Simple & Multiple Correspondence Analyses

Contingency, categorical, ordinal, continuous and mixed data

Derek Beaton

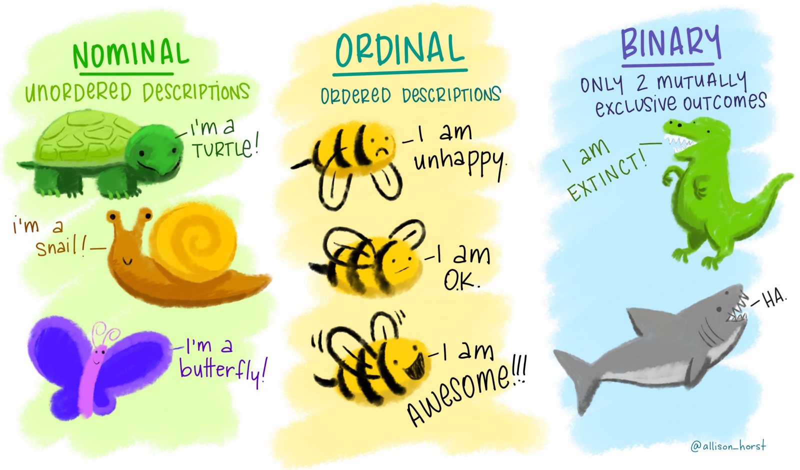
Rotman Research Institute

October 27, 2019

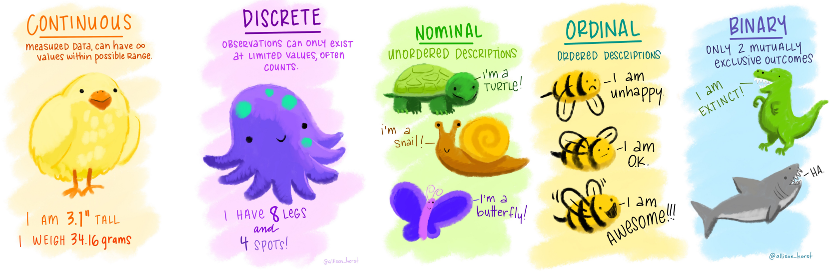
Before we get started

Our new best friends

via @allison_horst



via @allison_horst



via @allison_horst

Motivation for today

- Not everything is a number

Motivation for today

- Not everything is a number
- Sometimes numbers aren't numbers!

Motivation for today

- Not everything is a number
- Sometimes numbers aren't numbers!
- We need to recognize when this happens

Motivation for today

- Not everything is a number
- Sometimes numbers aren't numbers!
- We need to recognize when this happens
  - And know what to do

Typology

- SS Stevens (not a boat!)

Typology

- SS Stevens (not a boat!)
- Levels of measurement

Typology

- SS Stevens (not a boat!)
- Levels of measurement
- Excellent examples: https://en.wikipedia.org/wiki/Level_of_measurement

Where to find everything

- Generally: https://github.com/derekbeaton/workshops

Where to find everything

- Generally: https://github.com/derekbeaton/workshops
- Today:

Overview

- Revisit PCA

Overview

- Revisit PCA
- Looking at some data

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types
- A whole bunch of bonuses

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types
- A whole bunch of bonuses
  - Robustness

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types
- A whole bunch of bonuses
  - Robustness
  - PLS

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
  - and many of its connections
- Multiple correspondence analysis
  - generalizes CA (amongst many other things)
  - and how to handle various data types
- A whole bunch of bonuses
  - Robustness
  - PLS
  - Networks

Overview

- Revisit PCA
- Looking at some data
- Simple correspondence analysis
    - and many of its connections
- Multiple correspondence analysis
    - generalizes CA (amongst many other things)
    - and how to handle various data types
- A whole bunch of bonuses
    - Robustness
    - PLS
    - Networks
    - Software

Revisting PCA

What is PCA for?

- When we can compute a covariance or correlation matrix

What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components

What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components
  - Orthogonal

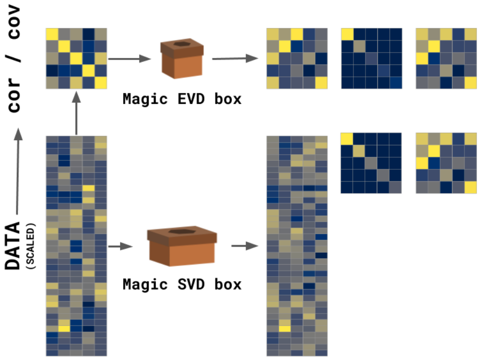
What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components
    - Orthogonal
    - Rank ordered

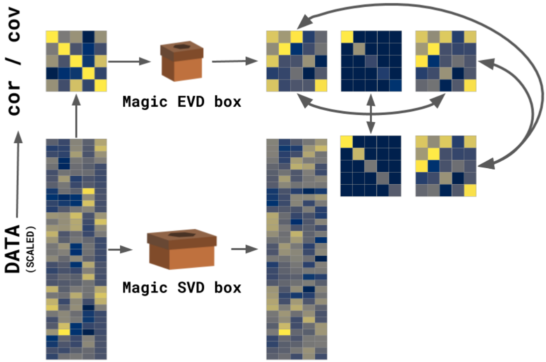
What is PCA for?

- When we can compute a covariance or correlation matrix
- Break data into components
  - Orthogonal
  - Rank ordered
  - Made of bits & pieces of original measures

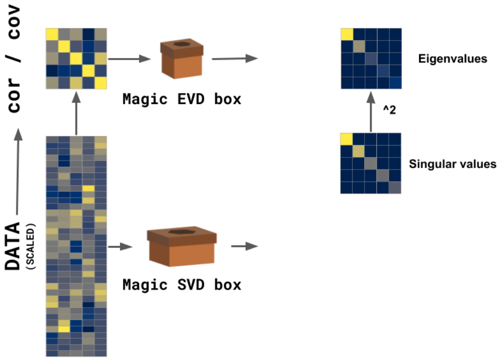
Eigen- and singular value decompositions



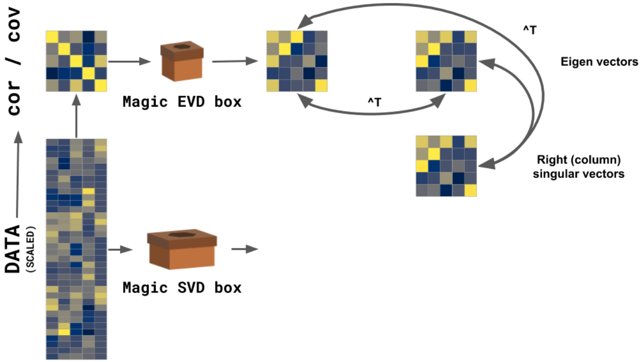
Eigen- and singular value decompositions



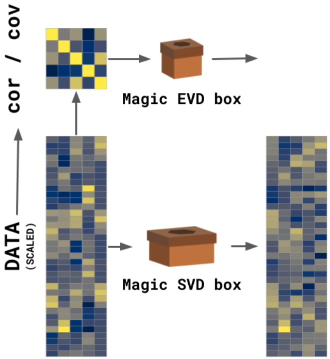
Eigen- and singular value decompositions



Eigen- and singular value decompositions



Eigen- and singular value decompositions



Left (row) singular vectors

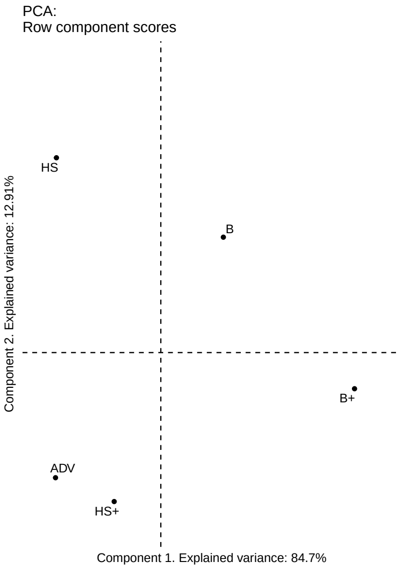
Some data

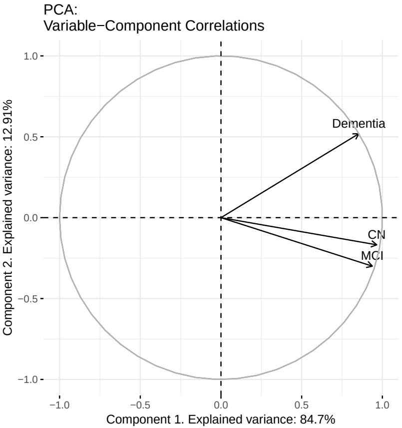
Diagnosis and education

	CN	Dementia	MCI
ADV	39	7	54
B	57	17	75
B+	75	19	113
HS	25	13	46
HS+	39	9	77

- Given you a table, and asked for a multivariate analysis

- Given you a table, and asked for a multivariate analysis
- We do what we know: PCA


PCA:
Row component scores
HS
B
B+
ADV
HS+
Component 2. Explained variance: 12.91%
Component 1. Explained variance: 84.7%


PCA:
Variable−Component Correlations
Dementia
CN
MCI
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Component 2. Explained variance: 12.91%
Component 1. Explained variance: 84.7%

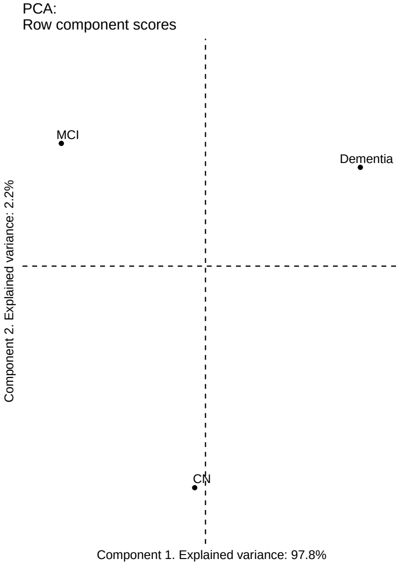
What did we analyze?

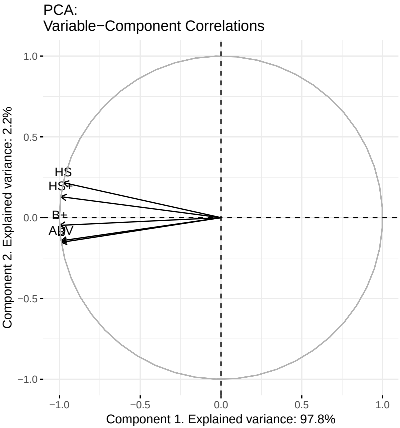
	CN	Dementia	MCI
CN	1.000	0.730	0.921
Dementia	0.730	1.000	0.652
MCI	0.921	0.652	1.000

What did PCA detect?

Let's try something different!

	ADV	B	B+	HS	HS+
CN	39	57	75	25	39
Dementia	7	17	19	13	9
MCI	54	75	113	46	77


PCA:
Row component scores
MCI
Dementia
CN
Component 2. Explained variance: 2.2%
Component 1. Explained variance: 97.8%


PCA:
Variable−Component Correlations
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
HS
HS+
B+
ABV
Component 2. Explained variance: 2.2%
Component 1. Explained variance: 97.8%

What did PCA analyze?

	ADV	B	B+	HS	HS+
ADV	1.000	1.000	0.995	0.935	0.963
B	1.000	1.000	0.994	0.932	0.960
B+	0.995	0.994	1.000	0.965	0.984
HS	0.935	0.932	0.965	1.000	0.996
HS+	0.963	0.960	0.984	0.996	1.000

What did PCA detect?

	ADV	B	B+	HS	HS+	**Row sums**
CN	39	57	75	25	39	**235**
Dementia	7	17	19	13	9	**65**
MCI	54	75	113	46	77	**365**

What is PCA for?

- When we can compute a *meaningful* covariance or correlation matrix

Let's take another look

```
## Warning in rbind(cbind(edu_dx_table, rowSums(edu_dx_tabl
## colSums(edu_dx_table)): number of columns of result is n
## vector length (arg 2)
```

NA

Simple correspondence analysis

History

- CA

History

- CA
  - Hirschfeld (1935)

History

- CA
  - Hirschfeld (1935)
  - Guttman (1941)

History

- CA
  - Hirschfeld (1935)
  - Guttman (1941)
  - Burt (1950)

History

- CA
  - Hirschfeld (1935)
  - Guttman (1941)
  - Burt (1950)
  - Benzecri (1964)

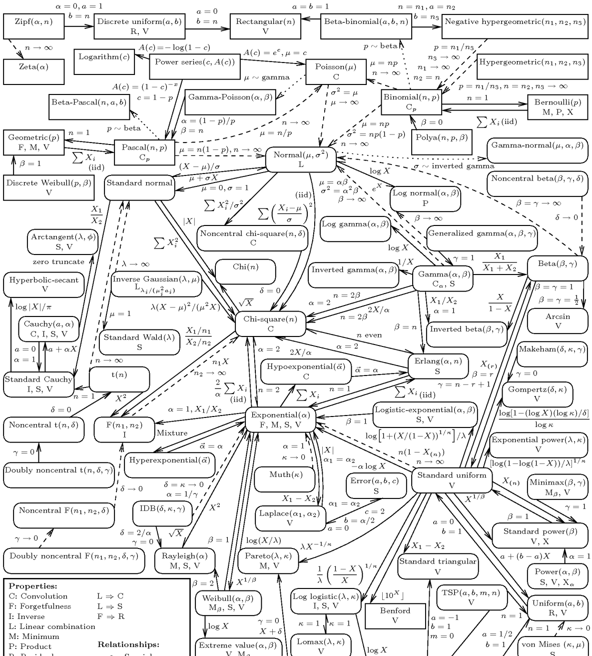
History

- CA
  - Hirschfeld (1935)
  - Guttman (1941)
  - Burt (1950)
  - Benzecri (1964)
  - Escofier (1965)

History

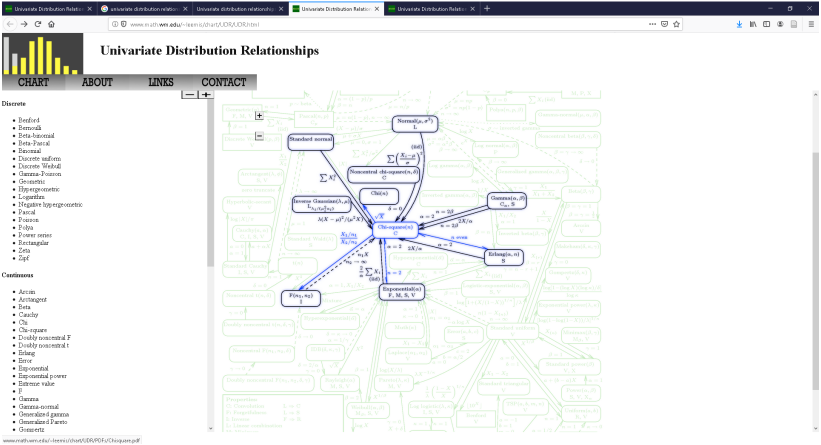
- CA
  - Hirschfeld (1935)
  - Guttman (1941)
  - Burt (1950)
  - Benzecri (1964)
  - Escofier (1965)
- See Lebart's History & Prehistory of CA: http://www.dtmvic.com/doc/About_the_History_of_CA.pdf

Chi-squared



Zipf(α, n)
Discrete uniform(a, b) R, V
Rectangular(n) V
Beta-binomial(a, b, n)
Negative hypergeometric(n_1, n_2, n_3)
Zeta(α)
Logarithm(c)
Power series($c, A(c)$)
Poisson(μ) C
Hypergeometric(n_1, n_2, n_3)
Beta-Pascal(n, a, b)
Gamma-Poisson(α, β)
Binomial(n, p) C$_p$
Bernoulli(p) M, P, X
Geometric(p) F, M, V
Pascal(n, p) C$_p$
Polya(n, p, β)
Normal(μ, σ^2) L
Gamma-normal(μ, α, β)
Discrete Weibull(p, β) V
Standard normal
Noncentral beta(β, γ, δ)
Log normal(α, β) P
Arctangent(λ, ϕ) S, V
Noncentral chi-square(n, δ) C
Log gamma(α, β)
Generalized gamma(α, β, γ)
Hyperbolic-secant V
Chi(n)
Inverse Gaussian(λ, μ)
Inverted gamma(α, β)
Gamma(α, β) C$_\alpha$, S
Beta(β, γ)
Cauchy(a, α) C, I, S, V
Chi-square(n) C
Arcsin V
Standard Wald(λ) S
Inverted beta(β, γ)
Makeham(δ, κ, γ)
Standard Cauchy I, S, V
t(n)
Hypoexponential($\vec{\alpha}$) C
Erlang(α, n) S
Gompertz(δ, κ) V
Noncentral t(n, δ)
F(n_1, n_2) I
Exponential(α) F, M, S, V
Logistic-exponential(α, β) S, V
Exponential power(λ, κ) V
Doubly noncentral t(n, δ, γ)
Hyperexponential($\vec{\alpha}$)
Muth(κ)
Error(a, b, c) S
Standard uniform V
Minimax(β, γ) M$_\beta$, V
Noncentral F(n_1, n_2, δ)
IDB(δ, κ, γ)
Laplace(α_1, α_2) V
Standard power(β) V, X
Doubly noncentral F(n_1, n_2, δ, γ)
Rayleigh(α) M, S, V
Pareto(λ, κ) M, V
Standard triangular V
Power(α, β) S, V, X$_\alpha$
Weibull(α, β) M$_\beta$, S, V
Log logistic(λ, κ) I, S, V
Benford V
TSP(a, b, m, n) V
Uniform(a, b) R, V
Extreme value(α, β)
Lomax(λ, κ) V
von Mises (κ, μ)
Properties:
C: Convolution
F: Forgetfulness
I: Inverse
L: Linear combination
M: Minimum
P: Product
L ⇒ C
L ⇒ S
F ⇒ R
Relationships:

Chi-squared



See here

Under the hood

- The eigenvalue decomposition (EVD)

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite
  - Generally correlation or covariance

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite
  - Generally correlation or covariance
- The singular value decomposition (SVD)

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite
  - Generally correlation or covariance
- The singular value decomposition (SVD)
  - Works with rectangular tables

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite
  - Generally correlation or covariance
- The singular value decomposition (SVD)
  - Works with rectangular tables
- The generalized SVD

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite
  - Generally correlation or covariance
- The singular value decomposition (SVD)
  - Works with rectangular tables
- The generalized SVD
  - Apply constraints (weights) to rows & columns of rectangular table

Under the hood

- The eigenvalue decomposition (EVD)
  - Requires squares, symmetric, and positive semi definite
  - Generally correlation or covariance
- The singular value decomposition (SVD)
  - Works with rectangular tables
- The generalized SVD
  - Apply constraints (weights) to rows & columns of rectangular table
  - Required for CA and fancier PCA-like techniques & extensions

The GSVD

Multiple correspondence analysis

Some many bonuses!

(Some) References

See the reference sections of these

- Beaton, D., Saporta, G., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2019). A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data. bioRxiv, 598888.

See the reference sections of these

- Beaton, D., Saporta, G., Abdi, H., & Alzheimer's Disease Neuroimaging Initiative. (2019). A generalization of partial least squares regression and correspondence analysis for categorical and mixed data: An application with the ADNI data. bioRxiv, 598888.
- Beaton, D., Sunderland, K. M., Levine, B., Mandzia, J., Masellis, M., Swartz, R. H., . . . & Strother, S. C. (2019). Generalization of the minimum covariance determinant algorithm for categorical and mixed data types. bioRxiv, 333005.

And these

- Abdi, H., Guillemot, V., Eslami, A., & Beaton, D. (2017). Canonical correlation analysis. Encyclopedia of Social Network Analysis and Mining, 1-16.

And these

- Abdi, H., Guillemot, V., Eslami, A., & Beaton, D. (2017). Canonical correlation analysis. Encyclopedia of Social Network Analysis and Mining, 1-16.
- Beaton, D., Dunlop, J., & Abdi, H. (2016). Partial least squares correspondence analysis: A framework to simultaneously analyze behavioral and genetic data. Psychological methods, 21(4), 621.

Techniques

- Greenacre, M. (2017). Correspondence analysis in practice. CRC press.

Techniques

- Greenacre, M. (2017). Correspondence analysis in practice. CRC press.
- Greenacre, M. J. (1984). Theory and Applications of Correspondence Analysis. Retrieved from http://books.google.com/books?id=LsPaAAAAMAAJ

Techniques

- Greenacre, M. J. (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114

Techniques

- Greenacre, M. J. (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114
- Lebart, L., Morineau, A., & Warwick, K. M. (1984). Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley.

Techniques

- Greenacre, M. J. (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5), 613–619. https://doi.org/10.1002/wics.114
- Lebart, L., Morineau, A., & Warwick, K. M. (1984). Multivariate descriptive statistical analysis: correspondence analysis and related techniques for large matrices. Wiley.
- Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLOS Computational Biology, 15(6), e1006907.

Data

- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée, 26(4), 29–37.

Data

- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée, 26(4), 29–37.
- Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Cahiers de l'Analyse Des Données, 4(2), 137–146.

Data

- Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. Revue de Statistique Appliquée, 26(4), 29–37.
- Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. Cahiers de l'Analyse Des Données, 4(2), 137–146.
- Greenacre, M. (2014). Data Doubling and Fuzzy Coding. In J. Blasius & M. Greenacre (Eds.), Visualization and Verbalization of Data (pp. 239–253). Philadelphia, PA, USA: CRC Press.