GLOBAL JOURNAL OF PURE AND APPLIED SCIENCES, VOL. 12 NO 1, JANUARY, 2006: 107-110
 ISSN 1118-0579

METHODOLOGICAL CONTRIBUTION TO CONTROL HETEROSCEDASTICITY IN DISCRIMINANT ANALYSIS STUDIES

R. G. KAKAI and R. PALM

(Received 29 June, 2004; Revision Accepted 30 September, 2004)

ABSTRACT

We describe for two groups, the process of establishing an heteroscedastic model in Monte Carlo discriminant a..alysis studies. The simple model proposed allows, by the linear transformation, to extend the results of discriminant analysis studies to a large variety of situations. The heteroscedasticity degree of the model is appreciated by a parameter defined in the study, which can be computed not only on populations but also on data samples. This model can then be used to express the results of Monte Carlo discriminant analysis studies as a function of the heteroscedasticity degree observed on data samples.

KEY WORDS: heteroscedasticity ; discriminant analysis ; two groups ; Monte Carlo studies.

1. INTRODUCTION

Discriminant analysis is a statistical method whose objective is to define an allocation rule to classify an unknown observation in one of the g groups known as *a priori*. The rule is established on p characteristics or variables observed on the g populations or samples related to the different populations. In many cases, this allocation rule can misclassify observations, so that an error rate is associated to each classification rule established. The error rate can be estimated in practice by several methods proposed in literature (McLachlan, 1974, 1992 ; Efron, 1983 etc.). One of the relevant topics in discriminant studies is the comparison of classification rules or error rates estimators for homoscedastic or heteroscedastic models.

The problem of how to simulate heteroscedastic model in two-group discriminant analysis has been addressed by many authors (Gilbert, 1969, Van Ness, 1979, Marks and Dunn, 1974, Snapinn and Knoke, 1989 etc.). However, the heteroscedasticity parameter proposed in these studies cannot be computed on data samples. So, these studies have limited use in practice because the effect of heteroscedasticity is related to the parameters of the populations, which are usually unknown to the user of discriminant analysis.

We propose here, for two-group heteroscedastic model, a simple parameter, which can be determined for the populations as well as for data samples.

2. Linear transformation

Let's define two *p*-variables populations, with mean vectors $\boldsymbol{\mu}_i (i=1,2)$ and covariance matrices $\boldsymbol{\Sigma}_i (i=1,2)$. Suppose $\mathbf{A}$, any *p*-symmetric matrix and let's $\mathbf{m}$ a *p*-vector and $\mathbf{V}$, a diagonal matrix so that:

$$\boldsymbol{\Sigma}_1 = \mathbf{A}\mathbf{A}' \quad , \quad \boldsymbol{\mu}_2 = \mathbf{A}\,\mathbf{m} + \boldsymbol{\mu}_1 \quad \text{and} \quad \boldsymbol{\Sigma}_2 = \mathbf{A}\mathbf{V}\mathbf{A}' \,. \qquad (2.1)$$

For any vector x belonging to population 1 or 2, let's consider the linear transformation:

$$y = \mathbf{A}^{-1}(x - \boldsymbol{\mu}_1) \,. \qquad (2.2)$$

The distribution of random vector y in population 1 has mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$. In population 2, the mean vector and covariance matrix are respectively $\mathbf{m}$ and $\mathbf{V}$.

Based on the invariability of the classification rule when applying the transformation (2.2) on observations, the conclusions of discriminant analysis results related to observations vector x are also valid for vectors y.

R. G. KAKAI, Faculte des Sciences Agronomiques, Universite d' Abomey Calavi (UAC), 01 BP 526 Cotonou, Benin
R. PALM, Faculte Universitaire des Sciences Agronomiques de Gembloux (FUSAGx), B-5030, Gembloux, Belgique

3. Heteroscedastic model

Let's consider the observations y so that, for population 2, we have:
$\mathbf{m} = (m,0,...,0)'$ and $\mathbf{V}$, a diagonal matrix with vector $\mathbf{v}$ of diagonal elements so that:

$$\mathbf{v}_{ii} = \lambda\ (>0)\ \text{for}\ i=1,\dots k\ \text{ and }\ \mathbf{v}_{ii} = 1\ \text{ for }\ i = k+1, \dots, p\quad (k \le p). \qquad (3.1)$$

In (3.1), λ is considered as an heteroscedasticity parameter of the model and $\lambda = 1$ corresponds to nomoscedasticity.

By considering the simple model study, constituted of vector y such as defined, it is possible to cover a large variety of real world problems with the inverse of the linear transformation (2.2).

The means vectors and covariance matrices of the distribution of random vector y when applying this linear transformation are linked by the relations (2.1). Some remarks come out of theses relations:

- since $\mathbf{A}$ can be any matrix, $\mathbf{\Sigma}_1$ can take all possible forms of covariance matrix;
- since λ and k, in $\mathbf{V}$ can take different values, $\mathbf{\Sigma}_2$ can also take different forms.

4. Appreciation of the heteroscedasticity degree of the model

To appreciate the heteroscedasticity of the model, the parameter λ is not useful in practice because it cannot be measured on data samples. So it is necessary to find another parameter, depending on λ, which will allow to appreciate the heteroscedasticity degree of the model.
We then define a parameter Γ for two covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, as:

$$\Gamma = -\sum_{i=1}^{2} \ln\left(|\mathbf{\Sigma}_i| / |\mathbf{\Sigma}|\right),$$

where $\mathbf{\Sigma}$ is the pooled covariance matrix of the model. For data samples, this parameter can be estimated by replacing the theoretical covariance matrices by their estimated values.
In the case of random vectors y or their linear transformations x, the parameters Γ, k and λ are linked by the relation:

$$\Gamma(\lambda,k) = k \ln\left[\frac{(\lambda+1)^2}{4\lambda}\right], \qquad (4.1)$$

where k and λ are the parameters defined in (3.1). In the case of homoscedasticity, $\Gamma(1,k) = 0$ for any value of k

Proof : for random vectors y, the covariance matrices in populations 1 and 2 are respectively $\mathbf{I}$ and $\mathbf{V}$ (defined in (3.1)). The pooled covariance matrix $\mathbf{\Sigma}$ of the model is then a p-diagonal matrix with diagonal vector $\boldsymbol{\tau}$ defined as :

$$\tau_{ii} = \frac{\lambda+1}{2}\ (>0)\ \text{for}\ i=1,\dots k\ \text{ and }\ \tau_{ii} = 1\ \text{ for }\ i = k+1, \dots, p\quad (k \le p).$$

We have then, $|\mathbf{I}| = 1$, $|\mathbf{V}| = \prod_{i=1}^{p} v_{ii} = \lambda^k$, $|\mathbf{\Sigma}| = \prod_{i=1}^{p} \tau_{ii} = \left(\frac{\lambda+1}{2}\right)^k$. (4.2)

Based on (4.2), $\Gamma(\lambda,k) = -\ln\frac{|\mathbf{I}|}{|\boldsymbol{\Sigma}|} - \ln\frac{|\mathbf{V}|}{|\boldsymbol{\Sigma}|} = k\ln\left[\frac{(\lambda+1)^2}{4\lambda}\right]$

Figure 1 gives different curves of Γ versus λ for some values of k. It can be noticed from this figure that, for each curve, the more the value of λ is far from 1, the more $\Gamma(\lambda)$ is high. It can be then noticed that:

$$\Gamma\left(\frac{1}{\lambda},k\right) = k\ln\left[\frac{\left(\frac{1}{\lambda}+1\right)^2}{\frac{4}{\lambda}}\right] = \Gamma(\lambda,k) \qquad (4.3)$$

Moreover, for a given value of k, the value $\Gamma(\lambda)$ is insensible to the different positions that can be taken by the k parameters λ on the diagonal of the matrix $\mathbf{V}$. For example, if $k=2$, the diagonal vectors $\mathbf{v}_1=(\lambda,\lambda,1\ldots,1)'$ and $\mathbf{v}_2=(\lambda,1\ldots\lambda,\ldots,1)'$ lead to the same value of $\Gamma(\lambda)$.

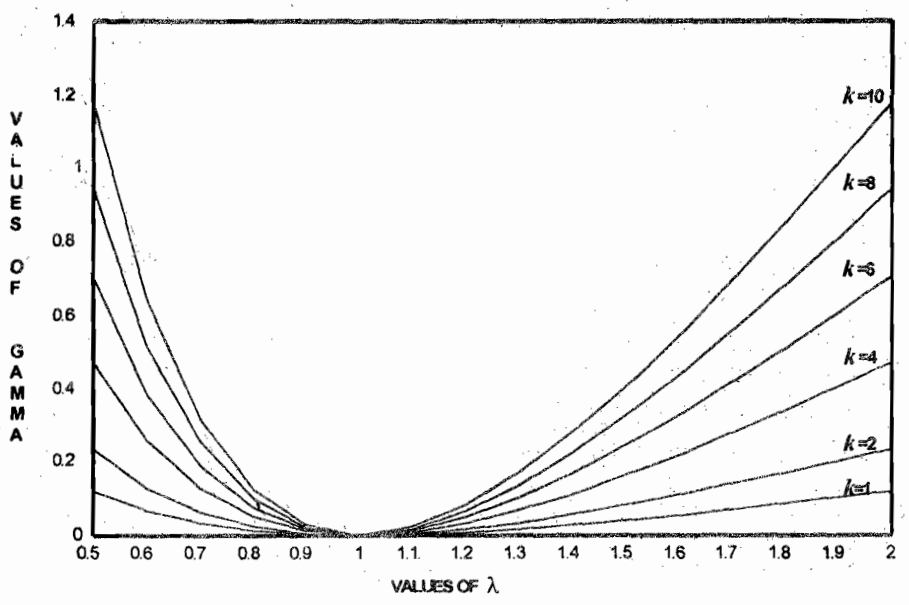

Figure 1. Curves of Γ versus λ for some values of k.

5. CONCLUSIONS AND LIMITATIONS

In discriminant analysis studies, the method used to obtain populations with control of mean vectors and covariance matrices depends on the homoscedasticity or the heteroscedasticity of the model. In the case of the homoscedasticity, the inverse of linear transformation (2.2) applied to observations vector y allows to cover all situations in practice. But, in the case of heteroscedasticity, we do not know a linear transformation that can transform any heteroscedastic model to the simple model constituted of $\mathbf{Y}_1(\mathbf{0},\mathbf{I})$ and $\mathbf{Y}_2(\mathbf{m},\mathbf{V})$ proposed above. The linear transformation applied to this simple model study leads to two populations whose mean vectors and covariance matrices are linked by the relation (2.1). Nevertheless, this model allows to extend the results of discriminant analysis studies to a large variety of real world problems, which are in our opinion sufficient for Monte Carlo experiments.

In Monte Carlo experiments, the choice of m and Γ values depends on the aims to be reached by discriminant analysis studies. The choice of Γ values can be done on the basis of some values of the power function of the homoscedasticity test related to Γ. This power function can be established by simulation. In the case of m, this choice can be done on the basis of MAHALANOBIS distance between the two populations when the model considered is homoscedastic. With heteroscedastic model, this choice can be done on the basis of the distribution of the first variable, the desired overlap of the two populations for this variable and the value of Γ.

Since it is possible to compute Γ on data samples, the Monte Carlo experiment related to discriminant analysis can take into account the effect of heteroscedasticity on a model and in empirical comparison of classification rules for example, the effect of heteroscedasticity on the performance of rules can be related to the estimated value of Γ determined on the data samples.

ACKNOWLEDGEMENTS

This work was supported in part by the International Foundation for Science (IFS) through research grant attributed to R. Glèlè Kakaï.

REFERENCES

Efron, B., 1983. Estimating the error of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.* 78 : 316-331.

Gilbert, E. S., 1969. The effects of unequal variance matrices on Fisher's linear discriminant function. *Biometrics*. 25 : 505-515.

Marks, S. and Dunn O. J., 1974. Discriminant functions when covariance matrices are unequal. *J. Amer. Statist. Assoc.* 69 : 555-559.

McLachlan, G.J., 1974. Estimation of the errors of misclassification on the criterion of asymptotic mean square error. *Technometrics* 16 : 255-260.

McLachlan, G.J., 1992. *Discriminant analysis and statistical pattern recognition*, New York, Wiley, 526p.

Snapinn, S. M. and Knoke, J. D., 1989. Estimation of error rates in discriminant analysis with selection of variables. *Biometrics* 45 : 289-299.

Van Ness, J.W., 1979. On the effects of dimension in discriminant analysis for unequal covariance populations. *Technometrics* 21 : 119-127.