Journal of the Korean
Statistical Society
Vol. 25, No. 1, 1996

Local Influence on Misclassification Probability

Myung Geun Kim [1]

Abstract

The local behaviour of the surface formed by the perturbed maximum likelihood estimator of the squared Mahalanobis distance is investigated. The study of the local behaviour allows a simultaneous perturbation on the samples of interest and it is effective in identifying influential observations.

Key Words : Local influence; Mahalanobis distance; Misclassification probability; Perturbation.

1. INTRODUCTION

In linear discriminant analysis, diagnostic methods based on the influence function and case deletion have been suggested for identifying influential observations. Campbell (1978) used the influence function for detecting outliers

[1] Department of Applied Statistics, Seowon University, 231, Mochung-Dong, Cheongju, Chung-Buk, 360-742

in linear discriminant analysis. Critchley and Vitiello (1991) considered the case deletion method to investigate the influence of observations on the misclassification probability.

Lawrance (1988) and Tsai and Wu (1992) used the local influence method inspired by Cook (1986) to get information about influential cases in a regression problem. They considered the maximum slope of a path on the surface formed by the perturbed maximum likelihood estimator of a parameter of interest and its associated direction vector. This method allows a simultaneous perturbation affecting all cases, which is totally different from the influence function approach and the case deletion method. Lawrance (1988) showed that the simultaneous perturbation is more effective than the individual perturbation.

In this work the local influence method suggested by Lawrance (1988) is adapted to linear discriminant analysis. The local behaviour of the surface formed by the perturbed maximum likelihood estimator of the squared Mahalanobis distance is investigated, and it leads us to investigating the influence of observations on the misclassification probability. This method allows a simultaneous perturbation on the samples of interest. An illustrative example is given.

2. A LOCAL INFLUENCE MEASURE

Two independent random samples $\mathbf{x}_1,\ldots,\mathbf{x}_{n_1}$ and $\mathbf{x}_{n_1+1},\ldots,\mathbf{x}_n$ $(n = n_1+n_2)$ are drawn from p-variate normal distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ with a common covariance matrix, respectively. The maximum likelihood estimators of μ_1, μ_2 and Σ are the usual sample mean vectors $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$ and the pooled sample covariance matrix $\mathbf{S}$ with its divisor n, respectively.

For computational convenience, we use a reparametrization of Σ. Let $\mathbf{A}$ be a lower triangular matrix with positive diagonal elements such that $\Sigma = \mathbf{A}\mathbf{A}^T$. Let $\mathbf{B}^T = \mathbf{A}^{-1}$. Then $\mathbf{B}$ is an upper triangular matrix satisfying the identity $\Sigma^{-1} = \mathbf{B}\mathbf{B}^T$. We denote by $\mathbf{b}_i$ the ith column of $\mathbf{B}$. The degree of separation between the two populations can be measured by the squared Mahalanobis distance

$$\Delta^2 = \sum_{i=1}^{p}\left[\mathbf{b}_i^T(\mu_1 - \mu_2)\right]^2.$$

The performance of the Fisher's linear discriminant function relies on Δ^2 through the misclassification probability $\Phi(-\frac{1}{2}\Delta)$. Let $\hat{\Delta}$ be the maximum likelihood estimator of Δ.

We consider a simultaneously perturbed model specified by a perturbation vector $\mathbf{w} = (w_1, ..., w_n)^T$ in which

$$\mathbf{x}_r \sim N(\mu, \Sigma/w_r),$$

where μ equals μ_1 or μ_2 according to the value of r. The n by 1 vector with all elements equal to 1 is written as $\mathbf{1}_n$. Then the perturbation vector can be represented by $\mathbf{w} = \mathbf{1}_n + a\mathbf{l}$, where a indicates the magnitude of the perturbation and $\mathbf{l} = (l_1, ..., l_n)^T$ of unit length its direction. When $a = 0$, that is $\mathbf{w} = \mathbf{1}_n$, the perturbed model reduces to the unperturbed model. The maximum likelihood estimator of Δ under the perturbed model is denoted by $\hat{\Delta}(\mathbf{w})$. Then the $n+1$ vector $(\mathbf{w}^T, \hat{\Delta}^2(\mathbf{w}))^T$ forms a surface in the $n+1$ dimensional Euclidean space. A path on the surface at $a = 0$ and its slope are considered for investigating the local behaviour of observations. The partial derivative of $\hat{\Delta}^2(\mathbf{w})$ with respect to a evaluated at $a = 0$ is the slope of the path given by

$$\left.\frac{\partial \hat{\Delta}^2(\mathbf{w})}{\partial a}\right|_{a=0} = \sum_{r=1}^{n} l_r \left[\sum_{i=1}^{q} \left.\frac{\partial \Delta^2}{\partial \theta_i}\right|_{\theta=\hat{\theta}} \left.\frac{\partial \hat{\theta}_i(\mathbf{w})}{\partial w_r}\right|_{\mathbf{w}=\mathbf{1}} \right], \qquad (2.1)$$

where the θ_i are the parameters μ_1, μ_2 and $\mathbf{b}_r$ (nonzero components of it) in the model, and the $\hat{\theta}_i(\mathbf{w})$ represent the maximum likelihood estimators of them under the perturbed model. Denoting the coefficient of l_r in (2.1) by c_r, the rth component of the direction vector $\mathbf{l}_{max}$ maximizing the slope is easily obtained as $c_r/(\sum_{i=1}^{n} c_i^2)^{1/2}$ by the Cauchy-Schwarz inequality and it yields information about influential observations. Observations corresponding to significantly large elements of $\mathbf{l}_{max}$ in their absolute values need special attention. In the next section, we briefly describe a procedure for obtaining $\mathbf{l}_{max}$.

3. A PROCEDURE FOR GETTING A LOCAL INFLUENCE MEASURE

We write as $\mathbf{q}_{(i+)}$ the i by 1 vector consisting of the first i elements of a vector $\mathbf{q}$. The partial derivatives of Δ^2 with respect to the parameters are

$$\begin{aligned}\frac{\partial\Delta^2}{\partial\mu_1} &= 2\sum_{i=1}^{p}[\mathbf{b}_i^T(\mu_1-\mu_2)]\mathbf{b}_i\\ \frac{\partial\Delta^2}{\partial\mu_2} &= -2\sum_{i=1}^{p}[\mathbf{b}_i^T(\mu_1-\mu_2)]\mathbf{b}_i\\ \frac{\partial\Delta^2}{\partial\mathbf{b}_{i(i+)}} &= 2[\mathbf{b}_i^T(\mu_1-\mu_2)](\mu_1-\mu_2)_{(i+)}.\end{aligned}$$

Let $\mathbf{w}_1$ be the n_1 by 1 vector formed by the first n_1 elements of $\mathbf{w}$. The maximum likelihood estimators of the parameters under the perturbed model are given by

$$\begin{aligned}\hat{\mu}_1(\mathbf{w}) &= \sum_{j=1}^{n_1} w_j\mathbf{x}_j/\sum_{j=1}^{n_1} w_j\\ \hat{\mu}_2(\mathbf{w}) &= \sum_{j=1}^{n_2} w_{n_1+j}\mathbf{x}_{n_1+j}/\sum_{j=1}^{n_2} w_{n_1+j}\end{aligned}$$

and $\hat{\mathbf{B}}(\mathbf{w})$ satisfies the following identity

$$\hat{\mathbf{B}}(\mathbf{w})^T\mathbf{S}(\mathbf{w})\hat{\mathbf{B}}(\mathbf{w}) = \mathbf{I}_p, \tag{3.2}$$

where $\mathbf{S}(\mathbf{w}) = (1/n)[\mathbf{X}_1\mathbf{H}_1\mathbf{X}_1^T+\mathbf{X}_2\mathbf{H}_2\mathbf{X}_2^T]$, $\mathbf{X}_1 = (\mathbf{x}_1, ..., \mathbf{x}_{n_1})$, $\mathbf{H}_1 = \text{diag}(w_1, ..., w_{n_1}) - \mathbf{w}_1\mathbf{w}_1^T/\mathbf{1}_{n_1}^T\mathbf{w}_1$, and $\mathbf{X}_2$ and $\mathbf{H}_2$ are similarly defined.

The partial derivatives of the perturbed maximum likelihood estimators with respect to w_r evaluated at $\mathbf{w} = \mathbf{1}$ are

$$\begin{aligned}\left.\frac{\partial\hat{\mu}_1(\mathbf{w})}{\partial w_r}\right|_{\mathbf{w}=\mathbf{1}} &= \frac{1}{n_1}(\mathbf{x}_r-\bar{\mathbf{x}}_1) \quad (1\le r\le n_1)\\ \left.\frac{\partial\hat{\mu}_2(\mathbf{w})}{\partial w_{n_1+r}}\right|_{\mathbf{w}=\mathbf{1}} &= \frac{1}{n_2}(\mathbf{x}_{n_1+r}-\bar{\mathbf{x}}_2) \quad (1\le r\le n_2)\end{aligned}$$

and the other cases for the $\hat{\mu}_r$ are zero. Finally, the partial differentiation of both sides in (3.2) with respect to w_r shows that $\partial\hat{\mathbf{B}}(\mathbf{w})/\partial w_r\big|_{\mathbf{w}=\mathbf{1}}$ satisfies the following identity

$$\left(\left.\frac{\partial\hat{\mathbf{B}}(\mathbf{w})}{\partial w_r}\right|_{\mathbf{w}=\mathbf{1}}\right)^T\hat{\mathbf{A}} + \hat{\mathbf{A}}^T\left(\left.\frac{\partial\hat{\mathbf{B}}(\mathbf{w})}{\partial w_r}\right|_{\mathbf{w}=\mathbf{1}}\right) = -\frac{1}{n}\hat{\mathbf{B}}^T(\mathbf{x}_r-\bar{\mathbf{x}})(\mathbf{x}_r-\bar{\mathbf{x}})^T\hat{\mathbf{B}},$$

where $\bar{\mathbf{x}}$ equals $\bar{\mathbf{x}}_1$ for $1 \leq r \leq n_1$ and $\bar{\mathbf{x}}_2$ for $n_1 + 1 \leq r \leq n$. The solution to the above equation is obtained by considering a system of linear equations for solving $\mathbf{U}$

$$\mathbf{U}^T\mathbf{V} \ + \ \mathbf{V}^T\mathbf{U} \ = \ \mathbf{C},$$

where $\mathbf{U} = (u_{ij})$ is an upper triangular matrix, $\mathbf{V} = (v_{ij})$ a lower one and $\mathbf{C} = (c_{ij})$ is a symmetric matrix. Then $\mathbf{U}^T\mathbf{V}$ is a lower triangular matrix and $\mathbf{V}^T\mathbf{U}$ an upper one so that the solution $\mathbf{U}$ can be found as follows.

STEP I Compute $u_{ii} = c_{ii}/(2v_{ii})$ for each $i = 1, ..., p$.

STEP II For each i from $p-1$ to 1, compute $u_{ij} = (c_{ij} - \sum_{r=i+1}^{j} v_{ri}u_{rj})/v_{ii}$ $(j = i + 1, ..., p)$.

Putting the results above into (2.1) leads to finding $\mathbf{l}_{max}$. A closed form of $\mathbf{l}_{max}$ is too complicated to write down.

4. NUMERICAL EXAMPLE

The method of local influence is applied to the Flea-Beetles data (Seber, 1984, p.295) which have measurements on four variables for two species of flea beetles. Nineteen observations are obtained for the first species and twenty observations for the second species. The observations are labelled as 1 to 19 for the first species and 20 to 39 for the second species.

The index plot of $\mathbf{l}_{max}$ is shown in Fig. 1. Observations 27 and 36 have high local influence on the misclassification probability, and they are possible candidates for outliers. They have the same sign in the direction vector $\mathbf{l}_{max}$. Note that observation 27 is misclassified using the Fisher's linear discriminant function.

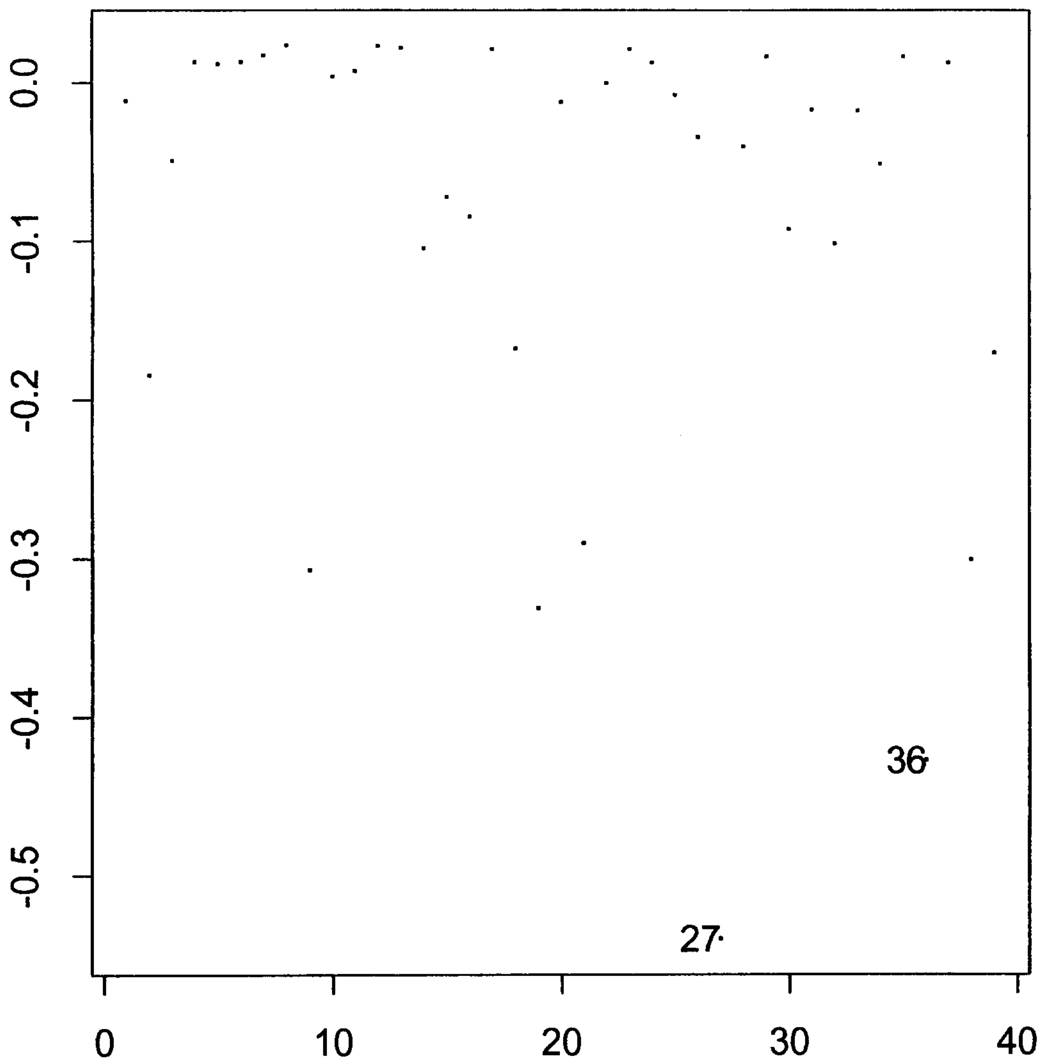

Figure 1. Index plot of the maximum direction vector

REFERENCES

(1) Campbell, N. A. (1978), The influence function as an aid in outlier detection in discriminant analysis, *Applied Statistics*, **27**, 251–258.

(2) Critchley, F. and C. Vitiello (1991), The influence of observations on misclassification probability estimates in linear discriminant analysis, *Biometrika* **78**, 677–690.

(3) Cook, R. D. (1986), Assessment of local influence (with discussions), *Journal of the Royal Statistical Society Sec. B* **48**, 133–169.

(4) Lawrance, A. J. (1988), Regression transformation diagnostics using local influence, *Journal of the American Statistical Association* **83,** 1067–1072.

(5) Seber, G. A. F. (1984), *Multivariate observations,* John Wiley & Sons.

(6) Tsai, C. H. and X. Wu (1992), Assessing local influence in linear regression models with first-order autoregressive or heteroscedastic error structure, *Statistics and Probability Letters* **14,** 247–252.