Communications for Statistical Applications and Methods
2015, Vol. 22, No. 5, 475–485

DOI: http://dx.doi.org/10.5351/CSAM.2015.22.5.475
Print ISSN 2287-7843 / Online ISSN 2383-4757

Graphical Methods for the Sensitivity Analysis in Discriminant Analysis

Dae-Heung Jang[a], Christine M. Anderson-Cook[b], Youngil Kim[1,c]

[a]Department of Statistics, Pukyong National University, Korea;
[b]Statistical Sciences Group, Los Alamos National Laboratory, USA;
[c]School of Business and Economics, Chung-Ang University, Korea

Abstract

Similar to regression, many measures to detect influential data points in discriminant analysis have been developed. Many follow similar principles as the diagnostic measures used in linear regression in the context of discriminant analysis. Here we focus on the impact on the predicted classification posterior probability when a data point is omitted. The new method is intuitive and easily interpretable compared to existing methods. We also propose a graphical display to show the individual movement of the posterior probability of other data points when a specific data point is omitted. This enables the summaries to capture the overall pattern of the change.

Keywords: discriminant analysis, influence detection measure, individual movement plot, influential observations

1. Introduction

Discriminant analysis is a collection of multivariate techniques that use statistical methods to characterize or separate two or more classes of objects or events. We often use this technique to allocate new observations to previously defined groups.

Using ideas from Cook (1977), the study of influential data points in discriminant analysis are used to investigate the effect on the estimated overall probability of misclassification if certain observations were deleted. Campbell (1978) considered the empirical influence curve, while Critchley and Vitiello (1991), Fung (1992, 1995), Lahiff and Whitcomb (1990), Moreno-Roldán *et al.* (2007) and Lee and Kim (2011) introduced relevant statistical measures based on Campbell (1978). Their papers are all based on the deletion principle.

An new modification of local influence approach suggested by Cook (1986) was recently introduced in discriminant analysis by Jung (1998) and Poon (2004). This approach utilizes local infinitesimal perturbation for a differential comparison of parameter estimates. In this case there is no need to delete an observation completely, which leads to an approach that has improved interpretability over deletion methods.

Section 2 contains a review of linear discriminant analysis and defines the statistical method proposed related to traditional approaches based on misclassification probabilities. Section 3 demonstrates the method with two examples. The approach is compared to the local influence method. Section 4 contains conclusions with some discussion.

This work was supported by a Research Grant of Pukyong National University 2015.

[1] Corresponding author: School of Business and Economics, Chung-Ang University, 84 Heukseok-ro, Dongjakgu, Seoul 156-756, Korea. E-mail: yik01@cau.ac.kr

Published 30 September 2015 / journal homepage: http://csam.or.kr
© 2015 The Korean Statistical Society, and Korean International Statistical Society. All rights reserved.

2. Review of Discriminant Analysis

We begin our discussion with some basic notation for the purpose of clarity. Given k populations $\Pi_1, \Pi_2, \ldots, \Pi_k$ suppose that each Π_i has a p.d.f. $f_i(\mathbf{x})$ for a set of p measurements, $\mathbf{x}$. In discriminant analysis it is important to find a partition of R^p into $R_1, R_2, \ldots, R_k$ disjoint regions with the following decision rule (2.1). When

$$\mathbf{x} \in R_j, \quad \text{allocate } \mathbf{x} \text{ to } \Pi_j. \tag{2.1}$$

We assume that $\{\Pi_i\}$ follows a multivariate normal distribution (MVN), $N_p(\boldsymbol{\mu}_p, \Sigma_i)$ for $i = 1, 2$. In case of equal $\Sigma_1 = \Sigma_2 = \Sigma$, the Fisher's linear discriminant rule results: allocate $\mathbf{x}$ to Π_1 if

$$h(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right) > 0 \tag{2.2}$$

and to Π_2 otherwise. Sometimes $h(\mathbf{x})$ is called the discriminant function, which is linear in $\mathbf{x}$. Furthermore, $\alpha^T = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the discriminant coefficient.

If we set up $\pi_i, i = 1, 2$ as the prior probability for each group and assume the simple case: $\pi_1 = \pi_2$, the misclassification probabilities, $p_{ij} = \Pr(\text{allocate to } \Pi_j \text{ when in fact from } \Pi_i)$ has the form (2.3) below

$$p_{ij} = \Pr(h(\mathbf{x}) > 0 | \Pi_2) = \Phi\left(-\frac{1}{2}\Delta \right), \quad \text{for 2 group case,} \tag{2.3}$$

where $\Delta^2 = (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \Sigma^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ is the Mahalanobis distance between $\boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_1$.

In practice, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$ are often estimated by $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, S_p$ where S_p is the pooled estimator of $\Sigma_i, i = 1, 2$ respectively.

This misclassification probability depends on the difference mean vector adjusted for Σ. By symmetry it is easy to see $p_{12} = p_{21}$. For the multiple group case with different Σ_i, this misclassification probability is difficult to calculate in closed form.

The actual error rate of misclassification, which is often reported in discriminant analysis output, is different from this probability.

In this paper, we focus on the two group case with equal prior probability and equal cost of misclassification for two reasons: (1) it is simple to explain the proposed new measure to detect influential observations. (2) The existing measures that rely on the probability of misclassification in more than 2 groups are usually hard to handle. However, the proposed graphical methods are not constrained to the $k = 2$ case.

As noted above in Section 1, most statistical measures to detect influential observations are based on the influence measure. Campbell (1978) applied the influence measure to discriminant scores and discriminant coefficients. Later Critichely and Vitiello (1991) used exact calculations to show that the influence of an observation on the misclassification probability is due to two sources: (a) the difference between the data points linear discriminant score and that of its sample mean and (b) the estimated Mahalanobis distance between the data point and its population mean.

Fung (1992, 1995) later developed diagnostic measures such as the expected change in discriminant scores due to the omission of specific observations for the two basic sources of misclassification. Lachenbruch (1997) showed that the leverage of a data point is an increasing function of (b) and decreasing function of (a).

Cook (1986) proposed a new approach without the need to delete one observation to detect influential observations in linear regression.

It is well known that the discriminant coefficient $\hat{\alpha}$ can be obtained equivalently using regression coefficients with a binary dummy dependent variable when $k = 2$. Let

$$Y = X\beta + \epsilon$$

be the usual linear regression where Y is $n \times 1$ vector of observation, X is $n \times q$ data matrix and ϵ follows the n-dimensional multi-variate $N(0, \sigma^2 I)$. Then the usual least squares coefficient estimate $\hat{\beta}$ is obtained by maximizing the log-likelihood function below.

$$L(\beta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 .$$

First define the following

$$L(\beta|\omega) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \omega_i \left(y_i - x_i^T \beta\right)^2$$

as the log-likelihood of the regression model with case-weight perturbation. The corresponding $\hat{\beta}$ is denoted as $\hat{\beta}_\omega$. Then the method focuses on the largest eigenvector of the normal curvature matrix of the following log-likelihood displacement (LD) (2.4) to determine the influential observations.

$$\mathrm{LD}(\omega) = 2\left[L\left(\hat{\beta}|\omega_0\right) - L\left(\hat{\beta}_\omega|\omega\right)\right], \tag{2.4}$$

where ω_0 is simply null perturbation, $\omega_i = 1$ for $i = 1, 2, \ldots, n$.

From Cook (1986), the normal curvature matrix C of for the case perturbation is of the following form (2.5)

$$C = \frac{2}{\sigma^2} D(r) H D(r), \tag{2.5}$$

where $D(r)$ is the diagonal matrix with the i^{th} diagonal element equal to i^{th} residual and $H = X(X^T X)^{-1} X^T$. The method identifies influential observations based on the magnitude of the elements of largest eigenvector $l_{\max}$ corresponding the normal curvature matrix. Furthermore, if C_i is the i^{th} diagonal element of C, the method considers $B_i = C_i / \sqrt{\mathrm{tr}(C^2)}$. The group of cases with large B_i values believed to be influential. For more details, readers are referred to the original paper. These two criteria, $l_{\max}$ and B_i, are displayed with the new graphical approach for comparison. This idea was exploited in the discriminant analysis by Poon (2004).

Cook's approach is free of the dominating and masking effects often encountered in the analysis using the *leave-one-out* diagnostics. Therefore this has been a welcomed addition to the existing literature despite the conceptual difficulty for practitioners. Our new statistical measure will be compared with this approach for this complimentary purpose.

Each statistical measure is developed by focusing on different aspects of the misclassification probability, and this means that each has its own subset of influential data points. A subset of influential observations can be different based on which measure is considered with a specific measure.

In Section 3 we propose another approach, which does not depend directly on the misclassification probability. The misclassification probability is an overall statistical measure like R^2 in linear regression. Rather a data-specific influential measure would be more meaningful in understanding the impact of different observations. The intuitive interpretation of the new measure should be appealing to practitioners since it provides additional information about the impact on the rest of observations when one data point is deleted.

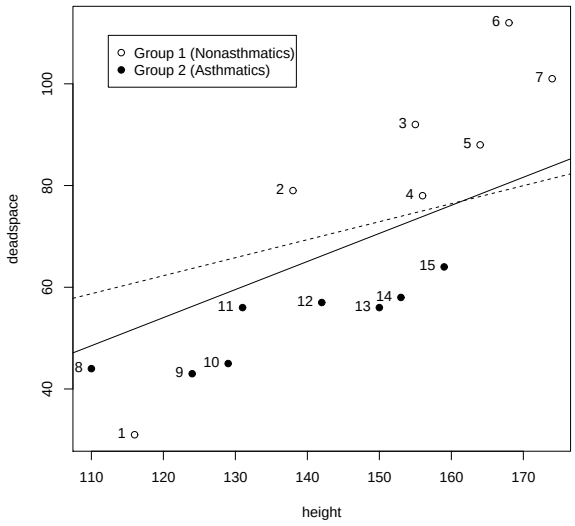


Figure 1: *Scatterplot for the lung function data set: The discriminant function with full data set (solid line) and when case 1 is excluded (dotted line).*

3. Graphical Methods for the Influence Detection

Before we introduce the new measure, we describe a real examples from the literature. The data set is taken from Campbell (2001). It is composed of two groups with 7 nonasthmatic children and 8 asthmatic children with two variables: children's pulmonary anatomical dead space, and their height measured in cm. From Figure 1, the solid line gives the sample discriminant function and it is clear that the child 1 has been misclassified. Furthermore, if the child 1 is excluded from the data set, the discriminant function shown with the dotted line changes significantly.

It is clear from the plot that we can obtain perfect separation into two groups for the data set when child 1 is excluded. The difference of estimated overall misclassification probability between with child 1 and without child 1 is also the largest among others, but the magnitude of the difference of misclassification probability is small

$$\Phi\left(-\frac{1}{2}\Delta\right) - \Phi\left(-\frac{1}{2}\Delta_{(1)}\right) = 0.420118 - 0.35108 = 0.069038,$$

where Δ is the sample Mahalanobis distance between two sample mean centroids and $\Delta_{(1)}$ is the corresponding value without the first observation.

The Mahalanobis distance depends on the difference between mean vectors for the two groups. It is interesting to check the magnitude of directional change of group mean if a certain observation is deleted. Figure 2 shows the group mean movement plot. We immediately see that the exclusion of the observation for child 1 leads to a relatively large size of change of the group mean centroid, especially the group mean centroid for nonasthmatic children.

Note that the misclassification probability still exists even when the child 1 is excluded. Nonetheless we conclude that observation 1 is influential since child 1 gives the largest change in the location of the group centroids. In this regard it seems that there is a need to have more data-specific statistical measures than one based on the overall misclassification probability. We now propose an alternate

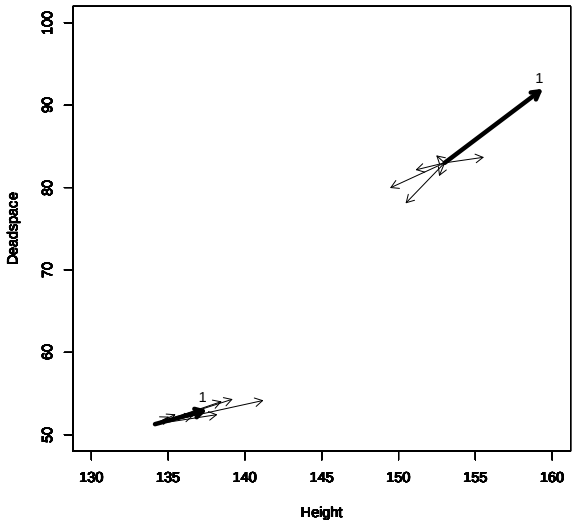


Figure 2: *Group mean movement plot for the lung function data set: The thick arrow corresponds to the group mean centroids for each group obtained by excluding child 1.*

measure which is simpler than the statistical measures previously developed and largely based on misclassification probability.

As we have mentioned in Section 1, the concept of the misclassification probability has been the cornerstone in detecting influential observations in the past.

Since the misclassification probability is an overall measure, the introduction of a data specific measure could be a beneficial complement. We consider the individual posterior classification as stated below

$$P(\Pi_1|x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)} = 1 - P(\Pi_2|x).$$

Note that the decision rule to classify the observation x_0 to Π_1 when $P(\Pi_1|x_0) > P(\Pi_2|x_0)$ is equivalent to finding the discriminant function that minimizes the overall misclassification probability and expected cost of misclassification under equal misclassification cost for both groups. Therefore it is comparable to deal with the change of posterior probabilities rather than the mathematical form of misclassification probability. Suppose we are interested in the effect of omission of an observation on the change of posterior probabilities. Now we are interested in the total of change of posterior probability in an absolute sense

$$D(i) = \sum_{k=1}^{2} \sum_{j}^{n} \left| P\left(\Pi_k|x_j\right) - P\left(\Pi_k|x_j(i)\right) \right|, \tag{3.1}$$

where $P(\Pi_k|x_j(i))$ is the posterior probability of the j^{th} observation when the i^{th} observation is omitted.

It is reasonable when $j = i$, to set $P(\Pi_k|x_j) - P(\Pi_k|x_j(i)) = 0$. This quantity is interpreted as the total of change of two posterior probabilities coming from a full and reduced data set. We have n reduced data sets to consider all of the leave-one-out cases. To complement the numerical summary, we suggest the following (3.2) for the $D(i)$ plot which shows a graphical summary of

$$(\bar{D}) + cS_D, \quad c = 1, 2, 3, \tag{3.2}$$

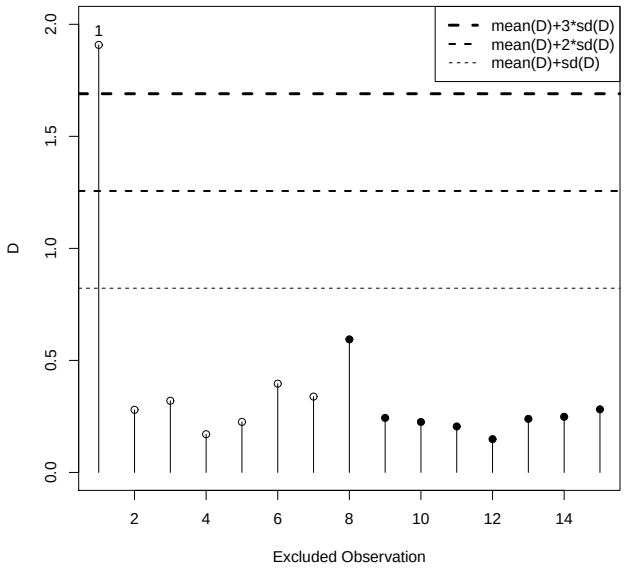


Figure 3: *D*(*i*) *plot for the lung function data set.*

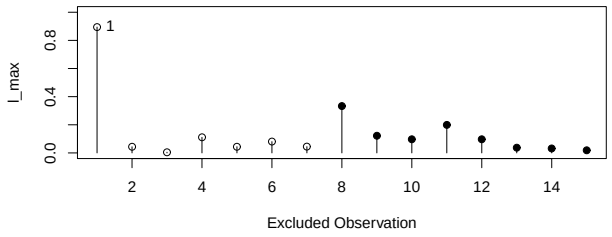
where ($\bar{D}$) and S_D is the mean and standard deviation of $D(i)$, respectively. Choosing $c = 1, 2, 3$ for reference cutoff lines may be natural to match with typical choices based on summaries involving the standard deviation. When $D(i)$ is inside the line set by $c = 1, 2$, we call the change in the posterior probability *weak* and *mild* respectively. If $D(i)$ is outside the line set by $c = 3$, we call it *extremely influential* in terms of the difference between the two posterior probabilities. These cutoff lines are intuitive and not new in the literature. See Moreno-Roldán *et al.* (2007).

In addition to the information about the possible influential and outlying observations, we illustrate the ability to look at the detailed movement of other individual data points. Figure 3 shows $D(i)$ plot for the lung data set. We conclude that child 1 is an extremely influential data point and child 8 may be weakly influential. We found a very similar pattern from both the l_{max} and B_i index plots in Figure 4. Our graphical approach is not strictly formal, but it does reflect similar information as the local influence approach.

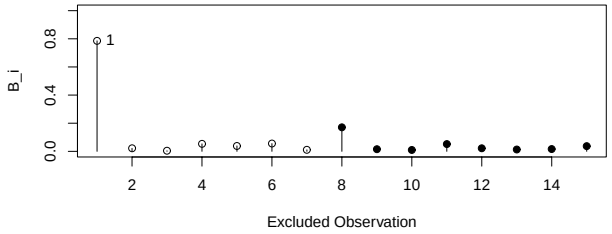
We also suggest another graphical display once observation 1 is identified as influential. We look at the movement of other data points in terms of their changing posterior probabilities. Figure 5 shows the Posterior probability movement plot, which illustrates that all of the other observations have a better separation direction once child 1 is removed. We would be unable to obtain this detailed information if we deal with the influential data points in terms of misclassification probabilities. This suggests why a more data-dependent influential measure than the overall influence measure based on $\Phi(-(1/2)\Delta)$ can be beneficial.

If observation 1 is omitted, observations 4, 5, 8, 11 show large changes in their posterior probabilities, as seen in Figure 5. For this example, no observation changes its membership, but this is not guaranteed in other cases. Even when a data point turns out to be very influential, it does not mean that its membership will automatically change. This depends on the incremental changes to the discriminant function.

The second data set considered is the Conn's syndrome data obtained from Aitchison and Dunsmore (1975) that was also analyzed by Critchley and Vitiello (1991) and Poon (2004). It consists of



(a) l_{max} index plot



(b) B_i index plot

Figure 4: *(a) l_{max} index plot and (b) B_i index plot for the lung function data set.*

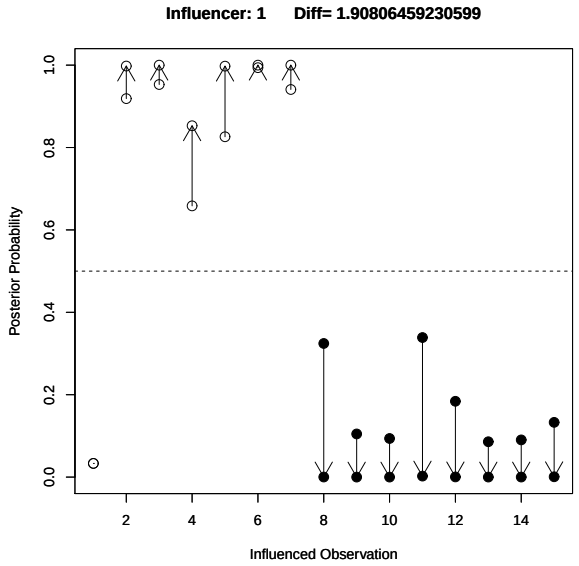


Figure 5: *Posterior probability movement plot with observation 1 omitting in the lung function data set.*

31 patients confirmed to have adenoma (group 1) and bilateral hyperplasia (group 2), $n_1 = 20, n_2 = 11$. It has two variables: the concentrations of potassium and renin in the patients' blood plasma. The scatter plot of the data in Figure 6 shows 4 observations (5, 18, 24, 27) are misclassified and several outlying values.

Figure 7 shows $D(i)$ plot for the Conn's syndrome data set. We could conclude that observations 5 and 27 are two mildly influential data points and observations 18, 21, 24 and 31 are the four weakly

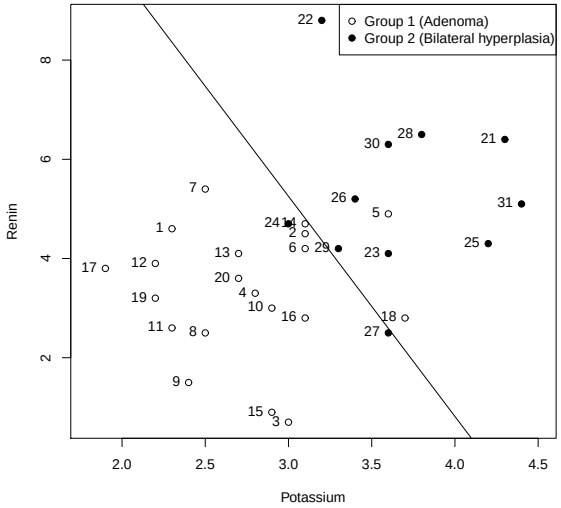


Figure 6: *Scatterplot for Conn's syndrome data set: (solid line) the discriminant function.*

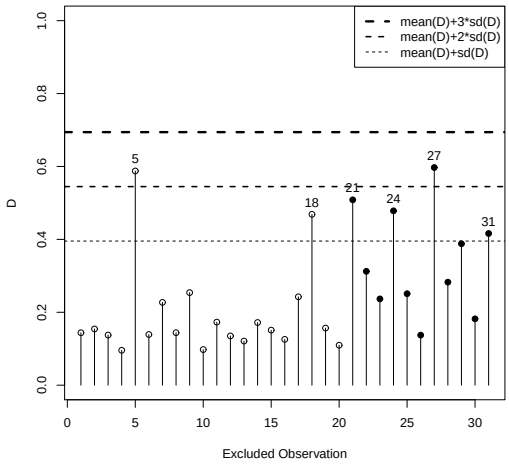
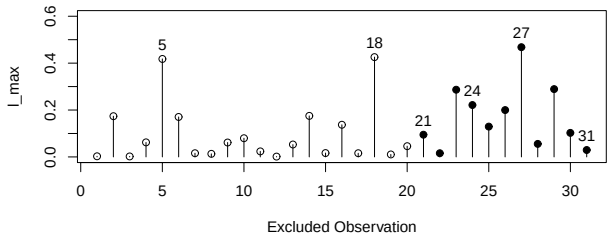


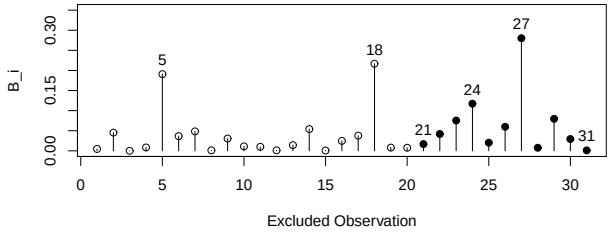
Figure 7: *D(i) plot for Conn's syndrome data set.*

influential. It is interesting that the two weak influential data points, observations 21 and 31, are extreme values of group 2 in Figure 6. Figure 8 shows the l_{max} and $B(i)$ index plots for the same data. There are different rankings of the observations based on these criteria and the plot. Figure 8 can be helpful to understand the relative contributions to the influence of the identified observations.

Figure 9 shows the corresponding posterior probability movement plots when observations 5, 18, 21, 24, 27, 31 are individually omitted from the Conn's syndrome data set. Figure 9 shows the magnitude of change and direction of other observation in terms of their posterior probabilities when a specific data point is omitted. When observations 5, 21 and 31 are omitted, observations 14 and 27 change their membership group. Figure 6 indicates that observations 14 and 27 are close to the discriminant function. Observations 21 and 31 are not influential using Cook's approach, but do prompt two observations to change their membership. Interestingly, deleting either of observations



(a) l_{max} index plot



(b) B_i index plot

Figure 8: *(a) l_{max} index plot and (b) B_i index plot for Conn's syndrome data set.*

21 and 31 results in an undesirable shift in direction for most observations in group 1, while omitting observation 5 has a desirable impact. Most observations in group 1 move outward toward 1.0 and most observations in group 2 move toward 0 with observation 5 omitted, while most observations in group 1 move inward toward 0.5 undesirably with each observations of 21, and 31 deleted.

When observation 18 is omitted, observation 27 changes its membership group with a large shift in value. When observation 24 is deleted, no observations change membership. For observation 27, observation 18 changes its membership with a large shift in value.

4. Conclusion

The local influence approach for discriminant analysis is becoming more popular since it is free of the dominating and masking effects often encountered with the leave-one-out approach in detecting influential observations. In this paper, we suggest another leave-one-out statistical measure to detect the influential data points in discriminant analysis. Despite the weakness of leave-one-out approaches, the newly proposed graphical summaries perform the similar mission as Cook's local influence approach with additional information obtained regarding the cut-off criterion. The $D(i)$ plot as well as the l_{max} and B_i index plots provide straightforward approaches to assess the impact of removing any observation in a dataset and quantifying the robustness of the results for leave-one-out subsets of the data. We think that the newly proposed plots and the existing local influence approach will complement each other.

Once an observation has been identified as potentially influential, the posterior probability movement plot provides a summary of the effect on the change of posterior probabilities for other data points. The plot also provides meaningful information about how the membership group for any observation changes and how misclassification is impacted when an observation is removed.

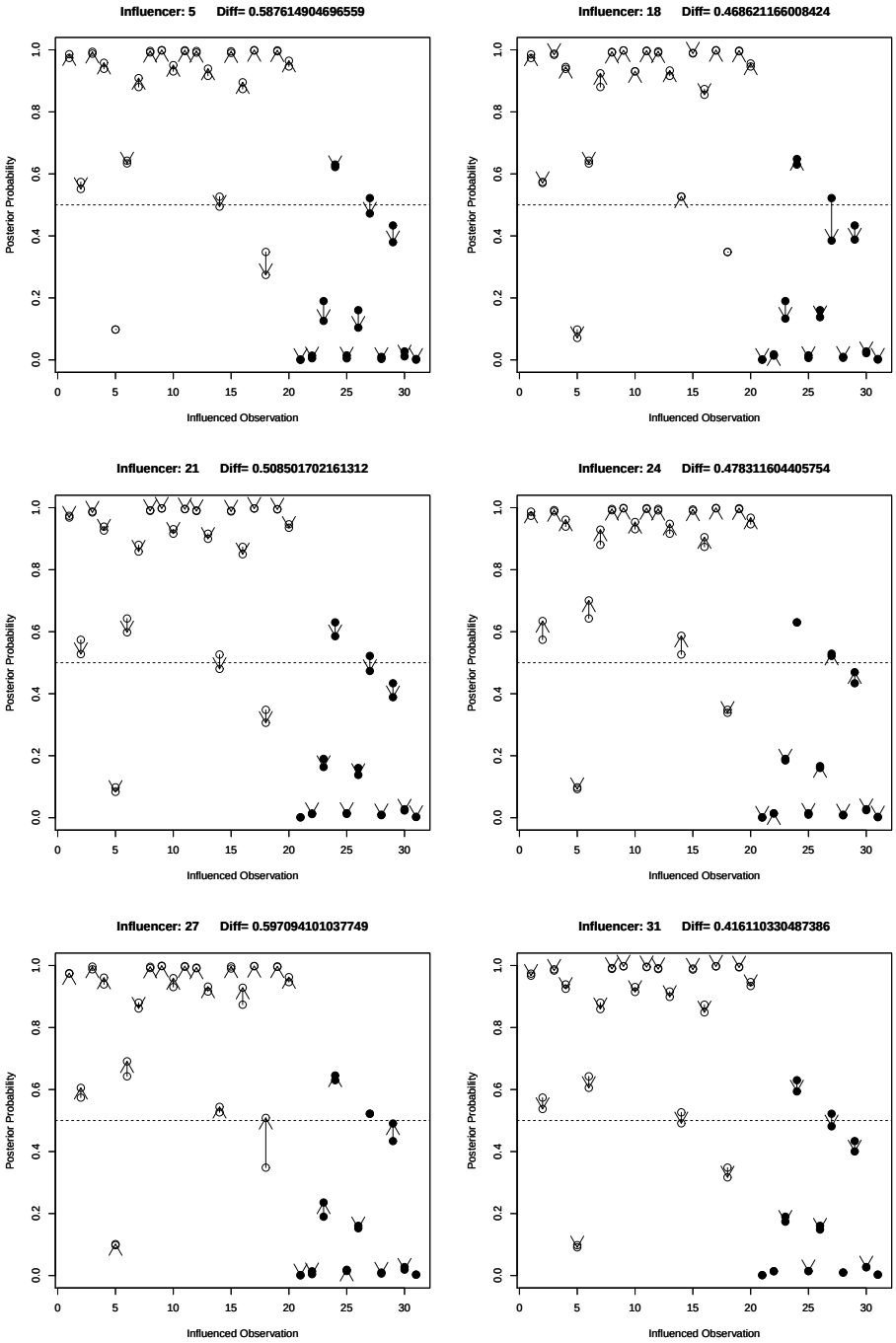


Figure 9: *Posterior probability movement plots when observations 5, 18, 21, 24, 27, 31 are omitted for the Conn's syndrome data set.*

References

Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*, Cambridge University Press, Cambridge, UK.

Campbell, M. J. (2001). *Statistics at Square Two: Understanding Modern Statistical Applications in Medicine*, Boston Medical Publishers Inc., London.

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

Cook, R. D. (1977). Detection of influential observations in linear regression, *Technometrics*, **19**, 15–18.

Cook, R. D. (1986). Assessment of local influence, *Journal of the Royal Statistical Society Series B (Methodological)*, **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.

Fung, W. K. (1995). Diagnostics in linear discriminant analysis, *Journal of the American Statistical Association*, **90**, 952–956.

Jung, K. M. (1998). Local influence assessment of the misclassification probability in multiple discriminant analysis, *Journal of the Korean Statistical Society*, **27**, 471–483.

Lachenbruch, P. A. (1997). Discriminant diagnostics, *Biometrics*, **53**, 1284–1292.

Lahiff, M. and Whitcomb, K. M. (1990). Empirical influence function for misclassification rates in discriminant analysis, *Communication in Statistics - Theory and Methods*, **19**, 2999–3009.

Lee, H. J. and Kim, H. G. (2011). Derivation and application of influence function in discriminant analysis for three groups, *The Korean Journal of Applied Statistics*, **24**, 941–949.

Moreno-Roldán, D., Muñoz-Pichardo, J. M. and Enguix-Gonzáles, A. (2007). Influence diagnostics in multiple discriminant analysis, *Test*, **16**, 172–187.

Poon, W. Y. (2004). Identifying influential observations in discriminant analysis, *Statistical Methods in Medical Research*, **13**, 291–308.

Received June 28, 2015; Revised July 21, 2015; Accepted July 28, 2015