Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v13n2p284

Discrimination and classification model from Multivariate Exponential Power Distribution

By Olosunde, Soyinka

Published: 14 October 2020

Discrimination and classification model from Multivariate Exponential Power Distribution

A.A. Olosunde*[a] and A.T. Soyinka[b]

[a] *Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria, akinolosunde@gmail.com*

[b] *Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria, soyinkataiwo@yahoo.ie*

Published: 14 October 2020

It is common to assume a normal distribution when discriminating and classifying a multivariate data based on some attributes. But when such data is lighter or heavier in both tails than the normal distribution, then the probability of misclassification becomes higher giving unreliable result. This study proposed multivariate exponential power distribution a family of elliptically contoured model as underlining model for discrimination and classification. The distribution has a shape parameter which regulate the tail of the symmetric distribution to mitigate the problem of both lighter and heavier tails data, this generalizes the normal distribution and thus will definitely gives a lower misclassification error in discrimination and classification. The resulting discriminant model was compared with fisher linear discriminant function when applying to real data.

keywords: Classification, discrimination, allocation strength, multivariate elliptical contoured distribution, Poultry feeds data data.
AMS Subject Classification: Primary 62F03; Secondary 62P10.

1 Introduction

The main purpose of discriminant analysis is to assign an unknown subject to one of K classes on the basis of multivariate observation $x = (x_1, ..., x_p)^T$, where p is the number

*Corresponding authors: akinolosunde@gmail.com

of features. For simplicity of notation y_i are defined to be integers ranging from 1 to K. We assume that there are n_k observations in class k with

$$x_{k,1}, ..., x_{k,n_k} \overset{i.i.d.}{\sim} N_p(\mu_k, \Sigma_k), \qquad k = 1, ..., K$$

where μ_k and Σ_k are the corresponding mean vector and covariance matrix of the $p-$ dimensional multivariate normal distribution. The total number of observations is $n = n_1 + .. + n_k$. Let π_k denote the prior probability of observing a class k member with $\pi_1 + ... + \pi_K = 1$.
Under the normal distribution assumption, we assign a new subject $\mathbf{x}$ to class k, which minimizes the following discriminant score

$$D_k(x) = (x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + ln\,|\Sigma_k| - 2ln\pi_k, \tag{1}$$

that is we assign $\mathbf{x}$ to $\widehat{k} = argmin_k D_k(x)$. This is the so-called quadratic discriminant analysis (QDA) since the boundaries that separate the disjoint regions belonging to each class are quadratic (Algamal, 2017). The first term on the right-hand side of equation (1) is known as the squared Mahalanobis distance between x and μ_k. When the covariance matrices are all the same, i.e., $\Sigma_k = \Sigma$ for all k, the discriminant score can be simplified as

$$d_k(x) = (x - \mu_k)^T \Sigma^{-1}(x - \mu_k) - 2ln\pi_k \tag{2}$$

This is referred to as linear discriminant analysis (LDA) [3]. LDA assigns a new subject to $\widehat{k} = argmin_k d_k(x)$ which uses linear boundaries. The mean vectors μ_k and covariance matrices Σ_k when not known are estimated by their maximum-likelihood estimates,
$\widehat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{k,i}$, $\widehat{\Sigma}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{k,i} - \widehat{\mu}_k)(x_{k,i} - \widehat{\mu}_k)^T$, $\widehat{\Sigma} = \frac{1}{n}\sum_{k=1}^{K} n_k \widehat{\Sigma}_k$.
The prior probabilities are usually estimated by the fraction of each class in the pooled training sample, i.e., $\widehat{\pi}_k = n_k/n$. The sample version rule for QDA is $\ell(x) = argmin_k \widehat{D}_k(x)$, where

$$\widehat{D}_k(x) = (x - \mu_k)^T \Sigma_k^{-1}(x - \mu_k) + ln\,|\Sigma_k| - 2ln\pi_k,$$

Similarly, the sample version rule for LDA is

$$\ell = argmin_k \widehat{d}_k(x),$$

where

$$\widehat{d}_k(x) = (x - \mu_k)^T \Sigma^{-1}(x - \mu_k) - 2ln\pi_k$$

Also for two populations, logistic discrimination was suggested by Day and Kerridge (1967) with the restriction that estimation of the discriminator was based on samples from the mixture of the populations. This method was extended (Andrews, 1972) to more than two populations and to more usual plan of sampling from each distribution separately, using Aitchison and Silvey (1958) method of constrained maximum likelihood estimation. Logistic discriminators can be used in a simple linear form and when the likelihood ratios of the populations are linear in the observations, they are optimal irrespective of the actual likelihoods. They are thus optimal for a much wider class of distributions than standard linear discriminators. The method of logistic discrimination can be extended to the case where the likelihood ratios are quadratic in the observations. Algamal (2017) provided model for classification of gene expression based on adaptive penalized logistic regression with the aim of identifying relevant genes and provided high classification accuracy of autism data. Algamal et al. (2018), proposed a new Bayesian Lasso method which employs a skewed Laplace distribution for the errors and a scaled mixture of uniform distribution for the regression parameters, he further used Bayesian MCMC for estimation. The classification rules are known to be sensitive to departures from basic model assumptions and normal and logistic are known to be fixed at the tail region. Given multivariate data with heavier or lighter tails than normal and logistic distribution, the resulting probability of misclassification will be higher in values and thus we obtain classification rule not suitable or reliable for future classification of any new entrant from the same population density. Since it is well known that, multivariate normal assumption are not realistic in many applications, hence there is need to study a generalized class of the family of this elliptical density in which multivariate normal is a special case.

2 Multivariate Exponential Power distribution

The family of elliptical density under consideration is the exponential power distribution with the pdf

$$f(x;\mu,\sigma,p)=\frac{1}{\sigma_p 2^{1+1/2p}\Gamma(1+\frac{1}{2p})}\exp\left\{-\frac{|y-\mu|^{2p}}{2\sigma_p^{2p}}\right\} \tag{3}$$

where $-\infty < x < \infty$, $-\infty < \mu < \infty$, $p > 0$ and $\sigma_p > 0$. (3) is called exponential power distribution with shape parameter p which regulates the tail region. The multivariate extension is

$$f(x;\mu,\Sigma,p)=\frac{n\Gamma(\frac{n}{2})}{\pi^{\frac{n}{2}}\sqrt{|\Sigma|}\Gamma\left(1+\frac{n}{2p}\right)2^{1+\frac{n}{2p}}}\exp\left\{-\frac{1}{2}\left[(y-\mu)^T\Sigma^{-1}(x-\mu)\right]^p\right\} \tag{4}$$

where the mean and variance are $E(Y)=\mu$, $var(Y)=\frac{2^{\frac{1}{p}}\left(\frac{n+2}{2p}\right)}{n\Gamma\left(\frac{n}{2p}\right)}\Sigma$ and p determines the kurtosis (Gómez et al., 1998). Thus, the correlation structure can be obtained directly from Σ in the usual way. However, when $p = 1$, we have a multivariate normal distribution; when $p = 1/2$, it becomes multivariate Laplace (double exponential) distribution; and when $p \to \infty$, a multivariate uniform distribution. Hence, when $p < 1$, the distribution has heavier tails than the multivariate normal distribution and this property can be useful in providing robustness against outliers. Parameters of the exponential power distribution were estimated using the method of maximum likelihood see. The resulting equations were not in close form, therefore, we developed code in R environment to estimate the Agro (1995) parameters of any given data. Similar code were developed for univariate case by Mineo (2007). Real data applications of multivariate exponential power distribution can be seen in Gómez et al. (1998) and Olosunde (2013) just to mention few.

3 Discrimination and Classification Under Multivariate Exponential Power Distribution

Extending (1) to (3) we can obtain that the mahalanobis distance $D_k(y)$ between y and μ_k, that will allocate y to π_k as

$$\begin{aligned} &\ln\left[\pi_k f_k(y|\mu_k, \Sigma_k, p_k)\right] \\ &= \ln\pi_k - \frac{1}{2}\left[(y-\mu_k)^T\Sigma^{-1}(y-\mu_k)\right]^{p_k} - \frac{1}{2}\ln|\Sigma_k| - \ln\Gamma\left(1+\frac{n}{2p_k}\right) - \frac{n}{2p_k}\ln 2 \\ &= max\left(\ln\left[\pi_i f_i(y|\mu_i, \Sigma_i, p_i)\right]\right) \quad \forall \quad i = 1,2,...,g \;\; groups. \end{aligned} \tag{5}$$

Which implies that we allocate y to group k if $D_k(y) =$ largest of $\{D_1(y), D_2(y), ..., D_g(y)\}$ $\forall \;\; i = 1,2,...,g$.

Comparing two groups (k,i) such that $k \neq i$ we can evaluate $D_k(y) - D_i(y) \geq 0$ as

$$\ln\left(\frac{\pi_k}{\pi_i}\right) - \frac{1}{2}\ln\left|\frac{\Sigma_k}{\Sigma_i}\right| - \frac{1}{2}\left(\left[(y-\mu_k)^T\Sigma_k^{-1}(y-\mu_k)\right]^{p_k} - \left[(y-\mu_i)^T\Sigma_i^{-1}(y-\mu_i)\right]^{p_i}\right) - \ln\left[\frac{\Gamma\left(1+\frac{n}{2p_k}\right)}{\Gamma\left(1+\frac{n}{2p_i}\right)}\right] - \frac{n}{2}\left[\frac{1}{p_k} - \frac{1}{p_i}\right]\ln 2 \tag{6}$$

Hence the overall mahalanobis discrimination and classification function $D_{ki}(y)$ in favour of group k for the multivariate exponential power distribution two group case is given as

$$-\frac{1}{2}\left\{\left[(y-\mu_k)^T\Sigma_k^{-1}(y-\mu_k)\right]^{p_k} - \left[(y-\mu_i)^T\Sigma_i^{-1}(y-\mu_i)\right]^{p_i}\right\} \geq H \tag{7}$$

where $H = \ln\left[\frac{\Gamma\left(1+\frac{n}{2p_k}\right)}{\Gamma\left(1+\frac{n}{2p_i}\right)}\right] + \frac{n}{2}\left[\frac{1}{p_k} - \frac{1}{p_i}\right]\ln 2 + \frac{1}{2}\ln\left|\frac{\Sigma_k}{\Sigma_i}\right| - \ln\left(\frac{\pi_k}{\pi_i}\right)$. This implies that y is allocated to group k if $D_{ki}(y) \geq H$ and to group i if $D_{ki}(y) < H$.
This further implies that if the covariance matrix $\Sigma_k = \Sigma_i$ then we have

$$H = \ln\left[\frac{\Gamma\left(1+\frac{n}{2p_k}\right)}{\Gamma\left(1+\frac{n}{2p_i}\right)}\right] + \frac{n}{2}\left[\frac{1}{p_k} - \frac{1}{p_i}\right]\ln 2 - \ln\left(\frac{\pi_k}{\pi_i}\right) \tag{8}$$

for the pooled variance $\Sigma = \frac{(n_k-1)\Sigma_k^2 + (n_i-1)\Sigma_i^2}{n_k+n_i-2}$.
Which subsequently indicate that when $\Sigma_k = \Sigma_i$ and $p_k = p_i$ then $D_{ki}(y)$ becomes

$$-\frac{1}{2}\left\{\left[(y-\mu_k)^T\Sigma_k^{-1}(y-\mu_k)\right]^{p_k} - \left[(y-\mu_i)^T\Sigma_i^{-1}(y-\mu_i)\right]^{p_i}\right\} \geq H \tag{9}$$

where $H = ln\left(\frac{\pi_i}{\pi_k}\right)$ and $H = 0$ when the prior probability are the same. So in (9) we have allocate y to group k if $D_{ki}(y) \geq H$ and to group i if $D_{ki}(y) < H$.

Table 1a: EPD with $\beta = 2$ versus normal distribution with unit variance in each case

pdf	n=10000	n=100	n=50
Normal	$P_1 = 0.35$, $P_2 = 0.23$	$P_1 = 0.39$, $P_2 = 0.27$	$P_1 = 0.36$, $P_2 = 0.36$
EPD	$P_1 = 0.38$, $P_2 = 0.29$	$P_1 = 0.41$, $P_2 = 0.25$	$P_1 = 0.33$, $P_2 = 0.25$

Table 1b: EPD with $\beta = 6$ versus normal distribution with unit variance in each case

pdf	n=10000	n=100	n=50
Normal	$P_1 = 0.28$, $P_2 = 0.27$	$P_1 = 0.29$, $P_2 = 0.27$	$P_1 = 0.43$, $P_2 = 0.37$
EPD	$P_1 = 0.25$, $P_2 = 0.23$	$P_1 = 0.10$, $P_2 = 0.05$	$P_1 = 0.05$, $P_2 = 0.01$

Note that μ_k is the mean vectors for each of the group $i = 1, 2, ..., g$ which can be obtained as the vector of the sample means from the n-variates in each group. The shape parameter, p cannot be obtained in closed form, so by scaling each of variables to make it dimensionless; and using the normalp r environment code '$paramp[(y/s)_i]$' on set of combined scaled observations, we can obtain the shape parameter for each group. Estimation of the parameters for any given data can be done using '*normalp*' software. Also some code were developed to extend it to multivariate cases.

4 Simulation Study

To simulate form MPED, we wrote a code using the rmvpowerexp from package MNM (Nordhausen and Oja, 2011) in Team et al. (2013) environment. This program generate data by making use of the stochastic representation for MPED proposed by Gómez et al. (1998). Two samples were generated from EPD (μ_i, Σ, β) $i = 1, 2$. It is common assumption in literature that the sample were from the normal population $N(\mu_i, \Sigma)$. Table 1a and 1b contains summary values of the numerical work, in particular misclassification probabilities P_i $i = 1, 2$ for each case of normal and exponential power distribution (EPD). Note that the same sample mean and covariance matrix were selected for both normal and EPD but the latter was simulated with additional shape parameter $\beta =$ 3 and 6. The resulting misclassification probabilities clearly showed that when β was chosen closed to normal distribution ($\beta = 1$) the P_i's were almost the same except when sample was large normal distribution have misclassification error lower than the EPD. But for large β, EPD performed better than normal distribution in both large and small sample. Clearly, we conclude that for optimum classification result, EPD should be preferred as discrimination and classification model when the tail of the data in consideration is thicker or thinner than the normal distribution. P_1 and P_2 respectively represent the probabilities of misclassification from population 1 and 2. Also, n is the number of sample taken in each trial.

Table 2: Summary of Data in Hand and Henley (1997).

Variables	Mean	Variance	Shape (β)
X_1	3.5182	0.5196364	2.8089
X_2	2.1	0.476	
Y_1	2.5	0.8514286	6.088
Y_2	3.24	1.548286	

5 Applications

Example 1 :The resulting discriminant model in this study was first applied to the data illustration in Hand and Henley (1997); values of urinary androsterone and etiocholanolone in healthy heterosexual and homosexual males in mg/24 hours. The data has been used in several studies we applied he EPD discriminant model (7) to this data owing to the fact that the data exhibited some level of departure from normal distribution at the tail, its shape parameter for the two groups is not unity ($\beta \neq 1$), for us to assume exact normal distribution. The summary of the preliminary analysis of the data is given in table 2, carried out under the assumption that the underlining distribution is exponential power distribution. Also table shows the confusion matrix resulting from classification using equal cost and equal prior probabilities which yields the estimated error rates based on table 1.

Table 3: (Confusion Matrix):

		Predicted	Membership
		Correct	Incorrect
Actual	Group X	11	0
Membership	Group Y	15	0
	Total	26	0

In Table 3, the apparent misclassification error rate using exponential power distribution is 0.00% compared to 3.85% obtained when normal distribution was assumed in Hand and Henley (1997); Mineo (2007). This show there was a departure from normality at the tail region of the data, hence the apparent misclassification error rate was higher when compared with the case of the exponential power distribution.

Example 2: Another application was based on the laboratory experiment carried out courtesy National Centre for Genetic Resources and Biotechnology (NACGRAB) Ibadan. These data were concerned about the discrimination and classification of two cowpea varieties (Ife Brown and Sampea 12) stored under different conditions (Ambient, Short term, medium term) and later planted at different seasons (May when the rain just started, August when the rain was at its peak, November during harmattan, February when there was no rain at all) of the year. These with the aim of obtaining increase in yields to meet up with the increase in demand by the consumers. The classification was based on the mean yields of the each variety, the sampea is expected to give a higher yields compare to local Ife brown. The hypothesis and the results from testing are as follows:

1. Test $H_0 : \Sigma_{sampea} = \Sigma_{ife\ brown}$. Using $-2 \ln \Lambda \sim \chi^2_{12}$. We reject H_0 and conclude that the variance-covariance matrices are the same for the two groups.
2. Next, we test H_0: μ_{sampea}=$\mu_{ife\ brown}$, given $\Sigma_{sampea} = \Sigma_{ife\ brown}$. The test show that the $\mu's$ are different. The summary of the data is presented in table 4.

Table 4: Summary of the data of cowpea varieties.

Variables	Mean	Variance	Shape
x_1	18.17	197.6667	10.4576861
x_2	27.0825	99.85316	
x_3	26.2525	120.7239	
y_1	29.415	48.3763	5.0540495
y_2	26.1675	75.30483	
y_3	24.085	58.51557	

Table 5: Confusion Matrix.

		Predicted	Membership
		Correct	Incorrect
Actual	Ife Brown (X)	4	0
Membership	Sampea 12 (Y)	3	1
	Total	7	1

Courtesy: National Centre for Genetic Resources and Biotechnology (NACGRAB) Ibadan.

We observed that the number of seed germination in Ife Brown can be discriminated from that of Sampea 12 with apparent misclassification error rate of 12.5% for both distributions. The result was the same when using Fisher's linear discriminant.

6 Conclusion

In the two examples given, it was observed that the exponential power performed better in the first example but same performance was noticed in the example 2. This affirmed our assertion, that exponential power serve as good substitute to the normal distribution, the worst scenario is to have the shape parameter resulting to normal when $\beta = 1$ or double exponential when $\beta = 1/2$. The discrimination model of exponential power distribution generalized the discrimination model of normal and double exponential distribution and also performed well than that of normal especially, when the data in question has either shorter or longer tail than the usual normal distribution. The code written in r environment extended to multivariate exponential power discrimination model use in classifying the training data can be used to classify future observations, although it was developed for bivariate and trivariate cases. The code can be extended to nth variates.

Acknowledgement

We would like to appreciate the useful suggestions of the anonymous reviewers. Those suggestions have improved the paper.

References

Agro, G. (1995). Maximum likelihood estimation for the exponential power function parameters. *Communications in Statistics-Simulation and Computation*, 24(2):523–536.

Aitchison, J. and Silvey, S. (1958). Maximum-likelihood estimation of parameters subject to restraints. *The annals of mathematical Statistics*, pages 813–828.

Algamal, Z. (2017). Classification of gene expression autism data based on adaptive penalized logistic regression. *Electronic Journal of Applied Statistical Analysis*, 10(2):561–571.

Algamal, Z. Y., Alhamzawi, R., and Ali, H. T. M. (2018). Gene selection for microarray gene expression classification using bayesian lasso quantile regression. 97:145–152.

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, pages 125–136.

Day, N. E. and Kerridge, D. F. (1967). A general maximum likelihood discriminant. *Biometrics*, pages 313–323.

Gómez, E., Gomez-Viilegas, M., and Marin, J. (1998). A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 27(3):589–600.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. 160:523–541.

Mineo, A. M. (2007). On the estimation of the structure parameter of a normal distribution of order p. *Statistica*, 63(1):109–122.

Nordhausen, K. and Oja, H. (2011). Multivariate l1 methods: the package mnm. *Journal of Statistical Software*, 43(5):1–28.

Olosunde, A. (2013). On exponential power distribution and poultry feeds data: a case study. *Journal of The Iranian Statistical Society*, 12(2):253–270.

Team, R. C. et al. (2013). R: A language and environment for statistical computing.