Geometric Discriminant Analysis (part I)

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

In these slides we discuss the approach originally proposed by Fisher. He formulated the classfication problem in a geometric way. He sought to find the linear combination of the predictors such that the between-group variance was maximized relative to the within-group variance.

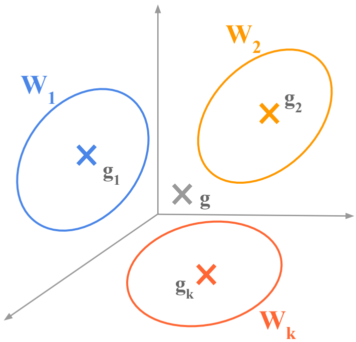
In other words, he wanted to find the combination of the predictors that gave maximum separation between the centroids of the data while at the same time minimizing the variation within each group of data.

Main Problem

How to find a representation of the objects which provides the best separation between groups (description emphasis)?

How to find the rules for assigning the objects to their groups (prediction emphasis)?

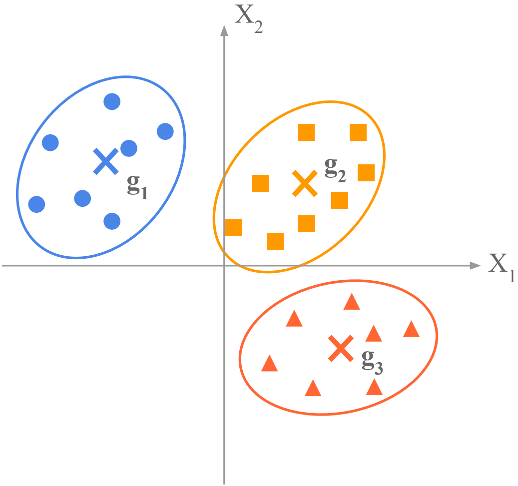
Between and Within Dispersion



Variance Matrices

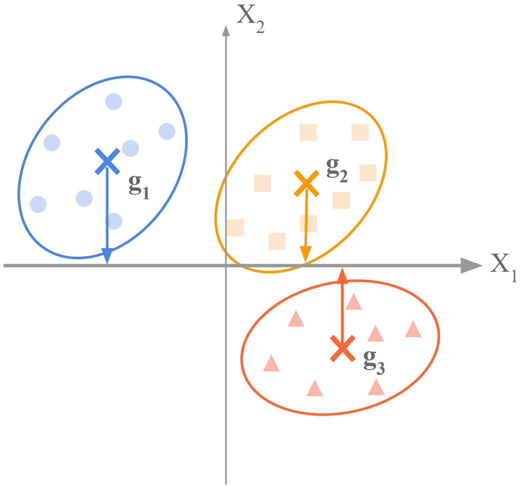
- Within-groups: $\mathbf{W}$
- Between-groups: $\mathbf{B}$
- Total: $\mathbf{V} = \mathbf{W} + \mathbf{B}$

Say we have 3 classes in 2-dim space



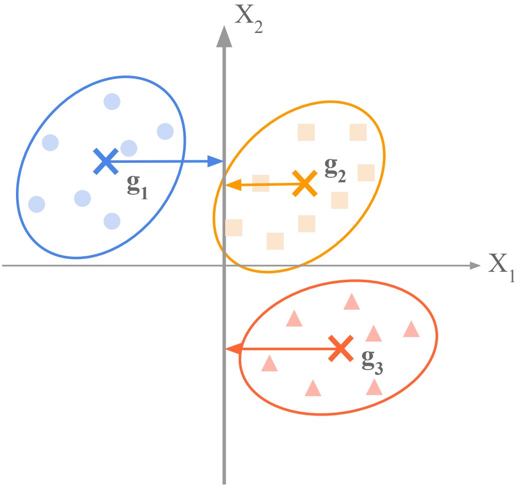
We look for the best representation separating the groups

Looking for optimal representation



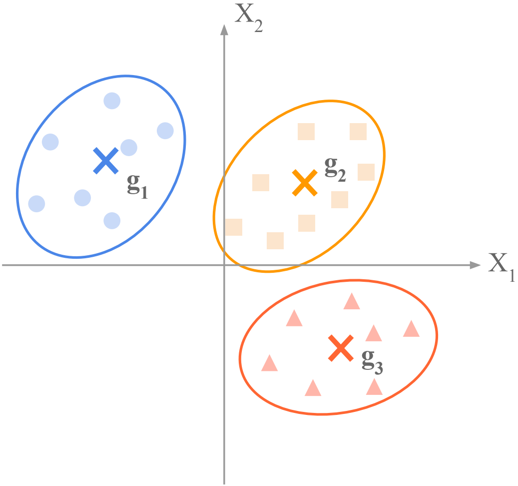
Axis X_1 separates group 1 from groups 2 and 3

Looking for optimal representation



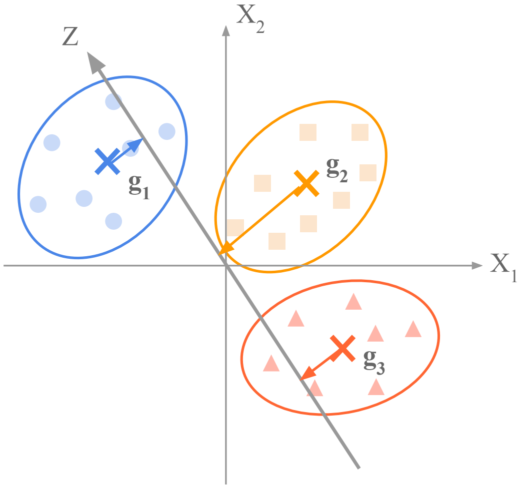
Axis X_2 separates group 3 from groups 1 and 2

Looking for optimal representation



Is there an axis that "best" separates the clouds?

Looking for a discriminant axis



Axis $Z = u_1X_1 + u_2X_2$ separates all three groups

Main Problem

How to find a low dimensional representation of the objects which provides the best separation between groups?

Double goal ideal

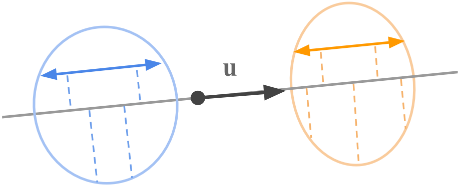
We look for a linear combination of the predictors, $\mathbf{z} = \mathbf{X}\mathbf{u}$, that *ideally* could:

- Minimize within-groups dispersion: $min\{\mathbf{u}^\mathsf{T}\mathbf{W}\mathbf{u}\}$

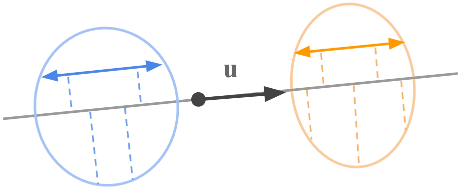
  and

- Maximize between-groups dispersion: $max\{\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}\}$

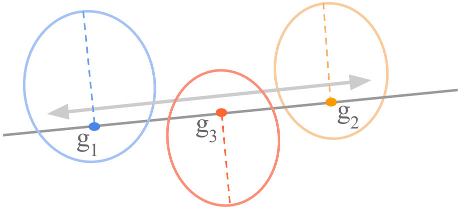
Minimize within-groups dispersion: $min\{\mathbf{u}^\mathsf{T}\mathbf{W}\mathbf{u}\}$



Minimize within-groups dispersion: $min\{\mathbf{u}^\top \mathbf{W} \mathbf{u}\}$



Maximize between-groups dispersion: $max\{\mathbf{u}^\top \mathbf{B} \mathbf{u}\}$



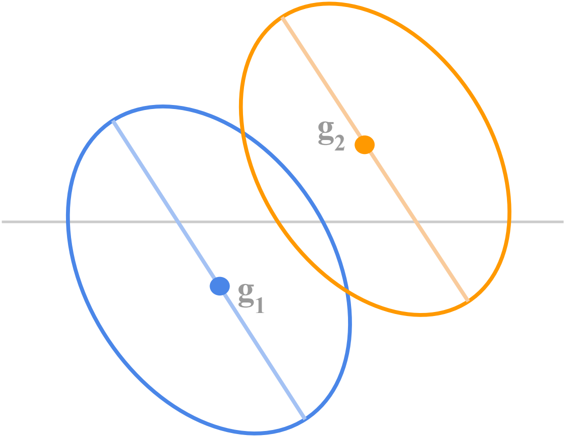
Two Incompatible Goals

Not so good news: It is generally impossible to find an axis Δ_u, spanned by $\mathbf{u}$, which in order to meet the objective of discriminant analysis, simoultaneously:

- maximizes the between-groups variance
- minimizes the within-groups variance

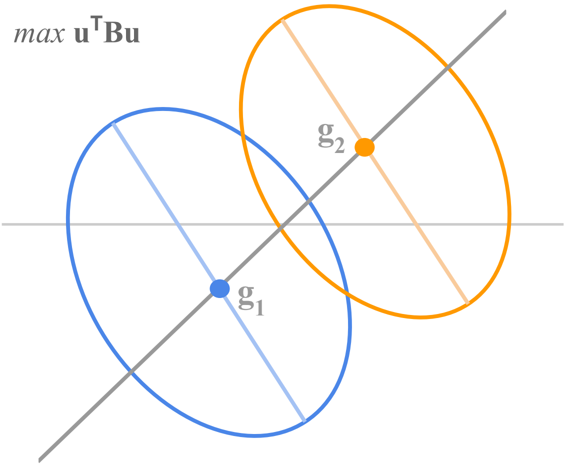
Let's see a picture of this issue

Double goal cartoon picture



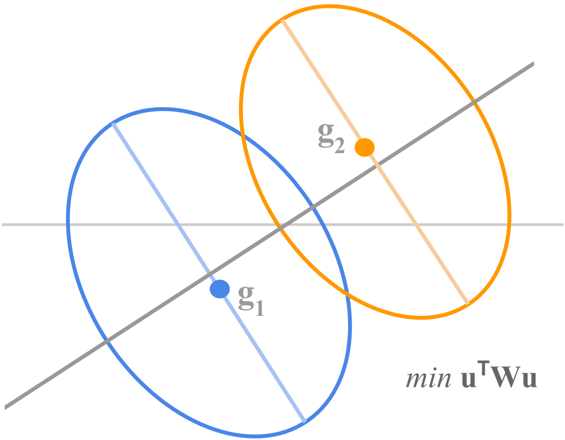
Double goal of discriminant analysis ... generally impossible

Double goal cartoon picture



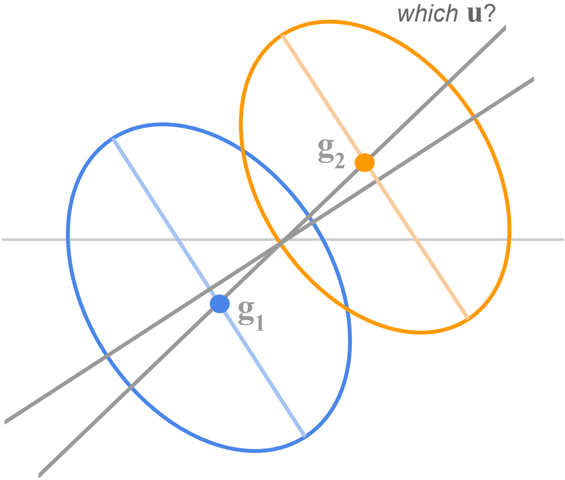
Double goal of discriminant analysis ... generally impossible

Double goal cartoon picture



Double goal of discriminant analysis ... generally impossible

Double goal cartoon picture



Double goal of discriminant analysis ... generally impossible

Double goal issue

If we are looking for the maximum between-groups dispersion, we will choose an axis $\mathbf{u}$ parallel to the segment linking the centroids.

If we are looking for the minimum within-groups dispersion, we will choose an axis $\mathbf{u}$ perpendicular to the principal axis of the ellipses.

Double goal issue

Impossible simultaneity (in general):

$$min\{\mathbf{u}^\mathsf{T}\mathbf{W}\mathbf{u}\} \Longrightarrow \mathbf{W}\mathbf{u} = \alpha\mathbf{u}$$

$$max\{\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}\} \Longrightarrow \mathbf{B}\mathbf{u} = \beta\mathbf{u}$$

Main Problem

We should look then for a compromise. This is where the variance decomposition comes handy: $\mathbf{V} = \mathbf{W} + \mathbf{B}$

$$\mathbf{u}^\mathsf{T}\mathbf{V}\mathbf{u} = \mathbf{u}^\mathsf{T}\mathbf{W}\mathbf{u} + \mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}$$

Main Problem

We should look then for a compromise. This is where the variance decomposition comes handy: $\mathbf{V} = \mathbf{W} + \mathbf{B}$

$$\mathbf{u}^\mathsf{T}\mathbf{V}\mathbf{u} = \underbrace{\mathbf{u}^\mathsf{T}\mathbf{W}\mathbf{u}}_{\text{minimize}} + \underbrace{\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}}_{\text{maximize}}$$

Main Problem

We have two options for the compromise:

$$max \left\{ \frac{\mathbf{u}^\top \mathbf{B} \mathbf{u}}{\mathbf{u}^\top \mathbf{V} \mathbf{u}} \right\} \quad \text{OR} \quad max \left\{ \frac{\mathbf{u}^\top \mathbf{B} \mathbf{u}}{\mathbf{u}^\top \mathbf{W} \mathbf{u}} \right\}$$

Solution

We look for $\mathbf{u}$ such that:

$$max\left\{\frac{\mathbf{u}^\intercal \mathbf{B}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{V}\mathbf{u}}\right\}$$

This criterion is scale invariant, meaning that we use any scale variation of $\mathbf{u}$: i.e. $\alpha\mathbf{u}$

Thus:

$$max\left\{\frac{\mathbf{u}^\intercal \mathbf{B}\mathbf{u}}{\mathbf{u}^\intercal \mathbf{V}\mathbf{u}}\right\} \iff max\{\mathbf{u}^\intercal \mathbf{B}\mathbf{u}\} \quad \text{s.t.} \quad \mathbf{u}^\intercal \mathbf{V}\mathbf{u} = 1$$

Lagrangian

Lagrangian:

$$L(\mathbf{u}) = \mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u} - \lambda(\mathbf{u}^\mathsf{T}\mathbf{V}\mathbf{u} - 1)$$

Deriving w.r.t $\mathbf{u}$ and equation to zero:

$$\frac{\partial L(\mathbf{u})}{\partial \mathbf{u}} = 2\mathbf{B}\mathbf{u} - 2\lambda\mathbf{V}\mathbf{u} = 0$$

The optimal vector $\mathbf{u}$ is such that:

$$\mathbf{B}\mathbf{u} = \lambda\mathbf{V}\mathbf{u}$$

Optimal $\mathbf{u}$

If the matrix $\mathbf{V}$ is inversible (which it is in general) then:

$$\mathbf{V}^{-1}\mathbf{B}\mathbf{u} = \lambda\mathbf{u}$$

that is, the optimal vector $\mathbf{u}$ is eigenvector of $\mathbf{V}^{-1}\mathbf{B}$

Equivalence between $\mathbf{V}^{-1}$ and $\mathbf{W}^{-1}$

The vector $\mathbf{u}$ verifies $\mathbf{Bu} = \lambda \mathbf{Vu}$ subject to $\mathbf{u}^\mathsf{T}\mathbf{Vu} = 1$

If we replace $\mathbf{V}$ by $\mathbf{B} + \mathbf{W}$ then we have:

$$\mathbf{Bu} = \rho \mathbf{Wu} \quad \text{with} \quad \rho = \frac{\lambda}{1 - \lambda}$$

It can be shown that $\mathbf{u}$ is also eigenvector of $\mathbf{W}^{-1}\mathbf{B}$

$$\mathbf{W}^{-1}\mathbf{Bu} = \rho \mathbf{u}$$

Keep in mind that $\mathbf{W}^{-1}\mathbf{B}$ is not symmetric.

Factorizing $\mathbf{B}$

Matrix $\mathbf{B}$ can be obtained as:

$$\mathbf{B} = \sum_{k=1}^{K} \frac{n_k}{n-1} (\mathbf{g_k} - \mathbf{g})(\mathbf{g_k} - \mathbf{g})^{\mathsf{T}}$$

where

- $\mathbf{g_k}$ is the centroid of group G_k
- $\mathbf{g}$ is the global centroid

Factorizing $\mathbf{B}$

Matrix $\mathbf{B}$ has general term (row j, column l):

$$b_{jl} = \sum_{k=1}^{K} \frac{n_k}{n-1} (g_{kj} - g_j)(g_{kl} - g_l)$$

where:

- g_{kj} if the j-th element of $\mathbf{g_k}$
- g_{kl} if the l-th element of $\mathbf{g_k}$
- g_j if the j-th element of $\mathbf{g}$
- g_l if the l-th element of $\mathbf{g}$

Factorizing $\mathbf{B}$

Matrix $\mathbf{B}$ has general term (row j, column l):

$$b_{jl} = \sum_{k=1}^{K} \frac{n_k}{n-1}(g_{kj} - g_j)(g_{kl} - g_l)$$

where:

- g_{kj} is the j-th element of $\mathbf{g_k}$
- g_{kl} is the l-th element of $\mathbf{g_k}$
- g_j is the j-th element of $\mathbf{g}$
- g_l is the l-th element of $\mathbf{g}$

Factorizing $\mathbf{B}$

We can factorize $\mathbf{B}$ as:

$$\mathbf{B} = \mathbf{C}\mathbf{C}^{\mathsf{T}}$$

where $\mathbf{C}$ has general term:

$$c_{jk} = \sqrt{\frac{n_k}{n-1}}(\bar{x}_{kj} - \bar{x}_j)$$

EVD solution

The $p \times p$ matrix $\mathbf{W}^{-1}\mathbf{B}$ and the $k \times k$ matrix $\mathbf{C}^{\mathsf{T}}\mathbf{W}^{-1}\mathbf{C}$ have the same eigenvalues.
Their eigenvectors are related by:

$$\mathbf{u} = \mathbf{W}^{-1}\mathbf{C}\mathbf{w}$$

Thus, we can diagonalize (EVD) the following symmetric matrix:

$$\mathbf{C}^{\mathsf{T}}\mathbf{W}^{-1}\mathbf{C}$$

and then use the eigenvector $\mathbf{w}$ to recover $\mathbf{u}$

Let's Recap

Instead of maximizing $\mathbf{u}^\intercal \mathbf{B} \mathbf{u}$ or minimizing $\mathbf{u}^\intercal \mathbf{W} \mathbf{u}$, we maximize $\mathbf{u}^\intercal \mathbf{B} \mathbf{u} / \mathbf{u}^\intercal \mathbf{V} \mathbf{u}$, which according to the Huygens theorem is equivalent to maximizing $\mathbf{u}^\intercal \mathbf{B} \mathbf{u} / \mathbf{u}^\intercal \mathbf{W} \mathbf{u}$

It can be shown that the solution $\mathbf{u}$ is the eigenvector of $\mathbf{V}^{-1}\mathbf{B}$ associated with λ, the largest eigenvector of $\mathbf{V}^{-1}\mathbf{B}$.

Moreover, it turns out that $\mathbf{u}$ is an eigenvector of $\mathbf{V}^{-1}\mathbf{B}$ if and only if $\mathbf{u}$ is an eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ with a corresponding eigenvalue of $\rho = \lambda/(1 - \lambda)$

Metrics

The metrics $\mathbf{V^{-1}}$ and $\mathbf{W^{-1}}$ are therefore called **equivalent**, but the metric $\mathbf{W^{-1}}$ (the Mahalanobis metric) is used more widely by software developers.

With the Mahalanobis metric, the square of the distance between two points p_1 and p_2 is

$$d^2(p_1, p_2) = (\mathbf{p_1} - \mathbf{p_2})^\mathsf{T} \mathbf{W^{-1}} (\mathbf{p_1} - \mathbf{p_2})$$

A special PCA

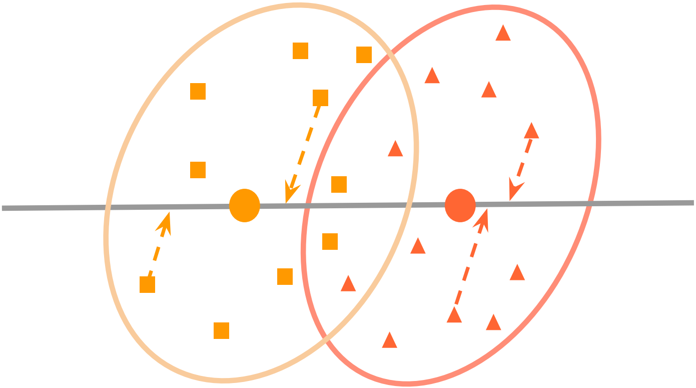
What do $\mathbf{u}$ and $\mathbf{W^{-1}}$ correspond in geometric terms?

$\mathbf{u}$ is the axis from the PCA on the cloud of centroids $\mathbf{g_k}$, but it is an axis on which the points are projected obliquely, not orthogonally.

Without this obliqueness, corresponding to the equivalent metrics $\mathbf{V^{-1}}$ and $\mathbf{W^{-1}}$, this would be a simple PCA, in which the groups would be less well separated.

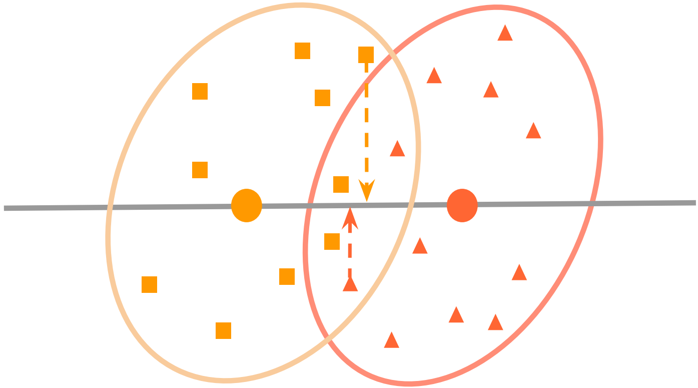
With $\mathbf{W^{-1}}$, the separation of two points depends not only on a Euclidean measurement, but also on the variance and correlation of the variables.

Oblique projection with $\mathbf{W}^{-1}$



Points are projected obliquely with $\mathbf{W}^{-1}$

Orthogonal projection without $\mathbf{W}^{-1}$



Without $\mathbf{W}^{-1}$, points would be orthogonally projected

Canonical Axes and Canonical Variables

- $\mathbf{u}$ is the vector associated to the so-called **canonical axis**
- When the first canonical axis has been determined, we search for a 2nd one
- The second axis should be the most discriminant and uncorrelated with the first one
- This procedure is repeated until the number of axis reaches the minimum of: $K-1$ and p

In fact, it is not the canonical axes that are manipulated directly, but the *canonical variables* or vectors associated to the canonical axes.

Canonical Axes and Canonical Variables

In the case of two classes ($K = 2$), the canonical axis is unique and it turns out that is proportional to $\mathbf{W}^{-1}(\mathbf{g_1} - \mathbf{g_2})$

Iris Data Again

Dataset `iris` in R

150 Observations

- 150 iris flowers

Four predictors

- `Sepal.Length`
- `Sepal.Width`
- `Petal.Length`
- `Petal.Width`

One response (qualitative)

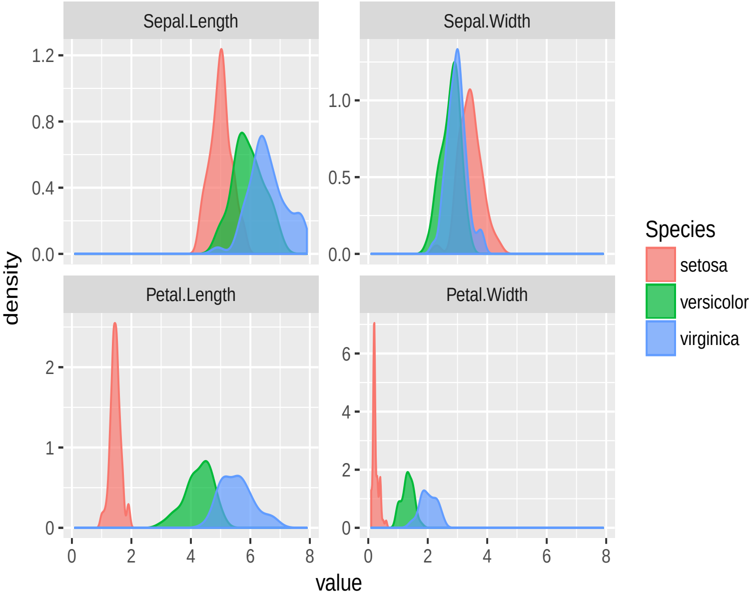
- `Species` (3 classes: setosa, versicolor, virginica)

Famous data set collected by Edgar Anderson (1935), and used by Ronald Fisher (1936) in his paper about Discriminant Analysis.

Dataset iris in R

```
head(iris)

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```



Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
density
value
Species
setosa
versicolor
virginica
1.2
0.8
0.4
0.0
1.0
0.5
0.0
2
1
0
6
4
2
0
0
2
4
6
8
0
2
4
6
8

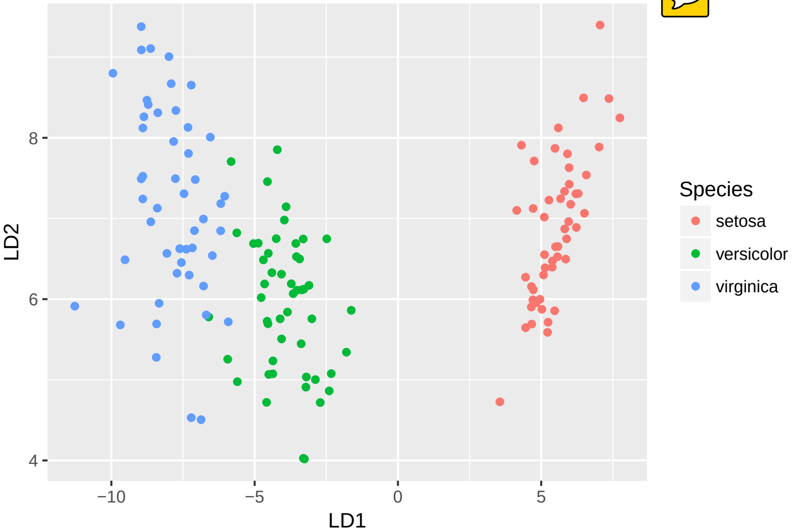
```r
# lda() from package "MASS"
geo_disc <- lda(Species ~ ., data = iris)
geo_disc$scaling
```

```
##                     LD1         LD2
## Sepal.Length  0.8293776  0.02410215
## Sepal.Width   1.5344731  2.16452123
## Petal.Length -2.2012117 -0.93192121
## Petal.Width  -2.8104603  2.83918785
```

```r
# canonical variables
Z <- as.matrix(iris[ ,1:4]) %*% geo_disc$scaling
iris_lda <- data.frame(Z)
iris_lda$Species <- iris$Species

head(iris_lda, n = 5)

##         LD1      LD2 Species
## 1 5.956693 6.961893  setosa
## 2 5.023581 5.874812  setosa
## 3 5.384722 6.396088  setosa
## 4 4.708094 5.990841  setosa
## 5 6.027203 7.175935  setosa
```



LD2
8
6
4
−10
−5
0
5
LD1
Species
setosa
versicolor
virginica

References

- **Principles of Multivariate Analysis: A User's Perspective** by W.J. Krzanowski (1988). *Chapter 11: Incorporating group structure: descriptive methods*. Oxford University Press.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*. Wiley.
- **Multivariate Analysis** by Maurice Tatsuoka (1988). *Chapter 7: Discriminant Analysis and Canonical Correlation*.
- **Practical Biostatistical Methods** by Steve Selvin (1995) *Chapter 6: Linear Discriminant Analysis*. Duxbury Press.

References

- **The use of multiple measurements in taxonomic problems** by R.A. Fisher (1936). *Annals of Eugenics, 7, 179-188*.
- **On the generalized distance in statistics** by P.C. Mahalanobis (1936). *Proceedings of the National Institute of Science, India, 12, 49-55*.
- **Discriminant Analysis** by Tatsuoka and Tiedeman (1954). *Review of Educational Research, 25, 402-420*.

References (French Literature)

- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3: Analyse factorielle discriminante*. Dunod, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique explicative appliquee** by Nakache and Confais (2003). *Chapter 1: Analyse discriminante sur variables quantitatives*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 10: L'analyse discriminante*. Dunod, Paris.