DE GRUYTER OPEN

DOI: 10.1515/bile-2015-0006

Biometrical Letters
Vol. 52 (2015), No. 2, 55 - 74

Directional representation of data in Linear Discriminant Analysis

Jolanta Grala-Michalak

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Umultowska 87, 61-614 Poznań, Poland, grala@amu.edu.pl

SUMMARY

Sometimes feature representations of measured individuals are better described by spherical coordinates than Cartesian ones. The author proposes to introduce a pre-processing step in LDA based on the arctangent transformation of spherical coordinates. This nonlinear transformation does not change the dimension of the data, but in combination with LDA it leads to a dimension reduction if the raw data are not linearly separated. The method is presented using various examples of real and artificial data.

Keywords LDA, pattern recognition, spherical coordinates, dimension reduction, PCA, directional statistics

1. Introduction

The published literature on pattern recognition and classification contains many algorithms for use on data defined by a pair of polar coordinates (R, α), with R as the distance between a sample element $\boldsymbol{x}$ and the centroid, and α the angular position of $\boldsymbol{x}$ with respect to a coordinate axis. There are several reasons for using this description:

- some sensors naturally produce datasets in the form of a distance and an angle (in optics, e.g. Matsushima and Marcus, 1995, in face recognition – Sajjanhar et al., 2007);
- classical Cartesian coordinates do not enable one to find discriminant surfaces that entirely enclose groups very closely without any misclassification for data that are not linearly separated (Duchene, 1987);

- we want to use a considerable amount of parameters, so first we reduce the dimension of the initial space without losing any significant information with the use of principal component analysis, and we are then interested in the angular position of $\boldsymbol{x}$ with respect to the principal axes (Duchene and Leclerq, 1988).

In this paper it is assumed that the nature of the data is not known.

The aim of the study is to present a pre-step in LDA which will improve the results if the data are ill-posed (not linearly dependent).

The paper contains a description of two systems of coordinates (Chapter 2), an overview of the most often used methods of reducing dimensionality (Chapter 3), a description of the methodology (Chapter 4) and datasets (Chapter 5), the results (Chapter 6), some comments (Chapter 7) and conclusions (Chapter 8).

2. Two systems of coordinates

One possibility is to analyse the angles between orthogonal projections onto planes and one of the axes spanning the plane, by means of spherical coordinates (see Fig. 1, left). A second approach allows us to look at closely lying points from another visual angle. Let us introduce spherical coordinates $(R, \alpha, \beta, \delta, \ldots, \psi, \omega)$ according to the formulae

$$x_1 = R\cos\omega\cos\psi \ \ldots \cos\delta\ \cos\gamma\ \cos\beta\ \cos\alpha$$

$$x_2 = R\cos\omega\cos\psi \ \ldots \cos\delta\ \cos\gamma\ \cos\beta\ \sin\alpha$$

$$x_3 = R\cos\omega\cos\psi \ \ldots \cos\delta\ \cos\gamma\ \sin\beta$$

$$x_4 = R\cos\omega\cos\psi \ \ldots \cos\delta\ \sin\gamma$$

$$x_5 = R\cos\omega\cos\psi \ \ldots \sin\delta$$

$$\ldots$$

$$x_{p-1} = R\cos\omega\ \sin\psi$$

$$x_p = R\sin\omega.$$

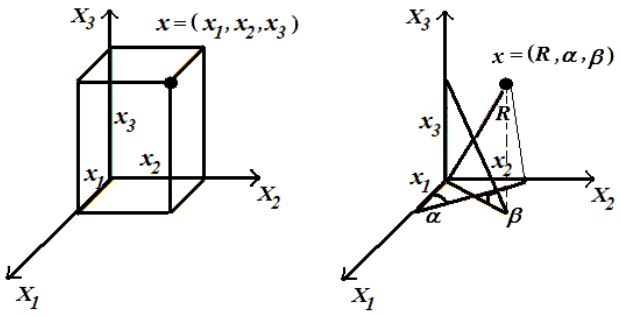

Figure 1. Two representations of a point x in $\Re^3$: in Cartesian coordinates (left) and in spherical coordinates (right)

These p coordinates $(R, \alpha, \beta, \delta, \ldots, \psi, \omega)$ may be derived from a set of p Cartesian coordinates $(x_1, x_2, \ldots, x_p)$ using a transformation that is locally invertible (a one-to-one map) in a neighbourhood of each point (Mardia, 1972, Mardia and Jupp, 2000).

This transformation will be called the arctangent transformation:

$$R^2 = x_1^2 + x_2^2 + \cdots + x_p^2,$$

$$\alpha = arctg\left(\frac{x_2}{x_1}\right) + \pi \cdot \mathrm{I}_{\{x_1 < 0\}},$$

$$\beta = arctg\left(\frac{x_3}{\sqrt{x_1^2 + x_2^2}}\right),$$

$$\gamma = arctg\left(\frac{x_4}{\sqrt{x_1^2 + x_2^2 + x_3^2}}\right),$$

$$\delta = arctg\left(\frac{x_5}{\sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2}}\right),$$

…

$$\psi = arctg\left(\frac{x_{p-1}}{\sqrt{x_1^2 + x_2^2 + \cdots + x_{p-2}^2}}\right),$$

$$\omega = arctg\left(\frac{x_p}{\sqrt{x_1^2 + x_2^2 + \cdots + x_{p-1}^2}}\right),$$

where the function I_A is the characteristic function of the set *A*. We should remember that the function arc tg is the inverse of the function tg on the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. The formula for the angle α is different from the others because of the possibility of negative values of x_1. In the other formulae the denominator cannot be negative.

This description, in some situations, may be more adequate than the standard Cartesian system of coordinates. If this is the case, we obtain better results for linear discrimination, which implies that we may have found the correct pattern in the data.

As it turns out, taking such a new view of well-known measurements can, in some cases, improve the quality of linear discrimination or enable a reduction of significant features without the use of Principal Component Analysis (PCA).

If our data are in some kind of nonlinear pattern, the classical Linear Discriminant Analysis (LDA) is not effective or does not work. Then it is reasonable to apply the kernel method, where the raw data are nonlinearly transformed to a high-dimensional feature space (Hilbert space), where the pattern should appear linear. However, we do not know how high the dimension should be and what kernel we should choose. The simplest method proposed here is to first transform the raw data in Cartesian coordinates in $\Re^p$ into spherical coordinates in $\Re^p$, when $p > 2$, or polar coordinates when $p = 2$. If the arctangent transformation reduces the data pattern to almost linear form, it will not be necessary to use such complicated methods as, for example, Kernel Uncorrelated and Orthogonal Discriminant Analysis.

3. The most often used methods of reducing the dimensionality of data in discriminant analysis

A range of different parametric and nonparametric methods of pattern recognition are presented by Aeberhard et. al. (1992). They describe two possible ways of dealing with ill-posed data: dimensionality reduction and regularization. Dimensionality reduction can be achieved by feature selection or by transforming the full feature space into a lower-dimensional space called reduced space. Regularization is the procedure of biasing parameters in order to reduce the variance of the estimators. They compared the classifiers in full feature space (quadratic discriminant analysis, linear discriminant analysis, regularized discriminant analysis, 1-nearest neighbour method), and after reducing the dimensionality (Fisher's discriminant plane, Fisher-Fukunaga-Koonz transformation, Fisher-radius plane transformation, Duchene and Leclerc's plane). The highest probabilities of correct classification were achieved for different methods depending on the data.

In linear discriminant analysis for several groups, we are concerned with finding linear combinations of variables that best separate groups of multivariate observations. Moreover, it happens that classical LDA fails to clearly recognize groups. For example, discriminant functions are not linear, or intergroup covariance matrices are significantly different, or dependences between variables are not linear and classical techniques (Pearson's correlation coefficients, LDA) are not adequate. In this situation a method of nonlinear dimension reduction is suggested. Such methods are based on low-dimensional surface embedding in high-dimensional space with the help of a nonlinear transformation.

A technique for exploratory analysis, referred to by Friedman and Tukey (1974) as 'projection pursuit' (Everitt at all, 2011) lets us apply lower-dimensional projections of multivariate data. This is very significant in any graphical presentation of data.

Kernel methods are a very good solution in the case of nonlinear pattern datasets. First, data items are embedded into a vector space called the feature

space, where the nonlinear pattern appears linear. The coordinates of the embedded points are not needed, only their pairwise inner products, which can be computed efficiently directly from the original data items using a kernel function (Shawe-Taylor and Cristianini, 2004).

Many authors suggest first performing the reduction of dimensionality and then carrying out any method of discriminant analysis; this gives rise to what are called two-stage methods. Principal Component Analysis (PCA) is the most often used method to prepare data for analysis. An overview of PCA over the last 60 years is presented by Trendafilov (2013).

Classical LDA lacks the capacity to capture a nonlinearly clustered structure in data. Nonlinear extensions of LDA use the kernel trick: mapping the data in the original space to a feature space where an inner product can be computed by a kernel function without requiring knowledge of the nonlinear mapping function explicitly. There is one disadvantage: the dimension of the feature space is often much larger than that of the original data space, and any singularity problem in the original data space becomes more severe. Ways of solving this problem are described by Xiong et al. (2006) as Kernel Uncorrelated and Orthogonal Discriminant Analysis – the unification of known discriminant methods.

4. The suggested methodology

Let us assume that $\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in_i}$ form a *p*-dimensional learning sample from the *i*-th population (where *i* = 1, 2, ..., *K* and $n_1 + n_2 + \cdots + n_K = n$). The mean vector of the *i*-th population is calculated as $\overline{\boldsymbol{X}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \boldsymbol{X}_{ij}$ and the variance-covariance matrix of the *i*-th population is $\boldsymbol{S}_i = (n_i - 1)^{-1}\sum_{j=1}^{n_i}\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i\right)\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_i\right)^T$ for *i* = 1,..., *K*. From the whole learning sample $\boldsymbol{X}$ of n elements we calculate $\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{K} n_i \overline{\boldsymbol{X}}_i$, $\boldsymbol{B} = \sum_{i=1}^{K} n_i(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^T$ and $\boldsymbol{W} = \sum_{i=1}^{K}(n_i - 1)\boldsymbol{S}_i$. In linear discriminant analysis we are looking for a set of

constant vectors $\{\boldsymbol{a}_i: i = 1, \ldots, k\}$ which maximizes the expression $\frac{(n-k)\boldsymbol{a'Ba}}{(k-1)\boldsymbol{a'Wa}}$ on condition that $\frac{\boldsymbol{a}_i\boldsymbol{'W}\boldsymbol{a}_j}{n-k} = \delta_{ij}$ (Krzyśko et al., 2008).

Let $\boldsymbol{a}^T = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)$ be a matrix built from the $\boldsymbol{a}_i$'s and δ_{ij} be the Kronecker delta (i.e. 1 if $i = j$ and 0 if $i \neq j$). This criterion means that the new variables (discriminant functions) $u_i = \boldsymbol{a}_i^T\boldsymbol{x}$ are uncorrelated with variances equal to 1 for every i. The first discriminant function u_1 is related to the first (i.e. the largest) eigenvalue λ_1 of the matrix $\left(\frac{\boldsymbol{W}}{n-k}\right)^{-1}\left(\frac{\boldsymbol{B}}{k-1}\right)$, the second is related to λ_2 , and so forth. The vectors $\boldsymbol{a}_i$ are the eigenvectors corresponding to the λ_i's, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s > 0$, and are the solutions of the determinant equation $\left(\boldsymbol{B} - \frac{k-1}{n-k}\lambda_i\boldsymbol{W}\right)a_i = \boldsymbol{0}$, for $i = 1, \ldots, s$.

In the classification process, the discriminant variable u_i is not useful if λ_i is not significantly different from zero. First we test the hypothesis that all eigenvalues are equal to 0, next that this holds for all apart from the first, etc. This procedure is continued until for the first time we fail to reject the hypothesis. Then we state that from that value the remaining (n - s) eigenvalues are equal to 0. The test statistic used is Wilks' lambda in the form $\Lambda_d = \prod_{i=d+1}^{s}\frac{1}{1+\lambda_i}$, where d = 0, 1, …, p-1, which has an asymptotic χ^2 distribution (more in Krzyśko et al., 2008).

In the next step the testing sample is classified according to the model based on the learning sample.

The cumulative percentage of variation is a good measure of the importance of a single variable. The level of correct classification according to the 1-NN method (in %) indicates the quality of classification.

Here the following methodology is proposed.

(1) Take the learning sample and perform classical LDA in order to find linear discriminant functions and verify the results of classification. If the results are satisfactory, go to (4); if not, go to (2),

(2) Spherically transform the raw data before LDA and perform the arctangent transformation. If the results are satisfactory, go to point (4).

(3) If the results are not sufficiently good, try to use some other methods of discrimination.
(4) Test the obtained discriminant functions on a testing sample and formulate conclusions.

There are several hints related to points above.
1) At first you had better try the linear method.
2) It is necessary to order p variables in order of ascending variance. This allows us to make smaller differences in values more significant in the arctangent transformation than in linear mapping. Centring of data is not necessary.

This analysis is possible to perform on all kinds of data, irrespective of their dimension, and in contrast to data in Cartesian coordinates, does not require pre-centring. For every observed value we must have $x_1 \neq 0$, because the point (0, ... ,0) has no polar/spherical representation. If this condition does not hold, some small number must be added to every value of x_1 in every observation to avoid this problem. Next we transform the data according to the formulae in chapter 2.
3) Other methods which may be used are mentioned in Chapter 3.
4) It is suggested that conclusions be drawn from the better model chosen according to certain criteria (for example, the percentage of correct classification).

5. Datasets

The proposed methodology is verified on real (A, B, C, D, E, F, G) and simulated (H) data of different dimensions.
A) Fisher's IRIS DATA – the best-known database. For three types of Iris flowers (50, 50 and 50 individuals) measurements of four variables were taken (from the UCI Machine Learning Repository).
B) BUGS DATA (Lubischew, 1962) from the UCI Machine Learning Repository, contains 6 variables and 3 classes (21, 22 and 31 individuals).

C) CRIMES DATA (Freedman 1975) from Rencher, Christensen, 2012, page 507 – 7 variables, 4 classes (6, 5, 3 and 2 individuals). The data represent crime rates per 100,000 population for US cities. Four clusters were identified according to several methods of clustering: complete linkage, average linkage, centroid clustering, median clustering, Ward's method, flexible beta method (the same clusters in all methods).

D) AIR POLLUTION DATA (Everitt, 2011) – 7 variables, 4 classes (6, 17, 15 and 4 individuals), with one outlier (Chicago) included in the fourth class. Four clusters were defined according to Ward's method (the most reasonable division).

E) WINES DATA – the result of a chemical analysis of wines grown in the same region in Italy but from three different cultivars. The data is 13-dimensional with 59, 71 and 48 training samples per class. (UCI Machine Learning Repository)

F) SEEDS DATA from the UCI Machine Learning Repository, 7 variables, 3 classes (70, 70 and 70 individuals).

G) THYROID DATA from the UCI Machine Learning Repository, 5 variables, 3 classes (165, 25 and 25 individuals).

H) BUTTERFLY DATA (in Duchene, 1987) – created from Gaussian distributions in 2d-space:

group I: 30 points from $N(\begin{bmatrix}-2\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix})$ and 30 points from $N(\begin{bmatrix}2\\0\end{bmatrix},\begin{bmatrix}1&0\\0&1\end{bmatrix})$,

group II: 30 points from $N(\begin{bmatrix}0\\-5\end{bmatrix},\begin{bmatrix}2&0\\0&1\end{bmatrix})$ and 30 points from $N(\begin{bmatrix}0\\5\end{bmatrix},\begin{bmatrix}2&0\\0&1\end{bmatrix})$.

6. Results

The calculations were done in STATISTICA 10. The learning and testing sample are the same, so the percentage of correct classification is very high. This is not important, because our interest is in comparing the quality of the two methods.

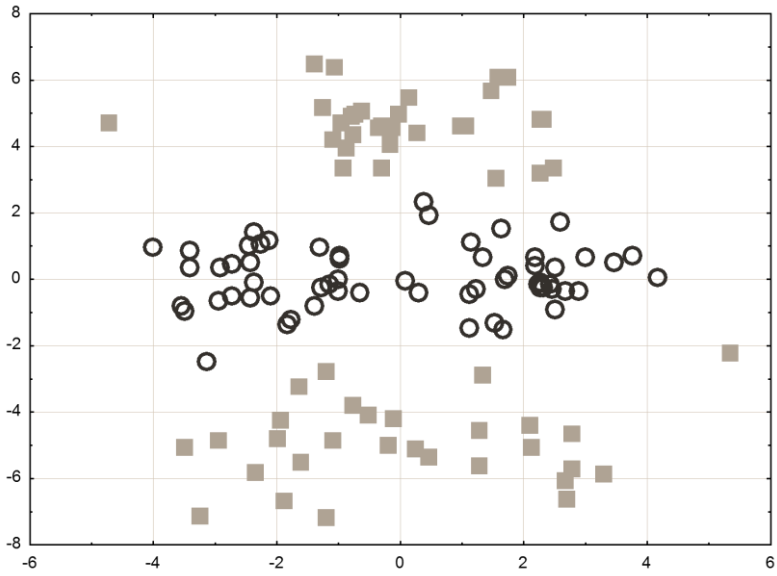

Figure 2. BUTTERFLY DATA in the Cartesian plane

A) IRIS DATA – see Table 1. There is no dimension reduction, because the number of significant variables is the same in both methods. The data projections on the first-and-second-discriminant-function plane in LDA are placed in separate regions according to their groups (see Fig. 3). Obviously, this is such a clear division that it cannot be improved by any other method. In fact, the first discriminant function explains 99% of the total variation. Hence the projections of the data on the abscissa allow classification into groups.

B) BUGS DATA – see Table 2 and Fig. 3. The reduction of dimension is visible. Instead of six variables it is enough to use only four for the linear classification functions after the arctangent transformation, and the result remains at the same (very good) level.

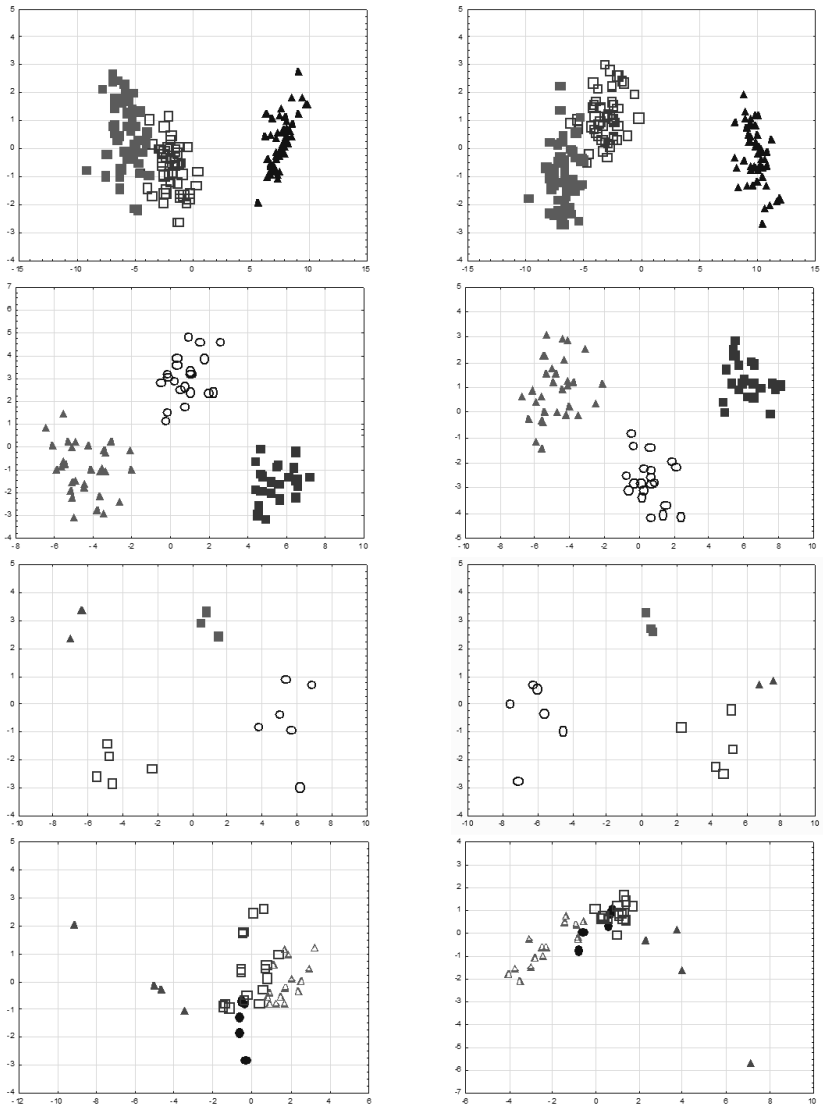

Figure 3. From top: IRIS, BUGS, CRIMES and AIR POLLUTION DATA on the (1,2)-discriminant plane in LDA (left) and after the arctangent transformation (right)

Table 1. Comparison of LDA without and following the arctangent transformation on IRIS DATA. Wilks' Λ test was used to determine significant (at level 0.05) variables. The table gives eigenvalues of the sample correlation matrix, tested by a χ^2 test at level 0.05, and the percentage of explained total variance

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
V1 (0.010)	-	R (0.000)	-
V2 (0.000)		alpha(0.000)	
V3 (0.000)		beta(0.000)	
V4 (0.000)		gamma(0.000)	
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
32.192 (0.000)	99	51.626 (0.000)	99
0.285 (0.000)	100	0.770 (0.000)	100
Class number	CC* (in %)	Class number	CC* (in %)
1	100	1	100
2	94	2	94
3	94	3	96

CC*Correct classification according to 1-NN method (in %)

C) CRIMES DATA – see Table 3 and Fig. 3. After the transformation, the radius makes it possible to distinguish groups, instead of the two significant variables in classical LDA.

D) AIR POLLUTION DATA – see Table 4 and Fig. 3. The proposed method is slightly worse than LDA because instead of two there are three significant variables. Moreover, the interpretation of the variables in classical LDA is more logical – the climate is negligible, the industrial level is important.

E) SEEDS DATA – see Table 6 and Fig. 4. A small dimension reduction (from six to five variables) and 100% correct classification after the arctangent transformation can be observed.

Table 2. Comparison of LDA without and following the arctangent transformation on BUGS DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
fjft (0.000) sjft (0.000) mwhbee (0.000) mwafp (0.000) faa(0.000) awfs (0.026)	-	R (0.000) alpha (0.000) gamma (0.000) epsilon (0.000)	beta (0.0.056) delta 90.345)
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
17.779 (0.000)	82	22.005 (0.000)	88
3.885 (0.000)	100	3.025 (0.000)	100
Class number	CC* (in %)	Class number	CC* (in %)
1	100	1	100
2	100	2	100
3	100	3	100

CC*Correct classification according to 1-NN method (in %)

Table 3. Comparison of LDA without and following the arctangent transformation on CRIMES DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
Burglary (0.001) Larceny (0.030)	Murder (0.663) Rape (0.207) Robbery (0.073) Assault (0.132) AutoTheft (0.6580)	R (0.001)	alpha (0.701) beta (0.592) gamma (0.294) delta (0.410) epsilon (0.403) z eta (0.278)
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
30.701 (0.000)	81	36.611 (0.000)	90
5.631 (0.000)	96	3.171 (0.095)	98
1.348 (0.15)	100	0.720 (0.397)	100
Class number	CC* (in %)	Class number	CC* (in %)
1	100	1	100
2	100	2	98
3	100	3	100
4	100	4	100

CC*Correct classification according to 1-NN method (in %)

Table 4. Comparison of LDA without and following the arctangent transformation on AIR POLLUTION DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
Manuf (0.031)	SO2 (0.522)	**R (0.000)**	alpha (0.117)
Pop (0.000)	Temp (0.982)	**delta (0.016)**	beta (0.632)
	Wind (0.545)	**epsilon(0.000)**	gamma (0.710)
	Precip (0.349)		zeta (0.058)
	Days (0.363)		
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
4.654 (0.000)	90	4.323 (0.000)	77
0.363 (0.222)	97	0.874 (0.000)	93
0.145 (0.459)	100	0.415 (0.035)	100
Class number	CC* (in %)	Class number	CC* (in %)
1	83	1	50
2	71	2	88
3	93	3	100
4	100	4	75

CC*Correct classification according to 1-NN method (in %)

F) WINES DATA – see Table 5 and Fig. 4. The new method operates on only seven instead of nine variables. The proposed method produces smaller and better separated clusters than LDA (see Fig. 4) and gives 100% correct classification.

G) THYROID DATA – see Table 7 and Fig. 4. The linear combination of the angles beta, gamma and delta in spherical representation has the same power of discrimination as a linear combination of five variables in classical LDA, but the effect in Fig. 4 is similar.

H) BUTTERFLY DATA – see Table 8 and Fig. 2. Cartesian coordinates cannot differentiate the groups, but polar coordinates (specifically the radius) do it very well. In Fig. 2 observations coming from one group are clustered in the centre, and those from the other far from the centre of the plane. Therefore there are no significant discriminant variables in classical LDA. The borders of discriminant regions are adequately described by two concentrically placed circumferences.

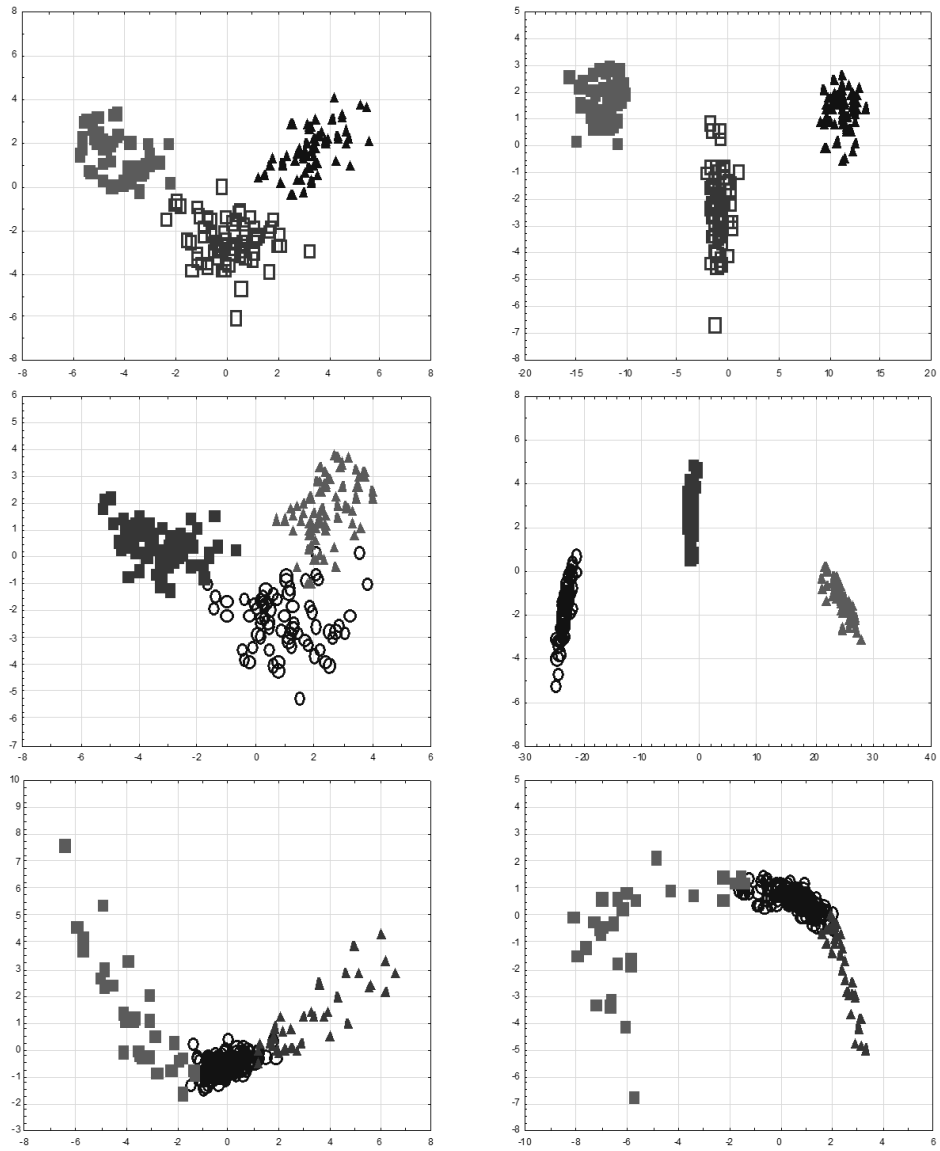

Figure 4. From top: WINES, SEEDS and THYROID DATA on the (1,2)-discriminant plane in LDA (left) and after the arctangent transformation (right)

Table 5. Comparison of LDA without and following the arctangent transformation on WINES DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
V1(0.000)	V5(0.949)	R(0.000)	beta (0.118)
V2(0.003)	V6(0.080)	alpha(0.000)	delta (0.254)
V3(0.000)	V8(0.055)	gamma(0.000)	epsilon (0.157)
V4(0.000)	V9(0.287)	zeta(0.000)	eta (0.423)
V7(0.000)		theta(0.000)	iota(0.536)
V10(0.000)		kappa(0,000)	lambda(0.069)
V11(0.018)		mu(0.022)	
V12(0.000)			
V13(0.000)			
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
9.082 (0.000)	69	85.704 (0.000)	96
4.128 (0.000)	100	3.395 (0.000)	100
Class number	CC* (in %)	Class number	CC* (in %)
1	100	1	100
2	100	2	100
3	98	3	100

CC*Correct classification according to 1-NN method (in %)

Table 6. Comparison of LDA without and following the arctangent transformation on SEEDS DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
V1(0.000)	V5(0.950)	R (0.000)	beta (0.214)
V2(0.000)		alpha (0.047)	epsilon (0.606)
V3(0.000)		gamma (0.027)	
V4(0.000)		delta (0.032)	
V6(0.000)		zeta (0.278)	
V7(0.000)			
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
6.237(0.000)	88	387.766 (0.000)	99
2.916 (0.000)	100	4.649 (0.000)	100
Class number	CC*(in %)	Class number	CC*(in %)
1	94	1	100
2	99	2	100
3	99	3	100

CC*Correct classification according to 1-NN method (in %)

Table 7. Comparison of LDA without and following the arctangent transformation on THYROID DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
V1(0.003) **V2(0.000)** **V3(0.000)** **V4(0.000)** **V5(0.000)**		**beta (0.000)** **gamma (0.000)** **delta (0.000)**	R (0.325) alpha (0.060)
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
3.839 (0.000)	84	5.647 (0.000)	84
0.732 (0.000)	100	0.870 (0.000)	100
Class number	CC*(in %)	Class number	CC*(in %)
1	98	1	99
2	88	2	88
3	96	3	96

CC*Correct classification according to 1-NN method (in %)

Table 8. Comparison of LDA without and following the arctangent transformation on BUTTERFLY DATA

Classical LDA		**Arctangent transformation+LDA**	
Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)	Significant variables (Wilks'Λ p-value)	Non-significant variables (Wilks'Λ p-value)
-	X (0.841) Y (0.778)	**R (0.000)**	alpha (0.409)
Eigenvalues (p-value)	Cumulative % of variation	Eigenvalues (p-value)	Cumulative % of variation
0.001 (0.943)	100	2.082 (0.000)	100
-	-	Class number	CC*(in %)
-	-	1	93
		2	88

CC*Correct classification according to 1-NN method (in %)

7. Comment

Let us assume that there are four points in a learning sample $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4$ with Cartesian coordinates (1, 2), (2, 2), (1, 1) and (2, 1) respectively. Points $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$ (or $\boldsymbol{x}_2$ and $\boldsymbol{x}_4$) differ in their second Cartesian coordinate, whereas $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

(or $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$) differ in the first. Moreover $\boldsymbol{x}_1$ and $\boldsymbol{x}_4$ have the same radius (the first coordinate in the polar coordinate system) and points $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ have the same second coordinate (the angle) in the polar coordinate system.

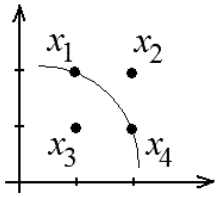

Figure 5. Example of four data points

Let us consider the different labelling of the four points representing the learning sample in the example. Table 1 shows the better model in discriminant analysis.

Table 9. Choice of model

Point→label	**Better method**	**The best discriminant variable**
$\boldsymbol{x}_1, \boldsymbol{x}_2 \rightarrow$ a $\boldsymbol{x}_3, \boldsymbol{x}_4 \rightarrow$ b	classical LDA	the second coordinate in the Cartesian system
$\boldsymbol{x}_1, \boldsymbol{x}_3 \rightarrow$ a $\boldsymbol{x}_2, \boldsymbol{x}_4 \rightarrow$ b	classical LDA	the first coordinate in the Cartesian system
$\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_4 \rightarrow$ a $\boldsymbol{x}_2 \rightarrow$ b	arctangent transformation+LDA	the radius in the polar coordinate system
$\boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4 \rightarrow$ a $\boldsymbol{x}_1 \rightarrow$ b	arctangent transformation+LDA	the angle in the polar coordinate system

8. Conclusions

According to Duchene (1987), classical Cartesian coordinates do not make it possible to find discriminant surfaces that entirely enclose groups very closely without any misclassification for data that are not linearly separated. If the discriminant surfaces based on the learning data are non-convex manifolds in

$\Re^p$, the suggested method is more adequate in data description. The discriminant regions in LDA have hyperplane borders, but the described transformation changes this into surfaces.

The simplicity of the suggested method consists in the application of a different description of a sample. We view data points as angles between projections on perpendicular axes and hyperplanes, not as projections on perpendicular Cartesian axes. This method works even for non-centred data. This description, in some situations, may be more suitable than one based on the Cartesian system of coordinates. If this is the case, we obtain better results from LDA, which means that we have found the correct pattern in the data.

By first applying the arctangent transformation to datasets we may:

- obtain a better quality of linear discrimination of ill-posed data (cf. BUTTERFLY DATA, SEEDS DATA);
- achieve a reduction of significant variables without using PCA (cf. BUGS DATA, CRIMES DATA, WINES DATA, SEEDS DATA, THYROID DATA);
- obtain the same quality by both methods (cf. IRIS DATA), if the raw data are linearly separated;
- obtain a worse result (AIR POLLUTION DATA).

The arctangent transformation does not make small differences between the values of variables (presented as Cartesian coordinates) negligible, because of the ordering in order of ascending variance. The dimension of the data does not increase as in kernel methods.

Finally, use of the suggested method is beneficial when:

- classical LDA is not useful because of the nonlinear pattern of data;
- there are large differences between the variances of variables;
- the data are not linearly separated;
- small differences in values of the data lead to classification in a different group.

REFERENCES

Aeberhard S., Coomans D., de Vel O. (1992): The performance of statistical pattern recognition methods in high dimensional settings. Tech. Rep. No 92-02, Dept. Of Computer Science and Depth. Of Mathematics and Statistics, James Cook University of North Queensland.

Duchene L. (1987): A New Form of Discriminant Surfaces Using Polar Coordinates. Pattern Recognition 20(4): 437- 442.

Duchene L., Leclerq S. (1988): An Optimal Transformation for Discriminant and Principal Component Analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 10(6): 978-983.

Everitt B.S., Landau S., Leese M., Stahl D. (2011): Cluster Analysis. Wiley.

Hartigan J.A. (1975) Clustering Algorithms. New York, Wiley.

http://ics.uci.edu/~mlearn/MLRepository.html (UCI Machine Learning Repository)

Krzyśko M., Wołyński W., Górecki T., Skorzybut M. (2008): Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości, WNT, Warsaw. (In Polish)

Matsushima T., Marcus P.S. (1995): *A Spectral Method for Polar Coordinates*. Journal of Computational Physics 120: 365-374.

Mardia K.V. (1972): Statistics of Directional Data. Academic Press, London.

Mardia K.V., Jupp P.E. (2000): Directional Statistics. Wiley Series in Probability and Statistics.

Rencher A.C., Christensen W.F. (2012): Methods of Multivariate Analysis, Third Edition. Wiley.

Sajjanhar A., Lu G., Zhang D. (2007): A Composite Descriptor for Shape Retrieval. Proceedings of the 6th IEEE/ACIS International Conference on Computer and Information Science, IEEE Computer Society, Melbourne, Australia: 795-800.

Shawe-Taylor J., Cristianini N. (2004): Kernel Methods for Pattern Analysis. Cambridge University Press.

Trendafilov N.T. (2013): From simple structure to sparse components: a review. Comput. Stat. (Online first article) 10.1007/500180-013-0434-5

Xiong T., Ye J., Cherkassky V. (2006): Kernel Uncorrelated and Orthogonal Discriminant Analysis: A Unified Approach. Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06).