COMPSTAT'2010

# Data Mining and Multiple Ordered Correspondence via Polynomial Transformations

Rosaria Lombardo

Second University of Naples, Via Gran Priorato di Malta, 81043 Capua (CE) -Italy-

rosaria.lombardo@unina2.it

# What will we consider?

- Data Mining and Customer Interaction System Data
- Exploring huge data sets $\Rightarrow$ Customer Satisfaction and Job Satisfaction studies
- Collecting ordered categorical variables
- Ordered multiple correspondence analysis -OMCA- $\Rightarrow$ Singular Value Decomposition and Hybrid Value Decomposition
- Applications of OMCA to customer satisfaction and job satisfaction data sets

# The Learning Management System Data

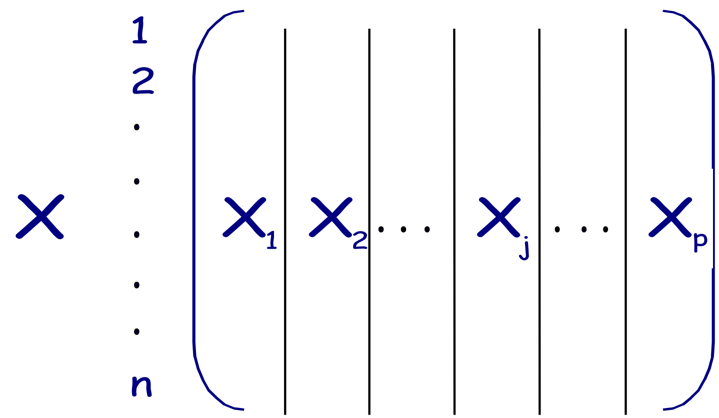
- The Learning Management System data and the subsequent **Customer Interaction System data** can help to provide "**Early Warning System data**" for **risk detection** in enterprises

- various **EWSs** have been established (Kim *et al.*, 2004): for detecting fraud, for credit-risk evaluation (Phua, *et al.*, 2009) , to detection of risks potentially existing in medical organizations, to support decision making in **customer-centric planning tasks** (Lessman & Vob, 2009)

- we focus on EWS of LMSD for customer-centric planning tasks, to develop **exploratory tools** that identify at-risk customers and allow for more timely interventions

# Multiple Correspondence Analysis

$\mathbf{X_k}$ $\Rightarrow$ indicator matrix of dimension $n \times J_k$ of the $k$.th variable




Aim: to analyse large survey data:

$\mathbf{X}=[\mathbf{X_1}|..|\mathbf{X_p}]$ complete disjunctive/ indicator matrices of *P* variables

- rows $\Rightarrow$ *individuals/observations /units*
- *columns* $\Rightarrow$ ordered categories $\Rightarrow$ preference data $\Rightarrow$*replying questionnaire*

Fisher (1940), Guttman (1941), Hayashi (1952), Benzecri (1973) Gifi(1981), Greenacre (1984), etc…

# Multiple CA via the Indicator Super-Matrix

$$SVD\left(\frac{1}{p\sqrt{n}}\mathbf{X}\mathbf{D}^{-1/2}\right) = \mathbf{\Phi}\,\mathbf{\Lambda}_X\,\mathbf{Y}'$$

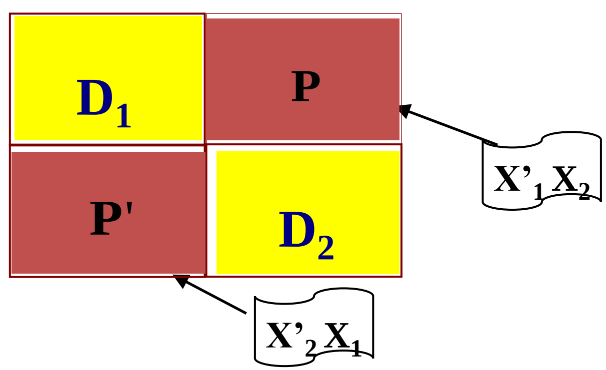
Column Singular Vectors $\mathbf{Y}'\mathbf{D}\mathbf{Y} = \mathbf{I}$

Row Singular Vectors $\mathbf{\Phi}'\mathbf{\Phi} = \mathrm{I}$

where D is the super-diagonal matrix $\mathbf{D} = \begin{pmatrix} \mathbf{D_1} & 0 \\ 0 & \mathbf{D_2} \end{pmatrix}$

We could also consider the **Burt matrix** constructed for two variables P=2

$\mathbf{B = X'X} \Rightarrow$




$$\mathbf{D_k} = diag\left(p_{\bullet 1_k}, \ldots, p_{\bullet J_k}\right)$$

$$Total \quad Inertia = trace\left(\Lambda_X^2\right)$$

Remember that the sum of squares of a non-diagonal sub-matrix equals the Pearson chi-squared statistic divided by *n* (Bekker & de Leeuw ,1988)

# Ordered MCA

- **Hybrid Value Decomposition** (Lombardo & Meulman, 2010, Lombardo & Beh, 2010)**– combining features of Singular Value Decomposition and Bivariate Moment Decomposition (Best & Rayner, 1996; Beh, 1997;1998)**
- Tools: **orthogonal polynomials** for ordered categorical variables by Emerson (1968), **singular vectors** of indicator super-matrix
- Visualising the relationships among ordinal-scale categories and *simultaneously* representing the **units in clusters**
- there is extra information to be obtained, concerning the **statistical significance** of the decomposed inertia

**Data trend** interpretation

COMPSTAT'2010

# Hybrid Decomposition for OMCA

$$HD\left(\frac{1}{p\sqrt{n}}\mathbf{XD}^{-1/2}\right)=\mathbf{\Phi Z \Psi}'$$

Orthogonal Polynomials (categories) $\mathbf{\Psi}'\mathbf{D\Psi}=\mathrm{I}$

Singular Vectors (for rows, or individuals) $\mathbf{\Phi}'\mathbf{\Phi}=\mathrm{I}$

where

$$\mathbf{Z}=\frac{1}{p\sqrt{n}}\mathbf{\Phi}'\mathbf{XD}^{-1/2}\mathbf{\Psi}$$

and D is the super-diagonal matrix consisting of orthogonal polynomials for the ordinal variables

$$Total\ \ Inertia = trace(\mathbf{Z}'\mathbf{Z}) = trace(\mathbf{ZZ}') = trace\left(\Lambda_X^2\right)$$

# Properties of OMCA

*OMCA* $\Rightarrow$ *permits to decompose the inertia in function of eigenvalues and of polynomial trasformations of different degree* associated to the ordered categorical variables

**Property 1** **the total inertia** *can be expressed in terms of squared z-values (bivariate moments) and eigenvalues*

$$\text{Total Inertia} = \sum_{m=1}^{M}\sum_{k=1}^{p}\sum_{v_k=1}^{(J_k-1)} z^2_{mv_k} = \sum_{m=1}^{M} \lambda^2_{X_m}$$

*Where M=J-p is the number of non-trivial solutions*

*We can compute the contribution of the linear component to the overall inertia*

**Property 2** *it is possible to identify which polynomial component (linear, quadratic or higher order) more contributes to the eigenvalue and so to the inertia of each axis.*

For example the first non trivial eigenvalue $\lambda^2_{X_1} = z^2_{11} + z^2_{12} + \ldots + z^2_{1,J-p}$

See also Beh (2001) for p = 2

# Graphical Displays in *OMCA*

1. Individual coordinates

$$\mathbf{F} = \mathbf{\Phi Z} = \frac{1}{p\sqrt{n}}\mathbf{X\Psi}$$

2. Category coordinates

$$\mathbf{G} = \frac{1}{p/\sqrt{n}}\mathbf{D}^{-1}\mathbf{\Psi Z}' = \frac{1}{p/\sqrt{n}}\mathbf{D}^{-1}\mathbf{X}'\mathbf{\Phi}$$

$$Total \quad Inertia = trace(\mathbf{F}'\mathbf{F}) = trace(\mathbf{G}'\mathbf{DG}) = trace\left(\mathbf{\Lambda}_X^2\right)$$

Category coordinates are identical to MCA coordinates

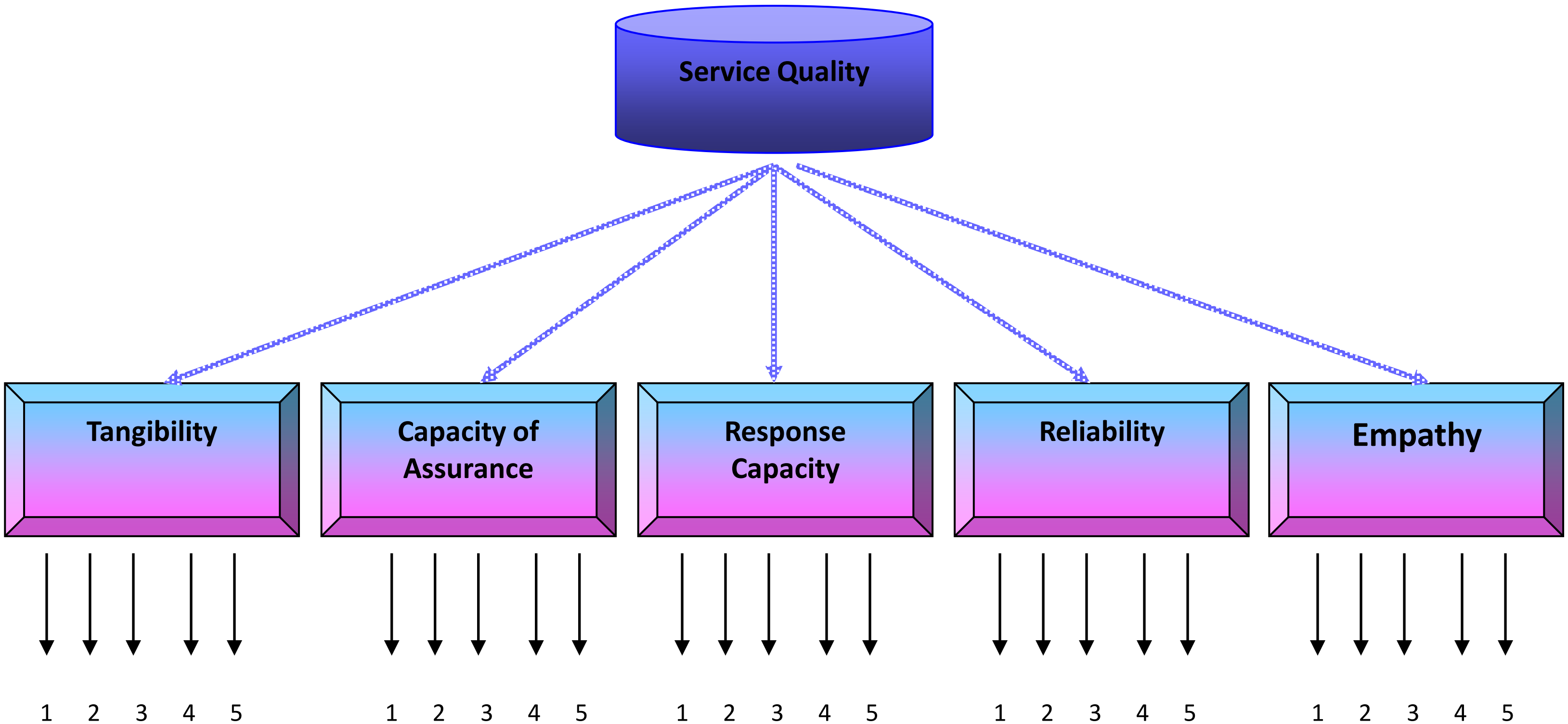
Individual coordinates computed by polynomials <u>are not the same</u> as the “classical” ones $\Rightarrow$ clusters of units in relation with the expressed ordered scores

# How can you consider nominal variables without destroying the ordered structure?

- Ordered multiple correspondence analysis and nominal variables
- Splitting the ordinal data using the nominal categories
- Apply OMCA to these data sub-sets

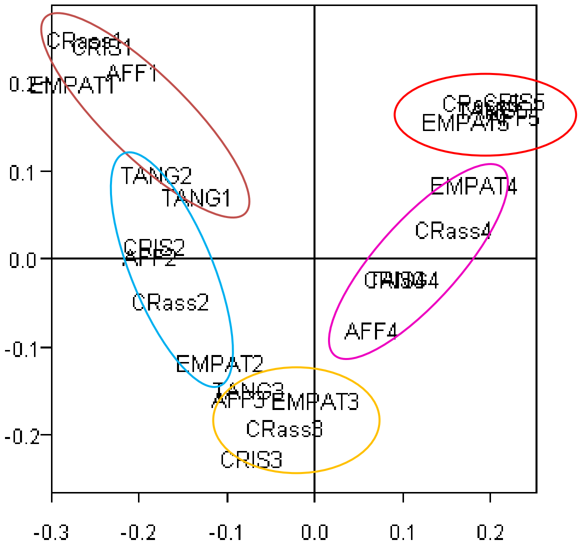
# The Evaluation of Customer Satisfaction in Health Care Services

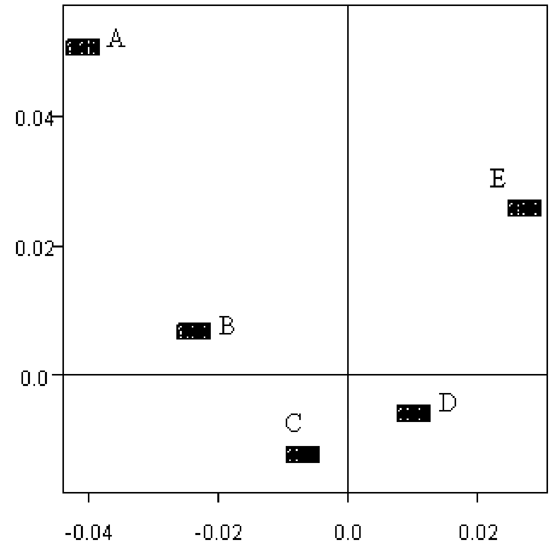
*To gauge the quality of **five** key characteristics of a Naples hospital based on a sample of 511 patients.*




*Ordered Responses*: 1 = Not satisfied, 5 = Very satisfied

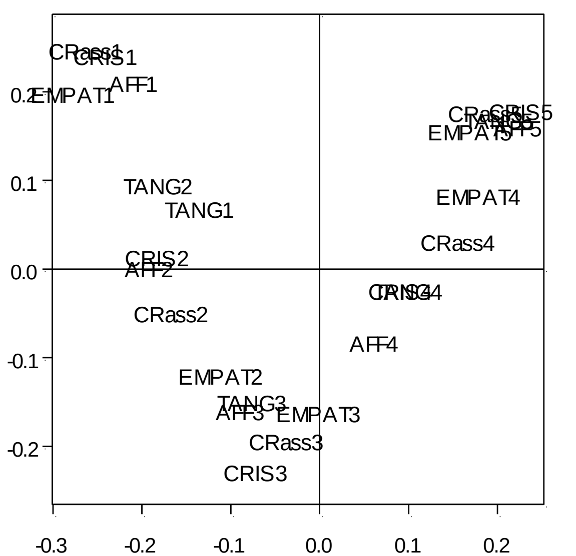
# Comparing OMCA and MCA in overall hospital

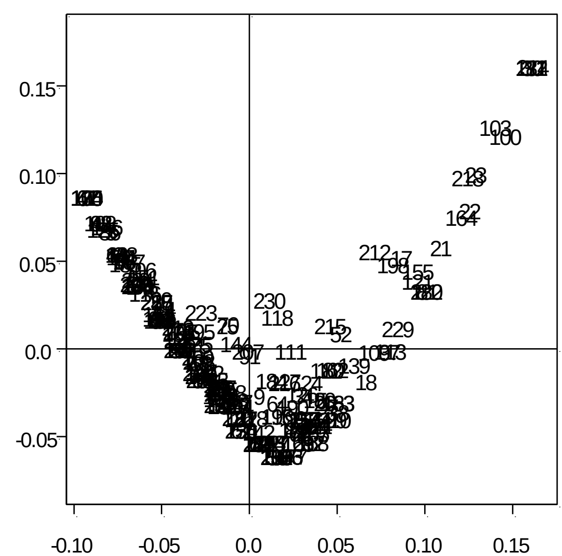



| Cluster | % of Patients in Cluster |
|---|---|
| E: very much satisfied | 13,6% |
| D: a lot satisfied | 41,7% |
| C: satisfied | 30,6% |
| B: little satisfied | 4,7% |
| A: not satisfied | 9,4% |

OMCA plots




MCA plots

# Ordered Multiple Analysis in overall hospital

*Table 1: Decomposition of the first two non-trivial eigenvalues and chi-square tests.*

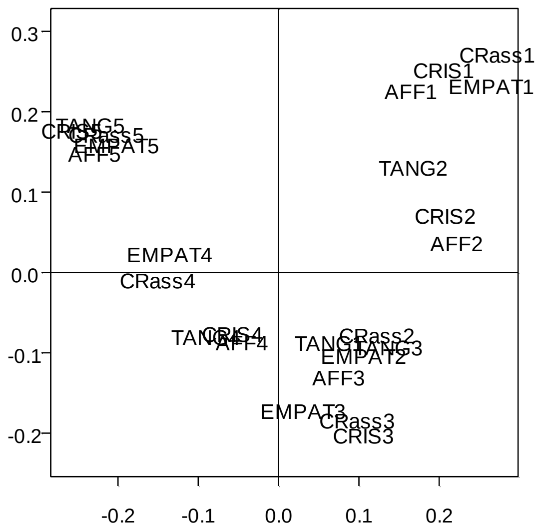
| Variable | Component | $z^2_{1(v_k)}=\lambda_1^2$ | $\chi^2$ | $z^2_{2(v_k)}=\lambda_2^2$ | $\chi^2$ | d.f. |
|---|---|---|---|---|---|---|
| Tangibility | Location | 0.104 | 73.230*** | 0.030 | 2.093 | 8 |
| | Dispersion | 0.000 | 0.328 | 0.051 | 35.956*** | 8 |
| | Skewness | 0.001 | 0.362 | 0.008 | 2.398 | 8 |
| | Kurtosis | 0.002 | 1.567 | 0.000 | 5.936 | 8 |
| Reliability | Location | 0.140 | 98.781*** | 0.000 | 0.282 | 8 |
| | Dispersion | 0.000 | 0.219 | 0.099 | 69.999*** | 8 |
| | Skewness | 0.001 | 0.368 | 0.003 | 2.217 | 8 |
| | Kurtosis | 0.000 | 0.038 | 0.000 | 0.033 | 8 |
| Capability of Response | Location | 0.153 | 107.539*** | 0.002 | 1.154 | 8 |
| | Dispersion | 0.003 | 1.950 | 0.131 | 92.568*** | 8 |
| | Skewness | 0.001 | 0.523 | 0.008 | 5.806 | 8 |
| | Kurtosis | 0.000 | 0.027 | 0.002 | 1.748 | 8 |
| Capability of Assurance | Location | 0.151 | 106.328*** | 0.002 | 1.106 | 8 |
| | Dispersion | 0.005 | 3.313 | 0.119 | 84.106*** | 8 |
| | Skewness | 0.001 | 0.529 | 0.013 | 9.315 | 8 |
| | Kurtosis | 0.001 | 0.454 | 0.000 | 0.011 | 8 |
| Empathy | Location | 0.143 | 101.009*** | 0.003 | 2.094 | 8 |
| | Dispersion | 0.003 | 2.242 | 0.093 | 65.398*** | 8 |
| | Skewness | 0.001 | 0.615 | 0.016 | 11.082 | 8 |
| | Kurtosis | 0.002 | 1.665 | 0.000 | 0.020 | 8 |
| | Total | 0.711 | 501.088*** | 0.558 | 393.320*** | 160 |

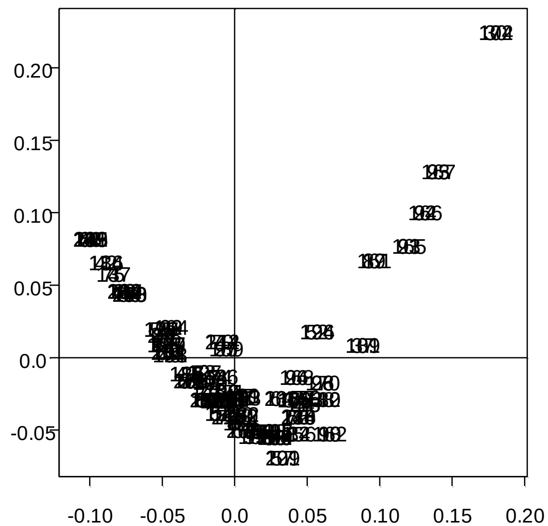
The statistically significant components are identified at three levels of significance:
0.01(***)
0.05 (**)
0.10 (*)

Tangibility, Reliability, Capability of response, Capability of assurance and Empathy account for 15.9%, 18.3%, 25.6%, 24.6% and 20.1% of the explained inertia

# Ordered Multiple Analysis in a division of the hospital

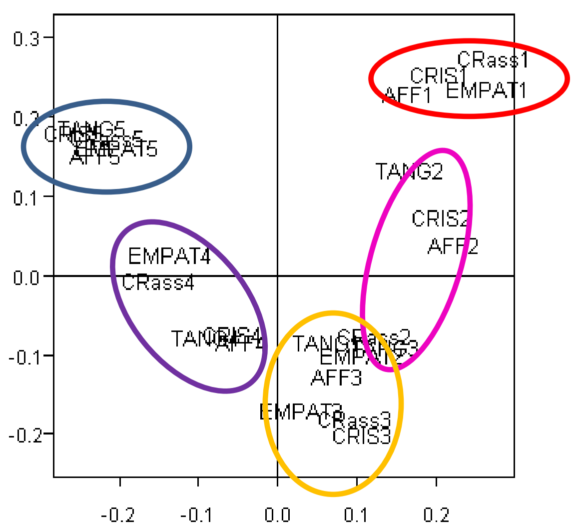
Plot of MCA

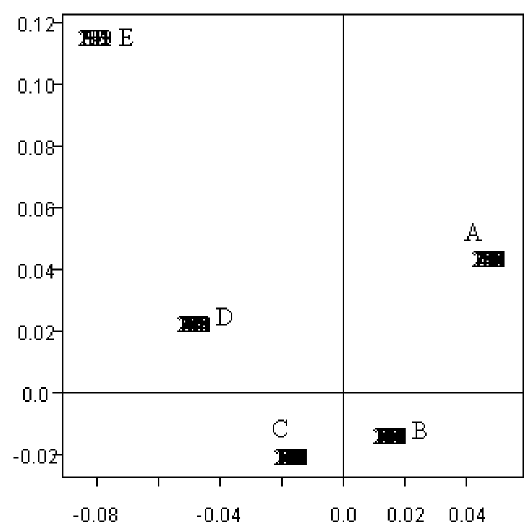



MCA plots

OMCA plots




| Cluster | % of Patients in Cluster |
|---|---|
| E | 15.3% |
| D | 36.1% |
| C | 36.1% |
| B | 2.8% |
| A | 9.7% |

# Ordered Multiple Analysis in gynaecology division

*Table 1: Decomposition of the first two non-trivial eigenvalues and chi-square tests.*

| Variable | Component | $z^2_{1(v_k)}=\lambda_1^2$ | $\chi^2$ | $z^2_{2(v_k)}=\lambda_2^2$ | $\chi^2$ | d.f. |
|---|---|---|---|---|---|---|
| Tangibility | Location | 0.11 | 22.76*** | 0.008 | 1.74 | 8 |
| | Dispersion | 0.01 | 1.52 | 0.019 | 4.16 | 8 |
| | Skewness | 0.00 | 0.26 | 0.033 | 7.22 | 8 |
| | Kurtosis | 0.00 | 0.10 | 0.013 | 2.79 | 8 |
| Reliability | Location | 0.13 | 28.26*** | 0.001 | 0.17 | 8 |
| | Dispersion | 0.00 | 0.28 | 0.088 | 19.06** | 8 |
| | Skewness | 0.00 | 0.87 | 0.009 | 1.92 | 8 |
| | Kurtosis | 0.00 | 0.04 | 0.002 | 0.47 | 8 |
| Capability of Response | Location | 0.16 | 35.38*** | 0.001 | 0.12 | 8 |
| | Dispersion | 0.00 | 0.17 | 0.141 | 30.42*** | 8 |
| | Skewness | 0.00 | 0.34 | 0.005 | 1.11 | 8 |
| | Kurtosis | 0.00 | 0.10 | 0.001 | 0.29 | 8 |
| Capability of Assurance | Location | 0.16 | 35.51*** | 0.000 | 0.00 | 8 |
| | Dispersion | 0.00 | 0.06 | 0.130 | 28.16*** | 8 |
| | Skewness | 0.00 | 0.12 | 0.012 | 2.65 | 8 |
| | Kurtosis | 0.00 | 0.47 | 0.001 | 0.32 | 8 |
| Empathy | Location | 0.14 | 29.84*** | 0.000 | 0.06 | 8 |
| | Dispersion | 0.00 | 0.27 | 0.107 | 23.02*** | 8 |
| | Skewness | 0.00 | 0.24 | 0.013 | 2.88 | 8 |
| | Kurtosis | 0.00 | 0.21 | 0.001 | 0.15 | 8 |
| | Total | 0.73 | 156.81*** | 0.587 | 126.69*** | 160 |

# Survey on Job satisfaction in Social Enterprises of Caserta – Italy-

1426 questionnaires
Ordered categorical variables with 4 categories

**Extrinsic Satisfaction**

E1 – organization and flexibility;
E2 – stability;
E3 – wage;
E4 –autonomy and independence.

**Intrinsic Satisfaction**

I1 – relationships with users;
I2 – relationships with managers;
I3 – recognized job
I4 – involvement in decisions
I5 – trasparency of relationships.

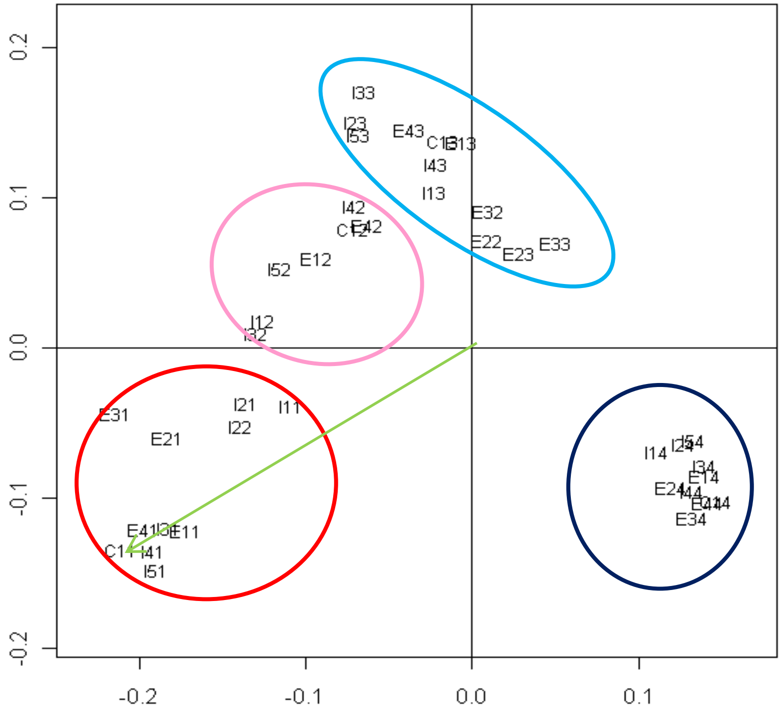
**Total Satisfcation**

C1- actual job

Nominal variables

- **Partner or not Partner**
- Title of study
- Job time
- Activity Areas
- ex-ante Motivation

# OMCA : Partner and not Partner in Social Enterprises



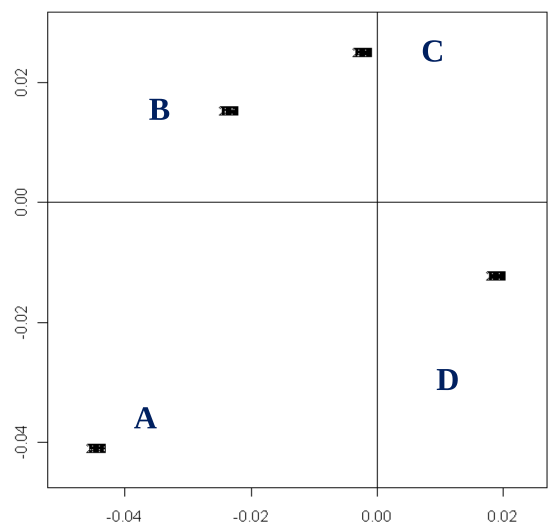

**Relationships with the general unsatisfaction** (C1):

- Intrinsic satisfaction I3 (recognition), I4 (involvement) e I5 (trasparency).

- Extrinsic Satisfcation: E1 (organization) e E4 (authonomy).

| | | |
|---|---|---|
| | partner | non-partner |
| A: not satisfied | 9,8 | 12,3 |
| B: little satisfied | 16,1 | 18,2 |
| C: satisfied | 28,1 | 40,4 |
| D: a lot satisfied | 46,0 | 29,1 |

- More satisfied workers are partners of social enterprises (46% against 29%)



| | Polynomial component | Inertia axis I | chi-2 | Inertia axis II | chi-2 | d.f. |
|---|---|---|---|---|---|---|
| **E1-Organization** | Location | 0,13 | **29,21**$^{***}$ | 0,00 | 0,57 | 6 |
| | Dispersion | 0,00 | 0,29 | 0,10 | **22,04**$^{***}$ | 6 |
| | Skewness | 0,00 | 0,11 | 0,00 | 0,14 | 6 |
| **E2-stability** | Location | 0,10 | **22,69**$^{***}$ | 0,00 | 0,55 | 6 |
| | Dispersion | 0,00 | 0,92 | 0,07 | **14,92**$^{**}$ | 6 |
| | Skewness | 0,02 | 3,46 | 0,00 | 0,00 | 6 |
| **E3-Wage** | Location | 0,13 | **28,49**$^{***}$ | 0,01 | 2,08 | 6 |
| | Dispersion | 0,01 | 1,64 | 0,09 | **19,63**$^{***}$ | 6 |
| | Skewness | 0,01 | 1,67 | 0,00 | 0,09 | 6 |
| **E4-autonomy** | Location | 0,12 | **25,77**$^{***}$ | 0,00 | 0,22 | 6 |
| | Dispersion | 0,00 | 0,88 | 0,10 | **21,40**$^{***}$ | 6 |
| | Skewness | 0,01 | 1,34 | 0,00 | 0,13 | 6 |
| **C1-Actual Job** | Location | 0,15 | **32,84**$^{***}$ | 0,00 | 0,25 | 6 |
| | Dispersion | 0,00 | 1,10 | 0,11 | **24,73**$^{***}$ | 6 |
| | Skewness | 0,00 | 1,08 | 0,00 | 0,00 | 6 |
| | Total | 0,68 | **151,49**$^{***}$ | 0,48 | **106,74**$^{***}$ | 90 |

# Conclusion and Perspectives

In customer satisfaction studies:
**Likert items** for the evaluation of quality aspects
and personal information,
the **splitting of individuals** with respect to the nominal categories and
the **automatic aggregation of individuals** in so many clusters as the number of the ordered categories provide an
**early warning system data** that help to identify at-risk customers/consumers/workers and suggest for more timely interventions **to improve quality in enterprises**.

In perspective: **External Information** in OMCA, **Stability** of OMCA.

# Main References


BABAKUS, E., and MANGOLD, G. (1992). Adapting the Servqual scale to hospital services: an empirical investigation. *Health Services Research Journal*, 767-786.

BEKKER P., & de LEEUW J., (1988). Relations between Variants of Nonlinear Principal Component Analysis. In: Component and Correspondence Analysis (J.L.A. van Rijckevorsel and J. de Leeuw, Eds.). Chichester: John Wiley & Sons.

BEH E. J., (1997). Simple correspondence analysis of ordinal cross-classifications using orthogonal polynomials. *Biometrical Journal*, 39, 589-613.

BEH E. J. , (1998). A comparative study of scores for correspondence analysis with ordered categories. *Biometrical Journal*, 40, 413-429.

BEH, E. J., (2001) . Partitioning Pearson's chi-squared statistic for singly ordered two-way contingency tables. *The Australian and New Zealand Journal of Statistics*, 43, 327-333.

BEST, D. J. & RAYNER, J. C. W., (1996). Nonparametric analysis for doubly ordered two-way contingency tables. *Biometrics*, 52, 1153-1156

EMERSON P. L., (1968) . Numerical construction of orthogonal polynomials from general recurrence formula. *Biometrics*, 24, 696-701.

GREENACRE M. J. , (1984). *Theory and Application of Correspondence Analysis*. Academic Press: London.

LEBART, L., MORINEAU A., & WARWICK K.M., (1984) . *Multivariate Descriptive Statistical Analysis*. Wiley: New York, 1984.

LOMBARDO, R. & MEULMAN, J. (2010). Multiple Correspondence Analysis via Polynomial Transformations of Ordered Categorical Variables. *Journal of Classifiation*, 10, 32-48.

LOMBARDO, R., BEH, E. J , D'AMBRA L.(2007). Non-symmetric correspondence analysis with ordinal variables using orthogonal polynomials. *Computational Statistics & Data Analysis*, 52, 566-577.

LOMBARDO, R. , BEH,. E. J . (2010) . Simple and multiple correspondence analysis for ordinal scale variables. *Journal of Applied Statistics*, in press.