SECOND ORDER LOCAL INFLUENCE IN LINEAR DISCRIMINANT ANALYSIS

Kang-Mo Jung*, Myung Geun Kim[†] and Byung Chun Kim[‡]

ABSTRACT

We adapt the local influence method to linear discriminant analysis for the purpose of investigating the influence of observations. A simultaneous perturbation on all observations coming from two populations is considered. We study the curvatures and the associated direction vectors of the surface formed by the perturbed maximum likelihood estimators of parameters of interest, in addition to the direction vector of the maximum slope. We show that the influence function method gives essentially the same information as the direction vector of the maximum slope. A numerical example illustrates that the local influence method gives valuable information about influential observations and outliers, even when the influence function method and the case deletion method are not adequate.

1. Introduction

In linear discriminant analysis, diagnostic methods based on the influence functions and case deletions have been suggested for detecting influential observations and outliers. Campbell (1978) used the influence functions. Critchley and Vitiello (1991) and Fung (1992) considered the case deletion method to investigate the influence of observations on the misclassification probability. These influence diagnostics examine the effect of deleting a single observation.

Lawrance (1988) used the local influence method inspired by Cook (1986) to investigate the effect of small perturbations in a regression problem, based on the direction vector of the maximum slope of a path on the surface formed by the perturbed maximum likelihood estimator of a parameter. It is called the first order local influence method. Kim (1996) studied the direction vector of the maximum slope for the misclassification probability. In Section 4, we will show that the i th component of the direction vector of the maximum slope is equivalent to the empirical influence function at the i th observation, up to constants, as noted in Tanaka (1994) for Cook's local influence. Wu and Luo (1993) extended Lawrance's approach to the so-called second order local influence method based on the direction vectors corresponding to relatively large local maximum curvatures.

* Department of Computer Science and Statistics, Kunsan National University, 68 Miryong-Dong, Kunsan, Chollapuk-Do 573-701, Korea.

[†] Department of Applied Statistics, Seowon University, 231 Mochung-Dong, Chongju, Chung-Buk 360-742, Korea.

[‡] Department of Industrial Management, Korea Advanced Institute of Science and Technology, 373-1 Kusong-Dong, Yusong-Gu, Taejon 305-701, Korea.

Key words: Diagnostics; influential observations; linear discriminant analysis; local influence; perturbation.

In this work the second order local influence method is adapted to linear discriminant analysis for the purpose of investigating the influence of observations. The first order local influence gives the same information about influence as the empirical influence function. A numerical example in Section 5 shows that the second order local influence identifies multiple influential observations and outliers, while the first order local influence method and case deletion methods do not. This numerical result is a motivation for using the second order method.

A brief review of influence measures in linear discriminant analysis is given in Section 2. In Section 3, we consider a simultaneous perturbation on all observations coming from two populations. This perturbation is totally different from the influence function method and the case deletion method. Lawrance (1988) showed that a simultaneous perturbation is more effective than the individual perturbation in a regression context. The direction vectors corresponding to the largest and second largest absolute curvatures form the main diagnostic. Section 4 develops the local influence method for parameters of interest in linear discriminant analysis – the squared Mahalanobis distance, the misclassification probability, the vector of discriminant coefficients and the discriminant means. A closed form of the direction vector of the maximum slope is derived by computing directly the partial derivatives of the perturbed maximum likelihood estimators. It can be also seen that the local influence method includes two statistics proposed by Critchley and Vitiello (1991) and Fung (1992). In Section 5, a numerical example illustrates that the local influence method gives valuable information about influential observations and outliers, even when the influence function method and the case deletion method are not adequate.

2. Some Influence Measures

We review some influence measures proposed by Campbell (1978) and Fung (1992), which will be compared with the local influence method for numerical examples in Section 5. Let $\mathbf{x}_1, \dots, \mathbf{x}_{n_1}$ and $\mathbf{x}_{n_1+1}, \dots, \mathbf{x}_{n_1+n_2}$ ($n = n_1 + n_2$) be two independent random samples from p -variate normal distributions $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with a common positive definite covariance matrix, respectively. The maximum likelihood estimators of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ 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. The popular Fisher's linear discriminant rule assigns an observation $\mathbf{x}$ to population 1 if

$$\psi = \boldsymbol{\alpha}^T(\mathbf{x} - \boldsymbol{\mu}_1) > -\Delta^2/2, \quad (1)$$

where $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the vector of discriminant coefficients and $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the squared Mahalanobis distance.

The performance of the Fisher's rule relies on Δ^2 through the misclassification probability $\Phi(-\Delta/2)$. With the assumption that population 1 is perturbed, Campbell (1978) derived the influence function for Δ^2 as $I(\mathbf{x}, \Delta^2) = v_1 \Delta^2 + 2\psi - v_1 \psi^2$, where $v_1 = n_1/n$ and $v_2 = 1 - v_1$. In what follows, the hat notation will be used for writing the maximum likelihood estimator. The empirical influence function for Δ^2 at $\mathbf{x}$ is obtained as

$$I(\mathbf{x}, \widehat{\Delta}^2) = v_1 \widehat{\Delta}^2 + 2\widehat{\psi} - v_1 \widehat{\psi}^2, \quad (2)$$

where $\widehat{\psi} = \widehat{\boldsymbol{\alpha}}^T(\mathbf{x} - \bar{\mathbf{x}}_1)$. Also the empirical influence function for the misclassification probability can be easily derived as

$$I(\mathbf{x}, \widehat{MP}) = -I(\mathbf{x}, \widehat{\Delta}^2)\phi(-\widehat{\Delta}/2)/(4\widehat{\Delta}), \quad (3)$$

where $\widehat{MP} = \Phi(-\widehat{\Delta}/2)$ and $\phi(\cdot)$ is the standard normal density function. Since $I(\mathbf{x}, \widehat{MP})$ is proportional to $I(\mathbf{x}, \widehat{\Delta}^2)$, they give essentially the same influence information.

It is known that the distribution of $I(\mathbf{x}, \Delta^2)$ is negatively skewed (Campbell, 1978). This property and the quadratic form of $I(\mathbf{x}, \widehat{\Delta}^2)$ in (2) give the maximum value of $I(\mathbf{x}, \widehat{\Delta}^2)$ that is small and positive. Since the maximum value occurs at $\widehat{\psi} = v_1^{-1}$, an observation located on the plane $\widehat{\psi} = v_1^{-1}$ would have small influence, even when it is away from its sample mean.

Critchley and Vitiello (1991) considered the case deletion method to investigate the influence of observations on the misclassification probability. The influence measure suggested by them can be expressed in terms of $\widehat{\psi}_i (= \widehat{\alpha}^T(\mathbf{x}_i - \bar{\mathbf{x}}_1))$ and the following distance,

$$d_i^2 = (\mathbf{x}_i - \tilde{\mathbf{x}}_i)^T \mathbf{S}^{-1} (\mathbf{x}_i - \tilde{\mathbf{x}}_i), \quad i = 1, \dots, n, \quad (4)$$

where $\tilde{\mathbf{x}}_i = \bar{\mathbf{x}}_1$ if $1 \leq i \leq n_1$, and $\tilde{\mathbf{x}}_i = \bar{\mathbf{x}}_2$ otherwise. Based on $\widehat{\psi}_i$ and d_i^2 , Fung (1992) suggested the influence measure DMP_i that evaluates the effect of the omission of observation i on the estimation of the misclassification probability. DMP_i can be approximately equal to

$$DMP_i \approx \frac{\phi(-\widehat{\Delta}/2)}{4\widehat{\Delta}(n_1 - 1)^2} \left[(1 - v_1 \widehat{\psi}_i)^2 (d_i^2 - \widehat{\psi}_i^2 / \widehat{\Delta}^2) + \widehat{\psi}_i^2 / 4 \right] \quad (5)$$

(refer to equations (8) to (13) in Fung(1992)).

Note that if a single observation j has a very large d_j^2 and the others have small, $DMP_k (k \neq j)$ may be relatively small to DMP_j . This case may occur if observation j is lying far away from its sample mean and is not located near the plane $\widehat{\psi} = v_1^{-1}$ (that is, easy to be classified), because an observation close to the plane has small value of $(1 - v_1 \widehat{\psi}_i)^2$ in (5). Then DMP_i may not give sufficient information about multiple influential observations and the influence of the other observations. In Section 5 a numerical example which reflects this phenomenon will be given.

3. Local Influence Measures

In this section the local influence method (Cook, 1986 ; Wu and Luo, 1992) will be adapted to linear discriminant analysis. We consider a simultaneous perturbation on all observations coming from two populations. It is totally different from the perturbation of the influence function method and the case deletion method, which can perturb a single observation. The perturbed model is specified by a perturbation vector $\mathbf{w} = (w_1, \dots, w_n)^T$, in which

$$\mathbf{x}_r \sim \mathbf{N}(\tilde{\boldsymbol{\mu}}_r, \boldsymbol{\Sigma}/w_r), \quad (6)$$

where $\tilde{\boldsymbol{\mu}}_r$ equals $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$ according to the index r . The perturbation vector can be written as $\mathbf{w} = \mathbf{1}_n + a\mathbf{l}$, where $\mathbf{l} = (l_1, \dots, l_n)^T$ denotes the direction vector of unit length, scalar a indicates the magnitude of perturbation, and $\mathbf{1}_n = (1, \dots, 1)^T$ of order n . When $a = 0$, that is $\mathbf{w} = \mathbf{1}_n$, the perturbed model reduces to the unperturbed model.

Let θ be a parameter of interest, for example Δ^2 or $\Phi(-\Delta/2)$. The maximum likelihood estimator of θ under the perturbed model is denoted by $\widehat{\theta}(\mathbf{w})$. Then the $(n+1)$ by 1 vector $\tau(\mathbf{w}) = (\mathbf{w}^T, \widehat{\theta}(\mathbf{w}))^T$ forms a surface in the $(n+1)$ -dimensional space as $\mathbf{w}$ varies over a certain space. The direction vector of the maximum slope of a path on the surface $\tau(\mathbf{w})$ at

$a = 0$ is considered for investigating the local behaviour of observations for the misclassification probability (Kim, 1996). However, it will be shown in Section 4 that the results from the direction vector of the maximum slope, say, l_{slope} , is essentially the same as the influence function method. The first and second direction vectors, say, l_{max} and l_{sec} , respectively, of the surface at $\mathbf{w} = \mathbf{1}_n$ corresponding to the largest and second largest absolute curvatures respectively yield information about influential observations. Observations corresponding to significantly large direction cosines of the two direction vectors above in its absolute value can be influential. Here the absolute value of the curvature is required, for the curvature can be negative. Usually it is sufficient to examine l_{max} and l_{sec} , even though other directions (that is, related to the third and the fourth absolute largest curvatures) may provide much information.

The influence of observations on the perturbed maximum likelihood estimator $\hat{\theta}(\mathbf{w})$ of a parameter θ can be investigated as follows. The first and second order partial derivatives of $\hat{\theta}(\mathbf{w})$ with respect to a evaluated at $a = 0$ are

$$\begin{aligned} \left. \frac{\partial \hat{\theta}(\mathbf{w})}{\partial a} \right|_{a=0} &= \sum_{r=1}^n \left(\left. \frac{\partial \hat{\theta}(\mathbf{w})}{\partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} \right) l_r, \\ \left. \frac{\partial^2 \hat{\theta}(\mathbf{w})}{\partial a^2} \right|_{a=0} &= \sum_{s=1}^n \sum_{r=1}^n \left(\left. \frac{\partial^2 \hat{\theta}(\mathbf{w})}{\partial w_s \partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} \right) l_r l_s. \end{aligned}$$

The curvature and its associated direction vector of the surface at $\mathbf{w} = \mathbf{1}_n$ (refer to equations (2.2) to (2.5) in Wu and Luo (1993)) are obtained by solving the generalized eigenvalue problem

$$|\mathbf{F} - \beta \mathbf{G}| = 0, \quad (7)$$

where $\mathbf{F}$ is the n by n matrix having $\partial^2 \hat{\theta}(\mathbf{w}) / \partial w_s \partial w_r |_{\mathbf{w}=\mathbf{1}_n}$ as its (r, s) th element, $\mathbf{G} = (1 + \dot{\theta}^T \dot{\theta})^{1/2} (\mathbf{I}_n + \dot{\theta} \dot{\theta}^T)$, and $\dot{\theta}$ is the n by 1 column vector whose r th element is $\partial \hat{\theta}(\mathbf{w}) / \partial w_r |_{\mathbf{w}=\mathbf{1}_n}$. The curvature of the surface is given by the eigenvalue in (7) and the direction vector l is its associated eigenvector of unit length. This comes from the fact that the curvature is equivalent to the value of $l^T \mathbf{F} l / l^T \mathbf{G} l$.

4. Derivation

Consider the linear discriminant function for two populations in (1). The parameters of interest are the squared Mahalanobis distance Δ^2 , the misclassification probability $\Phi(-\Delta/2)$, the vector of discriminant coefficients α and the discriminant means $\alpha^T \mu_k$ for $k = 1, 2$. They all involve Σ^{-1} and μ_k . Hence to solve the generalized eigenvalue problem (7), we need the first and second order partial derivatives of the perturbed maximum likelihood estimators $\mu_k(\mathbf{w})$, $\mathbf{S}(\mathbf{w})$ and $\mathbf{S}(\mathbf{w})^{-1}$ evaluated at $\mathbf{w} = \mathbf{1}_n$.

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 mean vectors and the covariance matrix under the perturbed model (6) are obtained as

$$\hat{\mu}_1(\mathbf{w}) = \sum_{j=1}^{n_1} w_j \mathbf{x}_j / \sum_{j=1}^{n_1} w_j, \quad (8)$$

$$\hat{\mu}_2(\mathbf{w}) = \sum_{j=1}^{n_2} w_{n_1+j} \mathbf{x}_{n_1+j} / \sum_{j=1}^{n_2} w_{n_1+j}, \quad (9)$$

$$\mathbf{S}(\mathbf{w}) = (\mathbf{X}_1 \mathbf{H}_1(\mathbf{w}) \mathbf{X}_1^T + \mathbf{X}_2 \mathbf{H}_2(\mathbf{w}) \mathbf{X}_2^T) / n, \quad (10)$$

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

Let $I_1(r)$ be one if $r \in \{1, \dots, n_1\}$ and zero otherwise, and $I_2(r)$ be one if $r \in \{n_1 + 1, \dots, n\}$ and zero otherwise. By differentiating equations (8) to (10) with respect to w_r and putting $\mathbf{w} = \mathbf{1}_n$, the first order partial derivatives of $\hat{\mu}_k(\mathbf{w})$ for $k = 1, 2$ and $\mathbf{S}(\mathbf{w})$ evaluated at $\mathbf{w} = \mathbf{1}_n$ are

$$\left. \frac{\partial \hat{\mu}_k(\mathbf{w})}{\partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} = \frac{1}{\tilde{n}_r} I_k(r) (\mathbf{x}_r - \tilde{\mathbf{x}}_r), \quad (11)$$

$$\left. \frac{\partial \mathbf{S}(\mathbf{w})}{\partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} = \frac{1}{n} (\mathbf{x}_r - \tilde{\mathbf{x}}_r) (\mathbf{x}_r - \tilde{\mathbf{x}}_r)^T, \quad (12)$$

where $\tilde{n}_r = n_1$ if $1 \leq r \leq n_1$, and $\tilde{n}_r = n_2$ otherwise. Let $I(r, s)$ be one if both $\mathbf{x}_r$ and $\mathbf{x}_s$ are observations from the same population, and zero otherwise. We also get

$$\left. \frac{\partial^2 \hat{\mu}_k(\mathbf{w})}{\partial w_s \partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} = -\frac{1}{\tilde{n}_r \tilde{n}_s} I(r, s) \{(\mathbf{x}_r - \tilde{\mathbf{x}}_r) + (\mathbf{x}_s - \tilde{\mathbf{x}}_s)\}, \quad (13)$$

$$\begin{aligned} \left. \frac{\partial^2 \mathbf{S}(\mathbf{w})}{\partial \mathbf{w}_s \partial \mathbf{w}_r} \right|_{\mathbf{w}=\mathbf{1}_n} &= -\frac{1}{n \tilde{n}_r} I(r, s) \{(\mathbf{x}_r - \tilde{\mathbf{x}}_r) (\mathbf{x}_s - \tilde{\mathbf{x}}_s)^T \\ &\quad + (\mathbf{x}_s - \tilde{\mathbf{x}}_s) (\mathbf{x}_r - \tilde{\mathbf{x}}_r)^T\}. \end{aligned} \quad (14)$$

Furthermore, the first order partial derivative of $\mathbf{S}(\mathbf{w})^{-1}$ is easily found by $\partial \mathbf{S}(\mathbf{w}) / \partial w_r = -\mathbf{S}(\mathbf{w})^{-1} (\partial \mathbf{S}(\mathbf{w}) / \partial w_r) \mathbf{S}(\mathbf{w})^{-1}$. From (12) and (14), we obtain

$$\left. \frac{\partial \mathbf{S}(\mathbf{w})^{-1}}{\partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} = -\frac{1}{n} \mathbf{S}^{-1} (\mathbf{x}_r - \tilde{\mathbf{x}}_r) (\mathbf{x}_r - \tilde{\mathbf{x}}_r)^T \mathbf{S}^{-1}, \quad (15)$$

$$\begin{aligned} \left. \frac{\partial^2 \mathbf{S}(\mathbf{w})^{-1}}{\partial w_s \partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} &= \frac{1}{n} \left\{ \frac{1}{n} (\mathbf{x}_r - \tilde{\mathbf{x}}_r)^T \mathbf{S}^{-1} (\mathbf{x}_s - \tilde{\mathbf{x}}_s) + \frac{1}{\tilde{n}_r} I(r, s) \right\} \\ &\quad \cdot \mathbf{S}^{-1} \{(\mathbf{x}_r - \tilde{\mathbf{x}}_r) (\mathbf{x}_s - \tilde{\mathbf{x}}_s)^T + (\mathbf{x}_s - \tilde{\mathbf{x}}_s) (\mathbf{x}_r - \tilde{\mathbf{x}}_r)^T\} \mathbf{S}^{-1}. \end{aligned} \quad (16)$$

4.1. Misclassification Probability

For getting the local influence measures of the misclassification probability $\Phi(-\Delta/2)$, we begin to derive the local influence measures of Δ^2 .

The maximum likelihood estimator of Δ^2 under the perturbation (6) can be written as

$$\hat{\Delta}^2(\mathbf{w}) = (\hat{\mu}_1(\mathbf{w}) - \hat{\mu}_2(\mathbf{w}))^T \mathbf{S}(\mathbf{w})^{-1} (\hat{\mu}_1(\mathbf{w}) - \hat{\mu}_2(\mathbf{w})). \quad (17)$$

The chain rule of differentiation, (11) and (15) yield

$$\hat{\Delta}_r^2 \equiv \left. \frac{\partial \hat{\Delta}^2(\mathbf{w})}{\partial w_r} \right|_{\mathbf{w}=\mathbf{1}_n} = (2\tilde{e}_r - \tilde{v}_r \tilde{e}_r^2) / \tilde{n}_r, \quad (18)$$

where $\tilde{v}_r$ is similarly defined as $\tilde{n}_r$, and $\tilde{e}_r$ is defined by

$$\tilde{e}_r = \begin{cases} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\mathbf{x}_r - \bar{\mathbf{x}}_1), & 1 \leq r \leq n_1, \\ -(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \mathbf{S}^{-1}(\mathbf{x}_r - \bar{\mathbf{x}}_2), & n_1 + 1 \leq r \leq n. \end{cases}$$

The direction vector of the maximum slope of a path on the surface $(\mathbf{w}^T, \hat{\Delta}^2(\mathbf{w}))^T$ at $a = 0$ becomes the vector with the r th component $\hat{\Delta}_r^2 / (\sum_{i=1}^n (\hat{\Delta}_i^2)^2)^{1/2}$ by the Cauchy-Schwarz inequality. Since $\hat{\Delta}_r^2 = (I(\mathbf{x}_r, \hat{\Delta}^2) - v_1 \hat{\Delta}^2) / n_1$ for $\mathbf{x}_r$ coming from population 1, $\hat{\Delta}_r^2$ is equivalent to $I(\mathbf{x}_r, \hat{\Delta}^2)$, up to constants. It implies that l_{slope} gives the same information as the influence function method.

Using the partial derivatives in equations (11) to (16), we get

$$\begin{aligned} \hat{\Delta}_{rs}^2 \equiv \frac{\partial^2 \hat{\Delta}^2(\mathbf{w})}{\partial w_s \partial w_r} \bigg|_{\mathbf{w}=\mathbf{1}_n} &= \frac{2}{\tilde{n}_r \tilde{n}_s} \{ (-1)^{I(r,s)+1} \tilde{e}_{rs} (1 - \tilde{v}_r \tilde{e}_r) (1 - \tilde{v}_s \tilde{e}_s) \\ &\quad + I(r, s) (\tilde{v}_r \tilde{e}_r \tilde{e}_s - \tilde{e}_r - \tilde{e}_s) \}, \end{aligned} \quad (19)$$

where $\tilde{e}_{rs} = (\mathbf{x}_r - \tilde{\mathbf{x}}_r)^T \mathbf{S}^{-1}(\mathbf{x}_s - \tilde{\mathbf{x}}_s)$. In what follows, such expressions as $\hat{\Delta}_r^2$ and $\hat{\Delta}_{rs}^2$ are interpreted as the first and second order partial derivatives of the perturbed maximum likelihood estimator evaluated at $\mathbf{w} = \mathbf{1}_n$, respectively. The statistics $\hat{\psi}_r (= \tilde{e}_r)$ and $d_r^2 (= \tilde{e}_{rr})$ are useful for detecting outliers and influential observations (Fung, 1992). From (19), we can see that $\hat{\Delta}_{rs}^2$ includes these two statistics. Besides $\hat{\psi}_r$ and d_r^2 , $\hat{\Delta}_{rs}^2$ involves $\tilde{e}_{rs} (r \neq s)$ which is interpreted as an outlyingness estimate of $\mathbf{x}_r$ and $\mathbf{x}_s$. Thus l_{max} and l_{sec} can provide information about multiple influential observations.

It is well known that Δ^2 is invariant under affine transformations. It is easy to show that $\tilde{e}_r$ and $\tilde{e}_{rs}$ are invariant, which lead to the invariance of the local influence measures of Δ^2 .

For the misclassification probability, the maximum likelihood estimator under the perturbation (6) is given by $\widehat{MP}(\mathbf{w}) = \Phi(-\hat{\Delta}(\mathbf{w})/2)$. By the invariance of the maximum likelihood estimator and the chain rule, we get

$$\widehat{MP}_r = -\hat{\Delta}_r^2 \phi(-\hat{\Delta}/2) / (4\hat{\Delta}), \quad (20)$$

$$\widehat{MP}_{rs} = \left\{ \hat{\Delta}_r^2 \hat{\Delta}_s^2 \left(\frac{1}{4} + \frac{1}{\hat{\Delta}^2} \right) - 2\hat{\Delta}_{rs}^2 \right\} \phi(-\hat{\Delta}/2) / (8\hat{\Delta}). \quad (21)$$

From (3) and (20), it can be easily shown that $\widehat{MP}_r$ and $I(\mathbf{x}_r, \widehat{MP})$ give the same influence information as in Δ^2 . Obviously the local influence measures of the misclassification probability are also invariant.

4.2. Vector of Discriminant Coefficients

Under the perturbation (6), we get the maximum likelihood estimator of the vector of discriminant coefficients $\boldsymbol{\alpha}$ as $\hat{\boldsymbol{\alpha}}(\mathbf{w}) = \mathbf{S}(\mathbf{w})^{-1}(\bar{\mathbf{x}}_1(\mathbf{w}) - \bar{\mathbf{x}}_2(\mathbf{w}))$. From equations (11) to (16), we get

$$\hat{\boldsymbol{\alpha}}_r = \frac{1}{\tilde{n}_r} (1 - \tilde{v}_r \tilde{e}_r) \tilde{\mathbf{u}}_r, \quad (22)$$

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_{rs} &= \frac{1}{\tilde{n}_r \tilde{n}_s} [(-1)^{I(r,s)+1} \tilde{e}_{rs} \{ \tilde{v}_s (\tilde{v}_r \tilde{e}_r - 1) \tilde{\mathbf{u}}_s + \tilde{v}_r (\tilde{v}_s \tilde{e}_s - 1) \tilde{\mathbf{u}}_r \} \\ &\quad + I(r, s) \{ (\tilde{v}_r \tilde{e}_r - 1) \tilde{\mathbf{u}}_s + (\tilde{v}_s \tilde{e}_s - 1) \tilde{\mathbf{u}}_r \}], \end{aligned} \quad (23)$$

where

$$\tilde{\mathbf{u}}_r = \begin{cases} \mathbf{S}^{-1}(\mathbf{x}_r - \bar{\mathbf{x}}_1), & 1 \leq r \leq n_1, \\ -\mathbf{S}^{-1}(\mathbf{x}_r - \bar{\mathbf{x}}_2), & n_1 + 1 \leq r \leq n. \end{cases}$$

A simple scalar of the vector of discriminant coefficients is the squared length of α . The first and second order partial derivatives of the squared length of α evaluated at $\mathbf{w} = \mathbf{1}_n$ are given by $2\hat{\alpha}_r^T \hat{\alpha}$ and $2(\hat{\alpha}_r^T \hat{\alpha}_s + \hat{\alpha}_{rs}^T \hat{\alpha})$, respectively, where $\hat{\alpha} = \hat{\alpha}(\mathbf{1}_n)$.

4.3. Discriminant Means

The perturbed maximum likelihood estimator of the discriminant mean $\alpha^T \mu_k$ for $k = 1, 2$ is $\widehat{DM}_k(\mathbf{w}) = \hat{\alpha}(\mathbf{w})^T \hat{\mu}_k(\mathbf{w})$. Proceeding exactly in the same way as in Subsections 4.1 and 4.2, we obtain

$$\widehat{DM}_{k,r} = \frac{1}{n_r} (1 - \tilde{v}_r \tilde{e}_r) \tilde{\mathbf{u}}_r^T \bar{\mathbf{x}}_k + \frac{\tilde{e}_r}{n_r} I_k(r) (I_1(r) - I_2(r)).$$

Furthermore, (11), (13), (22), (23), and some algebra yield

$$\widehat{DM}_{k,rs} = (\tilde{\delta}_{k,rs} + \tilde{\delta}_{k,sr}) / (\tilde{n}_r \tilde{n}_s),$$

where

$$\begin{aligned} \tilde{\delta}_{k,rs} = & (\tilde{v}_r \tilde{e}_r - 1) \left\{ \tilde{e}_{rs} \tilde{v}_s (-1)^{I(r,s)+1} + I(r, s) \right\} \tilde{\mathbf{u}}_s^T \bar{\mathbf{x}}_k \\ & - I_k(s) (I_1(r) - I_2(r)) (\tilde{v}_r \tilde{e}_r - 1) \tilde{e}_{rs} - I(r, s) (I_1(r) - I_2(r)) \tilde{e}_r. \end{aligned}$$

5. Numerical Examples

The local influence method is applied to the iris data (Mardia, *et al.*, 1979, pp. 6–7, Table 1.2.2), in which only two species, Iris versicolor and Iris virginica are considered for our illustration. For clarity only two variables, sepal width and petal width, are selected so that we can easily detect the sources of influence from the scatter plot of the data. The observations are labelled as 1 to 50 for Iris versicolor and 51 to 100 for Iris virginica. This data set has also been studied by Fung (1992). Our analysis will be confined to the misclassification probability for comparing the local influence method with the influence function method and the case deletion method.

First the local influence method is conducted based on the first order local influence. Since the index plot of l_{slope} exhibits essentially the same behaviour as that of $I(\mathbf{x}_r, \widehat{MP})$ (refer to Fig. 2 in Fung (1992)), we omit it. The scatter plot of l_{max} versus l_{sec} is given in Fig. 1. We can find that observations 19, 70 and 82 have high local influence as they are apart from the others. This result agrees with that of Fung (1992). Furthermore, it can be observed that the misclassified observations 19, 21, 80, 84 and 85 lie at the outer side in Fig. 1.

As pointed out in Section 2, there is a possibility that $DM P_i$'s are largely dominated by that of a single observation. To confirm this phenomenon, observation 1 such that (3.2, 1.4) is changed to (6.2, 1.4). The modified data together with each sample mean denoted by M and the discriminant line are given in Fig. 2. It reveals that observation 1 can be regarded as an outlier from population 1 and observations 21, 28, 70, 80, 84 and 85 are misclassified.

The misclassified observations for the modified data are detected by the index plot of l_{slope} given in Fig. 3, but observation 1 is not. Since $I(\mathbf{x}_r, \widehat{MP})$ gives the same information as l_{slope} , we omit the index plot of $I(\mathbf{x}_r, \widehat{MP})$.

Fig. 1. The scatter plot of $l_{\max}$ versus $l_{\sec}$ for the iris data.

Fig. 2. A modified data of sepal and petal width of Iris versicolor and Iris virginica.

Fig. 3. The index plot of l_{slope} for the modified iris data.

Fig. 4. The index plot of DMP_i for the modified iris data. The value of y coordinate is $DMP_i \times 10^4$.

Fig. 5. The scatter plot of l_{max} versus l_{sec} for the modified iris data.

For the modified data, the index plot of DMP_i is given in Fig. 4, using equations (9) to (12) in Fung (1992). It detects observation 1 only. Six misclassified observations are not identified by the index plot, because DMP_1 dominates the other DMP_i 's as mentioned in Section 2.

The scatter plot of l_{max} versus l_{sec} for the misclassification probability for the modified iris data is presented in Fig. 5. The largest and second largest absolute curvatures are 0.192 and 0.105, respectively. Observation 1 is clearly separated from the others in Fig. 5. Besides observation 1, it can be observed that observations 19, 21, 28, 34, 70, 80, 82, 84 and 85 have high local influence, and they are possible candidates for influential observations. Recall that observations 21, 28, 70, 80, 84 and 85 are misclassified. Observations 19, 34 and 82 may be influential observations, because they lie on the outer side from the corresponding population and near the discriminant line in Fig. 2 (Fung, 1992). Furthermore, observations 19 and 82 are influential observations for the unmodified iris data. In summary, the scatter plot of l_{max} versus l_{sec} identifies influential observations (observations 19, 34 and 82 (not detected by l_{slope} or DMP_i)) as well as an outlier (observation 1 (not detected by l_{slope})) and six misclassified observations (not detected by DMP_i).

This example shows that the scatter plot of l_{max} versus l_{sec} in linear discriminant analysis is very effective, in that it gives valuable information about misclassified observations,

influential observations and outliers. However, DMP_i is not adequate for getting influence information.

6. Concluding Remarks

The local influence method involves $\tilde{e}_{rs}(r \neq s)$, in addition to $\tilde{e}_r$ and $\tilde{e}_{rr}$ used in the influence function method and the case deletion method. The influence function method gives essentially the same information as l_{slope} , the result of the first order local influence method. The influence function method is not sufficient in getting information about influential observations. The scatter plot of l_{max} versus l_{sec} , the result of the second order local influence method, is very informative in identifying influential observations and outliers, even when DMP_i 's are largely dominated by that of an observation so that it does not give influence information about other observations and multiple influential observations. A numerical example shows it.

Acknowledgements

We are grateful to the referees for helpful comments. The first author's work was partially supported by funds granted by Post-Doctoral Program, Korea Science and Engineering Foundation in 1997.

REFERENCES

- Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant analysis. *Applied Statistics*, **27**, 251–258.
- Cook, R. D. (1986). Assessment of local influence (with discussions). *J. R. Statist. Soc.*, **B 48**, 133–169.
- Critchley, F. and Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis. *Biometrika*, **78**, 677–690.
- Fung, W.-K. (1992). Some diagnostic measures in discriminant analysis. *Statistics & Probability Letters*, **13**, 279–285.
- Kim, M. G. (1996). Local influence on misclassification probability. *Journal of the Korean Statistical Society*, **25**, 145–151.
- Lawrance, A. J. (1988). Regression transformation diagnostics using local influence. *J. Amer. Statist. Assoc.*, **83**, 1067–1072.
- Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.
- Tanaka, Y. (1994). Recent advance in sensitivity analysis in multivariate statistical methods. *J. Jpn. Soc. Comp. Statist.*, **7**, 1–25.
- Wu, X. and Luo, Z. (1993). Second-order approach to local influence. *J. R. Statist. Soc.*, **B 55**, 929–936.

(Received January 1997; Revised October 1997)