Influence analysis on discriminant coordinates

Juan M. Muñoz-Pichardo, Alicia Enguix-Gonzalez, Joaquin Muñoz García, Juan Luis Moreno-Rebollo

Influence analysis on discriminant coordinates

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2009-0325.R2
Manuscript Type:	Original Paper
Date Submitted by the Author:	10-Jan-2011
Complete List of Authors:	Muñoz-Pichardo, Juan M.; Universidad de Sevilla, Estadistica e I.O. Enguix-Gonzalez, Alicia; Universidad de Sevilla, Estadistica e I.O. Muñoz García, Joaquin; Universidad de Sevilla Moreno-Rebollo, Juan Luis; Universidad de Sevilla, Estadistica e I.O.
Keywords:	discriminant coordinates, influence analysis, deletion approach
Abstract:	Discriminant analysis (DA), particularly Discriminant Coordinates (DC), is broadly applied in the scientific literature and included in many statistical software packages. DC is used to analyze biomedical data, especially for differential diagnosis on the basis of laboratory profiles. Papers handling influence analysis in DA can be found in the literature; however this topic has been scarcely touched upon in DC. In this paper the case-deletion approach is followed to introduce a perturbation in the data and influence measures are proposed to assess the effect on three statistics of interest: the transformation matrix, canonical directions and configuration of the sample centroids

Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

dccisscs.tex
discanCIS.bib
tandfx.bst

Influence analysis on discriminant coordinates

J. M. Muñoz-Pichardo, A. Enguix-González,

J. Muñoz-García and J.L. Moreno-Rebollo

Departamento de Estadística e I.O. Universidad de Sevilla.

Avd. Reina Mercedes sn. 41012-Sevilla. Spain.

email: juanm@us.es; aenguix@us.es, joaquinm@us.es, jlmoreno@us.es

January 10, 2011

Abstract

Discriminant analysis (DA), particularly Discriminant Coordinates (DC), is broadly applied in the scientific literature and included in many statistical software packages. DC is used to analyze biomedical data, especially for differential diagnosis on the basis of laboratory profiles. Papers handling influence analysis in DA can be found in the literature; however this topic has been scarcely touched upon in DC. In this paper the case-deletion approach is followed to introduce a perturbation in the data and influence measures are proposed to assess the effect on three statistics of interest: the transformation matrix, canonical directions and configuration of the sample centroids.

Key Words: discriminant coordinates, influence analysis, deletion approach.

1 Introduction

Discriminant analysis (DA), also named "classification" and "statistical pattern recognition", is a methodology commonly used to classify a set of observations into predefined classes or to distinguish between a set of groups or sub-populations.

DA has been broadly applied in the scientific literature in almost all scientific fields. In particular, DA has been frequently applied in medical research in order to distinguish between several diseases or between health and disease states. Many medical research papers apply DA for the differential diagnosis on the basis of a laboratory profile.

DA includes several techniques. Linear discriminant analysis (LDA), DC (also named canonical variates), and logistic discrimination are probably the most frequently used in differential diagnosis. Mirkin *et al.* (2004), DC in a study of human endometrium; Galanaud *et al.* (2006), DA for a noninvasive diagnostic assessment of brain tumors; Struijk *et al.* (2006), a new method for discriminating between people with a normal genotype and those with the congenital long-QT syndrome; Guo *et al.* (2007) and Perelman *et al.* (2003), DA in microarray data, all illustrate relevant applications of DA in biomedical research.

Generally speaking, influence analysis (IA) deals with the study and assessment of the variations caused in statistical conclusions by perturbations. Several perturbation schemes can be considered, although case-deletion may be the most commonly used in IA. Papers handling IA in almost all statistical techniques can be found in the literature.

Obviously, accuracy of the estimation and the classification in DA might be affected by outliers and influential observations. This fact justifies the interest of IA in DA. The study of influence in linear discriminant analysis is normally carried out by using the common case-deletion approach, and usually by assessing its effect on the estimated total probability of misclassification. This approach is applied in the following papers: Campbell (1978), Critchley and Vitiello (1991), Fung (1992), Fung (1995c), Fung (1995b) and Fung (1996). Recently, Moreno-Roldán *et al.* (2007) proposed two case-deletion diagnostics based on the L_2-norm which evaluate the effect of the omission on the linear functions which determine Fisher's linear discriminant rule. Riani and Atkinson (2001) provided a unified approach to study influential observations and outliers in Quadratic Discriminant Analysis.

As far as we know, there is no paper which deal with influence in discriminant analysis focused on DC in the literature, perhaps because analysis based on DC and LDA

lead to identical results if the complete set of discriminant coordinates is considered. However, for convenience of the subsequent analysis, in particular for the graphical representation of the transformed feature data, only the first two discriminant coordinates are often considered in practice. The results obtained from DC and LDA might differ in such a case.

In this paper, case-deletion diagnostics are proposed on three statistics of interest in DC when only the first two discriminant coordinates are considered: the transformation or projection matrix, the directions of the projection matrix and the configuration of the sample centroids of the first two discriminant coordinates.

- *The transformation matrix.* Two diagnostic measures related to this statistic are proposed. The first one assesses the effect of the omission through a matrix norm, the classic norm of Frobenius, and the second one through a ratio of determinants.
- *The directions of the projection.* In this case, the effect of the perturbation is measured through the angle between the non-perturbed and perturbed directions of the projections.
- *The configuration of the sample centroids.* The Euclidean distance between the non-perturbed and perturbed centroids is considered as influence measure on this statistic.

The paper is organized as follows. In the next section, the notation, the discriminant rule and the basic statistics in canonical discriminant analysis are introduced. The influence measures that we propose are presented in Section 3. In Section 4 these influence measures are illustrated with two data sets. Finally, some conclusions are set out in Section 5.

2 Discriminant coordinates

First a short introduction to DC, which serves as a means to establish the notation, is given. Let $\{G_i,\ i = 1, \ldots, g\}$ be g mutually exclusive groups or populations and let $\underline{X} = (X_1, ..., X_p)^t$ be a p-dimensional random vector. We assume that the distribution

of $\underline{X}$ for G_i is $N_p\left(\underline{\mu}_i, \mathbf{\Sigma}\right)$, $i = 1, ..., g$.

The basic question in DC is to determine linear combinations of $X_1, ..., X_p$ (discriminant coordinate variables) which reflect the differences between groups, that is, the linear transformations which lead to the greatest separation among the mean vectors. To this end, a sample of size n_i ("training" data) is selected from each population, $\left\{\underline{\mathbf{x}}_{ij} \in \mathbb{R}^p \ : \ j = 1, \ldots, n_i\right\}$, $i = 1, \ldots, g$. The sample means of the groups are denoted by $\overline{\underline{\mathbf{x}}}_{i\bullet}$, $i = 1, \ldots, g$ and the global sample mean by $\overline{\underline{\mathbf{x}}}$, that is, $\overline{\underline{\mathbf{x}}} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\underline{\mathbf{x}}_{ij}$, where $n = \sum_i^g n_i$. $\mathbf{W}$ and $\mathbf{B}$ denote the within-group and between-group matrices, respectively, that is,

$$\mathbf{W} = \sum_{i=1}^{g}\sum_{j=1}^{n_i}(\underline{\mathbf{x}}_{ij} - \overline{\underline{\mathbf{x}}}_{i\bullet})(\underline{\mathbf{x}}_{ij} - \overline{\underline{\mathbf{x}}}_{i\bullet})^t \quad \text{and} \quad \mathbf{B} = \sum_{i=1}^{g} n_i(\overline{\underline{\mathbf{x}}}_{i\bullet} - \overline{\underline{\mathbf{x}}})(\overline{\underline{\mathbf{x}}}_{i\bullet} - \overline{\underline{\mathbf{x}}})^t.$$

The computation of the vectors (coefficients of the linear combinations), $\underline{\mathbf{c}}_k$, $k = 1, ..., d$ ($d = \min\{p, g-1\}$), that determine the discriminant coordinates is well-covered in standard textbooks on multivariate analysis (see, for instance, Seber (1984) (Chapter 5)).

The vectors $\underline{\mathbf{c}}_k$, $k = 1, ..., d$, are the solutions of the **optimization problems**

$$\sup_{\underline{\mathbf{c}} \neq \underline{0}} \frac{\underline{\mathbf{c}}^t\mathbf{B}\underline{\mathbf{c}}}{\underline{\mathbf{c}}^t\mathbf{W}\underline{\mathbf{c}}} \quad \text{for } k = 1 \text{ and}$$

$$\sup_{\underline{\mathbf{c}} \neq \underline{0},\ \underline{\mathbf{c}}_j^t\mathbf{W}\underline{\mathbf{c}} = 0,\ j=1,\ldots,k-1} \frac{\underline{\mathbf{c}}^t\mathbf{B}\underline{\mathbf{c}}}{\underline{\mathbf{c}}^t\mathbf{W}\underline{\mathbf{c}}} \quad \text{for } \ k = 2, \ldots, d. \tag{1}$$

If $\lambda_1 \geq \ldots \geq \lambda_d$ are the non-zero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ with associated unit eigenvectors $\{\underline{\mathbf{e}}_1, \ldots, \underline{\mathbf{e}}_d\}$, then the k^{th} discriminant coordinate variate is given by $\underline{\mathbf{c}}_k^t\underline{X}$ where $\underline{\mathbf{c}}_k = (\frac{n-g}{h_k^2})^{1/2}\underline{\mathbf{e}}_k$ and $h_k^2 = \underline{\mathbf{e}}_k^t\mathbf{W}\underline{\mathbf{e}}_k$. Then the discriminant coordinates have null sample correlations, $\underline{\mathbf{c}}_k^t\mathbf{W}\underline{\mathbf{c}}_r = 0$, $k \neq r$. Given $k \leq d$, $\mathbf{C}_k = [\underline{\mathbf{c}}_1 \ldots \underline{\mathbf{c}}_k]^t$ is called the transformation or projection matrix. The centroids of the groups in the transformed space are $\overline{\underline{\mathbf{z}}}_{\mathbf{i}\bullet} = \mathbf{C}_k\overline{\underline{\mathbf{x}}}_{i\bullet}$, $i = 1, \ldots, g$.

A future observation $\underline{\mathbf{x}}$ is assigned to the i^{th}-group if

$$d_E\left(\mathbf{C}_k\underline{\mathbf{x}}, \mathbf{C}_k\overline{\underline{\mathbf{x}}}_{i\bullet}\right) = \min_{l=1,\ldots,g} d_E\left(\mathbf{C}_k\underline{\mathbf{x}}, \mathbf{C}_k\overline{\underline{\mathbf{x}}}_{l\bullet}\right),$$

where d_E denotes the vector Euclidean distance.

At this point it should be borne in mind that the classification rule based on DC is derived under two assumptions: normality and homoscedasticity. Under these assumptions the discriminant coordinates are obtained by maximizing the F-ratio statistic of the analysis of variance. As this technique is moderately robust to longer-tailed symmetric distributions, the classification rule can be used under moderate violation of the normality assumption.

The discriminant coordinates are determined in order to emphasize the separation between groups, but with decreasing effectiveness. From a practical point of view, it is necessary to fix the number k of discriminant coordinates to be considered in the statistical analysis. The relative magnitudes of the eigenvalues $\lambda_1, ..., \lambda_d$ are frequently used to this end. For $k = 2$ and $k = 3$, plots of the transformed data are helpful to study the degree and nature of the separation between the groups. This option of the discrimination problem is provided by several statistical packages which allow DC to be broadly applied in statistical data analysis. Obviously, considering a dimension of the transformed space greater than 3 is not adequate to display the behavior of the groups and the data properly, although plots of pairs of coordinates still can be explored.

In this paper, only $k = 2$ is considered although the diagnostics that we propose can easily be generalized to any value of k.

Hence, the most relevant statistics in DC are the transformation matrix $\mathbf{C}_2$, the directions that determine the linear transformation, that is, the rows $\underline{\mathbf{c}}_1$ and $\underline{\mathbf{c}}_2$ of $\mathbf{C}_2$, and the sample means in the transformed space $\overline{\underline{\mathbf{z}}}_{\mathbf{i}\bullet}$, $i = 1, \ldots, g$.

3 Influence measures

In the following, influence diagnostics are proposed in order to assess the effect of this perturbation on the most relevant statistics in DC.

From now on, for any sample statistic T, $T_{(s,l)}$ denotes the statistic perturbed by deleting $\underline{\mathbf{x}}_{s,l}$, the l^{th} case of the s^{th} group, from the sample. In particular, $\overline{\underline{\mathbf{x}}}_{r\bullet(s,l)} = \overline{\underline{\mathbf{x}}}_{r\bullet}$ if $r \neq s$, $\overline{\underline{\mathbf{x}}}_{s\bullet(s,l)} = \overline{\underline{\mathbf{x}}}_{s\bullet} - \frac{1}{n_s - 1}(\underline{\mathbf{x}}_{s,l} - \overline{\underline{\mathbf{x}}}_{s\bullet})$, $\mathbf{C}_{2(s,l)}$ is determined by the eigenvectors of $\mathbf{W}^{-1}_{(s,l)}\mathbf{B}_{(s,l)}$ corresponding to the two largest eigenvalues, and $\overline{\underline{\mathbf{z}}}_{r\bullet(s,l)} = \mathbf{C}_{2(s,l)}\overline{\underline{\mathbf{x}}}_{r\bullet(s,l)}$.

3.1 Transformation matrix

Two diagnostics are proposed to quantify the effect caused on the transformation matrix $\mathbf{C}_2$ by the omission of $\underline{\mathbf{x}}_{s,l}$:

- The Frobenius norm of the difference matrix $\mathbf{C}_2 - \mathbf{C}_{2(s,l)}$,

$$M_{(s,l)} = \left\|\mathbf{C}_2 - \mathbf{C}_{2(s,l)}\right\|_F,$$

where $\|\mathbf{A}\|_F = \sum_i \sum_j a_{ij}^2$ for any matrix $\mathbf{A} = [(a_{ij})]$. Obviously, $M_{(s,l)}$ can be considered as an influence diagnostic of the effect of the omission of $\underline{\mathbf{x}}_{s,l}$ on $\mathbf{C}_2$. This diagnostic is called "M-measure". Large values of M correspond to cases that lead to large changes in the projection matrix upon which the discriminant rule is based.

- The aim of canonical discriminant analysis is to determine the linear transformations of the vector of the observed variables that lead to the greatest separation between the groups in the transformed space. Therefore, measuring the effect that the deletion of each observation exerts on discrimination ability of the discriminant coordinates is of interest.

  The ratio between the between- and within-group dispersion is a measure of discrimination ability, and the determinant of the dispersion matrix of the discriminant coordinates, $\frac{1}{n-g}\mathbf{C}_2\mathbf{W}\mathbf{C}_2^t$, is a measure of the within-group dispersion in the transformed space. Therefore, if the omission of an observation causes a significant change in the value of the determinant, it also will provoke a significant change in discrimination ability. Since DC leads to standardized discriminant coordinates, $\left|\frac{1}{n-g}\mathbf{C}_2\mathbf{W}\mathbf{C}_2^t\right| = 1$, we also propose

$$R_{(s,l)} = \left|\frac{1}{n-g}\mathbf{C}_{2(s,l)}\mathbf{W}\mathbf{C}_{2(s,l)}^t\right|$$

as an influence measure of the omission of $\underline{\mathbf{x}}_{(s,l)}$ on the dispersion matrix. We will call "R-measure" this diagnostic.

Taking into account that $\mathbf{W} = \mathbf{W}_{(s,l)} + (n_s/(n_s-1))(\underline{\mathbf{x}}_{s,l} - \overline{\underline{\mathbf{x}}}_{s\bullet})(\underline{\mathbf{x}}_{s,l} - \overline{\underline{\mathbf{x}}}_{s\bullet})^t$ R-measure can be expressed by

$$R = \left(1 - \frac{1}{n-g}\right)^2 \left(1 + \underline{\mathbf{u}}_{s,l}^t \underline{\mathbf{u}}_{s,l}\right)$$

where $\underline{\mathbf{u}}_{s,l} = \left(\frac{n_s}{(n-s-1)(n-g-1)}\right)^{1/2}\left(\underline{\mathbf{z}}_{s,l(s,l)} - \overline{\underline{\mathbf{z}}}_{s\bullet(s,l)}\right)$, with $\underline{\mathbf{z}}_{s,l(s,l)} = \mathbf{C}_{2(s,l)}\underline{\mathbf{x}}_{s,l}$. Therefore, values of R far from 1 are associated with influential observations. Those verifying $\left(R_{(s,l)} - 1\right) > 0.05$ or 0.10 are proposed as influential observations.

In should be pointed out here that the determinant ratio approach has previously been applied in the literature as an influence diagnostic and to identify outliers. For example, Barnett and Lewis (1978) and Munoz-García *et al.* (1990) used this approach to identify outliers in a sample from a q-dimensional normal population, while Belsley *et al.* (1980) proposed an influential diagnostic in linear regression.

3.2 Directions of the projection matrix

The M-measure quantifies the effect of the omission on the coefficients of the linear projection as a whole. However, separate conclusions about its effect on the directions $\underline{\mathbf{c}}_1$ and $\underline{\mathbf{c}}_2$ cannot be obtained using this measure. The directions $\underline{\mathbf{c}}_1$ and $\underline{\mathbf{c}}_2$ are not orthogonal ($\underline{\mathbf{c}}_1^t\ \underline{\mathbf{c}}_2 \neq 0$) but uncorrelated ($\underline{\mathbf{c}}_1^t\mathbf{W}\underline{\mathbf{c}}_2 = 0$). Therefore, the study of the effect of the omission on $\underline{\mathbf{c}}_1$ and $\underline{\mathbf{c}}_2$ has to be carried out separately.

We propose the angle between $\underline{\mathbf{c}}_j$ and $\underline{\mathbf{c}}_{j(s,l)}$, $j = 1, 2$, as influence measures of the omission of $\underline{\mathbf{x}}_{(s,l)}$ on the direction $\underline{\mathbf{c}}_j$,

$$A^{(1)}_{(s,l)} = \left\langle \underline{\mathbf{c}}_1, \underline{\mathbf{c}}_{1(s,l)} \right\rangle = \frac{100}{\pi}\arccos\frac{\underline{\mathbf{c}}_1^t\underline{\mathbf{c}}_{1(s,l)}}{\|\underline{\mathbf{c}}_1\|\left\|\underline{\mathbf{c}}_{1(s,l)}\right\|},$$
$$A^{(2)}_{(s,l)} = \left\langle \underline{\mathbf{c}}_2, \underline{\mathbf{c}}_{2(s,l)} \right\rangle = \frac{100}{\pi}\arccos\frac{\underline{\mathbf{c}}_2^t\underline{\mathbf{c}}_{2(s,l)}}{\|\underline{\mathbf{c}}_2\|\left\|\underline{\mathbf{c}}_{2(s,l)}\right\|}.$$

We will call "$A^{(1)}$-measure" and "$A^{(2)}$-measure", respectively, these diagnostics. For convenience, $A^{(1)}$ and $A^{(2)}$ are stated in hexadecimal degrees, and therefore $A^{(1)}, A^{(2)} \in [0, 100]$. This fact enables reference values to be fixed in such a way that those cases with "A-measures" greater than the reference values would be considered as influential observations. The reference values for high and moderate influence are fixed at 5 and 2.5 respectively.

It should be noted that $\underline{\mathbf{c}}_{1(s,l)}$ is an eigenvector corresponding to the largest eigenvalue of $\mathbf{W}^{-1}_{(s,l)}\mathbf{B}_{(s,l)}$, thereby verifying the condition $\underline{\mathbf{c}}^t_{1(s,l)}\mathbf{W}_{(s,l)}\underline{\mathbf{c}}_{1(s,l)} = n - g - 1$.

Obviously, the opposite vector $-\underline{\mathbf{c}}_{1(s,l)}$ also verifies this condition. Therefore $\underline{\mathbf{c}}_{1(s,l)}$ has to be chosen appropriately so that the effect of the omission can be accurately assessed. A similar consideration has to be made for $\underline{\mathbf{c}}_{2(s,l)}$.

3.3 Centroids

We propose assessing the effect of the omission of $\underline{\mathbf{x}}_{(s,l)}$ on the centroids through the sum of the Euclidean distances between the perturbed $\{\overline{\underline{\mathbf{z}}}_{r\bullet(s,l)}, r = 1, \ldots, g\}$, and the non-perturbed centroids $\{\overline{\underline{\mathbf{z}}}_{r\bullet}, r = 1, \ldots, g\}$,

$$D_{(s,l)} = \left[\sum_{r=1}^{g} \left(\overline{\underline{\mathbf{z}}}_{r\bullet(s,l)} - \overline{\underline{\mathbf{z}}}_{r\bullet}\right)^t \left(\overline{\underline{\mathbf{z}}}_{r\bullet(s,l)} - \overline{\underline{\mathbf{z}}}_{r\bullet}\right)\right]^{1/2}.$$

We will call "D-measure" this diagnostic. Obviously, a large value of $D_{(s,l)}$ means a significant change in the centroid's configuration and in consequence a significant change in the classification rule. Therefore, cases with large values in the $D-$measure can be considered as influential observations.

To conclude this section, it should be taken into account that the diagnostics here proposed can obviously be generalized if $k > 2$ canonical discriminant variates are necessary to obtain a discriminant rule with admissible error rates.

4 Applications

In this section, two examples with real data are presented to illustrate the diagnostics proposed. In the first example the Lubischew data set (Lubischew, 1962) is used. This data set has been considered by other authors in several studies on discriminant analysis. See for example, Moreno-Roldán *et al.* (2007), Bremner and Taplin (2002), Fung (1995a), Schott (1990), and McKay (1977). Two variables from the Lubischew data set have been selected, thereby enabling a simple display and facilitating the interpretation of the measures proposed. The second example is an application of discriminant analysis to a medical data set (Plomteux, 1980). This medical application attempts to determine a differential diagnosis of diseases of the liver on the basis of a laboratory profile determined by liver enzymes.

Table 1: Lubischew data set. Results of the classification rule.

	Predicted group			
Actual group	G_1	G_2	G_3	Total
G_1	16 (76.2%)	0	5 (23.8%)	21
G_2	0	31 (100%)	0	31
G_3	2 (9.1%)	0	20 (90.9%)	22

4.1 Lubischew data set

This application has been included due to two reasons. Firstly, the Lubischew data set has been widely used in the scientific literature. Secondly, it simplifies the interpretation of the measures here proposed. Lubischew (1962) analyzed three groups of genus of flea beetle: Chaetocnema Concinna (G1), Chaetocnema Heikertingeri (G2) and Chaetocnema Heptapotamica (G3). The groups consist of 21, 31 and 22 observations, respectively. They are labeled as cases 1 to 21, 22 to 52, and 53 to 74, respectively. Each observation consists of six variables from which two have been selected: the fourth variable (the maximal width of the aedeagus in the fore-part in microns) and the sixth variable (the aedeagus width from the side in microns).

Figures 1(a) and (b) display the original data set and the transformed data set through the discriminant coordinates, respectively.

The resultant eigenvectors of the analysis are $\underline{\mathbf{c}}_1^t = [0.1475, 0.0690]$ and $\underline{\mathbf{c}}_2^t = [-0.1650, 0.1268]$. Table 1 shows the results of the classification through the discriminant coordinates, whereby 90.95% of cases are correctly classified. Cases 6, 8, 9, 16 and 17 of G_1 are misclassified in G_3, and cases 53 and 66 of G_3 are misclassified in G_1 through DC. Obviously, these results reflect what Figures 1(a) and (b) show: cases in G_2 are clearly separated from those in G_1 and G_3; but some cases corresponding to G_1 and G_3 are mixed.

Figure 2 (a)-(d) display the index plots of the diagnostics proposed in Section 3: M-measure, D-measure, $A^{(1)}$-measure and R- measure. The $A^{(2)}$-measure plot is omitted

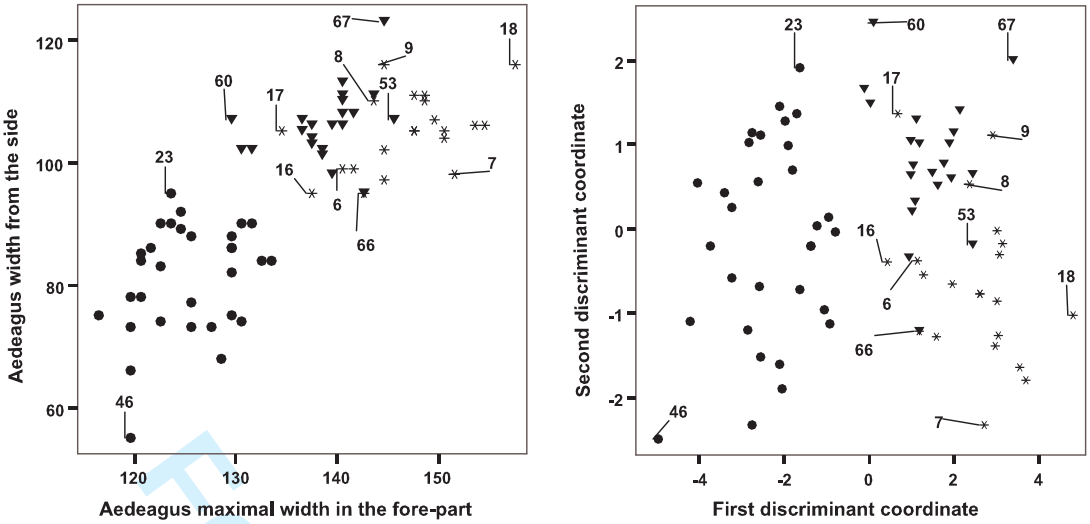
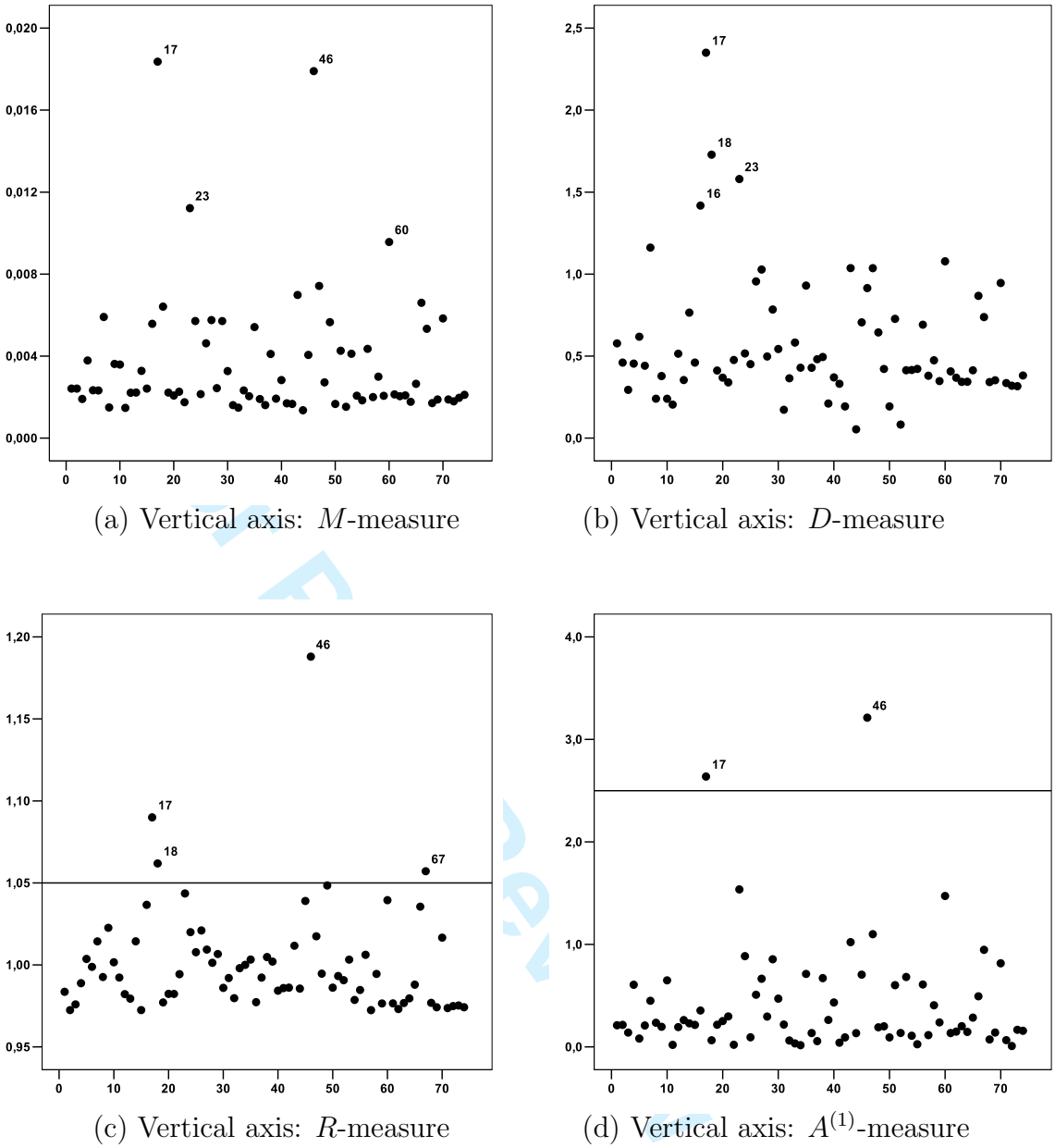


Figure 1: Lubischew data set: Scatter plots. $(*)$-G_1; $(\bullet)$-G_2; $(\blacktriangledown)$-G_3

since no case is identified as influential from this measure. Table 2 summarizes the results of the influence analysis through these diagnostics.

From the results obtained by applying the diagnostics proposed, it can be concluded **that**:

- Cases 17 and 46 are the most influential. Case 17 is identified as influential by the four diagnostics and case 46 by three diagnostics. From Figure 1 (a), it is observed that cases 17 and 46 are extreme values in G_1 and G_3, respectively. Cases 16, 18, 23, 60 and 67 can also be considered as influential observations.
- Different cases are identified by the measures proposed here. That is, the influence measures M, D, $A^{(1)}$, $A^{(2)}$ and R provide different information. Each one of these measures assesses the effect of the omission on a different statistic of interest. Hence each measure provides useful information to the researcher.
- The reader could consider the R and A measures as the most interesting diagnostics since they have reference values to determine influential cases. However the information that they provide may not be comprehensive. For example, cases 16, 23 and 60 are not identified by these measures, but they are identified by M and



(a) Vertical axis: M-measure (b) Vertical axis: D-measure

(c) Vertical axis: R-measure (d) Vertical axis: $A^{(1)}$-measure

Figure 2: Lubischew data set. Index plot of influence measures (in all graphics, horizontal axis: cases)

Table 2: Lubischew data set. Summary of the influence measures.

Case	M-measure	D-measure	$A^{(1)}$-measure	R-measure
16		*		
17	**	**	*	*
18		*		*
23	*	*		
46	**		*	**
60	*			
67				*

(**) High influence; (*) moderate influence.

D measures.

- It could be suspected that outliers and influential cases coincide. However, this fact is not necessarily true as the results of this example illustrate. There are influential cases which are not outliers. For example, case 23, identified as influential by M and D measures, cannot be considered as an outlier in G_2. On the other hand, there are also outliers (case 7) which cannot be considered as influential observations, see Fig.1(a).

- Associating influential observations with misclassified cases is erroneous. Cases 6, 8, 9, 53 and 66 are misclassified; but they are not identified as influential. On the other hand, there are influential cases, 18, 23, 46, 60 and 67, which are correctly classified.

As previously mentioned, the Lubischew data set has been used in the literature to illustrate influence diagnostics in discriminant analysis. At this point, it should be noted that cases 16, 17, 66 and 67 (high influence) and cases 9, 46 and 60 (moderate influence) were identified as influential observations on the estimated probability of misclassification by Fung (1995a); and cases 16, 17, 18, 46, and 67 were identified as influential observations on the linear discriminant rule by Moreno-Roldán *et al.*

(2007). Although diagnostics on a different discriminant technique are proposed, our results are similar, but not totally identical, to those obtained by Fung and Moreno-Roldán. This fact shows that the diagnostics proposed here can be considered as useful complementary influence measures in discriminant analysis.

4.2 Plomteux data set

In this example, the influence measures proposed are illustrated in a sample data set with 4 groups. The data consists of 218 patients with liver diseases (Plomteux, 1980). Four diseases are considered: acute viral hepatitis (AVH) ($n_1 = 57$ patients), persistent chronic hepatitis (PCH) ($n_2 = 44$), aggressive chronic hepatitis (ACH) ($n_3 = 40$) and post-necrotic cirrhosis (PNC) ($n_4 = 77$). The diagnosis of AVH was carried out by biological and clinical signs. PCH, ACH and PNC were diagnosed by laparoscopy and biopsy. For convenience, the cases corresponding to each group are labeled by 1 to 57, 58 to 101, 102 to 141, and 142 to 218, respectively. The data are reproduced in Albert and Harris (1987).

The aim of this study is to obtain a differential diagnosis of the four liver diseases considered by means of an enzyme profile.

Plomteux (1980) showed that good discrimination between the four diseases could be achieved on the basis of three liver function tests: aspartate aminotransferase (Y_1), alanine aminotransferase (Y_2), and glutamate dehydrogenase (Y_3) (all expressed in international units per litre). The observed variables Y_i have been transformed, $X_i = \ln Y_i,\ i = 1, 2, 3$, to verify the normality assumption. See Albert and Harris (1987) (**Chapter 5, p. 113**) for more details on the results of this research. The Plomteux data set has been considered in the literature to illustrate several statistical techniques, Lesaffre and Albert (1989), Lesaffre and Albert (1988), and Bull *et al.* (1994).

Figure (3) displays the scatter plots of the original variables in pairs and the scatter plot of the first two discriminant coordinates; cases that are identified as influential below have been labeled.

The eigenvectors corresponding to the first two discriminant coordinates are: $\underline{\mathbf{c}}_1^t = [1.8547, -2.7866, 0.4956]$ and $\underline{\mathbf{c}}_2^t = [1.3015, -0.3410, 0.7735]$.

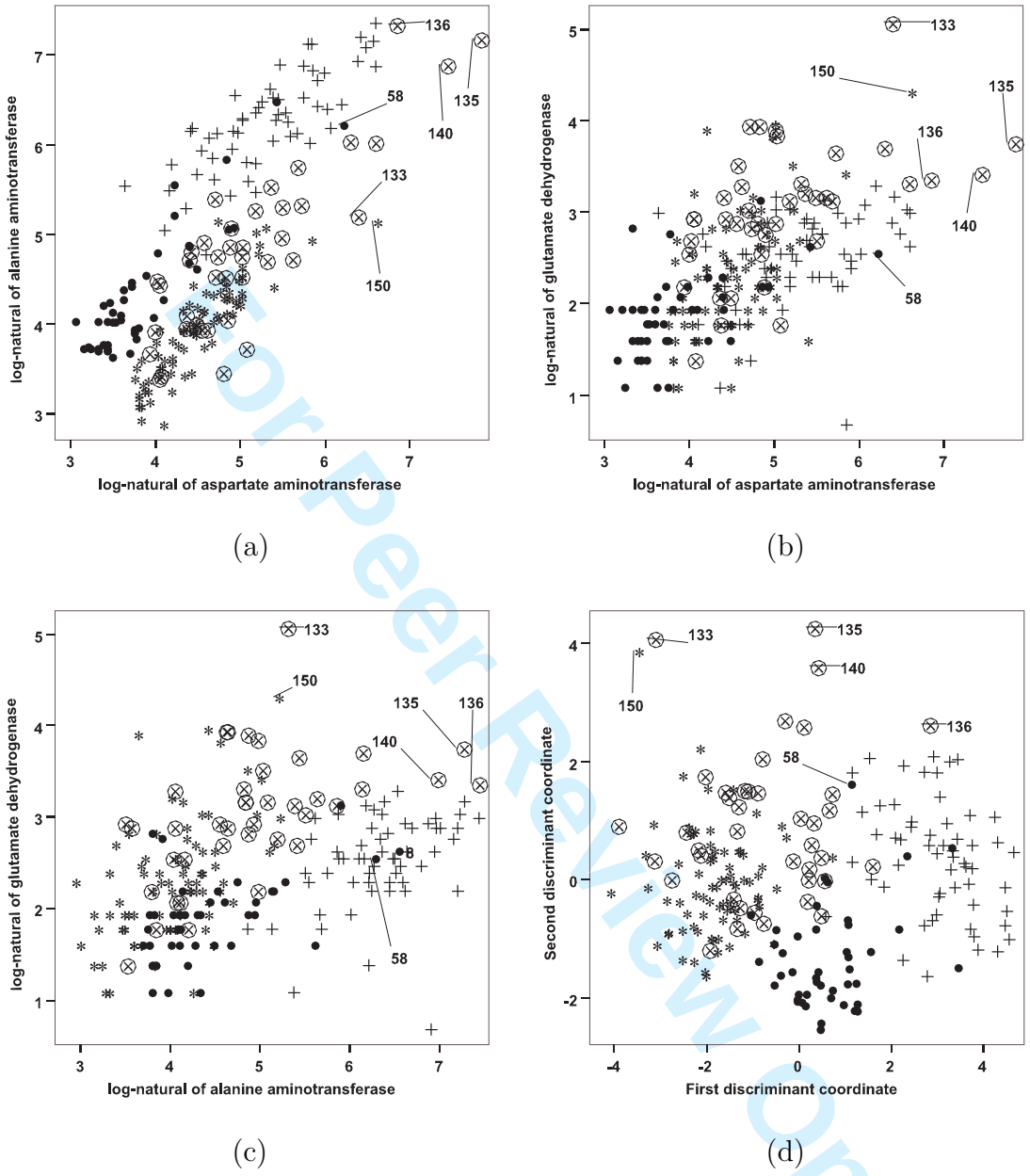


Figure 3: Plomteux data set. Scatter plots (+)-AVH; (•)-PCH; (⊗)-ACH; (∗)-PNC

Table 3: Plomteux data set. Descriptive statistics of log-transformation of liver enzyme levels

	Disease groups			
Logarithm of liver enzyme levels	AVH $n = 57$	PCH $n = 44$	ACH $n = 40$	PNC $n = 77$
X_1 (aspartate aminot.)	5.24 ± 0.67	3.79 ± 0.64	4.77 ± 0.93	4.44 ± 0.52
	$(5.28, 0.70)$	$(3.60, 0.40)$	$(4.76, 0.71)$	$(4.41, 0.55)$
X_2 (alanine aminot.)	6.26 ± 0.56	4.35 ± 0.69	4.82 ± 0.95	3.97 ± 0.54
	$(6.29, 0.50)$	$(4.09, 0.51)$	$(4.76, 0.91)$	$(3.99, 0.56)$
X_3 (glutamate dehyd.)	2.52 ± 0.51	1.88 ± 0.44	3.03 ± 0.71	2.32 ± 0.66
	$(2.56, 0.49)$	$(1.95, 0.51)$	$(3.09, 0.53)$	$(2.30, 0.60)$
	First line: mean±S.D. Second line: (median, standardized MAD)			

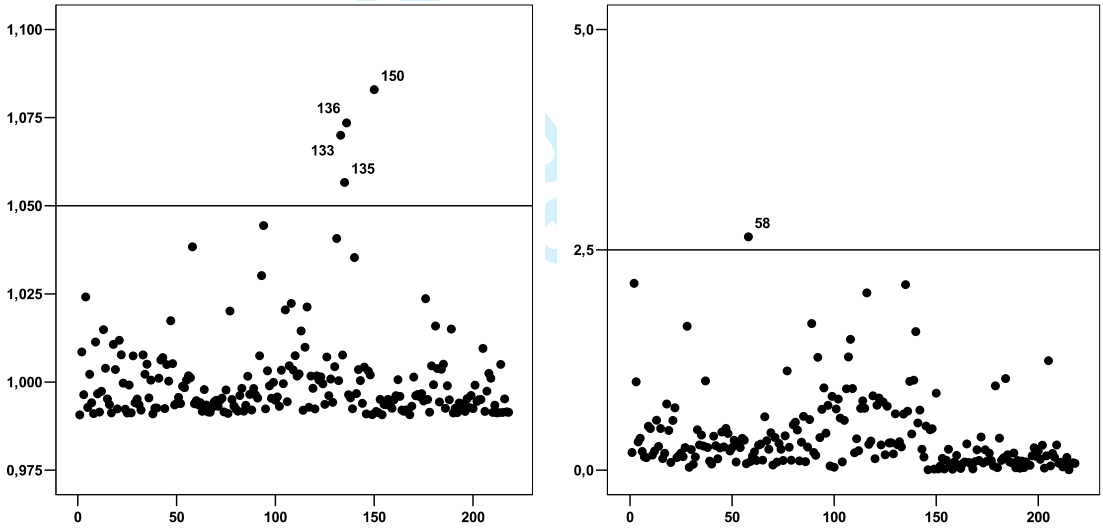
Table 4 summarizes the results obtained by DC through the canonical discriminant variates. The overall classification rate is 80.3%; there are 43 misclassified cases (19.7%), while 93.0% of AVH cases, 86.4% of PCH cases and 79.2% of PNC cases are correctly classified. However, only 57.5% of ACH cases are correctly classified. Aggressive chronic hepatitis and post-necrotic cirrhosis constitute two groups that are difficult to discriminate between. Albert and Harris (1987) stated on this point, "this is not surprising because even histological criteria cannot always clearly distinguish these two disorders".

Figure 4 (a)-(b) displays the index plot of $R-$ and $A^{(2)}-$ measures, respectively. It can be observed that cases 58, 133, 135, 136 and 150 are identified as influential observations. Omission of case 58 significantly affects the direction of the second canonical variate. Omission of cases 133, 135, 136 and 150 affects the determinant of the covariance matrix of the canonical discriminant variates. No case is identified as influential from $M-$, $D-$ or $A^{(1)-}$ measures.

As in the first example, influential cases have been related to outliers and mis-

Table 4: Plomteux data set. Results of the classification rule

	Predicted group				
Actual group	AVH	PCH	ACH	PNC	Total
AVH	53 (93.0%)	2 (3.5%)	1 (1.8%)	1 (1.8%)	57
PCH	4 (9.1%)	38 (86.4%)	1 (2.3%)	1 (2.3%)	44
ACH	1 (2.5%)	5 (12.5%)	23 (57.5%)	11 (27.5%)	40
PNC	0	1 (1.3%)	15 (19.5%)	61 (79.2%)	77



(a) Vertical axis: R-measure (b) Vertical axis: $A^{(2)}$-measure

Figure 4: Plomteux data set. Index plot of influence measures (in both graphics, horizontal axis: cases)

classified cases. Analogous conclusions are obtained. In particular, there are several misclassified cases that are not influential observations. On the other hand, there are influential observations (cases 133 and 135) that are correctly classified by the assignment rule.

Case 58 is a PCH patient whose enzyme profile $(6.16; 6.23; 2.56)$ is remarkably higher than those in his/her group with respect to the two first components (aspartate aminotransferase and alanine aminotransferase), see Figure 3(a), (b). Therefore, case 58 could be considered as an outlying observation in the PCH group with respect to X_1 and X_2. This enzyme profile is similar to those patients in the AVH group, see Figure 3 (a)-(c). This enzyme profile is closer to the enzyme profile mean of the AVH patients (5.24; 6.26; 2.52) than that of this case's own group (3.79; 4.35,1.88). Moreover, case 58 is misclassified in the AVH group. In short, case 58 is a PCH patient who is misclassified in the AVH group because the profile presents outlying values with respect to the aspartate aminotransferase and alanine aminotransferase levels in this group.

Case 133 is an ACH patient whose enzyme profile is $(6.29; 5.21; 5.08)$. It presents a considerably high glutamate dehydrogenase level (X_3) in its group, (see Figure 3 (b),(c)). Its glutamate dehydrogenase level (5.08) is greater that the mean (3.03) plus two times the standard deviation (0.71) in its group. However, no extreme values are presented with respect to the other two enzymes. In spite of its high glutamate dehydrogenase level, case 133 is correctly classified. That is, case 133 is a correctly classified ACH patient with an outlying glutamate dehydrogenase level.

Cases 135 and 136 are ACH patients whose enzyme profiles, $(7.74; 7.18; 3.76)$ and $(6.75; 7.35; 3.00)$, respectively, are somehow similar. Both of them present remarkably high aspartate aminotransferase (X_1) and alanine aminotransferase (X_2) levels in the ACH group whose mean profile level is $(4.77; 4.82; 3.03)$. Their aspartate aminotransferase and alanine aminotransferase levels are closer to the corresponding mean values in the AVH group, 5.24 and 6.26 respectively. However, it should be noted that case 135 is correctly classified and case 136 is misclassified in the AVH group. The reason for their different behavior with respect to the classification rule can be observed in

Figure 3 (d).

It can be concluded that all cases identified as influential observations present remarkably high levels in some liver enzymes with respect to their groups. However, no region exists that could be associated with influential observations. For example, cases 135, 136 and 140 are ACH patients who present similar enzyme levels; however, case 135 is an influential correctly classified observation, case 136 is an influential misclassified observation and case 140 is a non-influential correctly classified observation.

If cases 58, 133, 135, 136 and 150 were omitted, the overall classification rate would come to 84.0%, that is, it would be increased only by 3.7%. The ACH group would exhibit the largest increase in the classification rate, from 57.5% to 64.9%.

According to these results, a specific study of those cases identified as influential (laboratory profile, diagnosis and other clinical aspects) should be carried out before taking the decision of whether to include them in or omit them from the statistical analysis.

5 Conclusions

DC is applied in biology, medicine, economics, psychology, sociology and other sciences to classify a new case into one of several different groups, in statistical pattern recognition to select the characteristics or variables that enable discrimination between populations, etc.

The classification rule based on the discriminant coordinates might be strongly affected by the presence of several extreme observations in the sample data and the results might be substantially altered by some perturbation of the observations. Therefore the researcher should be able to identify influential cases and assess their effects on the main statistics of the analysis.

This paper presents influence measures on three relevant statistics on discriminant coordinates: the transformation matrix, the canonical directions and the centroid configurations. It is possible that the user of methods of discriminant analysis is primarily interested in the influence of cases on the classification probabilities. However it is

advisable to analyze all aspects of the influence on the discriminant analysis. That is, it is advisable to complement the information generated by the influence measures based on the probability of misclassification.

Hence, we can conclude that the diagnostic tools that we propose are useful in many statistical analyses. Although the diagnostics have been developed for the first two canonical directions, they can easily be generalized to any dimension greater than 2.

From the illustrative examples presented in the above section we conclude:

- The sets of outliers, misclassified cases and influential observations are different. Outliers and/or misclassified cases may or may not be influential observations. Therefore, influence diagnostics provide additional information to that provided by outlier detection methods and the determination of misclassified cases.

- Each of the diagnostics proposed in Section 3 is designed for a specific purpose. Each diagnostic has its own interpretation and provides different and complementary information.

We also highlight that the masking and swamping effects should be taken into account when using these diagnostics. We think a similar analysis would be necessary to that carried out by Lawrance (1995) on the regression model. **In order to overcome masking and swamping, it is also possible to add to the three suggested measures the diagnostics tools which come from the use of robust estimators in discriminant analysis (see Atkinson *et al.* (2004) and Hubert *et al.* (2008))**. However, due to the huge amount of work implied, it should be the aim of a future paper.

Finally, a large value for $M-$, $D-$ or $A-$ measures indicates that the corresponding observation can be considered as influential. However, the following question can be posed: how large is large? Cutpoints have been associated with several diagnostics in the literature in such a way that observations can thereby be identified as highly influential. We propose reference values for $A-$ and $R-$ measures. Nevertheless, as Hadi (1992) asserted, influence diagnostics are designed to detect observations whose influence results are greater than other observations in a data set. They are not de-

signed to be a formal test of hypothesis. Thus, the values of a given influence measure should be compared **with** each other. The reference values proposed can be modified according to the researcher's opinion.

Acknowledgements. This paper was partly supported by the Project MTM2008-00018 (National Plan 2008, Ministry of Science and Technology, Spain).

References

Albert, A. and Harris, E.K., 1987. *Multivariate interpretation of clinical laboratory data*, New York: Marcel Dekker.

Atkinson, A., Riani, M., and Cerioli, A., 2004. *Exploring Multivariate data with the Forward Search*, New York: Springer Verlag.

Barnett, V. and Lewis, T., 1978. *Outliers in Statistical Data*, New York: John Wiley and Sons, 1st ed.

Belsley, D.A., Kuh, E., and Welsch, R.E., 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*, New York: John Wiley and Sons.

Bremner, A.P. and Taplin, R.S., 2002. Modified classification and regression tree splitting criteria for data with interactions, *Australian and New Zealand Journal of Statistics*, 44 (2), 169–176.

Bull, S.B., Greenwood, C.M.T., and Donner, A., 1994. Efficiency of reduced logistic regression models, *The Canadian Journal of Statistics*, 22 (3), 319–334.

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

Critchley, F. and Vitiello, C., 1991. The influence of observations on misclassification probability estimates in linear discriminant analysis, *Biometrika*, 78 (3), 677–690.

Fung, W.K., 1992. Some influence measures in discriminant analysis, *Statistics & Probability Letters*, 13, 279–285.

Fung, W.K., 1995a. Detecting influential observations for estimated probabilities in multiple discriminant analysis, *Computational Statistics and Data Analysis*, 20 (5), 557–568.

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

Fung, W.K., 1995c. Influence on classification and probability of misclassification, *Sankhya, The Indian Journal of Statistics, Series B*, 57, 337–384.

Fung, W.K., 1996. The influence of observations on misclassification probability in multiple discriminant analysis, *Communications in Statistics. Theory and Methods*, 25, 1917–1930.

Galanaud, D., Nicoli, F., Chinot, O., Confort-Gouny, S., Figarella-Branger, D., Roche, P., Fuentes, S., Le Fur, Y., Ranjeva, J., and Cozzone, P.J., 2006. Noninvasive diagnostic assessment of brain tumors using combined in vivo MR imaging and spectroscopy, *Magnetic Resonance in Medicine*, 55 (6), 1236–1245.

Guo, Y., Hastie, T., and Tibshirani, R., 2007. Regularized discriminant analysis and its application in microarrays, *Biostatistics*, 8 (1), 86–100.

Hadi, A.S., 1992. A new measure of overall potential influence in linear regression, *Computational in Statististics and Data Analysis*, 14, 1–27.

Hubert, M., Rousseeuw, P., and Van Aelst, S., 2008. High-breakdown robust multivariate methods, *Statistical Science*, 23 (1), 92–119.

Lawrance, A.J., 1995. Deletion influence and masking in regression, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lesaffre, E. and Albert, A., 1988. An uncertainty measure in logistic discrimination, *Statistics in Medicine*, 7 (4), 525–533.

Lesaffre, E. and Albert, A., 1989. Multiple-group logistic regression diagnostics, *Applied Statistics*, 38, 425–440.

Lubischew, A.A., 1962. On the use of discriminant functions in taxonomy, *Biometrics*, 18, 455–476.

McKay, R.J., 1977. Simultaneous procedures for variable selection in multiple discriminant analysis., *Biometrika.*, 64, 283–290.

Mirkin, S., Nikas, G., Hsiu, J.G., Diaz, J., and Oehninger, S., 2004. Gene expression profiles and structural-functional features of the peri-implantation endometrium in natural and gonadotropin-stimulated cycles, *The Journal of Clinical Endocrinology and Metabolism*, 89 (11), 5742–5752.

Moreno-Roldán, D., Munoz-Pichardo, J.M., and Enguix-González, A., 2007. Influence diagnostics in multiple discriminant analysis, *Test*, 16 (1), 172–187.

Munoz-García, J., Moreno-Rebollo, J.L., and Pascual-Acosta, A., 1990. Outliers: a formal approach, *International Statistics Review*, 58 (3), 215–226.

Perelman, S., Mazzella, M.A., Muschietti, J., Zhu, T., and Casal, J.J., 2003. Finding unexpected patterns in microarray data, *Plant Physiology*, 133, 1717–1725.

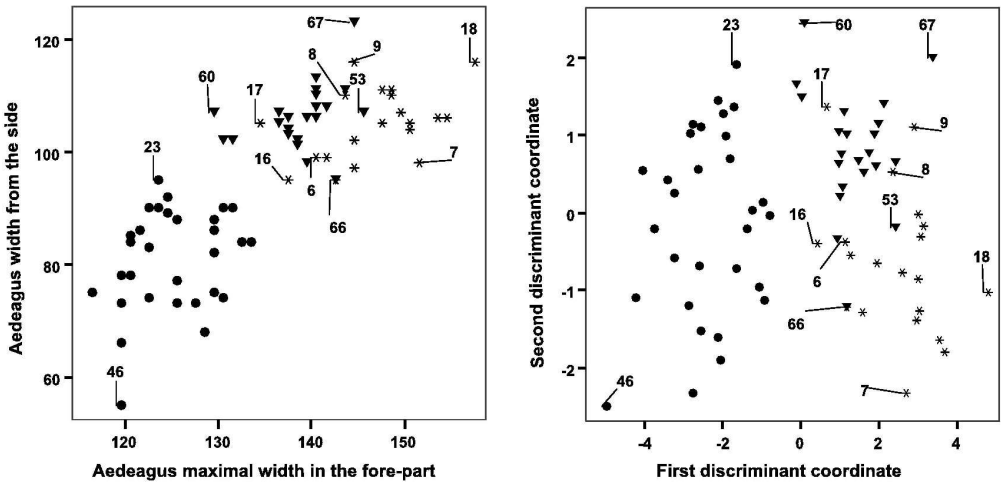
Plomteux, G., 1980. Multivariate analysis of an enzyme profile for the differential diagnosis of viral hepatitis, *Clinical Chemistry*, 26, 1987–1899.

Riani, M. and Atkinson, A., 2001. A unified approach to outliers, influence and transformations in discriminant analysis, *Journal of Computational and Graphical Statistics*, 10 (3), 513–544.

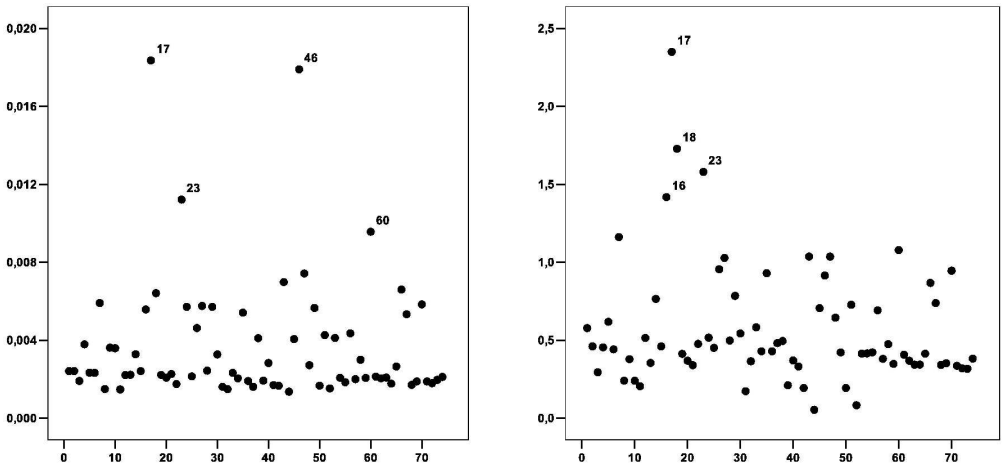
Schott, J.R., 1990. Canonical mean projections and confidence regions in canonical variate analysis., *Biometrika*, 77, 587–596.

Seber, G.A.F., 1984. *Multivariate Observations*, New York: John Wiley & Sons.

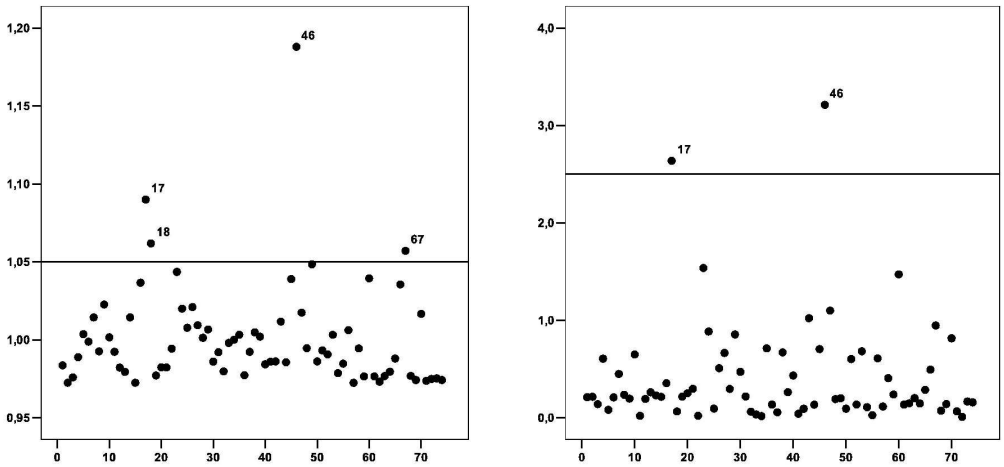
Struijk, J., Kanters, J.K., Andersen, M., Hardahl, T.B., Graff, C., Christiansen, M., and Toft, E., 2006. Classification of the long-qt syndrome based on discriminant analysis of t-wave morphology, *Medical and Biological Engineering and Computing*, 44 (7), 543–549.



Lubischew data set: Scatter plots. (*)-G1; (●)-G2; (▽)-G3
143x77mm (600 x 600 DPI)

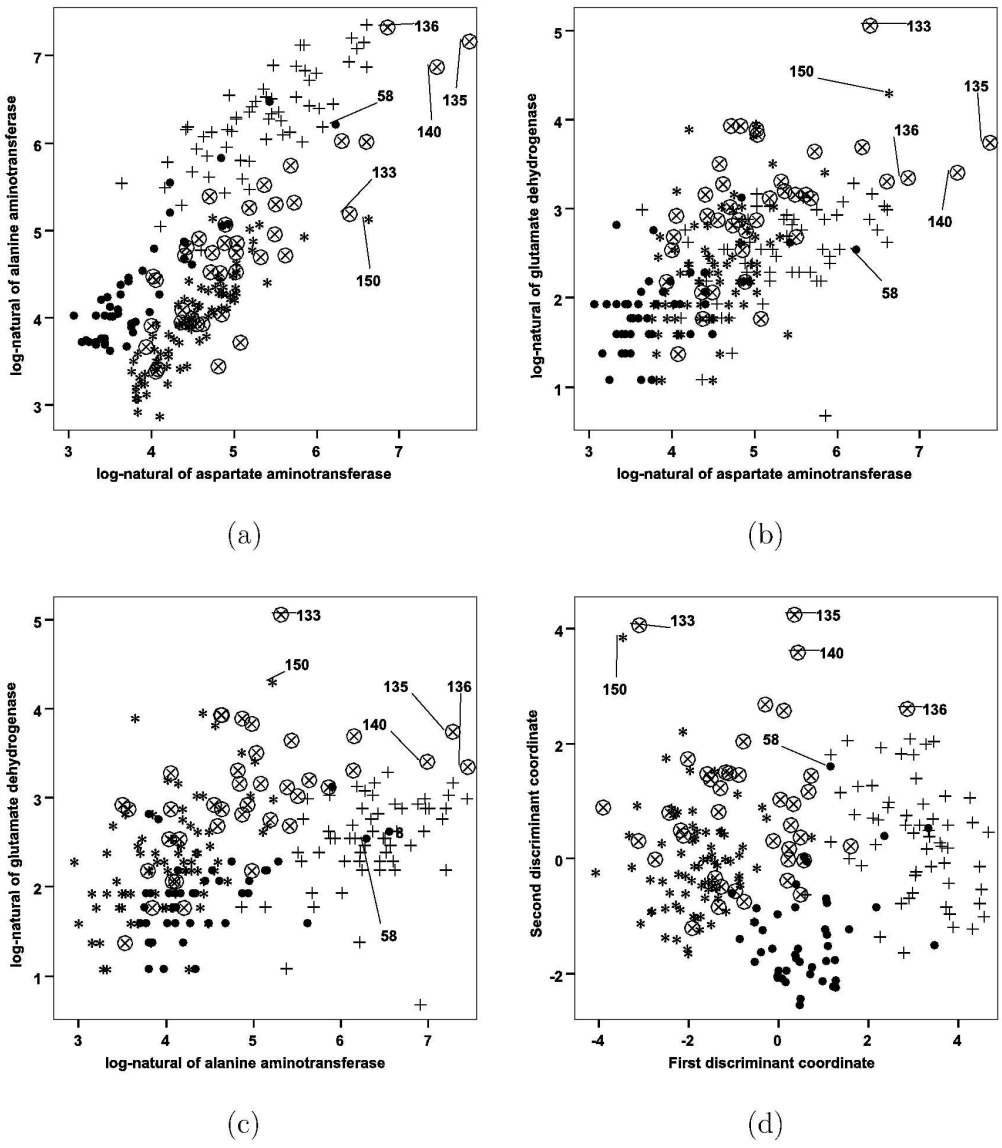


(a) Vertical axis: M-measure (b) Vertical axis: D-measure

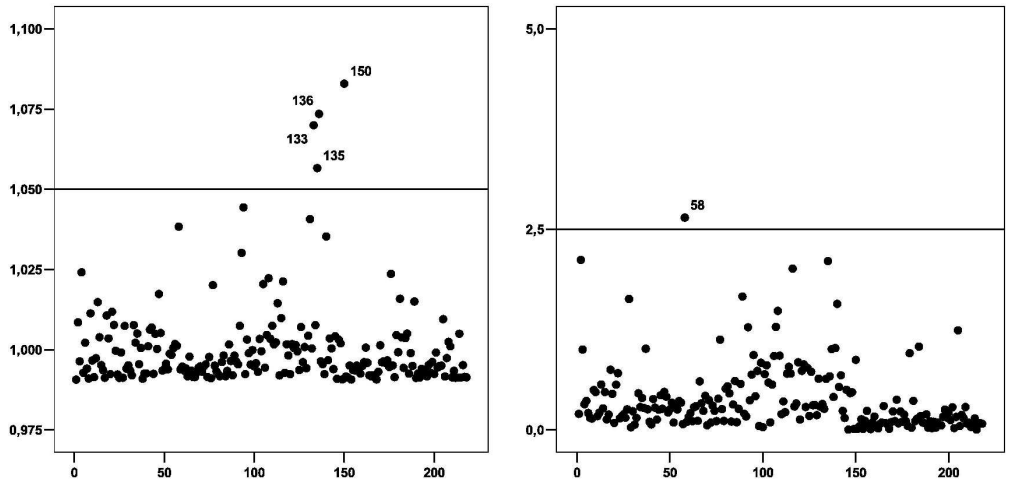


(c) Vertical axis: R-measure (d) Vertical axis: $A^{(1)}$-measure

Lubischew data set. Index plot of influence measures (in all graphics, horizontal axis: cases)
150x163mm (600 x 600 DPI)



Plomteux data set. Scatter plots (+)-AVH; (●)-PCH; (⊙)-ACH; (*)-PNC
144x164mm (600 x 600 DPI)



(a) Vertical axis: R-measure (b) Vertical axis: $A^{(2)}$-measure

Plomteux data set. Index plot of influence (in both graphics, horizontal axis: cases)
145x76mm (600 x 600 DPI)