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*.

Classification Idea

- 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

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

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

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

Caveat: messy notation

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

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

I will use the symbol x_{ijk} to represent the i-th observation, of the j-th variable, in the k-th group.

In turn, I'll use x_{jk} to represent values of the j-th variable in group k

I hope this doesn't create a lot of confussion

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

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

Assuming centered variables (mean = 0)

$$TSS_j = \frac{1}{n} \mathbf{x_j}^\mathsf{T} \mathbf{x_j}$$

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 variances:

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

What's this?

Decomposition of Variance

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

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

Formula from one-way analysis of variance (anova)

Iris Data

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

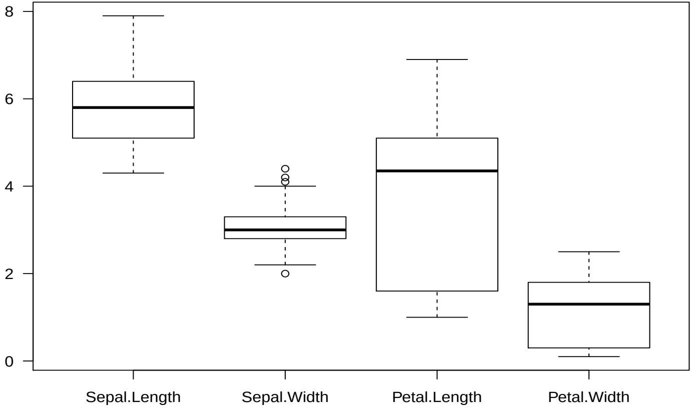
```
head(iris)

  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
```

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                  
```

Predictors in iris



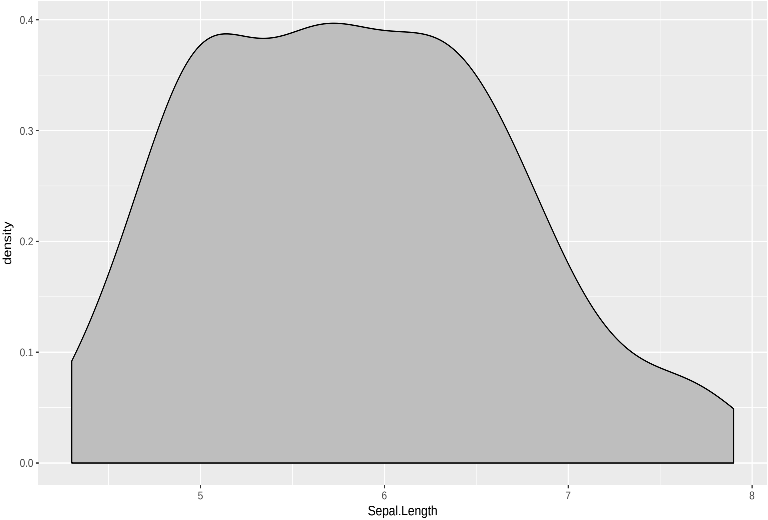

```r
boxplot(iris[,1:4], las = 1)
```

Toy demo

Let's check the formula $TSS = BSS + WSS$

For illustration purposes let's focus on predictor `Sepal.Length`, and response `Species`

Exploring Sepal.Length




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

TSS for Sepal.Length

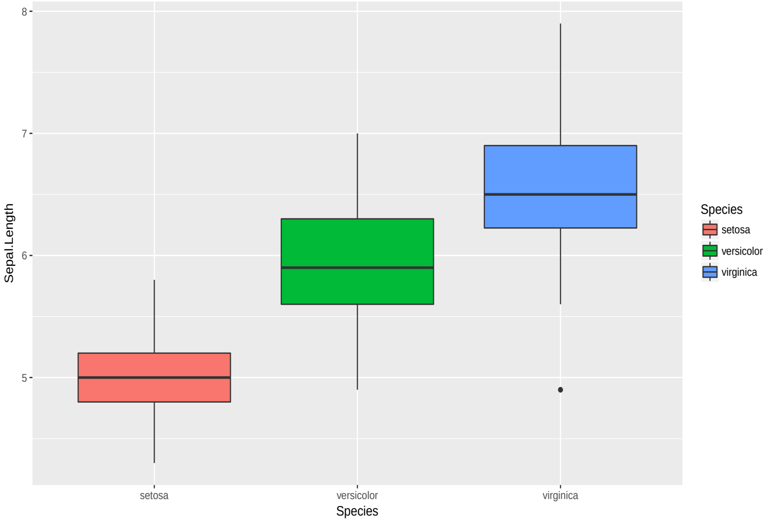
```r
x = iris$Sepal.Length

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

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

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

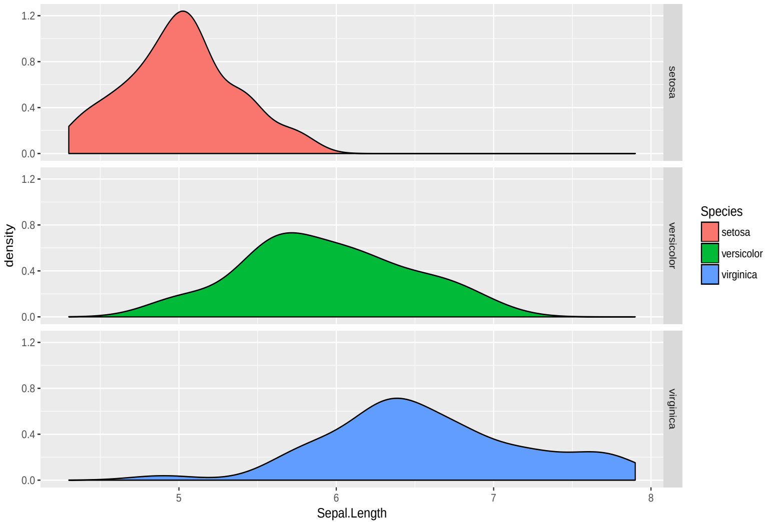
```
## [1] 102.1683
```

Let's consider the group structure




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

Sepal.Length by groups




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

Density curves

```
# group means
group_means <- tapply(x, iris$Species, mean)
group_means

##     setosa versicolor  virginica
##      5.006      5.936      6.588

group_num <- c(50, 50, 50)
```

Between and Within groups sum-of-squares

```
# between sums-of-squares
BSS = sum(group_num * (group_means - x_bar)^2)
BSS
```

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

```
# within sums-of-squares
w1 = sum((x[1:50] - group_means[1])^2)
w2 = sum((x[51:100] - group_means[2])^2)
w3 = sum((x[101:150] - group_means[3])^2)
WSS = (w1 + w2 + w3)
WSS
```

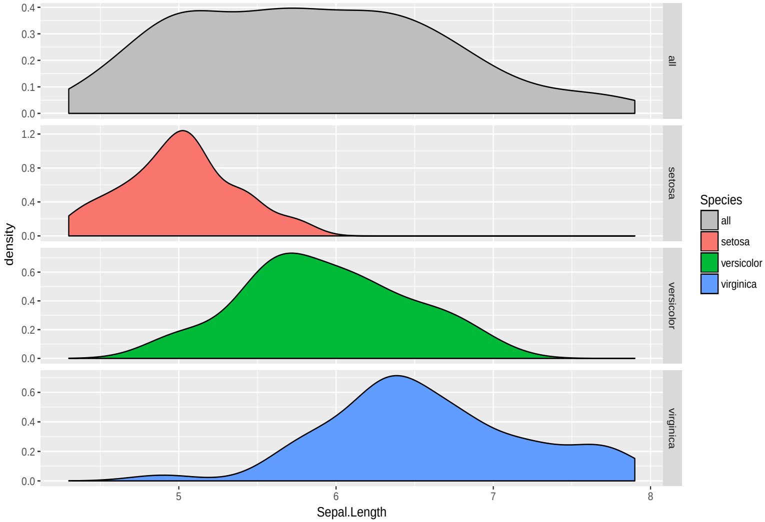
```
## [1] 38.9562
```

Dispersion Decomposition

Let's check the decomposition: $TSS = BSS + WSS$

```r
# the total sums-of-squares
TSS

## [1] 102.1683

# is equal to the sum of Between-groups SS
# plus the Within-groups SS
BSS + WSS

## [1] 102.1683
```

Dispersion in `Sepal.Length`



Decomposition of Variance

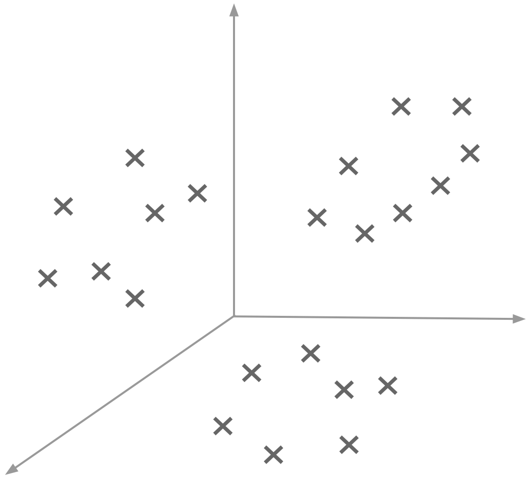
Variance Decomposition for one variable:

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

Let's see how this idea gets extended when we have more than one variable.

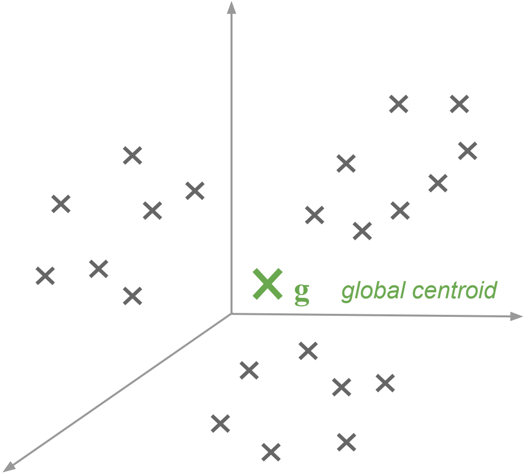
Geometric Perspective

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.

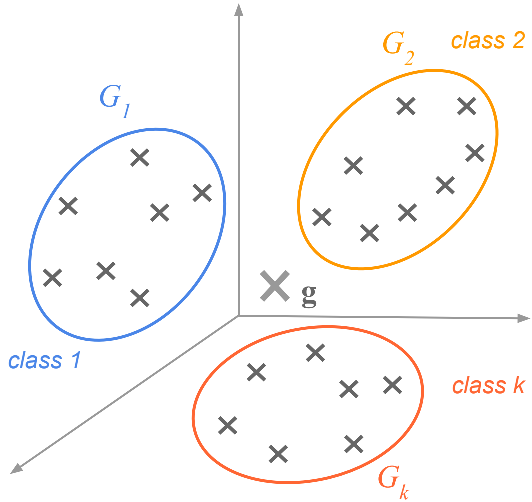
Assuming centered variables, a matrix of total dispersion is given by the *Total Sums of Squares* (TSS):

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

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

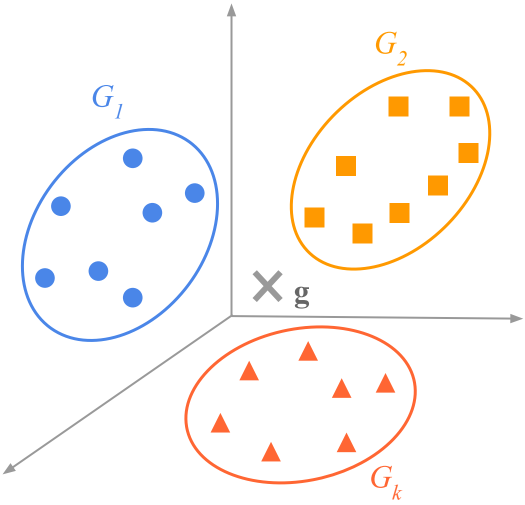
$$\mathbf{V} = \frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$$

Class (group) structure



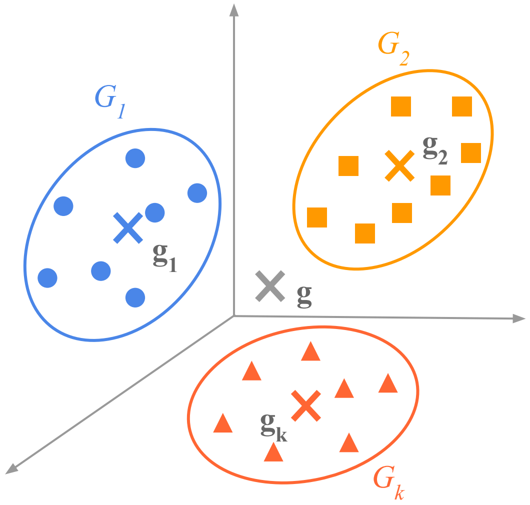
The objects are divided into classes or groups

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

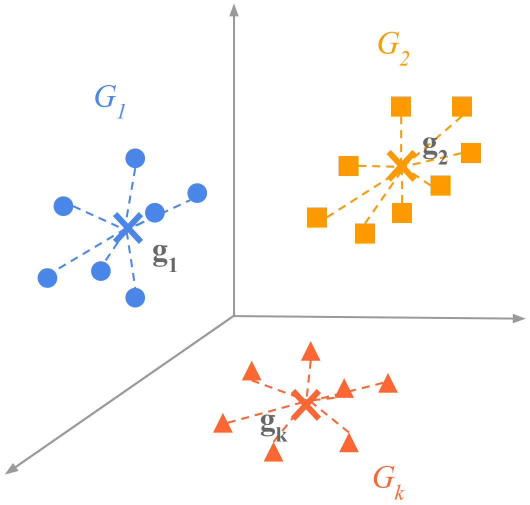
The group centroid $\mathbf{g_k}$ is the point of averages for those observations in group k:

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

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

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

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

Within-groups dispersion

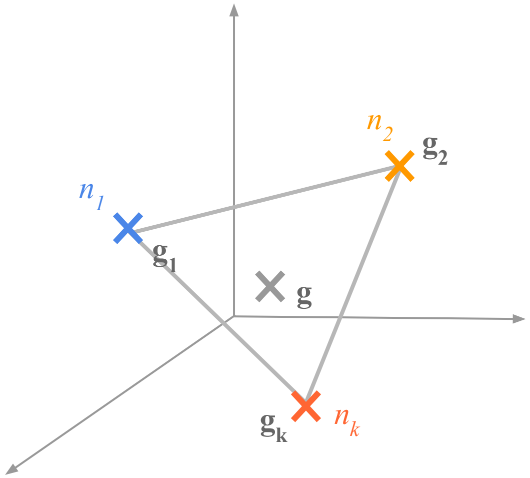
We can combine the groups dispersion to obtain a Within-groups Sums of Squares (WSS) matrix:

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

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

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

Global and Group Centroids



What if we focus on just the centroids?

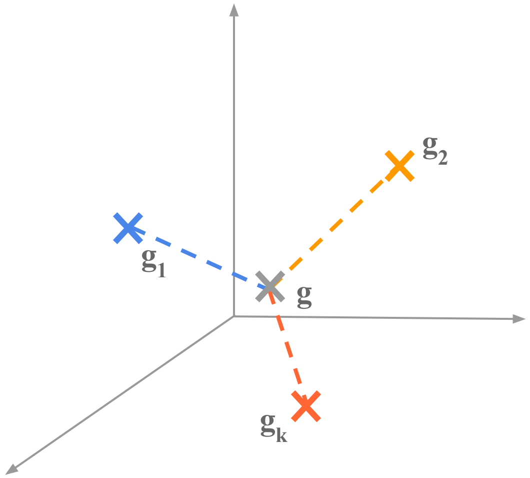
Global and Group Centroids

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):

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

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

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

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

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:

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

References

- **The use of multiple measurements in taxonomic problems** by R.A. Fisher (1936). *Annals of Eugenics, 7, 179-188*.
- **Principles of Multivariate Analysis: A User's Perspective** by W.J. Krzanowski (1988). *Chapter 11: Incorporating group structure: descriptive methods*. Wiley.
- **On the generalized distance in statistics**. by P.C. Mahalanobis (1936). *Proceedings of the National Institute of Science, India, 12, 49-55*.
- **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*.
- **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.