International Journal of Statistics and Applications 2015, 5(5): 208-212
DOI: 10.5923/j.statistics.20150505.04

Effect of Sampling Methods on Misclassification of Fisher's Linear Discriminant Analysis

Ghasem Rekabdar*, Bahare Soleymani

Department of Mathematics, Abadan Branch, Islamic Azad University, Abadan, Iran

Abstract In this study, the effect of stratified sampling design has been studied on the accuracy of Fisher's linear discriminant function or Anderson's $\hat{W}$. For this purpose, we put on weighted estimators in function $\hat{W}$ instead of simple random sampling estimators. The results of a simulation study indicated that the performance of $\hat{W}$ affected by alteration of sampling methods. The performance of proposed discriminant function $\hat{W}_{st}$ in comparison to the classical discriminant function is more appropriate. Specially, in case of the mean of strata have significant difference compared with the overall mean of each group.

Keywords Fisher's linear discriminant function, Multivariate normal distribution, Stratified sample design

1. Introduction

The discrimination between two groups using multivariate data has been recognized as an important problem that was firstly studied by Fisher (1936). The linear discriminant function (LDF) is a standard approach to yield optimal results when the two groups have a conditional multivariate normal distribution with distinct mean vectors and common covariance matrix (Mardia & et al, 1979). Computing the misclassification probabilities or error rates of the discriminant function are interesting issues. When competing groups have known parameters, the LDF distribution can be obtained exactly by univariate normal distribution (Johnson & Wichern, 1992). In practice, the parameters of the LDF are unknown. Then we estimate these parameters by means of independent random "training samples". The sample distribution of LDF has been studied by several authors. Anderson (1973) obtained the asymptotic expansion of the distribution of the sample Fisher's linear discriminant function $\hat{W}$ in terms of order $O(n^{-2})$. Atakan (2009) compared the performance of seven well known methods in literature to estimating probability of misclassification by bootstrap percentile confidence intervals. This research can provide a good literature review for more study.

In several researches, the sampling design effects on statistical methods have been studied. Especially, in regression analysis effect of sampling designs on least square estimator studied by some authors (DuMuchel & Duncan, 1981; Horton & Fitzmaurice, 2004). Also, in analysis of variance about mean difference of groups, effect of cluster sampling design on F ratio studied in social and psychological survey, frequently (Hegges & Rhoads, 2011). In multivariate statistical analysis, complex sampling design lead to complicated methods. However, little study has been dedicated to the effect sampling methods on LDF because analytical complexity. Nonetheless, some researchers examining the effect of sampling design on the misclassification probability of the LDF (Kao & McCabe, 1991; Leu & Tsui, 1997). In light of stratified random sampling, Tsui & Leu (1998) indicated that asymptotic expansion of LDF has an error of order $O(1)$. Therefore, using of LDF without correction can increases the probability of misclassification. Recently, Shahrokh Esfahani & Dougherty (2014) by simulation study showed *that separate sampling with an inappropriate sampling ratio can significantly reduce classification accuracy* of LDF.

The main contribution of the present paper is to approximate LDF probability of misclassification using weighted estimators. In some researches, we have auxiliary information about the groups and it is beneficial to use it to construct LDF. For example, we can be able to categorize each group on the basis of a qualitative variable. In this case, stratified sampling design can be used to draw data from each group. In this study, we substitute unbiased weighted estimators in LDF when the sample design is stratified. Also, a comparison between two linear discriminant functions is made by a simulation study.

* Corresponding author:
ghasem_rekabdar@yahoo.com (Ghasem Rekabdar)
Published online at http://journal.sapub.org/statistics

2. Preliminaries for the LDF

In this section, we introduce some preliminaries of the LDF. Suppose Π_1 and Π_2 denote two distinct groups whose known multivariate probability density functions of *p*-dimensional random vector $x = (x_1, x_2, \ldots, x_p)$ are denoted by $f_1(x)$ and $f_2(x)$, respectively. We use $P(i \mid j)$ to denote the probability of misclassification an observation x into group Π_i when, in fact, it belongs to the group Π_j. Let p_1 and p_2 be the prior probabilities of the groups, then the total probability of misclassification (*TPM*) is defined as

$$TPM = P(2 \mid 1)p_1 + P(1 \mid 2)p_2$$

According to the Bayes optimal classification rule, *TPM* is minimized when a new observation x is classified into group Π_1 by

$$\log \frac{f_1(x)}{f_2(x)} \geq k, \tag{1}$$

Where $k = 2\log(p_2/p_1)$. If the prior probabilities in each group are taken equal, then cut-off value is $k = 0$. Also, if the multivariate normal densities with common covariance matrices are used in previous equation, then the LDF is given by

$$W = (\mu_1 - \mu_2)'\Sigma^{-1}(x - \frac{1}{2}(\mu_1 + \mu_2)). \tag{2}$$

Using the Equation (2), a new observation x is assigned into the group Π_1 when $W \geq 0$. In the case of, $W < 0$, this observation is assigned into the group Π_2. Suppose that the prior probabilities are taken to be equal i.e. $p_1 = p_2$, then the *TPM* is defined as

$$\begin{aligned} TPM &= \frac{1}{2}(P(W < 0) + P(W \geq 0)) \\ &= \Phi(-\frac{\Delta}{2}), \end{aligned} \tag{3}$$

where Φ is the cumulative distribution function of standard normal random variable and Δ is Mahalanobis distance between the groups, i.e.,

$$\Delta^2 = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2). \tag{4}$$

3. Sample LDF

In this section, we illustrate the sample representation of the Fisher's linear discriminant function (2) under random sampling and stratified designs.

3.1. Random Sampling

Suppose we have n_1 observation $x_{11}, \ldots, x_{1n_1}$ drawn from Π_1 and n_2 observation $x_{21}, \ldots, x_{2n_2}$ drawn from Π_2, where $n_1, n_2 > p$. We estimate the parameters (2) by the unbiased sample means

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i},$$

$$\bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i},$$

and

$$S = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)S_1 + (n_2 - 1)S_2),$$

where

$$S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)',$$

$$S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)(X_{2i} - \bar{X}_2)',$$

respectively. Then, the discriminant functions (2) can be modified as $\hat{W}$ yields a *plug-in* discriminant function is given by

$$\hat{W}_1 = (\bar{x}_1 - \bar{x}_2)'S^{-1}(x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)). \tag{5}$$

In this case a natural estimate of (4) is

$$D_1^2 = (\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2), \tag{6}$$

and the estimated of the total misclassification probability is given by

$$TPM_{\hat{W}} = \Phi(-\frac{D_1}{2}). \tag{7}$$

3.2. Stratified Sampling

Suppose the groups Π_g where $g = 1, 2$ split into α_g parts Π_{gj} where $j = 1, \ldots, \alpha_g$. If the group size is denoted N_g then $N_g = \sum_{j=1}^{\alpha_g} N_{gj}$, where N_{gj} is denoted size of Π_{gj}. Also, we select a random sample n_{gj} of fixed size n_g from each group, where $n_g = \sum_{j=1}^{\alpha_g} n_{gj}$. We furthermore assume throughout that the designs are simple without replacement within each stratum. In light of this design, the unbiased estimation of means in each group is given by

$$\bar{X}_g^{st} = \sum_{j=1}^{\alpha_g} \frac{N_{gj}}{N_g} \bar{X}_{gj} = \sum_{j=1}^{\alpha_g} \pi_{gj} \bar{X}_{gj},$$

where

$$\bar{X}_{gj} = \frac{1}{n_{gj}} \sum_{i=1}^{n_{gj}} X_{gi},$$

is mean estimation of the jth stratum of group Π_g and weight of stratums are π_{gj}. Also, if we suppose the covariance matrix into each stratum is common then unbiased estimation of the covariance matrix Σ is defined by

$$S_{gj} = \frac{1}{n_g - 1} \sum_{i=1}^{n_{gj}} (X_{ig} - \bar{X}_{gj})(X_{ig} - \bar{X}_{gj})'.$$

If weighted estimation

$$\tilde{S}_g = \sum_{j=1}^{\alpha_g} \frac{N_{gj}}{N_g} S_{gj} = \sum_{j=1}^{\alpha_g} \pi_{gj} S_{gj},$$

is assumed in each group then the pooled covariance matrix is given by

$$\tilde{S} = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)\tilde{S}_1 + (n_2 - 1)\tilde{S}_2).$$

By substituting these unbiased estimators into (2), we obtain a new sample LDF

$$\hat{W}_{st} = (\bar{x}_1^{st} - \bar{x}_2^{st})'\tilde{S}^{-1}(x - \frac{1}{2}(\bar{x}_1^{st} + \bar{x}_2^{st})). \qquad (8)$$

Similar to (6) we define

$$D_2^2 = (\bar{x}_1^{st} - \bar{x}_2^{st})'\tilde{S}^{-1}(\bar{x}_1^{st} - \bar{x}_2^{st}), \qquad (9)$$

therefore, the total probability of misclassification is estimated by

$$TPM_{\hat{W}_{st}} = \Phi(-\frac{D_2}{2}). \qquad (10)$$

Clearly, in the case of Mahalanobis distance (9) is greater than (6), then the Equation (10) is less than (7). Thus, the stratified sampling designs can provide greater efficient estimates than corresponding random sampling in discriminant analysis.

4. Simulation Study

In this section, we examine the performance of sample discriminant function $\hat{W}_{st}$ in comparison $\hat{W}$ by conducting numerical experiments. It is further noted that Mathematica software was used to write program codes for numerical calculation. The package is available from the authors upon request.

Suppose the group sizes are equal i.e., $N_1 = N_2$ and each group is categorized into two stratums. The first group size of stratums are considered $N_{11} = 2N_{12}$ and the second group $N_{21} = N_{22}$. Therefore, the weights of stratums are $\pi_{11} = 2/3, \pi_{12} = 1/3, \pi_{21} = \pi_{22} = 1/2,$ respectively. The covariance matrix structure considered in this examination in each group and stratums by

$$\Sigma = \begin{pmatrix} 2 & .5 & .5 & .5 \\ .5 & 1 & .5 & .5 \\ .5 & .5 & 3 & .5 \\ .5 & .5 & .5 & 4 \end{pmatrix}.$$

The stratum means of each group are defined by

$$\mu_{11} = (2,1,3,0)', \ \mu_{12} = (2,1,3,m)'$$

and

$$\mu_{21} = (1,0,\text{-}1,0), \ \mu_{22} = (1,0,-1,m),$$

The parameter m controlling distance between two stratums and we consider its values 0, 2 and 5, respectively. Therefore, the vector mean of each group is given by

$$\mu_1 = (2,1,3,m/3)',$$

and

$$\mu_2 = (1,0,-1,m/2)'.$$

The exact total probability misclassification of population discriminant function (2) in terms of (3) is demonstrated in Table 1. From the table, we can see that the TPM of W is scale down when m is increasing.

Table 1. Exact TPM of population LDF

TPM_W		
m		
0	2	5
0.1170	0.1133	0.1059

In each simulation, we generate random samples from four normal populations conditional distributions $N_4(\mu_{gj}, \Sigma)$, where $g, j = 1,2$. In each simulation the size of samples considered $n_1 = n_2 = 30, 70, 150,$ respectively. The samples divided in each group equally. Also, each simulation was run 100 times. Thus, the results presented in Table 2 are the average of estimated total probability misclassification. When the parameter m increased then TPM of $\hat{W}_{st}$ decreased for all sample sizes. While, by increasing m the TPM of discriminant function $\hat{W}$ has been increased except for sample size 30. Also, when the sample size increased then TPM of

discriminant functions $\hat{W}$ and $\hat{W}_{st}$ tend to TPM of W obtained in Table 1. For $m = 0$ the TPM of $\hat{W}$ is closer to exact TPM while for $m = 2,5$, we can see from Table 2 the TPM of $\hat{W}_{st}$ are closer to the TPM of W than discriminant function $\hat{W}$.

Table 2. Estimated TPM of sample LDF

	$TPM_{\hat{W}}$			$TPM_{\hat{W}_{st}}$		
	m			m		
n	0	2	5	0	2	5
30	0.1110	0.1136	0.1129	0.1089	0.1069	0.1004
70	0.1124	0.1139	0.1177	0.1123	0.1096	0.1039
150	0.1156	0.1193	0.1169	0.1152	0.1136	0.1044

In Figure 1, we display the histogram of discriminant functions by performing 200,000 iterations of the Equations (5) and (8). As can be seen in figure, the histograms of discriminat function $\hat{W}$ are almost symmetrical for all values m but they aren't seem normally distributed. Nonetheless, the histogram of discriminat function $\hat{W}_{st}$ is symmetrical for $m = 5$. In other words, when strata of the groups are significantly diversity in means then the limited distribution of $\hat{W}_{st}$ is symmetric and unimodal.

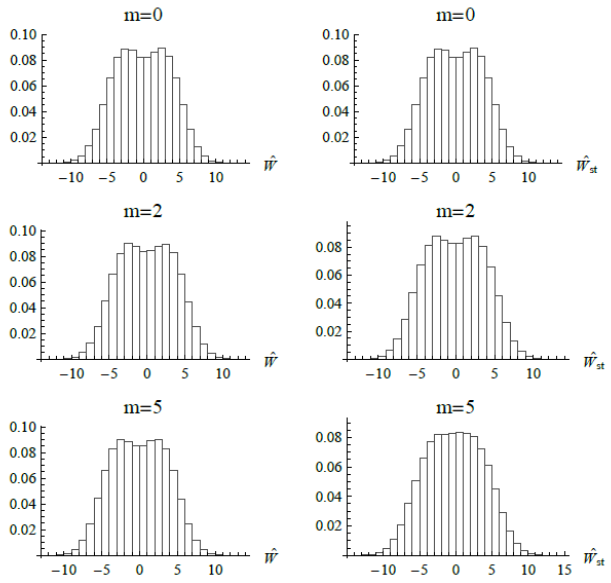

Figure 1. Histogram of the discriminant functions ($n_1 = n_2 = 200,000$)

5. Discussions

In many studies, particularly in the field of human sciences such as psychology, education, financial management and medical researches the sampling method is stratified. A common error in this type of research is the inadvertence of sampling designs and using analytical methods in statistical software in which the sampling method assumes that the simple random. In this study, in case of stratified sampling, we present a linear discriminant function by replacing the usual unbiased sample estimators with unbiased weighted estimators. In simulations, we demonstrate discriminant function $\hat{W}_{st}$ has better performance in comparison $\hat{W}$ when the groups consist of strata with distinct means. This discriminant function can be used to obtain error rate between groups that are categorized by an auxiliary variable such as gender, job, etc. An expansion of distribution $\hat{W}_{st}$ remains as open problem which it can study in future research.

ACKNOWLEDGEMENTS

This article is resulted from a research project which financed by Islamic Azad University Abadan branch.

REFERENCES

[1] Anderson, T. W. (1973). An asymptotic expansion of the distribution of the studentized classification statistics *W*. The Annals of statistics, 1, 964-972.

[2] Atakan, C. (2009). Bootstrap percentile confidence intervals for actual error rate in linear discriminant analysis. Hacettepe Journal Mathematics and Statistics, 38, 357- 372.

[3] DuMuchel, W. H. & Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, 535-543.

[4] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188

[5] Hedges, L. V. & Rhoads, C. H. (2011). Correcting an analysis of variances for clustering. British Journal of Mathematical and Statistical psychology, 64, 20-37.

[6] Horton, N. J. & Fitzmaurice, G. M. (2004). Regression analysis of multiple source and multiple informant data from complex survey samples. Statistics in Medicine, 23, 2911-2933.

[7] Johnson, R. A., Wichern, D. W. (1992). Applied Multivariate Statistical Analysis. New Jersey: Pearson Prentice Hall.

[8] Kao, T. C. & McCabe, G. P. (1991). Optimal Sample Allocation for Normal Discrimination and Logistic Regression under Stratified Sampling. Journal of the American Statistical Association, 86, 432-436.

[9] Leu, C. H. & Tsui, K. W. (1997). Discriminant analysis of survey data. Jornal of Statistical Planning and Inference, 60, 273-290.

[10] Mardia, K. V., Kent, J. T., & Bibby, J. (1979). Multivariate

Analysis. London: Academic Press.

[11] Shahrokh Esfahani, M., & Dougherty, E., R. (2014). Effect of separate sampling on classification accuracy. Bioinformatics, 30(2), 242-250.

[12] Tsui, K. W. & Leu, C. H. (1998). The Effect of Sampling Design on Anderson's Expansion of the Distribution of Fisher's Sample Discriminant Function. Statistica Sinica, 8, 1115-1130.