Preamble to Discriminant Analysis

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

In these slides I'll talk about the concept of Variance decomposition taking into account a group structure.

The idea is to layout a couple of foundational principles that should allow you to understand discriminant methods in a more comprehensive way.

BTW: this material is not in the textbooks *ISL* and *APM*.

Iris Data

Dataset `iris` in R

$n = 150$ Observations, i.e. iris flowers

$p = 4$ predictors

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

One response (categorical)

- `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.

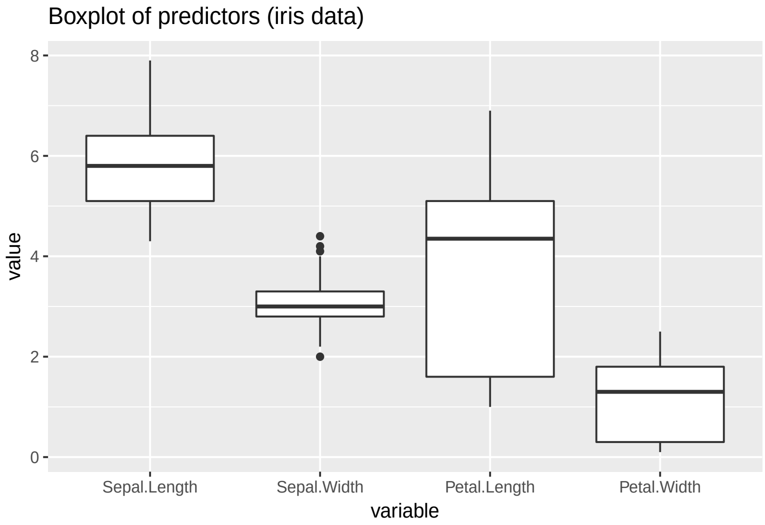
Dataset iris in R

```
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
```

Dataset iris in R

```r
summary(iris)
```

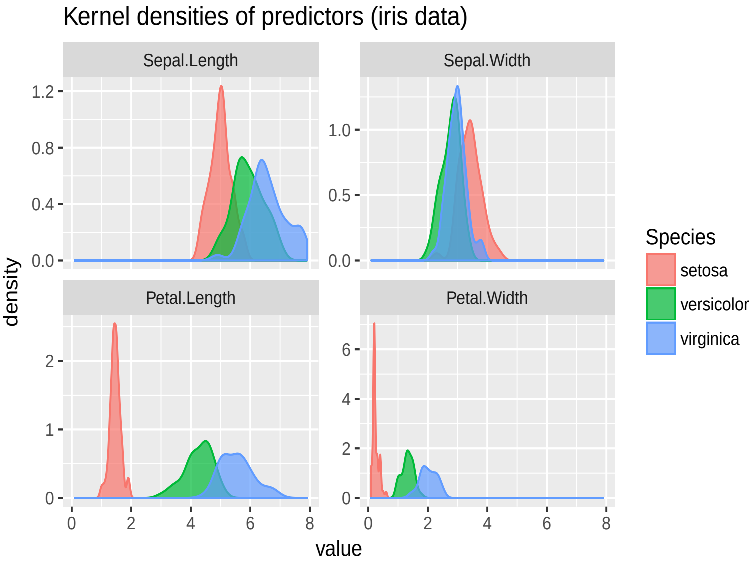
```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```



Boxplot of predictors (iris data)
value
8
6
4
2
0
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
variable

Let's take into account
the group structure

Boxplot of predictors (iris data)


Kernel densities of predictors (iris data)
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
library(reshape2)
library(ggplot2)

iris_melt <- melt(iris, id = "Species")

ggplot(data = iris_melt, aes(x = variable, y = value)) +
  geom_boxplot() +
  ggtitle("Boxplot of predictors (iris data)")

ggplot(data = iris_melt, aes(x = variable, y = value)) +
  geom_boxplot(aes(fill = Species, color = Species)) +
  ggtitle("Boxplot of predictors (iris data)")

ggplot(data = iris_melt, aes(x = value)) +
  geom_density(aes(fill = Species, color = Species),
               alpha = 0.7) +
  facet_wrap(~ variable, scales = 'free_y') +
  ggtitle("Kernel densities of predictors (iris data)")
```

Which predictor provides the
"best" distinction between Species?

Caveat: messy notation

In regression problems we've been using two indices i and j

- i for objects, $i = 1, \dots, n$
- j for predictors, $j = 1, \dots, p$

New index k

Now we have a new index k for groups or classes, $k = 1, \dots, K$.

Sum of Squares

Consider a single predictor X and a categorical response Y

Ignoring the respone, we can obtain the mean $\bar{x}$ and the total sum of squares (TSS) of X as:

$$\bar{x} = \sum_{i=1}^{n} x_i$$

$$\text{TSS} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Group (or class) structure

Let's take into account the group structure conveyed by Y

- Let G_k represent the k-th group in Y
- Let n_k be the number of observations in group G_k,
- Then:

$$n = n_1 + n_2 + \cdots + n_K = \sum_{k=1}^{K} n_k$$

Between Sum of Squares

Each group k will have its mean $\bar{x}_k$:

$$\bar{x}_k = \sum_{i \in G_k} x_{ik}$$

We can obtain the *Between-Groups Sum of Squares* (BSS)

$$\text{BSS} = \sum_{k=1}^{K} n_k (x_k - \bar{x})^2$$

Within Sum of Squares

Each group k will also have its own sum-of-squares SS_k

$$\mathsf{SS}_k = \sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2$$

We can obtain the *Within-Groups Sum of Squares* (WSS)

$$\mathsf{WSS} = \sum_{k=1}^{K} \sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2$$

Decomposition of sums-of-squares

An important aspect has to do with looking at the squared deviations: $(x_i - \bar{x})^2$ in terms of the group structure.

A useful trick is to rewrite the deviation terms $x_i - \bar{x}$ as:

$$\begin{aligned} x_i - \bar{x} &= x_i - (\bar{x}_k - \bar{x}_k) - \bar{x} \\ &= (x_i - \bar{x}_k) + (\bar{x}_k - \bar{x}) \end{aligned}$$

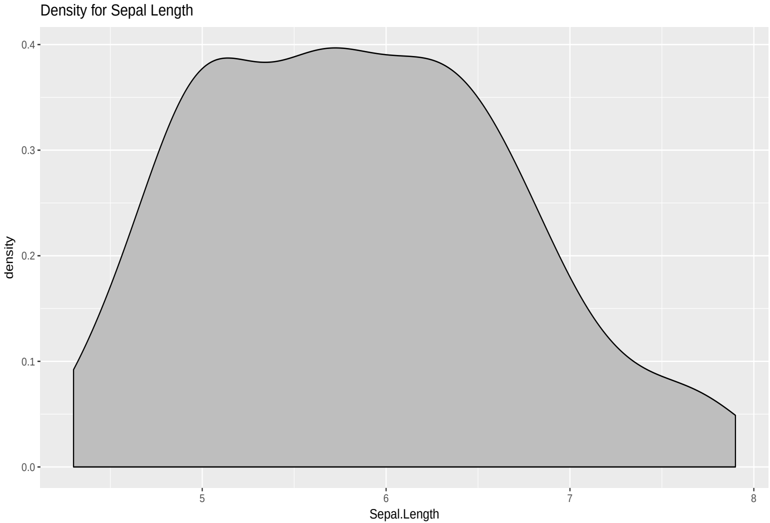
Sum of Squares Decomposition

We can decompose TSS in terms of BSS and WSS:

$$\underbrace{\sum_{k=1}^{K} \sum_{i \in G_k} (x_{ik} - \bar{x})^2}_{\text{TSS}} = \underbrace{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2}_{\text{BSS}} + \underbrace{\sum_{i \in G_k} (x_{ik} - \bar{x}_k)^2}_{\text{WSS}}$$

In summary:

$$\text{TSS} = \text{BSS} + \text{WSS}$$




```r
ggplot(data = iris, aes(x = Sepal.Length)) +
geom_density(fill = 'gray') +
ggtitle('Density for Sepal Length')
```

TSS for Sepal.Length

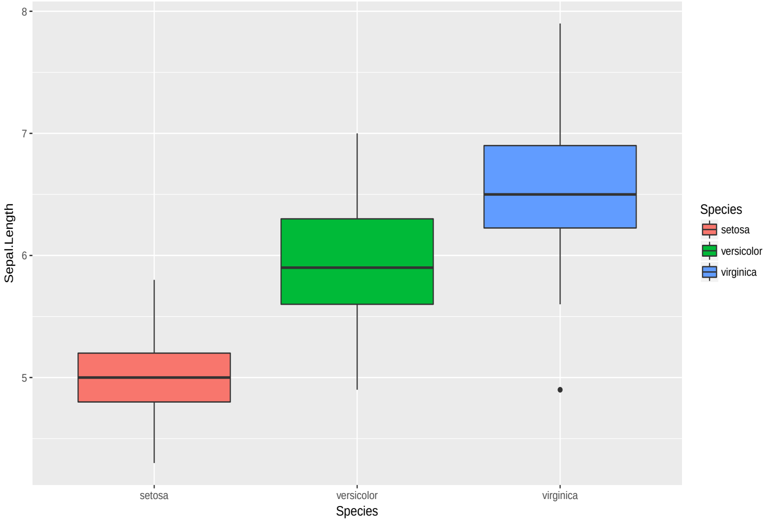
```r
x <- iris$Sepal.Length

# overall mean
x_bar <- mean(x)
x_bar
```

```
## [1] 5.843333
```

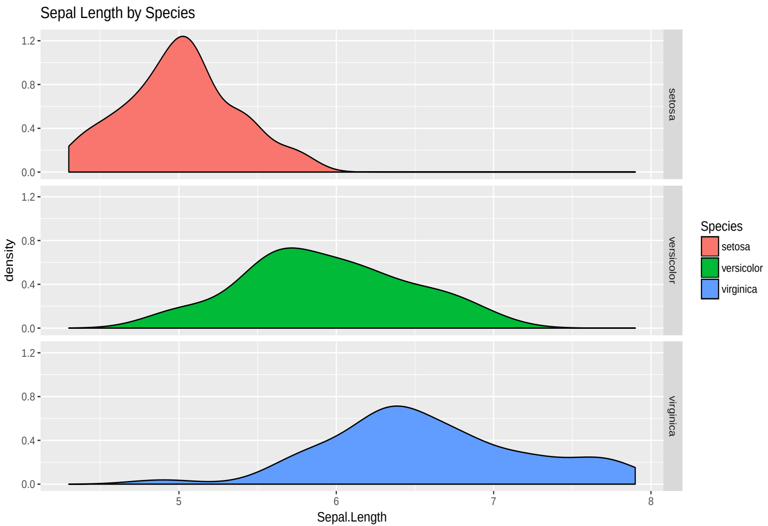
```r
# total sums-of-squares
tss <- sum((x - x_bar)^2)
tss
```

```
## [1] 102.1683
```

Let's consider the group structure




```r
ggplot(data = iris, aes(x = Species, y = Sepal.Length)) +
  geom_boxplot(aes(fill = Species))
```




```r
ggplot(data = iris, aes(x = Sepal.Length, group = Species)) +
  geom_density(aes(fill = Species)) +
  facet_grid(Species ~ .) +
  ggtitle('Sepal Length by Species')
```

BSS for Sepal.Length

```
# Sepal Length group means
x_means <- tapply(x, iris$Species, mean)

# between sums-of-squares
bss <- sum(50 * (x_means - x_bar)^2)
bss
```

```
## [1] 63.21213
```

WSS for Sepal.Length

```r
# Sepal Length group sum of squares
w1 <- sum((x[1:50] - x_means[1])^2)
w2 <- sum((x[51:100] - x_means[2])^2)
w3 <- sum((x[101:150] - x_means[3])^2)

# within sums-of-squares
wss <- w1 + w2 + w3
wss
```

```
## [1] 38.9562
```

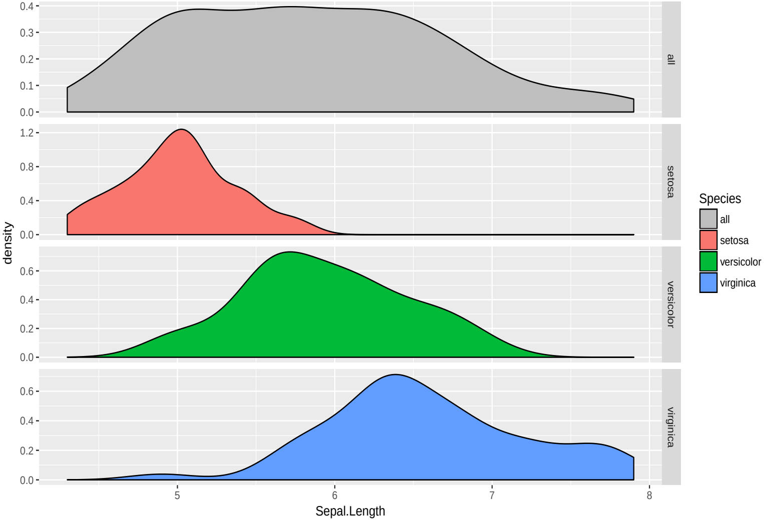
TSS Decomposition

Let's check that we have:

$$\text{TSS} = \text{BSS} + \text{WSS}$$

```
# tss
tss

## [1] 102.1683

# bss + wss
bss + wss

## [1] 102.1683
```

Dispersion in Sepal.Length



Derived Ratios from $\text{TSS} = \text{BSS} + \text{WSS}$

Correlation Ratio

Correlation ratio η^2 (proposed by K. Pearson):

$$\eta^2(X, Y) = \frac{\mathsf{BSS}}{\mathsf{TSS}}$$

- η^2 takes vaues between 0 and 1
- $\eta^2 = 0$ represents the special case of no dispersion among the means of the different categories
- $\eta^2 = 1$ refers to no dispersion within the respective categories.

The correlation ratio is a measure of the relationship between the dispersion within categories and the dispersion across all individuals.

F Ratio

With TSS $=$ BSS $+$ WSS, we can also calculate:

F ratio (proposed by R.A. Fisher):

$$F = \frac{\text{BSS}/(k-1)}{\text{WSS}/(n-k)}$$

The larger the value of both ratios, the more variability is there between groups than within groups.

Ratios for Sepal.Length

```
# correlation ratio
eta_sqr <- bss / tss
eta_sqr

## [1] 0.6187057

# F ratio
F_ratio <- (bss / (3 - 1)) / (wss / (150 - 3))
F_ratio

## [1] 119.2645
```

Ratios for all Variables

Let's compute the decompositions for all predictors, and obtain the correlation ratios and F ratios

```
etas

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    0.6187057    0.4007828    0.9413717    0.9288829

Fs

## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##    119.26450     49.16004   1180.16118    960.00715
```

More Notation: generalization for more than 1 predictor

Predictors and Response

- p predictors $X_1, X_2, \ldots, X_p$
- One categorical response Y with K categories
- Y introduces a group or class structure
- Observations divided in K groups or classes

Caveat: messy notation

Here's some notation that I'll be using while covering classification methods:

Let n_k be the number of observations in the k-th group

Let x_{ijk} represent the i-th observation, of the j-th variable, in the k-th group.

Let x_{ik} represent i-th observation in group k

Let x_{jk} represent j-th variable in group k

I hope this doesn't create a lot of confussion

Caveat: messy notation

Let n_k be the number of observations in the k-th group G_k, then:

$$n = n_1 + n_2 + \cdots + n_K = \sum_{k=1}^{K} n_k$$

Caveat: messy notation

For a given variable X_j, represented with vector $\mathbf{x_j}$, we have:
Total or global mean:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

Local mean of observations in group k:

$$\bar{x}_{jk} = \frac{1}{n_k} \sum_{i \in G_k} x_{ij}$$

where G_k represents the set of observations in group k

Caveat: messy notation

For a given variable X_j, representeded with vector $\mathbf{x_j}$, we have:
Total Sum of Squared deviations

$$\mathsf{TSS}_j = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

Assuming centered variables (mean $=$ 0)

$$\mathsf{TSS}_j = \mathbf{x_j}^\mathsf{T} \mathbf{x_j}$$

Decomposition of sums-of-squares

An important aspect has to do with looking at the squared deviations: $(x_{ij} - \bar{x}_j)^2$ in terms of the group structure.

A useful trick is to rewrite the deviation terms $x_{ij} - \bar{x}_j$, as:

$$\begin{aligned} x_{ij} - \bar{x}_j &= x_{ij} - (\bar{x}_{jk} - \bar{x}_{jk}) - \bar{x}_j \\ &= (x_{ij} - \bar{x}_{jk}) + (\bar{x}_{jk} - \bar{x}_j) \end{aligned}$$

Decomposition of Sums-of-Squares

Using the previous format of deviations, the sum of squared deviations can be decomposed as:

$$\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = \sum_{k=1}^{K} n_k(\bar{x}_{jk} - \bar{x}_k)^2 + \sum_{k=1}^{K}\sum_{i \in G_k}(x_{ijk} - \bar{x}_{jk})^2$$

What's this?

Decomposition of Sums-of-Squares

Using the previous format of deviations, the sum of squared deviations can be decomposed as:

$$\underbrace{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}_{\text{Total SS}} = \underbrace{\sum_{k=1}^{K} n_k(\bar{x}_{jk}-\bar{x}_k)^2}_{\text{Between-groups SS}} + \underbrace{\sum_{k=1}^{K}\sum_{i\in G_k}(x_{ijk}-\bar{x}_{jk})^2}_{\text{Within-groups SS}}$$

Decomposition of Variance

The sums-of-squares decompositions can be put in terms of **population** variances:

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}_{\text{Total variance}} = \underbrace{\sum_{k=1}^{K}\frac{n_k}{n}(\bar{x}_{jk}-\bar{x}_k)^2}_{\text{Between-groups variance}} + \underbrace{\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in G_k} n_k(x_{ijk}-\bar{x}_{jk})^2}_{\text{Within-groups variance}}$$

Formula from one-way analysis of variance (anova)

Decomposition of Variance

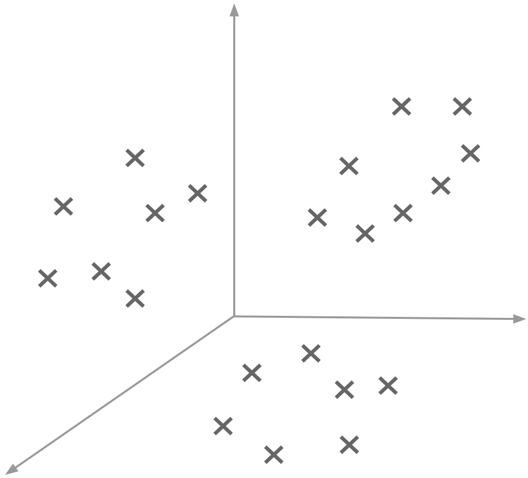
Alternatively, the sums-of-squares decompositions can also be put in terms of **sample** variances:

$$\mathsf{TSS} = \underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}_{\text{Total variance}} =$$

$$\underbrace{\sum_{k=1}^{K}\frac{n_k}{n}(\bar{x}_{jk}-\bar{x}_k)^2}_{\text{Between-groups variance}} + \underbrace{\frac{1}{n-1}\sum_{k=1}^{K}\sum_{i\in G_k}(n_k-1)(x_{ijk}-\bar{x}_{jk})^2}_{\text{Within-groups variance}}$$

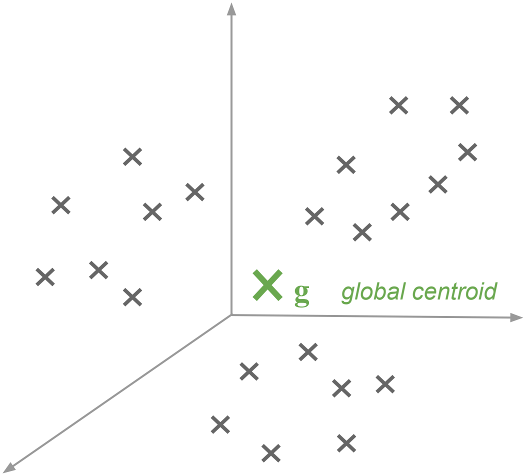
Geometric Perspective

Data as a cloud of points in p-dim space



Cloud of n points in p-dimensional space

Global centroid (center of gravity)



The *centroid* g is the point of averages

Global Centroid

The global centroid $\mathbf{g}$ is the point of averages which consists of the point formed with all the variable means:

$$\mathbf{g} = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]$$

where:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

If all variables are mean-centered, the centroid is the origin

$$\mathbf{g} = \underbrace{[0, 0, \dots, 0]}_{p \text{ times}}$$

Total Dispersion

Taking the global centroid as a point of reference, we can look at the amount of spread or dispersion in the data.

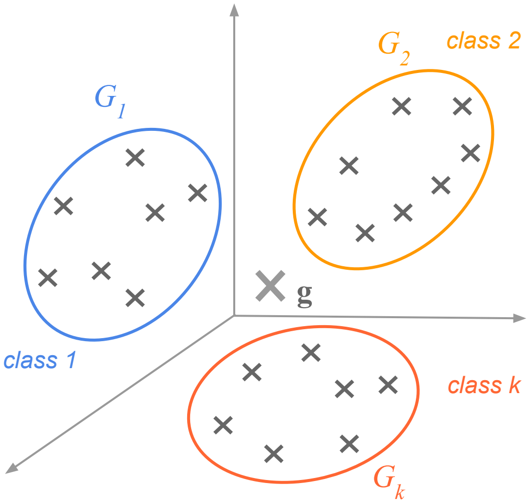
Assuming centered variables, a matrix of total dispersion is given by the *Total Sums of Squares* (TSS):

$$\mathsf{TSS} = \mathbf{X}^\mathsf{T}\mathbf{X}$$

Alternatively, we can get the variance-covariance matrix $\mathbf{V}$:

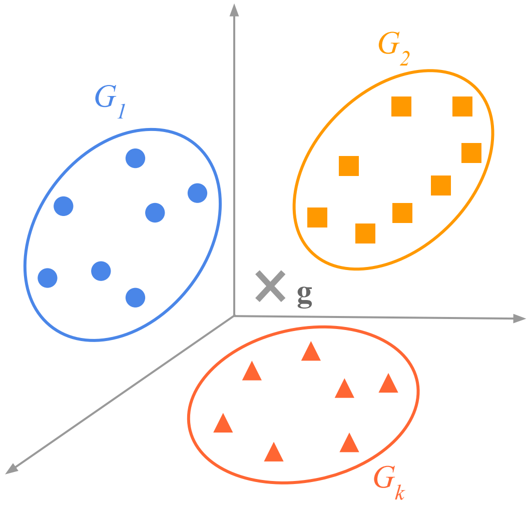
$$\mathbf{V} = \frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$$

Class (group) structure



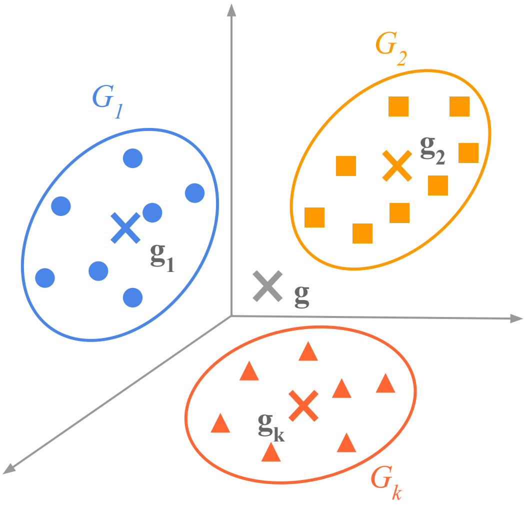
The objects are divided into classes or groups

Sub-cloud of points for each group



Each group G_k forms its own sub-cloud

Local or group centroids (one per class)



Each group G_k has its own centroid $\mathbf{g_k}$

Group Centroids

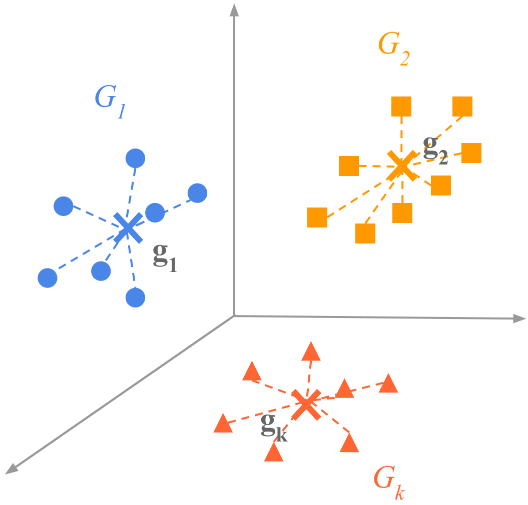
The group centroid $\mathbf{g_k}$ is the point of averages for those observations in group k:

$$\mathbf{g_k} = [\bar{x}_{1k}, \bar{x}_{2k}, \dots, \bar{x}_{pk}]$$

where:

$$\bar{x}_{jk} = \frac{1}{n_k} \sum_{i \in G_k} x_{ij}$$

Within-groups dispersion



We can focus on the dispersion within the clouds

Dispersion inside a group

Each group will have an associated spread or dispersion matrix given by a *Group Sums of Squares* (GSS):

$$\mathsf{GSS}_k = \mathbf{X_k}^\mathsf{T} \mathbf{X_k}$$

Equivalently, there is an associated variance matrix $\mathbf{W_k}$ for each group

$$\mathbf{W_k} = \frac{1}{n_k - 1} \mathbf{X_k}^\mathsf{T} \mathbf{X_k}$$

where $\mathbf{X_k}$ is the data matrix of the k-th group

Within-groups dispersion

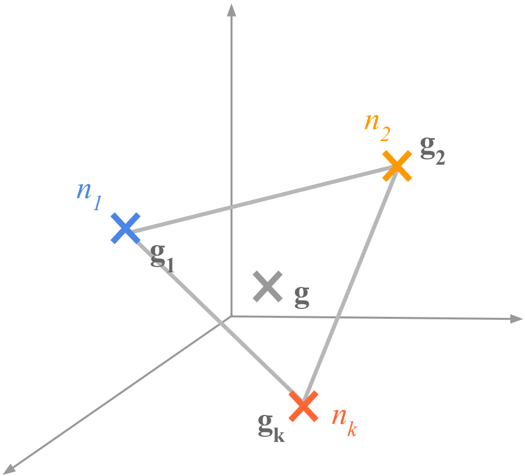
We can combine the groups dispersion to obtain a Within-groups Sums of Squares (WSS) matrix:

$$\mathsf{WSS} = \sum_{k=1}^{K} \mathbf{X_k}^\mathsf{T} \mathbf{X_k}$$

Likewise, we can combine the group variances $\mathbf{W_k}$ as a weighted average to get the Within-groups variance matrix $\mathbf{W}$:

$$\mathbf{W} = \sum_{k=1}^{K} \frac{n_k - 1}{n - 1} \mathbf{W_k}$$

Global and Group Centroids



What if we focus on just the centroids?

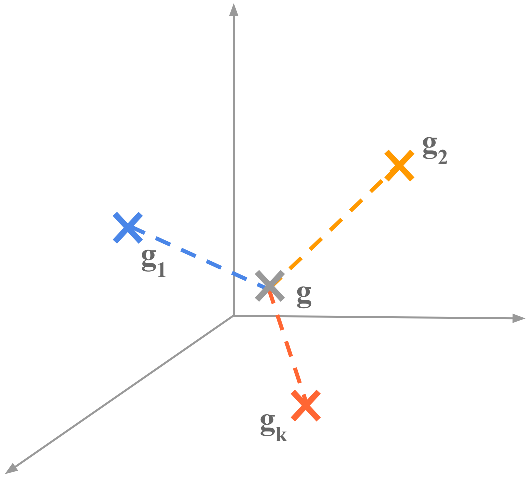
Global and Group Centroids

Note that the global centroid $\mathbf{g}$ can be expressed as a weighted average of the group centroids:

$$\mathbf{g} = \frac{n_1}{n}\mathbf{g_1} + \frac{n_2}{n}\mathbf{g_2} + \cdots + \frac{n_K}{n}\mathbf{g_K}$$

$$\mathbf{g} = \sum_{k=1}^{K} \left(\frac{n_k}{n}\right) \mathbf{g_k}$$

Between-groups dispersion



We can focus on the dispersion between the centroids

Dispersion between groups

Focusing on just the centroids, we can get its corresponding matrix of dispersion given by the *Between Sums of Squares* (BSS):

$$\text{BSS} = \sum_{k=1}^{K} (\mathbf{g_k} - \mathbf{g})(\mathbf{g_k} - \mathbf{g})^{\mathsf{T}}$$

Equivalently, there is an associated Between-groups variance matrix $\mathbf{B}$

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

Three types of Dispersions

Let's recap. We have three types of sums-of-squares matrices:

- TSS: Total Sums fo Squares
- WSS: Within-groups Sums fo Squares
- BSS: Between-groups Sums fo Squares

Alternatively, we also have three types of variance matrices:

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

Dispersion Decomposition

It can be shown (Huygens theorem) for both, sums-of-squares and variances, that the total dispersion (TSS or $\mathbf{V}$) can be decomposed as:

- $\text{TSS} = \text{BSS} + \text{WSS}$
- $\mathbf{V} = \mathbf{B} + \mathbf{W}$

Dispersion Decomposition

Let $\mathbf{X}$ be the $n \times p$ mean-centered matrix of predictors, and $\mathbf{Y}$ be the $n \times K$ dummy matrix of groups

- TSS $= \mathbf{X}^\mathsf{T}\mathbf{X}$
- BSS $= \mathbf{X}^\mathsf{T}\mathbf{Y}(\mathbf{Y}^\mathsf{T}\mathbf{Y})^{-1}\mathbf{Y}^\mathsf{T}\mathbf{X}$
- WSS $= \mathbf{X}^\mathsf{T}(\mathbf{I} - \mathbf{Y}(\mathbf{Y}^\mathsf{T}\mathbf{Y})^{-1}\mathbf{Y}^\mathsf{T})\mathbf{X}$

References

- **Principles of Multivariate Analysis: A User's Perspective** by W.J. Krzanowski (1988). *Chapter 11: Incorporating group structure: descriptive methods*. Wiley.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 11: Classification and prediction methods*.
- **Multivariate Analysis** by Maurice Tatsuoka (1988). *Chapter 7: Discriminant Analysis and Canonical Correlation*.

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.