Metodološki zvezki, Vol. 2, No. 2, 2005, 231-242

Robustness of the Fisher's Discriminant Function to Skew-Curved Normal Distribution

Maja Sever, Jaro Lajovic, and Borut Rajer[1]

Abstract

Discriminant analysis is a widely used multivariate technique with Fisher's discriminant analysis (FDA) being its most venerable form. FDA assumes equality of population covariance matrices, but does not require multivariate normality. Nevertheless, the latter is desirable for optimal classification. To test FDA's performance under non-normality caused by skewness the method was assessed with simulation based on a skew-curved normal (SCN) distribution belonging to the family of skew-generalised normal distributions; additionally, effects of sample size and rotation were evaluated. Apparent error rate (APER) was used as the measure of classification performance. The analysis was performed using ANOVA with (transformed) mean APER as the dependent variable. Results show the FDA to be highly robust to skewness introduced into the model via the SCN distributed simulated data.

1 Introduction

Discriminant analysis is a widely used multivariate statistical technique with two closely related goals: discrimination and classification. In its original form, proposed by Fisher, the method assumes equality of population covariance matrices, but does not explicitly require multivariate normality. However, optimal classification performance of Fisher's discriminant function can only be expected when multivariate normality is present as well, since only good discrimination can ensure good allocation.

The reported experiment aimed at examining the robustness of Fisher's discriminant function to violation of normality caused by skewness as exemplified by the skew-curved normal (SCN) distribution, a member of a new family of asymmetric normal distributions. Concomitantly, the influences of sample size and rotation were investigated. Allocation performance was assessed by calculation of the apparent classification error rate (APER).

[1] University of Ljubljana, Slovenia

2 Methods

2.1 Fisher's discriminant function

Discrimination is a separation procedure that tries to find a discriminant function whose numerical values are such that the observations from several populations are separated as much as possible. An allocation procedure that uses a discrimination function as a well-defined rule in order to optimally assign a new observation to the labelled classes is called classification. It is evident that only good discrimination leads to good classification.

Consider two populations π_1 and π_2, each with p-variate distribution having mean vectors μ_1 and μ_2 and common covariance matrix Σ. Let the sizes of two independent random samples be denoted by n_1 and n_2 respectively, the mean vectors by $\overline{x}_1$ and $\overline{x}_2$, the pooled sample covariance matrix by S and any observed unit by x. Fisher suggested finding a linear combination of multivariate observations x to create univariate observations y such that the ys derived from populations π_1 and π_2 are separated as much as possible. Fisher's discriminant function (DF) can be written as:

$$y = a^T x$$

where a is the vector of coefficients and x the vector of variable values for a particular case.

Fisher's DF assumes equality of population covariance matrices, but does not assume multivariate normality.

An allocation rule based on Fisher's DF is as follows: allocate observation x_0 to class π_1, if

$$\hat{y}_0 \geq \hat{m} \text{ or } \hat{y}_0 - \hat{m} \geq 0$$

where $\hat{m}$ is the cut-off point.

Else allocate observation x_0 to class π_2, if

$$\hat{y}_0 < \hat{m} \text{ or } \hat{y}_0 - \hat{m} < 0 .$$

Hills (Krzanowski, 1977: 191) has pointed out that Fisher's DF is a useful tool for discrimination under wide distributional conditions, but it may be a quite unsuitable technique for allocating particular observation to one of two multivariate non-normal populations. Therefore, classification based on Fisher's DF can be optimal only when multivariate normality holds.

2.2 Skew-Curved normal distribution

The skew-curved normal (SCN) distribution is a special case of a new class of asymmetric normal distributions, so-called skew-generalised normal (SGN) distributions introduced by Arrellano-Valle et al. (2004).

A random variable X has a SGN density with parameters λ_1 and λ_2 ($\lambda_1 \in \Re$ and $\lambda_2 \geq 0$) if:

$$f(x) = 2\phi(x)\Phi\left(\frac{\lambda_1 x}{\sqrt{1+\lambda_2 x^2}}\right),\ x \in \Re \tag{2.1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the N (0,1) pdf and cdf, respectively. It is denoted by X ~ SGN (λ_1, λ_2). SCN is a special case or subclass of the former for $\lambda_1 \in \Re$ and $\lambda_2 = \lambda_1^2$. Thus, a random variable X has a SCN density if:

$$\mathrm{f}(\mathrm{x}) = 2\phi(\mathrm{x})\Phi\left(\frac{\lambda_1 \mathrm{x}}{\sqrt{1+(\lambda_1 \mathrm{x})^2}}\right),\ \mathrm{x} \in \Re \tag{2.2}$$

and it is usually denoted by X ~ SCN (λ_1).

It should be noted that the well-established skew-normal (SN) density described by Azzalini and Capitanio (1999) also represents a special case of SGN family with $\lambda_1 \in \Re$ and $\lambda_2 = 0$:

$$\mathrm{f}(\mathrm{x}) = 2\phi(\mathrm{x})\Phi(\lambda_1 \mathrm{x}),\ \mathrm{x} \in \Re \tag{2.3}$$

Both SCN and SN distributions depend on only one asymmetry parameter (λ_1) and both include the classic normal distribution as a special case (when $\lambda_1 = 0$). However, the SCN distribution provides a wider range of kurtosis and less skewness compared to the SN. Besides, when $\lambda_1 \to \infty$, the asymptotic distribution of the SCN distribution does not become half-normal as does the SN distribution.

The fact that the SCN distribution provides »smoother skewness« was the motivation for its use in the experiment (Figure 1).

2.3 Simulation

The simulation involved the generation of 2 groups of 3 random, intercorrelated variables (essentially generating 2 groups in 3-dimensional space). Throughout the experiment the sample size, symmetry (normality) and position of one group (designated π_2) were kept constant (sample size 50, no skewness and no rotation), while the parameters of the other group (designated π_1) varied:

- sample size (n): 25, 50, 100;
- skewness (λ_1): 0, 1, 5, 10;
- rotation around the centroid (α): 0–180° in steps of 20°.

The experiment consisted of 100 simulations per combination of parameter values.

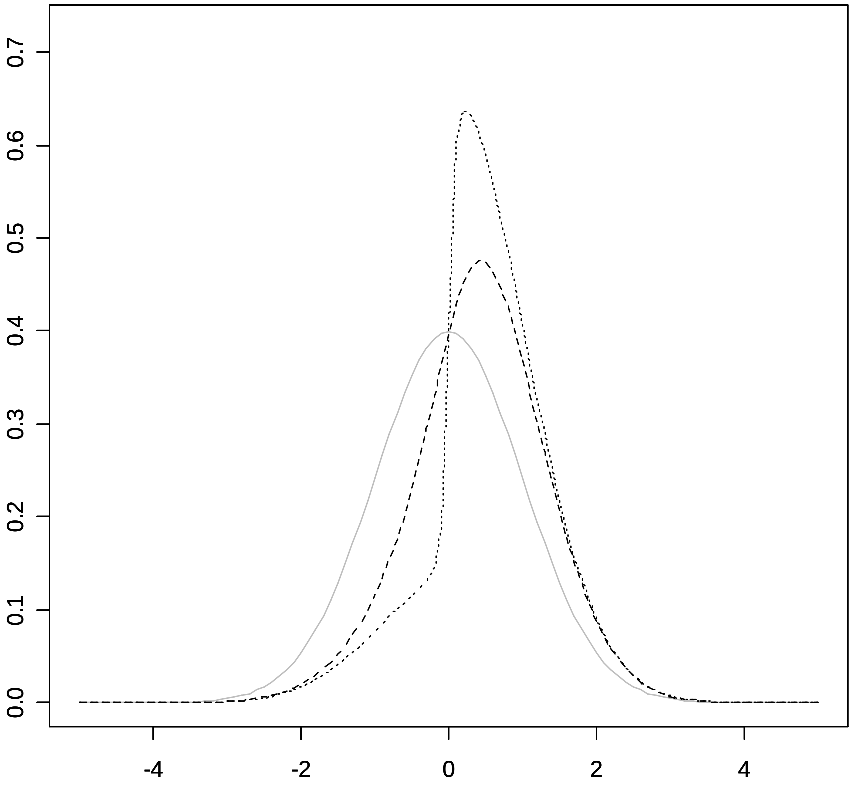

Figure 1: Examples of SCN density for $\lambda_1 = 1$ (dashed) and $\lambda_1 = 10$ (dotted), with $\lambda_1 = 0$ (i. e. standard normal) as a reference (solid grey).

As the use of completely fictitious data was not considered appropriate, group generation was based on parameters (means and covariance matrices) of a subset of the time-honoured Iris dataset. The Iris dataset consists of 50 measurements of 4 characteristics (sepal length, sepal width, petal length, petal width) in 3 Iris species (*I. setosa*, *I. versicolor*, *I. virginica*). Considering their spatial geometry, the *Iris setosa* (π_1) and *Iris virginica* (π_2) sepal length, petal length and petal width (labelled variable x, y and z, respectively) parameters (presented in Table 1) were used[2].

[2] Assignment of both groups (setosa as π_1 and virginica as π_2) was arbitrary.

Table 1: Means and covariances.

	Iris setosa (π_1)			**Iris virginica (π_2)**		
	x	y	z	x	y	z
μ	5.006	1.462	0.246	6.588	5.552	2.026
x	0.89	1.79	0.74	0.89	1.79	0.74
y	1.79	4.40	1.87	1.79	4.40	1.87
z	0.74	1.87	0.84	0.74	1.87	0.84

Figure 2 demonstrates the baseline geometry ($\lambda_1 = 0$, $\alpha = 0$) using a simulated instance; note that n = 100 for both groups to enhance the visualisation.

The simulation was programmed in R (v 2.1) and consisted of two main routines:

1. samples generation and
2. Fisher's discriminant analysis.

The key element of the generation routine was the SCN random number generator (RNG), built using the inversion method with utilisation of R's uniform RNG. Due to the fact that the SCN distribution function (see equation 2.5, above) and hence its inverse distribution function can not be obtained in closed form, a numerical technique with construction of cumulative values table (with appropriate parameters) in the range of –10 z to 10 z was applied, yielding a generator compliant with requirements for such a purpose (L'Ecuyer, 2004).

Although R offers several possibilities for discriminant analysis (e. g. functions lda, qda), we opted for implementation of the original Fisher's DF using our own function. The obtained confusion matrix was used for evaluation, i. e. for performance assessment.

2.4 Evaluation

The (predictive) performance of the Fisher's DF was assessed using the apparent classification error rate (APER). APER is an intuitively conceivable and easy to implement measure, albeit with some shortcomings, e.g. giving over-optimistic results if used for evaluation of the data used for construction of the classification rule (Pohar, 2004). To avoid the latter problem, two separate (independent) sample sets were generated at each parameter combination: one for the DF construction and the other for the DF evaluation.

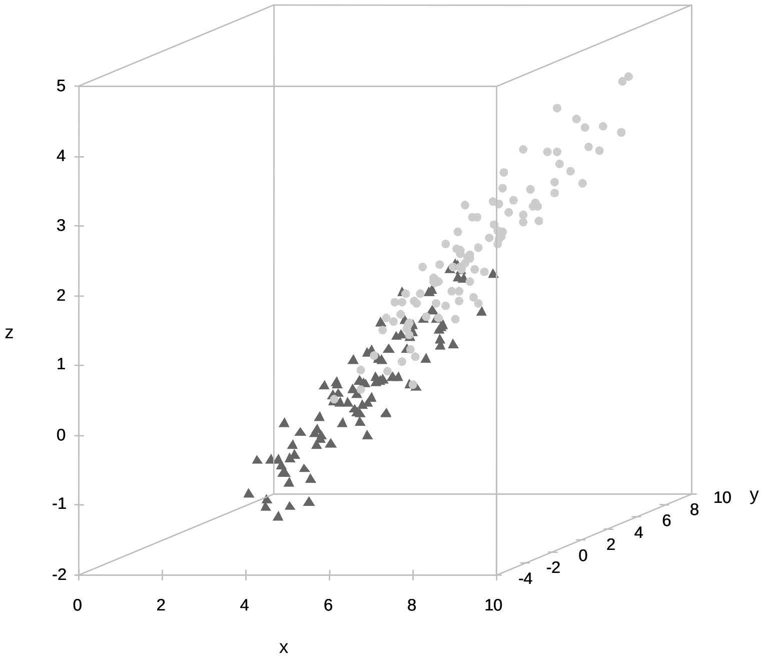

Figure 2: Baseline spatial geometry of the simulated groups: π_1 (dark grey triangles) and π_2 (light grey circles); n = 100, $\lambda_1 = 0$, $\alpha = 0$ for both groups.

The APER was calculated from the confusion matrix (contingency table of actual group membership versus predicted group membership) shown in Table 2.

Table 2: Confusion matrix.

		Predicted Group Membership		Total
		π_1	π_2	
Actual Group Membership	π_1	n_1^c	$n_1^m = n_1 - n_1^c$	n_1
	π_2	$n_2^m = n_2 – n_2^c$	n_2^c	n_2

π_1 and π_2 stand for populations of groups 1 and 2, while n_1 and n_2 stand for number of observations (items) in both groups, respectively. n_1^c is the number of items belonging to π_1 and classified as belonging to π_1, n_2^c is the same for π_2. n_1^m is the number of items belonging to π_1 and (mis)classified as belonging to π_2 while n_2^m is the number items belonging to π_2 and (mis)classified as belonging to

π_1. The APER is defined as the proportion of items misclassified by the allocation rule:

$$\text{APER} = \frac{n_1^m + n_2^m}{n_1 + n_2}$$

3 Results

The analysis was performed with statistical packages R (v 2.1) and SPSS (v 12.0). To meet the assumptions of ANOVA, the mean APER (defined as the mean value of classification error at each parameter combination in 100 simulations) was transformed to normal-score ranks by first ranking the mean APER values and then applying the Blom transformation (as implemented in SPSS). The results were analysed by ANOVA with normal-score ranks (nsAPER) as the dependent variable and with sample size, skewness (asymmetry coefficient) and rotation angle as independent variables. The main objective of the analysis was to determine the effect of independent variables upon the classification results.

The Table 3 summarises the results. The results show that group size does affect classification. This is not surprising regarding the fact that with increasing sample size lower misclassification rate is expected, since increasing the number of observations leads to more accurate discrimination function and thus to a better classification procedure.

Classification performance is also affected by rotation: with increasing rotation angle lower misclassification rate is expected. This is due to the fact that rotation diminishes group overlapping, facilitating formulation of a good discrimination function and, consequently, better classification (see Figure 3).

Table 3: Analysis of variance table.

Independent variables	Df	Sum Sq	Mean Sq	F value	Pr (>F)	Sig.
size	2	2596.5	1298.2	3067.88	$< 10^{-15}$	***
asymm	3	2.2	0.7	1.71	0.1624	
rotat	9	3923.0	435.9	1030.05	$< 10^{-15}$	***
size:asymm	6	1.9	0.3	0.75	0.6053	
size:rotat	18	252.9	14.1	33.20	$< 10^{-15}$	***
asymm:rotat	27	84.3	3.1	7.38	$< 10^{-15}$	***
size:asymm:rotat	54	26.3	0.5	1.15	0.2069	
residuals	11880	5027.3	0.4			

Dependent Variable: nsAPER

Moderate interactions are noted between the group size and rotation and between skewness and rotation. The former is plausibly explained by the combination of the group size and rotation angle effects above. The latter indicates that the major difference between the effects of the level of the rotation angle is also affected by the level of skewness. Again, this is not surprising: a small rotation angle in association with a high skewness may be expected to degrade classification performance.

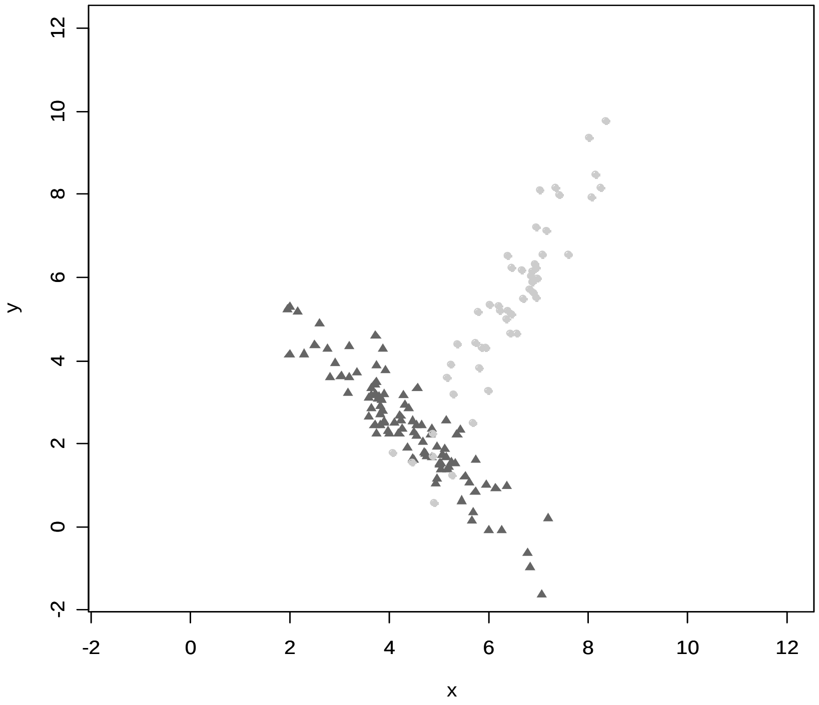

Figure 3: 2-dimensional (xy) geometry of a simulation instance with n = 100, λ_1 = 10, α = 80° for π_1 (dark grey triangles) and n = 50, λ_1 = 0, α = 0° for π_2 (light grey circles). Note skewness in π_1.

On the other hand, skewness by itself can not be demonstrated to have any significant effect on classification, indicating the Fisher's DF to be rather insensitive to skewness.

Figures 4–6 (below) illustrate the stated results. (Each box in the figures shows the median, quartiles and extreme values within a category.)

Figures 4 and 5 depict lowering of mean APER values with increasing sample size and increasing rotation angle. On the contrary, no such effect is noticeable in Figure 6, confirming the conclusion of no identifiable effect of skewness on APER. This visibly confirms that the performance of Fisher's DF does not seem to be affected by skewness, thus substantiating its robustness in the presence of skewness induced by using SCN distributed data.

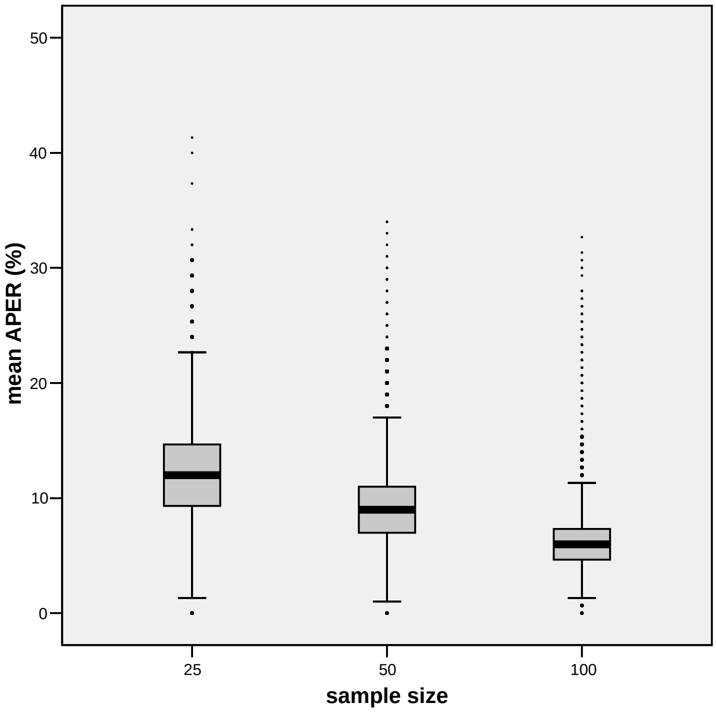

Figure 4: Box-plot of mean APER vs. sample size.

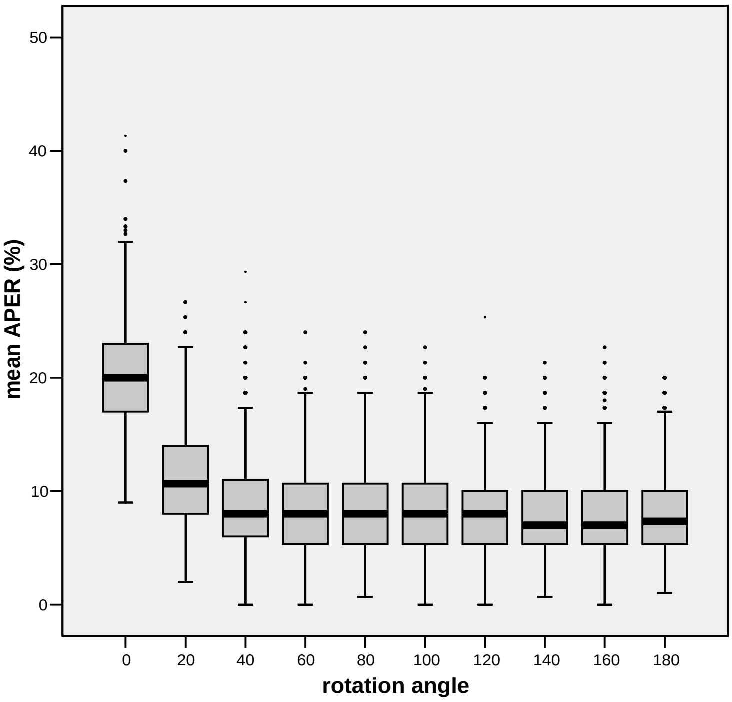

Figure 5: Box-plot of mean APER vs. rotation angle.

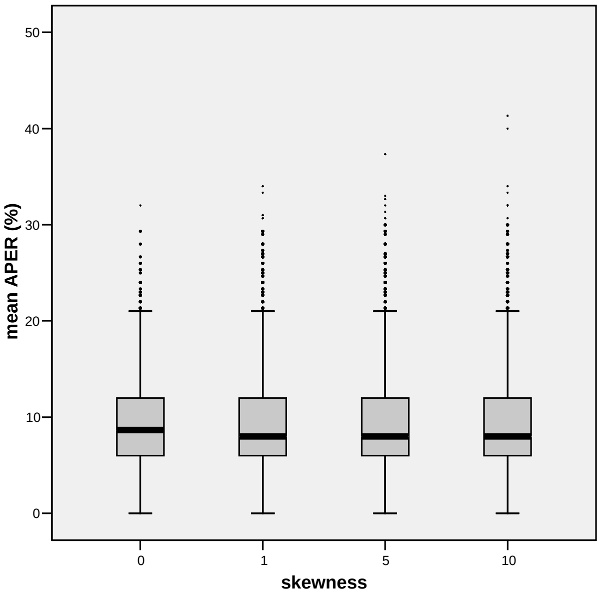

Figure 6: Box-plot of mean APER vs. skewness.

4 Conclusion

The purpose of the experiment described in this paper was to examine the robustness of the Fisher's discriminant function to skewness introduced by data following the skew-curved normal distribution. It was, *inter alia*, motivated by the still valid statement of Krzanowski (1977) who emphasised the value of information on performance of the Fisher’s discriminant function in non-optimal conditions. Since the size of groups and their spatial orientation might also be expected to be involved in multivariate settings, the effects of group size and rotation angle was analysed, too.

For generating multivariate non-normality we used a relatively new class of distributions, called skew-curved normal distributions, which include the normal distribution as a special case. Asymptotically, these distributions do not turn into the half-normal and thus provide more smooth skewness that Azzalini's skew-normal distribution.

Results indicate that Fisher's DF (as an allocation rule) is appreciably affected by sample size and rotation, while introduction of skewness could not be demonstrated to have any significant effect on classification. Therefore, Fisher’s DF seems to be rather insensitive/fairly robust to skewness introduced into the model via the SCN distributed simulated data.

A possible further step in the investigation could be to assess the predictive performance of Fisher’s DF using another performance measure.

Acknowledgement

The authors are indebted to Gaj Vidmar, M. Sc., Institute of Biomedical Informatics, University of Ljubljana, for his valuable advice on data transformation.

References

[1] Arellano-Valle, R.B., Gomez, H.W., and Quintana, F.A. (2004): A new class of skewnormal distributions. *Communications in Statistics – Theory and Methods*, **33**, 1465-1480.

[2] Ashikaga, T. and Chang, P.C. (1981): Robustness of Fisher's linear discriminant function under two-component mixed normal models. *Journal of American Statistical Association*, **76,** 676-680.

[3] Azzalini, A. and Capitanio, A. (1999): Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society*, **61**, 579-602.

[4] Balakrishnan, N. and Kocherlakota, S. (1985): Robustness of non-normality of the linear discriminant function: Mixtures of normal distributions. *Communications in Statistics – Theory and Methods*, **14(2)**, 465-478.

[5] Dillon, W.R. (1979): The performance of the linear discriminant function in non-optimal situations and the estimation of classification error rates: A review of recent findings. *Journal of Marketing Research*, **19**, 270-391.

[6] Härdle, W. and Simar, L. (2003): *Applied Multivariate Statistical Analysis*. Berlin: Springer.

[7] Johnson, R.A. and Wichern, D.W. (2002): *Applied Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

[8] Kumar, N. and Andreou, A.G. (1996): A generalization of linear discriminant analysis in maximum lokelihood framework. *Proceedings of the Joint Statistical Meeting*, *Statistical Computing Section*, 1-6.

[9] Krzanowski, W.J. (1977): The performance of Fisher's linear discriminant function under non-optimal conditions. *Technometrics*, **19**, 191-200.

[10] Lachenbruch, P.A., Sneeringer, C., and Revo, L.T. (1973): Robustness of the linear and quadratic discriminant function to certain types of non-normality. *Communications in Statistics*, **1**, 39-56.

[11] L'Ecuyer, P. (2004): Random number generation. In J.E. Gentle, W. Haerdle, and Y. Mori (Eds): *Handbook of Computational Statistics*, 35-70. Berlin: Springer.

[12] Mardia, K.V. (1970): Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519-530.

[13] Pohar, M., Blas, M., and Turk, S. (2004): Comparison of logistic regression and linear discriminant analysis: A simulation study. *Metodološki zvezki*, **1**, 143-161.

[14] Sharma, S. (1996): *Applied Multivariate Techniques*. New York: John Wiley & Sons.

[15] Tabachnick, B.G. and Fidell, L.S. (1996): *Using Multivariate Statistics*. New York: Harper Collins College Publishers.

[16] Velilla, S. and Hernández, A. (2002): On the consistency and robustness properties of linear discriminant analysis. *Statistics and Econometrics Series Working Papers*.