# UNIVERSIDADE FEDERAL FLUMINENSE

## Programa de Mestrado e Doutorado em Engenharia de Produção

Multivariate Data Analysis

# Multidimensional Scaling

Professor: Valdecy Pereira, D. Sc.

email: valdecy.pereira@gmail.com

# Outline

# MVDA – *Multidimensional Scaling*

The **MDS** (**Multidimensional Scaling**) or **PCoA** (**Principal Coordinates Analysis**), can be considered as an alternative to EFA, and it is a technique of interdependence that allows mapping distances between points (objects or events) in a multidimensional space. It reduces large amounts of data in structures easy to understand, the points that are close together represent similar objects as different objects are represented by points that are distant, so clusters can be interpreted. Characteristics:

- **Data Type**: The input data may be from similarities or dissimilarities (differences or distances). Many authors encourage the use of dissimilarities because their relationship with distances is direct and positive, that is, greater the inequality, greater the distance.
- **Analysis Type**: Metric MDS (when order between objects does not matter) and Non-Metric MDS (when order between objects does matter).
- **Typical Applications**: Consumer perception of a brand, Perception evaluation of process implementation (before and after effect), Object clustering, etc.

# MVDA – *Multidimensional Scaling*

**Metric MDS** produces a set of uncorrelated (orthogonal) axes to summarize the variability of a dataset, each axis has an eigenvalue that indicates a certain amount of variation. Each object has a coordinate along each axis.

**Non-Metric MDS** produces an ordination based on a dissimilarity matrix, representing as closely as possible, the pairwise dissimilarity between objects in a low-dimensional space. The original dissimilarity is substituted with ranks, therefore, the information about the magnitude of distances is lost but not its relative position, indicating that clusters possess similar objects.

**Limitations**:

- The dimensions change with time.
- An object may have bias.
- The interpretation of clusters can vary from researcher to researcher.
- The number of dimensions (between 1 and 3) may not be sufficient to identify the structure of relationship between objects.
- It is an **EXPLORATORY** technique.

# Assumptions

# MVDA – *Multidimensional Scaling*

## ASSUMPTIONS

- **Number of respondents**: At least 1 respondent must evaluate the objects. To achieve an inferential power, sample size must be calculated statistically.

- **Number of objects**: When the number of objects increases, the more accurate is the output in statistical terms. However, data interpretation can be difficult to achieve. Rule of thumb - For a MDS with $\mathbf{2}$ dimensions is advisable to have at least $\mathbf{10}$ objects. For a MDS with $\mathbf{3}$ dimensions it is advisable to have at least $\mathbf{15}$ objects.

- **Number of dimensions**: It is recommended that the number of dimensions could range between 1 and 3. For more than 3 dimensions interpretability is questionable and difficult to do.

# Distances and Dissimilarities

# MVDA – *Multidimensional Scaling*

For **similarities**, the following equation transforms **similarities** in **dissimilarities**.

$$DIS_{ij} = 1 - SIM_{ij}$$

Where:

$DIS_{ij} =$ degree of dissimilarity between two objects;

$SIM_{ij} =$ degree of similarity between two objects;

For **dissimilarities**, the following equation transforms **dissimilarities** in **similarities**.

$$SIM_{ij} = 1 - DIS_{ij}$$

# MVDA – *Multidimensional Scaling*

For **correlations**, the following equation transforms **correlations** in **disparity** (**estimated dissimilarity**).

$$\delta_{ij} = \sqrt{2(1 - \rho_{ij})}$$

$$or$$

$$\delta_{ij} = 1 - \rho_{ij}$$

Where:

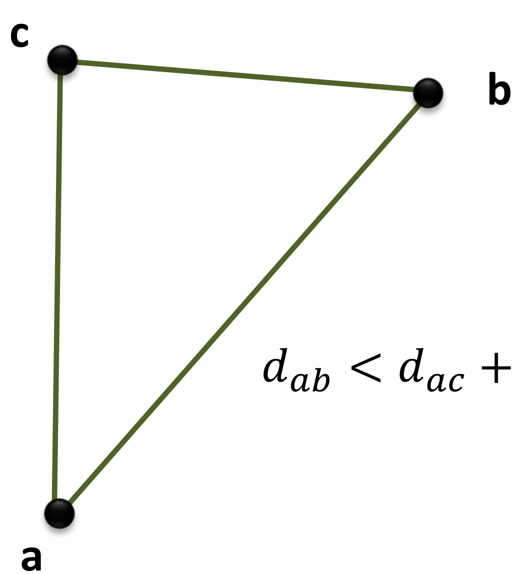
$\rho_{ij}$ = Correlation between two objects;

$\delta_{ij}$ = Disparity.

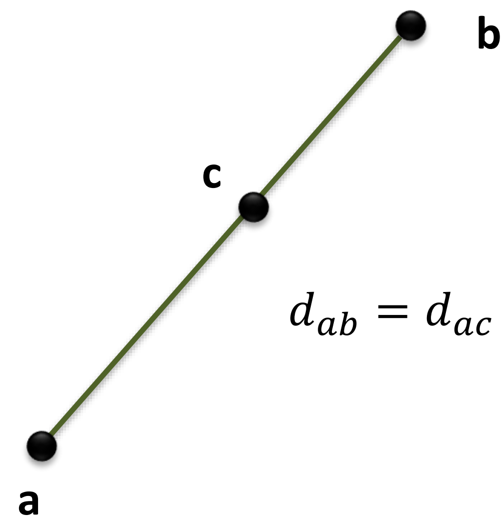
# MVDA – *Multidimensional Scaling*

A true measure of distance, follows three properties:

1. $d_{ab} = d_{ba}$
2. $d_{ab} \geq 0$ e $d_{ab} = 0 \leftrightarrow a = b$
3. $d_{ab} \leq d_{ac} + d_{cb}$




$$d_{ab} < d_{ac} + d_{cb}$$




$$d_{ab} = d_{ac} + d_{cb}$$

# MVDA – *Multidimensional Scaling*

Distance - Minkowski:

$$d_{mw} = \left(\sum |x_k - y_k|^p\right)^{\frac{1}{p}}$$

Depending on the value $p$, different distances metrics are obtained. The value of $p$ can not be less than 1, otherwise the triangle inequality is not obeyed.

Distance - Manhattan:

$$d_{mh} = \sum |x_k - y_k|$$

When $p$ = 1; we have the Manhattan distance.

# MVDA – *Multidimensional Scaling*

Distance - Euclidian:

$$d_{ec} = \left( \sum |x_k - y_k|^2 \right)^{\frac{1}{2}}$$

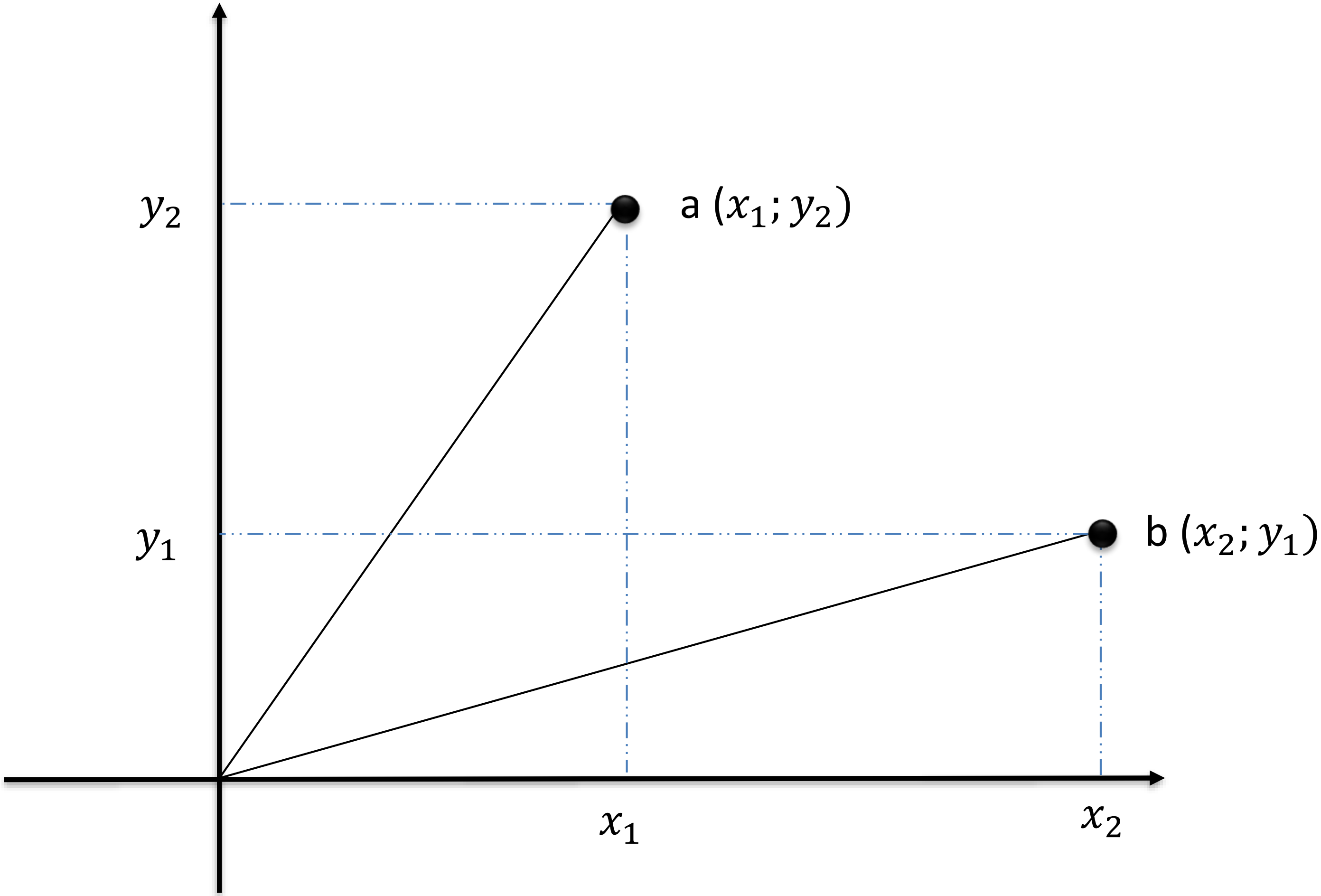
When $p$ = 2; we have the Euclidean distance.

Distance - Chebychev:

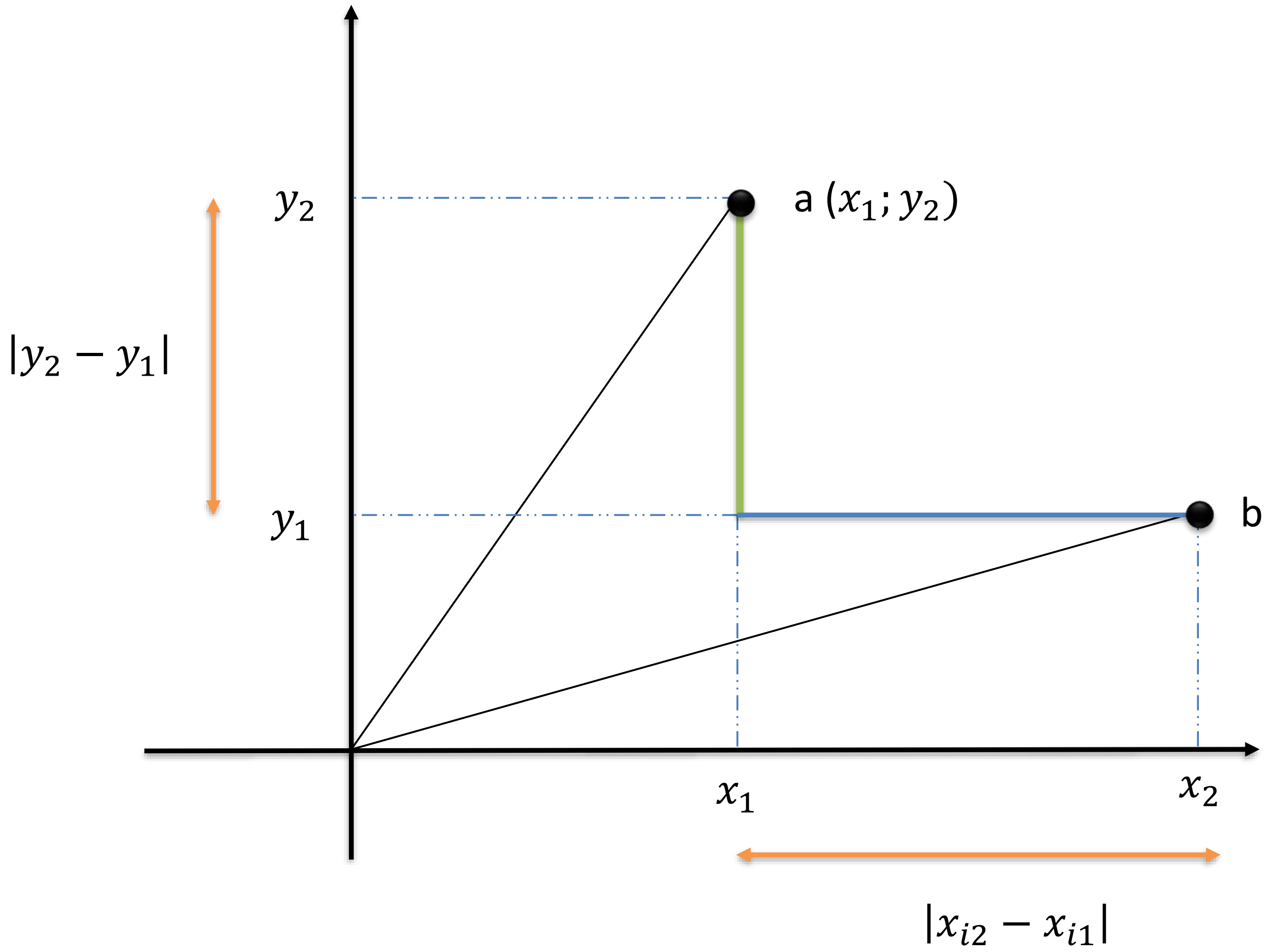
$$d_{cb} = \left( \sum |x_k - y_k|^{\infty} \right)^{\frac{1}{\infty}} = max|x_k - y_k|$$

When $p$ = 0; we have the Chebychev distance.

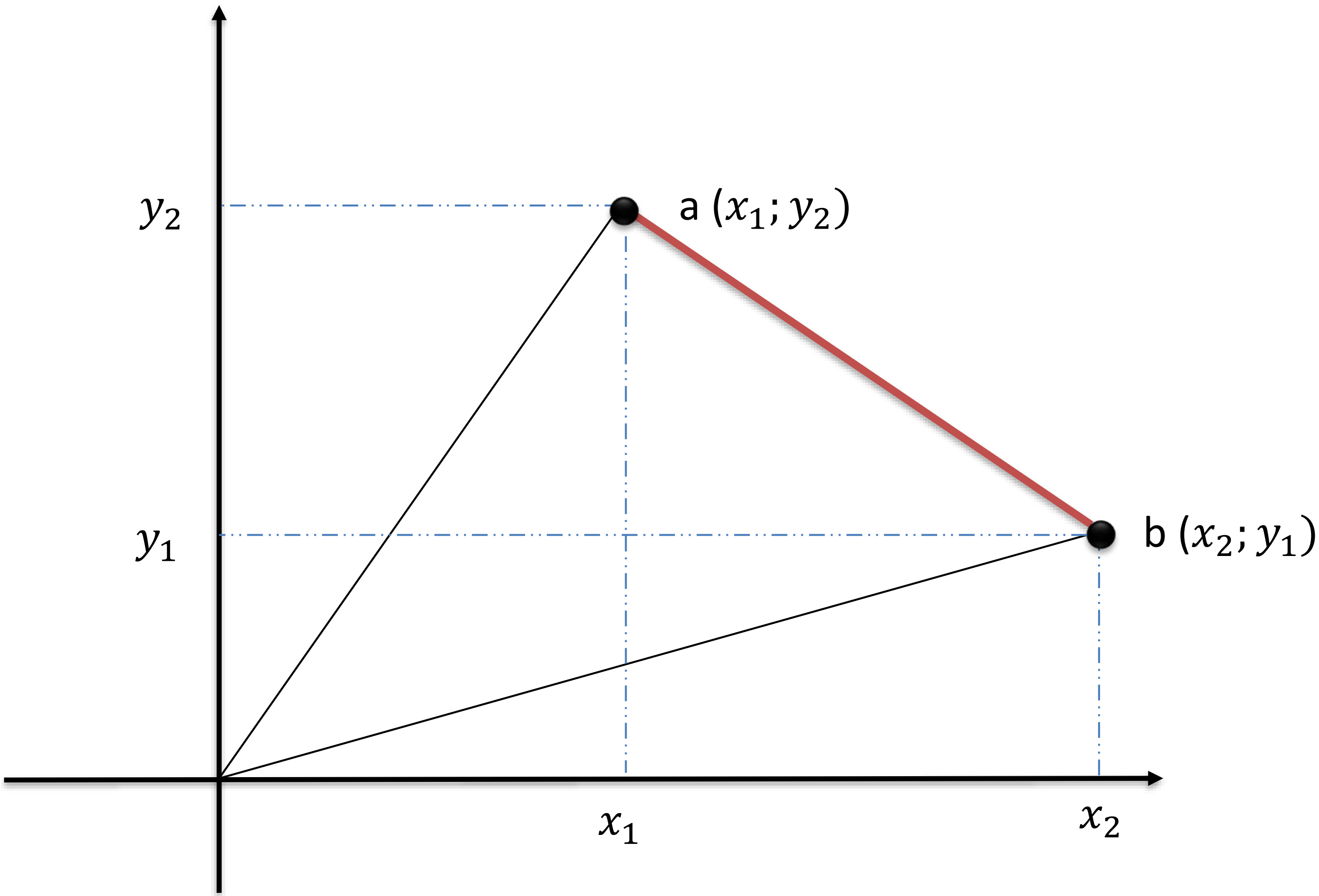
# MVDA – *Multidimensional Scaling*



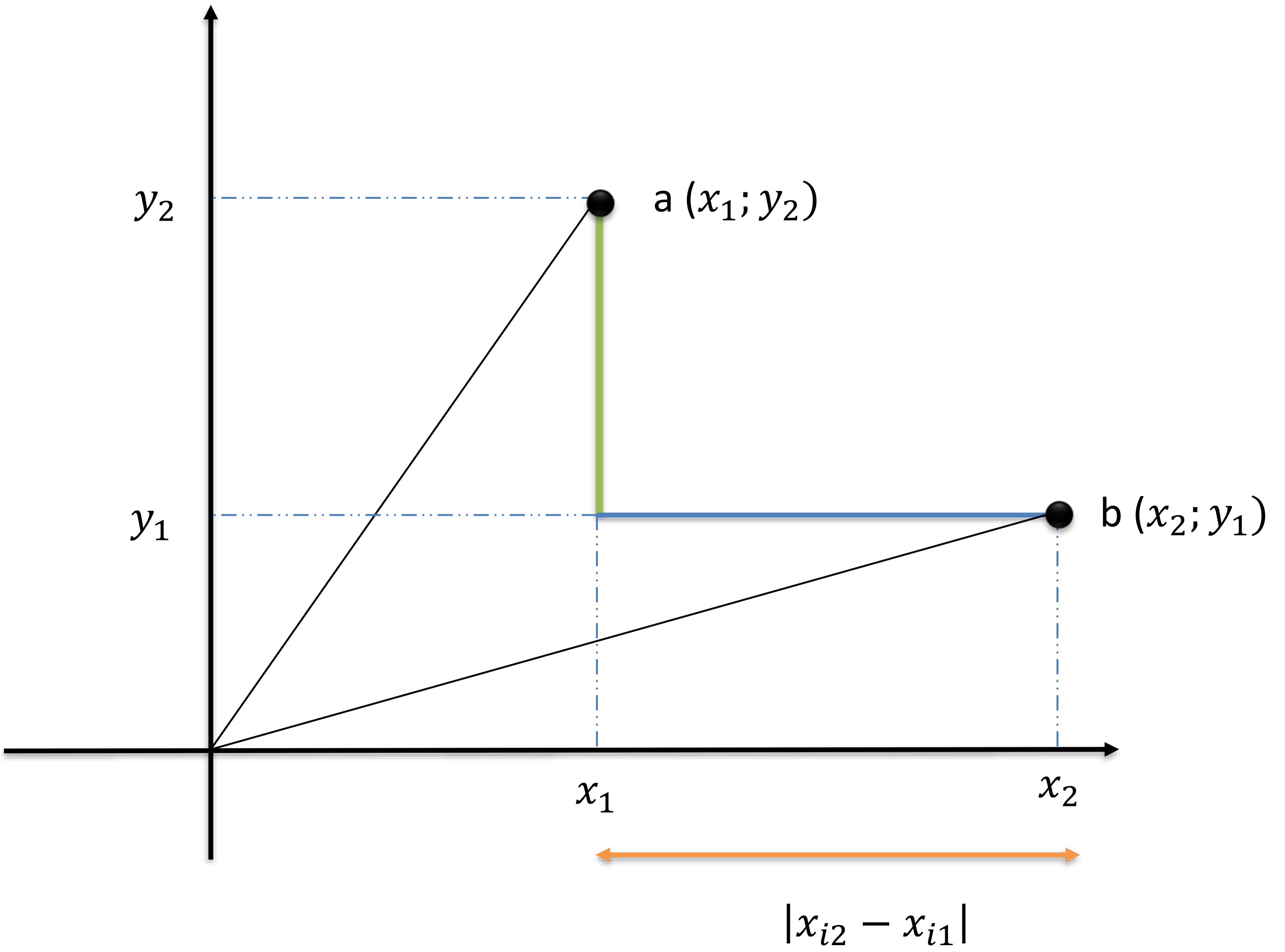
# MVDA – *Multidimensional Scaling*



# MVDA – *Multidimensional Scaling*



# MVDA – *Multidimensional Scaling*



# MVDA – *Multidimensional Scaling*

Counting Dissimilarity - $\chi^2$ (**Chi-square measure**: This measure is based on the chi-square test of equality for two sets of frequencies):

$$d_{\chi 2} = \sqrt{\sum_k \frac{(x_k - E(x_k))^2}{E(x_k)} + \sum_k \frac{(y_k - E(y_k))^2}{E(y_k)}}$$

Counting Dissimilarity - $\varphi^2$ (**Phi-square measure**: This measure is equal to the chi-square measure normalized by the square root of the combined frequency):

$$d_{\varphi^2} = \sqrt{\frac{\sum_k \frac{(x_k - E(x_k))^2}{E(x_k)} + \sum_k \frac{(y_k - E(y_k))^2}{E(y_k)}}{N}}$$

# MVDA – *Multidimensional Scaling*

Counting Dissimilarity - Bray-Curtis:

$$d_{bc} = \frac{\sum |x_k - y_k|}{\sum x_k + y_k}$$

Bray-Curtis is a rank order dissimilarity (used in Non-Metric MDS) between two objects. A value of 0 that the two objects are identical. The similarity can be calculated as:

$$s_{bc} = 1 - d_{bc}$$

A value of 1 means that the two objects are identical.

# MVDA – *Multidimensional Scaling*

Binary Dissimilarity:

| Case 1 | Case 2 | |
|---|---|---|
| | 0 | 1 |
| 0 | a | b |
| 1 | c | d |

Binary Dissimilarity – Binary Euclidean:

$$d_{\{0;1\}ec} = \sqrt{b + c}$$

Binary Dissimilarity – Squared Binary Euclidean:

$$d_{\{0;1\}eq} = b + c$$

# MVDA – *Multidimensional Scaling*

Binary Dissimilarity – Matching Coefficient:

$$d_{\{0;1\}mc} = \frac{b+c}{(a+b+c+d)}$$

Binary Dissimilarity – Jaccard Index:

$$d_{\{0;1\}jc} = \frac{b+c}{(a+b+c)}$$

Binary Dissimilarity – Pattern Difference:

$$d_{\{0;1\}pd} = \frac{bc}{(a+b+c+d)^2}$$

Binary Dissimilarity – Size Difference:

$$d_{\{0;1\}sd} = \frac{(b+c)^2}{(a+b+c+d)^2}$$

Binary Dissimilarity – Variance:

$$d_{\{0;1\}var} = \frac{b+c}{4(a+b+c+d)}$$

Binary Dissimilarity – Lance Williams:

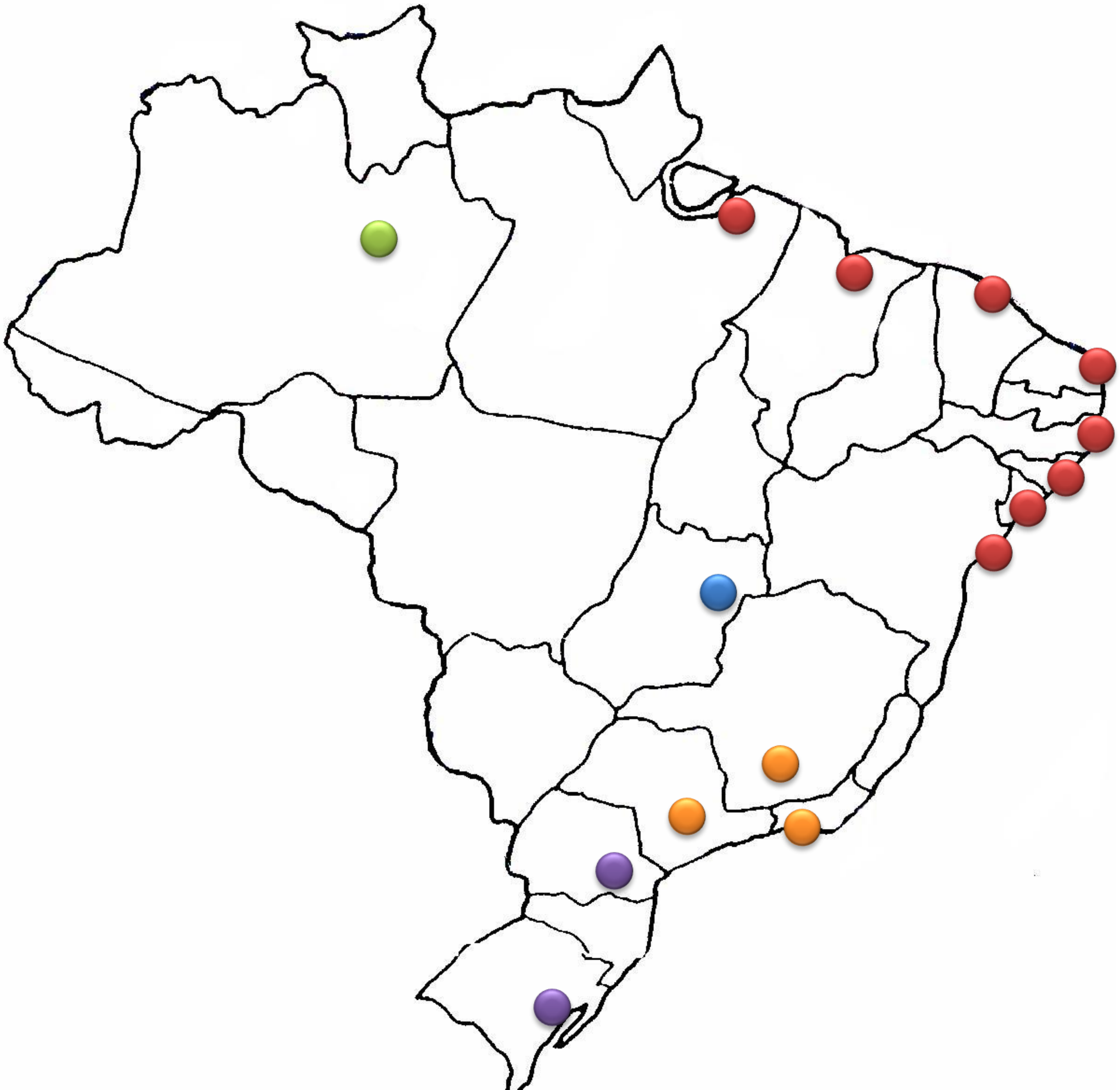
$$d_{\{0;1\}lw} = \frac{b+c}{2a+b+c}$$

# Metric MDS

# MVDA – *Multidimensional Scaling*

In order to explain a **Metric MDS** approach, the following dataset will be used: The collected data represents the Euclidean distance between 15 Brazilian capitals.

| | Manaus | Belem | Sao_Luis | Fortaleza | Natal | Maceio | Recife | Aracaju | Salvador | Brasilia | BH | RJ | SP | Curitiba | Porto_Al |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Manaus | 0 | | | | | | | | | | | | | | |
| Belem | 1288 | 0 | | | | | | | | | | | | | |
| Sao_Luis | 1752 | 493 | 0 | | | | | | | | | | | | |
| Fortaleza | 2295 | 1138 | 640 | 0 | | | | | | | | | | | |
| Natal | 2658 | 1550 | 1035 | 444 | 0 | | | | | | | | | | |
| Maceio | 2670 | 1632 | 1200 | 717 | 435 | 0 | | | | | | | | | |
| Recife | 2823 | 1680 | 1197 | 640 | 252 | 191 | 0 | | | | | | | | |
| Aracaju | 2574 | 1590 | 1237 | 810 | 606 | 202 | 386 | 0 | | | | | | | |
| Salvador | 2617 | 1695 | 1290 | 1018 | 870 | 464 | 654 | 267 | 0 | | | | | | |
| Brasilia | 1967 | 1627 | 1530 | 1682 | 1775 | 1566 | 1632 | 1271 | 1053 | 0 | | | | | |
| BH | 2569 | 2123 | 1848 | 1860 | 1800 | 1416 | 1632 | 1350 | 980 | 589 | 0 | | | | |
| RJ | 2854 | 2460 | 2271 | 2190 | 2122 | 1680 | 1865 | 1485 | 1220 | 900 | 340 | 0 | | | |
| SP | 3100 | 2490 | 2360 | 2238 | 2486 | 1940 | 2135 | 1740 | 1486 | 865 | 500 | 364 | 0 | | |
| Curitiba | 2634 | 2574 | 2514 | 2598 | 2580 | 2205 | 2400 | 2010 | 1734 | 1087 | 823 | 669 | 330 | 0 | |
| Porto_Al | 3987 | 3084 | 3042 | 3126 | 3069 | 2712 | 3083 | 2520 | 2241 | 1617 | 1370 | 1133 | 844 | 547 | 0 |



# MVDA – *Multidimensional Scaling*

```
# MDS
mds <- cmdscale(my_data, 2, eig = TRUE)
mds$points
```

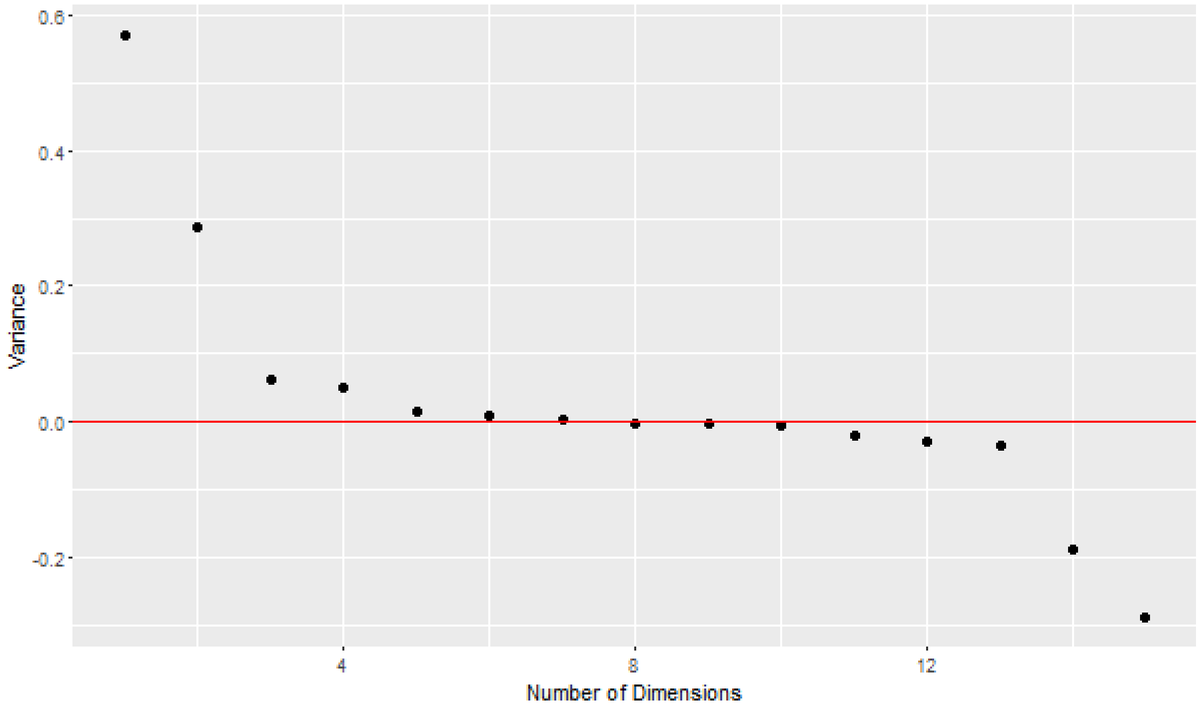
| | [,1] | [,2] |
|---|---|---|
| Aracaju | 291.3074 | 948.45993 |
| Belem | 981.8100 | -381.28651 |
| Belo_Horizonte | -754.9364 | 363.38085 |
| Brasilia | -433.8077 | -54.72129 |
| Curitiba | -1413.0517 | -232.91621 |
| Fortaleza | 985.5966 | 559.84226 |
| Maceio | 460.9759 | 940.98162 |
| Manaus | 1042.2848 | -1871.25345 |
| Natal | 936.4713 | 753.20862 |
| Porto_Alegre | -2034.4776 | 36.75666 |
| Recife | 884.6164 | 878.77954 |
| Rio_Janeiro | -933.8447 | 216.05265 |
| Sao_Luis | 1189.6683 | -159.71171 |
| Sao_Paulo | -1135.0267 | 185.68239 |
| Salvador | 274.0372 | 652.88174 |

# MVDA – *Multidimensional Scaling*

```
# MDS – Extracting 2 Dimensions
mds <- cmdscale(my_data, 2, eig = TRUE)
mds$eig

  [1]   15523283.41  7808551.77  1715852.55  1361090.96   459742.14   230271.54
  [7]   88970.33      -14663.68      -84346.22     -170203.71   -512328.51  -763414.61
[13]  -929346.59    -5096607.32  -7819836.34

mds$GOF[2] # Percentage of variance explained by 2 dimensions

[1] 0.8581742
```

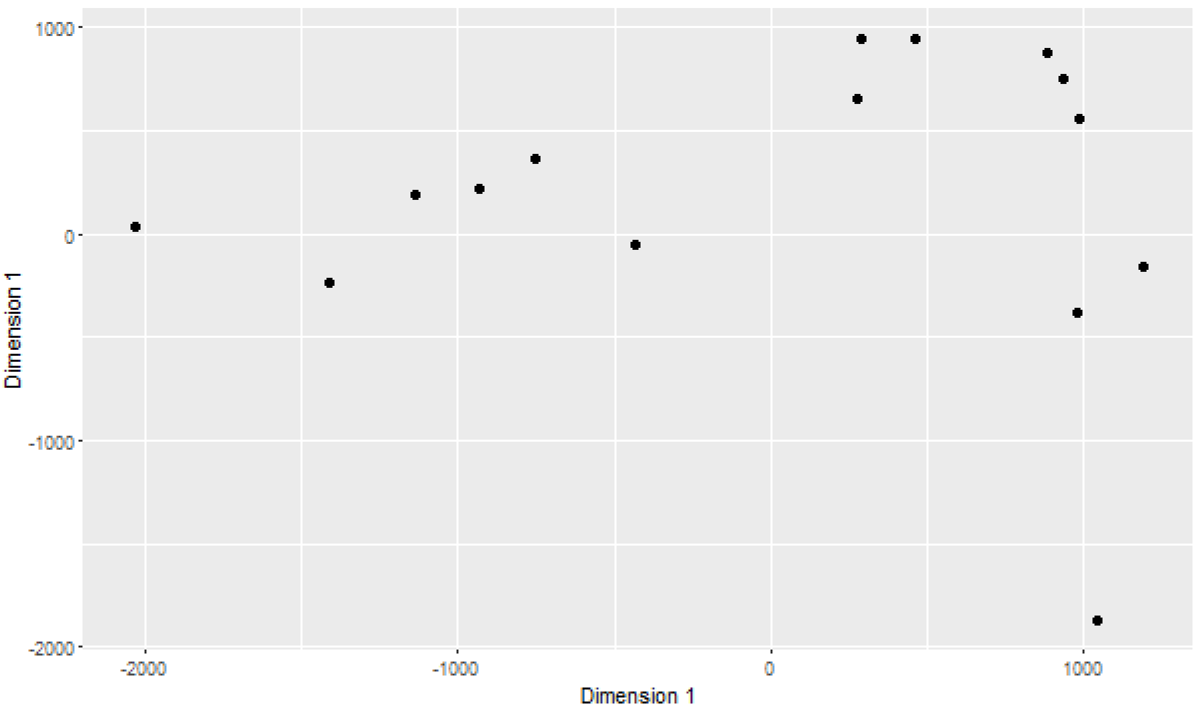
# MVDA – *Multidimensional Scaling*

```
# Variance Plot
library(ggplot2)
ggplot(data = my_data, aes(x = 1:15, y = mds$eig/sum(mds$eig[which(mds$eig > 0)]))) + geom_point(size = 2) +
labs(x = "Number of Dimensions", y = "Variance") + geom_hline(yintercept = 0, color = "red")
```



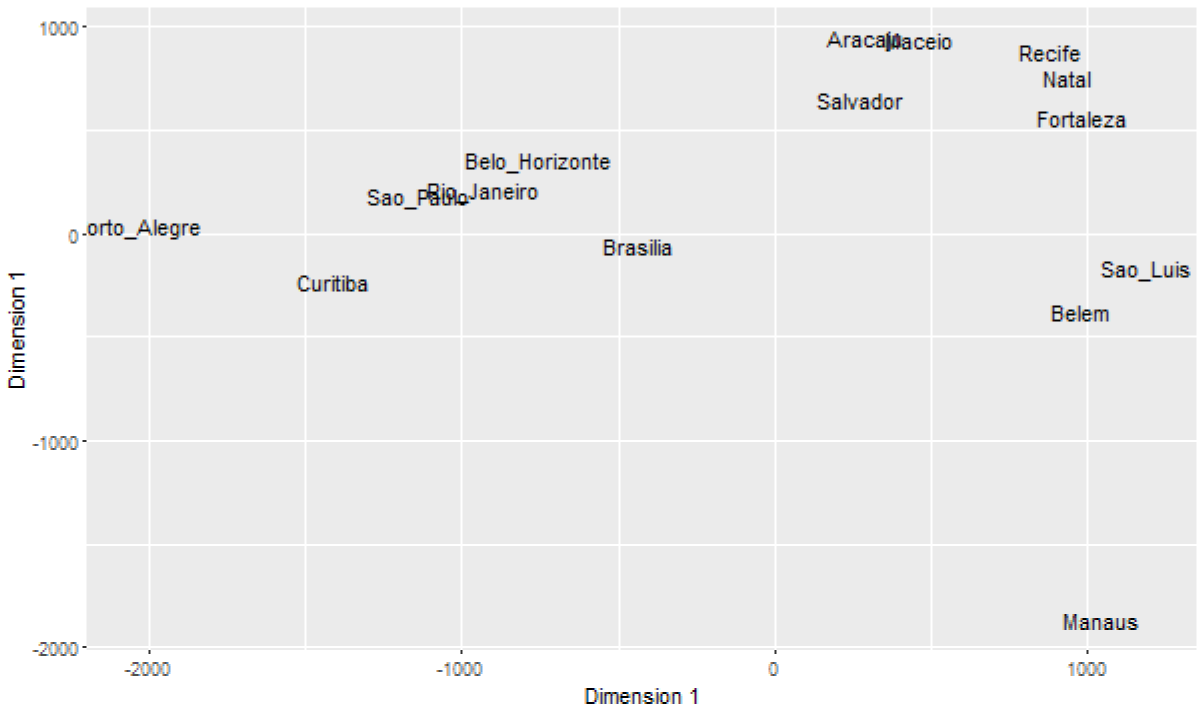
# MVDA – *Multidimensional Scaling*

```
# MDS Plot
dimension_x <- mds$points[, 1]
dimension_y <- mds$points[, 2]
library(ggplot2)
ggplot(data = my_data, aes(x = dimension_x, y = dimension_y, label = row.names(my_data))) + geom_point(size = 2) + labs(x = "Dimension 1", y = "Dimension 1")
```



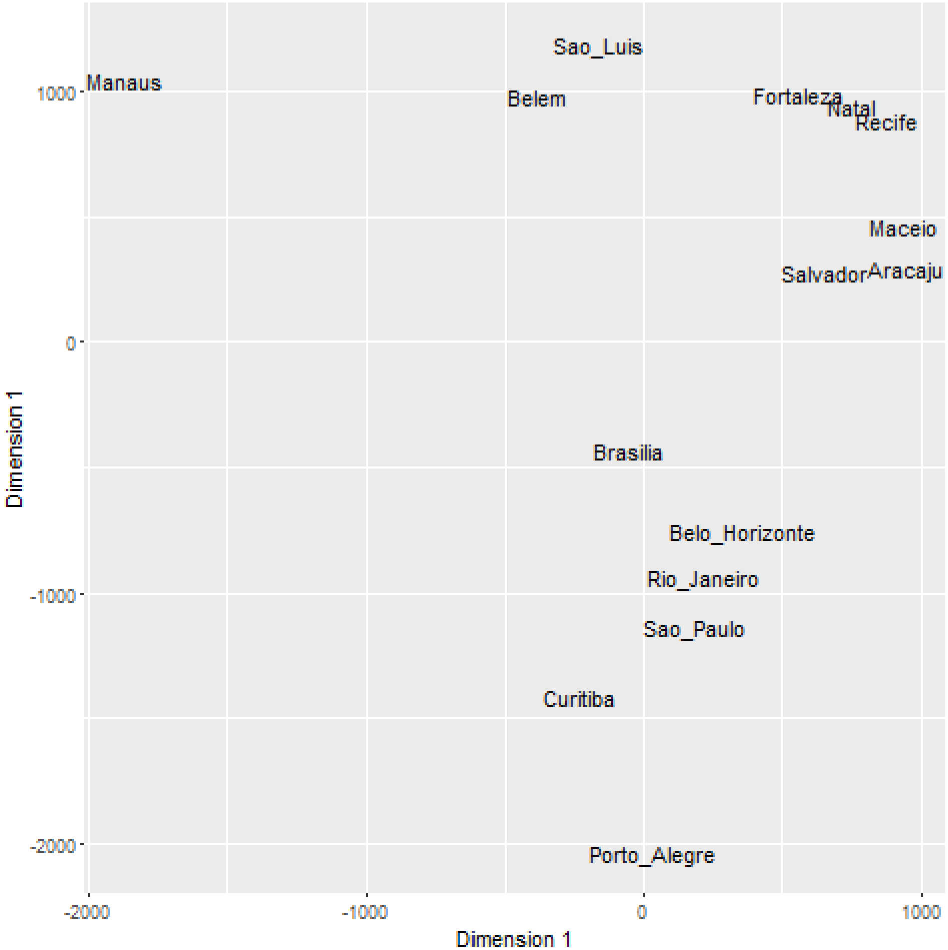
# MVDA – *Multidimensional Scaling*

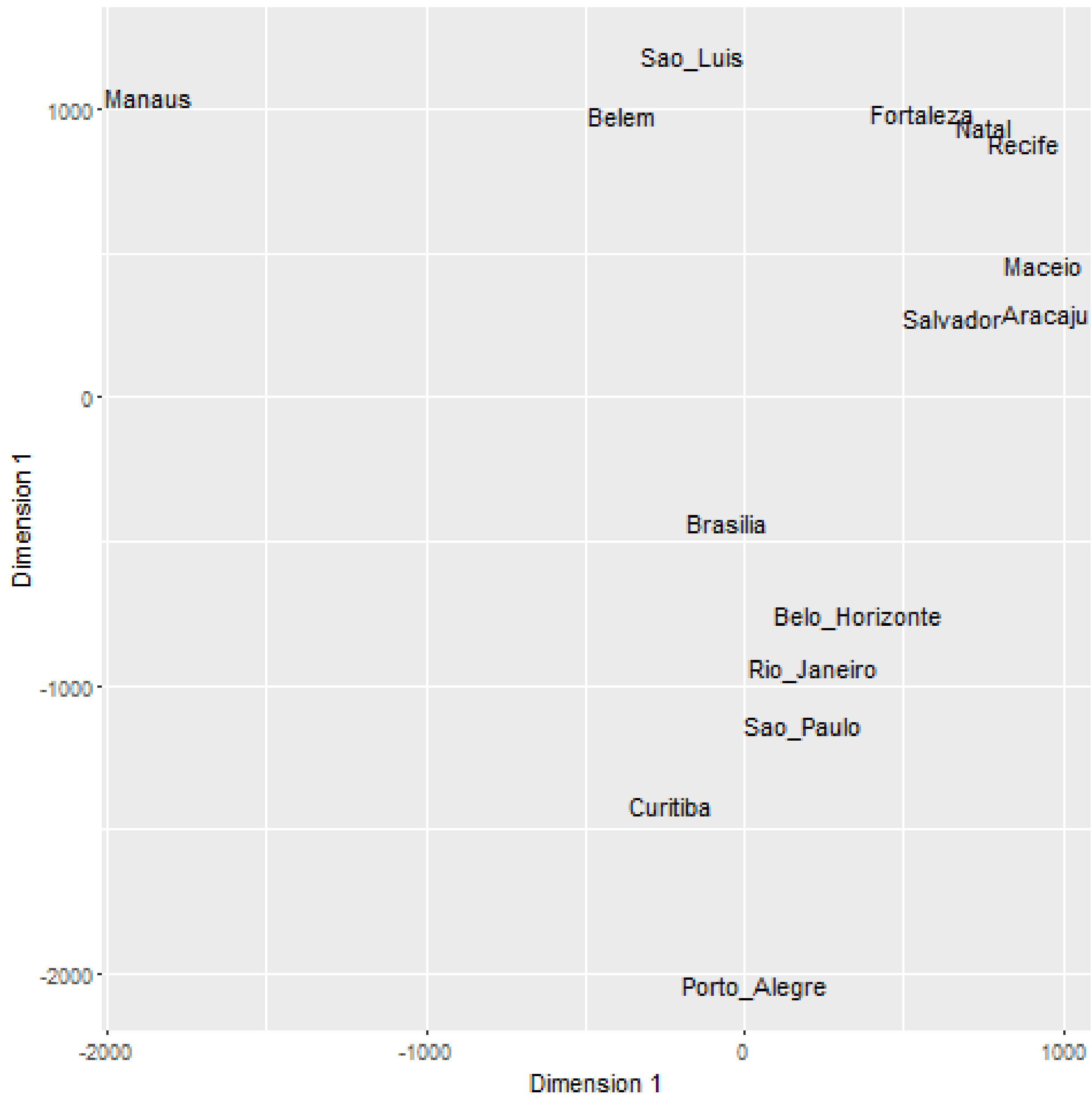
```
# MDS Plot
library(ggplot2)
ggplot(data = my_data, aes(x = dimension_x, y = dimension_y, label = row.names(my_data))) + labs(x = "Dimension 1", y = "Dimension 1") + geom_text(size = 4)
```



# MVDA – *Multidimensional Scaling*

```
# MDS Plot
library(ggplot2)
ggplot(data = my_data, aes(x = dimension_y, y = dimension_x, label = row.names(my_data))) + labs(x = "Dimension 1", y = "Dimension 1") + geom_text(size = 4)
```




Sao_Luis
Manaus
Belem
Fortaleza
Natal
Recife
Maceio
Salvador
Aracaju
Brasilia
Belo_Horizonte
Rio_Janeiro
Sao_Paulo
Curitiba
Porto_Alegre
1000
0
-1000
-2000
-2000
-1000
0
1000
Dimension 1
Dimension 1

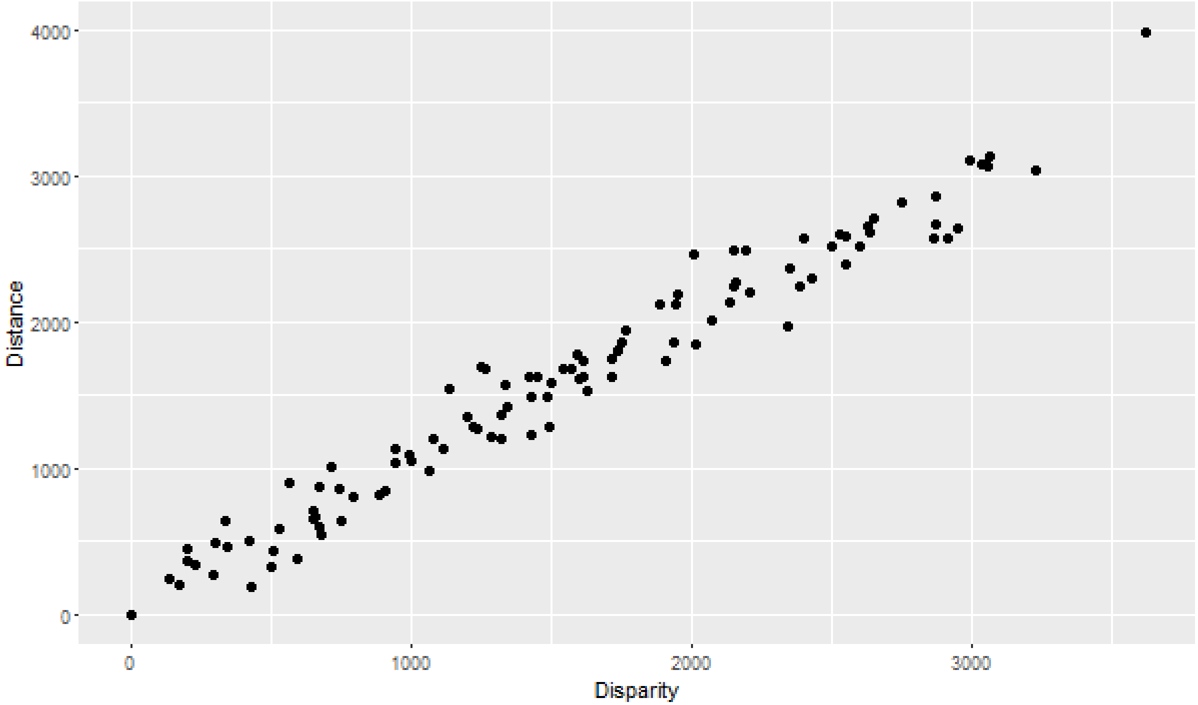
# Goodness of Fit

# MVDA – *Multidimensional Scaling*

```
# Diagnosis
d_hat <- dist(mds$points, method = "euclidean", diag = TRUE, upper = TRUE, p = 2)
cor (c(as.matrix(d_hat)),c(as.matrix(my_data)))
```

[1] 0.9843927

```
# An angle of 45º indicates a good fit.
qplot(x = c(as.matrix(d_hat)), y = c(as.matrix(my_data))) + geom_point(size = 2) + labs(x = "Disparity", y = "Distance")
```



# MVDA – *Multidimensional Scaling*

The Kruskal's $Stress$ (Standardized Residual Sum of Squares) is a dimensionless measure that indicates the model's goodness of fit (lower value = best model).

$p$ = Number of dimensions;
$d_{ij}$ = Distance between objects;
$\delta_{ij}$ = Estimated distance between objects (disparity).

$$Kruskal's\ STRESS\ 2 = \sqrt{\frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\left(d_{ij} - \delta_{ij}\right)^2}{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\left(\delta_{ij} - d_{..}\right)^2}}$$

# MVDA – *Multidimensional Scaling*

The Young's $Stress$ is also dimensionless measure that indicates the model's goodness of fit (lower value = best model).

$$Young's\ STRESS\ 2 = \sqrt{\frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\left({d_{ij}}^{2} - {\delta_{ij}}^{2}\right)^{2}}{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}({\delta_{ij}}^{2} - {d_{..}}^{2})^{2}}}$$

$$Young's\ S\_STRESS\ = \frac{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\left({d_{ij}}^{2} - {\delta_{ij}}^{2}\right)^{2}}{\sum_{i=1}^{p-1}\sum_{j=i+1}^{p}(\delta_{ij})^{4}}$$

# MVDA – *Multidimensional Scaling*

$$Raw\ STRESS = \sum \left(d_{ij} - \delta_{ij}\right)^2$$

$$Normalized\ Raw\ STRESS = \sqrt{\frac{STRESS\ Bruto}{\sum {d_{ij}}^2}}$$

ou

$$Normalized\ STRESS = \sqrt{\frac{STRESS\ Bruto}{\sum (d_{ij} - d_{..})^2}}$$

# MVDA – *Multidimensional Scaling*

The following interpretation is suggested:

| Stress | Goodness of Fit |
|---|---|
| 20.0% | Poor |
| 10.0% | Fair |
| 05.0% | Good |
| 02.5% | Excelent |
| 00.0% | Perfect |

# MVDA – *Multidimensional Scaling*

The following interpretation is suggested:

| Stress | Goodness of Fit |
|---|---|
| 20.0% | Poor |
| 10.0% | Fair |
| 05.0% | Good |
| 02.5% | Excelent |
| 00.0% | Perfect |

**Rule of Thumb: Excelent $\leq 10\%$ ; Acceptable $> 10\%$ & $< 15\%$; Inacceptable $\geq 15\%$**

# MVDA – *Multidimensional Scaling*

Other adjustment measures can also be checked.

- $RSQ$ ($r\text{-}squared$) = Values close to 1, indicates a good fit.

$$RSQ = \frac{\left(\sum_i \sum_j \left(\delta_{ij} - \delta_{..}\right).\left(d_{ij} - d_{..}\right)\right)^2}{\left(\sum_i \sum_j \left(\delta_{ij} - \delta_{..}\right)^2\right).\left(\sum_i \sum_j \left(\left(d_{ij} - d_{..}\right)\right)^2\right)}$$

# MVDA – *Multidimensional Scaling*

```
# Diagnosis
d_hat <- as.matrix(d_hat)


K_STRESS_2 <- ((sum((my_data - d_hat)^2))/sum(((d_hat - sum(my_data))^2)))^(1/2)

[1] 0.0004934325


Y_STRESS_2 <- ((sum(((my_data)^2 - d_hat^2)^2))/sum(((d_hat^2 - sum(my_data)^2)^4)))^(1/2)

[1] 4.776563e-17

Y_STRESS_S <-  (sum(((my_data)^2 - d_hat^2)^2))/sum(((d_hat^2 - sum(my_data)^2)^4))

[1] 2.281556e-33
```

# MVDA – *Multidimensional Scaling*

```
# Diagnosis
d_hat <- as.matrix(d_hat)

R_STRESS   <-   sum((my_data - d_hat)^2)

[1] 6345153

N_R_STRESS <-  (sum((my_data - d_hat)^2)/sum((my_data)^2))^(1/2)

[1] 0.09468024

N_STRESS   <-  (sum((my_data - d_hat)^2)/sum((my_data - sum(my_data))^2))^(1/2)

[1] 0.0004934924

RSQ <- ((sum((my_data - sum(my_data))*(d_hat - sum(d_hat))))^2)/(sum((my_data - sum(my_data))^2)*sum((d_hat - sum(d_hat))^2))

[1] 0.9999998
```

# Non-Metric MDS

# MVDA – *Multidimensional Scaling*

In order to explain a **Non-Metric MDS** approach, the following dataset will be used: A questionnaire consisting of 12 questions that were applied to 185 students in a college. Each question was answered by a 4-point Likert Scale (1 = Very Bad, 2 = Poor, 3 = Good, 4 = Very Good). The questions were:

| Code | Questions (Variables) |
|---|---|
| PR01 | How do you evaluate your commitment to the course? |
| PR02 | Attendance to classes? |
| PR03 | Punctuality? |
| PR04 | Frequency to Library? |
| PR05 | How do you evaluate your general reading habit (literature, magazines, newspapers, internet, etc.)? |
| PR06 | How do you evaluate your reading habit focused on the course? |
| PR07 | Involvement and participation in institutional events (plenary lecture, seminars, lectures, etc.)? |
| PR08 | Involvement and participation in the events of your course (seminars, lectures)? |
| PR09 | Involvement and participation in classroom discussions? |
| PR10 | Involvement and participation in student representation in the institution? |
| PR11 | Level of learning about the content taught by teachers? |
| PR12 | You commitment in doing the activities recommended by the teacher? |

# MVDA – *Multidimensional Scaling*

```
# nMDS
my_data2 <- R.MVDA.nMDS
library(vegan)
n_mds <- metaMDS(my_data2, k = 2, trymax = 100)

Call:
metaMDS(comm = my_data2, k = 2, trymax = 100)

global Multidimensional Scaling using monoMDS

Data:     my_data2
Distance: bray

Dimensions: 2
Stress:     0.2306176
Stress type 1, weak ties
No convergent solutions - best solution after 100 tries
Scaling: centring, PC rotation, halfchange scaling
Species: expanded scores based on 'my_data2'
```

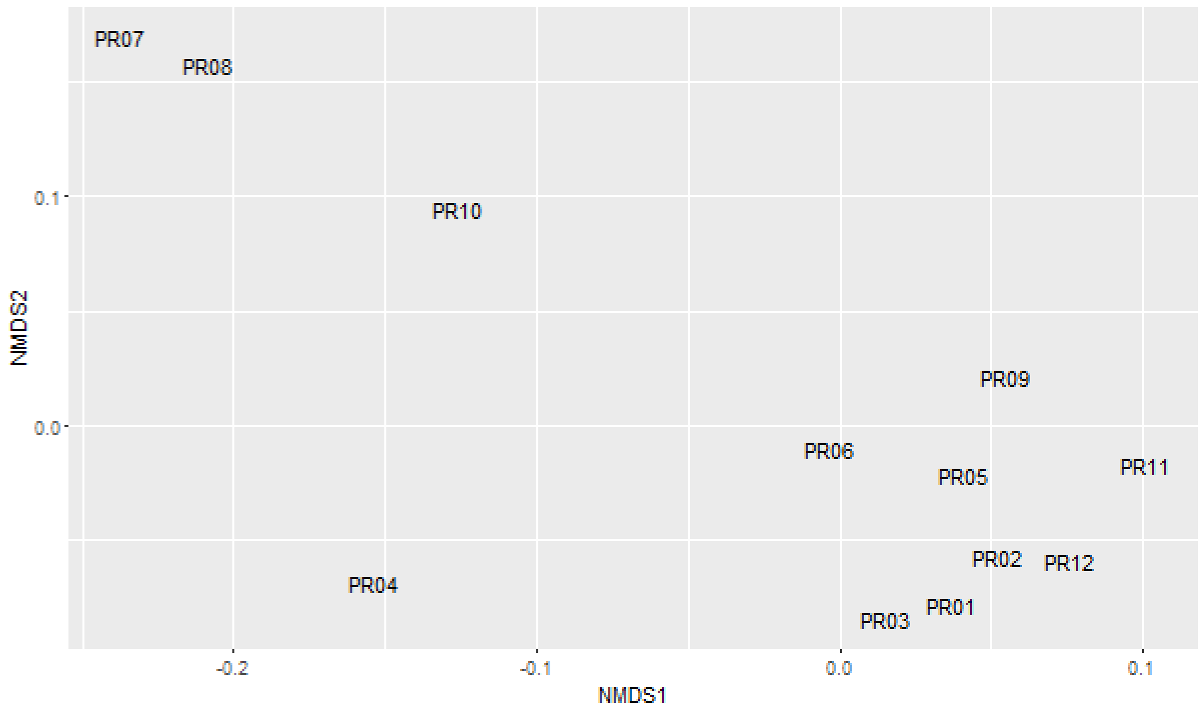
# MVDA – *Multidimensional Scaling*

The following interpretation is suggested:

| Stress | Goodness of Fit |
|---|---|
| $> 30.0\%$ | Poor |
| $> 20.0\% \ \& \le 30.0\%$ | Fair |
| $> 10.0\% \ \& \le 20.0\%$ | Good |
| $> 05.0\% \ \& \le 10.0\%$ | Great |
| $\le 05.0\%$ | Excelent |
| $00.0\%$ | Perfect |

# MVDA – *Multidimensional Scaling*

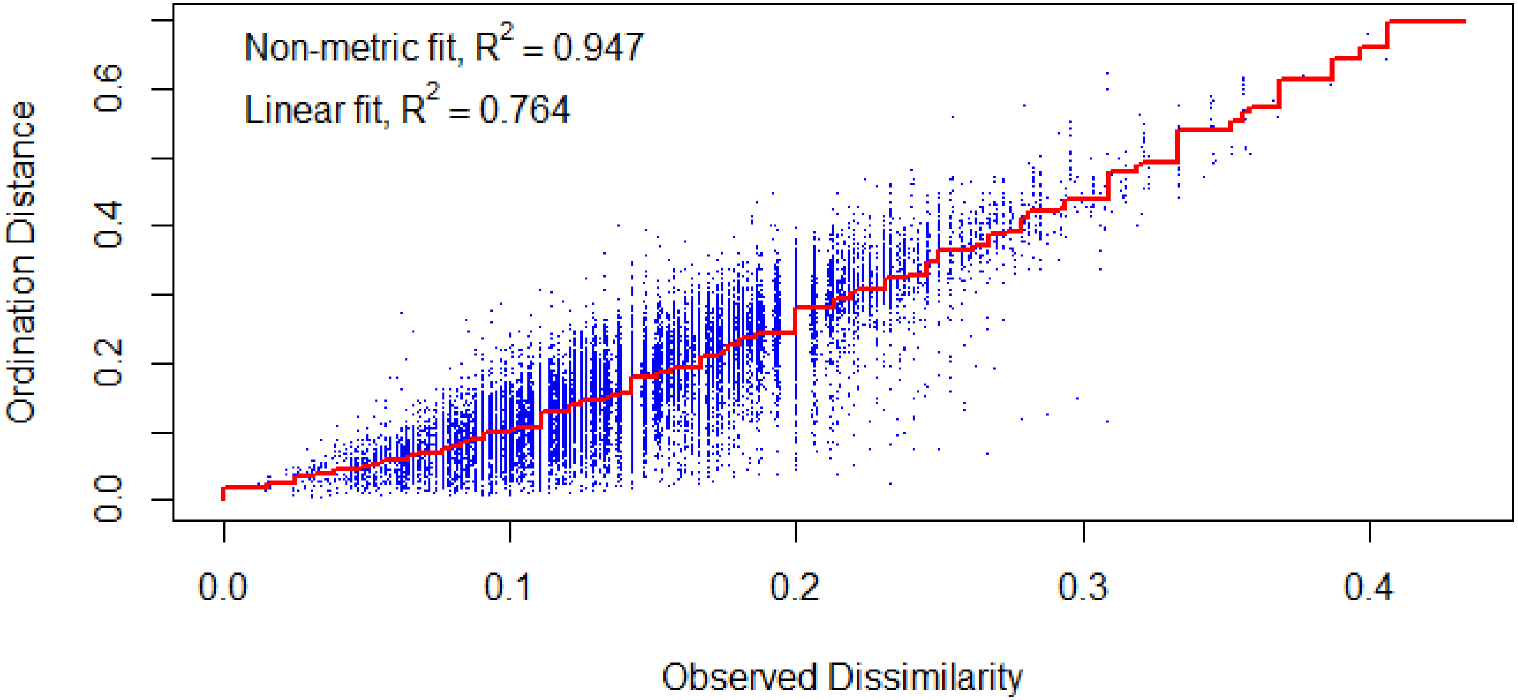
```
# Plot nMDS Variables
library(ggplot2)
n_mds_variables <- scores(n_mds, "spec")
n_mds_variables <- cbind.data.frame(n_mds_variables, label = rownames(n_mds_variables))
ggplot(data = n_mds_variables, aes(x = NMDS1, y = NMDS2)) + geom_text(aes(label = label))
```



# Goodness of Fit

# MVDA – *Multidimensional Scaling*

```
# Shepard Plot
stressplot(n_mds)
```



# 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: Exploratory Factor Analysis

Citation:
**PEREIRA, V. (2016). Project: Multivariate Data Analysis, File: R-MVDA-04-MDS.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.