ISSN: 1984-5529

Científica

Jaboticabal

v.46, n.4, p.344–352, 2018

http://dx.doi.org/10.15361/1984-5529.2018v46n4p344-352

Discrimination of populations under covariance matrix heterogeneity and non-normal random vectors in genetic diversity studies

Discriminação de populações na presença de heterogeneidade de matrizes de covariâncias e vetores aleatórios não normais em estudos de diversidade genética

Vitor Prado de CARVALHO[1,2]; Ithalo Coelho de SOUSA[1]; Moysés NASCIMENTO[1]; Ana Carolina Campana NASCIMENTO[1]; Cosme Damião CRUZ[3]

[1] Autor correspondência: Departamento de Estatística, Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2] Departamento de Estatística, Universidade Federal do Espírito Santo, Vitória, ES, Brasil
[3] Departamento de Biologia Geral, Universidade Federal de Viçosa, MG, Brasil.

Recebido em: 22-02-2018; Aceito em: 25-07-2018

Abstract

Genetic diversity analysis has guided the choice of appropriate parents in breeding programs. Multivariate statistical methods such as discriminant analysis are used to obtain the necessary results in these studies. However, to obtain reliable results, one must meet assumptions such as covariance matrix heterogeneity and multivariate normality of the observation vector. Artificial Neural Network (ANN), Support Vector Machine (SVM), Decision Tree (DT) and its refinements do not have these assumptions and may be used in the choice of appropriate parents. This study evaluates the robustness of the Fisher’s discriminant function under covariance matrix heterogeneity and multivariate non-normal random vectors. The results were compared with those obtained from Quadratic Discriminant Analysis (QDA), ANN, SVM and DT. Scenarios characterized by heterogeneous covariance matrices and multivariate non-normal random vectors were simulated. Considering the apparent error rate (APER), the SVM method (APER-Normal = 0.07; APER-Poisson = 0.13) and quadratic discriminant method (APER-Normal = 0.09; APER-Poisson = 0.09) presented better results for scenarios simulated with covariance matrix heteroscedasticity. For scenarios with multivariate normality and covariance matrix homoscedasticity, the SVM (APER = 0.15) and ANN (APER = 0.06) presented best results. For situations in which the data had multivariate Poisson distribution and covariance matrix homogeneity, the SVM (APER = 0.15), Fisher’s discriminant function (APER = 0.19) and ANN (APER = 0.19) presented better performances. Finally, DT refinements (Bagging, Random Forest and Boosting) presented APER values less than 0.25 and are shown to be alternatives.

Additional Keywords: quadratic discriminant function; multivariate analysis, simulation.

Resumo

Análises de diversidade genética têm orientado a escolha de genitores apropriados em programas de melhoramento. Métodos de Estatística Multivariada, como por exemplo, as análises discriminantes são utilizadas para obtenção dos resultados necessários nesses estudos. Entretanto, a obtenção de resultados confiáveis está associada ao atendimento de pressupostos, como por exemplo a heterogeneidade de matrizes de covariância e normalidade multivariada do vetor de observações. Redes Neurais Artificiais (RNA), Máquina de Vetor Suporte (MVS), Árvores de Decisão (AD) e seus refinamentos, não possuem pressupostos e podem ser utilizadas para esse fim. O objetivo desse trabalho foi avaliar a robustez da função discriminante de Fisher na presença de matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados. Os resultados foram comparados com aqueles provenientes da função discriminante quadrática (FDQ), RNA, MVS e AD. Foram simulados cenários caracterizados por matrizes de covariâncias heterogêneas e vetores aleatórios não normais multivariados. Considerando a Taxa de Erro Aparente média (TEA) a MVS (TEA–Normal=0,07; TEA–Poisson=0,13) e FDQ (TEA–Normal=0,09; TEA–Poisson=0,09) apresentaram melhores resultados para os simulados considerando heterocedasticidade de matrizes de covariância. Para os cenários com normalidade multivariada e homocedasticidade de matrizes de covariâncias a MVS (TEA=0,15) e RNA (TEA=0,06) apresentaram os melhores resultados. Já para as situações em que os dados apresentaram distribuição Poisson multivariada e homogeneidade de matrizes de covariância, a MVS (TEA=0,15), Função Discriminante de Fisher (TEA=0,19) e RNA (TEA=0,19) apresentaram melhores performances. Finalmente, os refinamentos da AD (*Bagging*, *Random Forest* e *Boosting*) apresentaram valores TEA inferiores a 0,25 e se apresentam como alternativas.

Palavras-chave adicionais: análise multivariada; função discriminante quadrática; simulação.

Introduction

Genetic diversity analysis has guided the choice of appropriate parents in breeding programs, leading to hybrids with higher heterosis and populations with greater variability. Moreover, such analysis allows quantification of the existing variability, facilitating the management of germplasm banks (Sant'Anna, 2015).

Currently, the literature presents several methods for quantification and evaluation of genetic diversity in population studies. In general, methods from Multivariate Statistics have been an effective alternative in these studies; for example, methods based on cluster analysis (Santos et al., 2017b; Rodrigues et al., 2017) and discriminant analysis (Santos et al., 2017a). However, to obtain reliable results, one must meet the assumptions of the method to be applied, which must be chosen according to the characteristics of the information set available and the objectives established by the researcher.

Clustering methods usually do not require data structure assumptions and generally involve hierarchical procedures, optimization analysis or graphical dispersion. The choice of the most appropriate method is not a simple task and, even within a class of analysis such as hierarchical clustering, one must choose carefully because different techniques use different statistical and biological concepts. Notwithstanding, in some cases, the choice of one of these methods is simply guided by measures such as the cophenetic correlation coefficient (Sokal & Rohlf, 1962).

In genetic diversity studies using discriminant analysis, various analysis options must also be considered so as to take advantage of the potentiality of each. Among these techniques, the linear discriminant functions of Anderson and Fisher deserve special mention. However, the use of linear discriminant functions (Fisher, 1936) requires homogeneous covariance matrices between populations (Ferreira, 2008) and, in some cases, multivariate normal distribution of the random vector. If the equality hypothesis is rejected, quadratic functions are recommended (Mingoti, 2007). Furthermore, if normality is not achieved, strategies such as data transformation are suggested. Despite these indications, the literature does not present studies evaluating the robustness of the technique regarding the breakdown of such assumptions. Moreover, studies without in-depth analysis make use of linear functions without specifying any criterion that has led to such a choice.

The literature also presents other methods that can be used to discriminate populations and do not require assumptions on covariance matrix heterogeneity and multivariate normality. Such methods, namely Artificial Neural Networks, Support Vector Machine, Decision Tree and its refinements are based on computer intelligence and statistical learning. They have been used in improvements to solve several problems. Nascimento et al. (2013) used artificial neural networks to classify alfalfa genotypes for phenotypic adaptability and stability. Sant'Anna et al. (2015) showed the superiority of neural networks in relation to discriminant analysis in genetic classification studies considering populations coming from backcrossings. Silva et al. (2017) carried out genomic prediction for orange rust resistance in arabica coffee by means of artificial neural networks. Despite the existence of these methods, there are no studies evaluating whether heterogeneous covariance matrices and normality affect the efficiencies thereof.

Considering the above, this study evaluates, through data simulation, the robustness of the linear discriminant function regarding the lack of homogeneity of covariance matrices and the presence of multivariate non-normal random vectors. These evaluations aim to guide researchers as to the appropriate method to be used in genetic diversity studies. The results will be compared with those from other methods commonly used for this purpose, such as quadratic discriminant analysis, Artificial Neural Networks, Support Vector Machine and Decision Tree.

Material and methods

Simulated Dataset

To evaluate the robustness of the discriminant function regarding covariance matrix heterogeneity and non-normal random vectors, datasets with different covariance structures and multivariate probability distribution were simulated. Method performance was evaluated considering two populations (A and B) and sample size n = 100. The number of variables (p) was established as p = 5, and the covariance matrix structures ($\boldsymbol{\Sigma}$) were defined as follows.

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0.9 \\ \vdots & \ddots & \vdots \\ 0.9 & \cdots & 1 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0.1 \\ \vdots & \ddots & \vdots \\ 0.1 & \cdots & 1 \end{bmatrix}; \qquad (1)$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0.9 \\ \vdots & \ddots & \vdots \\ 0.9 & \cdots & 1 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0.5 \\ \vdots & \ddots & \vdots \\ 0.5 & \cdots & 1 \end{bmatrix}; \qquad (2)$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0.9 \\ \vdots & \ddots & \vdots \\ 0.9 & \cdots & 1 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0.9 \\ \vdots & \ddots & \vdots \\ 0.9 & \cdots & 1 \end{bmatrix}; \qquad (3)$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 1 & \cdots & 0.1 \\ \vdots & \ddots & \vdots \\ 0.1 & \cdots & 1 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{bmatrix} 1 & \cdots & 0.1 \\ \vdots & \ddots & \vdots \\ 0.1 & \cdots & 1 \end{bmatrix} \qquad (4)$$

For the multivariate normal distribution, the parametric values of the mean vectors were considered as $\boldsymbol{\mu}_A = [0 \ \cdots \ 0]^T$ and $\boldsymbol{\mu}_B = [i \ \cdots \ i]^T$, where i = 0.5, 1, 2 and 3. On the other hand, for the parametric vectors of the multivariate Poisson distribution, we considered $\boldsymbol{\lambda}_A = [1 \ \cdots \ 1]^T$ and $\boldsymbol{\lambda}_B = [j \ \cdots \ j]^T$, where the mean values differ for j = 0.5, 2, 3 and 4. The difference between the means vectors aims to represent different levels of discrimination considering 0.5, 1, 2 and 3 standard deviations. The combination of different structures of covariance and probability distribution results in 32 distinct scenarios (Table1).

Table 1 - Scenarios evaluated for the robustness of the linear discriminant function regarding the lack of homogeneity of covariance matrices and the presence of random vectors not multivariate normal.

Scenario	Multivariated distribution	Structure of the covariance	Differences between mean vectors in standard deviations
1	Normal	(1)	0.5
2		(1)	1.0
3		(1)	2.0
4		(1)	3.0
5		(2)	0.5
6		(2)	1.0
7		(2)	2.0
8		(2)	3.0
9		(3)	0.5
10		(3)	1.0
11		(3)	2.0
12		(3)	3.0
13		(4)	0.5
14		(4)	1.0
15		(4)	2.0
16		(4)	3.0
17	Poisson	(1)	0.5
18		(1)	1.0
19		(1)	2.0
20		(1)	3.0
21		(2)	0.5
22		(2)	1.0
23		(2)	2.0
24		(2)	3.0
25		(3)	0.5
26		(3)	1.0
27		(3)	2.0
28		(3)	3.0
29		(4)	0.5
30		(4)	1.0
31		(4)	2.0
32		(4)	3.0

Specifically, simulated scenarios considering covariance matrices defined in (1) and (2) represent situations in which the covariance matrices are heterogeneous. Simulated scenarios considering structures (3) and (4) represent situations in which the covariance matrices are homogeneous. The combination of these covariance structures with the different multivariate probability distributions and degrees of discrimination, as a function of the parametric vectors, compose the whole set of simulated scenarios. To guarantee heterogeneity among covariance matrices, the hypothesis $H_0\colon \Sigma_A = \Sigma_B$ was evaluated by means of Box's M statistics (Morrison, 1976), derived from the likelihood ratio test. The whole simulation process was repeated 25 times.

Linear and Quadratic Discriminant Analysis

Consider the case where there are p > 1 variables measured in each sampling element of each population and coming from p-variate normal distributions. Assume that for population A, vector $\boldsymbol{X}$ is normal with mean vector $\boldsymbol{\mu}_A$ and covariance matrix $\boldsymbol{\Sigma}_A$; and for population B, $\boldsymbol{X}$ is normal with mean vector $\boldsymbol{\mu}_B$ and covariance matrix $\boldsymbol{\Sigma}_B$. For a fixed observation vector $\boldsymbol{x}^T = [x_1 x_2 \dots x_p]$, the ratio between the probability density functions of the two populations, in terms of neperian logarithm, will be:

$$-2\ln\big(\lambda(x)\big) = -2ln\left\{\frac{(2\pi)^{\frac{p}{2}}\left(|\Sigma_A|^{\frac{1}{2}}\right)^{-1}}{(2\pi)^{\frac{p}{2}}\left(|\Sigma_B|^{\frac{1}{2}}\right)^{-1}}\left[\frac{exp\left\{-\frac{1}{2}(x-\mu_A)'\Sigma_A{}'(x-\mu_A)\right\}}{exp\left\{-\frac{1}{2}(x-\mu_B)'\Sigma_B{}'(x-\mu_B)\right\}}\right]\right\} \tag{5}$$

Thus, a sampling element with observation vector $\boldsymbol{x}$ will be classified as belonging to population 1 when $-2\ln\big(\lambda(\boldsymbol{x})\big)$ is greater than zero; and to population 2, when less than zero. If $-2\ln\big(\lambda(\boldsymbol{x})\big)$ = 0, the sampling element can be classified in any of the two populations. If $\boldsymbol{\Sigma}_A \neq \boldsymbol{\Sigma}_B$, this function is called a quadratic discriminant function (Mingoti, 2007).

When matrices Σ_A and Σ_B are homogeneous, the function becomes equivalent to "Fisher's linear discriminant function" (Fisher, 1936), which is expressed as:

$$f(x) = (\mu_A - \mu_B)'\Sigma^{-1}x - \frac{1}{2}(\mu_A - \mu_B)'\Sigma^{-1}(\mu_A + \mu_B) \quad (6)$$

wherein Σ^{-1} is the inverse of the covariance matrix of the two populations, estimated by:

$$S_p = \left[\frac{n_A - 1}{(n_A - 1) + (n_B - 1)}\right]S_1 + \left[\frac{n_B - 1}{(n_A - 1) + (n_B - 1)}\right]S_2 \quad (7)$$

.

In this case, an individual will be classified as belonging to population 1 if *f(x)* is greater than zero, and to population 2 if *f(x)* is less than zero. If *f(x)* = 0, the sampling element can be classified into any of the two populations.

Artificial Neural Networks

An Artificial Neural Network (ANN) is formed by the combination of several artificial neurons, which are a logical structure that try to simulate the behavior and functions of a biological neuron. ANNs are usually structured into three layers: input, intermediate and output layers (Figure 1).

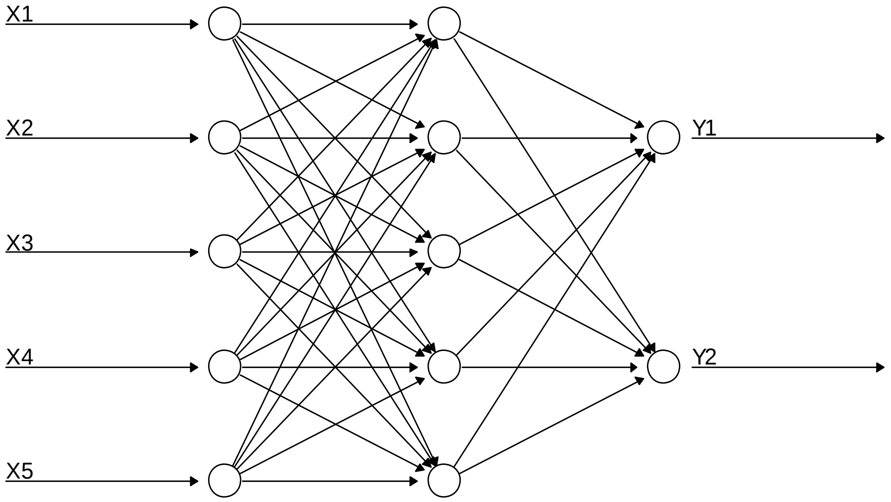

Figure 1 - Representation of existing layers in a model of Artificial Neural Networks (variables $\mathrm{X_i}$ in which i=1, 2, 3, 4, 5) and two outputs ($\mathrm{Y_1}$ and $\mathrm{Y_2}$).

In this study, we used a feed-forward network, which assumes that the output of any layer is not affected in that layer, i.e., there is no feedback. The input layer is fed by phenotypic values (simulated according to scenarios 1 to 24, already described), which will be used for diversity analysis. In other words, a dataset consisting of n = 100 individuals (accessions, genotypes, etc.) measured in p = 5 characters. The input layer is connected to the hidden layer, composed of T neurons (T ranging from 1 to 40) that are connected to the output layer, composed of a single neuron. These connections are directed by means of estimated weights, which measure the influence of the predictor variables on the response variable. In addition to weights, the bias (b_t), also known as intercept, is estimated (Glória et al., 2016).

Mathematically, in the ith intermediate layer, the jth neuron is formed by the weight vector, w_{ij}^T, added to the intercept. The resulting linear combination is then transformed by means of an activation function f(.), generating the output of said neuron, $a_i^T = \mathrm{f}\left(\sum_{j=1}^{5} w_{ij}^T x_{ij}\right) + b_t$. The activation function can be linear or nonlinear. For complex problems, however, Bishop (2006) state that nonlinear activation functions provide better results when compared to linear functions. In the last layer, considering, without loss of generality, an ANN with only one hidden layer, all outputs from the neurons that make up the intermediate layers are inputs in a new linear combination, which is again transformed by an activation function g(.). Thus, the ANN output, y_i, depends on a new weight vector and a scalar bias:

$$y_i = \mathrm{g}\left[\sum_{j=1}^{T} w_{2t}\,\mathrm{f}\left(\sum_{j=1}^{5} w_{1j}^T x_{ij} + b_t\right) + b\right] \quad (8)$$

In this study, we considered ANNs with one and two hidden layers, with Sigmoidal Tangent Hyperbolic and Logarithmic Sigmoid activation function. The number of neurons (T) ranged from 1 to 40, and the number of iterations was set at 100000.

Support Vector Machine

Support Vector Machine (SVM) (Lorena & Carvalho, 2007) is based on the theory of statistical learning, which aims to establish mathematical conditions that allow us to choose a classifier with good performance for the available dataset. The main idea of SVM is to create a separation hyperplane as decision surface, so that the separation between its positive and negative examples is maximal (Campbell, 2000; Lorena & Carvalho, 2007).

Mathematically, a hyperplane can be written as follows:

$$w \cdot x + b = 0, \tag{9}$$

wherein w is the adjustable weight vector; and b, as in ANN, is the bias term. From this equation, we divide the observation space X into two regions: $w \cdot x_i + b > 0$ and $w \cdot x_i + b < 0$ so that $g(x) = sgn(w \cdot x + b)$. Thus, the classification will be +1 if $f(x) > 0$; and -1 if $f(x) < 0$ (Lorena & Carvalho, 2007) (Figure 2).

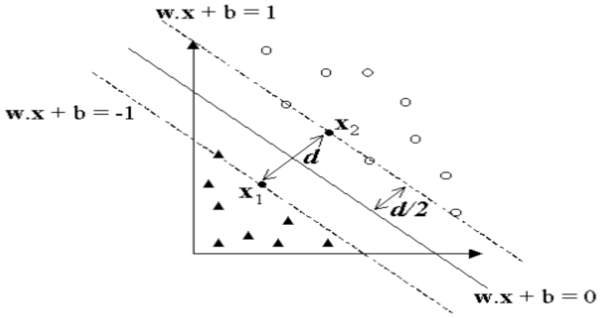

Figure 2 - Illustration of a data set linearly separable and the distance d between hyperplanes $\mathbf{w} \cdot \mathbf{x}_1 + b = -1$ and $\mathbf{w} \cdot \mathbf{x}_2 + b = +1$. Fonte: Lorena & Carvalho, 2007.

To deal with situations where the data cannot be satisfactorily divided by a linear hyperplane, the training sets are mapped to a new space with a greater dimensionality, obtaining a linear solution for the problem. However, an appropriate mapping function ϕ must be chosen. This can be done by simply applying the mapping function to each standard in equation $\mathrm{f}(x) = w \cdot \phi(x) + b$. Through this procedure, the information needed for mapping the function is defined by the internal product $\phi(x_i) \cdot \phi(x_j)$. This product is obtained by introducing the Kernels concept (Lorena & Carvalho, 2007), which comprise functions that receive two points from the input space, x_i and x_j, and compute the scalar product $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ in the feature space (Haykin, 1999). In practice, the most commonly used kernels are the Polynomials, RBF (Radial-Basis Function) and the Sigmoidal.

Another way to deal with data nonlinearity is to implement a smoothing constant (C), which determines the stiffness of the separation margin (Lorena & Carvalho, 2007). In this study, we used the Kernel RBF function (Shanthini et al., 2017), given by:

$$K(x_i, x_j) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) \tag{10}$$

Such function depends only on one parameter, sigma (σ), usually defined as the standard deviation of a Gaussian distribution. The search space of the sigma parameter was defined as $0.001 \leq \frac{1}{2\sigma^2} \leq 2$; and the smoothing constant (C) had a search space of 10^i, with i = 0, 1, 2 and 3.

Decision Tree and its Refinements

To construct the decision tree, regions R_1, R_2,..., R_M are aimed, which minimize the Gini index, given by (James et al., 2013):

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{11}$$

wherein $\hat{p}_{mk}$ represents the proportion of observations in the mth region belonging to the kth class. The Gini index decreases according to the growth of the tree, which occurs through recursive binary division. To avoid the overfitting, cost-complexity pruning is indicated (Hastie et al., 2009). In addition, no region should contain more than 5 individuals.

The construction of a single tree is not an indicated strategy, since this approach presents models with great variability. To circumvent the problem, the literature suggests the use of bootstrap aggregation (Bagging). Bagging consists of obtaining B samples with replacement (size equal to N) of the dataset, thus fitting B models [$\hat{f}^1(x)$, $\hat{f}^2(x)$,..., $\hat{f}^B(x)$] to be used as individual classifiers. A new individual will be ranked in the most common class among the predictions of individual B classifiers. Another approach that aims to increase accuracy in the classification of individuals is Random Forest (RF). In this procedure, in the same way as in bagging, B samples are drawn from the population. Notwithstanding, the number of predictor variables used in each partition is less than the total number of available variables (m<p). According to James et al. (2013), RF results in a process of decorrelating the generated trees, further improving the accuracy of the predictions.

Another refinement used to improve the decision tree result is Boosting. Unlike Bagging, which creates multiple independent trees, Boosting creates trees sequentially, using information from previous trees. The Boosting classifier has the form $\boldsymbol{H}(\boldsymbol{x}) = \sum_t \alpha_t \boldsymbol{h_t}(\boldsymbol{x})$, which seeks to minimize a loss function L by optimizing the scalar α_t (importance assigned to $\boldsymbol{h_t}(\boldsymbol{x})$) and the individual classifier $\boldsymbol{h_t}(\boldsymbol{x})$ (individual decision tree) at each iteration t (Freund & Schapire, 1999). Individual classifiers $\boldsymbol{h_t}(\boldsymbol{x})$ have low classificatory power, but when used together with $\boldsymbol{H}(\boldsymbol{x})$, they present good results (Appel et al., 2013).

Comparison of Methods

To access the predictive ability of the evaluated methods, the Apparent Error Rate (APER) was calculated according to the following expression:

$$APER(\%) = \frac{1}{N}\sum_{j=1}^{k} m_j, \qquad (12)$$

wherein m_j is the number of observations taken from a population, which by means of the evaluated technique was classified into another population. N is the total number of observations evaluated; and k is the number of populations considered. These values were obtained through a procedure considering 80% of the dataset for adjustment/training, and the remaining 20% for validation. The final APER value is given by the mean value obtained in the 25 replicates.

Computational Aspects

The entire process of population simulation and adjustment/training of the models was conducted using R software (R Development Core Team, 2017). Samples of the multivariate normal and Poisson distributions were obtained using the functions mvrnorm and gen.PoisBinOrd of the MASS (Venables & Ripley, 2002) and PoisBinOrd (Inan & Demitras, 2016) packages, respectively. The linear and quadratic discriminant functions, Neural Networks, Support Vector Machine, Tree, Bagging/Random Forest and Boosting were adjusted through the functions lda, qda, neuralnet, ksvm, tree, randomForest and gbm of the MASS, kernlab (Karatzoglou et al., 2004), neuralnet (Fritsch & Guenther, 2016) and tree (Ripley, 2016) packages, respectively.

Results and discussion

The homogeneity of variances was tested by means of the likelihood ratio test for all simulated scenarios and replicates from the multivariate normal distribution. As results, the hypothesis of homogeneity (P ≤ 0.01) was rejected in cases where matrices were simulated as heterogeneous, the opposite occurring in the other cases (P > 0.01). Since the Box-M test assumes the multivariate normality of the observation vector, the results presented refer to scenarios in which the multivariate normal distribution was used to generate data (Table 2).

Table 2 - Average P-values associated with a Box'M test.

Scenario	Multivariated distribution	Structure of covariance	p-valor
1	Normal	(1)	0.0001
2		(1)	0.0001
3		(1)	0.0001
4		(1)	0.0001
5		(2)	0.0001
6		(2)	0.0001
7		(2)	0.0001
8		(2)	0.0001
9		(3)	0.7782
10		(3)	0.8551
11		(3)	0.9186
12		(3)	0.8667
13		(4)	0.7132
14		(4)	0.2574
15		(4)	0.5422
16		(4)	0.4995

The APER values obtained for all the evaluated methods ranged from 0.00 to 0.47 (Table 3). In general, for scenarios that consider the heterogeneity of covariance matrices and multivariate normality (Scenarios 1, 2, 3, 4, 5, 6, 7 and 8), the Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM) presented lower mean values of APER (APER - QDA = 0.07 and APER - SVM = 0.09) compared to the other approaches evaluated. Specifically considering QDA, such scenarios represent ideal situations for the application of this method, since it requires that the covariance matrices are heterogeneous and that the random vector has a multivariate normal distribution (Ferreira, 2008).

Unlike QDA, SVMs do not have an assumption regarding data distribution and covariance matrices. SVMs have as principle to partition the points into predefined classes to maximize both the margin given by the support vectors and the separation hyperplane (Bridges et al., 2011). The other approaches, which also have no assumptions about the distribution of random vectors, presented mean APER values ranging from 0.12 (ANN with one hidden layer and Random Forest) to 0.23 (Fisher's linear discriminant function - FLD). The lowest performance of FLD for the classification, in terms of mean APER, can be attributed to the construction of the method, which requires that the covariance matrices are homogeneous (Ferreira, 2008). These results are expected since the QDA and other methods evaluated lead to nonlinear decision thresholds (Zhang et., 2000). Thus, the greater the heterogeneity of covariance matrices, the more nonlinear the classification thresholds and the better the methods that model this structure.

On the other hand, for scenarios that encompass situations in which the dataset was simulated considering multivariate normality of the random vector and homoscedasticity of the covariance matrices (Scenarios 9, 10, 11, 12, 13, 14, 15 and 16), the ANN with two hidden layers presented a lower mean APER (0.06) compared to the other methods evaluated, which ranged from 0.15 (SVM) to 0.24 (Decision Trees - Tree). FLD presented a reasonable performance (APER = 0.13), obtaining APER values close to (APER - Neural Networks with one hidden layer = 0.12, APER - SVM = 0.10, APER - Random Forest = 0.13, and APER - Boosting = 0.12) or higher (APER - Tree = = 0.18, APER - Pruning = 0.17, and APER - Bagging = = 0.17) than those observed by the other methods under study.

Table 3 - Apparent error rate (APER) obtained for 36 different scenarios by means of different techniques of classification.

MD	Scen.	FLD	QDA	SVM	ANN		DT	Prunning	Bagg	RanFor	Boost
					1 layer	2 layers					
Normal	1	0.39	0.07	0.16	0.15	0.12	0.26	0.25	0.16	0.14	0.37
	2	0.32	0.06	0.09	0.13	0.18	0.21	0.21	0.13	0.12	0.25
	3	0.14	0.03	0.04	0.08	0.31	0.12	0.11	0.07	0.06	0.10
	4	0.06	0.01	0.01	0.04	0.26	0.05	0.05	0.03	0.02	0.03
	5	0.39	0.15	0.21	0.22	0.11	0.38	0.37	0.26	0.25	0.40
	6	0.31	0.12	0.14	0.19	0.15	0.30	0.29	0.22	0.20	0.28
	7	0.15	0.06	0.05	0.11	0.26	0.16	0.15	0.11	0.10	0.13
	8	0.06	0.03	0.02	0.06	0.15	0.06	0.06	0.04	0.04	0.05
	9	0.41	0.42	0.35	0.35	0.05	0.47	0.44	0.43	0.42	0.39
	10	0.32	0.32	0.26	0.29	0.09	0.40	0.35	0.35	0.34	0.30
	11	0.16	0.16	0.13	0.16	0.13	0.20	0.18	0.18	0.17	0.14
	12	0.06	0.06	0.05	0.08	0.04	0.08	0.08	0.06	0.06	0.06
	13	0.34	0.33	0.26	0.29	0.03	0.39	0.38	0.36	0.34	0.32
	14	0.19	0.19	0.14	0.17	0.05	0.25	0.25	0.20	0.19	0.19
	15	0.03	0.03	0.03	0.04	0.07	0.10	0.11	0.06	0.04	0.04
	16	0.00	0.00	0.00	0.01	0.01	0.06	0.06	0.02	0.01	0.01
Poisson	17	0.40	0.21	0.16	0.20	0.19	0.29	0.29	0.22	0.21	0.35
	18	0.24	0.10	0.09	0.14	0.12	0.25	0.26	0.17	0.15	0.29
	19	0.15	0.04	0.04	0.08	0.06	0.15	0.15	0.10	0.08	0.15
	20	0.07	0.02	0.01	0.04	0.04	0.11	0.11	0.06	0.05	0.07
	21	0.41	0.36	0.21	0.32	0.29	0.37	0.37	0.36	0.33	0.36
	22	0.24	0.17	0.14	0.21	0.19	0.31	0.31	0.25	0.24	0.32
	23	0.14	0.09	0.05	0.13	0.15	0.20	0.20	0.15	0.13	0.21
	24	0.07	0.05	0.02	0.07	0.10	0.14	0.14	0.10	0.09	0.12
	25	0.44	0.38	0.35	0.34	0.32	0.39	0.38	0.41	0.40	0.38
	26	0.27	0.37	0.26	0.34	0.31	0.37	0.37	0.38	0.37	0.34
	27	0.15	0.23	0.13	0.21	0.20	0.26	0.25	0.28	0.27	0.24
	28	0.08	0.15	0.05	0.15	0.13	0.17	0.17	0.17	0.16	0.15
	29	0.36	0.32	0.26	0.27	0.24	0.33	0.32	0.32	0.31	0.30
	30	0.18	0.26	0.14	0.21	0.19	0.31	0.32	0.26	0.25	0.25
	31	0.05	0.10	0.03	0.09	0.07	0.16	0.17	0.11	0.10	0.10
	32	0.01	0.03	0.00	0.04	0.04	0.09	0.09	0.05	0.04	0.04

MD = multivariated distribution; Scen = scenario; FLD = Fisher's linear discriminant function; QDA = Quadratic Discriminant Analysis; ANN = Artificial Neural Network; SVM = Support Vector Machine; DT = Decision Tree; Prunning – Pruned tree of decision; Bagg = Bagging; RanFor = Random Forest; Boost = Boosting.

For the simulated scenarios considering multivariate Poisson distribution and heteroscedasticity of covariance matrices (Scenarios 17, 18, 19, 20, 21, 22, 23 and 24), SVM presented a lower mean APER (0.09). Considering the results for the case of homoscedasticity of covariance matrices and multivariate Poisson random vectors, FLD presented the same classificatory performance observed when data were simulated considering multivariate normal distribution (APER - FLD = 0.19). This equality of results corroborates with the literature, which shows that FLD derivation is based only on the assumption of homogeneity among covariance matrices, i.e., it does not require multivariate normality of the random vector (Ferreira, 2008). Again, SVM presented better results compared to the other methods (FLD, QDA, ANNs, Tree, Pruning, Bagging, Random Forest and Booting), with mean APER value equal to 0.15. The other methods (FLD, QDA, ANNs, Tree, Pruning, Bagging, Random Forest and Booting) presented mean APER values ranging from 0.19 (ANN) to 0.26 (Tree and Pruning).

The results indicated that among the techniques for classifying individuals, SVM showed better results in all situations evaluated. SVM presents the separation hyperplane between classes so as to maximize the margin defined as the distance between the classifier and the nearest sample (denoted by support vector). Thus, the method usually presents a good performance in test sets. Another interesting result is the good performance obtained by QDA in situations where the data presented heterogeneity of covariance matrices. This result is also supported by the literature, which states that such method is indicated in these cases (Ferreira, 2008). Another method that was highlighted is ANN. This method presented results close to those obtained by the best techniques in all scenarios evaluated. Finally, Decision Tree refinements (Bagging, Random Forest and Boosting) presented satisfactory performance (APER ranging from 0.12 to 0.25) and may be interesting alternatives in studies in which the assumptions are not met.

Nevertheless, it should be emphasized that the choice of the method depends on several characteristics of the set of observations under study. In this work, we evaluated situations involving different distributions of multivariate probability and the presence or not of homoscedasticity of variances. Other characteristics should be considered when choosing the method, such as type of variables, presence of outliers, ability to deal with missing values, and ability to extract linear patterns from data. For all these situations, except for the ability to extract linear patterns from data, the literature indicates the use of Decision Tree and its refinements (Hastie et al., 2009). To extract nonlinear patterns, in turn, ANN and SVM are indicated (Hastie et al., 2009).

Genetic diversity studies present data with different types of characters. In Vargas et al. (2015), for instance, the authors evaluated the genetic diversity of heirloom tomato accessions from the collection of the Department of Phytotechnology of the UFRRJ, through quantitative (e.g. fruit length and width) and qualitative descriptors (e.g. fruit shape and presence of pedicel knee). As regards data structuring, the literature usually does not present genetic diversity studies in which the authors are aware of the heteroscedasticity of covariance matrices. Nogueira et al. (2008) and Santos et al. (2017a) are examples of that, since they did not consider this hypothesis. In view of the above results, assessing the assumption of homogeneity of covariance matrices is important given the possible decrease in the performance of the applied technique.

Conclusions

For situations in which the data present heteroscedasticity of covariance matrices, the Support Vector Machine (SVM) and Quadratic Discriminant Analysis (QDA) presented better results regarding the Apparent Error Rate (APER).

For situations in which the data show multivariate normality and homoscedasticity of covariance matrices, the Support Vector Machine and Artificial Neural Networks presented better results regarding the Apparent Error Rate (APER).

- For situations where the data showed multivariate Poisson distribution and homogeneity of covariance matrices, the SVM, Fisher's Discriminant Function and Artificial Neural Networks showed lower APER values.

- Methods such as Decision Tree refinements (Bagging, Random Forest and Boosting) presented APER values lower than 0.25, being considered alternative techniques.

References

Bishop CM (2006) Pattern Recognition and Machine Learning (Information Science and Statistics). Springer. 738p.

Bridges M, Heron EA, O'Dushlaine C, Segurado R, ISC, Morris D, Corvin A, Gill M, Pinto C (2011) Genetic classification of populations using supervised learning. PloS one 6(5):e14802. doi: 10.1371/journal.pone.0014802

Campbell C (2000) An introduction to kernel methods. In Howlett RJ, Jain LC (ed) Radial Basis Function Networks: Design and Applications, Springer Verlag. p.155–192.

Ferreira DF (2008) Estatística multivariada. Editora UFLA. 662p.

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of human genetics, 7(2):179-188. doi:10.1111/j.1469-1809.1936.tb02137.x

Freund Y, Schapire RE (1999) A short introduction to Boosting. Journal of Japanese Society for Artificial Intelligence 14(5):771-780.

Fritsch S, Guenther F (2016) Neuralnet: Training of Neural Networks. R package version 1.33. Disponível em <https://CRAN.R-project.org/package=neuralnet> (Acesso em 26 dez 2017).

Glória LS, Cruz CD, Vieira RAM, de Resende MDV, Lopes PS, de Siqueira OHD, Silva FF (2016) Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. Livestock Science 191:91-96. doi: 10.1016/j.livsci.2016.07.015

Hastie T, Tibishirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. 745p.

Haykin S (1999) Neural Networks: A Comprehensive Foundation. Prentice Hall. 842p.

Inan G, Demirtas H (2016) PoisBinOrd: Data Generation with Poisson, Binary and Ordinal Components, R package version 1.2. Disponível em <https://CRAN.R-project.org/package=PoisBinOrd> (Acesso em 26 dez 2017).

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to Statistical Learning with Applications in R. Springer. 426p.

Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9):1-20. doi: 10.18637/jss.v011.i09

Lorena AC, de Carvalho AC (2007) Uma introdução às support vector machines. Revista de Informática Teórica e Aplicada 14(2):43-67. doi: 10.22456/2175-2745.5690

Mingoti SA (2007) Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Editora UFMG. 297p.

Morrison DF (1976) Multivariate Statistical Methods. McGraw-Hill. 415 p.

Nascimento M, Peternelli LA, Cruz CD, Nascimento ACC, Ferreira RDP, Bhering LL, Salgado CC (2013) Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. Crop Breeding and Applied Biotechnology 13(2):152-156.

Nogueira APO, Sediyama T, Cruz CD, Reis MS, Pereira DG, Jangarelli M (2008) Novas características para diferenciação de cultivares de soja pela análise discriminante. Ciência Rural 38(9):2427-2433. doi: 10.1590/S0103-84782008005000025

R Core Team. R (2017) A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ripley B (2016) tree: Classification and Regression Trees. R package version 1.0-37. Disponível em: <https://CRAN.R-project.org/package=tree> (Acesso em 26 dez 2017).

Rodrigues DL, Viana AP, Vieira HD, Santos EA, de Lima FH, Santos CL (2017) Contribuição de variáveis de produção e de semente para a divergência genética em maracujazeiro-azedo sob diferentes disponibilidades de nutrientes. Pesquisa Agropecuária Brasileira 52(8):607-614. doi: 10.1590/s0100-204x2017000800006

Sant'Anna IC, Tomaz RS, Silva GN, Nascimento M, Bhering LL, Cruz CD (2015) Superiority of artificial neural networks for a genetic classification procedure. Genetics and Molecular Research 14(3):9898-9906. doi: 10.4238/2015.August.19.24

Santos BWC, Ferreira FM, de Souza VF, Clement CR, Rocha RB (2017a) Análise discriminante das características físicas e químicas de frutos de pupunha (Bactris gasipaes Kunth) do alto Rio Madeira, Rondônia, Brasil. Científica 45(2):154-161. doi: 10.15361/1984-5529.2017v45n2p154-161

Santos MDS, Stancatte RS, Ferreira TC, Dorighello DV, Pazianotto RAA, de Melo IS, May A, Ramos, NP (2017b) Resistance to water deficit during the formation of sugarcane seedlings mediated by interaction with Bacillus sp. Científica 45(4):414-421. doi: 10.15361/1984-5529.2017v45n4p414-421

Shanthini D, Shanthi M, Bhuvaneswari MC (2017) A Comparative Study of SVM Kernel Functions Based on Polynomial Coefficients and V-Transform Coefficients. International Journal of Engineering and Computer Science (IJECS) 6(3):20765-20769. doi:10.18535/ijecs/v6i3.65

Silva GN, Nascimento M, Sant'anna IC, Cruz CD, Caixeta ET, Carneiro PC, Rosado R, Pestana K, Oliveira MS (2017) Artificial neural networks compared with Bayesian generalized linear regression for leaf rust resistance prediction in Arabica coffee. Pesquisa Agropecuária Brasileira 52(3):186-193. doi: 10.1590/s0100-204x2017000300009

Sokal RR, Rohlf FJ (1962) The comparison of dendrograms by objective methods. Taxon 11(2):33-40. doi: 10.2307/1217208

Vargas TO, Alves EP, Abboud ACS, Leal MAA, Carmo MGF (2015) Diversidade genética em acessos de tomateiro heirloom. Horticultura Brasileira 33(2):174-180. doi: 10.1590/S0102-053620150000200007

Venables WN, Ripley BD (2002) Modern Applied Statistics with S 4th Edition. Springer. 498p.

Zhang MQ (2000) Discriminant analysis and its application in DNA sequence motif recognition. Briefings in Bioinformatics 1(4):331-342. doi: 10.1093/bib/1.4.331