PLOS ONE

RESEARCH ARTICLE

LPDA: A new classification method based on linear programming

María J. Nueda*, Carmen Gandía, Mariola D. Molina

Mathematics Department, University of Alicante, Alicante, Spain

* mj.nueda@ua.es

OPEN ACCESS

Citation: Nueda MJ, Gandía C, Molina MD (2022) LPDA: A new classification method based on linear programming. PLoS ONE 17(7): e0270403. https://doi.org/10.1371/journal.pone.0270403

Editor: Eugene Demidenko, Dartmouth College Geisel School of Medicine, UNITED STATES

Received: January 12, 2022

Accepted: June 9, 2022

Published: July 7, 2022

Data Availability Statement: Regarding the datasets we have used in the paper to illustrate the method: - Palmdates data is within lpda package available in https://github.com/mjnueda/lpda that is flexible for new versions of packages. We would like also to include lpda package in https://cran.r-project.org/ as soon as possible. It was not the intention to publish the data alone, but in the package environment where the manual explains how to use it. - Default data is available in ISLR package as it is mentioned in the paper. - Cervical cancer data is an available dataset in GEO, as it is mentioned in the paper.

Abstract

The search of separation hyperplanes is an efficient way to find rules with classification purposes. This paper presents an alternative mathematical programming formulation to existing methods to find a discriminant hyperplane. The hyperplane *H* is found by minimizing the sum of all the distances to the area assigned to the group each individual belongs to. It results in a convex optimization problem for which we find an equivalent linear programming problem. We demonstrate that *H* exists when the centroids of the two groups are not equal. The method is effective dealing with low and high dimensional data where reduction of the dimension is proposed to avoid overfitting problems. We show the performance of this approach with different data sets and comparisons with other classifications methods. The method is called *LPDA* and it is implemented in a R package available in https://github.com/mjnueda/lpda.

Introduction

One of the main goals in many recent data analysis projects is the classification of samples or individuals into predefined groups, according to the characteristics available. Several approaches have been proposed to deal with this problem. Statistical methods, usually are based in the evaluation of a scoring function that needs distributional assumptions as Fisher Linear Discriminant Analysis (*LDA*) [1, 2] or Logistic Regression [3]. The high number of variables and the diverse type of distributional assumptions are challenging topics that researchers try to solve with non distributional approaches. Mathematical programming is a natural way of dealing with the classification problem regardless of distributional assumptions. In this sense, linear programming based methods look for a linear function that separates the classes avoiding parameters estimations. Support Vector Machine (*SVM*) [4, 5] is the most popular classification method based in hyperplanes, that can be extended to nonlinear separating functions, as polynomial or radial kernel. In [6] we find a discussion of mathematical optimization techniques proposed for SVM and [7] reviews and compares supervised classification methods related to optimization. These publications and other as [8] demonstrate the exinting interest of addressing the classification problem through mathematical programming. We can also mention the Machine Learning approach, where we find alternative methods as Decision Trees, CART or Random Forest, [9, 10] and Neural Networks approach [11]. This approach

Funding: This research has been partially supported by Generalitat Valenciana, Grant GV/2017/177.

Competing interests: The authors have declared that no competing interests exist.

tries to find a stepwise rule that combines the best ranking variables in a training set also ignoring distributional assumptions.

All these approaches could be considered complementary rather than competitive. Machine learning approaches are useful in classification when dealing with high dimensional data sets, but for interpreting variable influence it is preferable Logistic Regression or *LDA*. *SVM* is an effective method in different situations. When dealing with small dimension the flexibility of the separating function can help to find a perfect separation, however with high dimensional data over-fitted problems can emerge and, as mentioned in [12], there is not need of additional flexibility that give this models, being the linear function a good option.

In this work we propose an efficient alternative to the available classification methods in R without distributional assumptions. We formulate an optimization problem to find a discriminating hyperplane between two data sets that can be useful to classify new individuals. The method has been extended also to the case with more than two groups making paiwise comparisons. In addition, to avoid overfiting problems due to noisy data or high dimensional data sets, we consider Principal Components Analysis (PCA) to focus on the main sources of variation avoiding the noise. The method has been implemented in a R package named **lpda** available in github.

The paper is structured as follows. In the following section, the optimization problem is proposed on the basis of the general two-group classification approach and the PCA solution is presented. Then, it is described the evaluation strategy of the new technique: data and other approaches against which it is intended to be compared. In the Results section this evaluation is showed and finally, conclusions are presented in the last section.

Linear programming discriminant analysis method

The purpose of this section is to describe the problem we want to solve and to build the linear problem which will allow us to find the solutions. First, we present the approach for the case of two data sets and subsequently extend it to the case with more than two sets. Finally, we propose a strategy to avoid overfitting in data sets with more variables than individuals.

Model definition for two data sets

Let $\mathbf{X} = \{x_1, ..., x_{n_1}\}$ and $\mathbf{Y} = \{y_1, ..., y_{n_2}\}$ two sets whose elements are in $\mathbb{R}^p$, and $m^t = (m_1, m_2, ..., m_{n_1})$ and $w^t = (w_1, w_2, ..., w_{n_2})$ the vectors whose components are the weights of the elements of $\mathbf{X}$ and $\mathbf{Y}$ respectively, positive and such that $\sum_{i=1}^{n_1} m_i = \sum_{j=1}^{n_2} w_j = 1$.

Weights can be assigned depending on the importance of the individual in the sample. This could be of interest if the individuals are collectives; for example: cities or universities; that can be weighted by their size. If all the individuals are equally important, weights must be $m_i = 1/n_1 \forall i$ and $w_j = 1/n_2 \forall j$.

Definition 1. *A hyperplane H in $\mathbb{R}^p$ is an (p-1)-affine set and can be represented as $H = \{x \in \mathbb{R}^p | a^t x = b\}$, where $b \in \mathbb{R}$ and $a \in \mathbb{R}^p$, $a \neq 0_p$, and they are unique up to a common non-zero multiple.*

Initially, we look for a hyperplane H that strictly separates $\mathbf{X}$ from $\mathbf{Y}$ (Fig 1). If such hyperplane does not exist, we focus on a hyperplane that minimizes a measure of the deviation of this goal, called *separation error*.

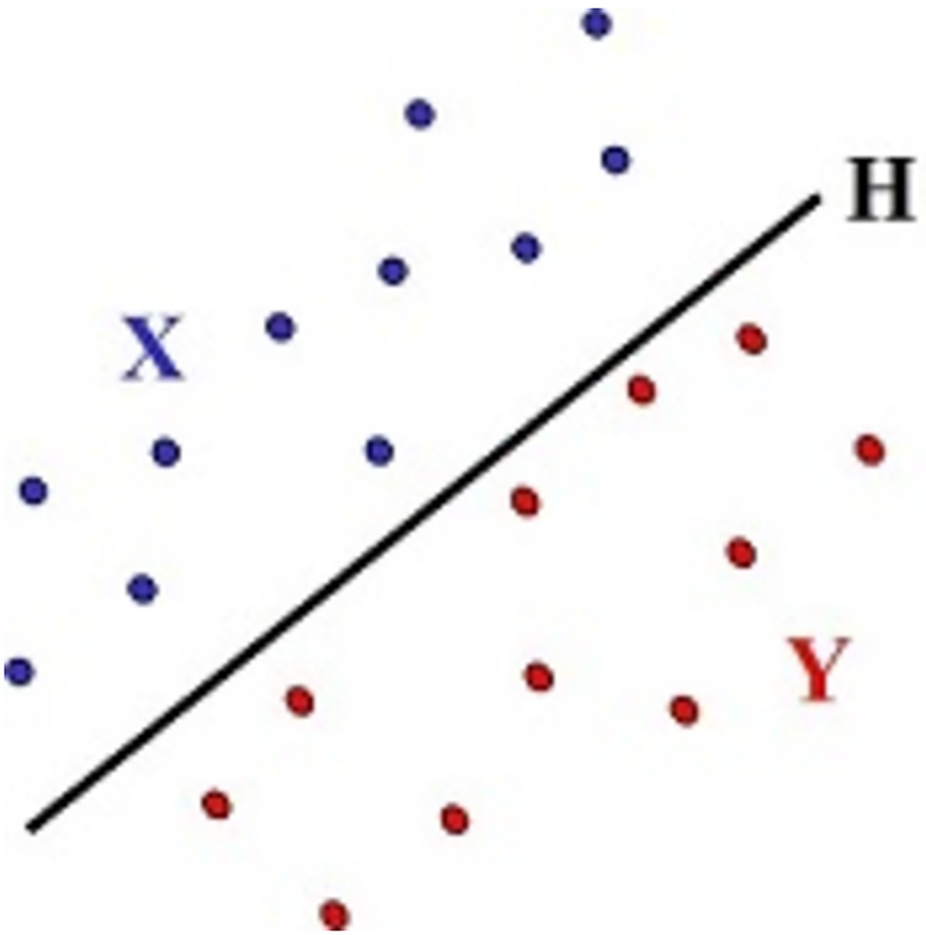

Fig 1. The objective is finding *H* that separates X from Y.

https://doi.org/10.1371/journal.pone.0270403.g001

Proposition 1. *X and Y are strictly separable if and only if there exists a hyperplane* $H = \{x \in \mathbb{R}^n | a^t x = b\}$ *such that the following system*:

$$\sigma = \begin{cases} a^t x_i \geq b + 1, & i = 1, \ldots, n_1 \\ a^t y_j \leq b - 1, & j = 1, \ldots, n_2 \end{cases} \tag{1}$$

is consistent. Such hyperplane is named separator hyperplane.

Proof.

If **X** and **Y** can be strictly separated, there exists $c \in \mathbb{R}^p$, $c \neq 0_p$ and $d \in \mathbb{R}$ such that:

$$\begin{aligned} c^t x_i > d, \quad i = 1, \ldots, n_1 \\ c^t y_j < d, \quad j = 1, \ldots, n_2 \end{aligned} \tag{2}$$

Let ε_i and δ_j the slacks of each constraint in (2):

$$\begin{aligned} \varepsilon_i = c^t x_i - d > 0, \quad i = 1, \ldots, n_1 \\ \delta_j = d - c^t y_j > 0, \quad j = 1, \ldots, n_2 \end{aligned}$$

and $\eta := \min\{\varepsilon_i, \delta_j\}$. We can define

$$(a^t, b) = \eta^{-1}(c^t, d), \tag{3}$$

where $a \in \mathbb{R}^p$, $a \neq 0_p$ and $b \in \mathbb{R}$.

Multiplying both sides of (3) by $(x_i^t, -1)^t$ we have:

$$a^t x_i - b = \eta^{-1}(c^t x_i - d) = \eta^{-1}\varepsilon_i \geq 1, i = 1, \ldots, n_1$$

Similarly, multiplying (3) by $(y_j^t, -1)^t$ we have:

$$a^t y_j - b = \eta^{-1}(c^t y_j - d) = -\eta^{-1}\delta_j \leq -1, j = 1, \ldots, n_2$$

Therefore, the pair (a, b) is a solution of the system (1).

Conversely, if the system (1) has a solution (a, b), then $H = \{x \in \mathbb{R}^p | a^t x = b\}$ verifies

$$\begin{cases} a^t x_i \geq b + 1 > b, & i = 1, \ldots, n_1 \\ a^t y_j \leq b - 1 < b, & j = 1, \ldots, n_2 \end{cases}$$

Moreover, $a \neq 0_p$ (otherwise, $b + 1 \leq 0 \leq -1$). Hence, H is a hyperplane separating strictly **X** and **Y**.

Such hyperplane will be referred to as a *separator hyperplane*. This proposition leads us to locate sets **X** and **Y** as it is showed in Fig 2, regarding the hyperplanes

$$H = \{x \in \mathbb{R}^p | a^t x = b\},$$

$$H_{+1} = \{x \in \mathbb{R}^p | a^t x = b + 1\}$$

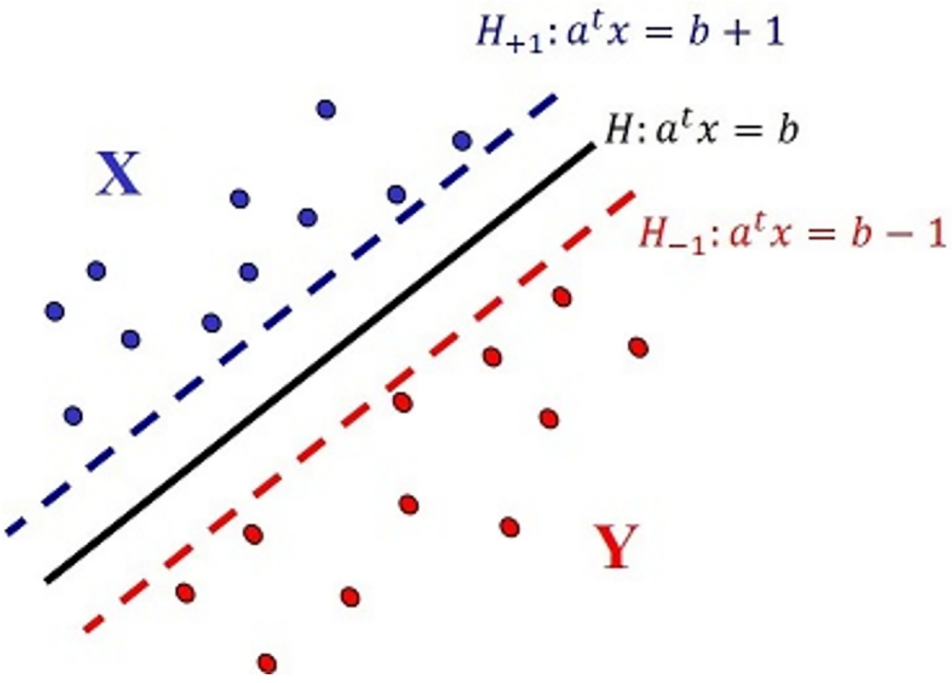

Fig 2. Situation of X and Y related to H, H_{+1} and H_{-1}.

https://doi.org/10.1371/journal.pone.0270403.g002

and

$$H_{-1} = \{x \in \mathbb{R}^p | a^t x = b - 1\}.$$

On the other hand, if (1) is an inconsistient system, there exists some $i \in \{1, \ldots, n_1\}$ for which $b + 1 - a^t x_i > 0$ or $j \in \{1, \ldots, n_2\}$ for which $a^t y_j - b + 1 > 0$. Therefore, we can take the following values as error measure of each element:

$$\max\{b + 1 - a^t x_i, 0\}, \quad i = 1, \ldots, n_1$$
$$\max\{a^t y_j - b + 1, 0\}, \quad j = 1, \ldots, n_2.$$

Adding all these measures weighted by m_i and w_j, respectively, we obtain the function f, called *separation error function*:

$$f(a, b) := \sum_{i=1}^{n_1} m_i \max\{b + 1 - a^t x_i, 0\} + \sum_{j=1}^{n_2} w_j \max\{-b + 1 + a^t y_j, 0\}. \tag{4}$$

The separation error function is a non-negative, convex and non-differenciable function and the aim is to solve the problem

$$(P_1) \min f(a, b).$$

Proposition 2. *σ is consistent if and only if $v(P_1) = 0$. In such case, any optimal solution of P_1 defines a separator hyperplane.*

Proof.

If **X** and **Y** can be strictly separated, there exists $(\bar{a}, \bar{b})$ solution of σ. So, $f(\bar{a}, \bar{b}) = 0$ and $(\bar{a}, \bar{b})$ is an optimal solution of (P_1).

Conversely, if $(\bar{a}, \bar{b})$ is an optimal solution of (P_1), each term in f will be equal to zero and by Proposition 1, **X** and **Y** can be separated strictly. Moreover, $\bar{a} \neq 0_p$ because $f(0_p, \bar{b}) = 2$ and it can not be an optimal solution of (P_1).

So, **X** and **Y** can be strictly separated if and only if $v(P_1) = 0$. But, in any case, the objective is translated in finding the solution $(\bar{a}, \bar{b})$ to the problem (P_1). We approach this task through a linear problem equivalent to (P_1), whose optimal solutions will define our discriminant hyperplane, that is, the hyperplane that minimizes the separation error function.

Proposition 3. (P_1) *is equivalent to the problem*

$$(P) \quad Min \quad \sum_{i=1}^{n_1} m_i u_i + \sum_{j=1}^{n_2} w_j v_j$$

s.t.

$$u_i \geq -a^t x_i + b + 1, \qquad i = 1, \ldots, n_1$$
$$u_i \geq 0 \qquad i = 1, \ldots, n_1$$
$$v_j \geq a^t y_j - b + 1, \qquad j = 1, \ldots, n_2$$
$$v_j \geq 0 \qquad j = 1, \ldots, n_2$$

where the objective is finding (a, b), that define the hyperplane, with the support of the variables u_i and v_j that identify potential errors to be minimized.

Proof.

Since each of the functions to maximize in each operand in (4) is convex, we have

$$\max\{b+1-a^t x_i, 0\} = \min\{u_i \in \mathbb{R}_+ \mid u_i \geq b+1-a^t x_i\}$$

and

$$\max\{a^t y_j - b + 1, 0\} = \min\{v_j \in \mathbb{R}_+ \mid v_j \geq a^t y_j - b + 1\}$$

It allows us to reformulate our initial problem as the equivalent problem (P) in the following sense [13]:

1. If $(\bar{a}, \bar{b})$ is an optimal solution of (P_1), then taking

$$\bar{u}_i = \max\{-\bar{a}^t x_i + \bar{b} + 1, 0\},\ i = 1, \ldots, n_1$$

and

$$\bar{v}_j = \max\{-\bar{b} + 1 + \bar{a}^t y_j, 0\},\ j = 1, \ldots, n_2,$$

then $(\bar{a}, \bar{b}, \bar{u}, \bar{v})$ is an optimal solution of (P).

2. If $(\bar{a}, \bar{b}, \bar{u}, \bar{v})$ is an optimal solution of (P), then $\bar{u}_i = \max\{\bar{b} + 1 - \bar{a}^t x_i, 0\}$, for $i = 1, \ldots, n_1$ and $\bar{v}_j = \max\{-\bar{b} + 1 + \bar{a}^t y_j, 0\},\ j = 1, \ldots, n_2$, and $(\bar{a}, \bar{b})$ is an optimal solution of (P_1).

(P) is a solvable problem and every optimal solution will provide a discriminant hyperplane that minimizes the separation error as long as $a \neq 0_p$. We can state that if $v(P) = 0$, this situation is guaranteed by Proposition 2 but, if $v(P) > 0$, we need to add a very weak condition on the data sets (in the sense that it will usually be verified), what we will prove in the following proposition. Let us remember that in a linear problem, a necessary and sufficient condition for $\bar{x}$ to be an optimal solution is that the objective vector can be written as a non-negative linear combination of the active constraints on $\bar{x}$.

Proposition 4 *If* $v(P) > 0$ *and* $\bar{x} \neq \bar{y}$, *with* $\bar{x} = \sum_{i=1}^{n_1} m_i x_i$ *and* $\bar{y} = \sum_{j=1}^{n_2} w_j y_j$ *there exist an optimal solution of* (P) *that gives a discriminant hyperplane.*

Proof.

Let us suppose that $(\bar{a}, \bar{b}, \bar{u}, \bar{v})$ is an optimal solution of (P) with $\bar{a} = 0_p$. Then,

$$u_i \geq b + 1, \text{ for all } i = 1, \ldots, n_1 \tag{5}$$

$$v_j \geq 1 - b, \text{ for all } j = 1, \ldots, n_2 \tag{6}$$

and all of them will be active (otherwise, the solution is no an optimal solution). So we can consider $\bar{u} = (\bar{b} + 1)\mathbf{1}_{n_1}$ and $\bar{v} = (1 - \bar{b})\mathbf{1}_{n_2}$, where $\mathbf{1}_{n_1}$ and $\mathbf{1}_{n_2}$ are vectors with all its elements equal to 1 in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively. Then,

$$v(P) = \sum_{i=1}^{n_1} m_i(\bar{b} + 1) + \sum_{j=1}^{n_2} w_j(1 - \bar{b}) = 2.$$

We will distinguish the different values of b in order to determine the active constraints in each case and apply the condition that characterizes the optimality.

(a). $|b| \neq 1$. Now, the unique active constraints are (5) and (6) and hence,

$$\begin{pmatrix} 0_p \\ 0 \\ m \\ w \end{pmatrix} = \sum_{i=1}^{n_1} \lambda_i \begin{pmatrix} x_i \\ -1 \\ I^i_{n_1} \\ 0_{n_2} \end{pmatrix} + \sum_{j=1}^{n_2} \mu_j \begin{pmatrix} -y_j \\ 1 \\ 0_{n_1} \\ I^j_{n_2} \end{pmatrix}, \tag{7}$$

where λ_i and μ_j belong to $\mathbb{R}_+$, for all $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$; $I^i_{n_1}$ and $I^j_{n_2}$ are the ith and jth vectors of the canonical basis in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively. Then,

$$\lambda_i = m_i, \text{for all } i = 1, 2..., n_1,$$

$$\mu_j = w_j, \text{for all } j = 1, 2..., n_2$$

whereas

$$0_p = \sum_{i=1}^{n_1} \lambda_i x_i + \sum_{j=1}^{n_2} \mu_j (-y_j) = \bar{x} - \bar{y}.$$

(b). $b = 1$. Now, in addition to (5) and (6), constraints

$$v_j \geq 0, j = 1, \ldots, n_2$$

are active too. Hence,

$$\begin{pmatrix} 0_p \\ 0 \\ m \\ w \end{pmatrix} = \sum_{i=1}^{n_1} \lambda_i \begin{pmatrix} x_i \\ -1 \\ I^i_{n_1} \\ 0_{n_2} \end{pmatrix} + \sum_{j=1}^{n_2} \mu_j \begin{pmatrix} -y_j \\ 1 \\ 0_{n_1} \\ I^j_{n_2} \end{pmatrix} + \sum_{j=1}^{n_2} \delta_j \begin{pmatrix} 0_n \\ 0 \\ 0_{n_1} \\ I^j_{n_2} \end{pmatrix}, \tag{8}$$

where λ_i, μ_j and δ_j belong to $\mathbb{R}_+$, for all $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$. Hence,

$$\lambda_i = m_i, \text{for all } i = 1, 2..., n_1,$$

$$\mu_j + \delta_j = w_j, \text{for all } j = 1, 2..., n_2,$$

whereas

$$0 = \sum_{i=1}^{n_1} \lambda_i (-1) + \sum_{j=1}^{n_2} \mu_j = -1 + \sum_{j=1}^{n_2} \mu_j,$$

which implies

$$\sum_{j=1}^{n_2} \mu_j = 1.$$

Then,

$$\sum_{j=1}^{n_2}(\mu_j+\delta_j)=1+\sum_{j=1}^{n_2}\delta_j=\sum_{j=1}^{n_2}w_j=1,$$

and, consequently,

$$\delta_j=0 \text{ and } \mu_j=w_j, \text{ for all } j=1,\cdots,n_2$$

And, as in the first case,

$$0_p=\sum_{i=1}^{n_1}\lambda_i x_i+\sum_{j=1}^{n_2}\mu_j(-y_j)=\bar{x}-\bar{y}.$$

(c). $b=-1$. Reasoning as in the case (b), we arise the same conclusion.

Model definition for more than two data sets

In the case of more than two groups, we could proceed in two different ways:

1. Obtain the discriminant hyperplanes for each set with respect to the rest.
2. Obtain the discriminant hyperplanes that separate the given sets by pairs. In this case, if we have k different sets, we would obtain $\binom{k}{2}$ equations corresponding to the discriminant hyperplanes. For each group we will consider the subgroups of equations that separate it from the rest.

   In **lpda** package we have implemented the second option.

Overfitting problem

In nowadays it is very usual being involved in projects where the number of measured variables is much higher than the number of samples. In such cases, the high dimension allows statistical methods were succesfull separating groups. However, the hyperplane can overfit the training data and as a result a bad evaluation in the data test is obtained. To avoid this problem we propose obtaining the hyperplane from Principal Components (PCs) instead of the original variables. In general, when managing large amounts of noisy but correlated data, data analysis can greatly benefit from the application of dimensionality reduction methods, such as PCA, which allows the identification of the main patterns of variability avoiding residual or non-structural variation (examples in [14, 15]). Such approaches are effective in providing global understanding of most relevant information that can help to detect the differences between the studied groups.

PCA reduces the dimension of a set of individuals measured in a p-dimensional basis, taking advantage of the relationship between the variables. The method consists of projecting the individuals on a subspace of dimension $q<p$ extracting the major information. The solution of this problem is the subspace defined by the q eigenvectors associated with the q higher eigenvalues of the variance-covariance matrix of the data. The selected number of PCs, q, is typically obtained on the basis of the percentage of the explained variability or by a cross-validation criterion. The PCA model corresponding to a data matrix $\mathbf{X}$, of dimensions $n\times p$, gives

us the following decomposition:

$$\mathbf{X} = \mathbf{1}_n \mu^t + \mathbf{T}\mathbf{P}^t + \mathbf{E} \tag{9}$$

where $\mathbf{1}_n$ is a size n column vector of ones, $\boldsymbol{\mu}^t$ is a size p row vector containing estimates of de average for each variable, scores of the individuals in each PC are collected in the matrix **T**, the loadings (eigenvectors) are given by the matrix **P** and the residuals are collected in **E**.

The aplication of *LPDA* to the scores, or **T** matrix, will provide a classification hyperplane that avoids the undesirable noise focussing in the signal of interest. For more details about the PCA model and other projection techniques see [16].

The evaluation strategy

To evaluate the performance of *LPDA* we first consider a data set with few variables to graphically inspect the behaviour of the hyperplane compared to *SVM*. Second, we consider an example of unbalanced and overlapping data between classes, with few variables but many individuals. Here the interest is to evaluate *LPDA* against other popular techniques such as *SVM*, *LDA* and Logistic Regression. Finally, we address a gene expression RNA-Seq data set, as example from the bioinformatics field, to show results with high-dimensional data. In this case, the method is compared with three classification techniques: *SVM* and two specific classification methods for RNA-Seq data. We describe the data and methods discussed below.

Data sets

Palmdates. A data set with scores of 21 palm dates including their respective Raman spectra and the concentration of five compounds covering a wide range of concentrations: fibre, glucose, fructose, sorbitol and myo-inositol. The first 11 dates are Spanish (from Elche, Alicante) with no well-defined variety and the last 10 are from other countries and varieties, mainly Arabian. The data set is available in **lpda** package including two data.frames: conc with 5 variables and spectra with 2050. In this paper we use only conc data.

Default. A simulated data set containing information on 10.000 customers of which only 333 are default. It is an example of unbalanced data. The aim here is to predict which customers will default on their credit card debt, the minority class. This data set is in ISLR package [12].

Cervical cancer. A data set quantifying the expression of 714 microRNAs measured to 29 samples of tumor and 29 nontumor cervical tissue samples. This data set is available in Gene Expression Omnibus (GEO) Datasets with access number GSE20592 [17] and we normalized with *Quantile normalizaton* method described in [18].

Classification methods

SVM is a hyperplane-based classification method, as said in the introduction. This method tries to find the hyperplane with the maximum margin that separates two classes, allowing some errors in the training set to avoid overfitting [4, 5]. Although *SVM* can also perform a non-linear classification, when dealing with so many variables there is no need of additional flexibility that will give polynomial or radial kernel models. For this reason, in next section we use linear classifiers, also called Support Vector Classifiers, for RNA-Seq example.

From the different packages avaible in **R** to apply *SVM* [19] we use the *SVM* implementation called **e1071**. The needed parameters in each application were computed with the crossvalidation proccess available in this package.

Logistic Regression considers a linear model where the response, a binary variable representing the class, is modelled with a logistic transformation. It is considered a specific case of

Generalised Linear Models that are a generalization of classical Linear Models, which can accommodate a wider class of distributions named as exponential family, providing great flexibility for modeling different types of response variables. Normal, Poisson, Binomial and Gamma are examples of this family of distributions. In Logistic Regression, Binomial distribution is considered to model the response. More details in [3].

LDA computes the probabilities of belonging to each of the groups according to the available variables using Bayes Theorem (posteriori probability) and Normal distribution. The predicted class will be the one whose posteriori probability is maximum [1, 2].

Poisson Discriminant Analysis (*PDA*) [20] and Negative Binomial Discriminant Analysis (*NBDA*) [21] are specific methods for RNA-Seq samples classification. They can be considered as an extension of the *LDA* because they are Bayes rule-based classifiers taking into account the discrete count distribution inherent in these data.

Results

We begin this section with palmdates data set to show a comparison between *LPDA* and *SVM* graphically. Then we show the results with Default data that is an unbalanced overlapped data set with a high number of samples where the separation is not possible. Finally, we present the application of *LPDA* and other methods to Cervical cancer RNA-Seq data.

Palmdates data

As *SVM* and *LPDA* are methods based in hyperplanes separation, it is worth taking a closer look at this comparison. By comparing results of *SVM* and *LPDA* to different data sets we have seen that working with a high number of variables or having clear differences between groups, both methods are succesfull separating groups. However when having few variables or existing overlaps between groups, we find some differences. As example we show pairwise variables comparison of palmdates concentration data. We consider 4 variables: fibre, fructose, sorbitol and myo-inositol, avoiding glucose because it is highly correlated with fructose and gives repeated results. Fig 3 shows cases where both methods are successfull but the hyperplanes are slight different and Fig 4 shows cases where *LPDA* gets less separation errors than *SVM*: only one predicted error with *LPDA* in the third comparison meanwhile there are 1, 2 and 7 errors respectively with *SVM*.

Default data

Another advantage we have found in *LPDA* with respect other techniques in several data sets is the good treatment of unbalanced data. These data is frequently encountered in biomedical

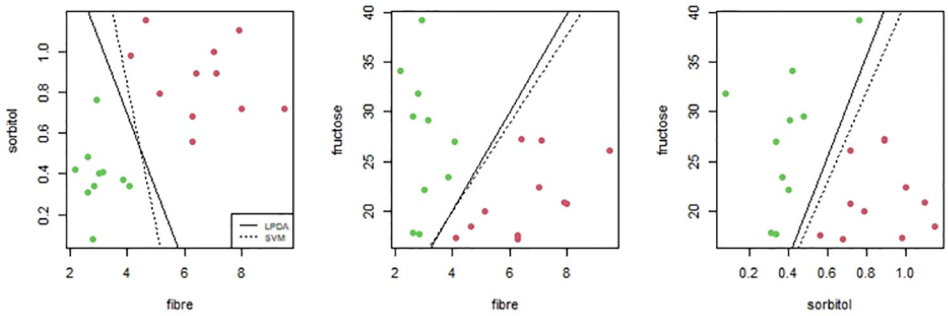

Fig 3. Examples where both methods are successfull but the hyperplanes are slight different.

https://doi.org/10.1371/journal.pone.0270403.g003

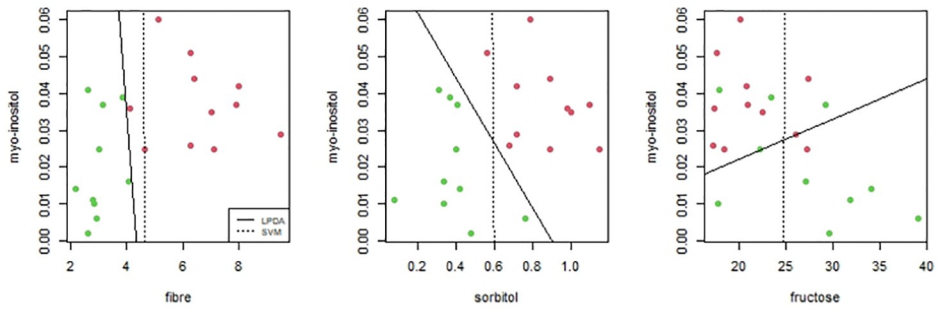

Fig 4. Examples where *LPDA* gets less separation errors than *SVM*.

https://doi.org/10.1371/journal.pone.0270403.g004

Table 1. Sensitivity, specificity and classification error for default data.

	Sensitivity	Specificity	Classification error
LPDA	0.9009	0.8646	0.1342
weighted-*SVM*	0.9039	0.8555	0.1429
Logistic	0.3153	0.2372	0.0267
LDA	0.2372	0.9977	0.0276

https://doi.org/10.1371/journal.pone.0270403.t001

and bioinformatics studies, where the group it is desired to predict is much smaller than the other one. General methods try to minimice the global error thus disadvantaging the minority class. Specific techniques are emerging for dealing with this problem [22]. Weights m_i and w_j considered by *LPDA* inside each group, mitigate this problem, meanwhile other techiques as *SVM* need to specify additional arguments when this situation arrises [23].

Default data is an example of unbalanced data that illustrates the problem clearly because only 0.3% of the data belongs to the group of interest (default class) that is desired to predict with low error. Table 1 shows sensitivity, especificity and the clasification error obtained with *LPDA*, weighted-*SVM*, Logistic Regression and *LDA*. We call weighted-*SVM* results of *SVM* applied considering as weights for each class the inverse of their sizes. As the interest is the good prediction in the default class, identified as the positive class, we must focuss in the sensitivity or percentage of True Positives detected. We observe that *LDA* and Logistic Regression give low sentitivity meanwhile *LPDA* gives a sensitivity very near the obtained with weighted-*SVM* and higher specificity, thus less global error.

Cervical cancer RNAseq data

We applied *LPDA*, *SVM*, *POlda* and *NBlda*, to the cervical cancer data described before. Firstly, all the data was considered to compute the number of classification errors as a training set. None error was detected with *LPDA* and *SVM*. However, *POlda* and *NBlda* gave 3 and 4 classification error respectively, therefore, *LPDA* and *SVM* give a separate hiperplane meanwhile methods based in distributional assumptions do not.

We also evaluated the methods in test sets with a cross-validation strategy where the model was obtained 1000 times in different training and test sets. Table 2 shows the classification error rates average jointly to their confidence intervals. First, we notice the importance of the dimension reduction (*LPDA-PCA*) in this case, and in general when dealing with high dimensional data as RNA-Seq, which significantly reduces the error rate. We also observe that

Table 2. Classification error test average and confidence interval in cervical cancer dataset.

Method	8 samples	10 samples	12 samples
LPDA	0.102 (0.096, 0.108)	0.106 (0.101, 0.112)	0.100 (0.095, 0.104)
LPDA-PCA	0.078 (0.072, 0.084)	0.078 (0.072, 0.083)	0.081 (0.076, 0.086)
SVM	0.076 (0.070, 0.082)	0.078 (0.073, 0.083)	0.081 (0.076, 0.085)
POlda	0.102 (0.096, 0.109)	0.105 (0.099, 0.110)	0.106 (0.101, 0.111)
NBlda	0.076 (0.071, 0.082)	0.082 (0.077, 0.087)	0.079 (0.075, 0.084)

https://doi.org/10.1371/journal.pone.0270403.t002

LPDA-PCA results are very similar to *SVM* and *NBlda* meanwhile *LPDA* without PCA results are similar to the *POlda* approach.

Conclusions

In this work, we propose a classification method based in a linear programming problem that is efficient in multiple scenarios. First, we show the basis of the method defining an optimization problem from the idea of separating two data sets in $\mathbb{R}^p$. Then we consider the aplication of PCA when having overfitting problems due to high dimensional data and also usefull for correlated data. The method has been applied to different data sets and compared with popular techniques as *SVM*, Logistic Regression and *LDA*. One of these data sets is a real RNA-Seq data for which we considered the comparison with specific methods developed for the specific problematic of this type of data (*NBlda* and *POlda*).

Results show that *LPDA* is efficient in different situations. We have demonstrated its effectiveness in unbalanced experiments where it is able to classify minority classes without adding additional considerations. Moreover, its performance in high-dimensional data sets, such as RNA-Seq data, is similar to the popular *SVM* and also to *NBlda*, developed specially for the specific problematic of this type of data considering distributional hypothesis.

In this paper we have applied the method only in experiments where individuals are classified in two groups, but the method is extrapolated to three or more classes making pairwise comparisons in the available R-package.

In conclusion, *LPDA* is an efficient classification method for general multivariate data.

Author Contributions

Formal analysis: María J. Nueda, Carmen Gandía, Mariola D. Molina.

Investigation: María J. Nueda.

Resources: María J. Nueda.

Software: María J. Nueda, Carmen Gandía.

Writing – original draft: María J. Nueda, Mariola D. Molina.

Writing – review & editing: María J. Nueda, Mariola D. Molina.

References

1. Fisher RA. The use of multiple measurements in taxonomic problems. Eugen. 1936; 7:179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
2. Rao CR. Linear Statistical Inference and its Applications. New York: Wiley; 1973.
3. Nelder JA, Wedderburn RWM. Generalized linear models. Journal of the Royal Statistical Society. 1972; Series A(135):370–384. https://doi.org/10.2307/2344614
4. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1996.

5. Vapnik V. Statistical Learning Theory. New York: Wiley; 1998.
6. Carrizosa E, Morales DR. Supervised classification and mathematical optimization. Computers & Operation Research. 2013; 40:150–165. https://doi.org/10.1016/j.cor.2012.05.015
7. Duarte-Silva AP. Optimization approaches to Supervised Classification. European Journal of Operational Research. 2017; 261:772–788. https://doi.org/10.1016/j.ejor.2017.02.020
8. Bal H, Örkcü HH. A new mathematical programming approach to multi-group classification. Computers & Operation Research. 2011; 38:105–111. https://doi.org/10.1016/j.cor.2010.04.003
9. Breiman L, Friedman J, Stone CJ, Olshen R. Classification and Regression Trees. Boca Raton: Chapman & Hall/CRC; 1984.
10. Breiman L. Random forests. Machine Learning. 2001; 45:5–32. https://doi.org/10.1023/A:1010933404324
11. Efron B, Hastie T. Computer Age Statistical Inference. Algorithms, Evidence and Data Science. pp.351–374. Cambridge University Press; 2016.
12. James G. An Introduction to Statistical Learning with applications in R. New York: Springer-Verlag; 2013. Available from: https://www.statlearning.com.
13. Bertsimas D, Tsitsiklis J. Introduction to Linear Optimization. Athena Scientific; 1998.
14. Nueda MJ, Conesa A, Westerhuis J, Hoefsloot H, Smilde A, Talón M, Ferrer A. Discovering gene expression patterns in Time Course Microarray Experiments by ANOVA-SCA. Bioinformatics. 2007; 23(14):1792–1800. https://doi.org/10.1093/bioinformatics/btm251 PMID: 17519250
15. Nueda MJ, Ferrer A, Conesa A. ARSyN: a method for the identification and removal of systematic noise in multifactorial time course microarray experiments. Biostatistics. 2012; 13(3):553–566. https://doi.org/10.1093/biostatistics/kxr042 PMID: 22085896
16. Smilde A, Bro T, Geladi P. Multi-way Analysis. England: Wiley; 2004.
17. Witten DM, Tibshirani R, Gu SG, Fire A, Lui W. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. BMC Biol. 2010; 8(58). https://doi.org/10.1186/1741-7007-8-58 PMID: 20459774
18. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. https://doi.org/10.1186/1471-2105-11-94 PMID: 20167110
19. Karatzoglou A, Meyer D. Support Vector Machines in R. Journal of Statistical Software. 2006; 15(9). https://doi.org/10.18637/jss.v015.i09
20. Witten DM. Classification and clustering of sequencing data using a Poisson model. The Annals of Applied Statistics. 2011; 5(4):2493–2518. https://doi.org/10.1214/11-AOAS493
21. Dong K, Zhao H, Tong T, Wan X. NBLDA: negative binomial linear discriminant analysis for RNA-Seq data. BMC Bioinformatics. 2016; 17:369. https://doi.org/10.1186/s12859-016-1208-1 PMID: 27623864
22. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE. 2017; 12(6):e0177678. https://doi.org/10.1371/journal.pone.0177678 PMID: 28574989
23. Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. In: Machine learning: ECML 2004. p. 39-50. Springer; 2004.