UNIVERSIDADE FEDERAL FLUMINENSE

Programa de Mestrado e Doutorado em Engenharia de Produção

Multivariate Data Analysis

Discriminant Analysis

Professor: Valdecy Pereira. D. Sc.

email: valdecy.pereira@gmail.com

Outline

1. DA & MDA
2. Discriminant Functions
3. Assumptions
4. GOF
5. Visualization
6. Bibliography

MVDA - *Discriminant Analysis*

The **DA** (Discriminant Analysis) is a technique used to analyze the relationship between a dependent non-metric variable and metric or dichotomous independent variables. The **DA** uses the metric independent variables to be able to distinguish clusters (groups or categories) of non-metric dependent variable.

When the non-metric dependent variable has 2 clusters (groups or categories) we use the Simple **DA**, and when it has 3 or more clusters (groups or categories) we use the **MDA** (Multiple Discriminant Analysis).

MVDA - *Discriminant Analysis*

The **DA** generates m discriminant functions ($\boldsymbol{Z_m}$) - linear combinations of the independent variables - that enhance the discrimination of clusters. Not every discriminant function is significant, usually the first two are the most important and a discriminant function is always orthogonal to the previous ones. The maximum of discriminant functions is calculated by:

$$\min = \{g - 1; k\}$$

g = Number of clusters;
k = Total number of independent variables.

The discriminant function is defined by:

$$Z_m = B_{m0} + B_{m1}X_1 + \cdots + B_{mk}\,X_k$$

B_0 = Constant;
X_k = Independent variable k (Predictor k);
B_{mk} = m-th discriminant coefficient that maximizes the distance between the means of clusters and minimizes the variance within the same.

Assumptions

MVDA - *Discriminant Analysis*

ASSUMPTIONS

- Multivariate Normality: Statistics are improved if the dataset has Multivariate Normal Distribution. Relaxation – Most of the variables in the dataset are Normally Distributed (Univariate Normality).
- Multicolinearity: Statistics are improved if the independent variables are not significantly correlated. High correlation values indicates that the independent variables are redundant and some of them should be discarded.
- Relationship between cases and clusters: $\frac{n}{g} \geq 20 \rightarrow$ at least 20 cases (n) for each cluster (g). Relaxation – 20 cases (n) for each independent variable.
- Homoscedastic Clusters: Statistics are improved if clusters have similar variances.
- Outliers: Statistics are not robust in the presence of discrepant values.

MVDA - *Discriminant Analysis*

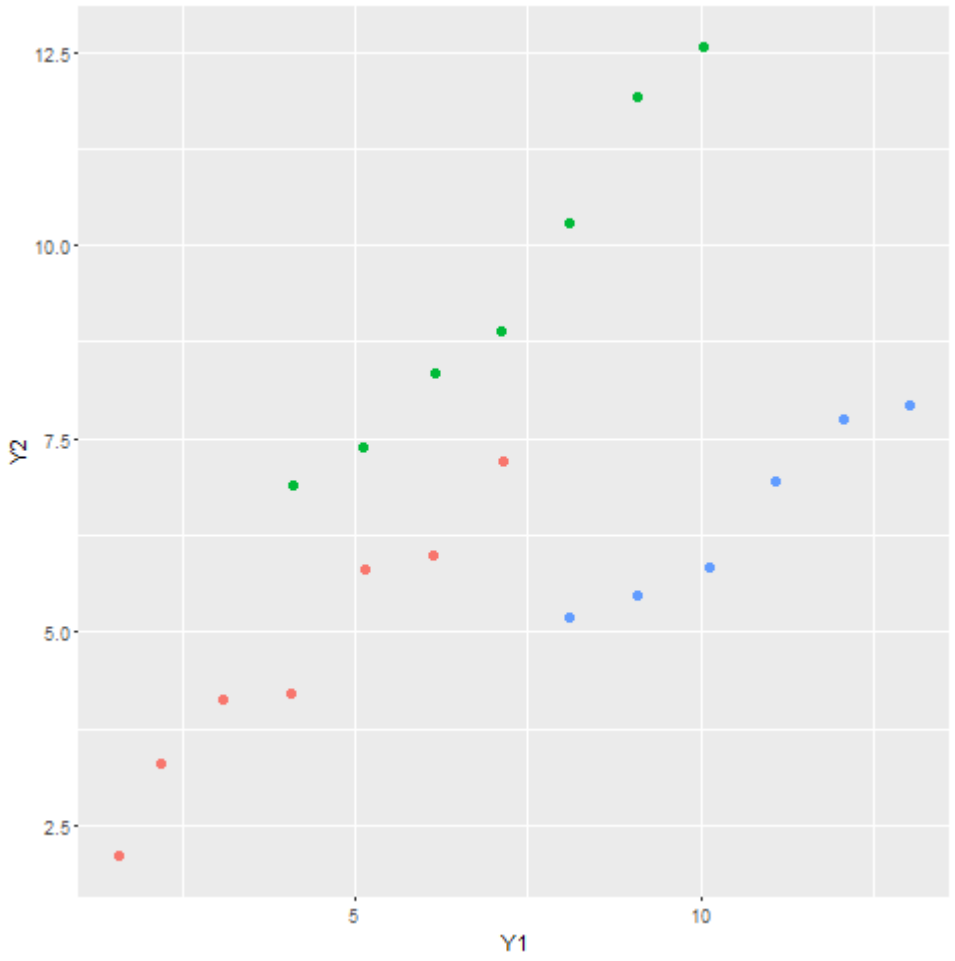
X1	X2	Cluster
1.57	2.1	1
2.16	3.3	1
3.08	4.12	1
4.07	4.2	1
5.13	5.8	1
6.12	6	1
7.14	7.2	1
4.1	6.9	2
5.11	7.38	2
6.16	8.34	2
7.1	8.88	2
8.09	10.3	2
9.08	11.92	2
10.05	12.56	2
8.09	5.2	3
9.07	5.48	3
10.13	5.84	3
11.1	6.96	3
12.06	7.74	3
13.04	7.92	3

In order to explain a **DA** approach, the following dataset will be used: The simulated dataset of 20 observations and 3 clusters.

```
# Graph
library("ggplot2")
ggplot(my_data, aes(x = Y1, y = Y2)) + geom_point(aes(colour = ifelse(my_data[,1] == 1, "blue", ifelse(my_data[,1] == 2,"green", "red" ))), size = 2) + theme(legend.position = "none")
```

MVDA - *Discriminant Analysis*

X1	X2	Cluster
1.57	2.1	1
2.16	3.3	1
3.08	4.12	1
4.07	4.2	1
5.13	5.8	1
6.12	6	1
7.14	7.2	1
4.1	6.9	2
5.11	7.38	2
6.16	8.34	2
7.1	8.88	2
8.09	10.3	2
9.08	11.92	2
10.05	12.56	2
8.09	5.2	3
9.07	5.48	3
10.13	5.84	3
11.1	6.96	3
12.06	7.74	3
13.04	7.92	3

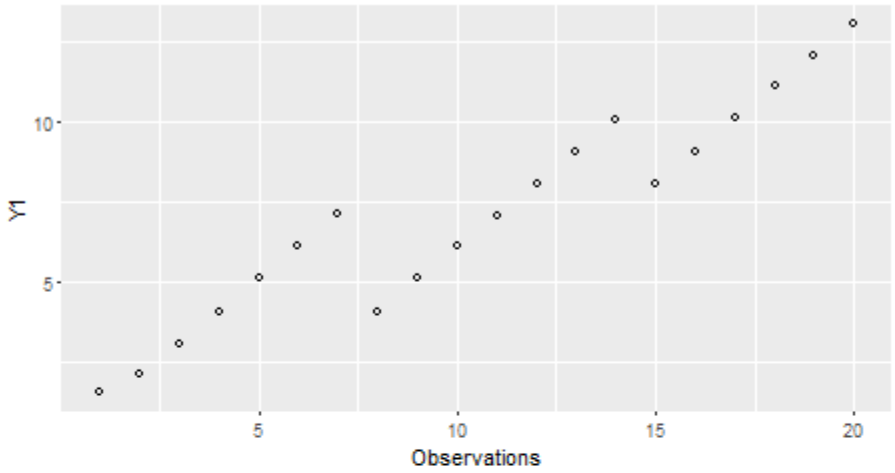


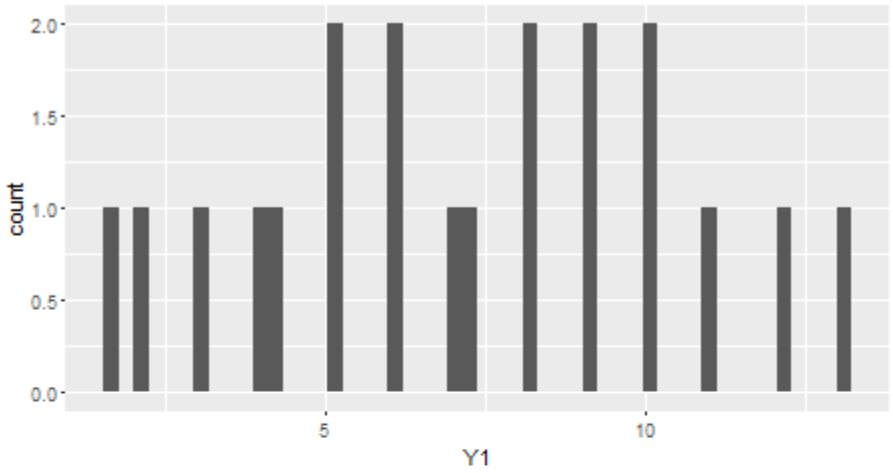
Assumptions – Univariate Normality

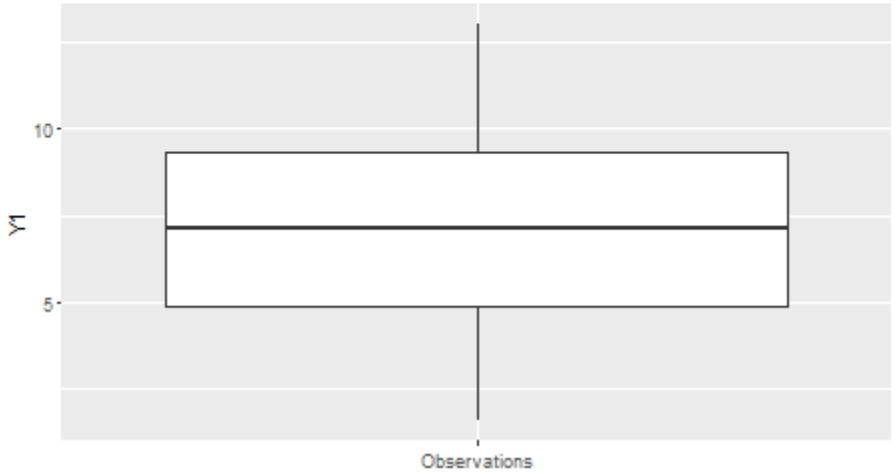
MVDA - *Discriminant Analysis*

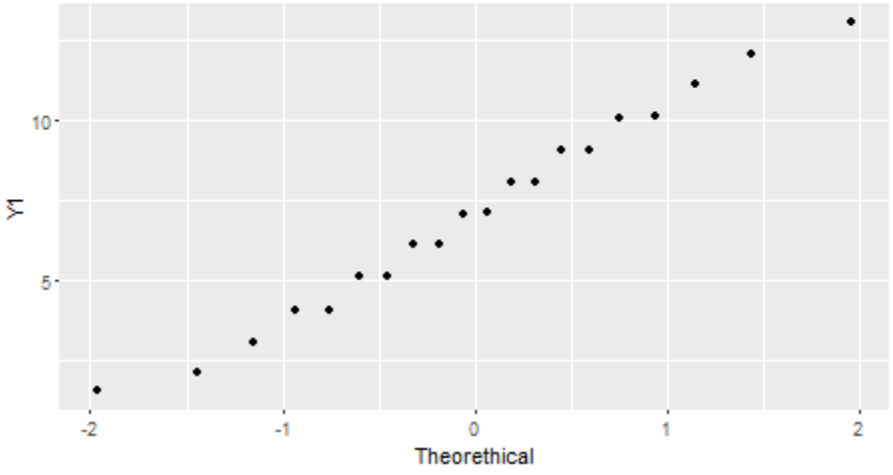
```
library("ggplot2")
ggplot(data = my_data, aes(x = 1:20, y = my_data$Y1)) + geom_point(shape = 1) + labs(x = "Observations", y = "Y1")
ggplot(data = my_data, aes(x = "Observations", y = my_data$Y1)) + geom_boxplot() + theme(axis.title.x = element_blank()) + labs(y = "Y1")
ggplot(my_data, aes(my_data$Y1)) + geom_histogram(bins = 50) + labs(x = "Y1")
ggplot(data = my_data, aes( sample = my_data$Y1)) + stat_qq()+ xlab("Theorethical") + ylab("Y1")
ggplot(data = my_data, aes(x = 1:20, y = my_data$Y2)) + geom_point(shape = 1) + labs(x = "Observations", y = "Y2")
ggplot(data = my_data, aes(x = "Observations", y = my_data$Y2)) + geom_boxplot() + theme(axis.title.x = element_blank()) + labs(y = "Y2")
ggplot(my_data, aes(my_data$Y2)) + geom_histogram(bins = 50) + labs(x = "Y2")
ggplot(data = my_data, aes( sample = my_data$Y2)) + stat_qq()+ xlab("Theorethical") + ylab("Y2")
```

MVDA - *Discriminant Analysis*

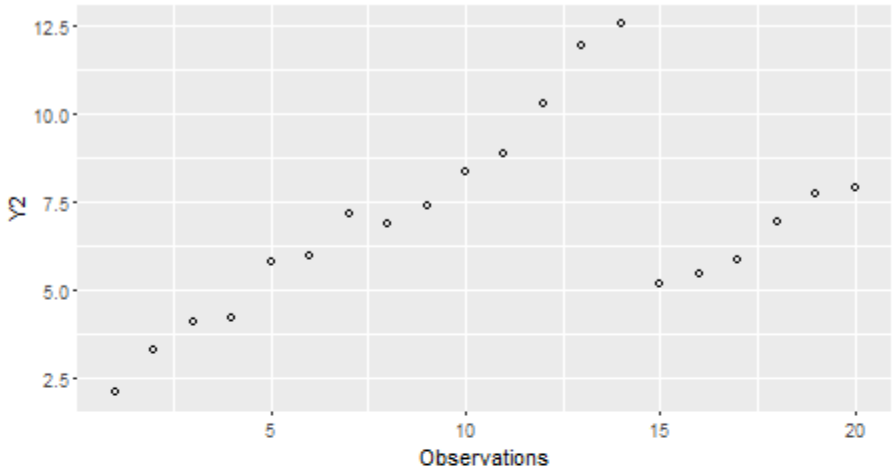


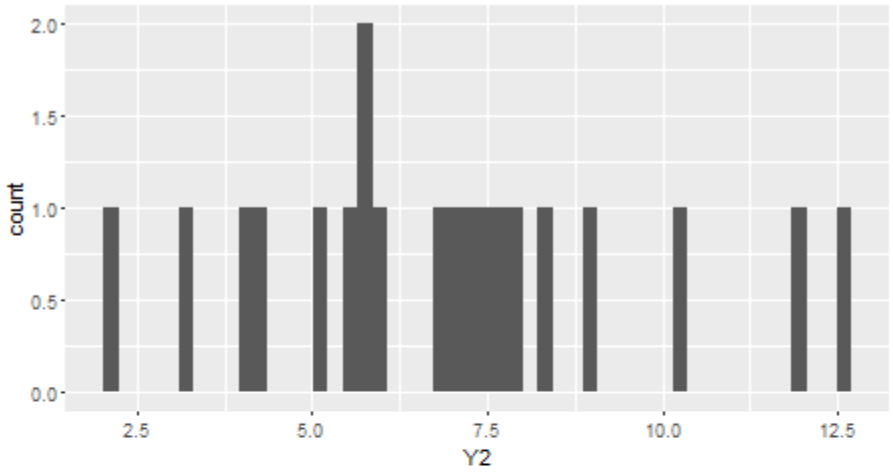


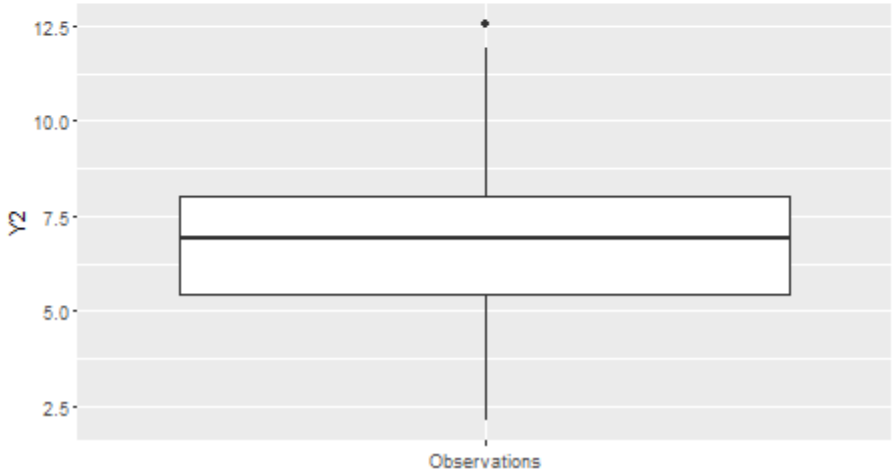


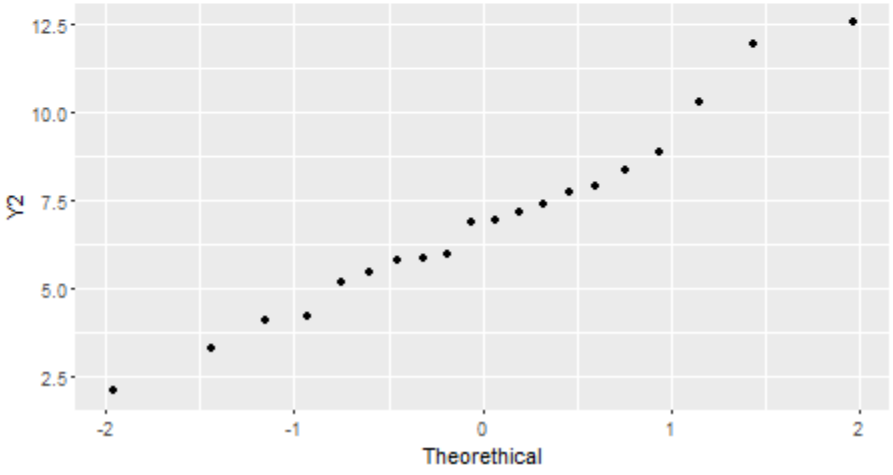


MVDA - *Discriminant Analysis*









MVDA - *Discriminant Analysis*

```
# Univariate Normality
shapiro.test(my_data$Y1)

Shapiro-Wilk normality test

data:  my_data$Y1
W = 0.97936, p-value = 0.9255

shapiro.test(my_data$Y2)

Shapiro-Wilk normality test

data:  my_data$Y2
W = 0.97184, p-value = 0.7931
```

Assumptions – Multivariate Normality

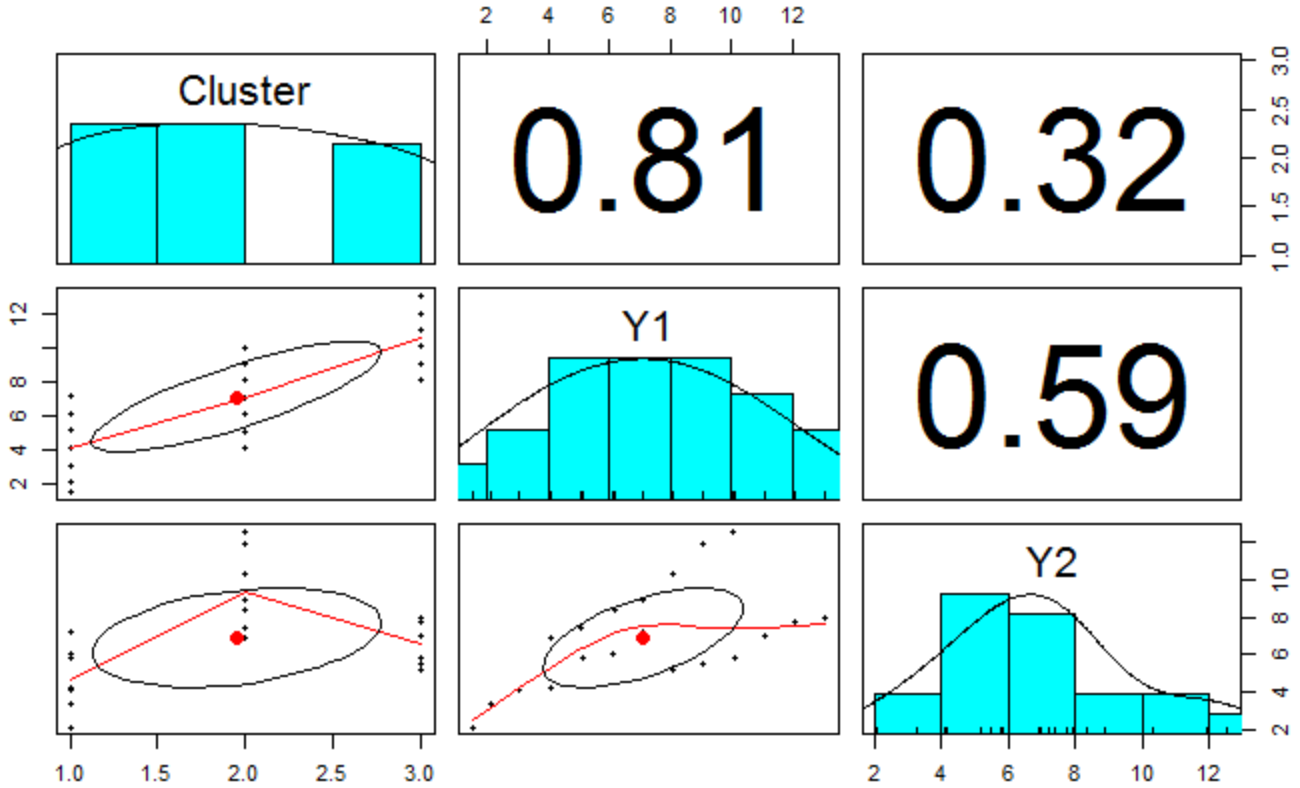
MVDA - *Discriminant Analysis*

```
# Multivariate Normality
library("MVN")
mardiaTest(my_data, qqplot = FALSE)

 Mardia's Multivariate Normality Test
---------------------------------------
   data : my_data

   g1p              : 2.261483
   chi.skew         : 7.538276
   p.value.skew     : 0.6738367

   g2p              : 10.20148
   z.kurtosis       : -1.95899
   p.value.kurt     : 0.050114

   chi.small.skew   : 9.335865
   p.value.small    : 0.5005529

   Result           : Data are multivariate normal.
---------------------------------------
```

Assumptions – Multicolinearity

MVDA - *Discriminant Analysis*

```
# Multicolinearity
library ("psych")
pairs.panels(my_data)
```



Assumptions – Ratio

MVDA - *Discriminant Analysis*

```
#  Relationship between cases and clusters
20/3

[1] 6.666667

#  Relationship between cases and independent variables
20/2

[1] 10
```

Assumptions – Homoscedasticity

MVDA - *Discriminant Analysis*

- Box's M Test: This test checks whether each cluster variance-covariance matrix is equal. by the following hypothesis test:

$$H_0: The\ variance\text{–}Covariance\ Matrices\ are\ equal$$
$$H_a: The\ variance\text{–}Covariance\ Matrices\ are\ not\ equal$$

```
# Box's M Test
library("biotools")
boxM(my_data[,2:3], my_data$Cluster)

Box's M-test for Homogeneity of Covariance Matrices

data:  my_data[, 2:3]

Chi-Sq (approx.) = 9.3375, df = 6, p-value = 0.1555
```

Discriminant Analysis

MVDA - *Discriminant Analysis*

```
#  Discriminant Analysis
library("MASS")
lda <- lda(Cluster ~ ., data = my_data)


# Prediction
lda_values <- predict(lda)

# Z Values
lda_values$x
```

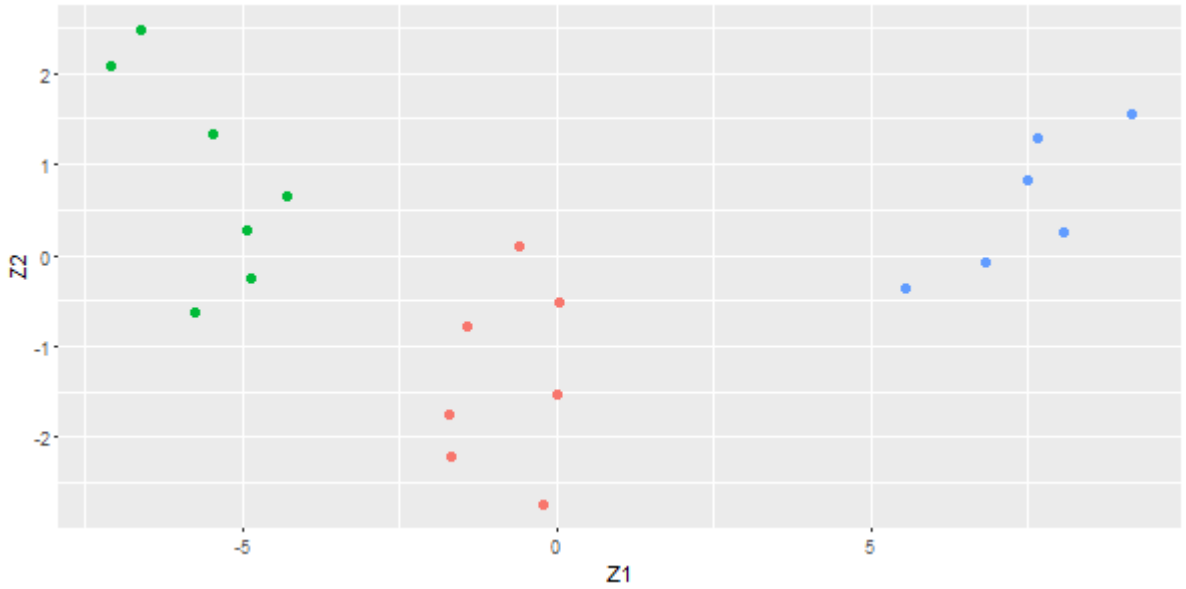
MVDA - *Discriminant Analysis*

Z1	Z2	Cluster
-0.22312	-2.73056	1
-1.69101	-2.21076	1
-1.70495	-1.7505	1
0.01658	-1.52257	1
-1.41749	-0.77394	1
0.04433	-0.50603	1
-0.60062	0.10119	1
-5.76932	-0.61681	2
-4.8752	-0.25154	2
-4.94333	0.28181	2
-4.31301	0.65284	2
-5.4915	1.32726	2
-7.10283	2.06833	2
-6.63152	2.47878	2
5.54588	-0.37209	3
6.81544	-0.07955	3
8.06496	0.2559	3
7.49746	0.8263	3
7.64664	1.28137	3
9.13261	1.54058	3

```
# Z Graph
ggplot(my_data, aes(x = lda_values$x[,1], y = lda_values$x[,2])) +
geom_point(aes(colour = ifelse(my_data[,1] == 1, "blue",
ifelse(my_data[,1] == 2,"green", "red" ))), size = 2) +
theme(legend.position = "none") + xlab("Z1") + ylab("Z2")
```

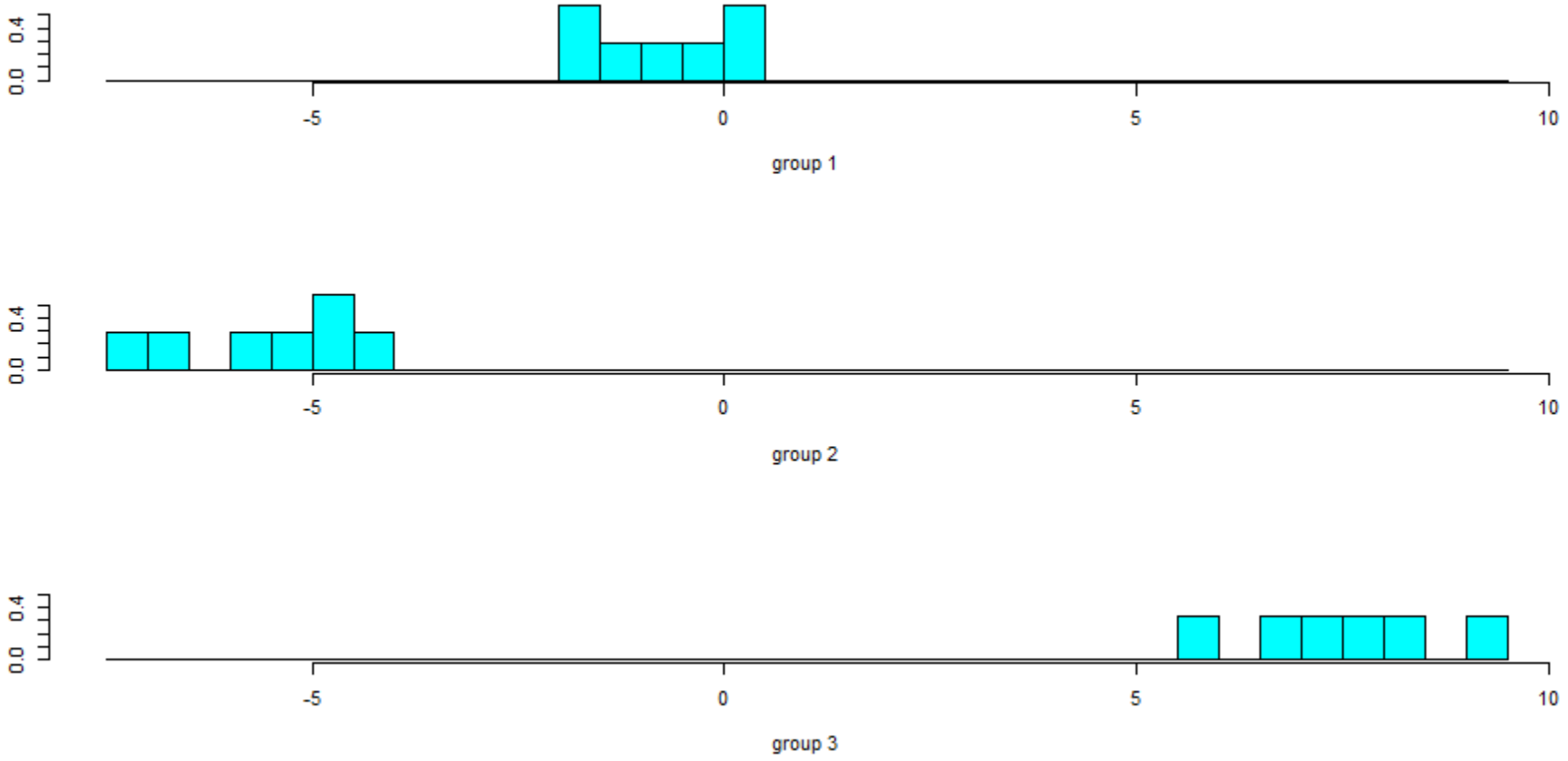
MVDA - *Discriminant Analysis*

Z1	Z2	Cluster
-0.22312	-2.73056	1
-1.69101	-2.21076	1
-1.70495	-1.7505	1
0.01658	-1.52257	1
-1.41749	-0.77394	1
0.04433	-0.50603	1
-0.60062	0.10119	1
-5.76932	-0.61681	2
-4.8752	-0.25154	2
-4.94333	0.28181	2
-4.31301	0.65284	2
-5.4915	1.32726	2
-7.10283	2.06833	2
-6.63152	2.47878	2
5.54588	-0.37209	3
6.81544	-0.07955	3
8.06496	0.2559	3
7.49746	0.8263	3
7.64664	1.28137	3
9.13261	1.54058	3



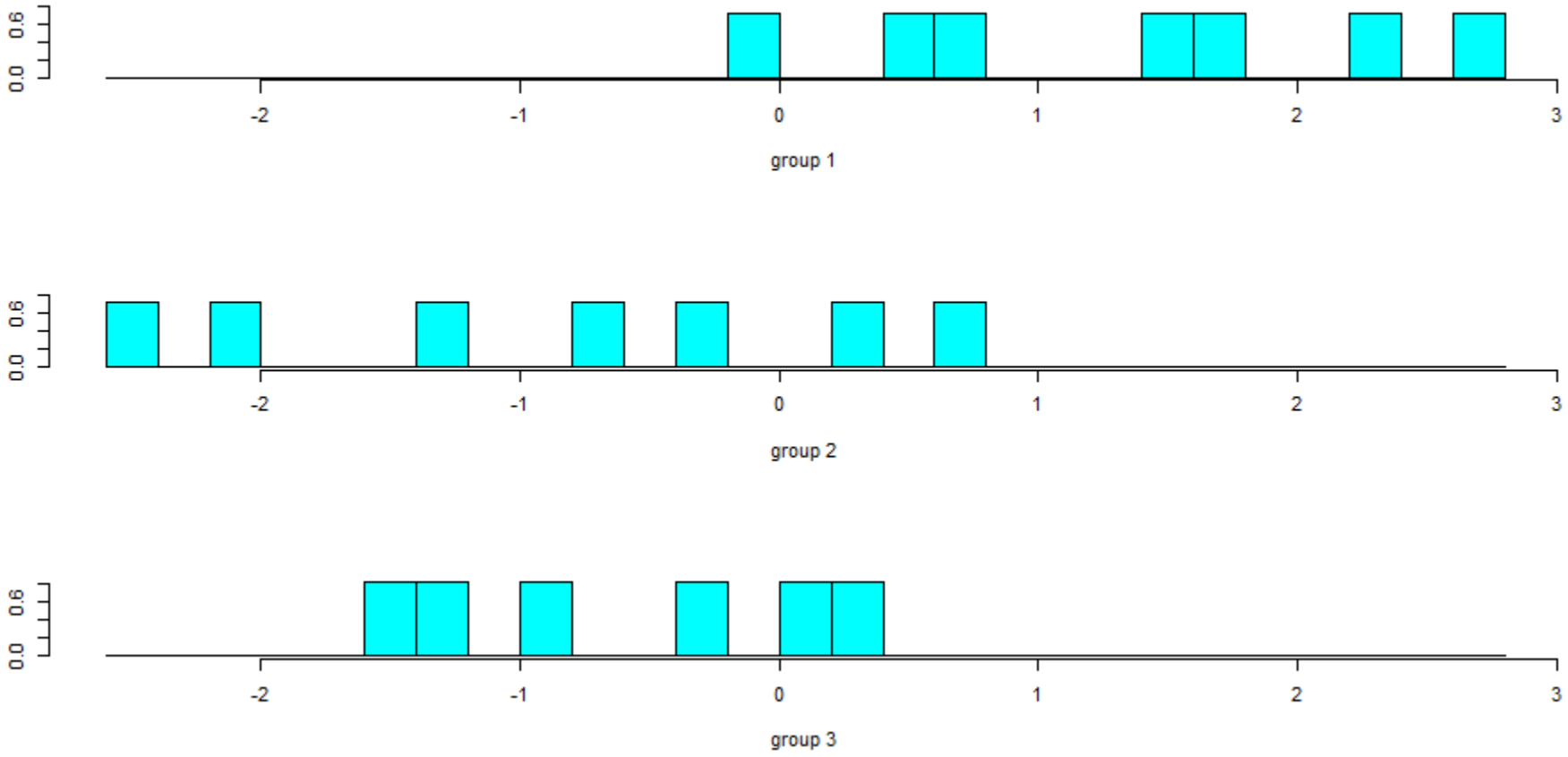
MVDA - *Discriminant Analysis*

```
# Cluster Histograms
ldahist(data = lda_values$x[,1], g = my_data$Cluster)
```



MVDA - *Discriminant Analysis*

```
# Cluster Histograms
ldahist(data = lda_values$x[,2], g = my_data$Cluster)
```



Goodness of Fit

MVDA - *Discriminant Analysis*

DIAGNOSIS

- U-Statistic - It varies between 0 and 1, and for each independent variable it verifies the existence of differences between clusters mean. Values closer to 0, indicates that the independent variable is very discriminant. Its value can be transformed to known distribution (Distribution F), and the following hypothesis can then tested:

$$H_0: The\ average\ of\ clusters\ is\ equal$$
$$H_a: The\ average\ of\ the\ clusters\ is\ not\ equal$$

```
# U-Statistics
library("DiscriMiner")
discPower(my_data[,2:3], my_data$Cluster)
```

	correl_ratio	wilks_lambda	F_statistic	p_value
Y1	0.6534675	**0.3465325**	16.02872	**0.000122412**
Y2	0.6011284	**0.3988716**	12.81011	**0.000404652**

- The Correlation Ratio measures the association level between the independent variable and the dependent variable.

MVDA - *Discriminant Analysis*

DIAGNOSIS

Eigenvalue ($\boldsymbol{\lambda}_z$): The eigenvalue indicates for each discriminant function the total percentage of explained variance, that is, how important is the function to discriminate clusters. It can be calculated as:

$$\boldsymbol{\lambda}_z = \frac{SS_{bz}}{SS_{wz}}$$

```
# Eigenvalue
eigen_values <- betweenSS(lda_values$x[,1:2], my_data$Cluster)/withinSS(lda_values$x[,1:2], my_data$Cluster)
```

	LD1	LD2
LD1	**32.71774**	0.50000
LD2	0.50000	**1.15487**

```
# Proportion of trace - The percentage of variance achieved by each discriminant function.
p_ev_1 <- eigen_values[1]/(eigen_values[1] + eigen_values[4])

[1] 0.9659055

p_ev_2 <- eigen_values[4]/(eigen_values[1] + eigen_values[4])

[1] 0.03409449
```

MVDA - *Discriminant Analysis*

DIAGNOSIS

- Canonical Correlation (R_{cz} ou R^{*z}): The canonical correlation measures the correlation between each discriminant function and the clusters. The closer to 1 is the value of canonical correlation, the more discriminating the function is. The squared value of the canonical correlation is analogous to the coefficient of determination and indicates the total variance explained by the discriminant function. It is calculated as:

$$R_{cz} = \sqrt{\frac{SS_{bz}}{SS_{wz} + SS_{bz}}}$$

MVDA - *Discriminant Analysis*

```
# Canonical Correlation
rc1 <- ((betweenSS(lda_values$x[,1:2], my_data$Cluster)[1])/(withinSS(lda_values$x[,1:2], my_data$Cluster )[1] +
betweenSS(lda_values$x[,1:2], my_data$Cluster)[1]))^(1/2)

[1] 0.9850594

rc2 <- ((betweenSS(lda_values$x[,1:2], my_data$Cluster)[4])/(withinSS(lda_values$x[,1:2], my_data$Cluster)[4] +
betweenSS(lda_values$x[,1:2], my_data$Cluster)[4]))^(1/2)

[1] 0.7320757
```

MVDA - *Discriminant Analysis*

DIAGNOSIS

χ^2 Test or Lambda Wilks (Λ_z): The χ^2 Test globally tests the discriminant functions by the following hypothesis, with k (g-1) degrees of freedom:

$$H_0: The\ function\ (s)\ is\ (are)\ not\ significant\ to\ discriminate\ clusters$$
$$H_a: The\ function\ (s)\ is\ (are)\ significant\ \ to\ discriminate\ clusters$$

The total rejection of the hypothesis test indicates that at least the first discriminant function is significant.

```
# Lambda Wilks
library("rrcov")
Wilks.test(Cluster ~ ., data = my_data, method = "c")

One-way MANOVA (Bartlett Chi2)

data:  x
Wilks' Lambda = 0.013763, Chi2-Value = 70.715, DF = 4.000, p-value = 1.599e-14
```

Classification

MVDA - *Discriminant Analysis*

The classification process is done scoring each object in all clusters. The object is allocated to the cluster that has the highest score.
linDA(my_data[,2:3], my_data$Cluster)$scores

X1	X2	Cluster	1	2	3	Previsto
1.57	2.1	1	-0.47	-20.15	-34.41	1
2.16	3.3	1	3.88	-7.62	-41.16	1
3.08	4.12	1	5.02	-5.41	-39.26	1
4.07	4.2	1	1.93	-16.25	-27.71	1
5.13	5.8	1	6.76	-2.90	-33.27	1
6.12	6	1	4.32	-11.77	-23.14	1
7.14	7.2	1	7.14	-4.52	-24.47	1
4.1	6.9	2	16.34	27.88	-59.28	2
5.11	7.38	2	15.33	23.38	-52.22	2
6.16	8.34	2	16.76	26.30	-50.33	2
7.1	8.88	2	16.32	23.66	-44.86	2
8.09	10.3	2	20.43	34.90	-49.17	2
9.08	11.92	2	25.62	49.43	-55.85	2
10.05	12.56	2	25.61	48.06	-51.18	2
8.09	5.2	3	-6.99	-49.15	11.18	3
9.07	5.48	3	-8.97	-56.57	20.23	3
10.13	5.84	3	-10.80	-63.66	29.35	3
11.1	6.96	3	-8.23	-57.12	28.33	3
12.06	7.74	3	-7.45	-56.05	31.21	3
13.04	7.92	3	-9.96	-65.13	41.45	3

Accuracy

MVDA - *Discriminant Analysis*

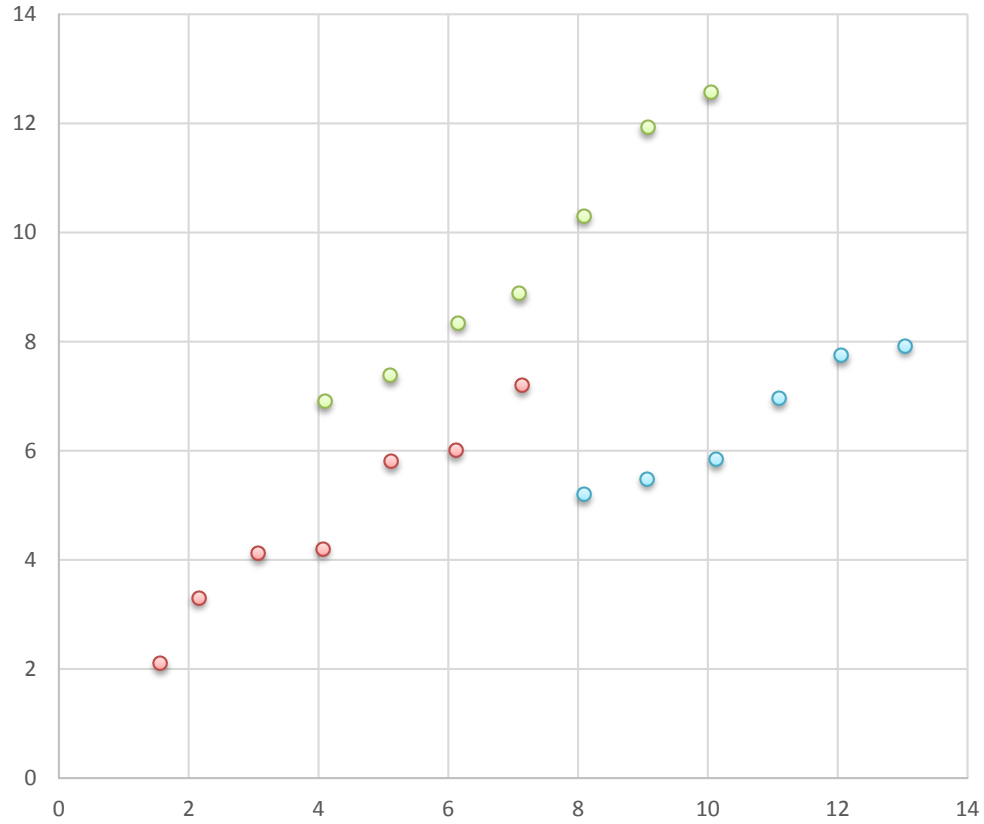
```
# Accuracy of the prediction
class_table   <- table(my_data$Cluster, lda_values$class)

   1 2 3
 1 7 0 0
 2 0 7 0
 3 0 0 6

class_table_p <- (prop.table(class_table, 1))

   1 2 3
 1 1 0 0
 2 0 1 0
 3 0 0 1

acc          <- sum(diag(prop.table(class_table)))

[1] 1
```

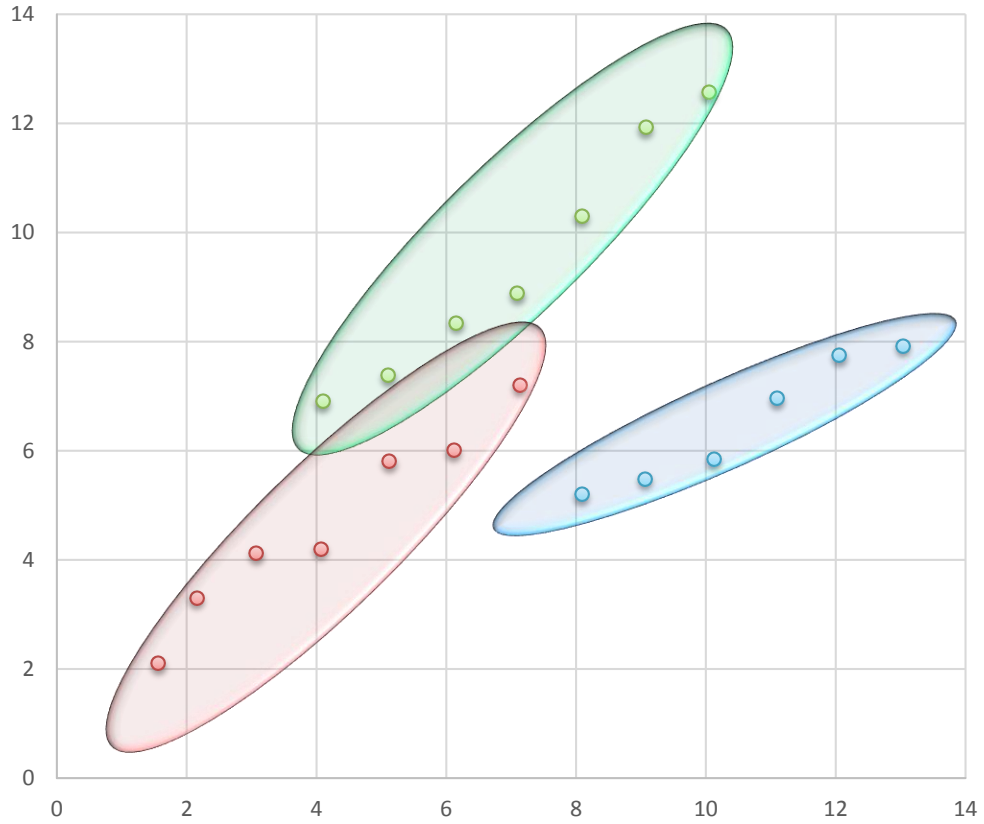
Prediction

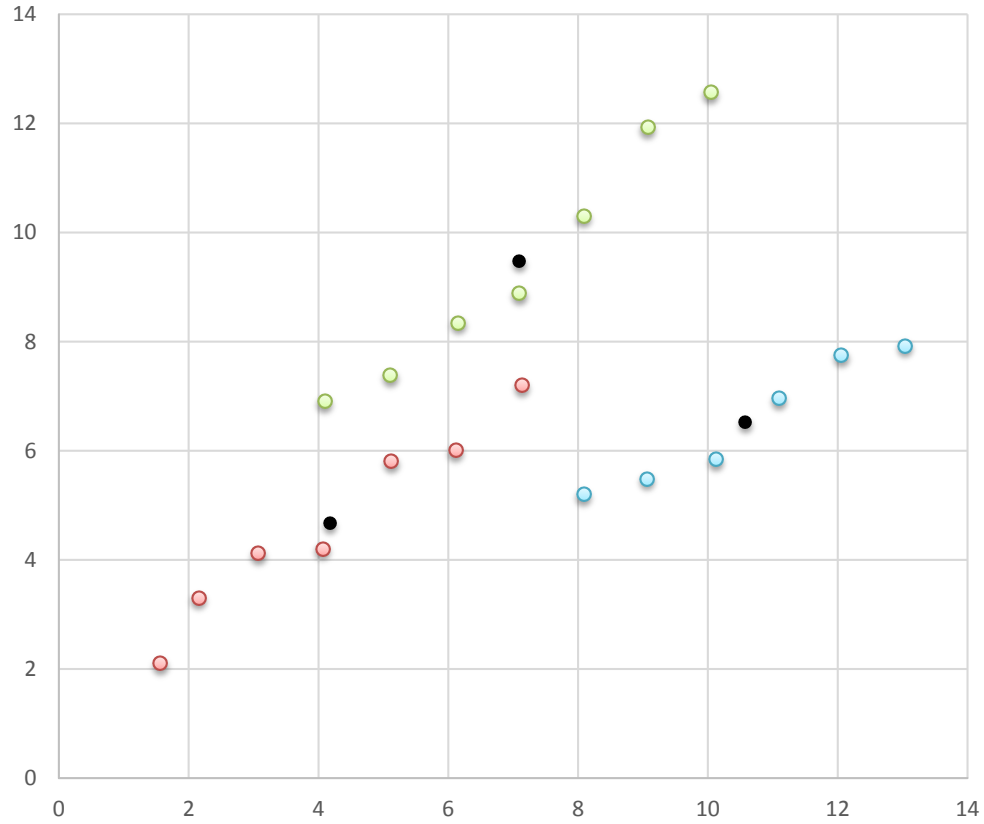
MVDA - *Discriminant Analysis*

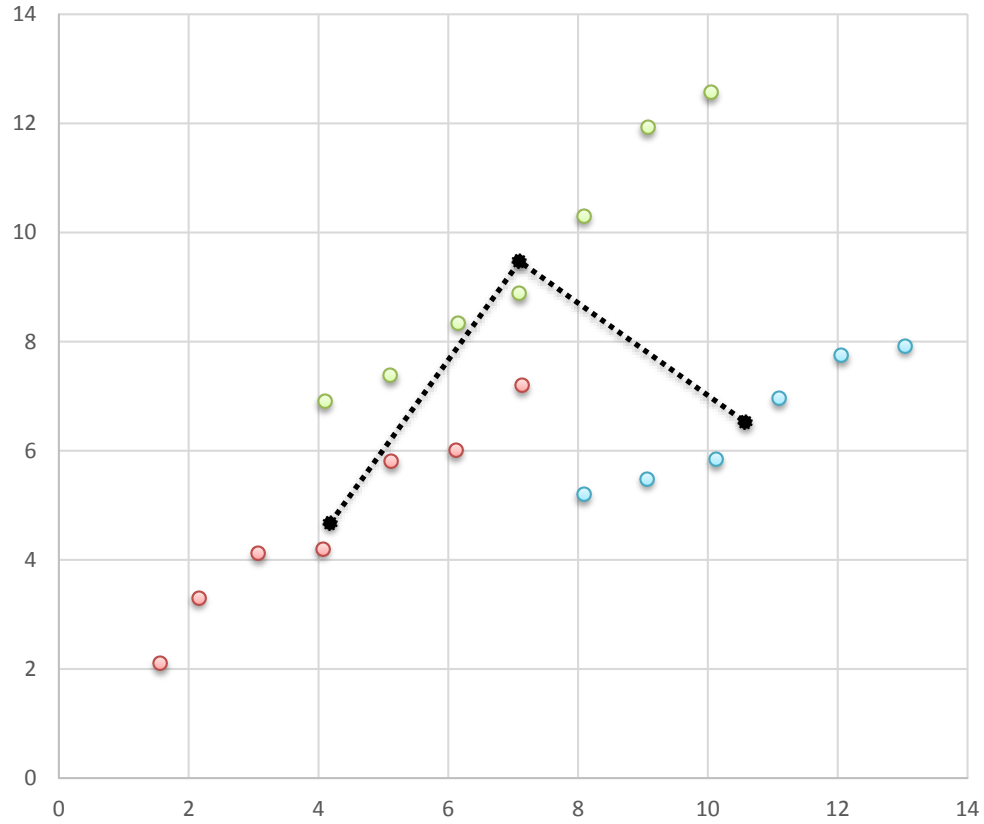
```
# Prediction
new_data <- as.data.frame(cbind(7.10,8.88))
colnames(new_data) <- c("Y1", "Y2")
predict(lda, new_data)$class

[1] 2
Levels: 1 2 3
```

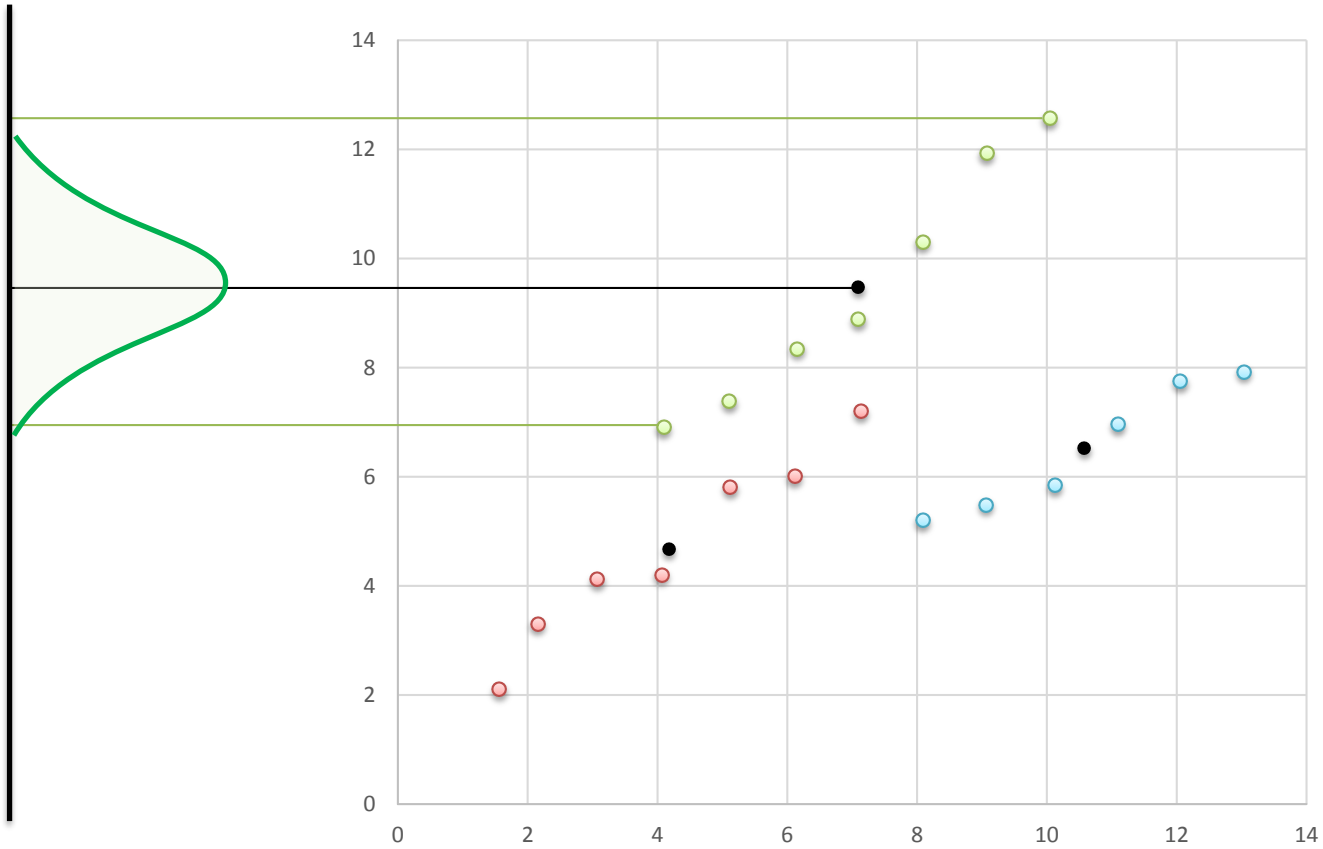
Visualization

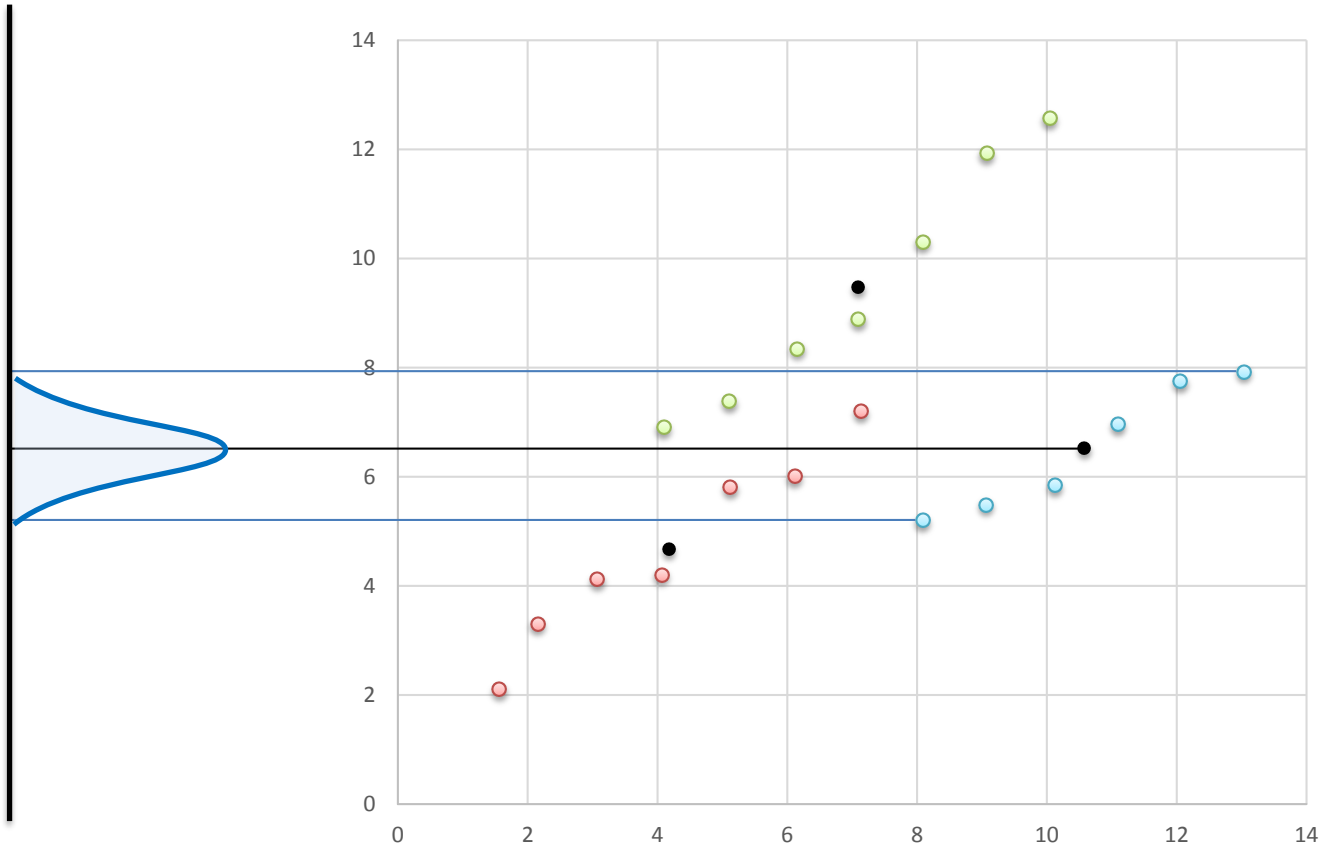

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

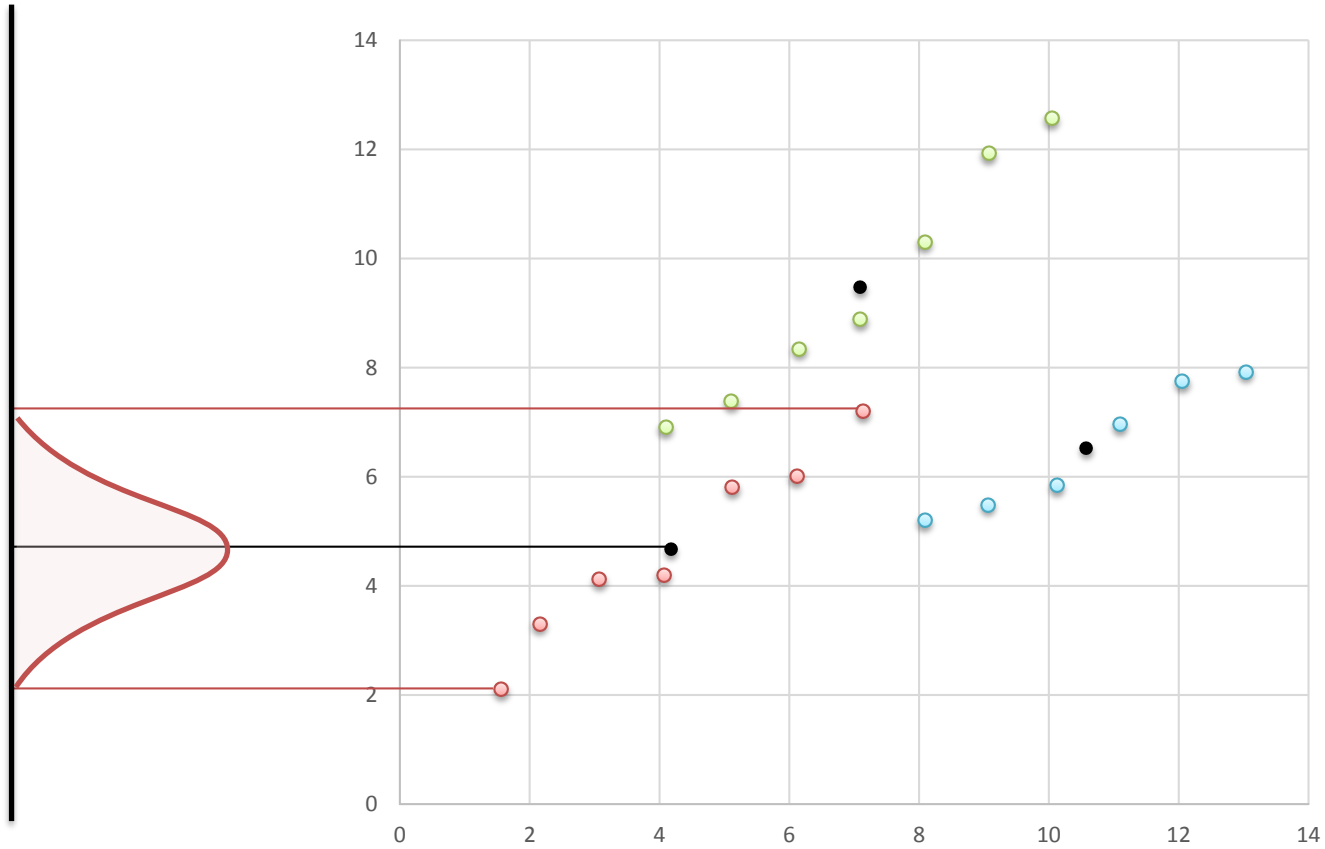

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

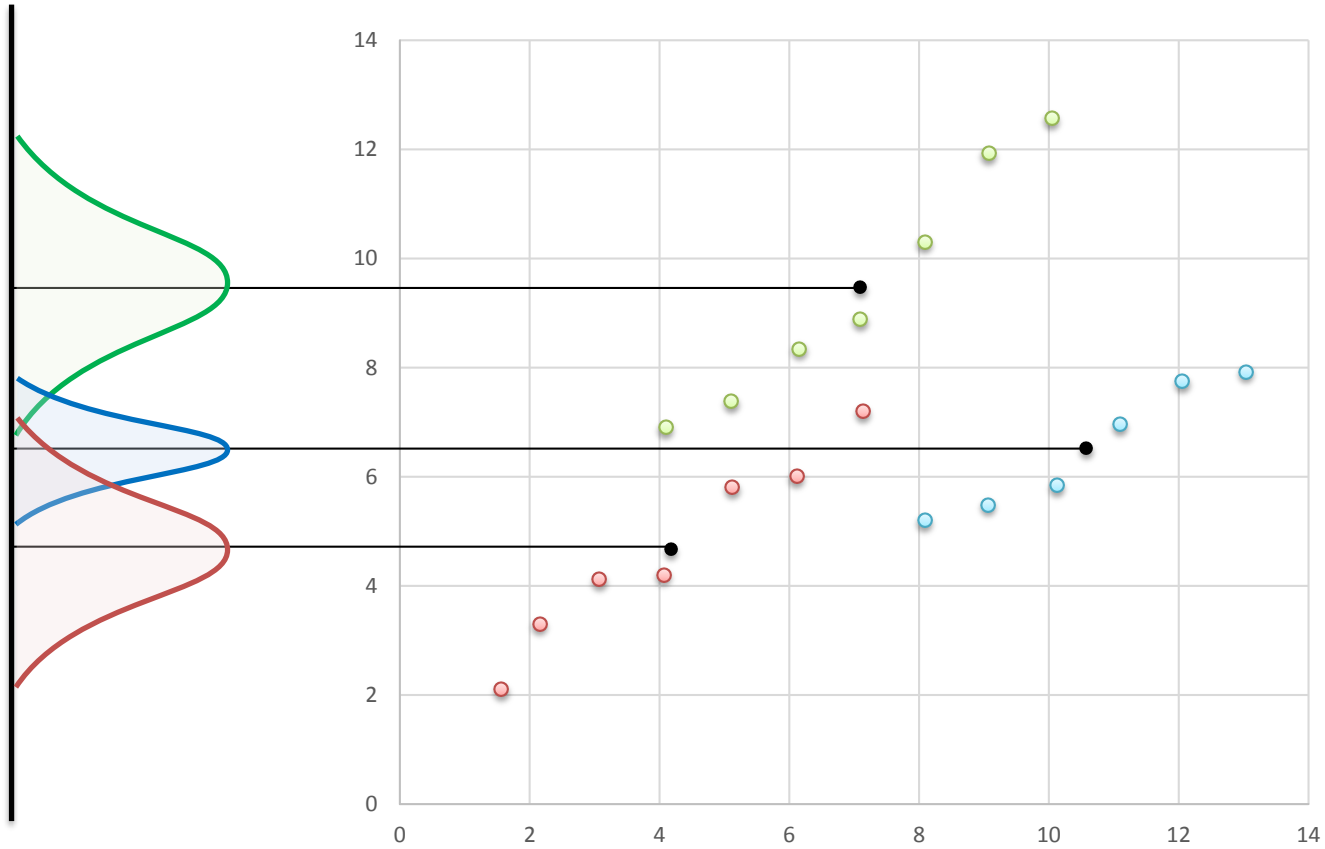

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

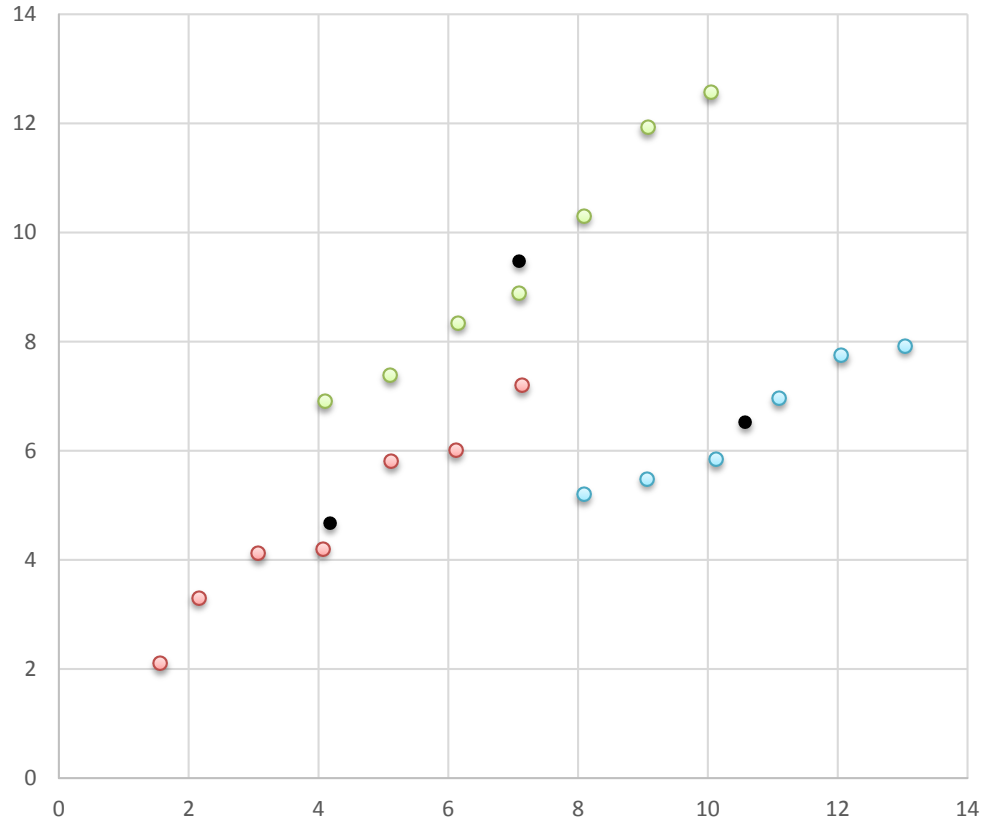

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

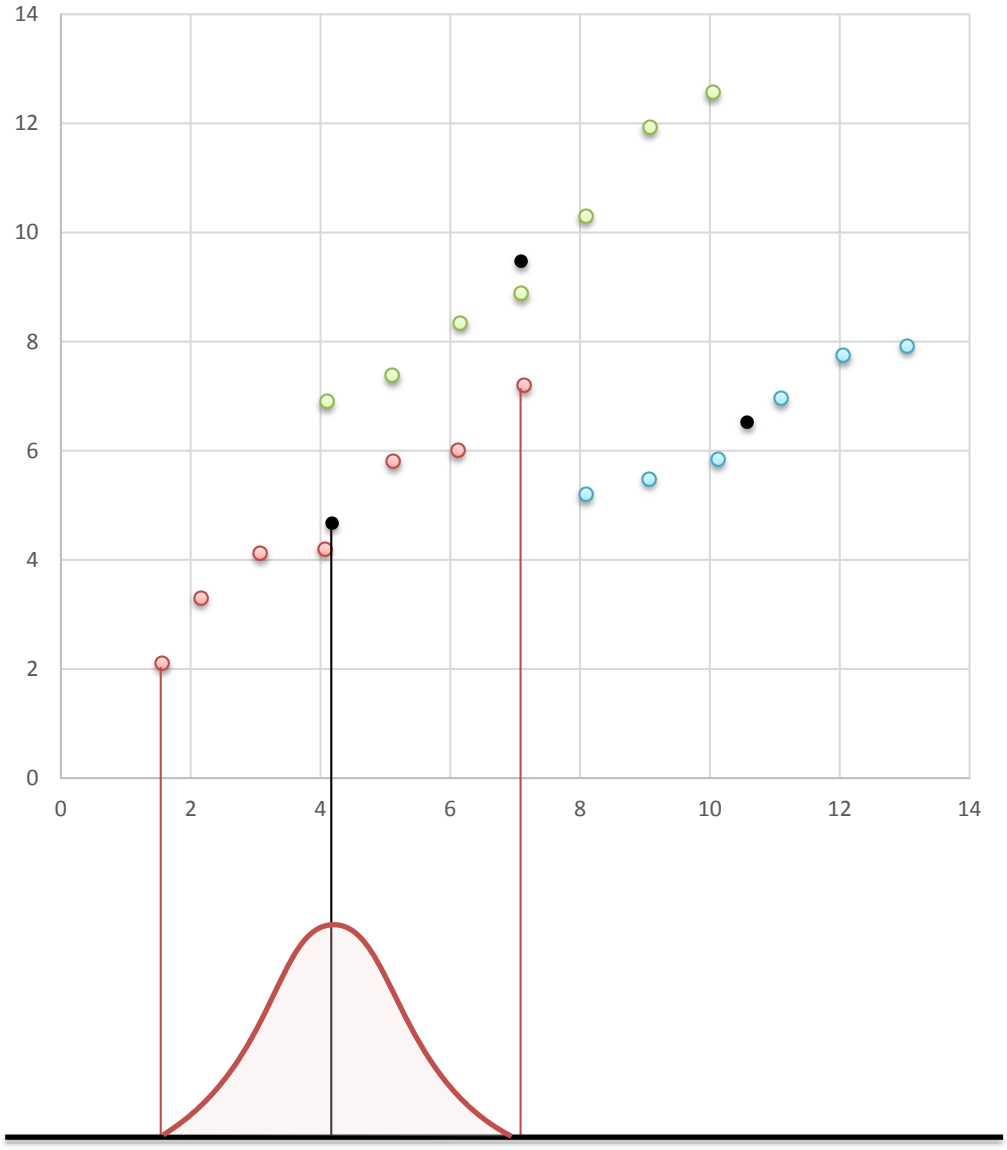

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

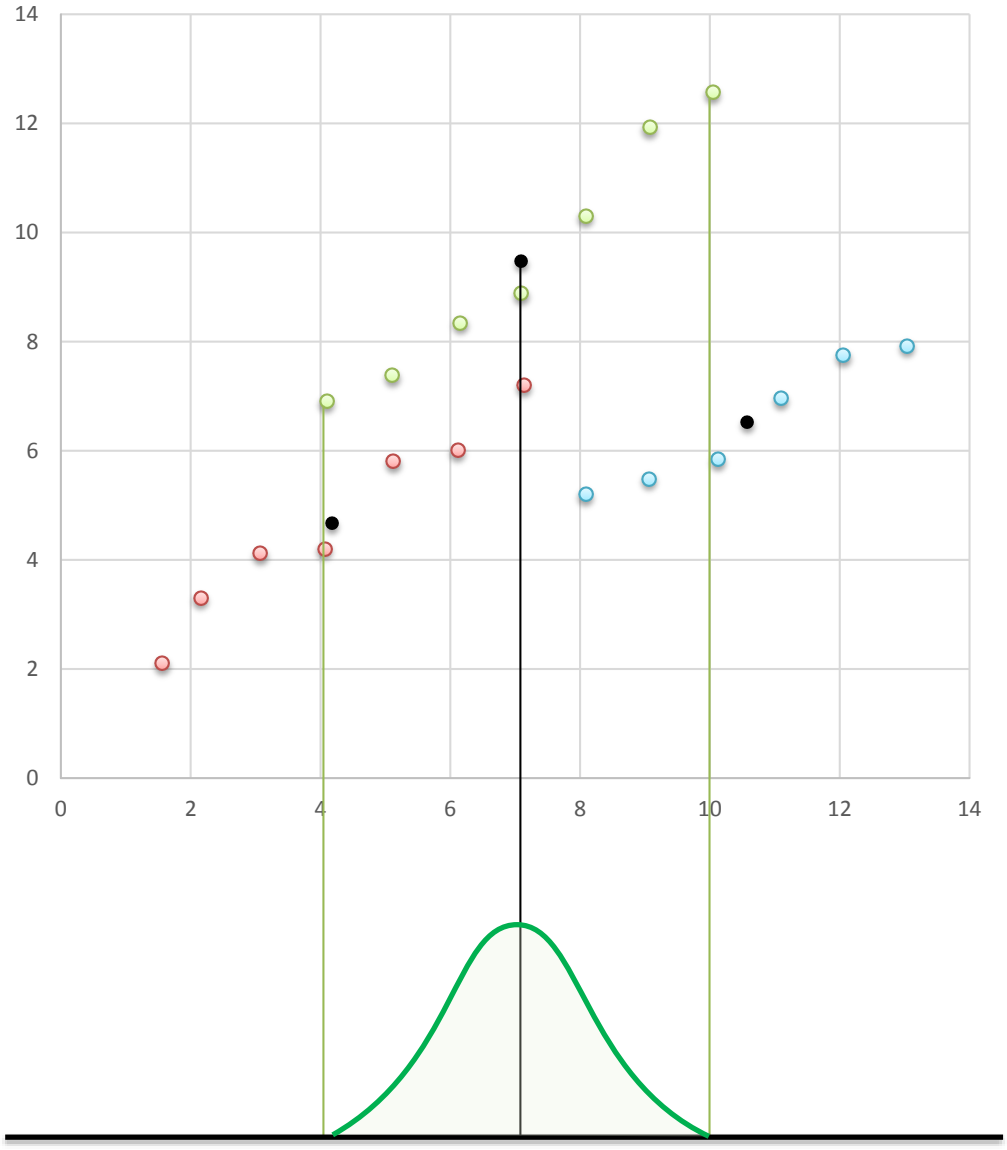

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

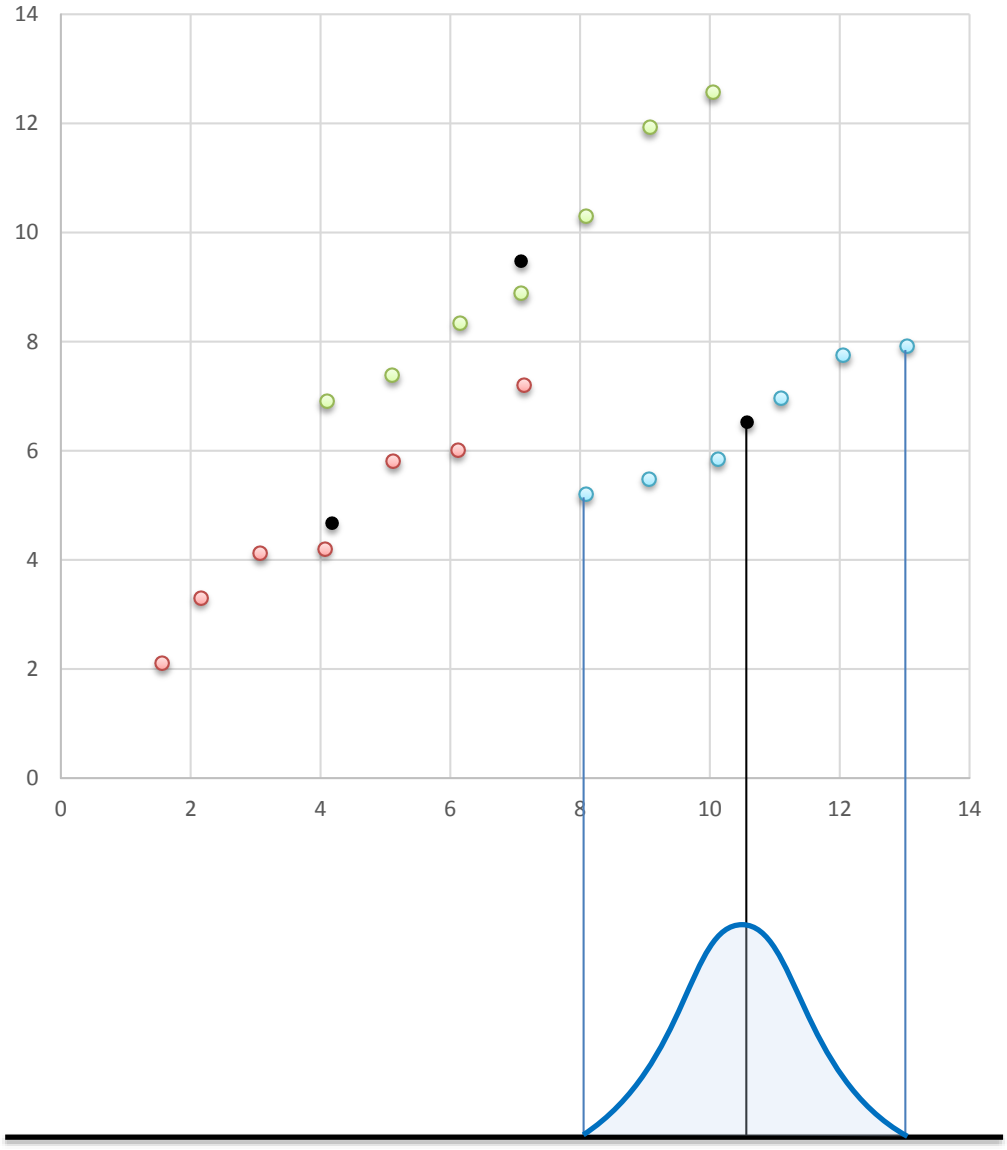

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

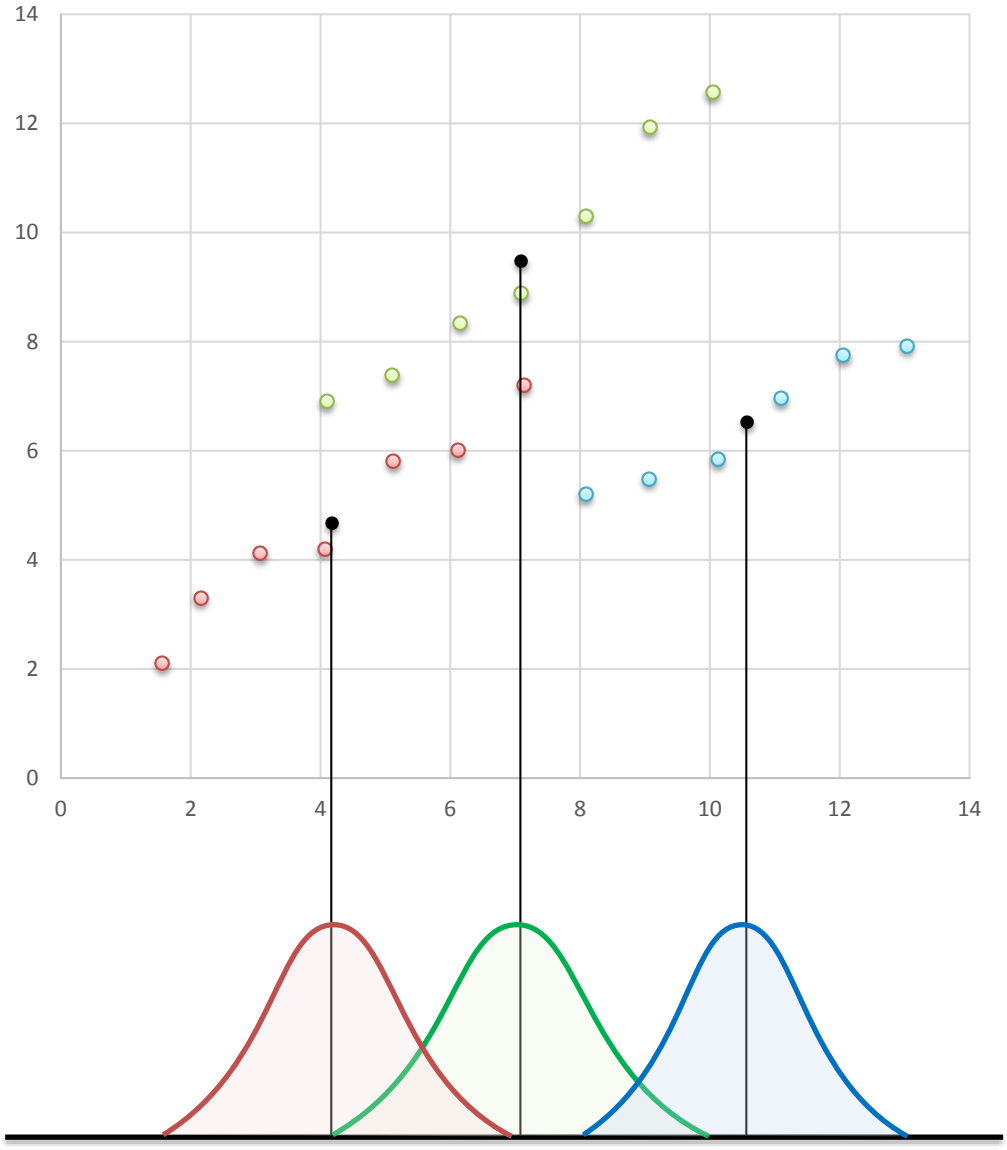

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

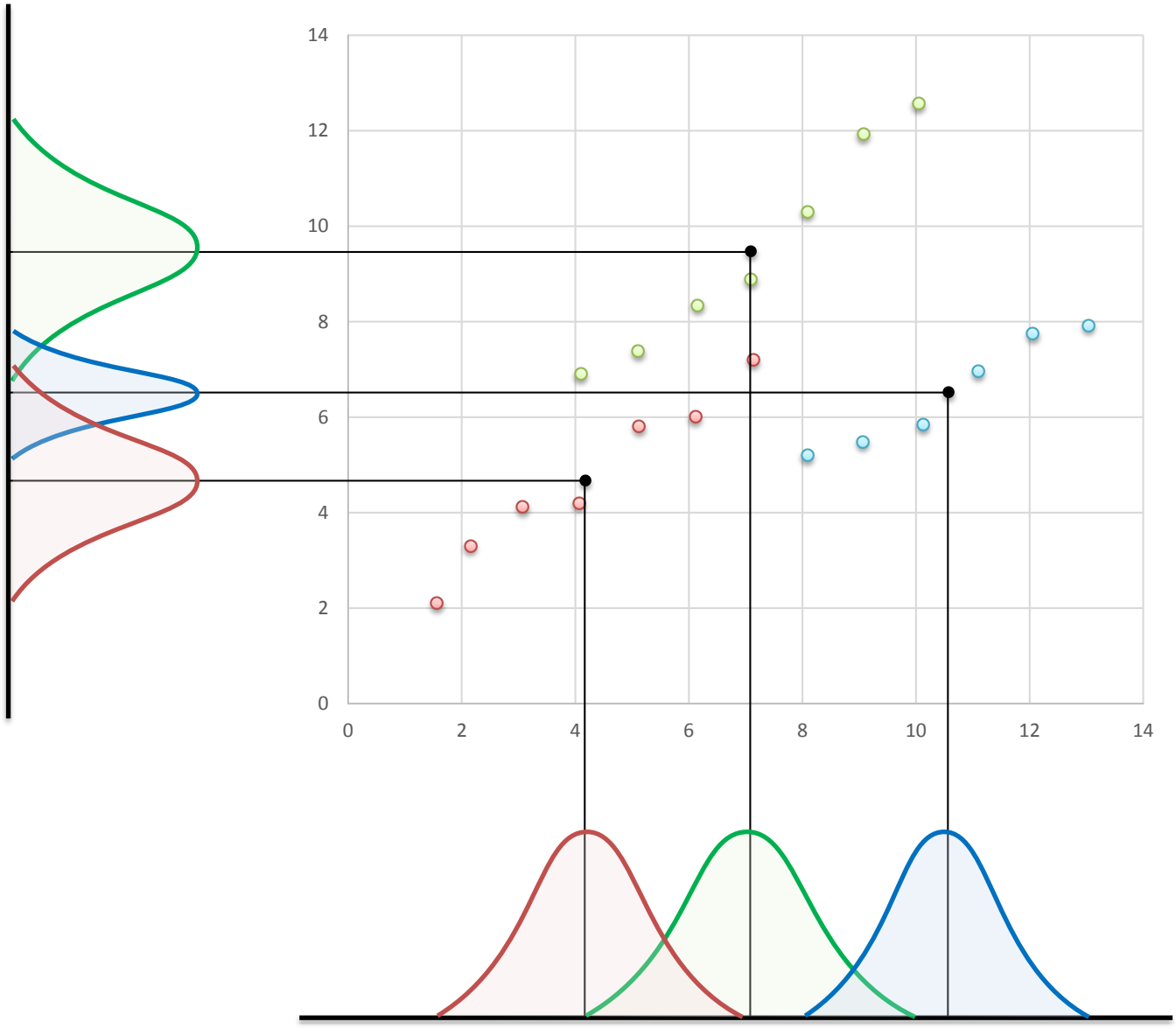

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

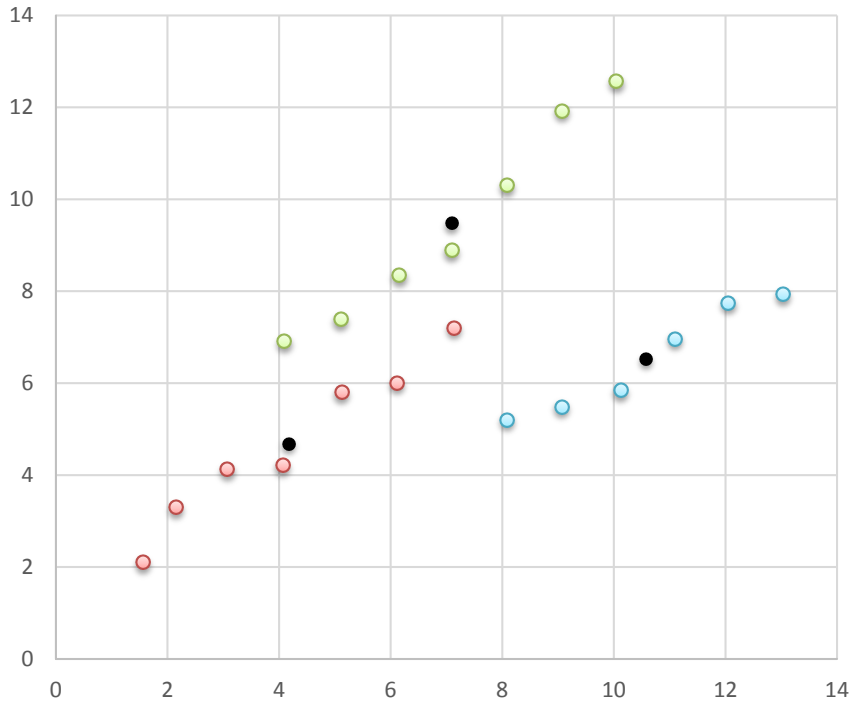

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

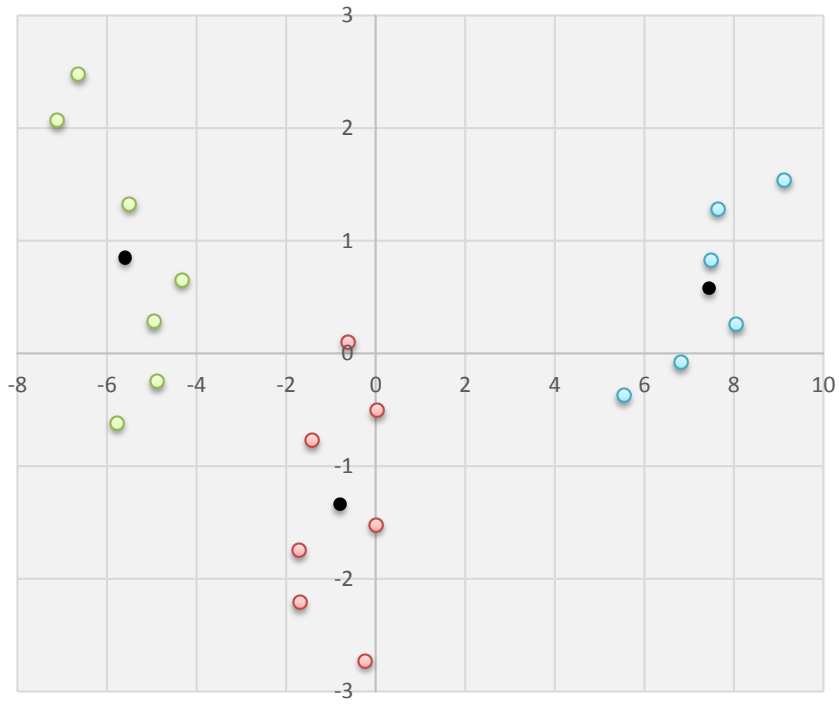

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

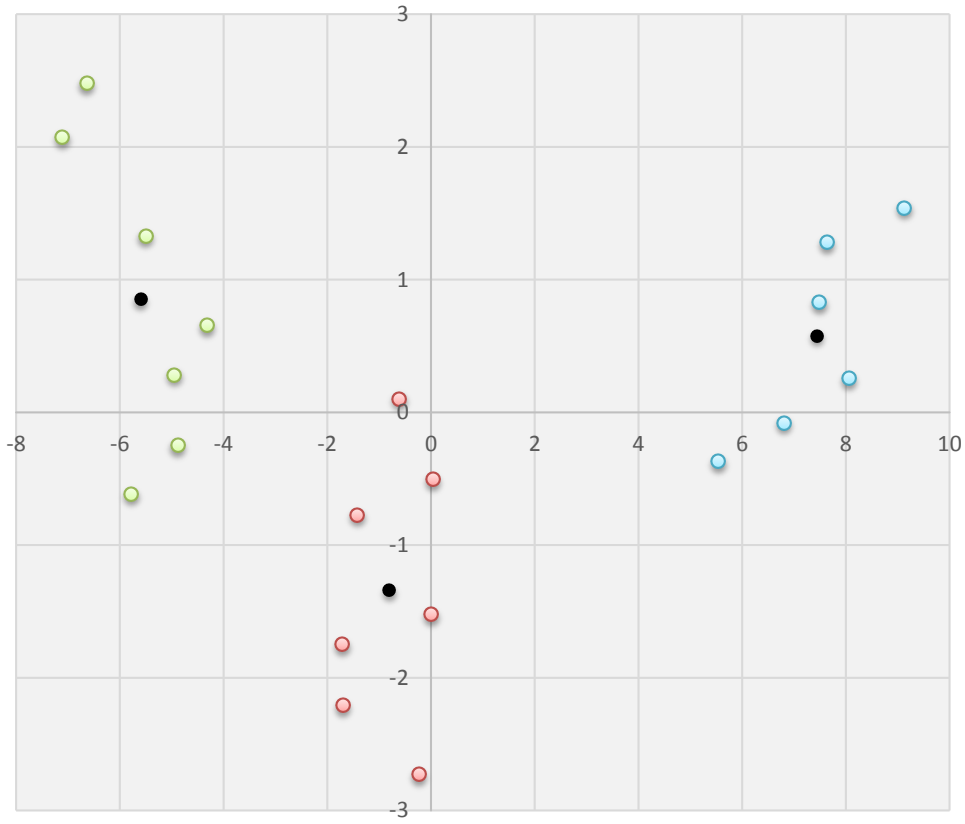

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

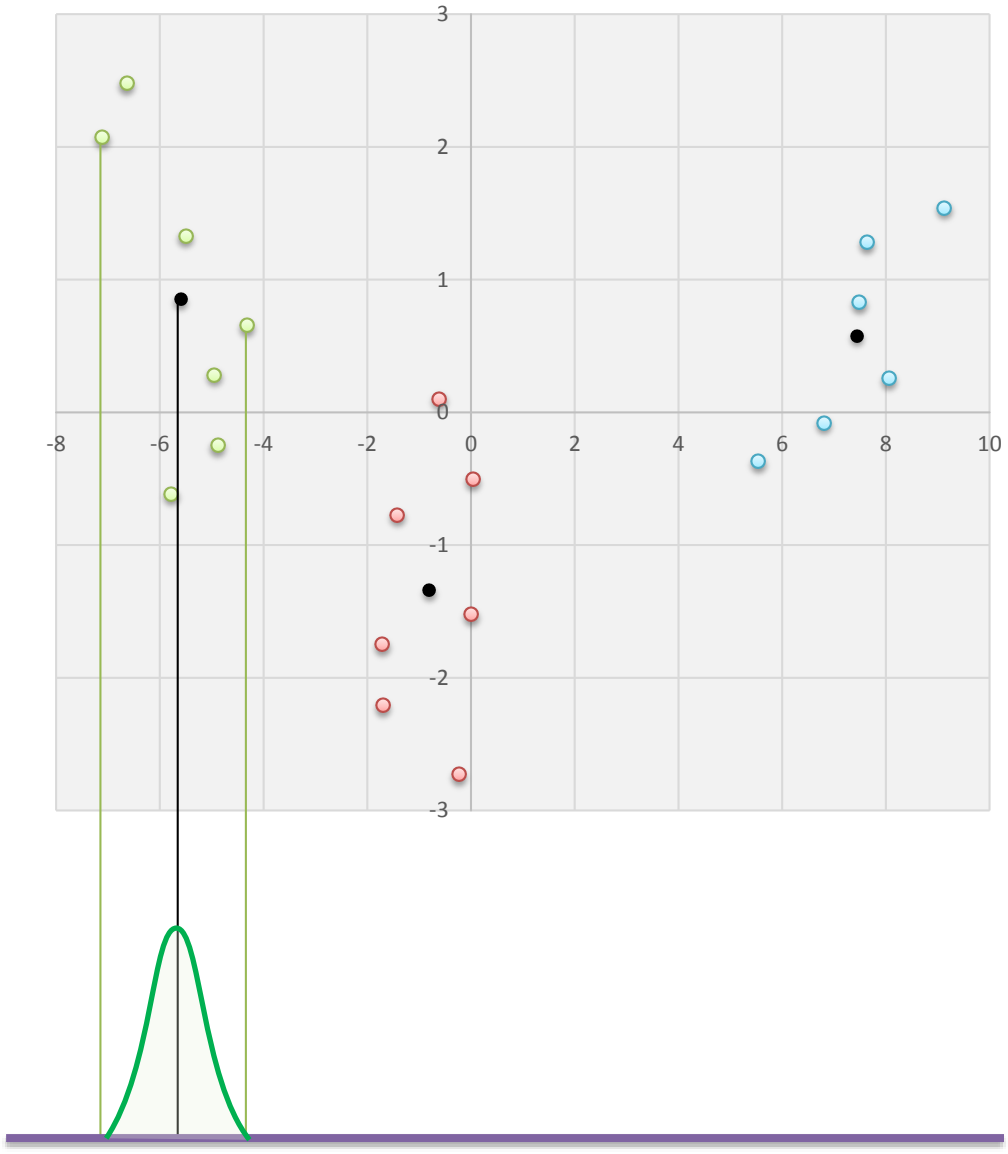

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

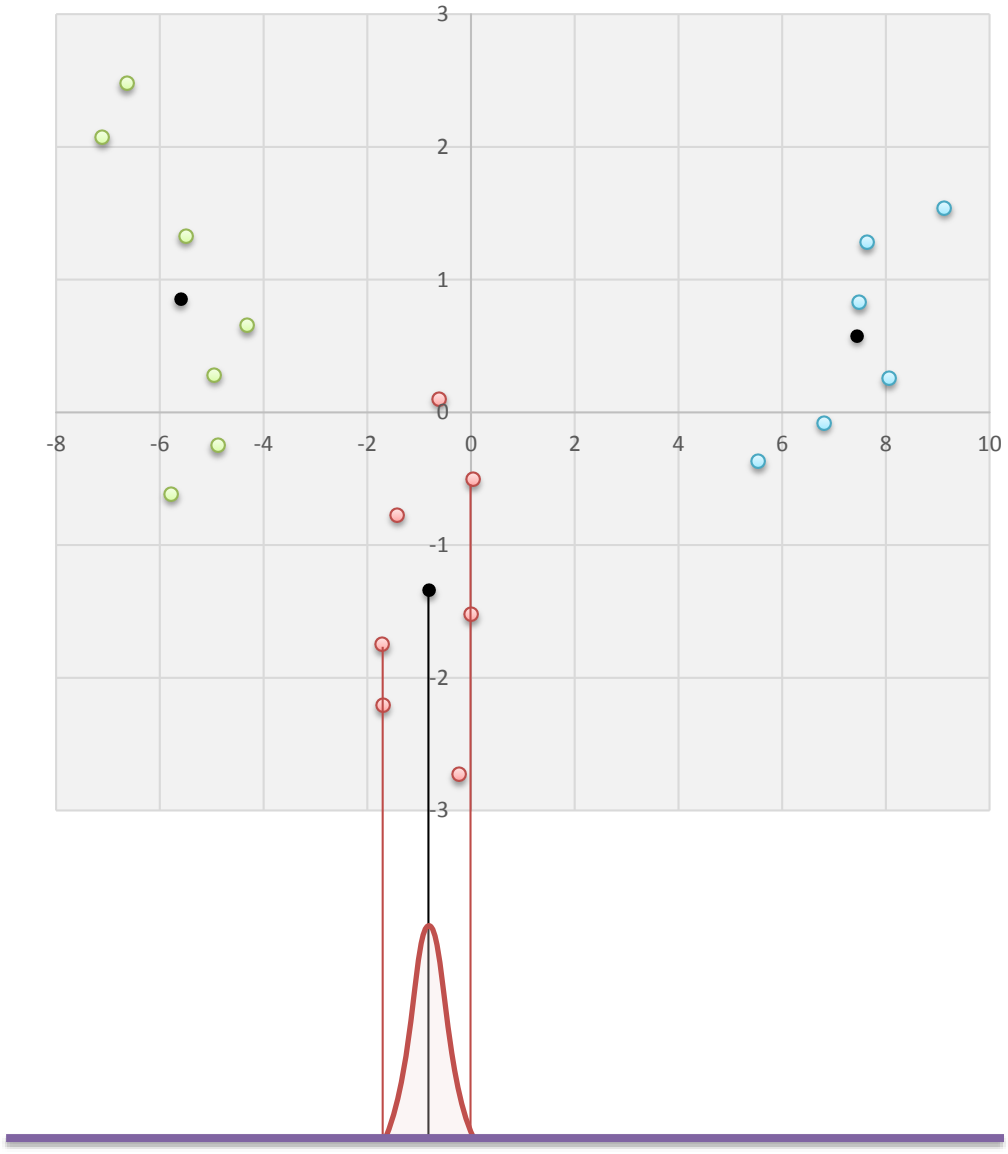

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

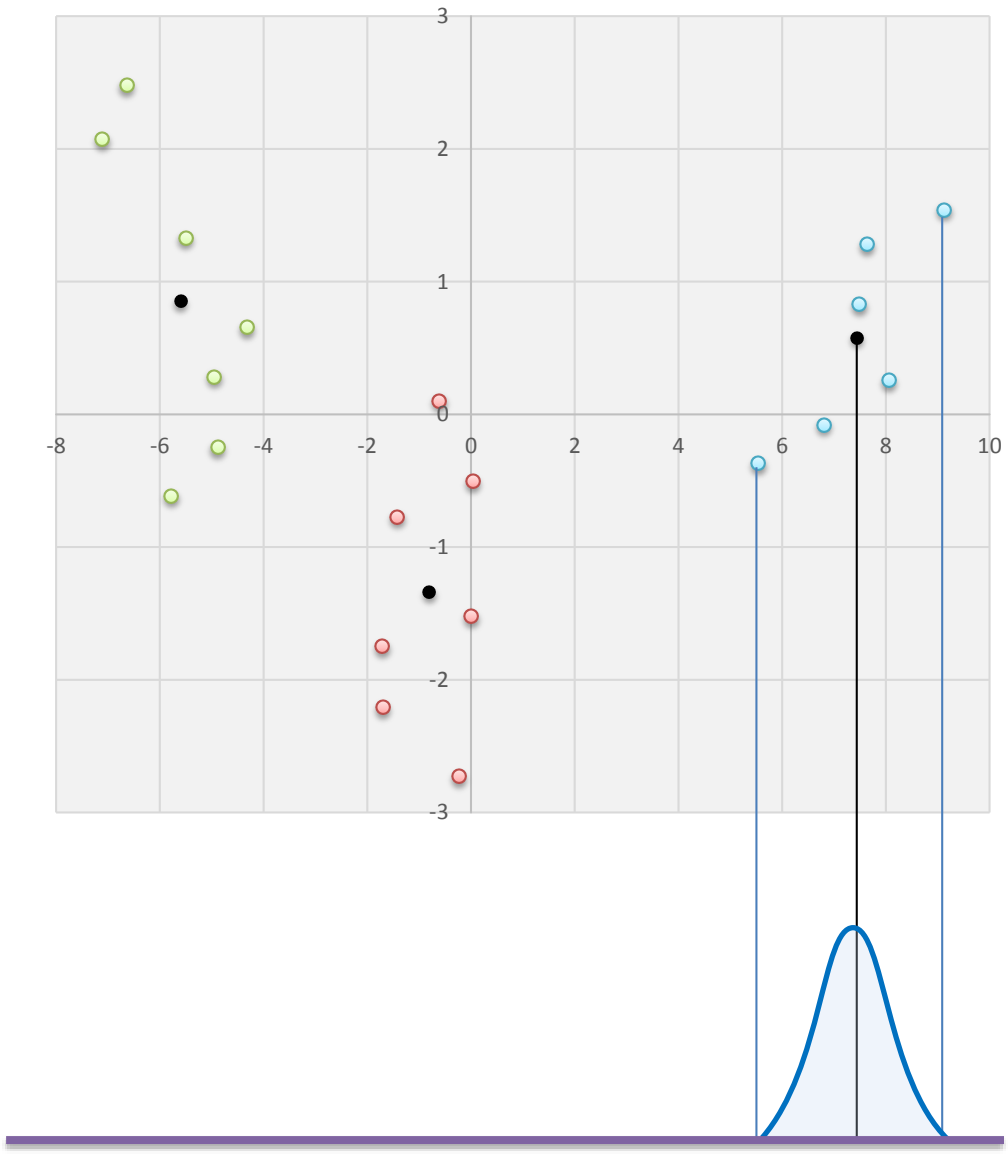

14
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14

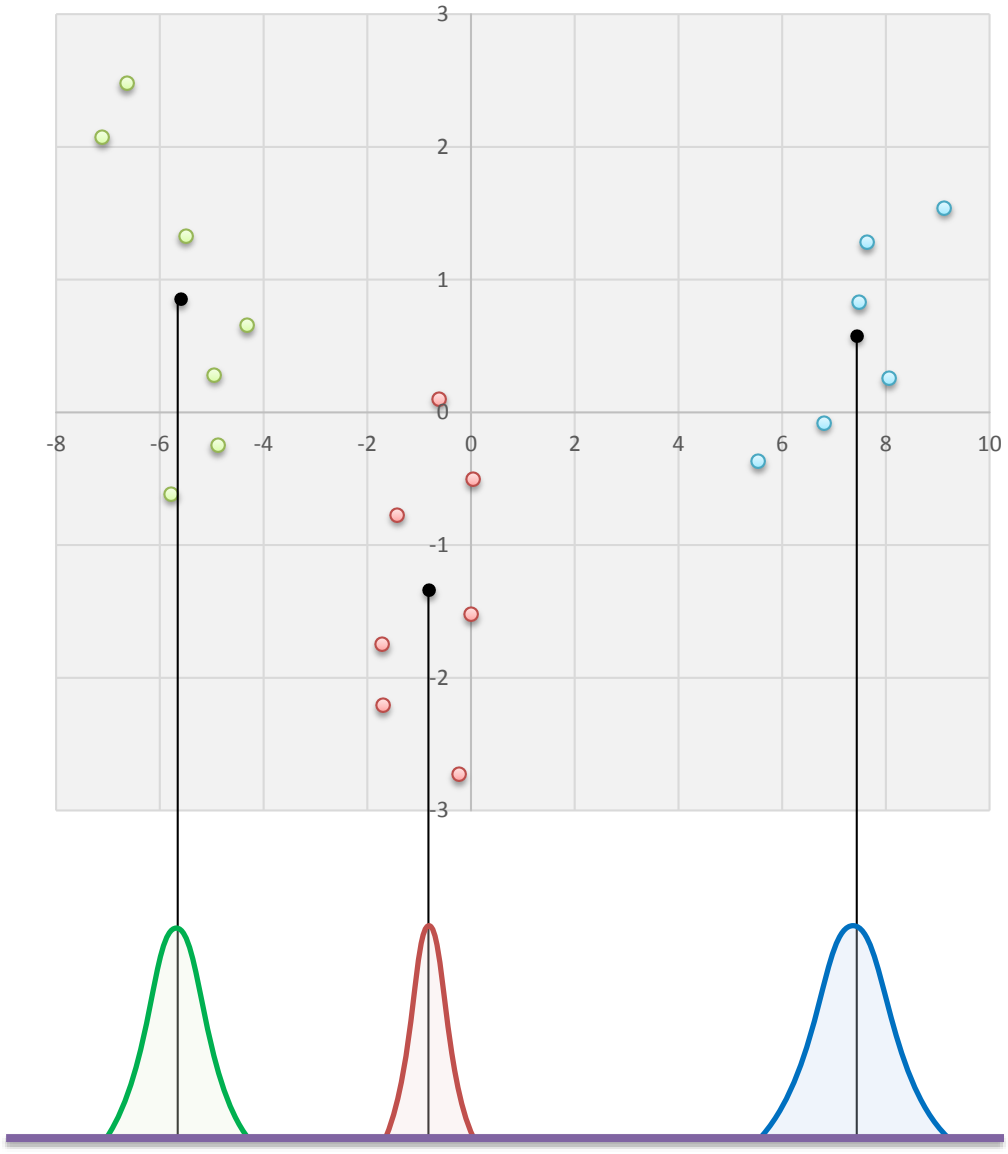

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

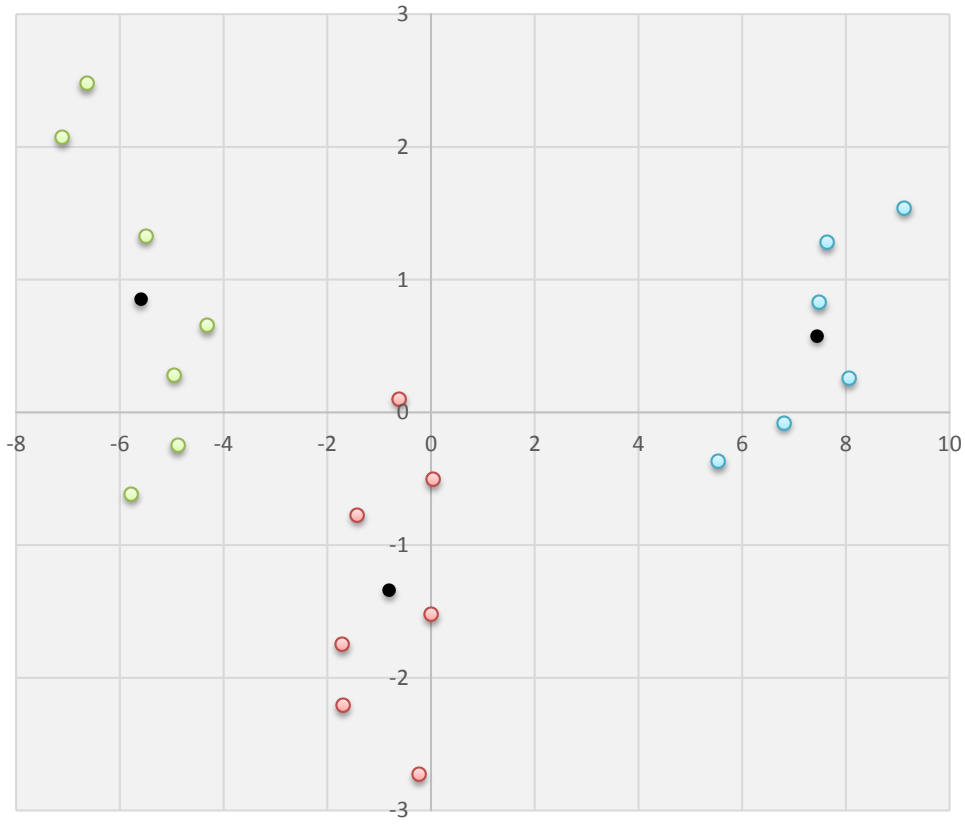

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

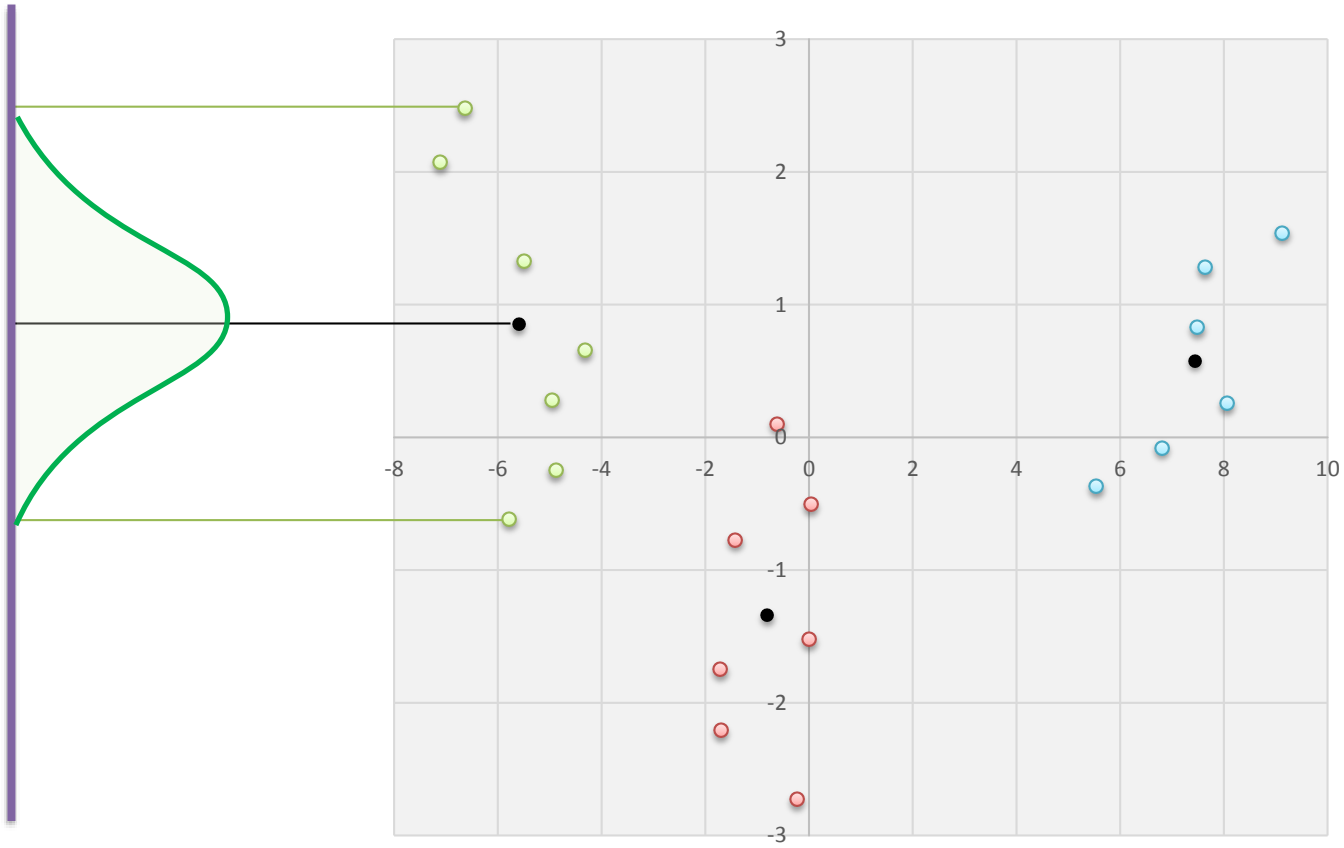

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

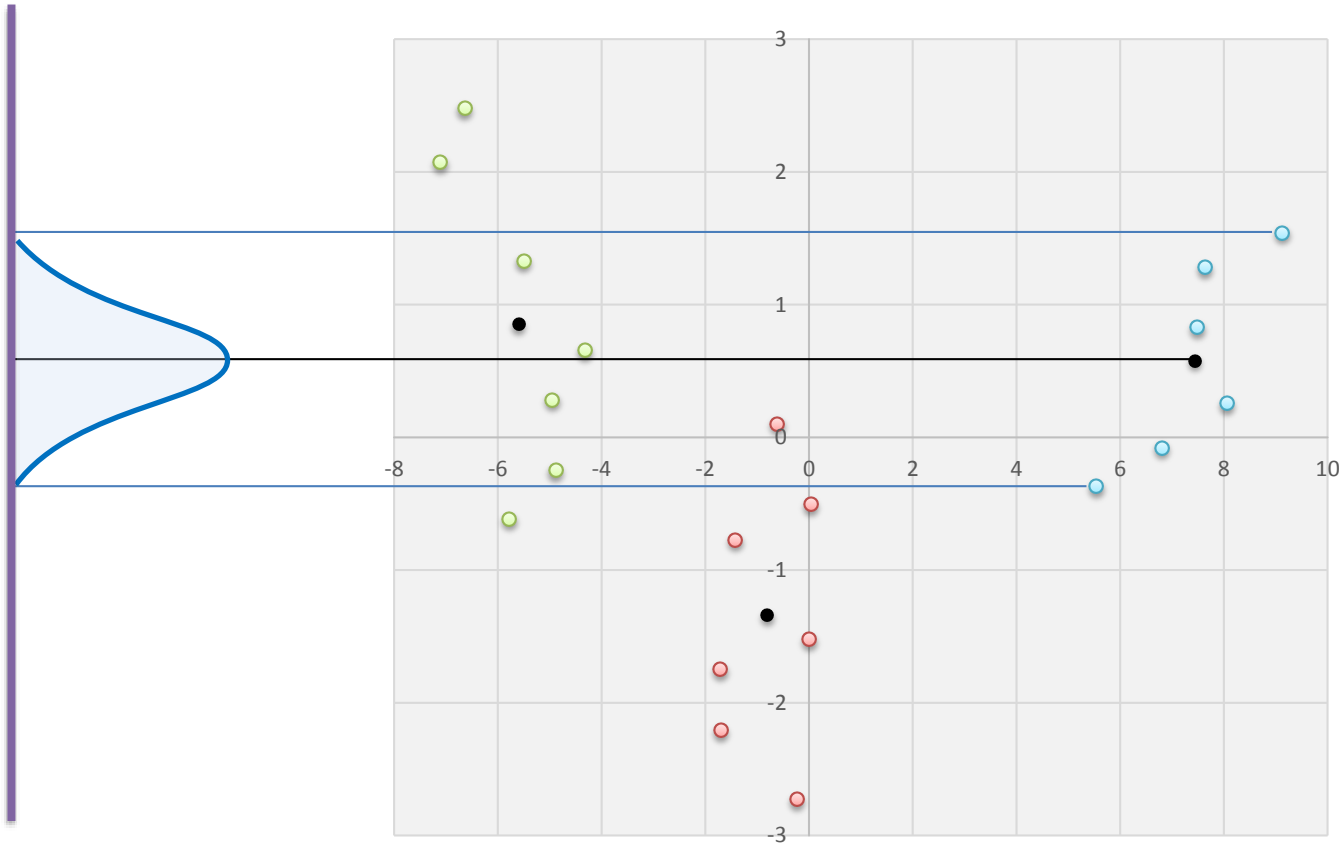

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

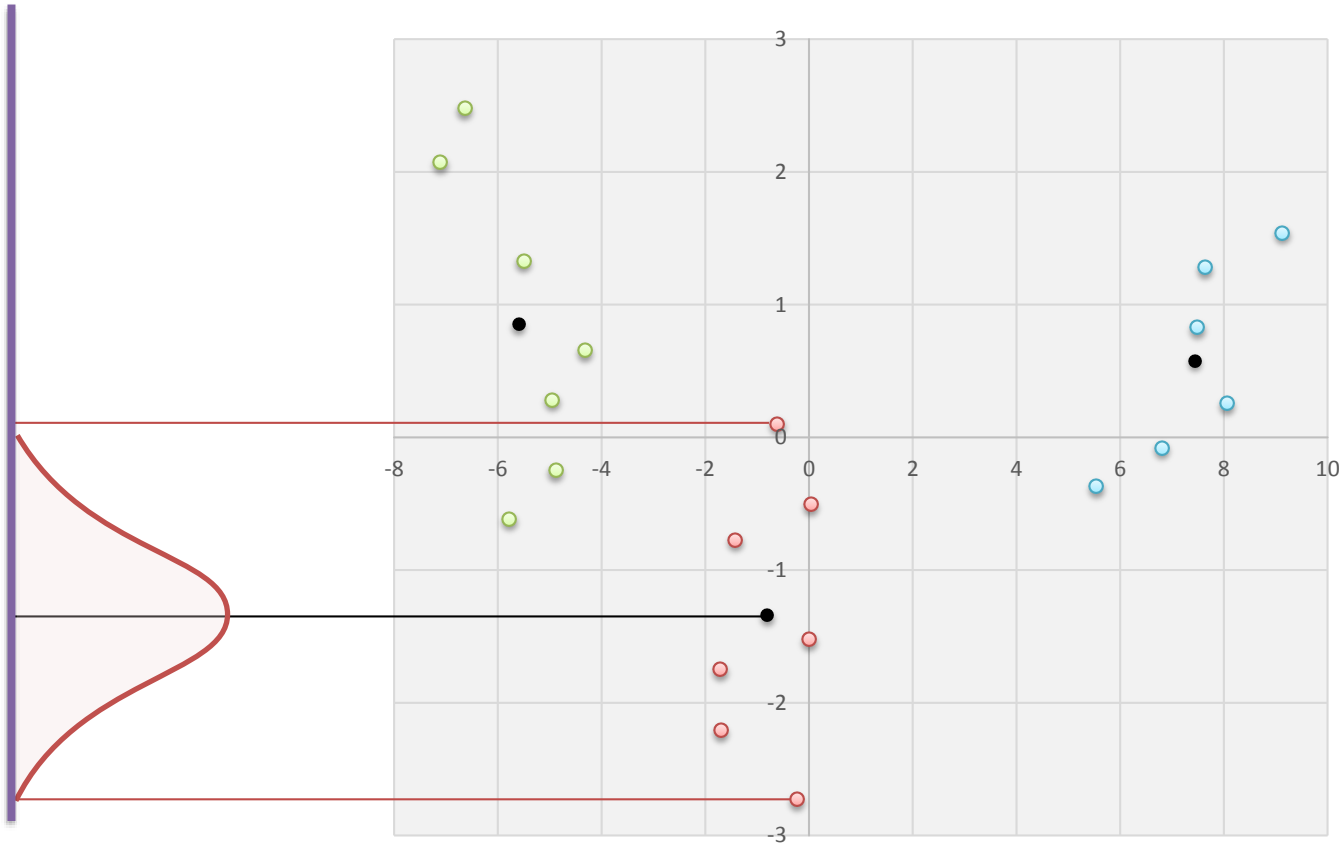

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

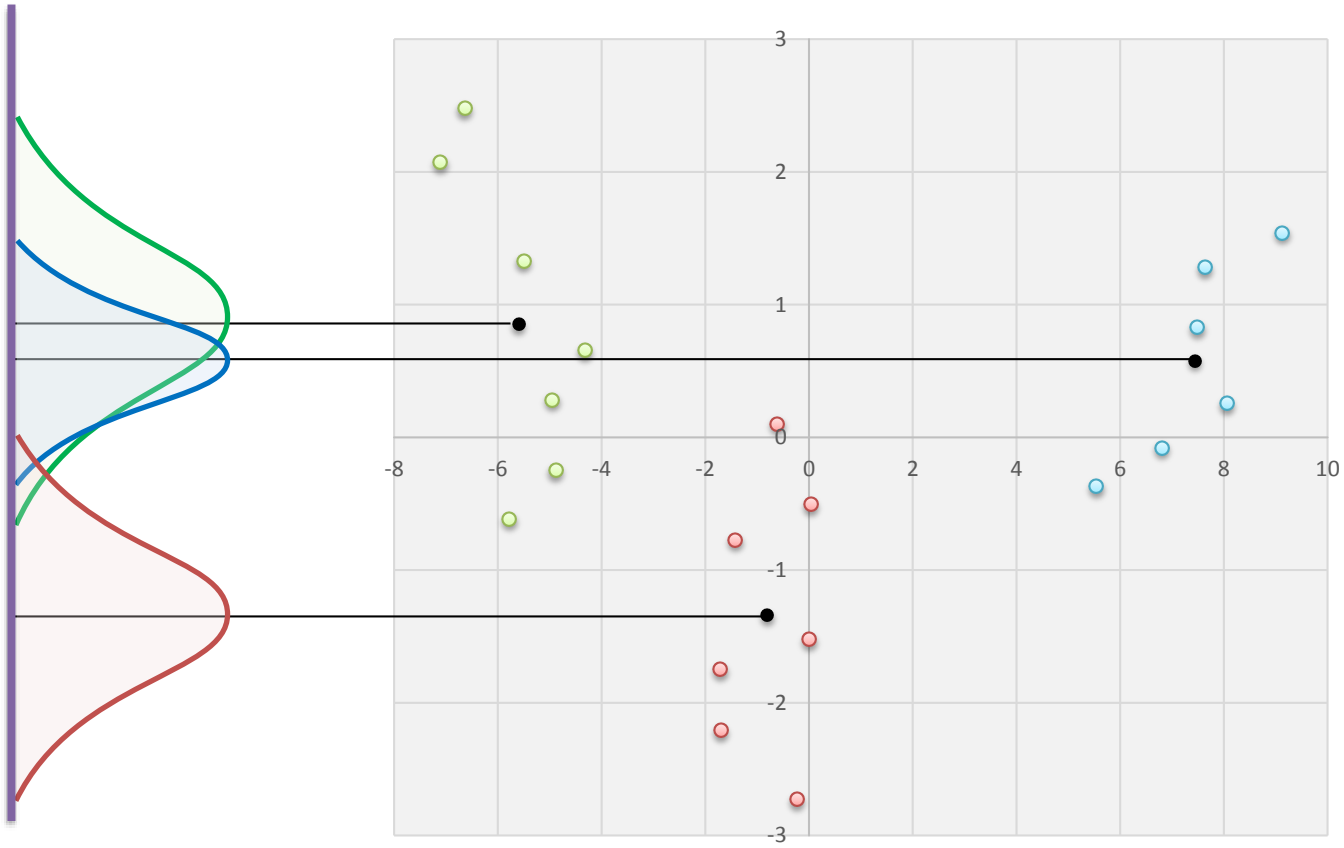

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

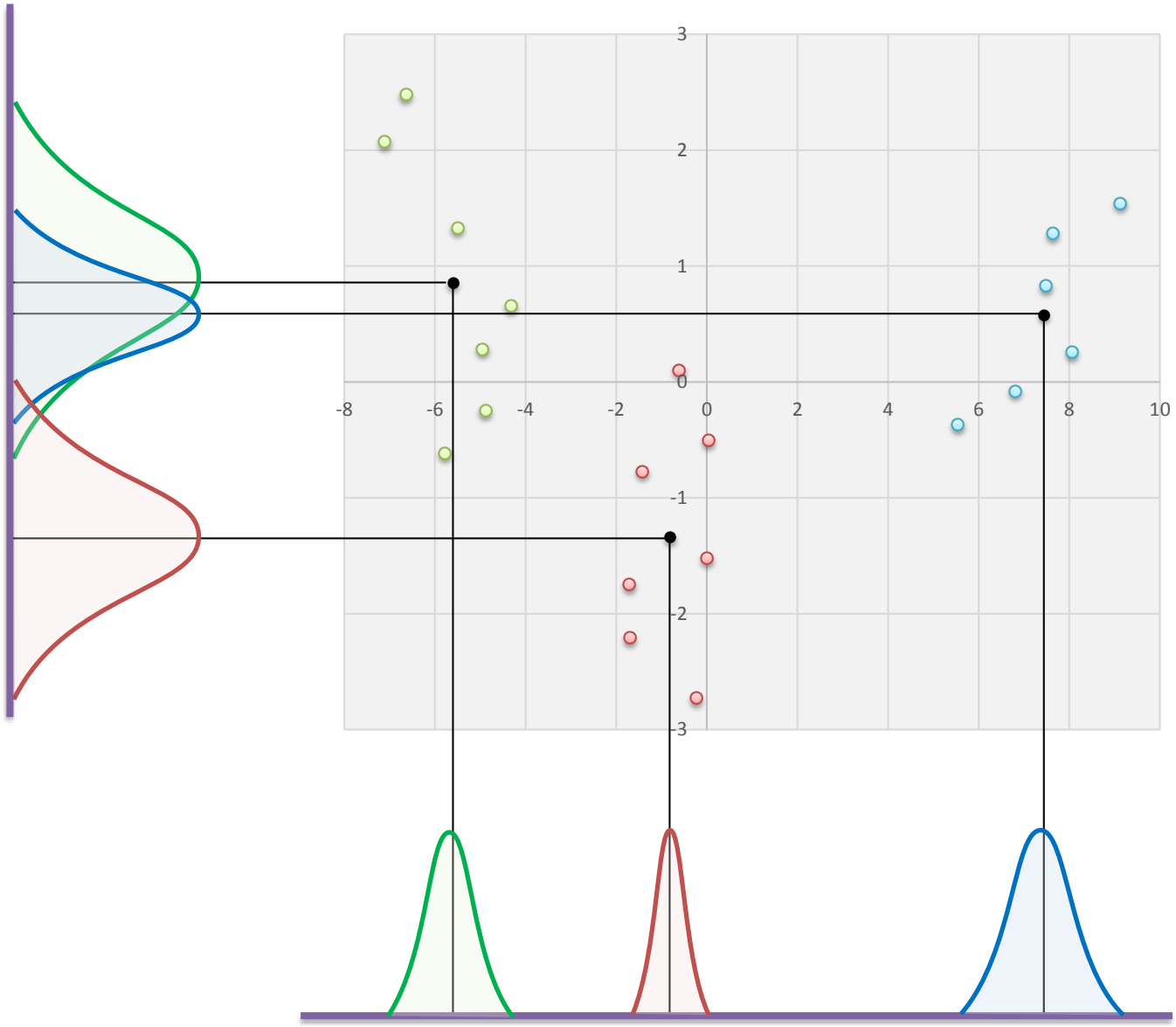

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

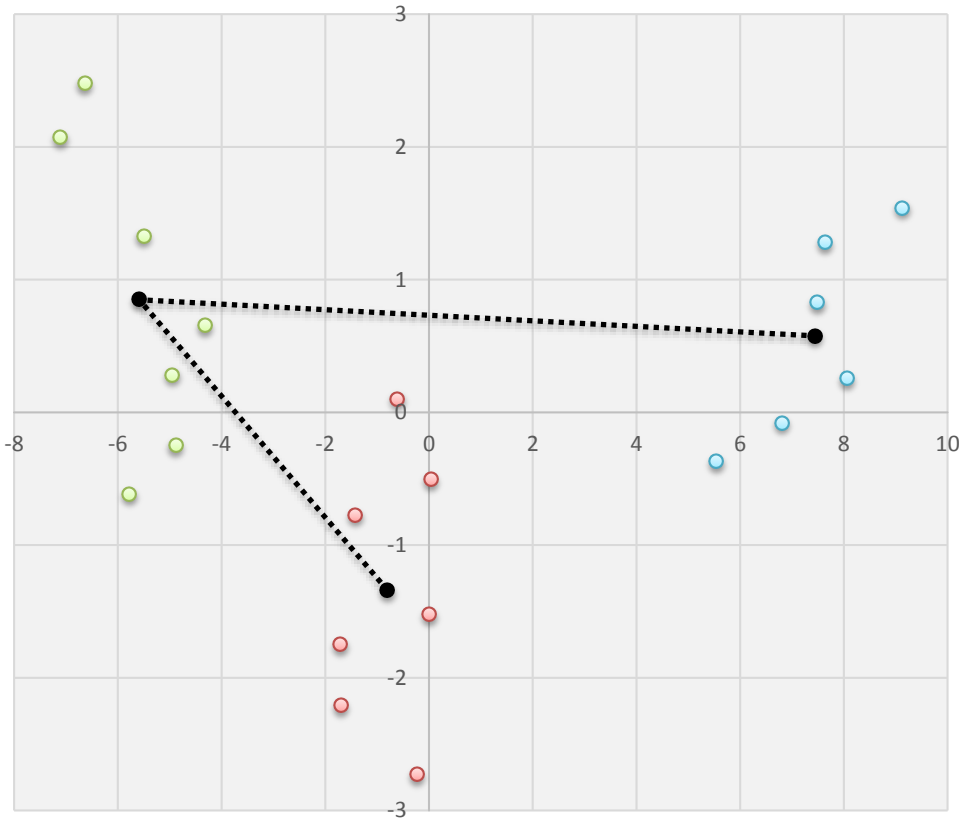

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

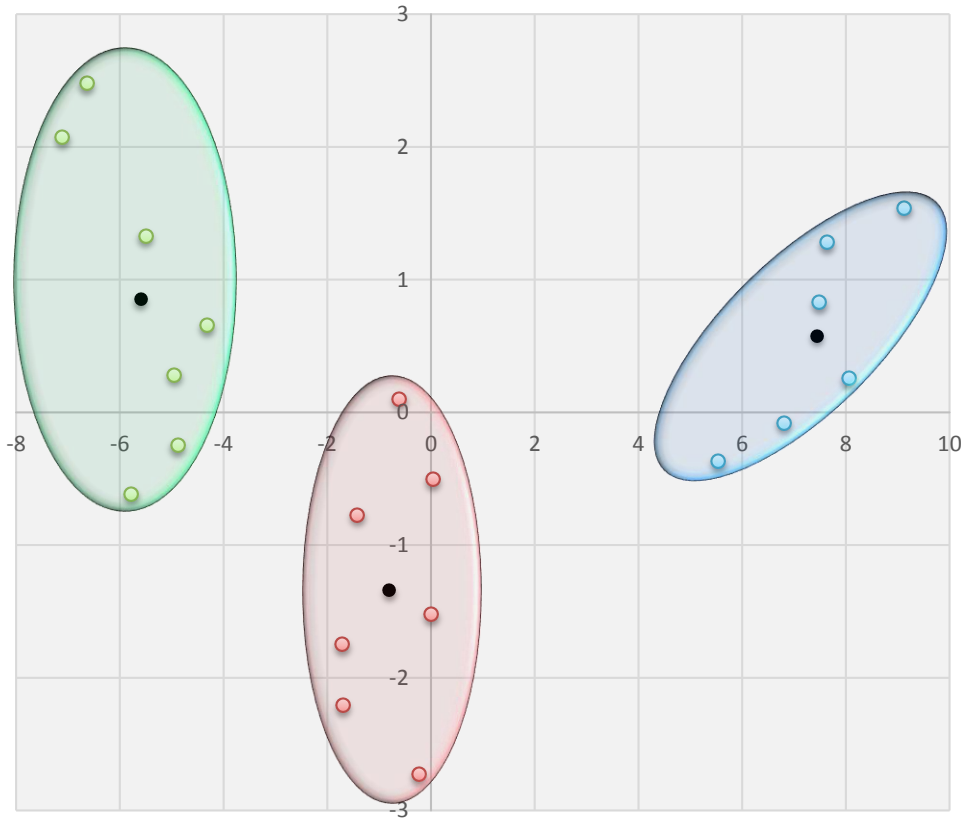

3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10


3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10


3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10


3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10


3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10


3
2
1
0
-1
-2
-3
-8
-6
-4
-2
0
2
4
6
8
10

MVDA

https://github.com/Valdecy/Multivariate_Data_Analysis

```
#############################################################################
#############################################################################

# Created by: Prof. Valdecy Pereira. D.Sc.
# UFF - Universidade Federal Fluminense (Brazil)
# email:  valdecypereira@yahoo.com.br
# Course: Multivariate Data Analysis
# Lesson: Discriminant Analysis
```

Citation:

PEREIRA. V. (2016). Project: Multivariate Data Analysis. File: R-MVDA-06-DA.pdf. GitHub repository: <https://github.com/Valdecy/Multivariate_Data_Analysis>

```
#############################################################################
#############################################################################
```

Bibliography

CORRAR. L.J.; PAULO. E.; DIAS FILHO. J. M. **Análise Multivariada para Cursos de Administração. Ciências Contábeis e Economia**. ATLAS. 2009.

FÁVERO. L. P.; BELFIORE. P.; SILVA. F. L.; CHAN. B. **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. CAMPUS. 2009.

HAIR. J. F.; BLACK. W. C.; BABIN. B. J.; ANDERSON. R. E.; TATHAM. R. L. **Análise Multivariada de Dados**. BOOKMAN. 2009.

LATTIN. J.; CARROLL. J. D.; GREEN. P. E. **Análise de Dados Multivariados**. CENGAGE Learning. 2011.

LEVINE. D. M.; STEPHAN. D. F.; KREHBIEL. T. C.; BERENSON. M. L. **Estatística - Teoria e Aplicações - Usando Microsoft Excel**. LTC. 2012.