# Old and New Perspectives on Optimal Scaling

Hervé Abdi, Agostino Di Ciaccio, Gilbert Saporta

# Old and New Perspectives on Optimal Scaling

Hervé Abdi, Agostino Di Ciaccio, & Gilbert Saporta

**Abstract** Processing in machine learning qualitative variables having a very large number of modalities is an opportunity to revisit the theory of optimal scaling and its applications. This revisitation starts with the pioneers of scaling in statistics, psychometrics, and psychology before moving on to more contemporary treatments of scaling that fall within the realm of machine learning and neural networks.

## 1 Introduction

Qualitative variables are ubiquitous in many fields but genetic and human sciences (especially psychology) have been some of the first disciplines to routinely incorporate qualitative variables in their practice. This importance of qualitative variables prompted Stevens (a psychologist) to create in 1946 the now classic typology of measurement scales. In this typology, qualitative (also called categorical) variables come in two varieties:

Hervé Abdi
The University of Texas at Dallas, e-mail: herve@utdallas.edu.
Orcid ID: 0000-0002-9522-1978.

Agostino Di Ciaccio
University of Rome La Sapienza, Rome, Italy, e-mail: agostino.diciaccio@uniroma1.it.
Orcid ID: 0000-0001-7998-1195.

Gilbert Saporta
CNAM, Paris, e-mail: gilbert.saporta@cnam.fr.
Orcid ID: 0000-0002-3406-5887.

In Beh, E.J., Lombardo, R., & Clavel, J.G. (Eds.), *Analysis of Categorical Data from Historical Perspectives*. 2023. Springer: Singapore.
https://doi.org/10.1007/978-981-99-5329-5_9

- Nominal variables, so called because the modalities—also named levels or categories—of a nominal variable are "names." Formally, a nominal variable corresponds to a partition of a set.
- Ordinal variables (a nominal variable whose modalities are ordered); formally, an ordinal variable corresponds to a pre-order on a set.

Because most multivariate statistical methods are designed for quantitative variables (in Stevens's typology: interval and ratio scales), an obvious problem is to *optimally* transform a qualitative variable into a quantitative variable. This problem being relevant for several disciplines, similar procedures to solve it were independently developed multiple times and therefore come under different names with *scaling, quantification, coding*, and *encoding* being favourites. So, a nominal or ordinal variable is *quantified*, *(en)coded*, or *scaled* when its modalities are replaced by numbers having at least the properties of an interval scale.

Note that the terms *coding*, and *encoding* are ambiguous because they can refer either to the transformation of a qualitative variable into a numerical variable (quantification) or to a way of representing a qualitative variable such as, for example, disjunctive coding.

The problem of transforming qualitative variables into quantitative variables has a long history. In statistics, its history goes back to the early contributions of major figures such as Hirschfeld (1935), Horst (1935), who coined the named "reciprocal averaging," Fisher (1940), and Hayashi (1950). In psychology (and of course psychometrics) early contributions of other major figures include Guttman (1941, 1944), Festinger (1947), and even Coombs in his classic work *a Theory of Data* (1964, see also, Coombs, 1948). The statisticians were mostly interested in maximising the (squared) correlation between sets of variables; but the psychologists (influenced by factor analytic models) were concerned about *scaling* (i.e., estimating a *quantitative* latent variable or factor from qualitative measurements). The maximisation approach of the statisticians would lead to (*simple*) correspondence analysis whereas the factorial approach of the psychologists would lead to *multiple* correspondence analysis (see, for details, the historical review of Lebart & Saporta, 2014).

This early work matured in the 1970s and early 1980, which were the years of the search for optimal codes (called factor scores or scaling scores) in supervised or unsupervised contexts, an endeavour where researchers such as de Leeuw (1973), Nishisato (1980), Takane (1980), Tenenhaus (1988), and Young (1976, 1978, 1981, see also Tenenhaud and Young, 1981) distinguished themselves. This research was then implemented by commercial software with procedures such as PRINQUAL and TRANSREG for SAS, or CATEGORIES for SPSS.

In the next 30 years or so, after this first foray in the theory of optimal scaling, the topic did not generate much research: routine applications involved computing predictive scores, such as risk scores in banking and insurance. However, recent interest in the scaling problem was reignited by the availability of massive data sets. Nowadays, machine learning researchers and practitioners need to handle categorical data (which are ill-suited for most machine learning algorithms such as neural networks) that often have large numbers of modalities (e.g., from dozens or even

hundreds of modalities, such as postal codes; for details, see, e.g., Hancock & Khoshgoftaar, 2020).

This new interest in qualitative data stimulated the development of several coding methods—mostly developed in the ignorance of the early work of statisticians and psychometricians. As an illustration of this trend, Di Ciaccio (2023) recently reported that the popular Python package `scikit-learn` offers seventeen different methods that he categorised into three groups:

- methods where the encoding of a variable does not depend on the other variables, in particular the response (e.g., hash encoding),
- methods where the encoding only depends on the response (e.g., conditional mean), and
- *One-Hot Encoding* (OHE), which is nothing more than the usual disjunctive representation with as many indicators as modalities (see Equation **??**).

The large size of certain categorical data sets raises problems of stability and overfitting—problems that were neglected in classical statistical applications where the number of modalities was typically small and the learning-testing methodology rarely used. Because of their different view points, the confrontation of the early approach of the statisticians and psychometricians with the newer approach from data scientists could foster a renewal of coding methods for qualitative data (for details, see Meulman et al., 2019).

The rest of the chapter is organized as follows: Sections 2 and 3 are devoted to notations and to the mathematical structures of quantifications. Section 4 describes early works from 1935 till the 1960s. Section 5 is devoted to the "golden seventies" dominated by optimal scaling (performed with alternating least squares) and Nishisato's dual scaling. Section 6 describes how machine learning has taken over the problem of encoding, with its connection to multivariate statistics and how this can foster a re-interpretation of correspondence analysis from a non-linear point of view.

## 2 Matrix representation of categorical encoding and notations

When dealing with $I$ observations it is often practical to represent a nominal variable as a binary *group* matrix (called a complete disjunctive coding matrix) denoted by $\mathbf{X}$ whose rows are observations and whose columns represent the modalities of the nominal variable[1].

For example, consider a sample with $I = 5$ observations, denoted $\{S_1, \ldots, S_5\}$, and a nominal scale with $J = 3$ modalities: $\{1, 2, 3\}$ that could be, for example, {disagree, neutral, agree}, with the following answers for these five observations

$$X = [1,\ 2,\ 3,\ 1,\ 2]^{\mathrm{T}}, \tag{1}$$

[1] As noted above, and developed later on, this is a procedure rediscovered in machine learning under the name of *one hot encoding*.

then the group matrix would be equal to

$$\mathbf{X} = \begin{matrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{matrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} \mathbb{1}_1, & \mathbb{1}_2, & \mathbb{1}_3 \end{bmatrix} . \tag{2}$$

where, for example, $\mathbb{1}_1 = [1, 0, 0, 1, 0]^{\mathrm{T}}$ is the indicator variable for the first category.

In this chapter, the following notations are used:

- $I$ is the number of units/observations $\{1, 2, \ldots, i, \ldots, I\}$,
- $X$ is a nominal variable, namely a sequence of $I$ modalities,
- $\mathbf{x}$ is a quantification of $X$ (i.e., a real vector of length $I$),
- $K$ is the number of nominal variables,
- $J$ is the number of modalities, of a variable, $\{1, 2, \ldots, j, \ldots, J\}$,
- $J_k$ is the number of modalities of the $k$th variable (when $K > 1$),
- $\mathbf{X}$ is the disjunctive matrix, (of dimensions $I \times J$) for variable $X$,
- $L$ is the dimension of a vector space $\{1, 2, \ldots, \ell, \ldots, L\}$,
- $\mathbf{a}_k$ is the single category quantification of variable $k$ (i.e., a real vector of lenght $J_k$),
- $\mathbf{A}_k$ is the category quantification array on $L$ dimensions (of dimensions $J_k \times L$),
- $\mathbf{q}_k$ is the vector of a single quantified variable $k$, (a real vector of Lenght $I$),
- $\mathbf{Q}_k$ is the quantified array of variable $\mathbf{X}_k$ (of dimensions $I \times L$) for $L$ dimensions.

## 3 The structure of quantifications

Quantifying or encoding a categorical variable can be written using simple transformations that we explicitly define in the following sections.

### 3.1 Categorical encoding

Let $X$ be a nominal variable with $J$ unordered modalities $\{1, \ldots, j, \ldots J\}$ and $\mathbf{x}$ a quantification of $X$ using at most $J$ distinct values $\{a_1, \ldots, a_j, \ldots a_J\}$. Then, if $\mathbb{1}_j$ denotes the indicator variable of the $j$th category, we have:

$$\mathbf{x} = \sum_{j=1}^{J} a_j \mathbb{1}_j . \tag{3}$$

Quantifying $X$ boils down to defining a linear combination (with the weights $a_j$ being called the code or scale values) of the indicator variables. When there is no

constraint on the $a_j$ weights, the set of possible quantifications $\mathbf{x}$ is a vector subspace $\mathcal{W}$ with dimension $J$.

Because

$$\sum_{j=1}^{J} \mathbb{1}_j = \mathbf{1}, \tag{4}$$

(with $\mathbf{1}$ begin a commensurable vector of 1s) the set $\Delta$ of constant variables (which is a one-dimensional subspace) is included into $\mathcal{W}$. If $\mathbf{x}$ is required to have zero mean, then

$$\mathbf{x} \in \left\{\Delta^{\perp} \cap \mathcal{W}\right\} \ . \tag{5}$$

Note that the encoding from Equation **??** is redundant because the value of any $\mathbb{1}_j$ variable can be deduced from the values of the other $(J-1)$ variables. Another possibility could be to use only $J-1$ indicator variables as done, for example, with the dummy coding scheme used in the general linear model and logistic regression. We will not use this coding scheme here so that all modalities play the same role.

## 3.2 Ordinal encoding

If there is a natural order between the modalities (i.e., a pre-order on the set of responses) it is natural to require that

$$a_1 \leq a_2 \leq \cdots \leq a_J.$$

Let us consider the following reparametrisation:

$$a_1 = b_1,\ a_2 = b_1 + b_2, \ldots, a_J = b_1 + \cdots + b_j + \cdots + b_J \text{ with } \begin{cases} b_1 \in \mathbb{R} \\ b_2, \ldots, b_J \geq 0; \end{cases} \tag{6}$$

then

$$\begin{aligned} \mathbf{x} &= \sum_{j=1}^{J} a_j \mathbb{1}_j \\ &= b_1 \mathbb{1}_1 + (b_1 + b_2)\mathbb{1}_2 + \cdots + (b_1 + b_2, \ldots)\mathbb{1}_J \\ &= b_1(\mathbb{1}_1 + \mathbb{1}_2 + \cdots + \mathbb{1}_J) + b_2(\mathbb{1}_2 + \cdots + \mathbb{1}_J) + \cdots + b_J \mathbb{1}_J \\ &= b_1 + b_2(\mathbb{1}_2 + \cdots + \mathbb{1}_J) + \cdots + b_J \mathbb{1}_J \\ &= b_1 + \sum_{j=2}^{J} b_j \mathbb{z}_j \end{aligned} \tag{7}$$

where

$$\mathbb{z}_j = \sum_{\ell=j}^{J} \mathbb{1}_\ell . \tag{8}$$

The variable $\mathbf{x}$ is thus a linear combination of $J-1$ variables with non negative coefficients, which is the definition of a convex polyhedral cone (see, e.g., Tenenhaus, 1988), plus one unconstrained constant term. In other words, $\mathbf{x}$ belongs to the direct sum of $\Delta$ and an $(J-1)$ convex polyhedral cone $C_{J-1}$, and so:

$$\mathbf{x} \in \{\Delta \oplus C_{J-1}\} \ . \tag{9}$$

Note: if we also require that $\mathbf{x}$ has zero mean, the constant $b_1$ will be negative.

### 3.3 Two simple optimal scaling problems

Let $Y$ be a numerical response variable. What is the optimal way to quantify a qualitative variable $X$ in order to best predict $Y$ in the least-squares sense?

If $X$ is categorical, the solution[2] is given by the projection of $Y$ onto the subspace $\mathcal{W}$ spanned by the set of the indicator variables $\mathbb{1}_j$. In other words, the optimal solution is obtained by performing a multiple regression without intercept of $Y$ onto the set of the $\mathbb{1}_j$. Because the $\mathbb{1}_j$ are orthogonal, the solution is easily found: The optimal scores $a_j$ are the conditional means for each modality $\bar{y}_j$.

If $X$ is ordinal, the solution is less straightforward because we have to project $Y$ onto a polyhedral cone instead of a vector subspace. However, because the cone is convex (cf. Equation **??**), the solution is unique and boils down to computing a multiple regression

$$\widehat{Y} = b_1 + \sum_{j=2}^{J} b_j \mathbb{z}_j \tag{10}$$

with positivity constraints for the $b_j$ coefficients (for $j > 1$, cf. Equation **??**). The solution of this constrained optimisation problem can be found using some efficient numerical methods such as the *pool adjacent violators algorithm* (see, e.g., Kruskal, 1964; Tenenhaus, 1988; de Leeuw et al., 2009).

### 3.4 Crisp coding, fuzzy coding, spline coding

Transforming a numerical variable into a qualitative variable by splitting it into classes, and then recoding this variable according to the previously mentioned principles, is a low cost way of non-linearly transforming a numerical variable.

Coding with Equation **??**—called here *crisp-coding*—has the disadvantage of introducing discontinuities that can loose some information from the original vari-

[2] Called *target encoding* in machine learning.

able. To alleviate this problem, various kinds of fuzzy encodings can be used—a procedure equivalent to defining membership functions for neighbouring intervals. Crisp-coding and piecewise-linear encoding (which is a fuzzy coding) are particular cases of linear combinations of spline functions as illustrated in Figure **??** that shows examples of splines of, respectively, Degrees 0 and 1 associated to (discontinuous) crisp-coding and piecewise continuous linear transformations.

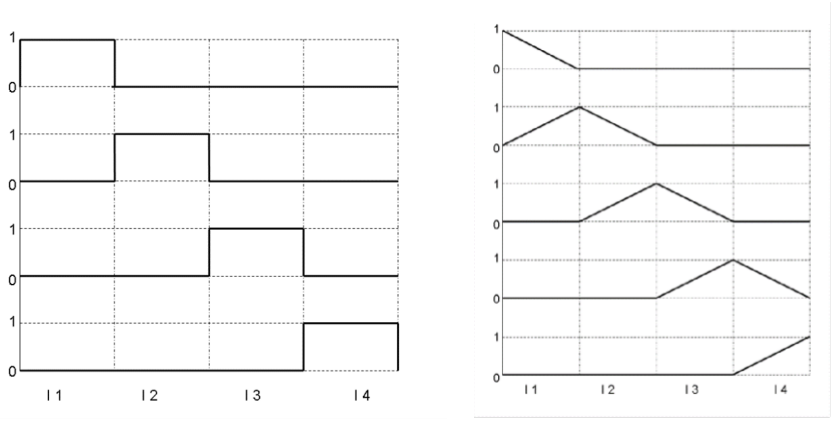


**Fig. 1** Basis spline functions of 0 and 1 degrees.

An additional example of spline function is suggested by Ramsay (1988) who advocates the use of monotonous spline functions. Gallego (1982), who also considers fuzzy coding, uses trapezoidal encodings as illustrated in Figure **??**.

**Fig. 2** Trapezoidal encoding (from Gallego, 1982).

## 4 Early works

Quantifying a qualitative variable on its own makes little sense if it is not linked to a goal, such as explaining another variable. Statisticians were concerned very early on with the search for non-arbitrary quantifications by seeking to optimise specific criteria (which were, most of the time, expressed as maximising squared

scalar products such as correlations). The early works were naturally concerned with the case of two categorical variables and their associated contingency table.

## 4.1 The case of bivariate distributions

Hirschfeld (1935, p. 520)—better known under his American identity of Hartley—is apparently the first researcher to ask the following question (and to answer it):

> It is well known that the correlation theory for such a distribution gives much better results if both regressions are linear [. . . ]. Given a discontinuous distribution $p_{vq}$, is it always possible to introduce [. . . ] new values for the variates $x_v$, $y_q$, such that *both* regressions are linear?

Later on (and without reference to Hirschfeld), as summarised by Lancaster (1957, pp. 289–290):

> 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).

Lancaster was referring to the algorithm described by Fisher (1940, p. 426), and now considered as an early example of alternating least squares or dual scaling, applied to the (now) famous table cross-tabulating the eye and hair colours of Scottish schoolchildren (from the county of Caithness):

> . . . 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.

This "optimal coding" algorithm converges to the solution given by the coordinates of the rows and columns along the first axis of the correspondence analysis of the contingency table.

Maung (1941, p. 200)—who was interested in the higher order encodings corresponding to the successive pairs of canonical variables—attributes to Fisher a formula giving the value of each cell in the contingency table from the margins, the canonical correlations, and the successive codings. This formula—also called the RC canonical correlation model—is none other than the well-known reconstitution formula of correspondence analysis. Williams (1952) is also a notable reference about the development of significance tests for canonical correlations.

Further details on the relationship between optimal scaling and correspondence analysis are given in Saporta (1975), Nishisato (2006, Chapter 3), Lebart and Saporta (2014), and many others, including Hill (1974), and Beh and Lombardo (2014).

## 4.2 Lancaster's theorem

The search for optimal scores is unexpectedly related to the problem of transforming a given probability distribution into a normal distributions. Lancaster (1957) showed that the (squared) correlation coefficient between the two components of a bivariate normal vector cannot be increased regardless of the (non-linear) transformations that can be applied to them. This result inspired the following comments to Kendall and Stuart (1961, pp. 568–569):

> We may ask: What scores should be allotted to the categories in order to maximize the correlation coefficient between the two variables? Surprisingly enough, it emerges that these 'optimum' scores are closely connected with the transformation of the frequencies in the table to bivariate normal frequencies [. . . ] And the theoretical implication of the [Lancaster's] 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 operations upon the margins of the table.

## 4.3 Quantifying more than two attributes: Guttman, Hayashi

Guttman (1941), in a famous paper, referred to the method of reciprocal averaging (as described by Horst, 1935), and proposed to simultaneously quantify $K$ categorical variables in such a way that they are as similar as possible and that their means are as dispersed as possible. The rationale behind this criterion was that such an approach would be optimal when the $K$ variables, collected in a multiple choice questionnaire, measured more or less the same construct (as in a factor analysis model with only one latent variable). When the total variance is fixed, this amounts to maximizing the measure of internal consistency as described below.

Let $\mathbf{X} = [\mathbf{X}_1|\ldots|\mathbf{X}_k|\ldots|\mathbf{X}_K]$ be the supermatrix of all $K$ disjunctive matrices, $\mathbf{a}_k$ the category quantification vector of variable $X_k$, $\mathbf{a}$ the supervector concatenating all category quantifications, $\mathbf{z}_k = \mathbf{X}_k\mathbf{a}_k$ the corresponding vector of object scores and

$$\bar{\mathbf{z}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_k = \frac{1}{K}\mathbf{X}\mathbf{a} \tag{11}$$

the vector of average object scores.

Guttman (1941) showed that the scores, which maximise the variance of $\bar{\mathbf{z}}$ under a scaling constraint for $\mathbf{a}$, are given by the coordinates of the modalities of the $K$ variables along the first axis of what will later be called Multiple Correspondence Analysis (MCA). On this occasion, Guttman coined the term "*chi-square* metric" now routinely associated to correspondence analysis.

Independently, Hayashi (1950) developed an approach similar to Guttman's under the name of Type III quantification. Three other types of quantification using (or not) an external response variable were also developed by Hayashi. Tanaka (1979) and Takeuchi et al. (1982, Chapter 8) are useful references for the Japanese contributions. A bit later Slater (1960) proposed a method to analyse personal preference data that

represents these data in a multi-dimensional space where observations and stimuli can be represented simultaneously, and, as noted by Nishisato (1978, p. 263), his approach was "essentially the same as Guttman's but the close relationship between them was apparently left unnoticed."

## 5 The golden seventies

The 1970s were a particularly fertile period for the development of optimal scaling and the journal *Psychometrika* was the privileged venue for publishing on this topic with no less than 145 articles published between 1968 and 1982 using the keywords "Optimal Scaling" (799 using the same keywords without dates and 199 using only the keywords "Dual Scaling"). It is therefore impossible to be exhaustive.

### 5.1 The Alternating Least Squares (ALS) approach for optimal scaling

In his 1981 Presidential Address to the Psychometric Society's Spring Meeting, Young (1981) returned at length to his work carried out in collaboration with, on one hand de Leeuw and Takane and, with, on the other hand, Tenenhaus. He reflected that these collaborations constituted an important new stream, because:

> 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 (Young 1981, p. 358).

A large number of algorithms were then developed using the alternating least squares (ALS) approach, which consists in separating the parameters of the problem into two sets:

1. the *model* parameters, and
2. the *data* parameters (the codings).

The optimisation then proceeds by obtaining the least squares estimates of the model parameters while assuming that the data parameters are constant. One then switches to the other set: obtaining the least squares estimates of the data parameters given the model parameters and so on until convergence. Even though convergence to a local optimum is guaranteed, convergence to a global optimum is *not* guaranteed because convergence depends upon the initial values (i.e., there are multiple local optima where the search could converge). Note that the ALS approach can also be applied to regression or predictive type problems which are now called *supervised* approaches, whereas the pioneers were not particularly interested in these methods.

MORALS type algorithms (Young et al., 1976) make it possible to carry out multiple regressions by transforming both a response $Y$ and the predictors $X_1, \ldots, X_k, \ldots, X_K$ with monotonic or non-monotonic optimal transformations according to

the nature of the variables by using successions of projections on vector subspaces or cones. Denoting by $\psi$ and $\varphi_1, \ldots, \varphi_K$ the transformations of the original variables, the optimisation problem is the following:

$$\max_{\psi,\varphi_1,\varphi_2,\ldots,\varphi_K} R^2 \left[\psi(Y); \varphi_1\left(X_1\right), \varphi_2\left(X_2\right), \ldots, \varphi_K\left(X_K\right)\right] \ . \tag{12}$$

Transformed variables are usually constrained to be standardized in order to avoid degeneracy.

The PRINQUAL (Bouroche et al., 1977) and PRINCALS (Young et al., 1978) algorithms implement a principal component analysis of $K$ coded qualitative variables while respecting the nominal or ordinal nature of these variables. However, the optimality criterion is not as obvious as is the maximisation of the (squared) multiple correlation in multiple regression, because this is an unsupervised problem. The most commonly used criterion maximises the percentage of variance explained by the first $L$ principal components $C_1, \ldots, C_L$ (the default value is $L = 2$ in the PRINQUAL procedure of SAS because two-dimensional displays are the ones most frequently used). Formally the maximisation problem can be expressed as the solution of:

$$\max_{\substack{\varphi_1,\varphi_2,\ldots,\varphi_k \\ C_1,\ldots,C_L}} \sum_{k=1}^{K}\sum_{\ell=1}^{L} r^2\left(\varphi_k\left(X_k\right), C_\ell\right) \ . \tag{13}$$

Note that if $L = 1$, the solution for $K$ nominal variables is identical to the solution provided by the first dimension of multiple correspondence analysis, (i.e., this is the solution of the problem from Guttman, 1941). However, there is a fundamental difference between the alorithms of the PRINQUAL-type—which look for unique codings of categorical variables—and the algorithms of the MCA and HOMALS types—which look for as many codings as the number of dimensions of the data (for more, see Gifi, 1990).

In the late 1980's, Van Buuren and Heiser (1989) developed GROUPALS, a method for optimising simultaneously a clustering of units and quantifications of categorical variables, which was taken up almost 30 years later by van de Velden et al. (2017) for their development of *cluster* correspondence analysis.

## 5.2 Dual scaling: Nishisato's synthesis

In the 1970's Nishisato (originally a psychologist, later turned into a psychometrician) revisits the problem of the quantification of qualitative variables (both nominal or ordinal) and integrates the two quantification traditions (i.e., statistics and psychometrics). Faced with so many names for equivalent methods, Nishisato preferred the appellation of *dual scaling*. In his early book, Nishisato (1980) presents an early synthesis of these two branches in the first chapter dedicated to the history of the "scaling" problem for qualitative variables—a review that remains one of the best sources for its origins and early efforts but that also often suggests future

developments. Nishisato anchors dual scaling in the early psychometric approach of Horst (1935) and Guttman (1941), but also integrates Maung's (1941) and Fisher's contributions (i.e., "additive scoring," 1940). Nishisato describes dual scaling as a maximisation problem as previously defined by Bock (1960) as an approach that:

> assign[s] numerical values to alternatives, or categories, so as to discriminate optimally among the objects (Bock, 1960; p. 1).

From this definition, Nishisato generalised and adapted the dual scaling methodology to a wider set of data types whose extension can only be compared to the, then, contemporary, French developments. For the specific problem of quantifying a set of nominal variables, Nishisato uses the super matrix approach described in Equation **??**, and derives from there the equations and properties of multiple correspondence analysis.

## 5.3 A success story: credit scoring

Credit scoring techniques are used to check if a loan applicant is worthy of credit. Using historical data on whether or not debtors have correctly repaid their instalments, the problem reduces for numerical predictors to an application of a supervised classification method such as discriminant analysis or logistic regression.

However, for individual applicants, most of the predictors are categorical variables such as gender, marital, and employment status. Scoring methods assign a score to each modality of a variable so that the addition of these partial scores best separates the two groups. Because the quantification of each predictor is equivalent to defining a linear combination of the indicators of its modalities, the optimal solution is obtained from a discriminant analysis using the columns of the associated disjunctive table as predictors:

$$\mathbf{X} = [\mathbf{X}_1 | \ldots | \mathbf{X}_k | \ldots | \mathbf{X}_K] \ . \tag{14}$$

Because $\mathbf{X}$ is not of full rank, Bouroche et al. (1977) proposed to replace it by the $P$ best components $\mathbf{z}_p$ of the multiple correspondence analysis of $\mathbf{X}$. Here "best components" means the components that best predict the target, instead of the ones with the largest eigenvalues. Fisher's linear discriminant function is then computed as and re-decomposed as a linear combination of all indicator variables which gives the optimal scores—a procedure similar to "principal component regression" for qualitative instead of quantitative variables. The previous method known as DISQUAL (see Niang & Saporta, 2006, for a detailed illustration of DISQUAL) as well as logistic regression (which eliminates an indicator in each $\mathbf{X}_k$) are routinely used by banks, insurance companies, and so on: Optimal coding has become transparent!

The interest of scores compared to black box approaches is to lead to easily interpretable decision rules—a feature now socially required.

# 6 Machine learning and variable encoding

In the machine learning terminology the modality quantification (or encoding) can be obtained by "embedding" the modalities in a low-dimensional space. For neural networks, a well-known embedding is called word-embedding (see, e.g., Bengio et al., 2003). Embedding in Natural Language Processing (NLP which is the set of techniques that use machine learning to analyse textual data) is a vector representation of the words in such a way that words which frequently appear in similar contexts are close to each other. It is possible to use the same approach for representing modalities in a vector space, in order to use models that require numerical data. Using neural networks, interesting connections appear with the optimal scaling methods described in the previous paragraphs. One of the advantages of the approach showed here is the ability to analyse categorical variables with hundreds of modalities, as long as the number of observations is adequate.

It is convenient to distinguish the supervised case, in which we need to predict a quantitative target $Y$, from the unsupervised case, in which we do not have a target variable. In the supervised case, quantification is only a tool for applying the model to qualitative data and generally has no interest in itself: The best quantification is the one that best predicts the target. By contrast, in the unsupervised case, the interest is precisely in the quantifications of the modalities: here the embedding of the modalities, and eventually of the units, should best represent the information present in the data.

## 6.1 Traditional encoding methods

In addition to the approaches described in the previous paragraphs, other methods have been proposed to encode categorical variables (for details, see the review by Hancock & Khoshgoftaar, 2020). These are simple and popular methods because they can be used for qualitative data with both classical models and machine learning algorithms. These methods either:

1. only use the target,
2. consider the target and other variables, or
3. do not consider any other data than the variable to be quantified.

In the latter case (i.e., ignoring the data), a criterion is chosen that does not use other data and the result is usually a single numeric variable. This way, there is no risk of overfitting, but the encodings obtained cannot be unambiguously interpreted. Such methods include: The *label encoder*—which assigns a different integer to each modality—and the *ordinal encoder*—which constrains the assignments to respect the natural modality order. The *hash encoder* uses a hash function to embed the $J$ modalities of a variable into a small number of dimensions, but multiple values can be represented by the same hash value—an effect known as a *collision*. Because this encoder is extremely efficient, it is sometimes used with big data sets when the

number of modalities of some variables is very high. But, in these cases, it is not possible to perform a reverse lookup to determine what the input was and so the quantifications provided by collision could be meaningless.

There are many methods that use the target to obtain a numerical coding of the modalities in such a way that the availability of other explanatory variables does not influence the coding. The result of such a procedure can be either

1. a single numeric variable for regression tasks (whose dimensionality would be the same as the dimensionality of the original data) or
2. multiple numerical variables that can then be used for classification.

Applying target-based encoding often produces *data leakage*—a problem leading to overfitting and poor predictive performance. To correctly work, this method needs large amounts of data, a small number of categorical variables, and the same target distribution in training and test data sets. To overcome data leakage, it has been suggested to add noise, or to use cross-validation techniques, or other forms of regularisation. The *simple target encoder*—a popular method for regression tasks—belongs to this group. This method assigns the conditional mean target value to each modality of the explanatory variable.

For classification tasks, where the target is also categorical, the explanatory categorical variable is encoded with $J$ new variables (where $J$ is the number of classes of the target). These variables contain the relative conditional frequencies of each class given the modality of the categorical variable.

Other methods in this approach are based on the *contrast* between some modalities and other modalities of the variable, these methods are called *contrast encoders* (an approach often used in the general linear model framework for testing specific predictions). For example, the *Helmert encoder* requires a quantitative target and ordered levels of the categorical variable; this encoder generates a set of contrasts where each modality is compared in turn to all the subsequent ones. This method is also routinely used in multiple regression and analysis of variance.

A favourite method to analyse qualitative variables is the, previously mentioned, *one hot encoding* which assigns one indicator matrix to each variable. Note that OHE differs from *dummy coding* that excludes one modality of the variable (to avoid multicollinearity). But, when applying machine learning models it is necessary to include all the modalities, otherwise the omitted modality disappears—a standard problem (called "the dummy variable trap") in multiple regression when using dummy coding (see, e.g., Darlington & Hayes, 2017).

In fact, one hot encoding is not a real quantification method, but just a binary transformation of the original data. Using OHE makes is possible to take into account the other explanatory variables because the quantifications are obtained as parameters of a model. The main drawback of OHE follows from the tendency of indicator variables to cause overfitting. Moreover, if a variable has many modalities, OHE generates a large number of new features and a sparse array in which the new indicator variables are perfectly independent—an unrealistic assumption. OHE is used in the optimal scaling approach (see MORALS in Section **??**) but is also widely used in machine learning.

## 6.2 Non-linear encoding in the supervised case

In the supervised case, modality quantifications are generally just a tool for applying a predictive model. The best quantification will therefore be the one that gives the best predictions for the model used.

As shown in Section **??**, MORALS makes it possible to perform a multiple regression considering optimal transformations of the variables. Let $\mathbf{X}_k$ (of dimensions $I \times J_k$) be the indicator matrix of variable $k$, and $Y$ a numerical response variable. If we have $K$ categorical explanatory variables, MORALS defines the *residual sum of squares* (RSS) as:

$$\text{RSS} = \left\| \mathbf{y} - \sum_{k=1}^{K} \beta_k \mathbf{X}_k \mathbf{a}_k \right\|^2 = \left\| \mathbf{y} - \sum_{k=1}^{K} \beta_k \mathbf{q}_k \right\|^2 \tag{15}$$

where $\mathbf{q}_k = \mathbf{X}_k \mathbf{a}_k$ is the vector of the quantified $k$th variable, $\mathbf{a}_k$ is the vector with the (single) quantification of the modalities of the $k$th variable, with the centering and normalisation constraints:

$$\mathbf{1}^{\mathrm{T}} \mathbf{q}_k = 0, \quad \frac{1}{I} \mathbf{q}_k^{\mathrm{T}} \mathbf{q}_k = 1, \quad k = (1, 2, ..., K). \tag{16}$$

The algorithm then defines the following optimisation problem, solved by an alternating least squares algorithm:

$$\min_{\substack{\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_K \\ \beta_1, \beta_1, \ldots, \beta_K}} \left\| \mathbf{y} - \sum_{k=1}^{K} \beta_k \mathbf{X}_k \mathbf{a}_k \right\|^2 . \tag{17}$$

With only explanatory nominal variables—unless a different normalisation of the parameters is used—MORALS essentially corresponds to a linear regression with OHE. This approach is likely to overfit data sets with few observations or when variables have many modalities. It is also possible to obtain multiple quantifications by creating copies of the variables (see, e.g., Gifi, 1990). However, this approach would increase the number of free parameters and having more parameters to fit the data would worsen the overfitting problems of MORALS.

In machine learning, and specifically for neural networks, OHE encoding is often used to analyse categorical variables. All the dummies of all the variables, put together, constitute the input of the network. However, this method is not an optimal choice because it greatly increases the size of the dataset by adding orthogonal binary variables.

A different and more adequate strategy (proposed by Di Ciaccio, 2020) is described below. Let $L$ be the chosen dimensionality of the embedding space. To explicitly introduce the quantification of modalities in a neural network, it is possible to define an architecture which provides a distinct input for each categorical variable. Each input will be of the OHE type and will be followed by a "dense layer" (the classical fully connected layer) with $L$ neurons without bias and with a linear

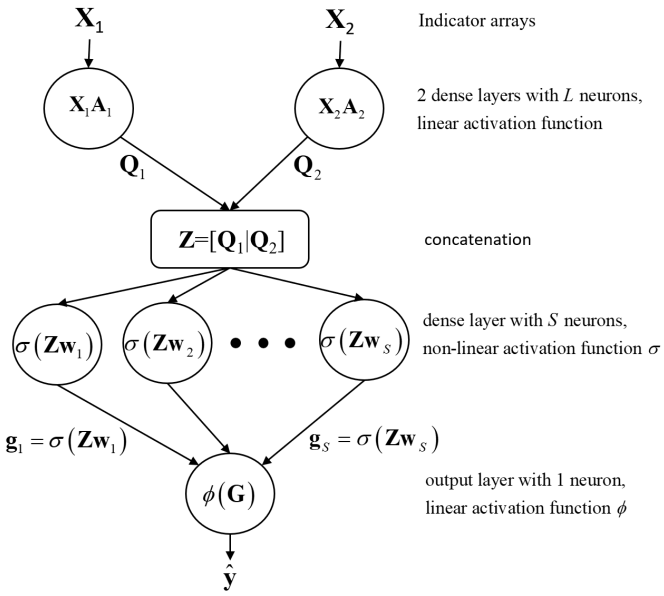


**Fig. 3** Supervised neural network for two nominal explicative variables

activation function. Layers and activation functions are the basic elements of a neural network (for definition of these terms see, e.g., Abdi et al., 1999; or Bengio et al., 2003). The output of this step is an array $\mathbf{Q}_k$ (of dimensions $I \times L$) for each variable, which gives the $L$-dimensional quantification of $\mathbf{X}_k$, while the modality quantifications are given by $\mathbf{A}_k$. In the next layer, the outputs, coming from all the variables, must be concatenated. At this point, we can add the classical layers of a neural network, for example, one dense layer with $S$ neurons and activation function $\sigma$ (usually non-linear, chosen by the researcher), and one output dense layer with only one neuron and a linear activation function $\varphi$ (if $Y$ is quantitative). The final network architecture is showed in Figure **??**. The corresponding neural network can be defined as:

$$\widehat{\mathbf{y}} = \beta_0 + \sum_{s=1}^{S} \beta_s \sigma \left( \sum_{k=1}^{K} \sum_{\ell=1}^{L} \mathbf{X}_k \mathbf{a}_{k\ell} w_{k\ell s} + w_{0s} \right). \tag{18}$$

Conversely, in the classical OHE encoding:

$$\widehat{\mathbf{y}} = \beta_0 + \sum_{s=1}^{S} \beta_s \sigma \left( \sum_{k=1}^{K} \mathbf{X}_k \mathbf{w}_{ks} + w_{0s} \right). \tag{19}$$

The function $\sigma$ is the activation function of the dense layer with $S$ neurons and is usually non-linear. The embedding dimension is given by $L$, while $S$ is the number of neurons which determines the adaptive capacity of the network. In Equation **??**, $\mathbf{X}_k \mathbf{a}_{k\ell}$ is equal to $\mathbf{q}_{k\ell}$, which is the $\ell$th column of $\mathbf{Q}_k$.

A relevant difference between the two expressions is the different number of parameters. If the qualitative variables have more than two modalities and if $L = 2$, there are fewer parameters in Equation **??**. Even if the variables have many modalities (e.g., 100 or 200), the embedding of Equation **??** makes it possible to perform the analysis without difficulty because it involves a smaller number of parameters. Di Ciaccio (2023) showed how this approach—compared to OHE or *target encoding*—leads, with neural networks, to much better predictions. Other works that consider a comparison between different techniques in the supervised approach are, for example, Di Ciaccio (2023) and Potdar et al. (2017).

## 6.3 Non-linear encoding in the unsupervised case

In the unsupervised case, the quantifications can be the true goal of the analysis and must therefore highlight the information present in the data. The modalities can be represented in a vector space obtaining multiple quantifications, as in the case for HOMALS and MCA.

With HOMALS or MCA, the modalities are "optimally" encoded by using the eigenvectors with the largest eigenvalues of the correlation matrix. In MCA, the problem is solved analytically, while in HOMALS, the problem is solved numerically. This numerical variant offers great flexibility in machine learning. The MCA / HOMALS approaches are linear methods that give a map where both units and variables are represented in a low $L$-dimensional Euclidean space in such a way that an observed unit is relatively close to the modalities that characterise it and away from the modalities that do not. In this representation, the modality embeddings are the centres of gravity of the units that share the same modality.

Let $\mathbf{Z}$ (of Dimensions $I \times L$) be the score matrix (the observations coordinates on the vector space), $\mathbf{X}_k$ (of dimensions $I \times J_k$) the indicator matrix of variable $k$, $\mathbf{A}_k$ (of dimensions $J_k \times L$) the multiple quantification of the modalities, and $\mathbf{U}_k$ the unitary matrix (of dimensions $L \times L$). The HOMALS loss finds the object scores $\mathbf{Z}$ and the quantifications $\mathbf{A}_k$ so that:

$$\min_{\substack{\mathbf{A}_1,\mathbf{A}_2,\ldots,\mathbf{A}_K \\ \mathbf{Z}}} \text{LOSS} = \sum_{k=1}^{K} \|\mathbf{Z} - \mathbf{X}_k\mathbf{A}_k\|^2 \tag{20}$$

with the centring and normalisation constraints $\mathbf{u}'\mathbf{Z} = \mathbf{0}$, $\mathbf{Z}'\mathbf{Z} = I\mathbf{U}$, to avoid the trivial solutions: $\mathbf{Z} = \mathbf{0}$, $\mathbf{A}_k = \mathbf{0}$. The LOSS function in Equation **??** can be written as:

$$\text{LOSS} = \sum_{k=1}^{K} \|\mathbf{Z} - \mathbf{X}_k\mathbf{A}_K\|^2 = \sum_{k=1}^{K} \left\|\mathbf{X}_k - \mathbf{Z}\mathbf{A}_k^+\right\|^2$$

$$= \sum_{k=1}^{K} \left\|\mathbf{X}_k - \widehat{\mathbf{X}}_k\right\|^2 = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k} \left(x_{ikj} - \widehat{x}_{ikj}\right)^2 \tag{21}$$

where $\widehat{\mathbf{X}}_k$ is the best "reconstruction" of $\mathbf{X}_k$ and $\mathbf{A}_k^+$ the Moore-Penrose inverse of $\mathbf{A}_K$. Considering that, to minimize this loss, $\mathbf{Z}$ has to be the mean of the $K$ matrices $\mathbf{X}_k\mathbf{A}_K$, the modality quantifications $\mathbf{A}_K$ are the only parameters to estimate.

The previous expression suggests an alternative formulation as an autoencoder neural network (an autoencoder, also called an autoassociator, associates a pattern to itself, often as a way of de-noising a signal; an autoencoder can also be seen as a non-linear version of principal component analysis; for more, see Bengio et al., 2003). Within our framework, an autoencoder is a particular neural network able to minimise the LOSS:

$$\min_{\sigma,\varphi} L(\mathbf{X}, \sigma(\varphi(\mathbf{X}))) \tag{22}$$

where $\varphi$ and $\sigma$ introduce some constraints in the reconstruction of $\mathbf{X}$ and the LOSS penalises the difference between $\mathbf{X}$ and $\widehat{\mathbf{X}}$. Using the Residual Sum of Squares (RSS), Equation **??** becomes:

$$\min_{\sigma,\varphi} \|\mathbf{X} - \sigma(\varphi(\mathbf{X})\|^2 \tag{23}$$

where $\varphi$ maps the indicator array $\mathbf{X}$ to an $L$-dimensional latent space (the bottleneck), $\sigma$ maps this representation to the output, which is the same as the input. Considering only linear $\varphi, \sigma$, and a low embedding of dimension $L$, the architecture of the corresponding autoencoder for only two nominal variables is shown in Figure **??**. This neural network includes only dense layers (also called standard or fully connected layers).

The first layer is composed by two dense sub-layers with $L$ neurons for each variable and linear activation function. The output layer has two dense sub-layers with as many neurons as the number of modalities of the corresponding variable and a linear activation function. The autoencoder produces the modality quantification $\mathbf{A}_1$ and $\mathbf{A}_2$ on $L$ dimensions (usually $L = 2$ or 3). The score matrix $\mathbf{Z}$ is the mean of the quantified variables $\mathbf{Q}_1$ and $\mathbf{Q}_2$ on $L$ dimensions. To obtain the same results as HOMALS, the score matrix $\mathbf{Z}$ needs to be orthonormalised and column centred. Of course, actually performing all these computations would not make sense, because with much less effort we can use the elegant analytical solution provided by MCA or the alternating least squares algorithm of HOMALS.

The neural network architecture shown in Figure **??** highlights two constraints:

1. the weights of the output layer are the inverse weights of the first layer, and
2. for all layers, the activation function is linear.

Moreover, the LOSS function of HOMALS is based on the classical RSS, which may not be the best choice to compare $\widehat{\mathbf{X}}_k$ to $\mathbf{X}_k$. It is possible to extend the previous approach

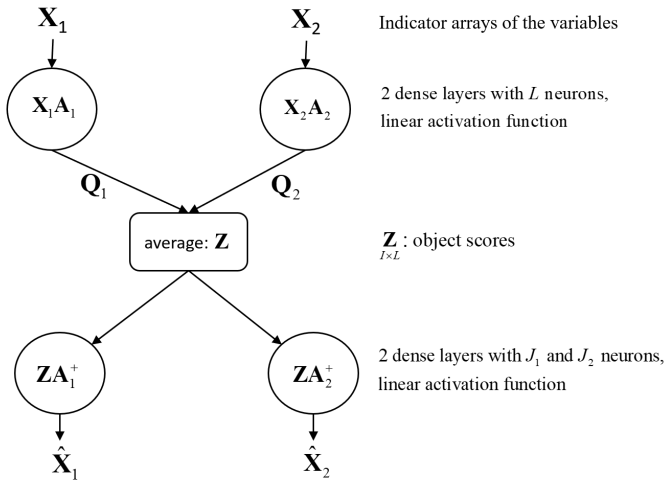


**Fig. 4** An autoencoder that reproduces HOMALS

by eliminating these two constraints and introducing a better LOSS function. The new architecture of the autoencoder for only two nominal variables and dimension $L$ is shown in Figure **??**.

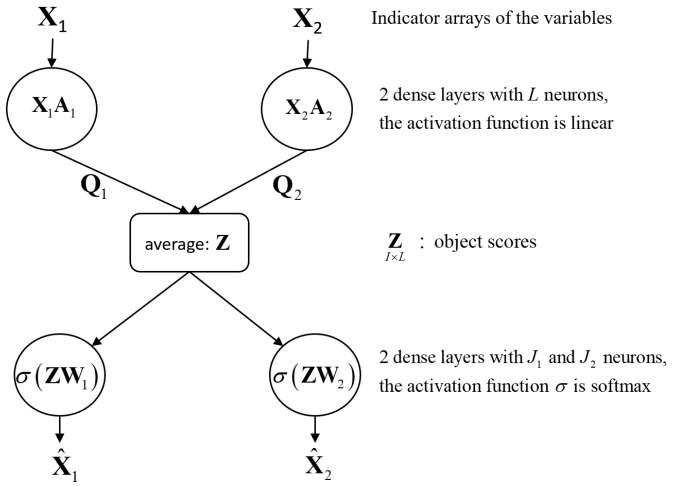


**Fig. 5** Autoencoder to extend HOMALS to non-linear encoding

Note that in the output layer there is a new parameter matrix $\mathbf{W}_k$ (of dimensions $L \times J_k$) and the activation function is now *Softmax* (see Bengio et al., 2003)—the same function as used in multinomial logistic regression. Specifically, Softmax is a

function, denoted $\sigma$:$\mathbb{R}^J \rightarrow (0,1)^J$, defined as

$$\sigma(\mathbf{v})_j = \frac{e^{v_j}}{\sum_{m=1}^{J} e^{v_m}}, \qquad j = 1, \ldots, J \quad \text{and} \quad \mathbf{v} = (v_1, v_2, \ldots, v_J)\,. \tag{24}$$

This way, $\widehat{\mathbf{X}}_k$ contains, for each unit, the estimated probability of assuming the different modalities of variable $k$. Then, the categorical cross-entropy $H\left(\mathbf{X}_k, \widehat{\mathbf{X}}_k\right)$ (also called logistic loss) is more appropriate to compare the reconstructed array to the indicator array $\mathbf{X}_k$:

$$\sum_{k=1}^{K} H\left(\mathbf{X}_k, \widehat{\mathbf{X}}_k\right) = -\sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{j=1}^{J_k} x_{ikj} \log \widehat{x}_{ikj} = -\sum_{k=1}^{K}\sum_{i=1}^{I} \log\left(\sigma\left(\mathbf{z}_i \mathbf{W}_k\right)\right)\mathbf{x}_{ik}^{\mathrm{T}} \tag{25}$$

where $\mathbf{z}_i$ is the $i$th row vector of $\mathbf{Z}$ (with length $L$). Then the minimisation problem becomes:

$$\min_{\substack{\mathbf{W}_1, \mathbf{W}_k, \ldots, \mathbf{W}_K \\ \mathbf{Z}}} \sum_{k=1}^{K} H\left(\mathbf{X}_k, \sigma\left(\mathbf{Z}\mathbf{W}_k\right)\right)\,. \tag{26}$$

Considering that, by definition, $\mathbf{Z}$ is the mean of $\mathbf{X}_k\mathbf{A}_K$, the modality quantifications $\mathbf{A}_K$ and the weights $\mathbf{W}_K$ are the parameters to estimate. The non-linear encoding achieved in this way can be much more effective than the encoding provided by homals / mca. Note that both methods (i.e., homals and its non-linear extension) use the same ohe coding of the categorical variables as input. However, the parametrisation is different and the extension includes more parameters, a non-linear transformation, and a different objective function.

As a simple example, consider only two categorical variables, $X$ and $Y$, each with 5 modalities denoted (respectively) by ($A$, $B$, $C$, $D$, $E$) and ($a$, $b$, $c$, $d$, $e$), which, together, produce the contingency table shown in Table **??** (from Di Ciaccio, 2023). The strong associations of the pairs of modalities ($A$, $a$), ($B$, $b$), ($C$, $c$), ($D$, $d$), ($E$, $e$) are evident because of the dominant cell frequencies that appear in the main diagonal of the table.

**Table 1** A contingency table showing the association between variables $X$ and $Y$.

| $X/Y$ | $a$ | $b$ | $c$ | $d$ | $e$ | 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 |

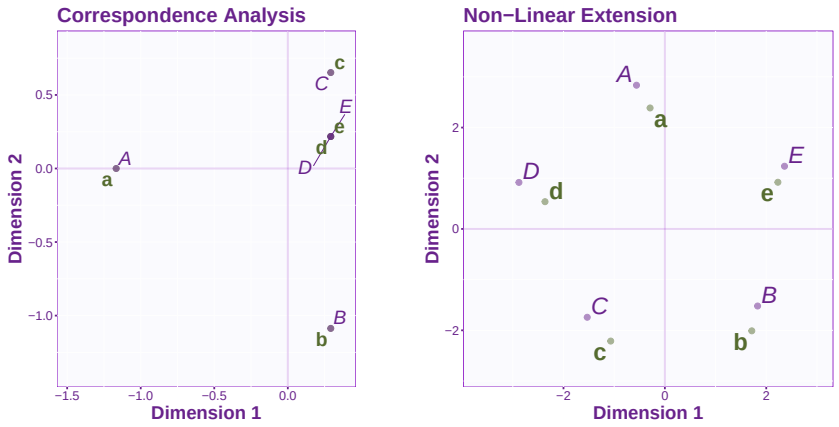


**Fig. 6** Categorical encoding for CA (left) and non-linear extension (right) on data of Table.**??**, first two components.

We would therefore expect a representation on two components that highlights these associations: a representation where strongly associated pairs are close to each other and equally far away from the other modalities. By applying MCA, the first four components have the same eigenvalue and are all necessary to obtain a satisfactory representation of the modalities. This is a feature of the matrix being symmetric; see Beh and Lombardo (2022). Figure **??** shows the result obtained from the first two components of MCA (on the left) and with the non-linear version just described (on the right). Note how—with the presence of only one more unit for the pair $(A, a)$—MCA creates, on the first two dimensions, a configuration hard to interpret. By contrast, non-linear extension shows, with only two axes, a representation of the associations very consistent with the data in the table.

## 7 Conclusion and perspectives: towards a renewal of optimal coding methods

Transforming qualitative variables into numerical variables is once again a hot topic in part because the profusion of (qualitative) variables with a large number of modalities often found in big data analytics applications.

The statisticians who developed optimal scaling methods were not very concerned about the overfitting and instability issues that could arise from the use of a large number of indicators because these statisticians often worked with low dimensional data (they, however, developed very efficient algorithms in the linear case). The DISQUAL method was certainly a method of regularisation by projection onto a low-dimensional subspace, but this aspect remained secondary to the objective of calculating scores. Similarly, the work of Russolillo (2012) uses optimal scaling to

be able to apply PLS regression and PLS path modeling to qualitative data without really focusing on the regularising effect of projection onto the PLS components.

It is only very recently (see Meulman et al, 2019) that regularisation by Ridge, LASSO, or Elastic Net has been combined with MORALS-type optimal scaling regression—a combination that opens up many new opportunities.

Largely independently, machine learning practitioners confronted with these high-dimensional problems have developed—without always being concerned with optimality or robustness—a large number of techniques, some of them arbitrary, or some of them being a rediscovery of known techniques. However, we have noticed that an approach based on neural networks leads to satisfactory results not only in supervised but also in unsupervised approaches. In the latter case, an autoencoder network minimising the cross-entropy with the consideration of non-linear links may give better results than the least-squares minimisation at the origin of the alternating least-squares methods.

## References


1. Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Sage.
2. Beh, E., & Lombardo, R. (2014). *Correspondence analysis: theory, practice and new strategies.* Wiley.
3. Beh, E., & Lombardo, R. (2022). Visualising departures from symmetry and Bowker's $X^2$ statistic. *Symmetry 14*, 1103.
4. Bengio Y., Ducharme R., Vincent P., Janvin C. (2003). A neural probabilistic language model, *The Journal of Machine Learning Research, 3*, 1137–1155.
5. Bock, R. D. (1960). *Methods and applications of optimal scaling*. The University of North Carolina Psychometric Laboratory Research Memorandum No. 25.
6. Bouroche, J. M., Saporta, G. & Tenenhaus, M. (1977). Some methods of qualitative data analysis. In J. R. Barra, (Ed.), *Recent developments in statistics, proceedings of the European meeting of statisticians* pp. (749–755). North-Holland. `https://hal-cnam.archives-ouvertes.fr/hal-03059983`
7. Coombs, C. H. (1948). Some hypotheses for the analysis of qualitative variables. *Psychological Review, 55*, 167–174.
8. Coombs, C. H. (1964). *A theory of data*. Wiley.
9. Darlington, R. B., Hayes, A. F. (2017). *Regression analysis and linear models.* Guilford.
10. de Leeuw, J. (1973). *Canonical analysis of categorical data* [Doctoral Dissertation, The University of Leiden].
11. de Leeuw, J., Hornik, K., & Mair, P. (2009). Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of Statistical Software, 32*, 1–24.
12. Di Ciaccio, A. (2020). Categorical Encoding for Machine Learning, in A. Pollice et al. (Eds.), *Book of short papers SIS 2020* (pp. 1048–1053). Pearson.
13. 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. CLADAG2021. Studies in Classification, Data Analysis and Knowledge Organization*. (pp. 39–41). Springer.
14. Festinger, L. (1947). The treatment of qualitative data by scale analysis. *Psychological Bulletin, 44*, 149–161.
15. Fisher, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics, 10*, 422–429.

16. Gallego, F. J. (1982). Codage flou en analyse des correspondances. *Cahiers de l'Analyse des Données, 7*, 413–430.
17. Gifi, A. (1990). *Nonlinear multivariate analysis*. John Wiley & Sons.
18. Guttman L. (1941). The quantification of a class of attributes: a theory and method of a scale construction. In P. Horst (Ed.), *The prediction of personal adjustment* (pp. 321–348). SSCR.
19. Guttman L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.
20. Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data, 7*(1), 1–41.
21. Hayashi, C. (1950). On the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statistical Mathematics, 2*, 35–47.
22. Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 23*(3), 340–354.
23. Hirschfeld, H. O. (1935). A connection between correlation and contingency. *Mathematical Proceedings of the Cambridge Philosophical Society 31*, 520–524.
24. Horst, P. (1935). Measuring complex attitudes *The Journal of Social Psychology, 6*, 369–374.
25. Hotteling, H. (1936). Relations between two sets of variates. *Biometrika, 28*, 321–377.
26. Kendall, M. G., & Stuart, A. (1961). *The advanced theory of statistics Volume II.* Charles Griffin.
27. Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika, 29*, 115–129.
28. Lancaster, H. O. (1957). Some properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika, 44*, 289–292.
29. Lebart, L., & Saporta, G. (2014). Historical elements of correspondence analysis and multiple correspondence analysis. In J. Blasius & M. Greenacre (Eds.), *Visualization and verbalization of data* (pp. 73–86). Chapman and Hall/CRC.
30. Maung, K. (1941). Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. *Annals of Eugenics, 11*, 189–223.
31. Meulman, J. J., van der Kooij, A .J., & Duisters, K. L. (2019). ROS regression: Integrating regularization with optimal scaling regression. *Statistical Science, 34*, 361–390.
32. Nishisato, S. (1978). Optimal scaling of paired comparison and rank order data: An alternative to Guttman's formulation. *Psychometrika, 43*, 263–271.
33. Nishisato, S. (1980). *Analysis of categorical data: dual scaling and its applications*. University of Toronto Press.
34. Nishisato, S. (2006). *Multidimensional nonlinear descriptive analysis*. Chapman and Hall/CRC.
35. Potdar, K., Pardawala, T. S., & Pai, C. D. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications, 175*, 7–9.
36. Ramsay, J. O. (1988). Monotone regression splines in action, *Statistical Science, 3*, 425–441.
37. Russolillo, G. (2012). Non-metric partial least squares. *Electronic Journal of Statistics, 6*, 1641–1669.
38. Saporta, G. (1975). Dépendance et codages de deux variables aléatoires. *Revue de Statistique Appliquée, 23*, 4–63.
39. Saporta, G. & Niang-Keita, N. (2006). Correspondence analysis and classification. in M. Greenacre, & J. Blasius (Eds.), *Multiple correspondence analysis and related methods* (pp. 371–392). Chapman and Hall/CRC.
40. Slater, P. (1960). The analysis of personal preferences. *The British Journal of Statistical Psychology, 13*, 119–135.
41. Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677–680.
42. Takane, Y. (1980). Analysis of categorizing behavior by a quantification method. *Behaviometrika, 8*, 57–67.
43. Takeuchi, K., Yanai, H., & Mukherjee, B. N. (1982). *The foundations of multivariate analysis*. Wiley.

44. Tanaka, Y. (1979). Review of the methods of quantification. *Environmental Health Perspectives, 32*, 113–123.
45. Tenenhaus, M. (1988). Canonical analysis of two convex polyhedral cones and applications. *Psychometrika, 53*, 503–524.
46. Tenenhaus, M., & Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical data. *Psychometrika, 50*, 91–119.
47. van Buuren, S., & Heiser, W. J. (1989). Clustering $N$ objects into $K$ groups under optimal scaling of variables. *Psychometrika, 54*, 699–706.
48. van de Velden, M., D'Enza, A. I., & Palumbo, F. (2017). Cluster correspondence analysis. *Psychometrika, 82*, 158–185.
49. Williams, E. J. (1952). Use of scores for the analysis of association in contingency tables. *Biometrika, 39*(3/4), 274–289.
50. Young, F. W., de Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: alternating least squares methods with optimal scaling features. *Psychometrika, 41*, 505–529.
51. Young, F. W., Takane, Y., & de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika, 43*, 279–281.
52. Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika, 46*, 357–388.