le cnam

Optimal Scaling: New Insights into an Old Problem

Gilbert Saporta

CEDRIC-CNAM, Paris, France

In collaboration with:

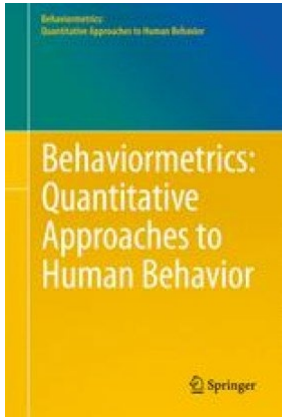
Hervé Abdi, UT Dallas

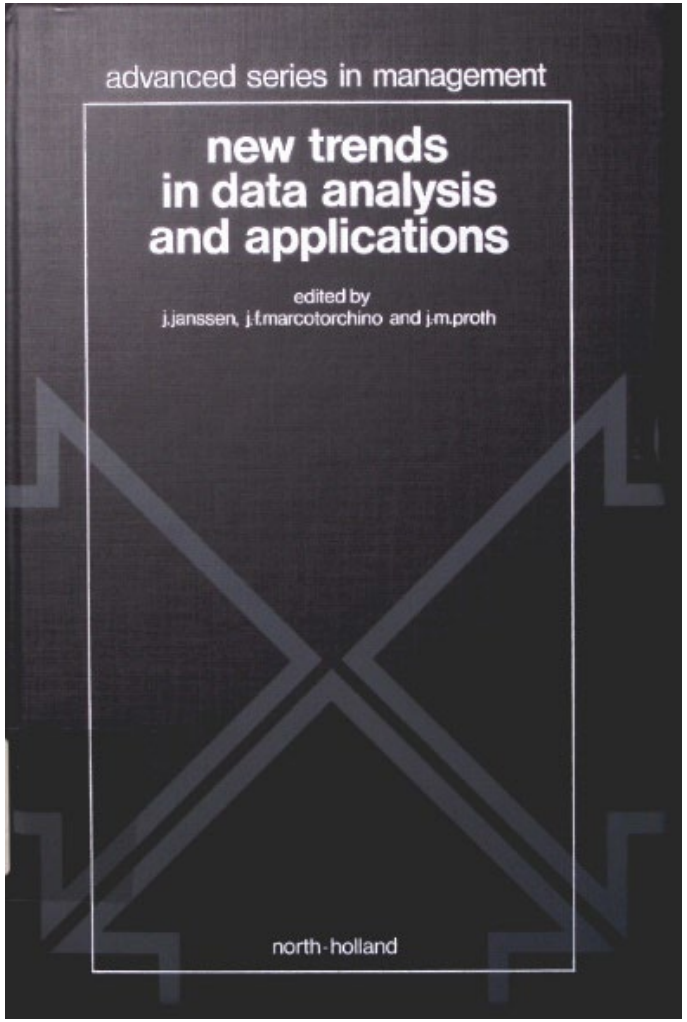
Agostino Di Ciaccio, La Sapienza, Roma

To be published in:

“Analysis of Categorical Data from Historical Perspectives – Essays in Honour of Shizuhiko Nishisato", Springer, 2023

Part of the book series:





Papers presented at the Symposium on New Trends in Data Analysis, held Dec. 7 - 9, 1981 at the European Inst. for Advanced Studies Management, Brussels, and at a second congress held Mar. 11 – 12, 1982 at the Université Libre de Bruxelles, published in 1983

New Trends in Data Analysis and Applications
J. Janssen, J.-F. Marcotorchino, J.-M. Proth (eds.)

MULTIDIMENSIONAL DATA ANALYSIS AND QUANTIFICATION OF CATEGORICAL VARIABLES

Gilbert SAPORTA

IUT de Paris
143 Avenue de Versailles, 75016 Paris

Quantification is a powerful tool for dealing with qualitative variables in data analysis : it is both an efficient way for description and prediction with nominal data and a conceptual framework for presenting many techniques. This paper attempts to be a survey of this topic.

I. USE AND DEFINITION OF QUANTIFICATION

By "quantification" (the words "scaling" or "scoring" are also often used) we mean a transformation of one or several categorical variables into numerical ones. Before describing the way to achieve a quantification (and the problems it involves) it is necessary to argue about the rationale of this approach.

1) Quantification : a device or a fundamental method ?

a) Numerical methods for data which are not numerical.

Outline

1. Introduction
2. Definitions and initial properties
3. Pioneering times
4. The 70s and the domination of alternating least squares
5. Categorical encoding and Machine Learning
6. Conclusion and perspectives

1. Introduction

- Why transform qualitative variables into numerical variables?
  - Apply methods reserved for numerical variables
  - Search for underlying factors
  - Perform non-linear analysis
- A history going back almost a century
  - The beginnings of CA and MCA
  - With surprising links to normal distribution
- Rediscovered in Machine Learning: « categorical data encoding »

2. Definitions and initial properties

- Forrest W. Young (1981):

Optimal scaling is a data analysis technique which assigns numerical values to observation categories in a way which maximizes the relation between the observations and the data analysis model while respecting the measurement character of the data.

- No coding in itself, without a goal or a model
- Retrieves latent scale, groups categories, validates ordinal character

2.1 A coding is a linear combination of indicator variables

$$X \in \{1, 2, .., M\}$$

$$\tilde{X} \in \{a_1, a_2, .., a_M\}$$

$$\tilde{X} = \sum_{m=1}^{M} a_m \mathbf{1}_m$$

- Matrix Formulation

$$\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ . \\ m \end{pmatrix} \quad \tilde{\mathbf{x}} = \begin{pmatrix} a_1 \\ a_2 \\ a_2 \\ a_1 \\ . \\ a_m \end{pmatrix} = \begin{pmatrix} 1000..0 \\ 0100..0 \\ 0100..0 \\ 1000..0 \\ . \\ 0000..1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ . \\ a_m \end{pmatrix}$$

$$\tilde{\mathbf{x}} = \mathbf{X}\mathbf{a}$$

- Subspace of dimension M $\quad \tilde{\mathbf{x}} \in W$
- Redundancy (?): $\sum_{m=1}^{M} \mathbf{1}_m = \mathbf{1}$
- Centered solutions: $\tilde{\mathbf{x}} \in \left\{ \Delta^{\perp} \cap W \right\}$

Variables with ordered modalities

- Require $a_1 \leq a_2 \leq \ldots \leq a_M$
- Reparametrization

$$a_1 = b_1, \;\; a_2 = b_1 + b_2, \;\; .., \;\; a_M = b_1 + b_2 + \ldots + b_M$$

$$b_1 \in \mathbb{R} \quad b_2, \ldots, b_M \geq 0$$

$$\begin{pmatrix} a_1 \\ a_2 \\ . \\ a_M \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ . \\ b_M \end{pmatrix}$$

$$\tilde{\mathbf{x}} = \sum_{m=1}^{M} a_m \mathbf{1}_m = b_1 \mathbf{1}_1 + (b_1 + b_2)\mathbf{1}_2 + \ldots (b_1 + b_2 + \ldots b_M)\mathbf{1}_M$$

$$= b_1(\mathbf{1}_1 + \mathbf{1}_2 + \ldots + \mathbf{1}_M) + b_2(1_2 + \ldots + \mathbf{1}_M) + \ldots + b_M \mathbf{1}_M$$

$$= b_1 \mathbf{1} + b_2(\mathbf{1}_2 + \ldots + \mathbf{1}_M) + b_3(\mathbf{1}_3 + \ldots + \mathbf{1}_M) + \ldots + b_M \mathbf{1}_M$$

- Linear combination with positive coefficients of *M*-1 variables and a constant term

$$\tilde{\mathbf{x}} = b_1\mathbf{1} + \sum_{m=2}^{M} b_m \mathbf{z}_m$$

- An element of the direct sum of the space of constants and a convex polyhedral cone

$$\tilde{\mathbf{x}} \in \left\{ \Delta \oplus \mathcal{C}_{M-1} \right\}$$

2.2 Two simple optimal coding problems

What is the optimal way to quantify a qualitative variable X in order to best predict Y in the least squares sense?

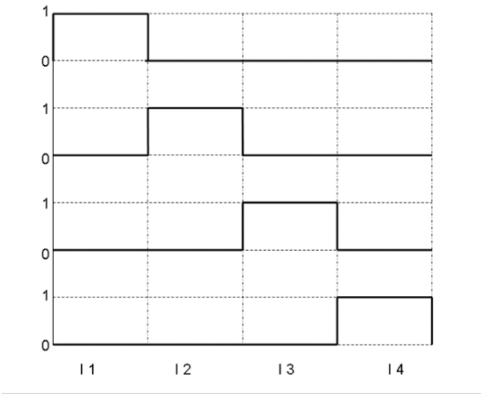
- $\boldsymbol{X}$ **categorical**. Projection of Y on W: multiple regression without constant of Y on $\mathbf{1}_m$ indicators. By orthogonality of indicators, *M* simple regressions. Coding by conditional averages (percentages if *Y* binary) :

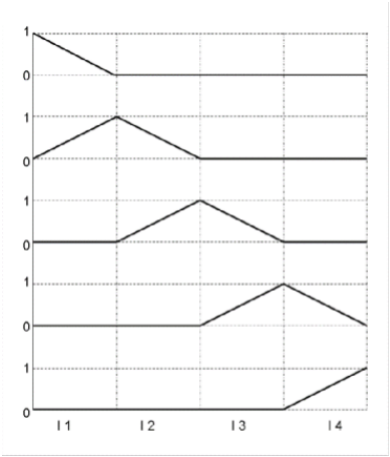
$$a_m = \overline{y}_m$$

- $\boldsymbol{X}$ **ordinal**. Projection on a cone: multiple regression with *M*-1 positivity constraints. PAVA for *Pool Adjacent Violators Algorithm* (Kruskal, 1964) or backward regression with elimination of variables with negative coefficients and iteration (Tenenhaus, 1988).

2.3 Coding and non-linear transformations

- Transforming a numerical variable X into a categorical variable $\mathcal{X}$ by splitting into classes, then recoding $\mathcal{X}$ into a numerical variable allows $\tilde{X}$ non-linear effects to be studied. « Crisp coding »
- But: discontinuities and loss of information.
- Remedy: fuzzy coding





- Generalization: splines and monotone splines (Ramsay,1988)

3. Pioneering times

- Maximize the correlation between two coded variables

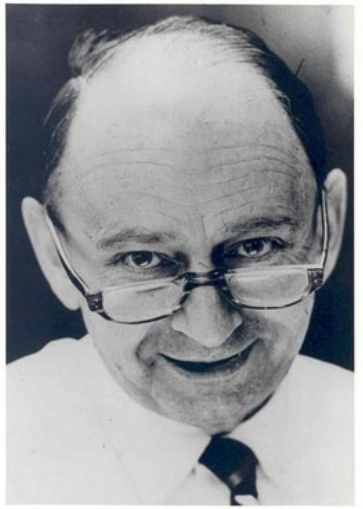
A CONNECTION BETWEEN CORRELATION AND CONTINGENCY

By H. O. HIRSCHFELD, Fitzwilliam House

[Communicated by Mr J. Wishart]

[*Received* 7 May, *read* 28 October 1935]

Hermann Otto Hirschfeld
aka Hartley



1912-1980

Introduction

Let us consider a discontinuous bivariate distribution. That is, let us consider $N \times Q$ non-negative values $p_{\nu q}$ ($\nu = 1, 2, \ldots, N$; $q = 1, 2, \ldots, Q$), being the theoretical probabilities of the νth value of a variate X_μ ($\mu = 1, 2, \ldots, N$) concurring with the qth value of a second variate Y_s ($s = 1, 2, \ldots, Q$).

It is well known that the correlation theory for such a distribution gives much better results, if both regressions are linear, and that these regressions are transformed by a change of the variates X_ν, Y_q. On the other hand the original scales or better values assigned to the X_ν and Y_q are often chosen in a conventional or artificial way and, if a distribution of characteristics is treated, they are not known at all. Thus the following question naturally arises: Given a discontinuous distribution $p_{\nu q}$, is it always possible to introduce (instead of the original variates, if there are any) new values for the variates x_ν, y_q, such that *both* regressions are linear?

- Maximizing discrimination (Fisher)

2. In 1940 Fisher considered contingency tables from the point of view of discriminant analysis. Suppose that 'scores', i.e. arbitrary variate values, are assigned to the rows and also to the columns of a contingency table: what are the best scores to assign to the rows so that a linear function of them will best differentiate the classes determined by the columns, and vice versa? This turns out to be a problem in maximizing the correlation between the scores and the required correlations are those known as 'canonical' in the sense of Hotelling (1936). The work was continued and developed by Maung (1941). In particular, Maung quotes a result by Fisher which gives the observed frequency in terms of the canonical correlations; in fact, if the frequency is a_{ij} with marginal *totals* $a_{i.}$, $a_{.j}$ and total $a_{..}$, and if the canonical correlations are $R_1, R_2, \ldots, R_{m-1}$, we have

$$a_{ij} = \frac{a_{i.}a_{.j}}{a_{..}}\left\{1 + \sum_{1}^{m-1}(x_k y_k R_k)\right\}, \qquad (4)$$

where x and y are the assigned scores corresponding to the given cell.

Lancaster, 1957

For example, in a contingency table individuals are cross classified in two categories, such as eye colour and hair colour, as in the following example (Tocher's data for Caithness compiled by K. Maung of the Galton Laboratory).

Eye colour	Hair colour					
	Fair	Red	Medium	Dark	Black	Total
Blue	326	38	241	110	3	718
Light	688	116	584	188	4	1580
Medium	343	84	909	412	26	1774
Dark	98	48	403	681	85	1315
Total	1455	286	2137	1391	118	5387

Variation among the four eye colours may be regarded as due to variations in three variates defined conveniently in some such way as the following:

Eye colour	x_1	x_2	x_3
Blue	0	0	0
Light	1	0	0
Medium	0	1	0
Dark	0	0	1

We may then ask for what eye colour scores, i.e. for what linear function of x_1, x_2, x_3, are the five hair colour classes most distinct. The answer may be found in a variety of ways. For example, by starting with arbitrarily chosen scores for eye colour, determining from these average scores for hair colour, and using these latter to find new scores for eye colour.

Apart from a contraction of scale by a factor R^2 for each completed cycle, this form tends to a limit, and yields scores such as the following:

Eye colour	x	Hair colour	y
Light	−0·9873	Fair	−1·2187
Blue	−0·8968	Red	−0·5226
Medium	0·0753	Medium	−0·0941
Dark	1·5743	Dark	1·3189
		Black	2·4518

The particular values given above have been standardized so as to have mean values zero, and mean square deviations unity. In the sample from which they are derived each score has a linear regression on the other, the regression coefficient being 0·44627; this is, of course, equal to the correlation coefficient between the two scores regarded as variates. Hotelling has called pairs of functions of this kind canonical components. It may be noticed that no assumption is introduced as to the order of the classes of each category. In Tocher's schedule Light eyes come between Blue and Medium, but the discriminant function puts Blue between Medium and Light, though near the latter.

- At the origin of correspondence analysis of a contingency table $\mathbf{N}$
  - Hirschfeld : transition formulas $\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = \sqrt{\lambda}\mathbf{a} \quad \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \sqrt{\lambda}\mathbf{b}$
  - Fisher: maximization of : $\operatorname{cov}(\mathbf{X}_1\mathbf{a}, \mathbf{X}_2\mathbf{b}) = \frac{1}{n}\mathbf{a}'\mathbf{N}\mathbf{b}$

$$\begin{cases} \mathbf{D}_1^{-1}\mathbf{N}\mathbf{D}_2^{-1}\mathbf{N}'\mathbf{a} = \lambda\mathbf{a} \\ \mathbf{D}_2^{-1}\mathbf{N}'\mathbf{D}_1^{-1}\mathbf{N}\mathbf{b} = \lambda\mathbf{b} \end{cases} \qquad \frac{1}{n}\mathbf{a}'\mathbf{D}_1\mathbf{a} = \frac{1}{n}\mathbf{b}'\mathbf{D}_2\mathbf{b} = \lambda$$

  - Only one dimension!

- Optimal scaling and normal distribution: Lancaster's theorem (1957)

A MAXIMAL PROPERTY OF THE BIVARIATE NORMAL DISTRIBUTION

4. We may consider the variables standardized so as to have unit variance and take $0 < \rho < 1$.

THEOREM. *Let x and y be jointly distributed in the bivariate normal distribution with correlation ρ. If now a transformation, $x' = x'(x)$, $y' = y'(y)$, is made to any new variables x' and y' such that*

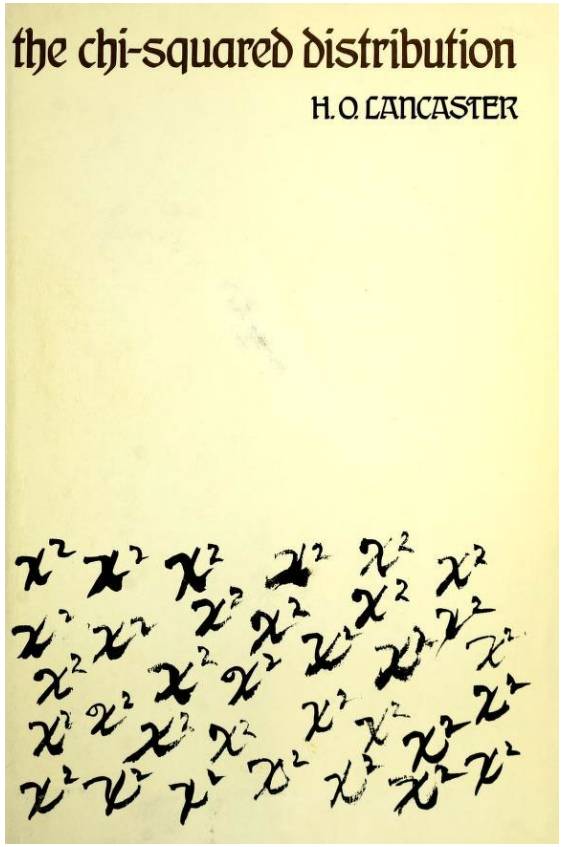
$$(2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} x'^2 \exp -\tfrac{1}{2}x^2 dx \quad \textit{and} \quad (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} y'^2 \exp -\tfrac{1}{2}y^2 dy$$

are finite, then the correlation of the new variables is less in absolute value than ρ. That is, ρ is the maximum canonical correlation.

- Kendall & Stuart , 1961:

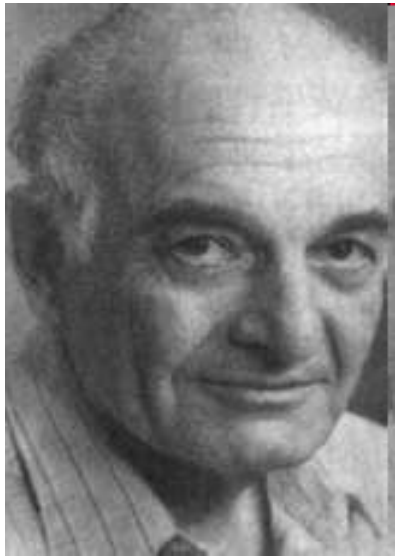
The theoretical implication of the result is clear: if we seek separate scoring systems for the two categorized variables such as to maximize their correlation, we are basically trying to produce a bivariate normal distribution by operation upon the margins of the table.

Henry Oliver Lancaster (1913-2001)



1969, John Wiley

- Guttman, inventor of MCA as the optimal simultaneous coding of *p* categorical variables



Louis Guttman (1916-1987)

Professor of Sociology at Cornell University, Founder of the Israel Institute of Applied Social Research, renamed *Guttman Center of Public Opinion and Policy Research until 2021*

Known for “Guttman effect” or “horseshoe effect”

Supplementary Study B-3

The Quantification of a Class of Attributes: A Theory and Method of Scale Construction

Thus, we are given the responses of a population of U individuals to a set of m items which have a common content that is desired to be thought as a single class of behavior. These responses can be represented by check marks as in the following table (with hypothetical entries):

SUBCATEGORY	INDIVIDUAL					
	1	2	3	4	...	U
A_1	✓		✓		...	
A_2		✓			...	✓
A_3				✓	...	
B_1			✓		...	✓
B_2	✓	✓		✓	...	
...	...	...	...	...	...	...
Z_1		✓			...	✓
Z_2	✓		✓		...	
Z_3				✓	...	
Z_4					...	

in P. Horst et al., *The Prediction of Personal Adjustment*, SSRC, 1941, pp. 321-348

- **Principle of internal consistency** :

  - Assign each modality a score such that the ξ_i variables thus created are as homogeneous as possible (minimize the within variance) and the average score as dispersed as possible (minimize the between variance)
  - The optimal scores are the coordinates of the categories on the first axis of Multiple Correspondence Analysis (MCA).
  - Multiple codings if we consider the following principal axes
  - At the origins of Dutch school (HOMALS)

- Similar to Hayashi's quantification method n°III (1950)

(1918-2002)

4. The 70s and the domination of Alternating Least Squares (ALS)

- 145 articles on optimal scaling published in Psychometrika between 1968 and 1982
- Supervised and unsupervised
- ALS approach
  - Split parameters into two groups: model and coding
  - Optimize those of one group, knowing those of the other
  - Alternate until convergence (local optimum)

ALSOS Programs

Program	Analysis	Data	Source	Primary Reference
ADDALS	Additivity analysis (analysis of variance)	Two or three way tables. Nonorthogonal and incomplete designs permitted.	UNC	de Leeuw, Young & Takane (1976)
WADDALS	Weighted additivity analysis	Same as ADDALS	UNC	Takane, Young & de Leeuw (1980)
MANOVALS	Multivariate analysis of variance	Multi-way tables	RUL	Gifi (1981)
MORALS CORALS & CANALS	Multiple and canonical analysis	Mixed measurement level multivariate data	UNC or RUL	Young, de Leeuw & Takane (1976)
OVERALS	Canonical analysis	Multiple set mixed measurement level multivariate data	RUL	Gifi (1981)
CRIMINALS	Multiple group discriminant analysis	Mixed measurement level predictors	RUL	Gifi (1981)
PATHALS	Path analysis	Mixed measurement level multivariate data	RUL	Gifi (1981)
PRINCALS & PRINCIPALS	Principal components analysis	Mixed measurement level multivariate data	UNC or RUL	Young, Takane & de Leeuw (1978)
HOMALS	Principal components analysis	Multivariate nominal data	RUL	de Leeuw & van Rijkevorsel (1976)
ALSCOMP & TUCKALS	Three-mode factor analysis	Mixed measurement level multivariate data	UNC or RUL	Sands & Young 1978 de Leeuw & van Rijkevorsel (1976)
FACTALS	Common-factor analysis	Mixed measurement level multivariate data	UNC	Takane, Young & de Leeuw (1978)
ALSCAL	Two or three-way multidimensional scaling	Similarity data	UNC	Takane, Young & de Leeuw (1977)
GEMSCAL	Two or three-way multidimensional scaling	Similarity data	UNC	Young, Null & De Soete (Note 5)

Young, 1981

- PRINQUAL (SAS), PRINCALS, PRINCIPALS…

$$\max_{\substack{\varphi_1,\varphi_2,\ldots,\varphi_p \\ C_1,..C_K}} \sum_{p=1}^{P}\sum_{k=1}^{K} r^2\left(\varphi_p\left(X_p\right), C_k\right)$$

  - Equivalent to maximizing the sum of the first *K* eigenvalues of the correlation matrix
  - Fundamental difference with MCA: looking for unique encodings, whereas MCA provides different encodings for each dimension..

- Multiple regression : MORALS, TRANSREG (SAS)

$$\max_{\psi,\varphi_1,\varphi_2,..,\varphi_P} R^2\left(\psi(Y);\varphi_1\left(X_1\right),\varphi_2\left(X_2\right),..,\varphi_P\left(X_P\right)\right)$$

  - Alternating regression and optimal variable-by-variable coding
  - With all categorical predictors and no tranformation of Y, equivalent to regress $\mathbf{y}$ on the complete disjunctive table of predictors

$$\mathbf{X} = \left[\mathbf{X}_1 \,|\ldots|\, \mathbf{X}_p \,|\ldots|\, \mathbf{X}_P\right]$$

  - Rank problems solved by centering constraints

- Example : risk scoring in automobile insurance using DISQUAL methodology (Bouroche, Saporta, Tenenhaus, 1977 and Saporta, Niang, 2006):

  - 1106 automobile insurees from Belgium observed in 1992 belonging to 2 groups.

    Those without claim n_1=556 (the “good” ones).

    Those with more than one claim (the “bad” ones) n_2=550.

  - 9 categorical predictors with a total of 20 categories

    1. Use type (2): professional, private
    2. Insuree type (3): male, female, companies
    3. Language (2): French, Flemish
    4. Birth cohort (3): 1890–1949, 1950–1973, unknown
    5. Region (2): Brussels, other regions
    6. Level of bonus-malus (2): B-M+, other B-M (–1)
    7. Horsepower (2): 10–39, 40–349
    8. Year of subscription (2): <86 , others
    9. Year of vehicle construction (2): 1933–1989, 1990–1991

Fisher's LDA using MCA components

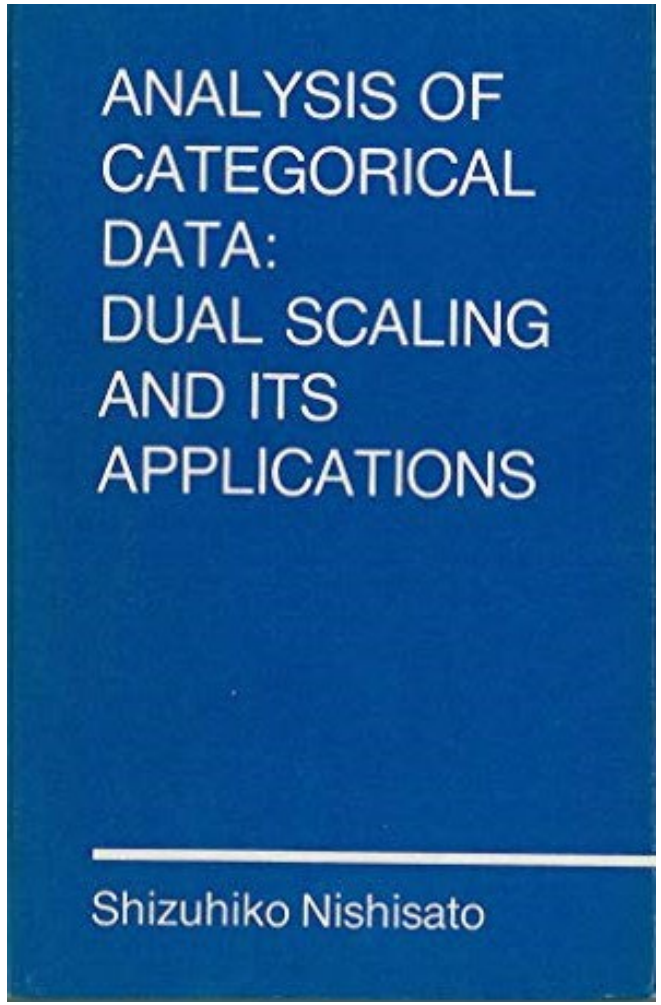
COMPONENTS	CORRELATIONS	COEFFICIENTS
1 F 1	0.719	6.9064
2 F 2	0.055	0.7149
3 F 3	-0.078	-0.8211
4 F 4	-0.030	-0.4615
5 F 5	0.083	1.2581
6 F 6	0.064	1.0274
7 F 7	-0.001	0.2169
8 F 8	0.090	1.3133
9 F 9	-0.074	-1.1383
10 F 10	-0.150	-3.3193
11 F 11	-0.056	-1.4830

Global Score= 6.90 F1 - 0.82 F3 + 1.25 F5 + 1.31 F8 - 1.13 F9 - 3.31 F10

Then back to the indicators, since each component is a linear combination of them.

Scorecard

CATEGORIES	COEFFICIENTS DISCRIMINANT FUNCTION	TRANSFORMED COEFFICIENTS PARTIAL SCORE)
2 . Use type		
USE1 - Profess.	-4.577	0.00
USE2 - private	0.919	53.93
4 . Gender		
MALE - male	0.220	24.10
FEMA - female	-0.065	21.30
OTHE - companies	-2.236	0.00
5 . Language		
FREN - French	-0.955	0.00
FLEM - flemish	2.789	36.73
24 . Birth date		
BD1 - 1890-1949 BD	0.285	116.78
BD2 - 1950-1973 BD	-11.616	0.00
BD? - ???BD	7.064	183.30
25 . Region		
REG1 - Brussels	-6.785	0.00
REG2 - Other regions	3.369	99.64
26 . Level of bonus-malus		
BM01 - B-M 1 (-1)	17.522	341.41
BM02 - Others B-M (-1)	-17.271	0.00
27 . Duration of contract		
C<86 - <86 contracts	2.209	50.27
C>87 - others contracts	-2.913	0.00
28 . Horsepower		
HP1 - 10-39 HP	6.211	75.83
HP2 - >40 HP	-1.516	0.00
29 . year of vehicle construction		
YVC1 - 1933-1989 YVC	3.515	134.80
YVC2 - 1990-1991 YVC	-10.222	0.00


ANALYSIS OF
CATEGORICAL
DATA:
DUAL SCALING
AND ITS
APPLICATIONS
Shizuhiko Nishisato

1980

More recently

- Optimal coding and clustering

  - GROUPALS (van Buuren & de Leeuw, 1989)

  Alternating K-means and codings resulting from the crossing between the partition and each variable.

  - CLUSTER-CA (van de Velden, D'Enza, & Palumbo, 2017)

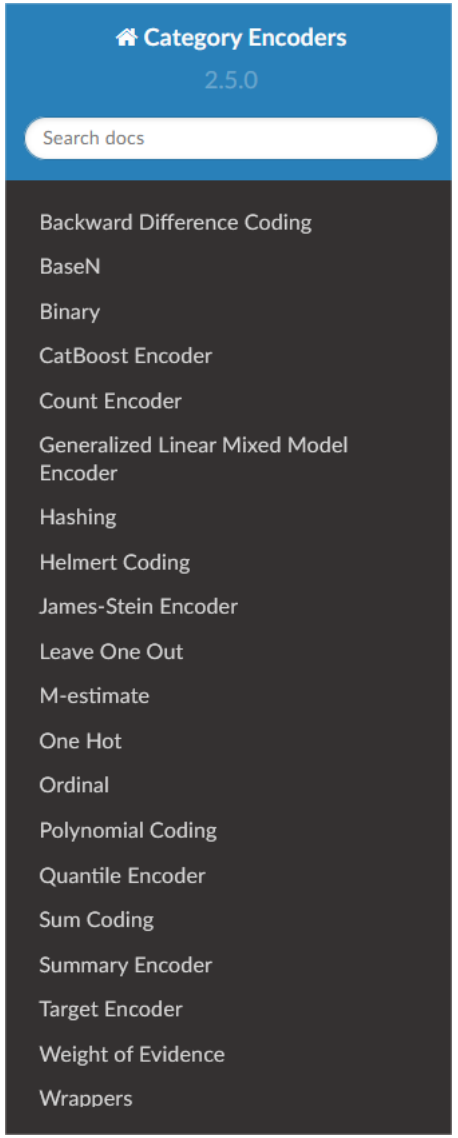
  Equivalent method with correspondence analysis of the table concatenating the cross-classification tables between the partition and each variable (Burt subtable).

$$\mathbf{Y}'\mathbf{X} = \left[\mathbf{Y}'\mathbf{X}_1 | \ldots | \ \mathbf{Y}'\mathbf{X}_p | \ldots | \ \mathbf{Y}'\mathbf{X}_P\right]$$

- Non metric PLS including path modelling (Russolillo, 2012)

5. Categorical encoding and Machine Learning

- A profusion of methods stemming from the need to process variables with a large number of categories (zip code, etc.)
- Reinventing classical methods
  - Target encoder : conditional mean coding
  - One Hot Encoding: indicator matrix
- Others more or less arbitrary
  - Label encoder, ordinal encoder, hash encoder



https://contrib.scikit-learn.org/category_encoders/index.html

- A typology of encodings for supervised methods
  - Methods using only the response variable Y

    Target encoder, Leave One Out, CatBoost
  - Methods using Y and X_p

    Morals
  - Methods that don't use data

    Label encoding, Hash encoding

    Specific issues : when new categories not observed during the learning process (label encoding) appear

    “Collision” when several values can be represented by the same hash value. Does not take into account periodicities such as the day of the week.

    Neither interpretation nor optimality for any model

- Risk of overfitting with disjunctive coding (One Hot Encoding) when the number of categories is high

Regularize!

1. Ridge, lasso, elastic-net for supervised problems


Statistical Science
2019, Vol. 34, No. 3, 361–390
https://doi.org/10.1214/19-STS697

ROS Regression: Integrating Regularization with Optimal Scaling Regression

Jacqueline J. Meulman, Anita J. van der Kooij and Kevin L. W. Duisters

- Linear setup $\varphi_p\left(X_p\right) = \mathbf{X}_p\mathbf{b}_p$
- Elastic net: $\min_{T,b_p}\left\{\left\|T(\mathbf{y}) - \sum_{p=1}^{P}\mathbf{X}_p\mathbf{b}_p\right\|^2 + \lambda_1\sum_{p=1}^{P}\sum_{j=1}^{J_p}\left|b_{pj}\right| + \lambda_2\sum_{p=1}^{P}\sum_{j=1}^{J_p}\left(b_{pj}\right)^2\right\}$
- The advantage of the lasso in cancelling coefficients can become a disadvantage for categorical variables.
- Regularization by projection onto subspaces :
  - PLS, principal component regression
  - DISQUAL (versus logistic regression)
- Systematic use of the triptych learning, test, validation

2. Draw inspiration from word embedding (or vectorization)

- Representing a variable with a large number of categories in a low-dimensional space.
  - 101 dimensions to represent French departments? **Two GPS coordinates are enough!**

- Identify the dimensions needed to correctly represent the similarity between variable categories (category embedding).

5.1 Neural regression with encoding (Di Ciaccio, 2023)

- A model with *S* fully connected neurons and a non-linear activation function *σ* with *One Hot Encoding*

$$\hat{\mathbf{y}} = \beta_0 + \sum_{s=1}^{S} \beta_s \sigma\left(\sum_{p=1}^{P} \mathbf{X}_p \mathbf{w}_{ps} + w_{0s} \right)$$

- Far fewer parameters with low-embedding encoding (dimension L)

$$\hat{\mathbf{y}} = \beta_0 + \sum_{s=1}^{S} \beta_s \sigma\left(\sum_{p=1}^{P} \sum_{l=1}^{L} \mathbf{X}_p \mathbf{a}_{pl} w_{pls} + w_{0s} \right)$$

and better predictions

5.2 An unsupervised auto-encoder approach

- HOMALS criterium (similar to MCA) for a L-dimensional representation :

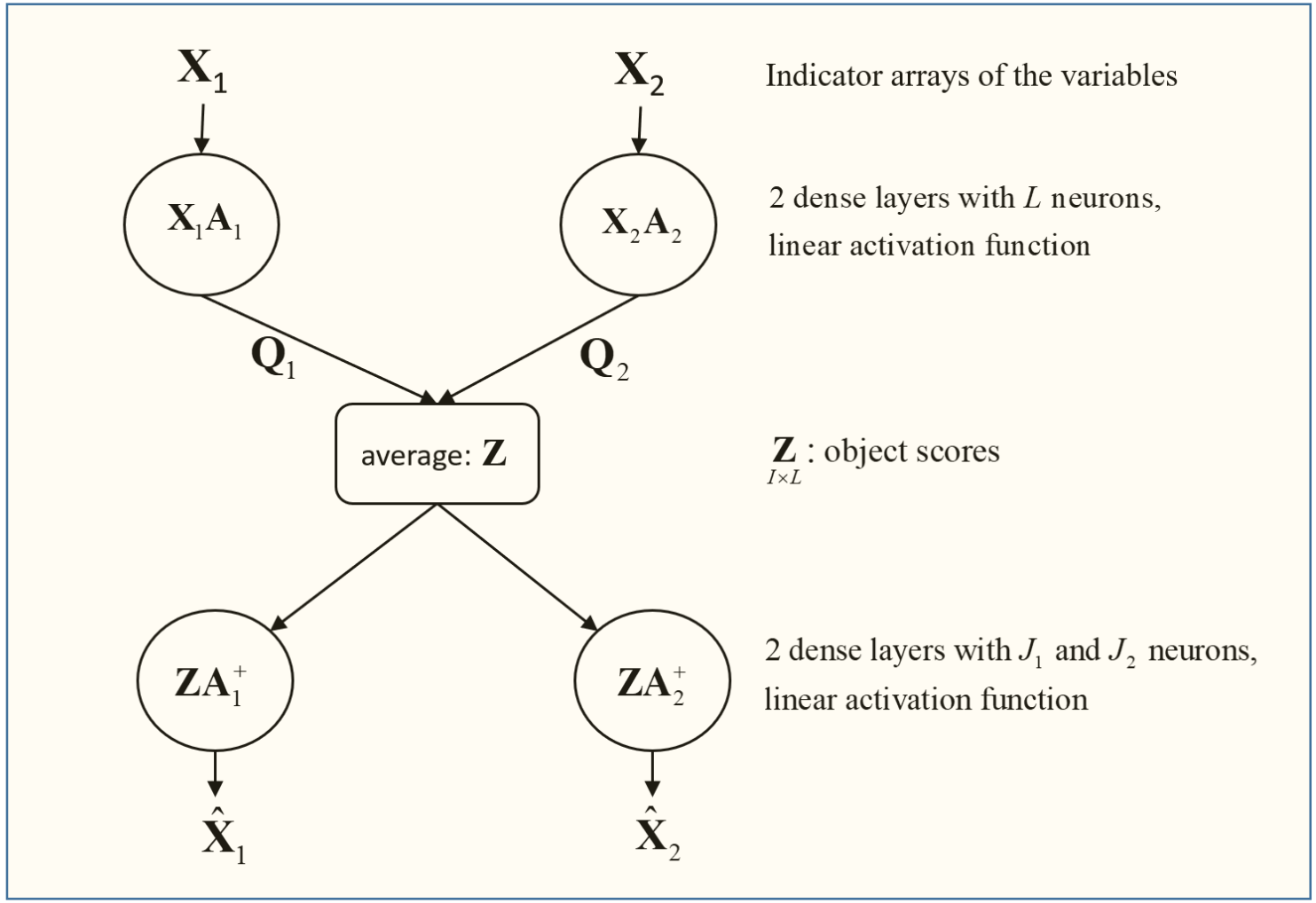
$$\min_{\substack{\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_P \\ \mathbf{Z}}} \left(\sum_{p=1}^{P} \left\| \mathbf{Z} - \mathbf{X}_p \mathbf{A}_p \right\|^2 \right) = \sum_{p=1}^{P} \left\| \mathbf{X}_p - \hat{\mathbf{X}}_p \right\|^2 = \sum_{i=1}^{I} \sum_{p=1}^{P} \sum_{j=1}^{J_p} \left(x_{ipj} - \hat{x}_{ipj} \right)^2$$

$\mathbf{Z}$ IxL matrix of units scores,

$\mathbf{A}_p$ J_pxL matrix of multiple quantifications of $\mathbf{X}_p$

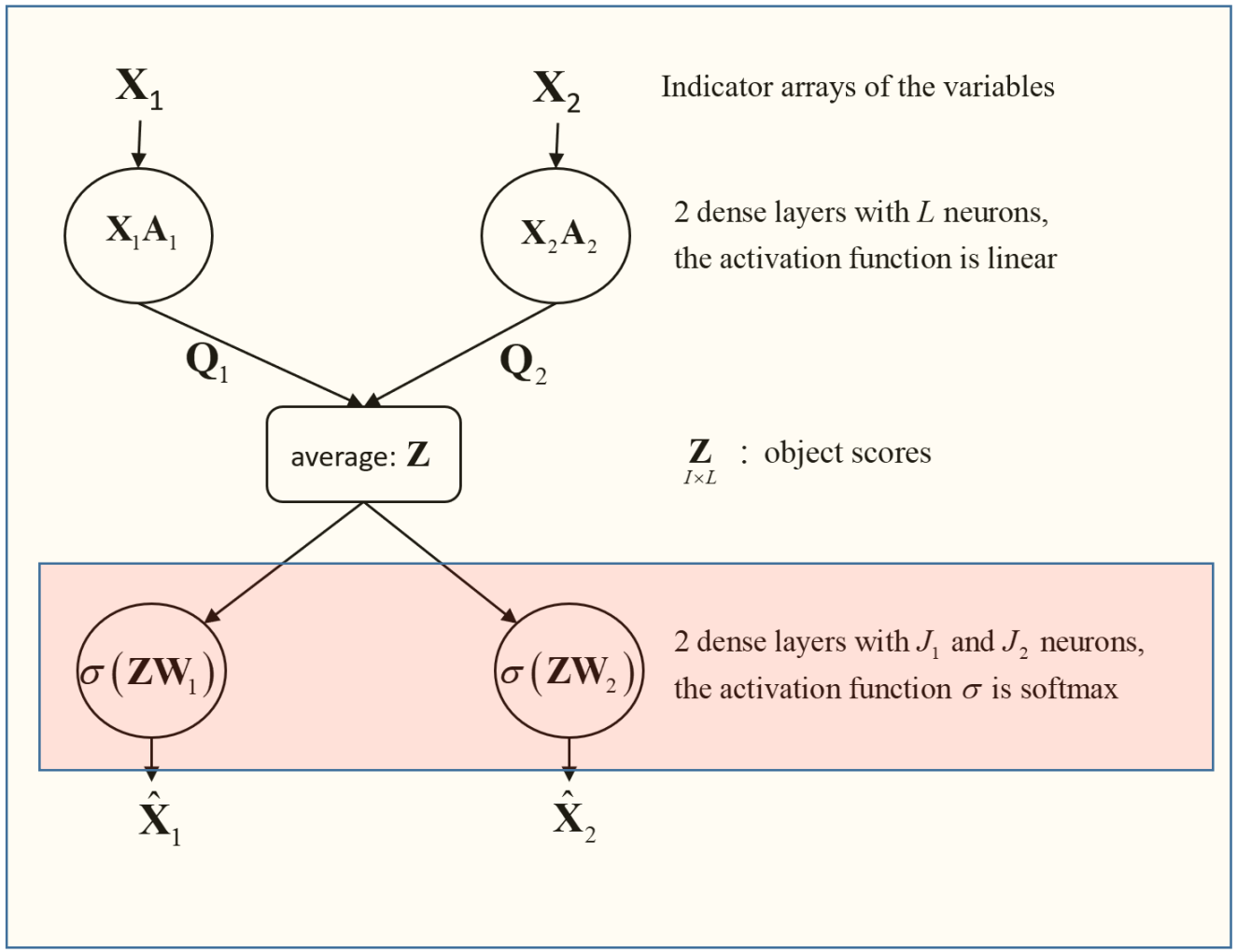
- Achievable with an autoencoder where category embeddings are the centroids of units sharing the same category.

Two categorical variables



- $\mathbf{A}_1$, $\mathbf{A}_2$ quantification of categories in L dimensions.
- NB: Algorithm far less efficient than MCA or HOMALS

A non-linear variant (Di Ciaccio, 2023)



- Softmax function

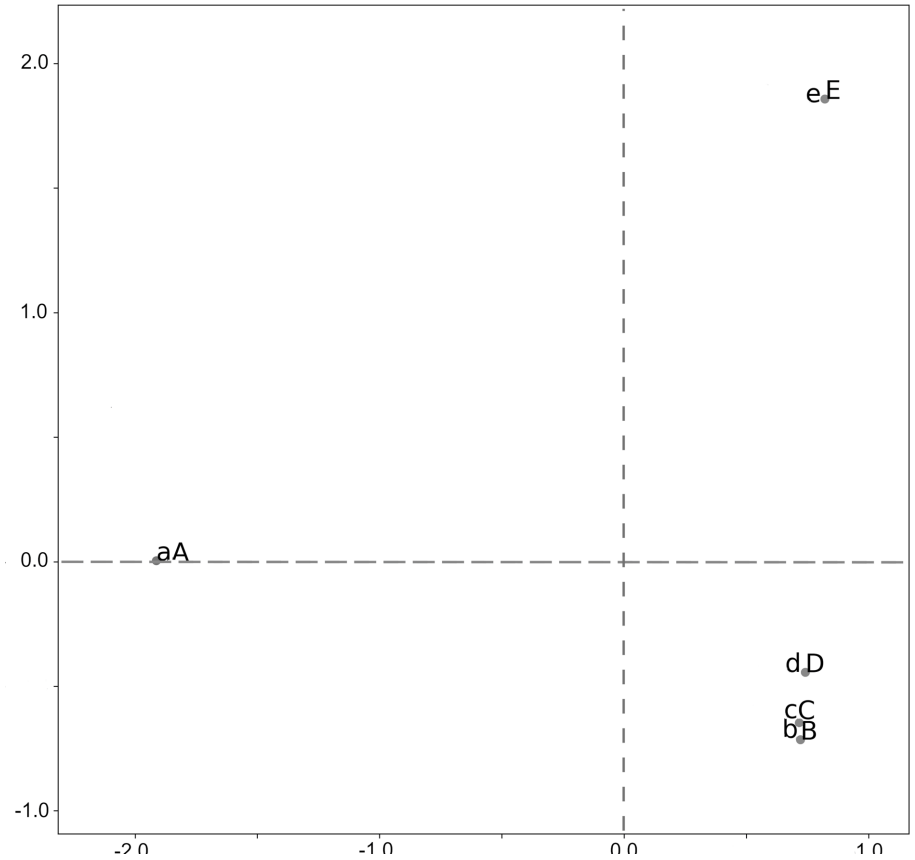
$$\sigma(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^{J}\exp(z_j)}$$

- Minimize cross-entropy instead of squared deviations

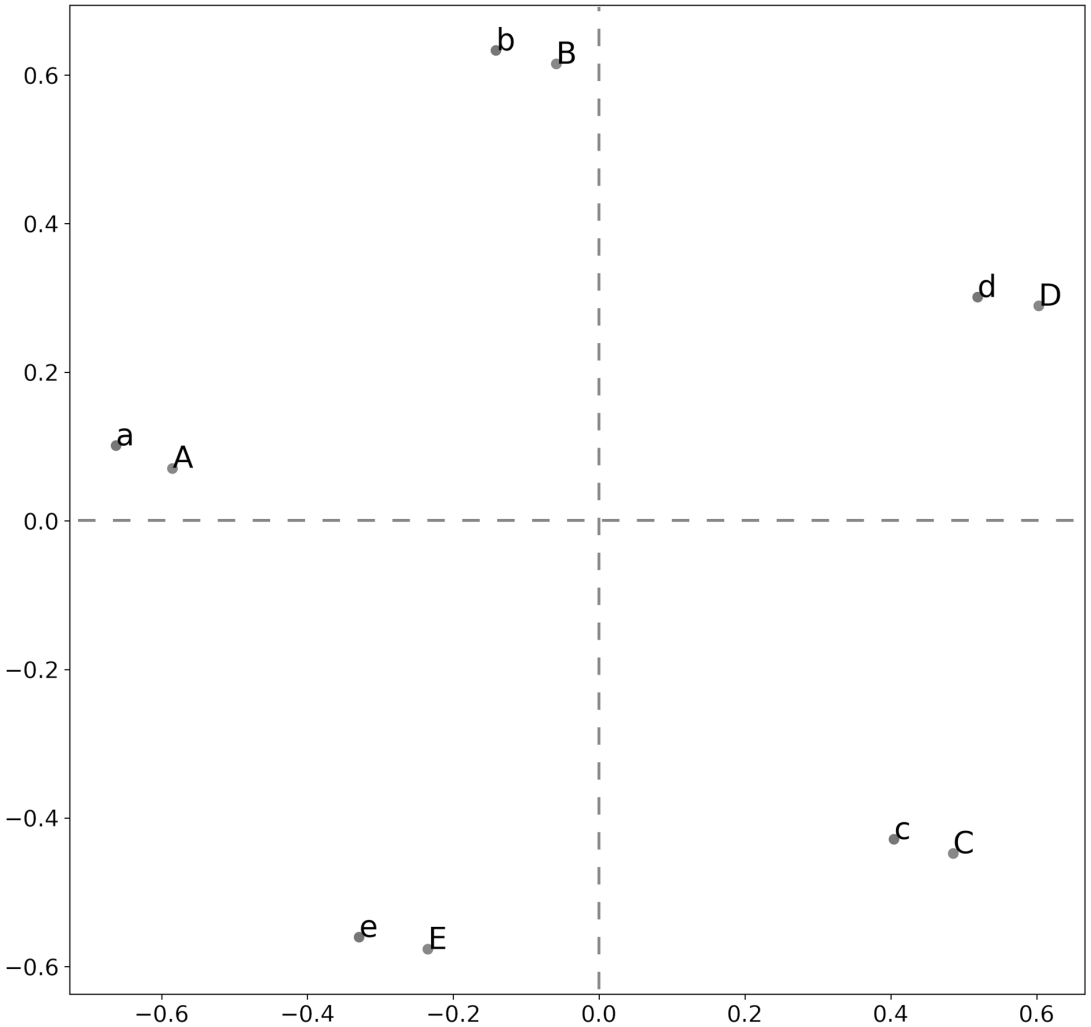
$$-\sum_{i=1}^{I}\sum_{p=1}^{P}\sum_{j=1}^{J_p} x_{ipj}\log\hat{x}_{ipj}$$

Example

X/Y	a	b	c	d	e	Total
A	801	100	100	100	100	1201
B	100	800	100	100	100	1200
C	100	100	800	100	100	1200
D	100	100	100	800	100	1200
E	100	100	100	100	800	1200
Total	1201	1200	1200	1200	1200	6001



MCA equivalent here to CA provides 4 identical eigenvalues and a poorly balanced representation



Non-linear version and cross-entropy criterion

6. Conclusion and perspectives

- Optimal coding: not just a trick for applying numerical methods to categorical variables.
- Letters of nobility for almost a century, with strong links to correspondence analysis.
- Renewal with high-dimensional data processing :
  - Regularization
  - Dialogue to be developed between statistics and Machine Learning
  - New criteria

References

- Bouroche, J.M., Saporta, G. & Tenenhaus, M. (1977). Some methods of qualitative data analysis. In J.R. Barra, (Ed.) *Recent Developments in Statistics,* North-Holland, pp.749–755. https://hal-cnam.archives-ouvertes.fr/hal-03059983
- Di Ciaccio A. (2023). Optimal Coding of categorical data in machine learning. In: Grilli, M., Lupparelli, M., Rampichini, C., Rocco, E., Vichi, M. (eds), *Statistical Models and Methods for Data Science. CLADAG 2021*. Springer
- Fisher, R.A. (1940). The precision of discriminant functions. *Annals of Eugenics, 10*, 422–429.
- Gifi, A. (1990). *Nonlinear Multivariate Analysis*. John Wiley & Sons.
- Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society* 31, 520–524.
- Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29, 115–129.
- Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika, 44*, 289–292.

- Meulman, J. J., van der Kooij, A. J., & Duisters, K. L. (2019). ROS Regression: Integrating Regularization with Optimal Scaling Regression. *Statistical science*, *34*(3), 361-390.
- Nishisato, S. (1980). *Analysis of Categorical Data: Dual Scaling and its Applications*. University of Toronto Press
- Russolillo, G. (2012). Non-Metric Partial Least Squares. *Electronic Journal of Statistics*, *6*, 1641-1669.
- Saporta,G., Niang- Keita, N. (2006). Correspondence analysis and classification. In Greenacre, M. & Blasius, J. *Multiple Correspondence Analysis and Related Methods*, Chapman and Hall/CRC, 371-392
- Tenenhaus, M. (1988). Canonical analysis of two convex polyhedral cones and applications. *Psychometrika, 53*, 503–524.
- van Buuren, S., & Heiser, W. J. (1989). Clustering N objects into K groups under optimal scaling of variables. *Psychometrika*, 54, 699-706.
- van de Velden, M., D'Enza, A. I., & Palumbo, F. (2017). Cluster correspondence analysis. *Psychometrika*, 82, 158-185.
- Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46*, 357–388.