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
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

\[ X \in \{1, 2, \ldots, M\} \]
\[ \tilde{X} \in \{a_1, a_2, \ldots, a_M\} \]
\[ \tilde{X} = \sum_{m=1}^{M} a_m 1_m \]

- Matrix Formulation

\[ x = \begin{pmatrix} 1 \\ 2 \\ 2 \\ 1 \\ \vdots \\ m \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} a_1 \\ a_2 \\ a_2 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} = \begin{pmatrix} 1000..0 \\ 0100..0 \\ 0100..0 \\ 1000..0 \\ \vdots \\ 0000..1 \end{pmatrix} \]

\[ \tilde{x} = Xa \]
• Subspace of dimension $M$ \( \tilde{x} \in W \)

• Redundancy (?):
\[
\sum_{m=1}^{M} 1_m = 1
\]

• Centered solutions:
\( \tilde{x} \in \left\{ \Delta^\perp \cap W \right\} \)
Variables with ordered modalities

- **Require** \( a_1 \leq a_2 \leq \ldots \leq a_M \)
- **Reparametrization**
  \[
  a_1 = b_1, \ a_2 = b_1 + b_2, \ldots, a_M = b_1 + b_2 + \ldots + b_M
  \]
  \[
  b_1 \in \mathbb{R} \quad b_2, \ldots, b_M \geq 0
  \]
  \[
  \tilde{x} = \sum_{m=1}^{M} a_m \mathbf{1}_m = b_1 \mathbf{1}_1 + (b_1 + b_2) \mathbf{1}_2 + \ldots + (b_1 + b_2 + \ldots + b_M) \mathbf{1}_M
  \]
  \[
  = b_1 (\mathbf{1}_1 + \mathbf{1}_2 + \ldots + \mathbf{1}_M) + b_2 (\mathbf{1}_2 + \ldots + \mathbf{1}_M) + \ldots + b_M \mathbf{1}_M
  \]
  \[
  = b_1 \mathbf{1} + b_2 (\mathbf{1}_2 + \ldots + \mathbf{1}_M) + b_3 (\mathbf{1}_3 + \ldots + \mathbf{1}_M) + \ldots + b_M \mathbf{1}_M
  \]
• Linear combination with positive coefficients of \( M-1 \) variables and a constant term

\[
\tilde{x} = b_1 1 + \sum_{m=2}^{M} b_m z_m
\]

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

\[
\tilde{x} \in \{ \Delta \oplus C_{M-1} \}
\]
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?

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

- **$X$ 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]
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 $\nu$th value of a variate $X_\mu (\mu = 1, 2, \ldots, N)$ concurring with the $q$th 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 $X_\nu, Y_q$. On the other hand the original scales or better values assigned to the $X_\nu$ 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 $x_\nu, y_q$, such that both regressions are linear?
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

\[
a_{ij} = \frac{a_i a_j}{a_{..}} \left( 1 + \sum_{k=1}^{m-1} (x_k y_k R_k) \right),
\]

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

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

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:

<table>
<thead>
<tr>
<th>Eye colour</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Light</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medium</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Dark</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Eye colour</th>
<th>$x$</th>
<th>Hair colour</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light</td>
<td>$-0.9873$</td>
<td>Fair</td>
<td>$-1.2187$</td>
</tr>
<tr>
<td>Blue</td>
<td>$-0.8968$</td>
<td>Red</td>
<td>$-0.5226$</td>
</tr>
<tr>
<td>Medium</td>
<td>$0.0753$</td>
<td>Medium</td>
<td>$-0.0941$</td>
</tr>
<tr>
<td>Dark</td>
<td>$1.5743$</td>
<td>Dark</td>
<td>$1.3189$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Black</td>
<td>$2.4518$</td>
</tr>
</tbody>
</table>

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 $N$

• Hirschfeld: transition formulas

$$D_1^{-1}Nb = \sqrt{\lambda}a \quad D_2^{-1}N'a = \sqrt{\lambda}b$$

• Fisher: maximization of:

$$\text{cov}(X_1a, X_2b) = \frac{1}{n}a'Nb$$

\[\begin{cases}
D_1^{-1}ND_2^{-1}N'a = \lambda a \\
D_2^{-1}N'D_1^{-1}Nb = \lambda b
\end{cases}\]

\[\begin{align*}
\frac{1}{n}a'D_1a &= \frac{1}{n}b'D_2b = \lambda
\end{align*}\]

• 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 $\rho$. If now a transformation, $x' = x'(x)$, $y' = y'(y)$, is made to any new variables $x'$ and $y'$ such that

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

are finite, then the correlation of the new variables is less in absolute value than $\rho$. That is, $\rho$ 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

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”

Louis Guttman (1916-1987)
Supplementary Study B-3

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

1. The Problem
2. Characteristics of the Method
3. The Correlation Ratio for Weights
4. Maximizing the Correlation Ratio
5. The "Chi-Square" Metric
6. The Number of Independent Solutions
7. The Correlation Ratio for Scores
8. The Bivariate Distribution of Weights and Scores and the Correlation Coefficient
9. Maximizing the Correlation Coefficient
10. The Identity of the Results
11. The Equations of Internal Consistency and Linearity of Regression
12. Criticism of Certain Practices
13. The Reconstruction and Prediction of Behavior
14. Computational Procedures
A Bibliographical Note
A Note on a Machine Method for the Quantification of Attributes

Ledyard R. Tucker

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):

<table>
<thead>
<tr>
<th>Subcategory</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>...</th>
<th>$U$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>...</td>
<td>✓</td>
</tr>
<tr>
<td>$A_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$B_1$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>...</td>
<td>✓</td>
</tr>
<tr>
<td>$B_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Z_1$</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>...</td>
<td>✓</td>
</tr>
<tr>
<td>$Z_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 $\xi_i$ 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)
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)
<table>
<thead>
<tr>
<th>Program</th>
<th>Analysis</th>
<th>Data</th>
<th>Source</th>
<th>Primary Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADDALS</td>
<td>Additivity analysis</td>
<td>Two or three way tables. Nonorthogonal and incomplete designs permitted.</td>
<td>UNC</td>
<td>de Leeuw, Young &amp; Takane (1976)</td>
</tr>
<tr>
<td>WADDALS</td>
<td>Weighted additivity analysis</td>
<td>Same as ADDALS</td>
<td>UNC</td>
<td>Takane, Young &amp; de Leeuw (1980)</td>
</tr>
<tr>
<td>MANOVA</td>
<td>Multivariate analysis of variance</td>
<td>Multi-way tables</td>
<td>RUL</td>
<td>Gifi (1981)</td>
</tr>
<tr>
<td>CORALS &amp; CANALS</td>
<td>Multiple and canonical analysis</td>
<td>Mixed measurement level multivariate data</td>
<td>UNC or RUL</td>
<td>Young, de Leeuw &amp; Takane (1976)</td>
</tr>
<tr>
<td>OVERALS</td>
<td>Canonical analysis</td>
<td>Multiple set mixed measurement level multivariate data</td>
<td>RUL</td>
<td>Gifi (1981)</td>
</tr>
<tr>
<td>CRIMINALS</td>
<td>Multiple group discriminant analysis</td>
<td>Mixed measurement level predictors</td>
<td>RUL</td>
<td>Gifi (1981)</td>
</tr>
<tr>
<td>PATHALS</td>
<td>Path analysis</td>
<td>Mixed measurement level multivariate data</td>
<td>RUL</td>
<td>Gifi (1981)</td>
</tr>
<tr>
<td>PRINCALS &amp; PRINCIALS</td>
<td>Principal components analysis</td>
<td>Mixed measurement level multivariate data</td>
<td>UNC or RUL</td>
<td>Young, Takane &amp; de Leeuw (1978)</td>
</tr>
<tr>
<td>NOMALS</td>
<td>Principal components analysis</td>
<td>Multivariate nominal data</td>
<td>RUL</td>
<td>de Leeuw &amp; van Rijkvorsel (1976)</td>
</tr>
<tr>
<td>ALSCOMP &amp; TUCKALS</td>
<td>Three-mode factor analysis</td>
<td>Mixed measurement level multivariate data</td>
<td>UNC or RUL</td>
<td>Sands &amp; Young (1978)</td>
</tr>
<tr>
<td>FACTALS</td>
<td>Common-factor analysis</td>
<td>Mixed measurement level multivariate data</td>
<td>UNC</td>
<td>Takane, Young &amp; de Leeuw (1978)</td>
</tr>
<tr>
<td>ALSCAL</td>
<td>Two or three-way multidimensional scaling</td>
<td>Similarity data</td>
<td>UNC</td>
<td>Takane, Young &amp; de Leeuw (1977)</td>
</tr>
<tr>
<td>GEMSCAL</td>
<td>Two or three-way multidimensional scaling</td>
<td>Similarity data</td>
<td>UNC</td>
<td>Young, Null &amp; De Soete (Note 5)</td>
</tr>
</tbody>
</table>
• PRINQUAL (SAS), PRINCALS, PRINCIPALS...

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

• 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)

$$\max_{\psi, \phi_1, \phi_2, \ldots, \phi_P} R^2(\psi(Y); \phi_1(X_1), \phi_2(X_2), \ldots, \phi_P(X_P))$$

• Alternating regression and optimal variable-by-variable coding

• With all categorical predictors and no transformation of Y, equivalent to regress y on the complete disjunctive table of predictors

$$X = \left[ X_1 \mid \ldots \mid X_p \mid \ldots \mid X_p \right]$$

• 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
  5. Region (2): Brussels, other regions
  8. Year of subscription (2): <86, others
**Fisher’s LDA using MCA components**

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

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

\[
Y'X = \begin{bmatrix}
Y'X_1 & \cdots & Y'X_p & \cdots & Y'X_p
\end{bmatrix}
\]

• 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
• Linear setup \[ \varphi_p \left( X_p \right) = X_p b_p \]

• Elastic net:
\[
\min_{T,b_p} \left\{ \left\| T(y) - \sum_{p=1}^P X_p b_p \right\|^2 + \lambda_1 \sum_{p=1}^P \sum_{j=1}^{J_p} |b_{pj}| + \lambda_2 \sum_{p=1}^P \sum_{j=1}^{J_p} (b_{pj})^2 \right\}
\]

• 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 \( \sigma \) with *One Hot Encoding*

\[
\hat{y} = \beta_0 + \sum_{s=1}^{S} \beta_s \sigma \left( \sum_{p=1}^{P} X_p w_{ps} + w_{0s} \right)
\]

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

\[
\hat{y} = \beta_0 + \sum_{s=1}^{S} \beta_s \sigma \left( \sum_{p=1}^{P} \sum_{l=1}^{L} X_p a_{pl} w_{pls} + w_{0s} \right)
\]

and better predictions
5.2 An unsupervised auto-encoder approach

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

\[
\min_{Z, A_1, A_2, \ldots, A_p} \left( \sum_{p=1}^P \| Z - X_p A_p \|^2 \right) = \sum_{p=1}^P \| X_p - \hat{X}_p \|^2 = \sum_{i=1}^L \sum_{p=1}^P \sum_{j=1}^{J_p} (x_{ij} - \hat{x}_{ij})^2
\]

\( Z \quad IxL \) matrix of units scores,

\( A_p \quad J_p xL \) matrix of multiple quantifications of \( X_p \)

• Achievable with an autoencoder where category embeddings are the centroids of units sharing the same category.
Two categorical variables

- $A_1, A_2$ quantification of categories in $L$ dimensions.

- NB: Algorithm far less efficient than MCA or HOMALS
A non-linear variant (Di Ciaccio, 2023)

- **Softmax function**

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

- **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

<table>
<thead>
<tr>
<th>X/Y</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>801</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>1201</td>
</tr>
<tr>
<td>B</td>
<td>100</td>
<td>800</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>1200</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
<td>100</td>
<td>800</td>
<td>100</td>
<td>100</td>
<td>1200</td>
</tr>
<tr>
<td>D</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>800</td>
<td>100</td>
<td>1200</td>
</tr>
<tr>
<td>E</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>800</td>
<td>1200</td>
</tr>
<tr>
<td>Total</td>
<td>1201</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>1200</td>
<td>6001</td>
</tr>
</tbody>
</table>

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


