# On Multiple-Case Diagnostics in Multivariate Methods


**Yutaka Tanaka[1], Shingo Watadani[2], Yoshihiro Yamanishi[3]**

[1] Department of Environmental and Mathematical Science, Okayama University, Tsushima, Okayama 700-8530, Japan.<tanaka@stat.ems.okayama-u.ac.jp>

[2] Department of Liberal Arts and Science, Kurashiki University of Science and the Arts, 2640 Nishinoura Tsurajima-cho, Kurashiki, 712-8505, Japan. <wat@las.kusa.ac.jp>

[3] Graduate School of Natural Science and Technology, Okayama University, Tsushima, Okayama 700-8530, Japan.<yoshi@stat1.stat.ems.okayama-u.ac.jp>


## 1 Introduction

Methods of detecting singly influential observations have been well established not only in regression type analysis but also other multivariate methods (see, e.g., Cook and Weisberg, 1982; Critchley, 1985; Tanaka and Odaka, 1989; Tanaka and Watadani, 1992; Wang and Lee, 1996). However, only limited results are available on detecting jointly influential observations in multivariate methods other than regression analysis.

We have proposed so far to apply principal component analysis (PCA) to the influence functions for detecting influential subsets of observations in multivariate methods where influence functions are available (e.g., Tanaka, Castano-Tostado and Odaka, 1990; Tanaka, 1994; Tanaka and Zhang, 1999; Mori et al., 2000). It works well in ordinary circumstances, but it may sometimes fail, if the so-called masking effect is serious. Tanaka and Watadani (1994) proposed a method for protecting from the masking effect. In the present paper we try to improve the method, and study precisely in what conditions the phenomenon of masking happens and in what conditions our method works well.

## 2 General Procedure based on PCA of influence functions

Here we consider the influence of a set of $k$ observations $A = \{\underline{x}_{i_1}, \cdots, \underline{x}_{i_k}\}$ 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 $\tilde{F} = (1-\epsilon)F + \epsilon G$, 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 of $A$ as the limit $TIF(A;\underline{\theta}) = \lim_{\epsilon\to 0}[\underline{\theta}(\tilde{F}) - \underline{\theta}(F)]/\epsilon$. Then it can be verified that $TIF(A;\underline{\theta}) = k^{-1}\sum_{\underline{x}_i \in A} TIF(\underline{x}_i;\underline{\theta})$, where $TIF(\underline{x}_i, \underline{\theta})$ is the ordinary influence function of $\underline{x}_i$. The similar relation holds for the empirical influence function ($EIF$), which is defined by replacing the cdf $F$ by the empirical cdf $\hat{F}$ in the definition of $TIF$. Hence the parameter estimate based on the sample with a subset $A$ omitted can be approximated as $\hat{\underline{\theta}}_{(A)} \cong \tilde{\underline{\theta}}_{(A)} \equiv \underline{\theta} - (n-k)^{-1}\sum_{\underline{x}_i \in A} EIF(\underline{x}_i;\hat{\underline{\theta}})$, where symbol$\tilde{()}$ indicates the linear approximation based on the $EIF$. Thus there exists an additivity property of the influence of observations, when up to the first order derivatives are taken into account. This relation suggests that we can detect influential subsets of observations by searching for observations which have relatively large EIF vectors with similar directions from the origin. For this purpose we can apply PCA to the EIF vectors. Here we should introduce metric $V^-$ to adjust the correlations among the components of $\hat{\underline{\theta}}$, where $V$ is the estimated asymptotic covariance matrix of $\hat{\underline{\theta}}$.

**Step 1.** Compute the $EIF$ vectors, $EIF(\underline{x}_i;\hat{\underline{\theta}}), i = 1, \cdots, n$

**Step 2.** Summarize the $EIF$ vectors into scalar influence measures from various aspects such as the influence on the estimate $\hat{\underline{\theta}}$, on its precision and on the goodness of fit. Find observations which are individually influential.

**Step 3.** Search for subsets of obervations whose members have relatively large $EIF$ vectors with similar directions using PCA with metric $V^-$.

Among possible influence measures are the generalized Cook's $D$ defined by $D_i = (\underline{\tilde{\theta}}_{(i)} - \underline{\hat{\theta}})^T V^- (\underline{\tilde{\theta}}_{(i)} - \underline{\theta})$ and the $COVRATIO$ like measure defined by $CVR_i = |V_{(i)}|/|V|$, where $\underline{\tilde{\theta}}_{(i)}$ and $V_{(i)}$ are approximate estimates for $\underline{\theta}$ and its covariance matrix based on the sample without the $i$-the observation.

The above general procedure is closely related to Cook(1986)'s local influence and Lu, Ko & Chang(1997)'s PCA of standardized influence matrix. As discussed by Tanaka (1994) and Tanaka and Zhang (1999) and also mentioned in the companion paper in this session, the PC scores of the above PCA are equal to the eigenvectors of Cook's local influence up to the normalizing constants and, in particular, the PC scores associated with the largest eigenvalue give the most influential direction in the sense of Cook's local influence. Moreover, it can be proved that the Cook's distance $D_i$ for the $i$-th individual is decomposed as $D_i = z_{1i}^2 + z_{2i}^2 + \cdots + z_{mi}^2$, where $z_{ki}$ is the $k$-th PC score of the $i$-th individual. In this sense PCA with metric $V^-$ provides the information on multivariate structure of the influence whose overall influence is measured with Cook's D (see, Tanaka and Zhang, 1999).

# 3 Robust procedure

The general procedure in the previous section works well as demonstrated in the companioin paper (Mori et al.,2000) in ordinary circumstances. However, there exists a possibility of the so-called masking effect.

Generally in most multivariate mothods the $EIF$ for $\underline{\hat{\theta}}$ at $\underline{x}_i$ is obtained as a linear function of the elements of $EIF(\underline{x}_i, S)$, where $S$ is the sample covariance matrix defined by $S = n^{-1}\sum_i(\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})^T$ and $EIF(\underline{x}_i, S)$ is given by $EIF(\underline{x}_i, S) = (\underline{x}_i - \underline{\bar{x}})(\underline{x}_i - \underline{\bar{x}})^T - S$. We can easily imagine that neither $EIF(\underline{x}_i, S)$ nor $EIF(\underline{x}_i, \underline{\hat{\theta}})$ reflect correctly the influences on the covariance matrix and on the estimated parameters, if $\underline{\bar{x}}$ and/or $S$ are quite different from the corresponding population values, and therefore we may meet the phenomenon of masking effect.

It is obvious that this effect can be protected by using an appropriate robust method for estimating the mean vector and covariance matrix. Tanaka and Watadani(1994) applied the minimum volume ellipsoid (MVE) method with one-step improvement (Rousseeuw and Leroy(1987), p.260). In the present paper we try to improve the procedure by incorporating adding back procedure in the confirmatory stage following the idea of Atkinson(1986) and Fung(1993). The procedure is summarized as follows.

**Step 1.** Using the MVE method followed by the iterative process of one-step improvement (Rousseeuw and Leroy, 1987; Rousseeuw and van Zomeren, 1990), assign unit or zero weight to each observation depending whether $(\underline{x}_i - \underline{T}(X))(\underline{x}_i - \underline{T}(X))^T < c$, or not, where $c$ is taken to be, say, 97.5 percent point of a chi-squared distribution of $m$ degrees of freedom. Regard the observations with unit weight as observations which do not contain any influential observations. We call those observations as "active observations" and the remaining ones as "supplementary observations".

**Step 2.** Based only on the active observations compute the mean vector $\underline{\bar{x}}_R$ and the covariance matrix $S_R$. Then, obtain the parameter estimate $\underline{\hat{\theta}}_R$.

**Step 3.** Based only on the active observations compute the coefficient matrix in the relation between $EIF(\underline{x}_i; S_R)$ and $EIF(\underline{x}_i; \underline{\hat{\theta}}_R)$.

**Step 4.** Compute the $EIF$ for the covariance matrix not only for the active observations but also for the supplementary observations in such a way that $EIF(\underline{x}_i; S) = (\underline{x}_i - \underline{\bar{x}}_R)(\underline{x}_i - \underline{\bar{x}}_R)^T - S_R$, and compute the corresponding $EIF$ for $\underline{\hat{\theta}}$ using the linear relation obtained in Step 3.

**Step 5.** Summarize the obtained $EIF$ vectors into Cook's $D$ and find candidates for influential observations. Apply a bootstrap test to the computed Cook's $D$. If there exists any non-significant observation among the supplementary observations, we add them back from the set of supplementary observations to that of active observations. On the other hand, if there is any significant observation among the active observations, we discard them from the set of active observations to that of the supplementary observations. Then go back to Step 2.

## 4 Masking in PCA

Lawrance(1995) studied precisely the notion of masking in regression analysis. He considered conditional as well as joint influences and introduced masking/boosting effects from the perspective of conditional influence and enhancing/reducing/swamping effects from the perspective of joint influence.

In the present paper we discuss those effects in multivariate methods, in particular, in PCA. Our interest also extends to the conditions when the additivity of influence holds to a degree of approximation we can utilize in practice for detecting influential subsets of observations.

Example. Let us analyze the soil composition data (Kendall, 1975, Tables 2.1, 2.4), which have been used so far for demonstrating the performance of influence analysis by several authors including Critchley(1985) and Tanaka(1988). The proportions of eigenvalues of the covariance matrix are 0.9161, 0.0750, 0.0049, 0.0040 in the order of their magnitudes, and it is suggested the data points lie almost on a 2-dimensional subspace. There are three observation ♯4, ♯8 and ♯9, which have large 1st PC scores. Among them ♯8 and ♯9 have masking effect with each other and ♯4 has boosting effect to ♯8 and ♯9 in the sense of Lawrance(1995). However, we can find that the linear approximation based on the $EIF$ gives good approximates to the effects of deleting every two observations by looking at the scatter diagram. We will modify the data set to study the conditions when the additive relation does not hold.

## References


[1] Atkinson, J. (1986). Masking unmasked. *Biometrika, 73, 533-541.*

[2] Barrett, B. E. and Gray, J. B. (1997). Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression. *Computational Statistics & Data Analysis 26 (1997) 39-52.*

[3] Cook, R. D. (1986). Assessment of local influence. *J. R. Statist. Soc., B48, 133-169.*

[4] Cook, R.D. and Weisberg, S. (1982). Residuals and influence in regression. *Chapman and Hall.*

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

[6] Fung, W. K. (1993). Unmasking outliers and leverage points: a confirmation. *J. Am. Statist. Assoc., 88, 515-519*

[7] Lawrance, A. J.(1995). Deletion influence and masking in regression. *J.R. Statist. Soc. B57, 181-189.*

[8] Lee, A. H. and Fung, W. K. (1997). Confirmation of multiple outliers in generalized linear and nonlinear regressions. *Computational Statistics & Data Analysis 25 (1997) 55-65.*

[9] Lu, J., Ko, D., and Chang, T. (1997). The standardized Influence matrix and its applications. *J. Am. Statist. Assoc., 92, 1572-1580.*

[10] Mori,Y., Watadani,S., Yamamoto,Y., Tarumi,T., and Tanaka,Y. (2000). Statistical software SAMMIF for sensitivity analysis in multivariate methods. *International Conference on Measurement and Multivariate Analysis (ICMMA).*

[11] Rousseeuw, P. J., and van Zomeren, B. C.(1990). Unmasking multivariate outliers and leverage points. *J. Am. Statist. Assoc., 85, 633-639.*

[12] Tanaka, Y. (1988). Sensitivity analysis in principal component analysis: Influence on the subspace spanned by principal components. *Comm. Statist., A17, 3157-3175. (Corrections, A18 (1989), 4305).*

[13] Tanaka, Y. (1994). Recent advance in sensitivity analysis in multivariate methods. *J. Jpn. Soc. Comp. Statist., 7, 1-25.*

[14] Tanaka, Y., Castano-Tostado, E. and Odaka, Y. (1990). Sensitivity analysis in factor analysis: Methods and software. In: *COMPSTAT90 Proceedings in Computational Statistics (ed. Momirovic, K. & Mildner, V.) , 205-210. Physica-Verlag.*

[15] Tanaka, Y. and Odaka, Y.(1989). Influential observations in principal factor analysis. *Psychometrika. 54, 475-485.*

[16] Tanaka, Y. & Watadani, S. (1994). Unmasking influential observations in multivariate methods. In: *COMPSTAT94 Proceedings in Computational Statistics (ed. R.Dutter & W.Grossman), 292-297. Physica-Verlag.*

[17] Wang, S.-J. & Lee, S.-Y.(1996). Sensitivity analysis of structural equation models with equality functional constraints. *Comp. Statist. & Data Analysis, 23, 239-256.*