
$$E(ff^T) = I, \quad E(ee^T) = \Psi, \quad E(fe^T) = 0,$$

Ψ denoting a diagonal matrix. Let us partition a $p \times k$ factor leading matrix Λ as $\Lambda = \lambda_1, \Lambda_2^T)^T$. The common factors $f_1, ..., f_k$ form a vector variate f where Λ_1^T is a row vector whose elements $\lambda_{11}, ..., \lambda_{1k}$ are the loadings of x_1 on the factors in f. The elements of the $(p-1) \times k$ matrix Λ_2 are the loadings of the variates in x_2 on f. Also, we express the corresponding diagonal matrix of unique variances in the form of $\Psi = \mathrm{diag}(\psi_{11}, \Psi_{22})$. Then, the covariance matrix derived from the FA model is expressed in terms of the loading matrix Λ and the unique variance matrix Ψ as

$$\Sigma = \Lambda\Lambda^T + \Psi \tag{6}$$

which is known as the common factor decomposition.

Using the factor decomposition the matrix Σ may be partitioned in the form of

$$\Sigma = \begin{pmatrix} \sigma_{11} & \lambda_1^T \Lambda_2^T \\ \Lambda_2 \lambda_1 & \Sigma_{22} \end{pmatrix}$$

where

$$\sigma_{11} = \lambda_1^T \lambda_1 + \psi_{11} \tag{7}$$

represents the variance of x_1 and

$$\Sigma_{22} = \Lambda_2 \Lambda_2^T + \Psi_{22} \tag{8}$$

is the covariance matrix of x_2.

Factor Analysis Regression

Consider the case in which it is not possible to measure the independent variates x_2 without error in which the dependent and independent variated x_1, x_2 follow a factor analysis model (5). Then, it is derived by Lawley and Maxwell (1973) and Browne (1988) that the regression weights of the best predictor in the sense to minimize $E(x_1 - \hat{x}_1)^2$ is given by the least squares estimate as

$$\hat{\beta}_F = \hat{\Sigma}_{22}^{-1} \hat{\sigma}_{21}^T \tag{9}$$

where $\hat{\Sigma}_{22}^{-1}$ and $\hat{\sigma}_{11}$ are obtained from equations (7) and (8) with the parameters replaced by their estimates.

Suppose that $\hat{\Lambda}$ and $\hat{\Psi}$ are obtained by the standard maximum likelihood method under the inequality constraints that the diagonal elements of $\hat{\Psi}$ are non negative and that the condition $\hat{\psi}_{11} \neq 0$ is satisfied. Then the residual variance is given by

$$s_e^2(\hat{\beta}_F) = \hat{\sigma}_{11} - \hat{\sigma}_{21}^T \hat{\Sigma}_{22}^{-1} \hat{\sigma}_{21} \tag{10}$$

and the squared multiple correlation coefficient between $x1$ *and* x_2 is expressed in the form of

$$r^2(\hat{\beta}_F) = 1 - \frac{s_e^2(\hat{\beta}_F)}{\hat{\sigma}_{11}}. \tag{11}$$

Sensitivity analysis

For sensitivity analysis, we make use of the influence functions related with FA regression. Our concern is to focus our attention on the two aspects of influence, namely influence on the FA regression weights $\hat{\beta}_F$ and influence on the squared multiple correlation coefficient $r^2(\hat{\beta}_F)$. Let us consider a perturbation on the empirical distribution $\hat{F}$ to $(1 - \varepsilon)\hat{F} + \varepsilon\delta_x$, where δ_x is the cdf of a unit point mass at x. Then, using the so called chain rule, we can obtain the empirical influence function for, $\hat{\beta}_F$ which is equivalent to the first derivative of $\hat{\beta}_F$ with respect to ε, as

$$\hat{\beta}_F^{(1)} = - \hat{\Sigma}_{22}^{-1} \hat{\Sigma}_{22}^{(1)} \hat{\Sigma}_{22}^{-1} \hat{\sigma}_{21} + \hat{\Sigma}_{22}^{-1} \hat{\sigma}_{22}^{(1)}, \tag{12}$$

where $\hat{\Sigma}_{22}^{(1)}$ and $\hat{\sigma}_{22}^{(1)}$ are obtained from the corresponding parts of $(\hat{\Lambda}\hat{\Lambda}^T)^{(1)} + \hat{\Psi}^{(1)}$, $(\hat{\Lambda}\hat{\Lambda}^T)$ and $\hat{\Psi}^{(1)}$ being the influence functions in the maximum likelihood factor analysis derived by Tanaka and Odaka (1989b). As a measure of the influence on the estimate $\hat{\beta}_F$, the vector-valued influence function $\hat{\beta}_F^{(1)}$ should be summarized into a scalar-valued quantity. The simplest war to summarize it is to tale the Eucidean norm of a vector $\hat{\beta}_F^{(1)}$, i.e., $\| \hat{\beta}_F^{(1)} \|$. Instead of taking Euclidian norm without any scaling we may scale or standardize before taking the norm. Here we consider two kinds of scaled norms such as univariately-scaled squared distance D_{us} and the multivariately-scaled squared distance D_{ms}.

$$D_{us}(\hat{\beta}_F) = (n-1)^{-1} \hat{\beta}_F^{(1)T} [\hat{V}_D(\hat{\beta}_F)]^{-1} \beta_F^{(1)}. \tag{13}$$

$$D_{ms}(\hat{\beta}_F) = (n-1)^{-2} \hat{\beta}_F^{(1)T} [\hat{V}_D(\hat{\beta}_F)]^{-1} \beta_F^{(1)}. \tag{14}$$

The matrices $\hat{V}$ and $(\hat{\beta}_F)$ and $\hat{V}_D$ are an estimated asymptotic covariance matrix of $\hat{\beta}_F$ and a diagonal matrix whose diagonal part consists of the diagonal part of $\hat{V}(\hat{\beta}_F)$, respectively. Actually, the quantity D_{ms} is a generalized version of the Cook's distance which is frequently applied in regression diagnostics. The asymptotic covariance matrix of $\hat{\beta}_F$, i.e., $\hat{V}(\hat{\beta}_F)$ can be derived from the asymptotic variances and covariances of $\hat{V}$ and $\hat{\Psi}$, which are evaluated in the paper of Tanaka and Watadani (1992), using the delta method. However, for simplicity to obtain $\hat{V}(\hat{\beta}_F)$ we utilize the jackknife technique in which $\hat{\beta}_{F(i)}$, the estimate based on the sample without the *i-th* observation, is replaced by its linear approximate using the empirical influence function. The covariance matrix based on the standard jackknife technique is given by (see, e.g., Fox, Hinkley and Larntz, 1980)

$$V_p = n(n-n)^{-1} \sum_{i=1}^{n} (P_i - P)(P_i - \bar{P})^T, \tag{15}$$

where the matrix V_p is the jackknife estimate of $\hat{V}(\hat{\beta}_F)$ and P_i are the vectors which represent the pseudo-values expressed in the form of

$$P_i = n\hat{\beta}_F - (n-1)\hat{\beta}_{F(i)} \tag{16}$$

with an average

$$\bar{P}_i = \hat{\beta}_J = n^{-1} \sum_{i=1}^{n} P_i, \; i = 1, 2, \ldots n. \tag{17}$$

As a measure of the influence on the goodness of fit, we consider the influence function for squared multiple correlation coefficient $r^2(\widehat{\beta}_F)$, which is derived in the following form.

$$[r^2(\widehat{\beta}_F)]^{(1)} = - \frac{[s_e^2(\widehat{\beta}_F)]^{(1)}}{\widehat{\sigma}_{11}} + \frac{s_e^2(\widehat{\beta}_F)\,\widehat{\sigma}_{11}^{(1)}}{(\widehat{\sigma}_{11})^2} \tag{18}$$

where $\widehat{\sigma}_{11}^{(1)}$ is obtained from the (1, 1) element of $(\widehat{\Lambda}\widehat{\Lambda}^T) + \widehat{\Psi}^{(1)}$ and

$$[s_e^2(\widehat{\beta}_F)]^{(1)} = \widehat{\sigma}_{11}^{(1)} - 2\widehat{\sigma}_{21}^{(1)T}\widehat{\Sigma}_{22}^{-1}\widehat{\sigma}_{21} + \widehat{\sigma}_{21}^T\widehat{\Sigma}_{22}^{-1}\widehat{\Sigma}_{22}^{(1)}\widehat{\Sigma}_{22}^{-1}\widehat{\sigma}_{21}. \tag{19}$$

Numerical study

To investigate the performance of our sensitivity analysis procedure, we apply our procedure to an artificial data set, which was generated on the basis of an example given in the paper of Lawley and Maxwell (1973). The example discusses the analysis of a sample of 200 students taking certain public examinations. The independent variables relate the examination obtained in the six subjects namely Gaelic, English, History, Arithmetic, Algebra and Geometry. The dependent variable represents the algebra examination score taken after two years. So the main purpose of the analysis is to see how this final algebra examination score (dependent variable) can be predicted from the scores obtained in the six earlier examinations (independent variables). The data of 100 individuals were generated by using the result of analysis with a two factor model of the correlation matrix given in the example of Lawley and Maxwell as

$$\begin{bmatrix} x_1 \\ x_2 \\ x_2 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{bmatrix} = \begin{bmatrix} 0.6929 & -0.2855 \\ 0.5405 & 0.4490 \\ 0.5576 & 0.3103 \\ 0.3825 & 0.4474 \\ 0.6971 & -0.1852 \\ 0.7376 & -0.2049 \\ 0.6411 & -0.1697 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} 0.4384e_1 \\ 0.5062e_2 \\ 0.5928e_3 \\ 0.6535e_4 \\ 0.4797e_5 \\ 0.4140e_6 \\ 0.5602e_7 \end{bmatrix}$$

where each of $(f_1, f_2, e_1, \ldots, e_6)$ follows independently $N(0, 1)$ distribution. Note that all of the variables have positive loadings on the first factor. This factor reflects the overall response of the students to instruction and might by labeled as "general intelligence" factor. Half of the loadings are positive and half are negative on the second factor. This factor is not easily identified, but is such that the individuals who get above-average scores on the verbal tests get above-average scores on the factor. Individuals with above-average scores on the mathematical tests get below-average scores on the factor. This may be labeled as "math-nonmath"

factor. The correlation matrix of examination scores in six subject areas and the result of FA regression are shown in Table 1 and 2, respectively. It is clearly indicated from Table 2 that the first three regression weights are negligibly small, and therefore the examination scores on the three nonmathematical subjects do not contribute much towards the prediction of x_1.

To analyze some fundamental properties of the influence in FA regression,

Table 1. Correlation matrix.

Variables	X_1	X_2	X_3	X_4	X_5	X_6	Y
X_1	1.000	0.439	0.410	0.288	0.329	0.248	0.249
X_2		1.000	0.351	0.354	0.320	0.329	0.284
X_3			1.000	0.164	0.190	0.181	0.146
X_4				1.000	0.595	0.470	0.495
X_5					1.000	0.464	0.572
X_6						1.000	0.539
Y							1.000

Table 2. Result of the FA regression.

	Data of f1-perturbation	St. error	Data of e6-perturbation	St. error
Variables	$(\widehat{\beta}_F)$	$SE(\widehat{\beta}_F)$	$(\widehat{\beta}_F)$	$SE(\widehat{\beta}_F)$
1	-0.0521	0.0674	-0.0561	0.0726
2	0.0670	0.0378	0.1218	0.0444
3	0.0339	0.0426	0.0586	0.0406
4	0.1622	0.0444	0.2449	0.0954
5	0.5026	0.0973	0.2225	0.1419
6	0.1470	0.0443	0.2238	0.0741

we put into two types of unusual data and make two artificial data sets with different types of perturbation. The first data set contains 99 ordinary observations and one observation (No. 70) with extraordinarily large value of $f_1 (f_1 = 10)$. On the other hand, the second data set is composed of 99 ordinary observations and one observation (again No. 70) with extraordinarily large value of $e_6 (e_6 = 10)$.

At first, we applied FA regression and the sensitivity analysis procedure to the first data set with an observation of large f_1. To investigate the influence of every individual, we have computed influence measures $\|\widehat{\beta}_F^{(1)}\|$, scaled distances D D_{us} and D_{ms} as well as the measure on the squared multiple correlation coefficient

$[r^2(\widehat{\beta}_F)]^{(1)}$. The index plots of those measures are shown in Figure 1. As seen from the figure, individual No. 70 which possesses an extraordinarily large value of f_1 is not at all influential to the regression coefficients $\widehat{\beta}_F$, but is has extremely large influence on the squared multiple correlation coefficient. It is noticed that the effect of f_1 perturbation appears only in the $[r^2(\widehat{\beta}_F)]^{(1)}$ not in $\|\widehat{\beta}_F^{(1)}\|$, D_{us} and D_{ms}.

Next, we analyzed the second data set with an observation of large e_6. The index plots in Figure 2 show the influence measures of every observation. As shown in those index plots, there exists one observation, the 70-*th* individual, which has much extremely large influence on $\widehat{\beta}_F$, but the influence of the same observation on the squared multiple correlation coefficient is not so extreme. We also analyzed the case of e_3-perturbation, the results of which are not shown here for saving space. The effect of e_3-perturbation has appeared more strongly in $\|\widehat{\beta}_F^{(1)}\|$, D_{us} and D_{ms} and not so much in $[r^2(\widehat{\beta}_F)]^{(1)}$.

Discussion

For detecting influential observations, we have derived influence functions for the parameters in the FA regression and proposed some measures of influence on the estimate of regression weights $\widehat{\beta}_F$ and of influence on the squared multiple correlation coefficient $r^2(\widehat{\beta}_F)$. In the present paper, we illustrated the present procedure when some perturbations are introduced on the data set. In the numerical study, we have considered two types of perturbation to characterize the fundamental properties of influence from the sensitivity perspective.

In the case of f-perturbation the effect appeared strongly in the measure of influence on $r^2(\widehat{\beta}_F)$ but not at all appeared in the measures of influence on $\widehat{\beta}_F$. This can be explained by the model as follows. The observation with extremely large f_1 fits the FA regression model well. So as the correlation structure of the variables does not change essentially, the residual variance, the numerator of the second tern of the right side of (11), is not affected much. But $\widehat{\sigma}_{21}^{(1)}$, the denominator of the same part, obviously increases. Therefore, $\widehat{r^2}(\beta_F)$ improves very much if this type of perturbation is introduced.

On the other hand, the effect of e-perturbation appeared strongly in the measures of influence on $\widehat{\beta}_F$ and also appeared, though not so strongly, in the measure of influence on $r^2(\widehat{\beta}_F)$. In this case, the perturbed observation is located far from the regression hyperplane because of the large e-value. So it is obvious that the regression weights $\widehat{\beta}_F$ are affected strongly by this observation.

Now, among the possible sample versions of the influence functions the empirical influence function *EIF* is the easiest to compute. To check the validity of the proposed procedure based on *EIF*, we investigate numerically its relationship to the sample influence function *SIF*. For this purpose, we omit a single observation in turn and perform the FA regression analysis n times. Then, we calculate the following sample influence function *SIF* which is defined as

$$SIF_i(\widehat{\beta}_F) = (n-1)\,\widehat{\beta}_{F(i)} - \widehat{\beta}_F)\; i = 1, 2, \ldots n$$

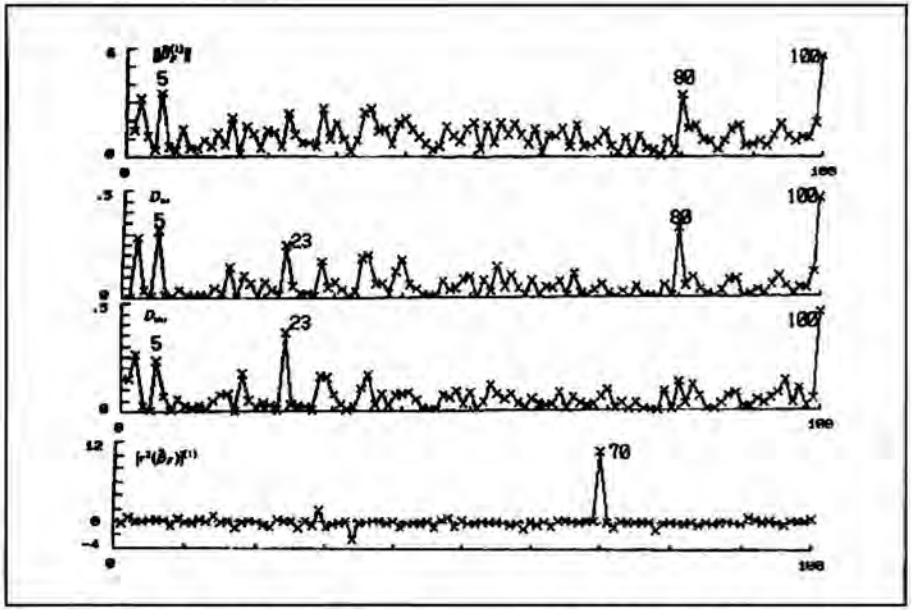

Figure 1. Index plot of four influence measures (Artificial data set of f_1-perturbation.)

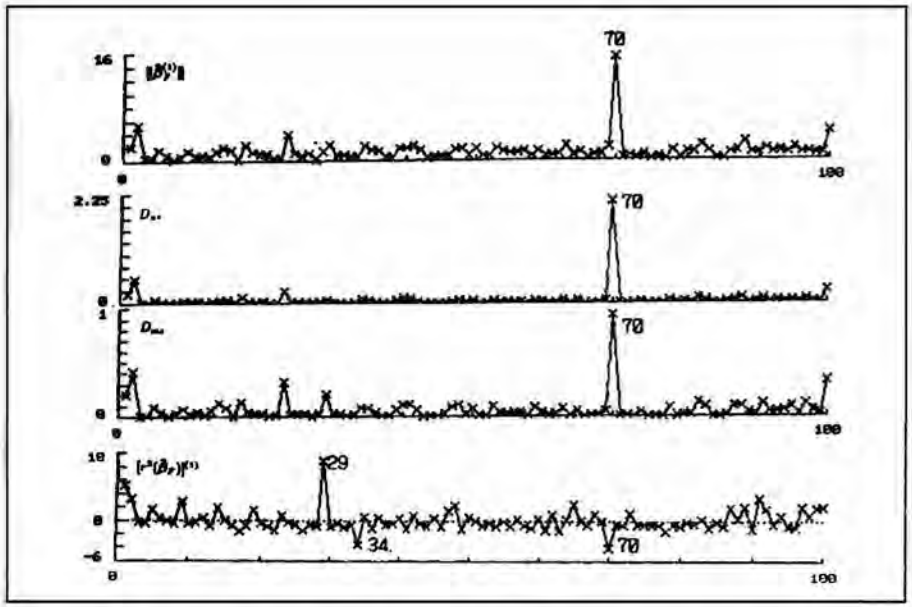

Figure 2. Index plot of four influence measures (Artificial data set of e_6-perturbation.)

where $\widehat{\beta}_F$ and $\widehat{\beta}_{F(i)}$ denote the estimated parameter vectors based on the sample with and without the *i-th* observation, respectively. Scatter diagrams of the statistics $\|\widehat{\beta}_F^{(1)}\|$, D_{us}, D_{ms} and $[r^2(\widehat{\beta}_F)]^{(1)}$ versus the corresponding statistics based on *SIF* are given in Figure 3 and Figure 4. In these scatter diagrams, most of the points are located near the straight line *SIF* = *EIF*. These results suggest that the correspondence is good enough to imply that the *EIF*, which is based on the differential coefficient, can be used practically for detecting influential observations instead of the *SIF*. Besides, in terms of computing time, *SIF* requires much time to compute than *EIF*.

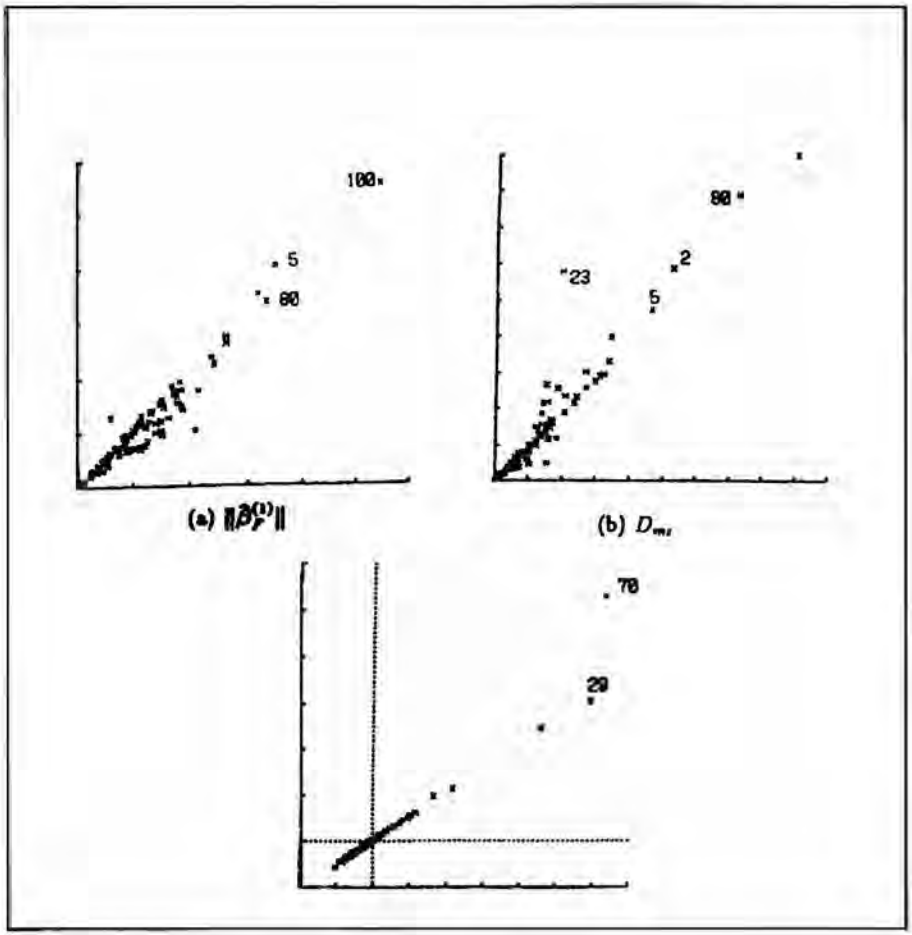

Figure 3. Scatter diagrams of three influence measures based on the EIF (horizontal) versus SIF (vertical) Artificial data set of f_1-perturbation

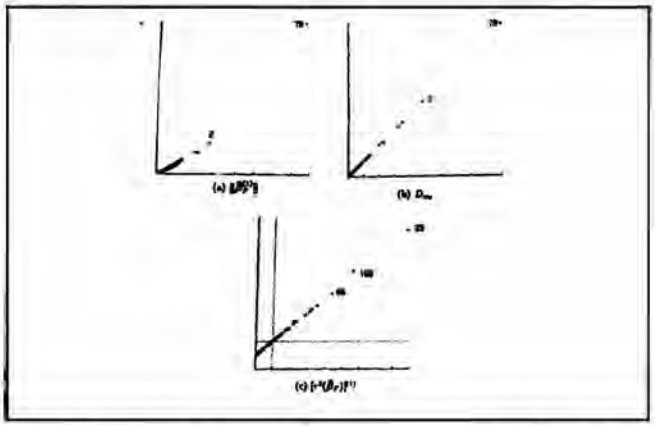

Figure 4. Scatter diagrams of three influence measures based on the EIF (horizontal) versus SIF (vertical) Artificial data set of e_6-perturbation

REFERENCES

Browne, M. W. (1988). Properties of the maximum likelihood solution in factor analysis regression. *Psychometrika*, **53**, 585-589.

Fox, T., Hinckley, D., and Larntz, K. (1980). Jackknifing in nonlinear regression. *Technometrics*, **22**, 29-33.

Horst, P. (1941). *Approximating a multiple correlation system by one of lower rank as a basis for deriving more stable prediction weights.* In P. Horst (Ed.), New York: Social Research Council.

Isogawa, Y., and Okamoto, M., (1980). Linear prediction in the factor analysis model, *Biometrika*, **67**, 482-484.

Joreskog, K. G., (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, **32**, 443-484.

King, B. (1969). Comment on factor analysis and regression. *Econometrica*, **37**, 538-540.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor analysis as a statistical method* (2nd ed.). London: Butterworth.

Lawley, D. N. and Maxwell, A. E. (1973). Regression and factor analysis. *Biometrika*, **60**, 331-338.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press.

Scott, J. T. (1966). Factor analysis regression revisited. *Econometrica*, **37**, 719.

Tanaka, Y., Castaño-Tostado, E. and Odaka, Y. (1989c). Sensitivity analysis in factor analysis: Methods and software. *COMPSTAT 1990* (Edited by Momirovic, K, and Mildner, V.), Physica-Verlag, 205-10,

Tanaka, Y. and Odaka (1989a). Influential observations in principal factor analysis. *Psychometrika*, **54**, 4067-84.

Tanaka, Y. and Odaka (1989b). Sensitivity analysis in maximum likelihood factor analysis. *Comm. Statist.*, **A18**, 4067-84.

Tanaka, Y., Tarumi, T. and Wakimoto, K. (1984). *Handbook of the Statistical Analysis Publishing with BASIC Programs for Personal Computers II Multivariate analysis.* Kyoritsu Publishing Company. (Text in Japanese).

Tanaka, Y. and Watadani, S. (1992) Sensitivity analysis in covariance structure analysis with equality constraints. *Comm. Stat.*, **A21**, 1501-1515.