International Meeting of the Psychometric Society – 2003, Chia, Cagliari, Italy
July 7-10, 2003

Visualization and Classification via Contiguity Analysis

Ludovic Lebart

CNRS-ENST, 46 rue Barrault, 75013, Paris, France
lebart@enst.fr

To represent

categorical variables,
in a high dimensional space,
with a high level of noise,

powerful tools of visualization are needed.

There is usually a **trade-off** between:

- the **comfort** of the representation,
- the **transparency** of the method
- the possibilities of **assessment**.

Our aim:

to fill the gap between:

- some **black box** methods leading to **clear** and **legible** results (such as *Kohonen Self Organizing Maps*),

- and the **more statistically transparent** and **less constrained techniques** of clustering (used together with principal axes techniques) whose outputs are often **difficult to handle**.

We stress here the contribution of contiguity analysis to such approaches.

The structure of symmetric graph can aptly describe some a priori structures of observations: chronological observations, geographic information,...

But such a structure can be generated by the multidimensional data themselves, or by parts of the data set (any binary relationship R) :

Graphs of k nearest neighbours (after « symmetrisation » !)
R(i,j) means that i is one of the k nearest neighbours of j
(k varies from 1 to n, number of observations).

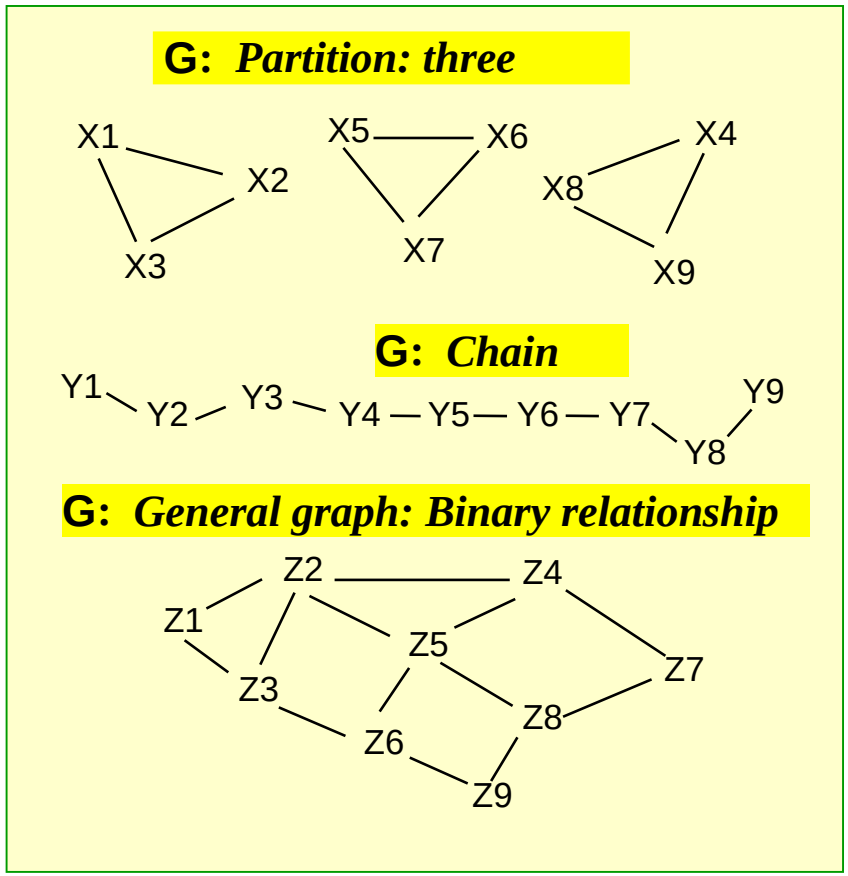
Graphs derived from distance thresholds t (generally derived from quantiles of distances). R(i,j) means that d(i,j) <= t.

We will show that such graphs constitute a (the?) **missing link** between **clustering and classification** on the one hand, and **principal axes techniques** on the other.

We consider a set of multivariate observations, ***n* objects** described by ***p*** variables, leading to a matrix **Y**, whose rows have an *a priori* **graph structure**.

The ***n* objects** are the vertices of a symmetric graph G whose associated (***n***, ***n***) matrix is **M** .

($\mathbf{m}_{ii'} = 1$ if vertices i and i' are joined by an edge, $\mathbf{m}_{ii'} = 0$ otherwise).



1) In a first step, the graph structure could be considered as **external** (geographic areas, time series).

2) Then we consider the situation in which the graph structure is **not external**, but derived from the matrix **Y** itself,

(e.g.: The series of graphs of **k-nearest neighbours**).

The idea of deriving from data a metric likely to highlight the existence of clusters dates back to:

ART, D., GNANADESIKAN, R., KETTENRING, J.R. (1982):
Data Based Metrics for Cluster Analysis, Utilitas Mathematica, 21 A, 75--99.

3) Eventually, we will mention the case of a graph derived from a part (some columns) of the data matrix Y, or from supplementary columns (instrumental variables).

The *n* **objects** (rows of Y) are the vertices of a symmetric graph G whose associated (*n*, *n*) matrix is **M** .

$m_{ii'} = 1$ if vertices i and i' are joined by an edge, $m_{ii'} = 0$ otherwise.

$m = \sum m_{ii'}$ (number of edges of G)

Local variance

$$v^c(y) = (1/2m)\sum m_{ii'}(y_i - y_{i'})^2$$

Global variance

$$v(y) = (1/2n(n-1))\sum (y_i - y_{i'})^2$$

(The classical empirical variance is a particular case of local variance when the graph is complete, i.e.: $m_{ii'} = 1$ for all i and i ')

Contiguity coefficient (Geary, 1954; after Moran and Von Neumann)

$$c(y) = v^*(y) / v(y)$$

« Corrected » definitions of local variance, new contiguity coefficient

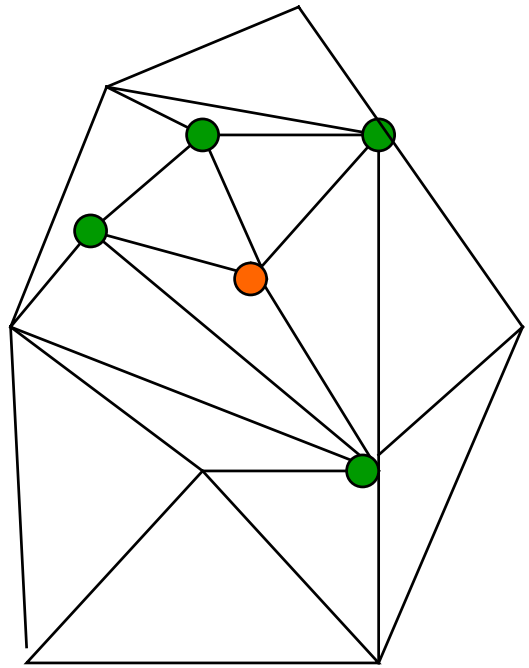
New Local variance

$$m_i^* = (1/n_i)\sum_k m_{ik} y_k$$

$$v^*(y) = (1/n)\sum (y_i - m_i^*)^2$$

New Contiguity coefficient

$$c(y) = v^*(y) / v(y)$$



With, as usual,

$$v(y) = 1/n \sum_{i=1}^{n} (y_i - m\ )^2$$

Defining the diagonal matrix **N** (matrix of degrees) such that

$$n_i = \Sigma_k m_{ik}$$

c(y) reads, in matrix form (**U** = matrix associated to a complete graph):

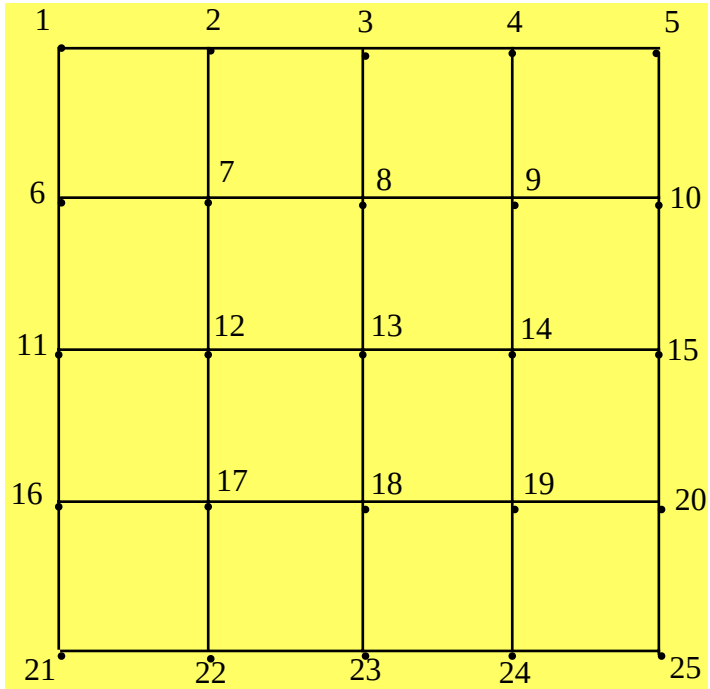
$$c(y) = \mathbf{y}' (\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})' (\mathbf{I} - \mathbf{N}^{-1}\mathbf{M}) \mathbf{y} \; / \; \mathbf{y}' (\mathbf{I} - (1/n)\mathbf{U}) \mathbf{y}$$

The *(p, p)* local covariance matrix **V*** is then defined

$$\mathbf{V}^* = (1/n)\, \mathbf{Y}'(\mathbf{I} - \mathbf{N}^{-1}\mathbf{M})' (\mathbf{I} - \mathbf{N}^{-1}\mathbf{M}) \mathbf{Y}$$

This matrix, together with the corresponding correlation matrix provide a powerful tool for studying partial correlation when the graph is derived from a set of « instrumental variables ».

Example of a graph **G** (***n*** = 25) associated with a squared lattice



… and its associated matrix **M** ➔

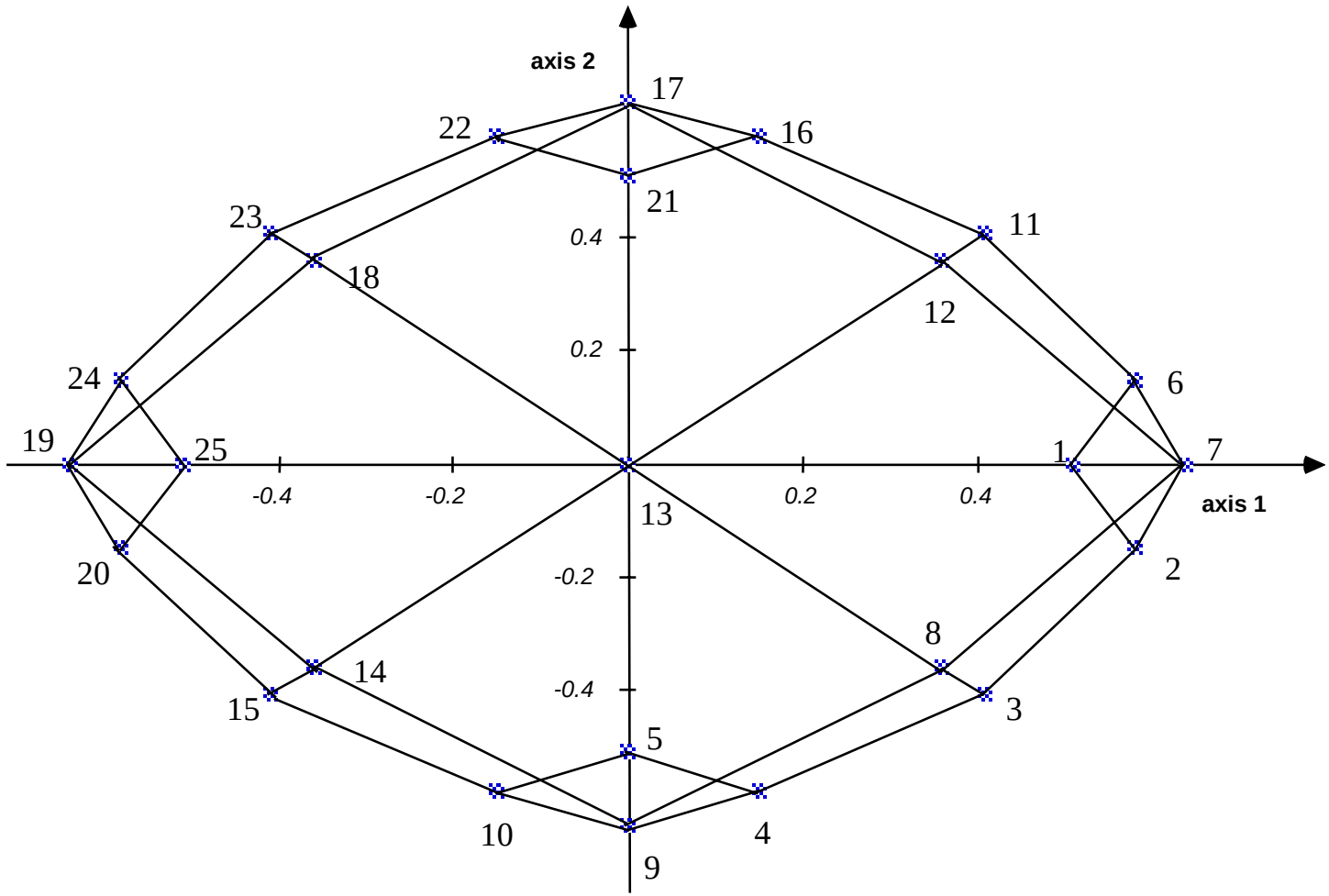
matrix:
M

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
r01	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r02	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r03	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r04	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r05	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r06	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r07	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
r08	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
r09	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
r10	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
r11	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0
r12	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0
r13	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0
r14	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0
r15	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
r16	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	0	0
r17	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0	0
r18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0	0
r19	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	1	0
r20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1
r21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0
r22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0
r23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0
r24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	1
r25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1

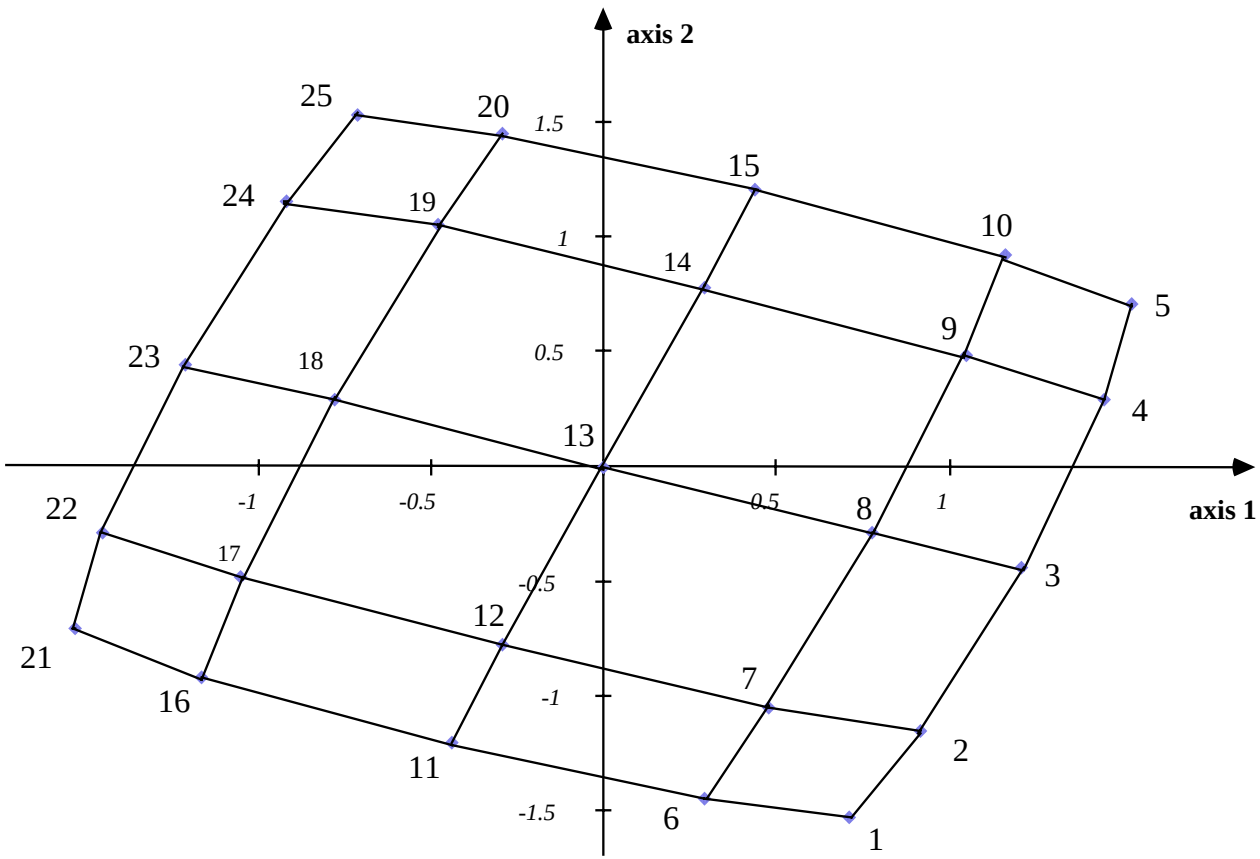
Condensed numerical coding of matrix **M**

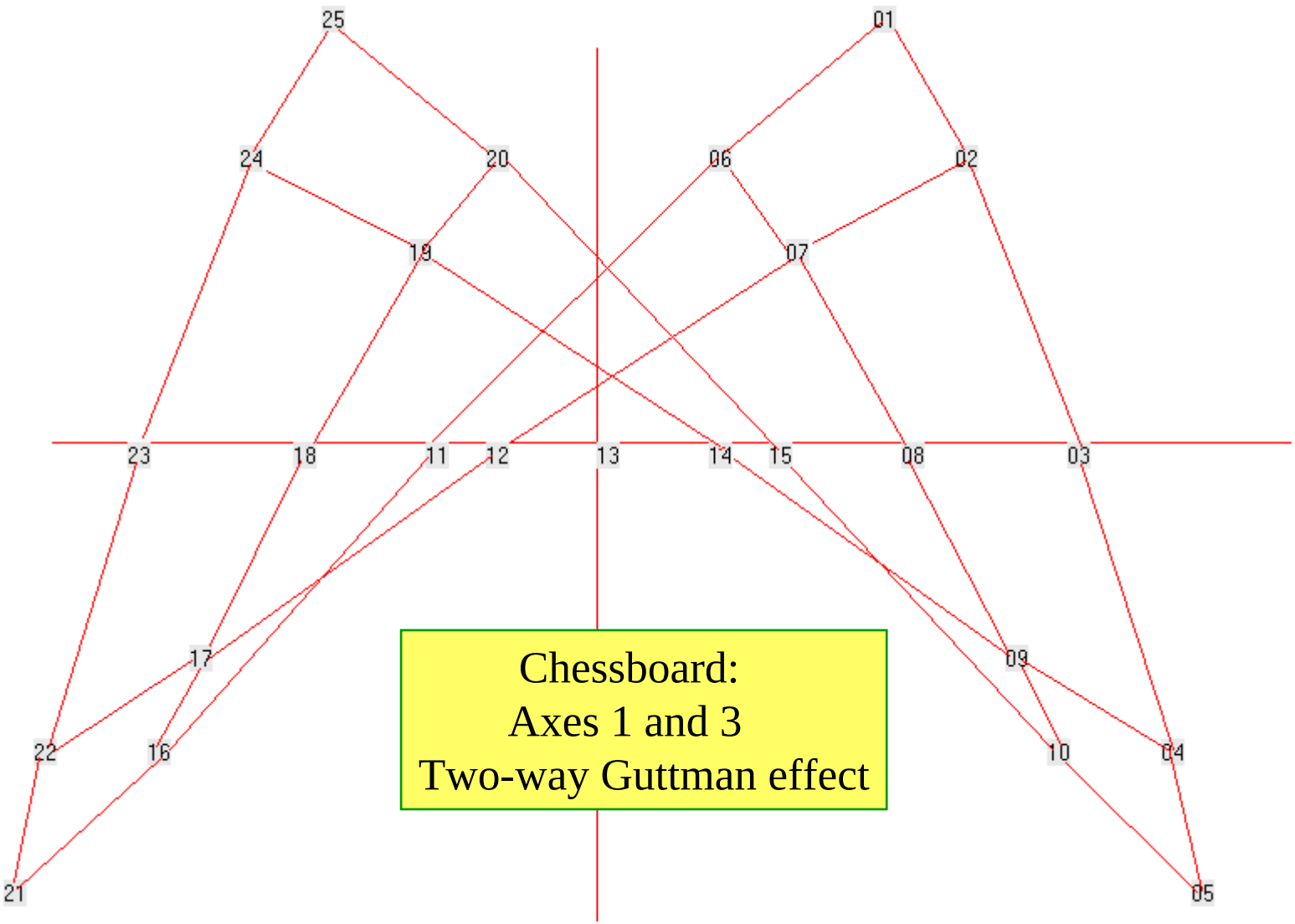
Vertex	*Adjacent vertices*				
1	1	2	6		
2	2	1	7	3	
3	3	2	8	4	
4	4	3	9	5	
5	5	4	10		
6	6	1	7	11	
7	7	2	6	8	12
8	8	3	7	9	13
9	9	4	8	10	14
10	10	5	9	15	
11	11	6	12	16	
12	12	7	11	13	17
13	13	8	12	14	18
14	14	9	13	15	19
15	15	10	14	20	
16	16	11	17	21	
17	17	12	16	18	22
18	18	13	17	19	23
19	19	14	18	20	24
20	20	15	19	25	
21	21	16	22		
22	22	21	17	23	
23	23	18	22	24	
24	24	19	23	25	
25	25	24	20		

Description of chessboard **G** through *Principal Component Analysis* of **M**

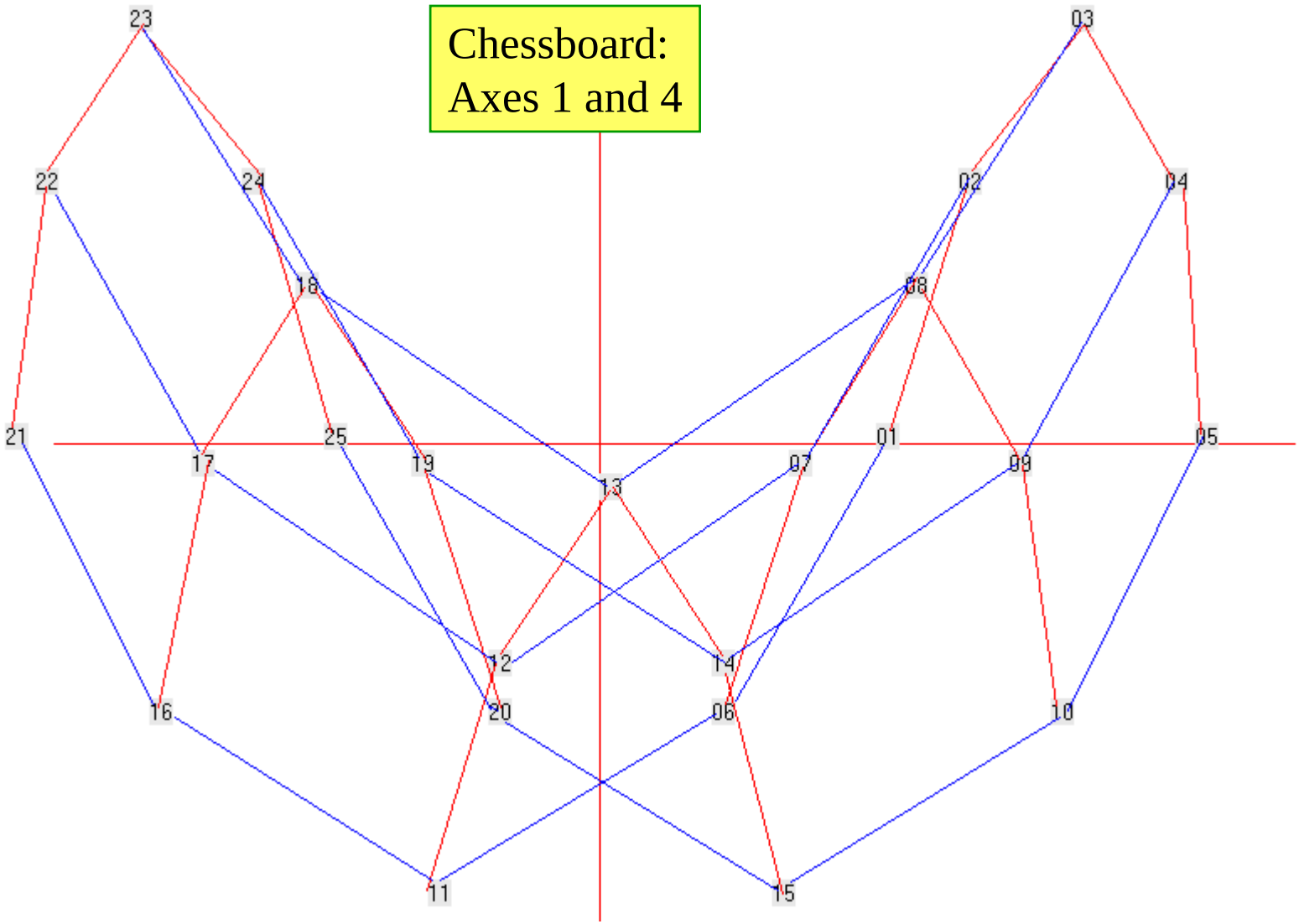


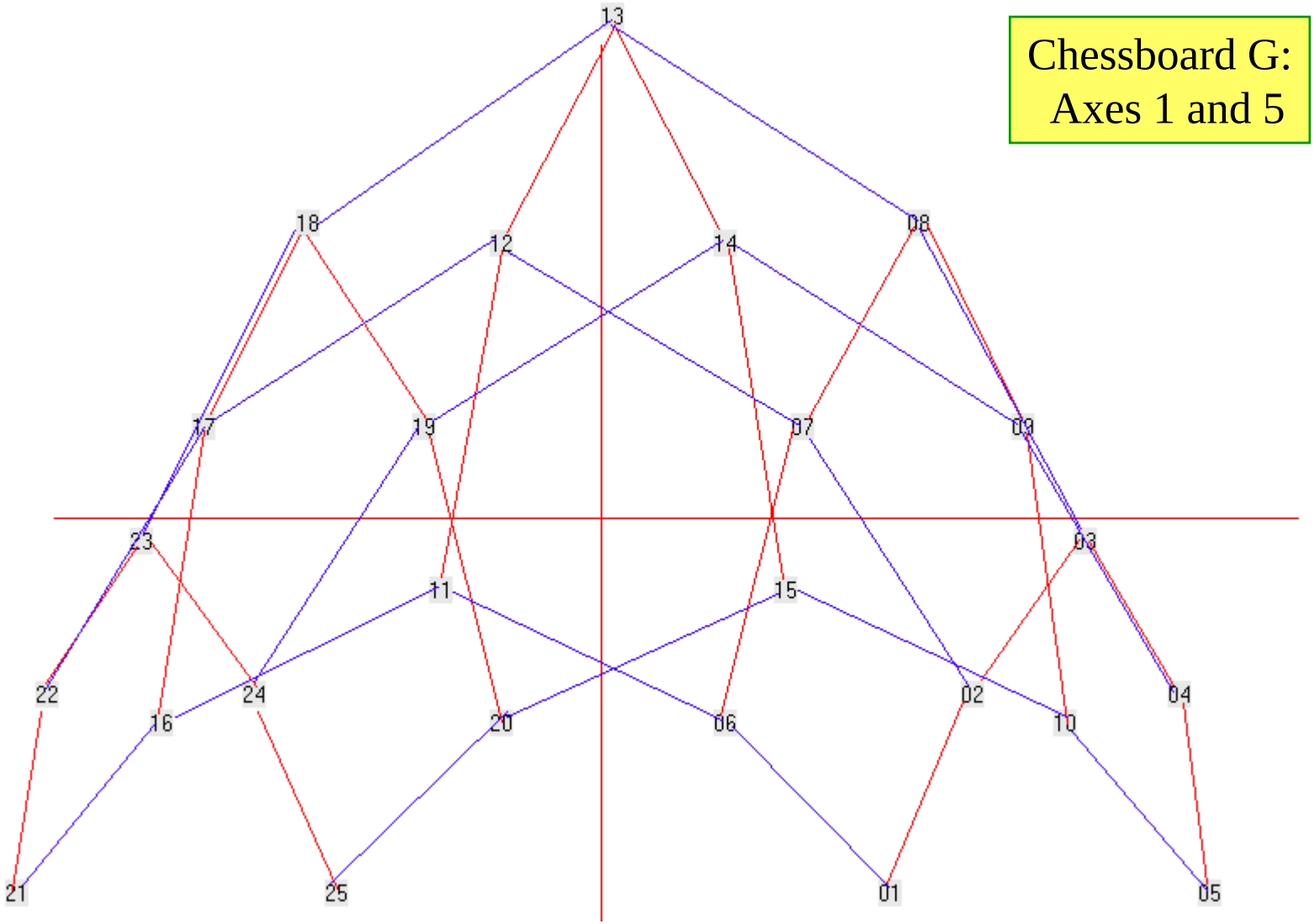
Description of chessboard **G** through *Correspondence Analysis* of **M**



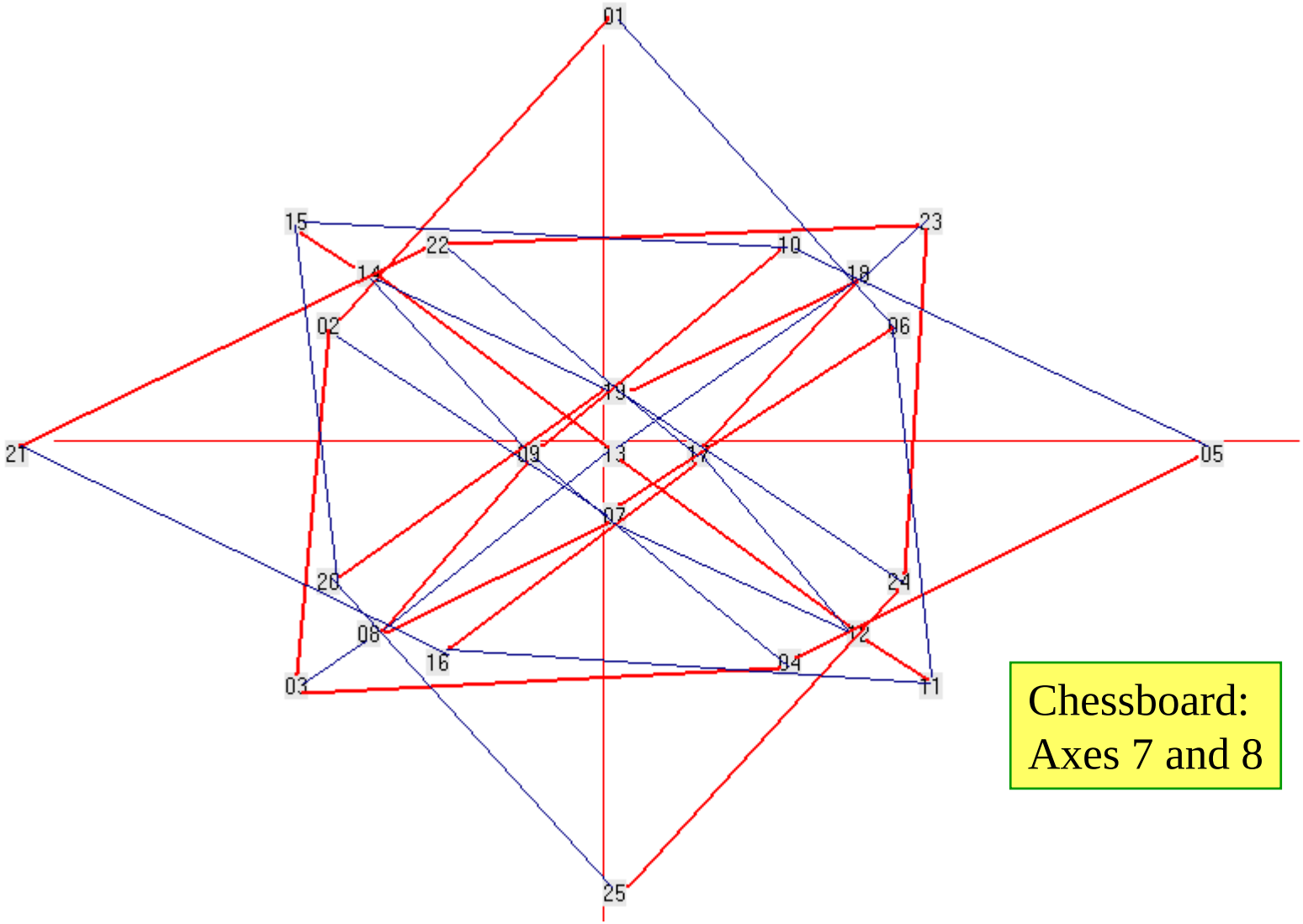


Chessboard:
Axes 1 and 3
Two-way Guttman effect


Chessboard:
Axes 1 and 4
23
03
22
24
02
04
18
08
21
25
01
05
17
19
07
09
13
12
14
16
20
06
10
11
15



Chessboard G:
Axes 1 and 5



Chessboard:
Axes 7 and 8

Note that the calculations involved in the CA of such typical graphs could be carried out directly, without the help of a computer.

In the case of a simpler graph (a chain) the description through CA leads to the relationship: $Ma = \lambda a$, which leads to a simple finite difference equation.

$$a_{i-1} + a_{i+1} = \lambda a_i$$

→ $$\varphi_\alpha(j) = \cos\left(\frac{2j\alpha\pi}{n}\right) \qquad \phi_\alpha(j) = \sin\left(\frac{2j\alpha\pi}{n}\right)$$

A chessboard can then be defined as a “ tensorial sum of chains ”, and the final results analytically derived from those of the chain (see : Benzécri, 1973). Note the analogy with the eigen-vectors of a Laplace operator.

Properties of the « Laplacian » (matrix: N - M) of a graph: Some references...

CHUNG F.R.K., *Spectral Graph Theory*. CBMS Reg. Conf. Ser. Math. 92, American Mathematical Society, 1997.

KOREN Y., CARMEL L., HAREL D., ACE: a Fast Multiscale Eigenvectors Computation for Drawing Huge Graphs, *Proceedings of IEEE Information Visualization, 2002*, p 137-144.

MOHAR B., The Laplacian Spectrum of Graphs, *Graph Theory, Combinatorics and Application*, 2, 1991, p 871-898.

MOHAR B., Some Applications of Laplace Eigenvalues of Graphs, *Graph Symmetry, Algebraic Methods and Application*, Hahn G., Sabidussi G., NATO Ser. C., 497, Kluwer, 1997, p 225-275.

Explaining the good quality of the visualization:

Local variance = $\mathbf{y'}(\mathbf{I} - \mathbf{N^{-1}M})' \; (\mathbf{I} - \mathbf{N^{-1}M}) \; \mathbf{y}$

Global variance $= \mathbf{y'y}$

Bounds for *c(y)* = contiguity coefficient**.**

$c(y) \;=\; \mathbf{y'}(\mathbf{I} - \mathbf{N^{-1}M})' \; (\mathbf{I} - \mathbf{N^{-1}M}) \; \mathbf{y} \; / \; \mathbf{y'} \; \mathbf{y}$

the minimum of *c(y)*, μ, is the square root of the smallest eigenvalue μ^2 of:

$$(\mathbf{I} - \mathbf{N^{-1}M})'(\mathbf{I} - \mathbf{N^{-1}M}) \; \psi = \mu^2 \psi$$

If the graph is regular, $\mathbf{N^{-1}} = (1/r) \; \mathbf{I}$

$$(\mathbf{I} - (1/r)\mathbf{M})^2 \; \psi = \mu^2 \psi$$

$$(\mathbf{I} - \mathbf{N^{-1}M}) \; \psi = \mu \psi$$

Equivalently: $\mathbf{N^{-1}M} \; \psi = (1 - \mu) \psi$

We have obtained, up to now: $\mathbf{N^{-1}M}\ \psi = (1-\mu)\ \psi$

Note the transition formula, (no: formulae, owing to symmetry!)

$$\mathbf{N^{-1}M}\ \phi = \varepsilon\sqrt{\lambda}\ \ \phi$$

(formula drawn *from the Correspondence Analysis of matrix* **M**)

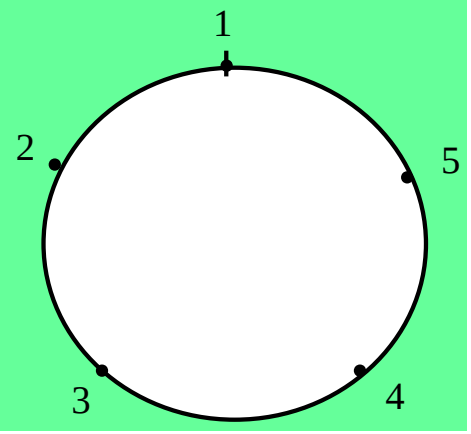
if $\varepsilon = +1$, direct factor,
if $\varepsilon = -1$, inverse factor.
(See, e.g., the analysis of confusion matrices).

Then: Min [c(y)] = Min μ = Max λ , [$= \lambda_{max}$ if ($\varepsilon = +1$).]

Thus: *Min [c(y)] = 1- $\sqrt{\lambda}_{max}$*

Paradoxical measurement of *information* : case of a *cycle*

$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$



$$\varphi_\alpha(j) = \cos\left(\frac{2j\alpha\pi}{n}\right) \qquad \phi_\alpha(j) = \sin\left(\frac{2j\alpha\pi}{n}\right)$$

$$\lambda_\alpha = \cos^2\left(\frac{2\alpha\pi}{n}\right) \qquad \tau_\alpha = \frac{2}{n}\cos^2\left(\frac{2\alpha\pi}{n}\right)$$

The contiguity ratio can be generalized :

a) to different distances between vertices in the graph,

b) to multivariate observations

(For Geary coefficient, both generalizations are dealt with in: Lebart, 1969).

a) The graph corresponding to the distance defined as " the shortest path of length k between two vertices " is associated to the matrix $\mathbf{M}^{(k)} - \mathbf{M}^{(k-1)}$, where $\mathbf{M}^{(k)}$ designates the k-th booleean power of the matrix $(\mathbf{I} + \mathbf{M})$.

Therefore, it is easy to test the significance of spatial auto-correlation, so long as these distances on the graph remain meaningful.

This approach provides a variant, in the graph case, of the *variogram* used in *geostatistics*, as presented in the seminal papers of Matheron (1963, 1965).

b) Generalization to multivariate observations

Let **Y'u** be the vector of the *n* values of a linear combination *u* of the *p* variables.

Then, its contiguity coefficient is defined likewise:

$$c(u) = \mathbf{u'\,V^*\,u} \;/\; \mathbf{u'\,V\,u}$$

… in which $\mathbf{V^*} = (1/n)\,\mathbf{Y'(I - N^{-1}M)'(I - N^{-1}M)Y}$

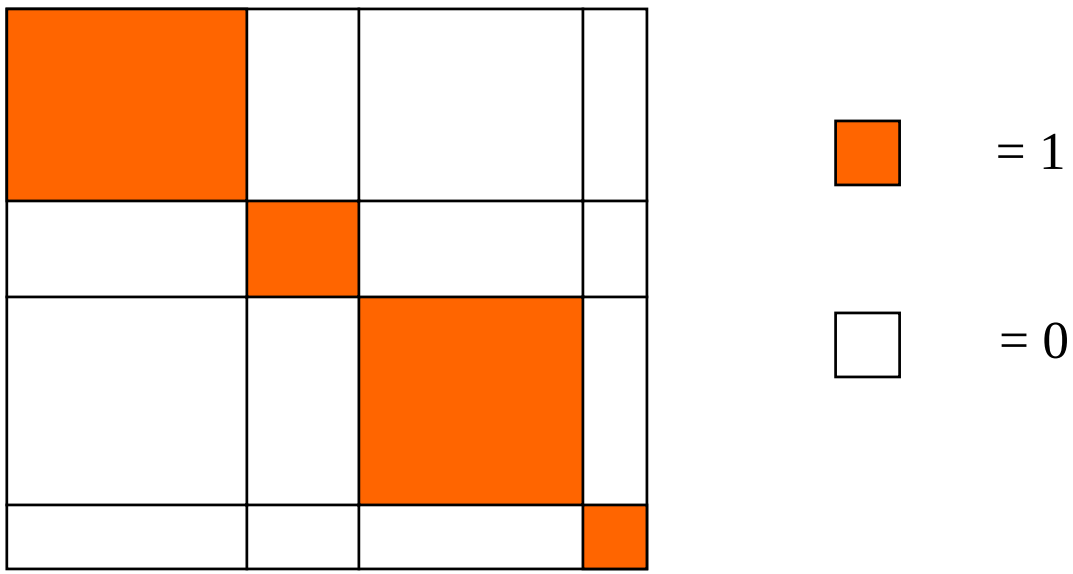
is the *(p, p)* local covariance matrix **V***

Contiguity analysis is the search for the minima of *c(u):*

$$c(u) = \mathbf{u'\,V^{*}\,u} \;/\; \mathbf{u'\,V\,u}$$

It reduced to Fisher Linear Discriminant Analysis if G is associated with a partition graph.

M =



C(u) allows one to deal with overlapping clusters, fuzzy partition, ...

External graphs: Partial correlation

The local covariance matrix and its related correlation matrix
can define a local context for any set of instrumental variables.

$$\mathbf{V^*} = \textit{(1/n)}\ \mathbf{Y'(\ I - N^{-1}M)'\ (\ I - N^{-1}M)\ Y}$$

M could be defined such that: $m(i,j) = 1$ iff $d2(z_i, z_j) < t_0$

(without “border effects”).

- See Kendall (1941), Mantel, (1967), Quade (1974), Hubert (1985).

Comparison with $\mathbf{V^*} = \textit{(1/n)}\ \mathbf{Y'(\ I - Z(Z'Z)^{-1}Z')\ (\ I - Z(Z'Z)^{-1}Z')\ Y}$

- « Quasi-Polynomial regression » instead of linear regression.

Internal graphs

The idea is to describe the local structure with a graph:

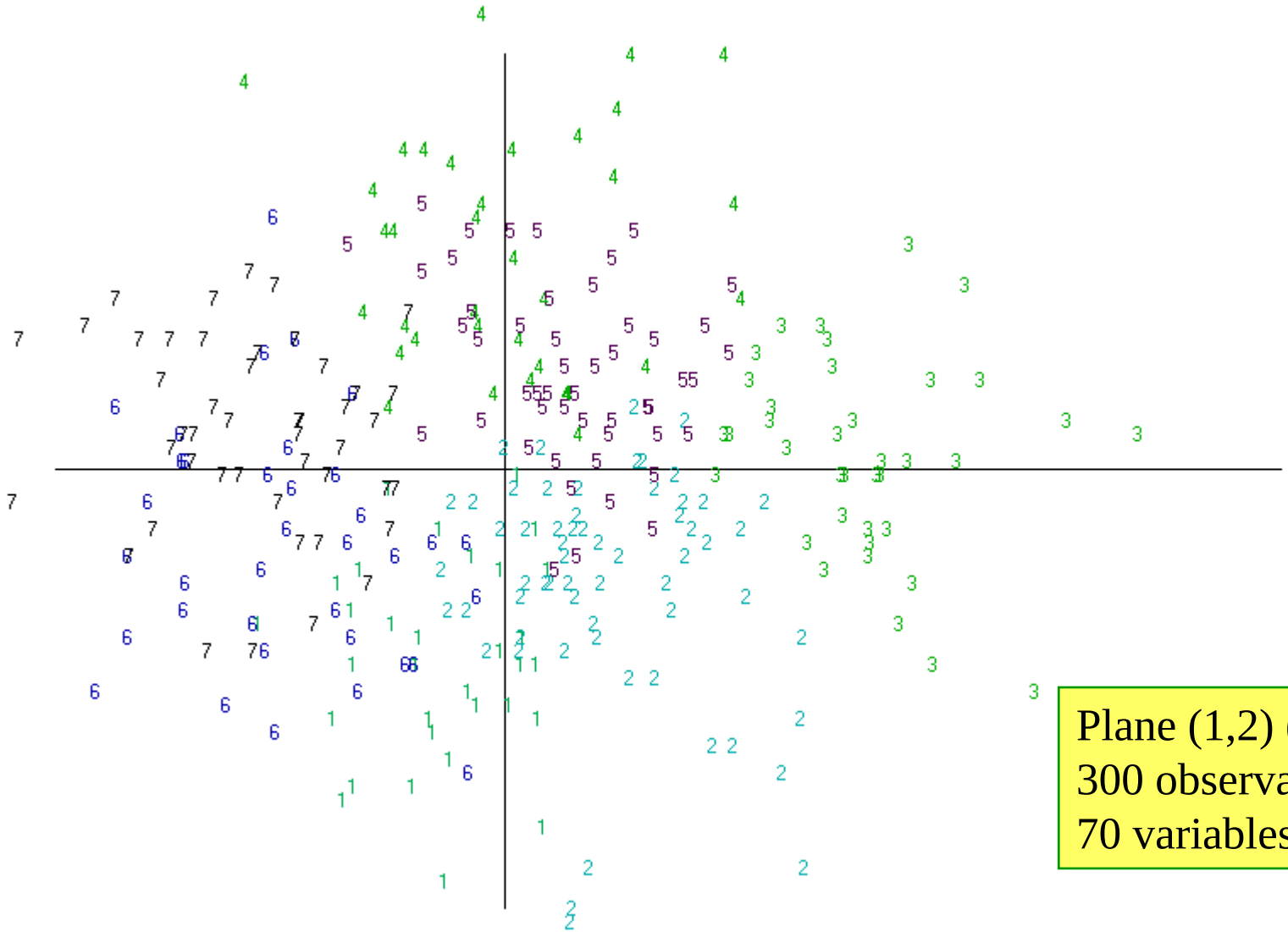
➔The k-nearest neighbours graph (1 ≤ k << n)

➔ A distance threshold graph

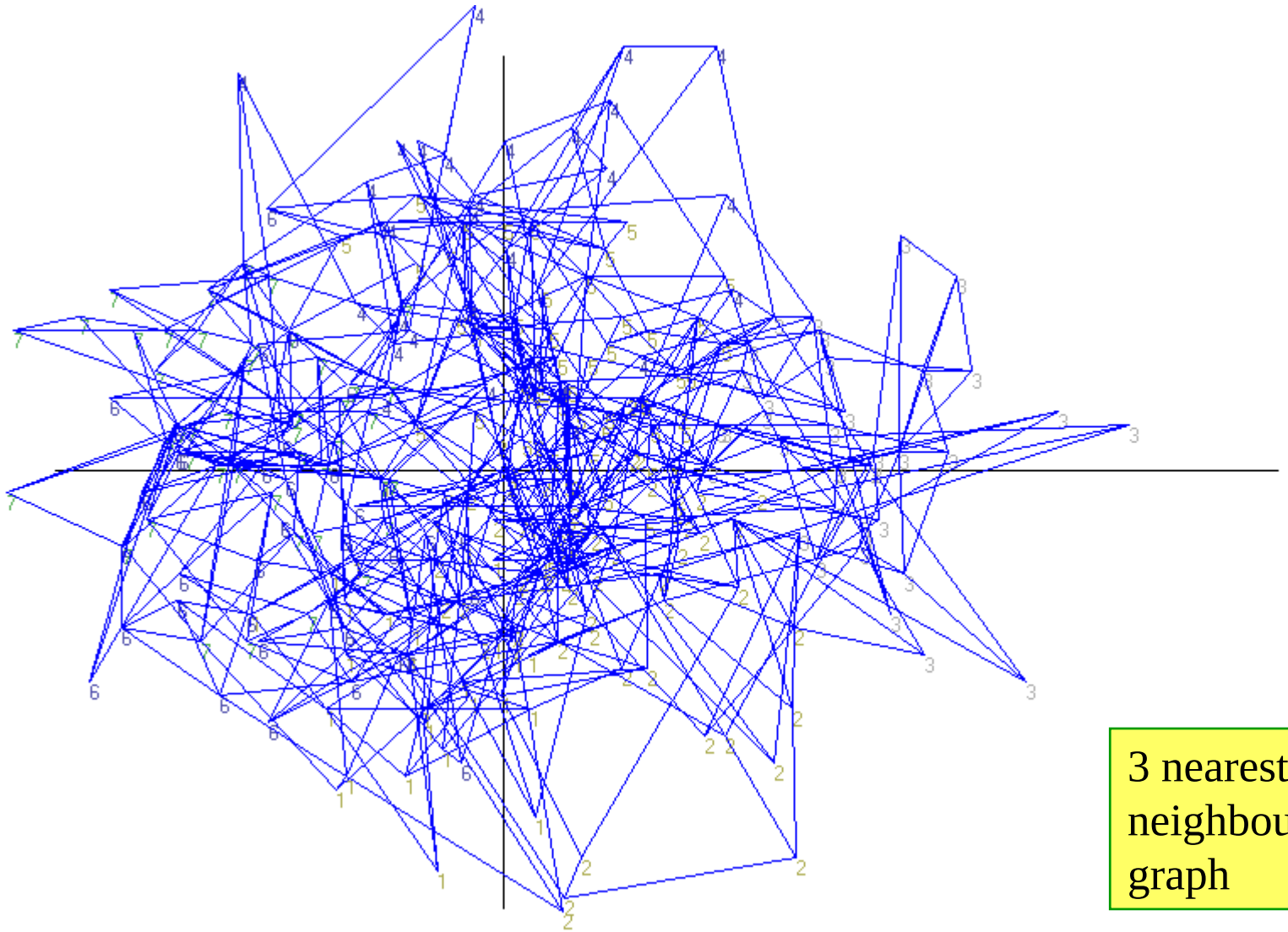
The local covariance matrix defined via such graphs, through the formula

$$\mathbf{V^*} = (1/n)\ \mathbf{Y'(\ I - N^{-1}M)'\ (\ I - N^{-1}M)\ Y}$$

could delinearize the search for the principal axe…

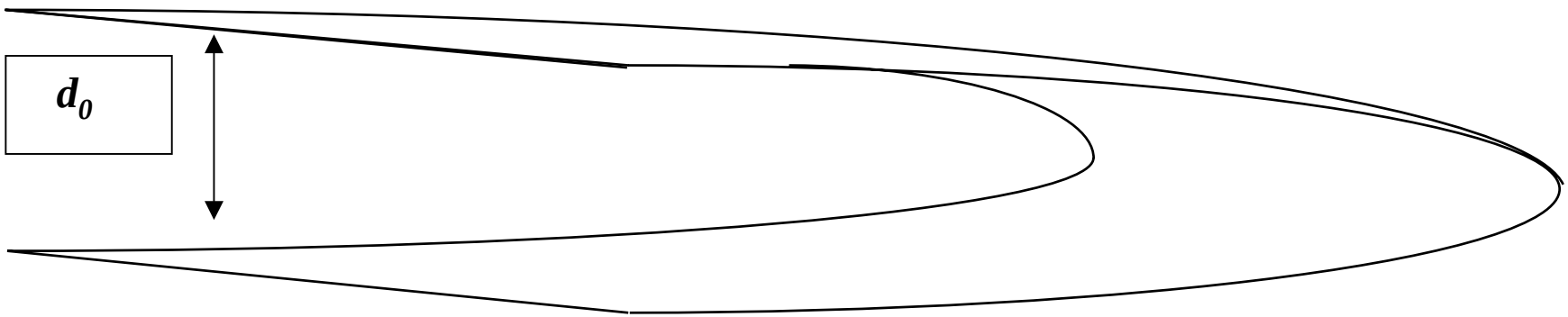


Plane (1,2) (PCA)
300 observations
70 variables



3 nearest neighbours graph

Example of unfolding



The distance threshold $\boldsymbol{d}$ *must be less than* $\boldsymbol{d_0}$ *to allows the contiguity analysis to unfold the scattering diagram*

Example 1: Anderson -Fisher Iris data set:

The 3 a priori categories could be ignored...

The behavior of the series of nearest neighbours graphs is stressed

```
  1  5.10  3.50  1.40   .20     1
  2  4.90  3.00  1.40   .20     1
  3  4.70  3.20  1.30   .20     1
  4  4.60  3.10  1.50   .20     1
  .........
 51  7.00  3.20  4.70  1.40     2
 52  6.40  3.20  4.50  1.50     2
 53  6.90  3.10  4.90  1.50     2
  .........
101  6.30  3.30  6.00  2.50     3
102  5.80  2.70  5.10  1.90     3
103  7.10  3.00  5.90  2.10     3
```

Processing Iris data

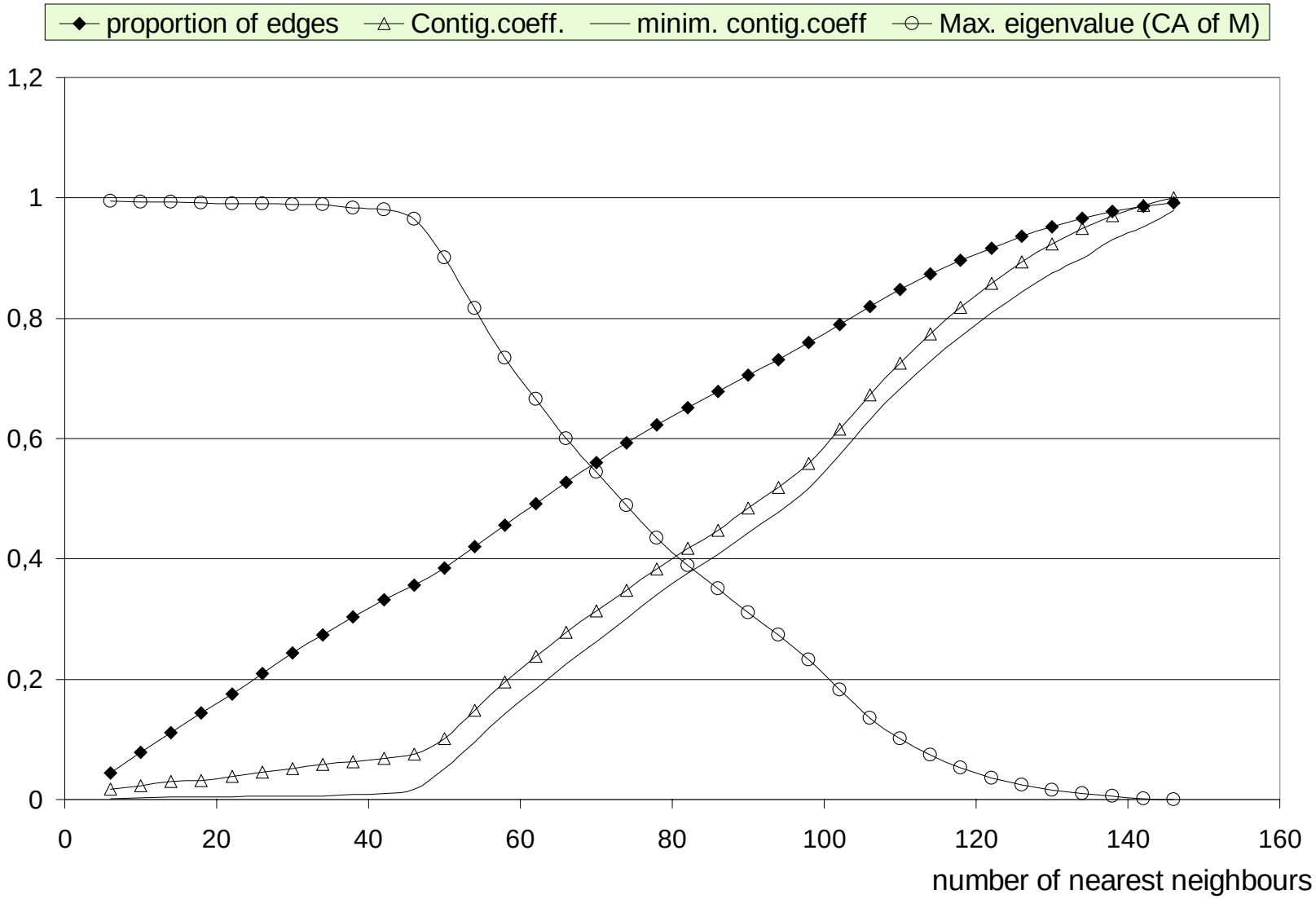
PCA
Visualization (V)

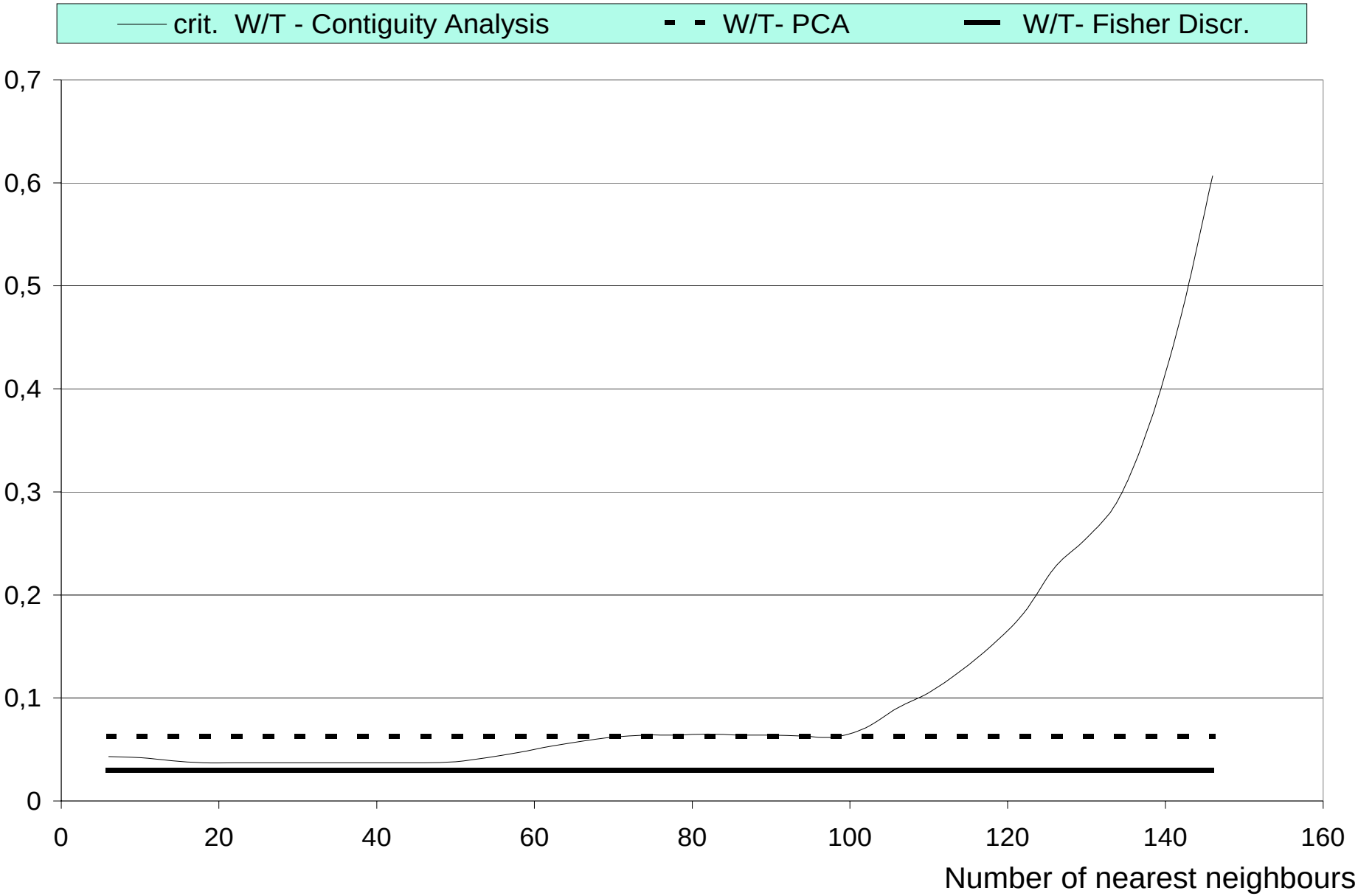
External specific contiguity graph: Fisher discrimination
(V)

Some nearest neighbours graphs (V)
Contiguity analysis (V)

Some threshold graphs
Contiguity analysis (V).

Numerical synthesis...


proportion of edges
Contig.coeff.
minim. contig.coeff
Max. eigenvalue (CA of M)
1,2
1
0,8
0,6
0,4
0,2
0
0
20
40
60
80
100
120
140
160
number of nearest neighbours


crit. W/T - Contiguity Analysis
W/T- PCA
W/T- Fisher Discr.
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
20
40
60
80
100
120
140
160
Number of nearest neighbours

Conclusion

- Transparency of the process (analytical solutions)

- Extension of linear disriminant analysis to more complex patterns

- Possibility of robust PCA and non linear partial correlation.

- Possibility of assessment.

ALUJA T. and LEBART L. (1984): Local and Partial Principal Component Analysis and Correspondence Analysis, *COMPSTAT Proceedings*. Physica Verlag, Vienna, 113-118.

ANSELIN L. (1995): Local indicators of spatial association . *Geog. Anal.*, 27, 2, 93-115.

ART D., GNANADESIKAN R., KETTENRING J.R. (1982): Data Based Metrics for Cluster Analysis, *Utilitas Mathematica*, 21 A, 75-99.

BENZECRI, J.P. (1973): *Analyse des Données: Correspondances*. Dunod, Paris.

BURTSCHY B., LEBART L. (1991): Contiguity analysis and projection pursuit. In*: Appl. Stoch. Mod. and Data Anal*. R. Gutierrez et al., Eds, World Scientific, Singapore, 117-128.

CAZES P. (1986) Correspondance entre deux ensembles et partition de ces deux ensembles, *Les Cahiers de l'Analyse des Données*, vol.XI, no.3, 335-340.

CHUNG F.R.K., *Spectral Graph Theory*. CBMS Reg. Conf. Ser. Math. 92, American Mathematical Society, 1997.

CLIFF A.D. et ORD J.K. (1981): *Spatial Processes: Models and Applications*. Pion, London.

COTTRELL M., ROUSSET P. (1997): The Kohonen Algorithm: a powerful tool for analysing and representing multidimensional qualitative and quantitative data. In: *Biological and Artificial Computation : From Neuroscience to Technology*. J. Mira, R. Moreno-Diaz, J. Cabestany, (eds), Springer, 861-871.

ESCOFIER B. (1989): Multiple correspondence analysis and neighboring relation. In: *Data Analysis, Learning Symbolic and Numeric Knowledge*. Diday E. (ed.), Nova Science Publishers, New York, 55-62.

FARAJ A. (1993): Analyse de contiguité: une analyse discriminante généralisée à plusieurs variables qualitatives. *Revue Statist. Appl*., 41, (3), 73-84.

GEARY R.C. (1954): The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician*, 5, 115-145.

GNANADESIKAN R., KETTENRING J.R. et LANDWEHR J.M. (1982): Projection Plots for Displaying Clusters, In: *Statistics et Probability*. G. Kallianpur *et al*., eds, North-Holland.

GOWER J. C. (1984): Procrustes analysis. In: *Handbook of Applicable Mathematics*. 6, Lloyd E.H. (ed.), J. Wiley, Chichester, 397-405.

HUBERT L. (1985): Combinatorial data analysis: association and partial association. Psychometrika, 50, 4, 449-467.

KLEIWEG P. (1996*): Een inleidende cursus met practica voor de studie Alfa-Informatica*. Master's thesis, Rijksuniversiteit Groningen, 1996.

KOHONEN T.(1989): *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 3rd edition, 1989.

KOREN Y., CARMEL L., HAREL D., ACE: a Fast Multiscale Eigenvectors Computation for Drawing Huge Graphs, *Proceedings of IEEE Information Visualization, 2002*, p 137-144.

LAFOSSE R. (1985): *Analyse Procustéenne de deux tableaux*. Thèse, Université de Toulouse.

LE FOLL Y. (1982): Pondération des distances en analyse factorielle. *Statist. et Anal. des Données*. 7, 13-31.

LEBART L. (1969): Analyse Statistique de la Contiguité, *Publ. de l'ISUP*. XVIII, 81-112.

LEBART, L. (2000): Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds): *Data Analysis*. Springer,Berlin, 233--244.

LEBART, L. (2001): Representing words and texts through contiguity analysis. In: *ASMDA 2001, 10th International Symposium on Applied Stochastic Models and Data Analysis*. G. Govaert, J. Janssen, N. Limnios (eds), UTC, Compiègne, 654-659.

LEBART, L., SALEM, A. and BERRY, L. (1998): *Exploring Textual Data*. Kluwer, Dordrecht.

LEBART L., MIRKIN B. (1993): Correspondence Analysis and Classification. In*: Multivariate Analysis: Future Directions 2*. Cuadras C.M. and Rao C.R., (eds), North-Holland, 341-357.

LEBART L., MORINEAU A. PIRON M., L. (1998): *Statistique Exploratoire Multidimensionnelle,*. Dunod, Paris.

LEBART L., PIRON M., STEINER J.F. (2003): *La Sémiométrie*. Dunod, Paris.

MATHERON G. (1963): Principles of geostatistics. *Economic Geology*. 58, 1246-1266.

MEOT A., CHESSEL D. et SABATIER R. (1993): Opérateur de voisinage et analyse des données spatio-temporelles. In *Biométrie et environnement*, Lebreton J.-D., Asselain B., (eds), Masson, Paris, 45-71.

MOM A. (1988): *Methodologie Statistique de la Classification des reseaux de transport*. Thèse, Université des Sciences et Techniques du Languedoc, Montpellier.

RIPLEY B. D. (1981): *Spatial Statistics*. J. Wiley, New York.

SCHONEMANN P. H. (1968): On two-sided orthogonal procrustes problems. *Psychometrika*. 33, 19-33.

STEINER J.-F. and AULIARD, O. (1992): La sémiometrie: un outil de validation des réponses. In *Qualité de l'Information dans les Enquêtes*, ASU (ed) Dunod, Paris, 241-274.

TUCKER, L. R. (1958): An inter-battery method of factor analysis. *Psychometrika*. 23, (2).

VON NEUMANN, J.(1941): Distribution of the ratio of the mean square successive differences to the variance. *Ann. of Math. Statistics*. 12, 367-395.

Danke

Thank You

Obrigado

Grazie

Domo Arigato

Gracias

Merci

Choukrane