SOM-based algorithms for qualitative variables

Roussos Ioannis

Motivation

- Data analysis (Factorial methods)

Find subspace and new axes ➔ *dimensionality reduction*

- **P**rincipal **C**omponents **A**nalysis (quantitative variables)
- **F**actorial **C**orrespondence **A**nalysis } *qualitative*
- **M**ultiple **C**orrespondence **A**nalysis } variables

- Analyze *qualitative* variables
- Use SOM…
  - SOM algorithm == PCA
  - FCA ➔ 2*PCA
  - Generalization to many variables ➔ MCA

Qualitative Variables

- variables are classified as
  - *(a) nominal, (b) ordinal, (c) interval or (d) ratio*
- Quantitative variables
  - Distance can be defined.
- Qualitative variables
  - No ordering.
  - No distance.
  - Each item just belongs to a category.
  - e.g. race, gender, education, city etc.

Data representation

- Quantitative variables

	x	y	z
v1	12	22	5
…	…	…	…
v_n	3	51	47

- Qualitative variables (Contingency Table)

	Athens	Patra	Hania
High	234	155	88
Normal	567	300	160
Low	200	120	74

Encoding modalities as quantitative variables has no meaning

Kohonen's Self Organizing Map (SOM)

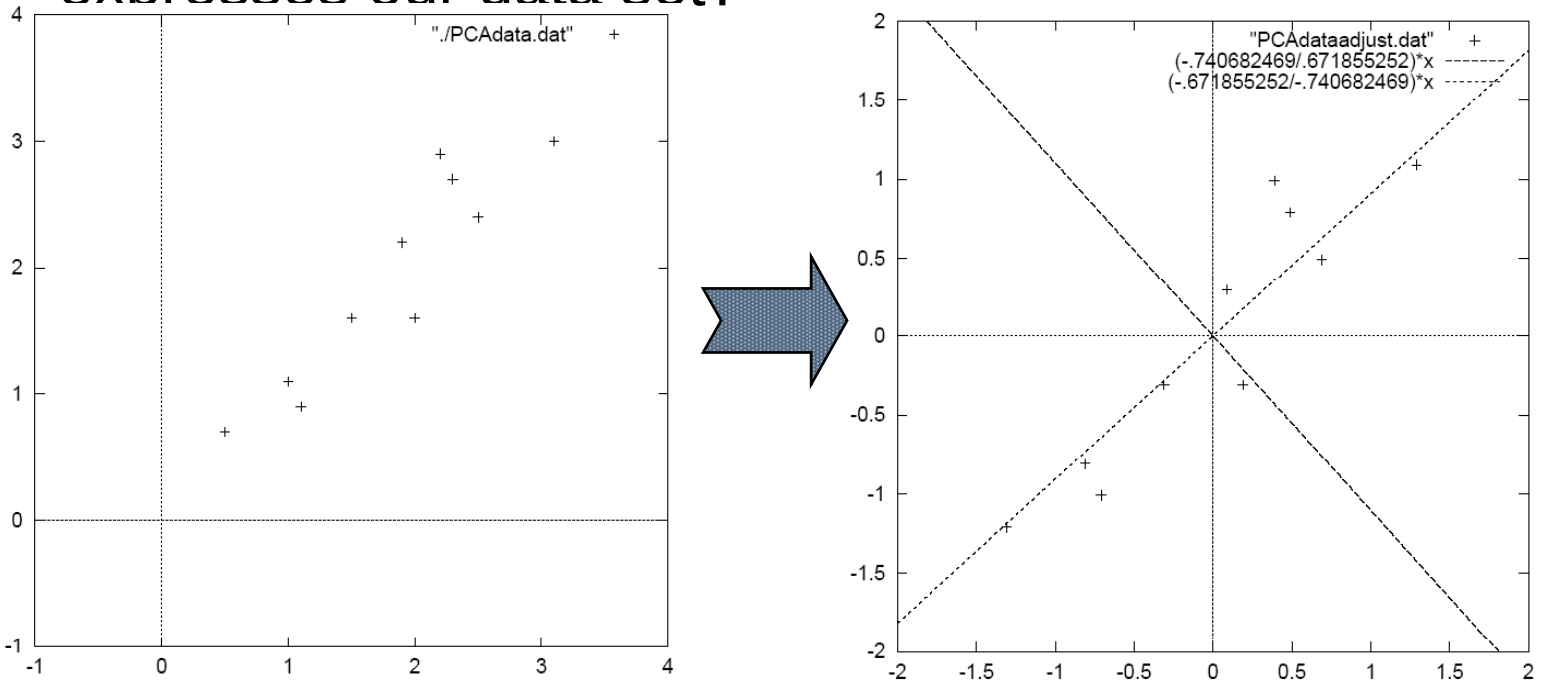
- Input: *quantitative* real-valued data.
- Dimensionality Reduction.
  - Projection of N-dimensional data to a 2D map.
- Topological relationships are maintained (topology-preserving map).
  - nearby outputs correspond to nearby inputs.
  - Classification of input data.
- Cluster data *without supervision*.
  - the network evolves via *competitive dynamics.*
- Easy to visualize/interpret the resulting map.

Training Algorithm

1. Each node's weights are initialized.
2. Randomly choose an input vector x.
3. Find node, with weights that are most like the input vector *(BMU).* $G_{u_0} = Arg\ \underset{G_u}{Min}\left(\left\|x - G_u\right\|\right)$
4. Each neighbouring node's weights are adjusted to make them more like the input vector.
   - The closer a node is to the BMU, the more its weights get altered. $G_u(s+1) = G_u(s) + \varepsilon(s)\sigma(s,u,u_0)\left(x - G_{u_0}(s)\right)$
5. Repeat…

Principal Component Analysis

- Is there another basis, which is a ***linear*** combination of the original basis, that *best* expresses our data set?



Background Mathematics

- Input data: n samples
- Variance (1-D): $s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$
- Covariance (2-D): $cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$
- Covariance Matrix (n-D):

$$C^{n \times n} = (c_{i,j},\ c_{i,j} = cov(Dim_i, Dim_j))$$

e.g. 3 dimensions (x,y,z)

$$C = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix}$$

Principal Component Analysis

- Data set:
  - m measurements (dimensions).
  - each sample is an m-dimensional vector.

    $x = (x_1, \ldots, x_m)^T$
- Subtract the mean. ($\mu_x = E\{x\}$)
- Calculate the covariance matrix C.
- Diagonalize C:
  - Calculate eigenvectors and eigenvalues of C.
  - Choose appropriate eigenvectors (variance is maximized) ➔ principal components

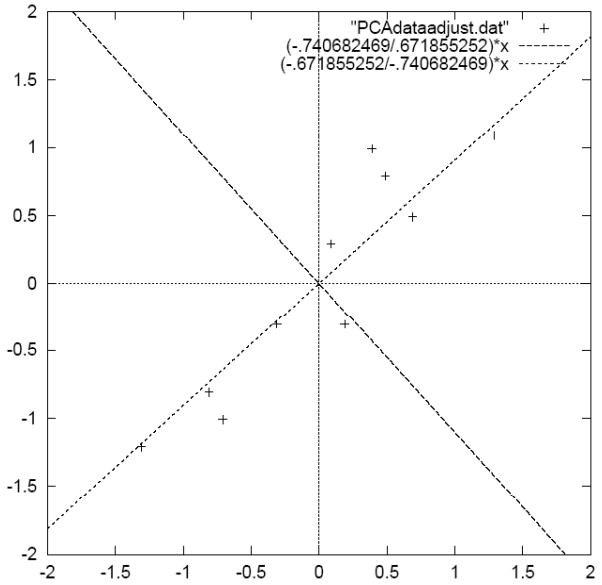
Principal Component Analysis *example*

- 2-D example:

$$cov = \begin{pmatrix} .616555556 & .615444444 \\ .615444444 & .716555556 \end{pmatrix}$$

$$eigenvalues = \begin{pmatrix} .0490833989 \\ 1.28402771 \end{pmatrix}$$

$$eigenvectors = \begin{pmatrix} -.735178656 & -.677873399 \\ .677873399 & -.735178656 \end{pmatrix}$$



Kohonen Map Vs PCA

- Data table X:
  - one row for each sample.
  - one column for each measure (dimension).
- *Kohonen map*
  - rows of table X as inputs to the SOM algorithm.
- *PCA*
  - Diagonalization of matrix X^TX
  - linear projections on the 2 principal axes.

➔ *Similar results*

Problem definition

- N-sample of individuals
- K variables or questions.
  - Each variable has m_k possible modalities.
- Individuals answer each question k by choosing *only one* modality among the m_k modalities.
- Total number of modalities: $M = \sum_{k=1}^{K} m_k$
- *Complete disjunctive table:* (N*M) data matrix D

	m_1			m_2		m_3		
Ind	1	2	3	1	2	1	2	3
1	0	1	0	0	1	1	0	0
2	1	0	0	1	0	0	1	0
⋮								
i	0	0	1	0	1	0	0	1
⋮								
N	0	0	1	1	0	0	0	1

Problem definition

- D contains *all* the information about the individuals.
- Relations between modalities: *Burt Matrix*

  $B = D^T D$
- B is a (M*M) symmetric matrix.
  - It is composed of K*K blocks B_{kl}.
- $\sum_l b_{jl} = b_{j.} = K b_j$
- Sum of all entries: $b = \sum_{j,l} b_{jl} = K \sum_j b_j = K^2 N$

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8
Z_1	b_1	0	0	b_{14}	b_{15}	b_{16}	b_{17}	b_{18}
Z_2	0	b_2	0	b_{24}	b_{25}	b_{26}	b_{27}	b_{28}
Z_3	0	0	b_3	b_{34}	b_{35}	b_{36}	b_{37}	b_{38}
Z_4	b_{41}	b_{42}	b_{43}	b_4	0	b_{46}	b_{47}	b_{48}
Z_5	b_{51}	b_{52}	b_{53}	0	b_5	b_{56}	b_{57}	b_{58}
Z_6	b_{61}	b_{62}	b_{63}	b_{64}	b_{65}	b_6	0	0
Z_7	b_{71}	b_{72}	b_{73}	b_{74}	b_{75}	0	b_7	0
Z_8	b_{81}	b_{82}	b_{83}	b_{84}	b_{85}	0	0	b_8

Factorial Correspondence Analysis

- Only for 2 *qualitative* variables, with I and J modalities.
- Contingency Table:

	Βόλος	Καλαμάτα	Χανιά	...	...	Αθήνα	...	...	...	...	...	...	Margin
Καμία μόρφωση	29	38	83	97	11	1900	62	18	164	230	260	35	2928
Λύκειο	606	800	1728	2016	236	39580	1300	375	3422	4783	5418	722	60987
Επαγγελματική εκπαίδευση	632	835	1802	2102	246	41271	1356	391	3568	4988	5650	753	63593
...	149	197	425	495	58	9729	320	92	841	1176	1332	178	14991
...	91	120	259	302	35	5924	195	56	512	716	811	108	9128
...	29	38	82	95	11	1872	61	18	162	226	256	34	2884
Πανεπιστήμιο	138	183	394	460	54	9027	297	297	86	780	1091	1236	13909
Άλλο	14	19	40	47	5	919	30	9	79	111	126	17	1416
Margin	1688	2229	4813	5613	656	110221	3621	1045	9529	13320	15089	2012	169'836

- Table F of the relative frequencies: $f_{ij} = (n_{ij}/n)$

Factorial Correspondence Analysis

- *Row profiles*

	Βόλος	Καλαμάτα	Χανιά	. . .	. . .	Αθήνα	. . .	. . .	. . .	. . .	. . .	. . .	Margin
Καμία μόρφωση	0,007	0,009	0,039	0,012	0,001	0,726	0,008	0,004	0,086	0,025	0,083	0,005	1,000
Λύκειο	0,006	0,011	0,036	0,023	0,002	0,659	0,016	0,005	0,057	0,060	0,115	0,009	1,000
. . .	0,011												
Πανεπιστήμιο	0,014	0,007	0,012	0,049	0,008	0,669	0,027	0,008	0,025	0,145	0,031	0,006	1,000
Άλλο	0,007	0,011	0,022	0,035	0,001	0,698	0,013	0,005	0,064	0,061	0,067	0,016	1,000
Margin	0,010	0,013	0,028	0,033	0,004	0,649	0,021	0,006	0,056	0,078	0,089	0,012	

- *Column Profiles*

	Βόλος	Καλαμάτα	Χανιά	. . .	Margin
Καμία μόρφωση	0,004	0,012	0,024		0,017
Λύκειο	0,204	0,304	0,461		0,359
Επαγκελματική εκπαίδευση	0,445	0,501	0,359		0,374
. . .	0,097	0,057	0,052		0,088
. . .	0,092	0,061	0,044		0,054
. . .	0,037	0,016	0,019		0,017
Πανεπιστήμιο	0,116	0,043	0,035		0,082
Άλλο	0,006	0,007	0,006		0,008
Margin	1,000	1,000	1,000	1,000	

$$f_{ij} = (n_{ij}/n)$$

$$p_{ij}^{R} = f_{ij} / \sum_{j} f_{ij}$$

$$p_{ij}^{C} = f_{ij} / \sum_{i} f_{ij}$$

Factorial Correspondence Analysis

- p^R_{ij} is the conditional probability that the first variable has value i, given that the second one is equal to j.
- It is usual to consider χ^2-distance between rows/columns:

$$\chi^2(i,i') = \sum_j \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

- Inertia (row profiles) = Inertia (column profiles)

$$\Im = \sum_i f_{i\cdot} \chi^2(i,\bar{i}) = \sum_{ij} \frac{(f_{ij} - f_{i\cdot} f_{\cdot j})^2}{f_{i\cdot} f_{\cdot j}} = \sum_{ij} \frac{f_{ij}^2}{f_{i\cdot} f_{\cdot j}} - 1$$

Factorial Correspondence Analysis

- In order to use the *Euclidean distance* instead of the χ^2-distance:
  - Replace table F by corrected F^c: $f_{ij}^c = \frac{f_{ij}}{\sqrt{f_{i\cdot}}\sqrt{f_{\cdot j}}}$
- FCA is a *double* PCA achieved on the *rows* and on the *columns* of this corrected data matrix F^c.
- eigenvalues and eigenvectors are computed by the diagonalization of matrix:
  - $F^{cT}F^c$ (Row profiles)
  - F^cF^{cT} (Column profiles)

  → same eigenvalues
  → strongly related eigenvectors
- total inertia J is equal to the sum of the eigenvalues of $F^{cT}F^c$ or F^cF^{cT}

Factorial Correspondence Analysis

- SOM algorithm and PCA have similar results on a given data set.
- *Key point* for defining the SOM algorithms adapted to qualitative variables:
  - The diagonalization of the data matrix $F^{c\,T}F^{c}$ can can be approximately replaced by a *SOM algorithm* in which the *row profiles* are used as inputs
  - The diagonalization of $F^{c}F^{c\,T}$ can be replaced by a *SOM algorithm* in which the *column profiles* are used as inputs.

Multiple Correspondence Analysis

(1) If we are interested in the *modalities* only:

- The data table (contingency table) is the *Burt table*.
- We consider the *corrected* Burt table B^c:

$$b_{jl}^c = \frac{b_{jl}}{\sqrt{b_{j\cdot}}\sqrt{b_{\cdot l}}} = \frac{b_{jl}}{K\sqrt{b_j}\sqrt{b_l}}$$

- B and B^c are symmetric, so the diagonalizations of $B^{cT}B^c$ or B^cB^{cT} are *identical*.
- Principal axes of usual PCA of B^c are the principal axes of MCA

Multiple Correspondence Analysis

(2) If we are interested in the *individuals:*

- Use the corrected matrix D^c of the Complete Disjunctive table D.

$$d_{ij}^c = \frac{d_{ij}}{\sqrt{d_{i\cdot}}\sqrt{d_{\cdot j}}} = \frac{d_{ij}}{\sqrt{K}\sqrt{b_j}}$$

- D^c is *not* symmetric.
  - Diagonalization of $D^{cT}D^c$ will provide a representation of the *individuals*.
  - Diagonalization of D^cD^{cT} will provide a representation of the *modalities*.
- Both representations can be superposed.

From MCA to SOM algorithm

(1) If we want to deal *only* with the modalities:

- Apply the SOM algorithm to the rows (or over the columns) of B^c.

(2) If we want to keep the *individuals:*

- Apply the SOM algorithm to the rows of D^c.
  - Problem: we will get a Kohonen map for the individuals only.
  - We want to simultaneously represent the modalities.
- Two techniques:
  - The modalities are assigned to the classes *after* training, as supplementary data.
  - Two SOM algorithms are used on the rows (individuals) and on the columns (modalities) of D^c.

Kohonen-based analysis of a Burt table (algorithm *KMCA*)

- Take into account only the modalities.
- Data matrix: corrected Burt Table B^c.
- n*n Kohonen network.
- each unit u is represented by code vector C_u in R^M.

Training algorithm:

- present at random an input r(j)
  - i.e. a row of the corrected matrix B^c.
- look for the winning unit u_0
  - i.e. that which minimizes $|| r(j) - C_u ||^2$.
- update the code vectors of the winning unit and its neighbors by

$$C_u^{\text{new}} - C_u^{\text{old}} = \varepsilon\sigma(u, u_0)(r(j) - C_u^{\text{old}})$$

algorithm KMCA example

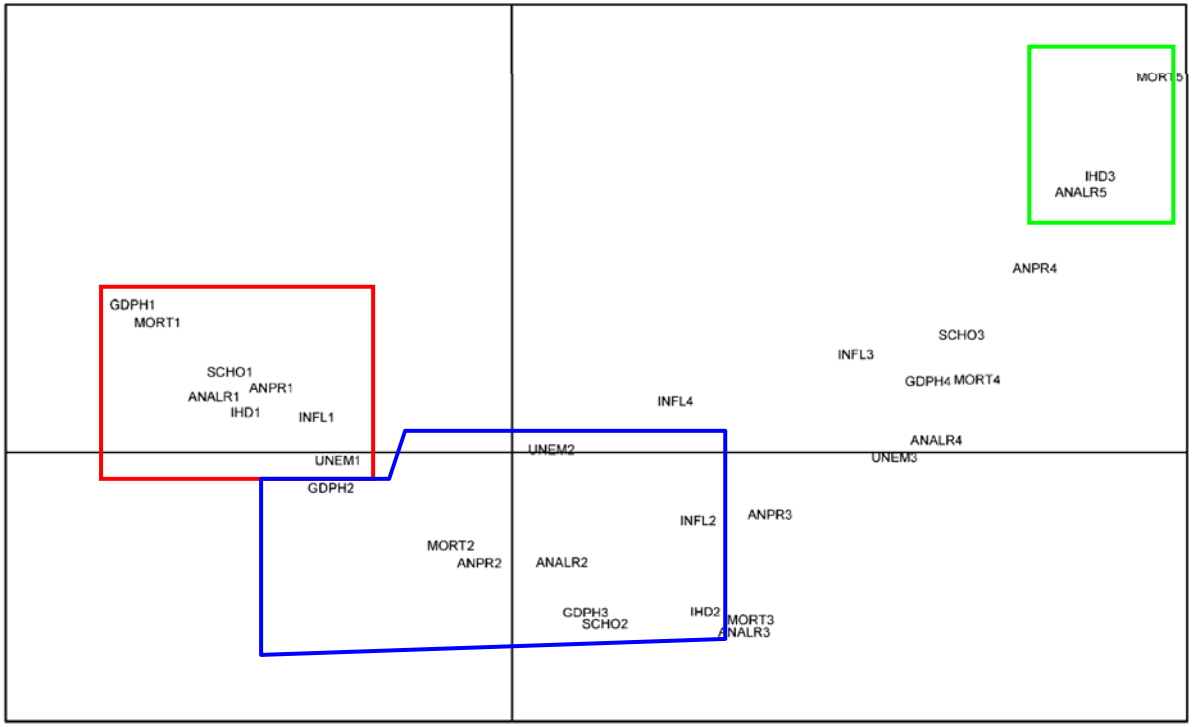
- Use POP_96 database
- 8 qualitative variables:

The qualitative variables for the POP_96 database

Heading	Modalities	Name
Annual population growth	[−1, 1[, [1, 2[, [2, 3[, ≥3	ANPR1, ANPR2, ANPR3, ANPR4
Mortality rate	[4, 10[, [10, 40[, [40, 70[, [70, 100[, ≥100	MORT1, MORT2, MORT3, MORT4, MORT5
Analphabetism rate	[0, 6[, [6, 20[, [20, 35[, [35, 50[, ≥50	ANALR1, ANALR2, ANALR3, ANALR4, ANALR5
High school	≥80, [40, 80[, [4, 40[	SCHO1, SCHO2, SCHO3
GDPH	≥10,000, [3000,10,000[, [1000, 3000[, <1000	GDPH1, GDPH2, GDPH3, GDPH4
Unemployment rate	[0, 10[, [10, 20[, ≥20	UNEM1, UNEM2, UNEM3
Inflation rate	[0, 10[, [10, 50[, [50, 100[, ≥100	INFL1, INFL2, INFL3, INFL4
IHD	1, 2, 3	IHD1, IHD2, IHD3

algorithm KMCA example

ANPR2 UNEM2		MORT2 ANALR2	SCHO2		MORT3 ANALR3
	GDPH2		GDPH3	INFL2 IHD2	ANPR3
UNEM1		INFL3	INFL4		UNEM3
IHD1	INFL1		ANALR4	ANPR4	
ANPR1 ANALR1 SCHO1			MORT4	IHD3	MORT5 ANALR5
	MORT1 GDPH1		GDPH4	FSCHO3	



KMCA_ind algorithm
qualitative variables + individuals

- More interesting and more valuable to cluster the *individuals* and the *modalities* at the same time.
  - Build a Kohonen map with the individuals.
  - After, project the modalities as supplementary data (with the suitable scaling).

Training algorithm:

(i) Kohonen map is trained with the rows of D^c.

(ii) Each modality j is represented by an M-vector:

$$\frac{b_{jl}}{b_j\sqrt{b_l}\sqrt{K}}, \quad \text{for } l = 1, \ldots, M$$

which is the *mean* vector of *all* the individuals who *share* this modality.

KMCA_ind algorithm
qualitative variables + individuals

- Each mean vector is assigned into the Kohonen class of its nearest code vector.
- Not always useful, especially when too many individuals.
- Nice representation, but breaks the symmetry between individuals and modalities

INFL4 Moldavia Romania Russia Ukraine	Bulgaria Poland	GDPH3 Costa Rica Ecuador Jamaica Lebanon	Colombia Fiji Panama Peru Thailand		MORT5 Afghanistan Angola Haiti Mozambique Pakistan Yemen
Brazil	ANPR2 MORT2	Croatia Venezuela	ANALR2 UNEM2	Ghana Mauritania Sudan	ANPR4 ANALR5 IHD3
GPDH2 Chili Cyprus S. Korea	Argentina Bahrain Malaysia Malta Mexico		INFL3 Macedonia Mongolia	SCHO3 GPDH4 UNEM3 Laos	MORT4 ANALR4 Cameroon Comoros Ivory Coast Nigeria
Greece Hungary Slovenia Uruguay	UNEM1 IHD1 Portugal	China Philippines Yugoslavia	Albania Indonesia Sri Lanka	INFL2 Guyana Vietnam	Bolivia Kenya Nicaragua
ANPR1 ANALR1 SCHO1 R. Czech	Germany, Sweden Australia, Israel USA, Iceland Japan, Norway New-Zealand Netherlands United Kingdom Switzerland	INFL1 U Arab Emirates	SCHO2 Egypt	ANPR3 IHD2 Morocco Paraguay	ANALR3 El Salvador Swaziland
Belgium Canada Denmark Spain, Italy Finland France Ireland	MORT1 GDPH1 Luxemburg	Singapore	Saudi Arabia Syria	MORT3 Algeria	South Africa Iran Namibia Tunisia Turkey Zimbabwe

KDISJ algorithm

- We want to keep together modalities and individuals in a more balanced way than KMC_ind.
- Each unit u of the Kohonen network has a code vector C_u that is comprised of (M+N) components:
  - first M components ➔ *individuals* (rows of D^c).
  - N final components ➔ *modalities* (columns of D^c).
- Kohonen algorithm:
  - double learning process.
  - At each step: *alternatively* draw a D^c row (individual) or a D^c column (modality).

KDISJ algorithm

(1) When we draw an *individual* i

(a) Associate the modality j(i)

$$j(i) = \text{argmax}_j \, d_{ij}^c = \text{argmax}_j \frac{d_{ij}}{\sqrt{K d_{\cdot j}}}$$

that maximizes the coefficien d_{ij}^c

- i.e. the *rarest* modality out of all.

(b) Create extended individual vector

$X = (i, j(i)) = (X_M, X_N)$

(c) Find closest unit u_0:

Euclidean distance *restricted* to first M components

KDISJ algorithm

(d) Move the code vectors of the unit u_0 and its neighbours closer to the extended vector X = (i, j(i)).

$$\begin{cases} u_0 = Arg\min_u ||X_M - C_{M,u}|| \\ C_u^{new} = C_u^{old} + \varepsilon\sigma(u, u_0)(X - C_u^{old}) \end{cases}$$

(2) When we draw a *modality* j

(a) Do not associate an individual with it.

- Many equally placed individuals ➔ arbitrary choice

(b) Find closest unit v_0:

Euclidean distance *restricted* to last N components

KDISJ algorithm

(c) Move the *last* N components of the code vectors of the unit u_0 and its neighbours closer to the corresponding components of the modality j.

- Do not modify the *first* M components.

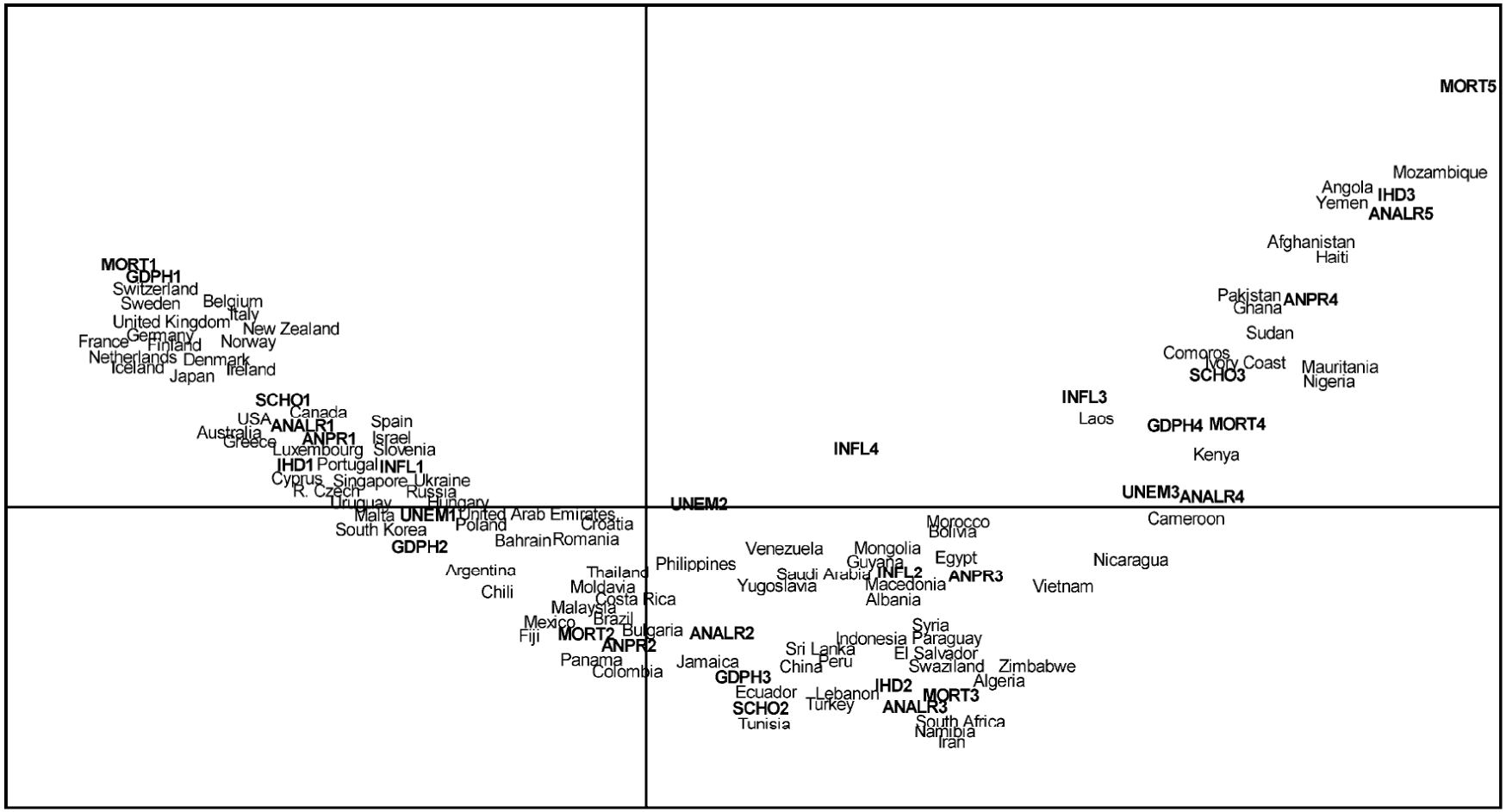
$$\begin{cases} v_0 = \text{argmin}_u \| Y - C_{N,u} \| \\ C_{N,u}^{\text{new}} = C_{N,u}^{\text{old}} + \varepsilon \sigma(u, v_0)(Y - C_{N,u}^{\text{old}}) \end{cases}$$

- Y is the N-column vector corresponding to modality j

KDISJ algorithm example-SOM

SCHO2 IHD2 Algeria Syria	Saudi Arabia Egypt Indonesia	Brazil Mexico	Argentina Chili Cyprus S. Korea	INFL1 Australia Canada USA	GDPH1 Belgium Denmark Finland France Ireland Italy
ANALR3 South Africa Iran Namibia El Salvador Zimbabwe	MORT3 Guyana Morocco Paraguay Tunisia Turkey		GDPH2	IHD1 Israel	MORT1 Germany, U Kingdom Iceland, Japan Lux, Singapore Norway, Sweden New Zealand Netherlands, Spain Switzerland
Kenya Nicaragua	Swaziland	UNEM2 U Arab Emirates Malaysia	Malta Portugal	UNEM2 Greece Hungary Slovenia Uruguay	ANPR1 ANALR1 SCHO1 R. Czech
ANALR4 Comoros Ivory Coast	MORT4 Cameroon Nigeria	ANPR3 Bolivia	INFL2 Bahrain Philippines	Poland	INFL4 Croatia Moldavia Romania Russia Ukraine
ANPR4 IHD3 Ghana	SCHO3 GDPH4 Laos Mauritania Sudan	UNEM3 Vietnam Yugoslavia		MORT2 ANALR2 Bulgaria Ecuador Jamaica Lebanon, Peru	GDPH3 Costa Rica
MORT5 ANALR5 Afghanistan Angola Pakistan Yemen	Haiti Mozambique	INFL3 Macedonia Mongolia	Albania China Sri Lanka	Colombia Panama	ANPR2 Fiji Thailand Venezuela

KDISJ algorithm example-MCA



Conclusion

- Approximation of FCA and MCA through SOM algorithm.
- Analysis of qualitative variables using SOM.

SOM-based algorithm	***Factorial method***
SOM algorithm on the rows of matrix X	PCA, diagonalization on X^TX
KCMA (clustering the modalities): SOM algorithm on the rows of B^c	MCA, diagonalization on $B^{cT}B^c$
KCMA_ind (clustering the individuals): SOM algorithm on the rows of D^c and setting of the Modalities.	MCA with individuals, diagonalization of $D^{cT}D^c$ and $D^c D^{cT}$
KDISJ (individuals+modalities): SOM on the rows and the columns of D^c	

References

- Cottrell, Letremy, ”Analyzing surveys using the Kohonen algorithm”, ESANN 2003.
- Cottrell, Ibbou, Letremy, “SOM-based algorithms for qualitative variables”, Neural Networks 2004
- Kaski, “Data exploration using self-organizing maps”, DSc thesis 1997
- Kohonen, “Self-organization and associative memory”, Springer 1984.
- Lindsay, “A tutorial on Principal Components Analysis”, 2002
- Micheloud, “Correspondence Analysis”, Advanced Econometrics Workshop, HEC 1996.
- Shlens, “A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS: Derivation, Discussion and Singular Value Decomposition”, 2003

Assumptions-Limits of PCA

- Linearity.
  - Linearity frames the problem as a *change of basis*.
- Mean and variance are sufficient statistics.
  - Only zero-mean probability distribution that is fully described by the variance is the Gaussian distribution.
  - In order for this assumption to hold, the probability distribution of xi must be *Gaussian*.
- Large variances have important dynamics.
  - data has a high SNR.
- The principal components are orthogonal.
  - intuitive simplification that makes PCA soluble with linear algebra decomposition techniques.