Input Data Coding in Correspondence Analysis

Topics:

• Case study: Financial time series analysis
• Fuzzy coding, example
• Coding terminology, complete disjunctive form
• Introduction and example of doubling

Input Data Coding in Correspondence Analysis – F. Murtagh
In other methods, standardization by dividing by data range is usual, or dividing
Euclidean distance.

Between profiles becomes, when a particular data coding is used, the classical
input data coding is inextricably linked to the analysis. Cf. how the X² distance
input data coding is open and flexible in regard to input data types. But
Correspondence analyses is open and flexible in regard to input data types.

Correspondence analyses is open and flexible in regard to input data types.

But... Velleman and Wilkinson (1993) criticized this approach on the grounds of
method like principal components analysis should not be used.
measurement. If data were not ratio level (and not real-valued), then a metric
interval or ratio. Appropriate analysis method depended on level of
psychophysics: a measurement value was of scale type nominal, ordinal,
measurement scales introduced by S.S. Stevens in the 1940s for use in
Data analysis = questionnaire analysis (1/3)

Homogeneity: the theme of the study delimits the domain from which one collects heterogeneous sets of variables collected on different levels: qualitative, quantitative, word counts, etc. However, in practice, it is often necessary to analyze the data at the level at which one describes reality: spatio-dimensional, chemical, biological, etc.

By taking each variable as a question containing a finite set of response modes (which is strictly the case for a qualitative variable, and will be also for a quantitative variable if the interval of variation is partitioned into classes), we end up with a quasi-universal coding format: the questionnaire.
Data analysis = questionnaire analysis (2/3)

Exhaustivity: to determine thoroughly analyses how a certain level of reality is

Inertia as \( N(I) \), but with weaker moments of inertia,

Input Data Coding in Correspondence Analysis – F Murtagh
Fidelity of the geometric representation: algorithmic calculations yield tables of values and planar maps on the basis of which we recognize, as far as this is possible, the structure of a multi-dimensional object. In addition, this object has to be a faithful geometric representation of the system of properties and of the observed relations.

Universality of processing: by coding all data according to the same format, i.e., a correspondence table, one can in very different domains apply the same algorithms ...

Stability of results: in the same study different approaches seem to be possible. ...

<table>
<thead>
<tr>
<th>Subject</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSc</td>
<td>19</td>
<td>91</td>
<td>19</td>
<td>73</td>
<td>47</td>
<td>85.2</td>
</tr>
<tr>
<td>CPg</td>
<td>15</td>
<td>76</td>
<td>20</td>
<td>74</td>
<td>54</td>
<td>67.0</td>
</tr>
<tr>
<td>CGr</td>
<td>12</td>
<td>20</td>
<td>42</td>
<td>39</td>
<td>42</td>
<td>34.0</td>
</tr>
<tr>
<td>CNw</td>
<td>64</td>
<td>14</td>
<td>33</td>
<td>20</td>
<td>56</td>
<td>36.6</td>
</tr>
<tr>
<td>DbM</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>SwE</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Scores for each subject:
- CSc: 19, 91, 19, 73, 47
- CPg: 15, 76, 20, 74, 54
- CGr: 12, 20, 42, 39, 42
- CNw: 64, 14, 33, 20, 56
- DbM: 0, 0, 0, 0, 0
- SwE: 64, 0, 0, 0, 0

Mean scores: 85.2, 67.0, 34.0, 36.6, 0, 0
Scores 5 students in 6 subjects (Cont'd.)

Correspondence analysis highlights the similarities and the differences in the profiles.

Both D and E show a positive association with CNW (computer networks) and a negative association with SwE (software engineering) because in comparison with the mean profile, D and E have, in their profile, a relatively larger component of CNW and a relatively smaller component of SwE.

Note that all the scores of D and E are in the same proportion (E's scores are one-third those of D).

Note also that E has the lowest scores both in absolute and relative terms in all the subjects.

D and E have identical profiles: without data coding they would be located at the same location in the output display.

Without data coding, profiles would be located at the same location in the output display.

Correspondence analysis highlights the similarities and the differences in the profiles.
We need to clearly differentiate between the profiles of D and E, which we do by doubling the data.

Doubling: we attribute two scores per subject instead of a single score. The "score awarded" $\text{score awarded} = \frac{\text{score not awarded}}{\text{mass of the profile of } j}$ is equal to the initial score. The "score not awarded" $\text{score not awarded} = 100 - \text{score awarded}$, i.e., \(100 - k(j)\).

Lever principle: a "+" variable and its corresponding "-" variable lie on the opposite sides of the origin and collinear with it.

\(100\), the point \(j^+\) is closer to the origin than \(j^-\).

And: if the mass of the profile of \(j^+\) is greater than the mass of the profile of \(j^-\), which means that the average score for the subject \(j\) was greater than 50 out of 100, the point \(j^+\) is closer to the origin than \(j^-\).

We will find that except in CPg, the average score of the students was below 50.
Data coding: Doubling

Doubled table of scores derived from previous table. Note: all rows now have the same total.

<table>
<thead>
<tr>
<th></th>
<th>CSc+</th>
<th>CSc-</th>
<th>CPg+</th>
<th>CPg-</th>
<th>CGr+</th>
<th>CGr-</th>
<th>CNw+</th>
<th>CNw-</th>
<th>DbM+</th>
<th>DbM-</th>
<th>SwE+</th>
<th>SwE-</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>54</td>
<td>46</td>
<td>55</td>
<td>45</td>
<td>46</td>
<td>54</td>
<td>46</td>
<td>36</td>
<td>69</td>
<td>31</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>B</td>
<td>35</td>
<td>65</td>
<td>56</td>
<td>44</td>
<td>20</td>
<td>80</td>
<td>20</td>
<td>49</td>
<td>01</td>
<td>74</td>
<td>32</td>
<td>17</td>
</tr>
<tr>
<td>C</td>
<td>47</td>
<td>53</td>
<td>73</td>
<td>27</td>
<td>39</td>
<td>61</td>
<td>30</td>
<td>07</td>
<td>04</td>
<td>85</td>
<td>25</td>
<td>04</td>
</tr>
<tr>
<td>D</td>
<td>54</td>
<td>46</td>
<td>55</td>
<td>45</td>
<td>46</td>
<td>54</td>
<td>46</td>
<td>36</td>
<td>69</td>
<td>31</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>E</td>
<td>18</td>
<td>82</td>
<td>24</td>
<td>76</td>
<td>11</td>
<td>89</td>
<td>14</td>
<td>38</td>
<td>42</td>
<td>67</td>
<td>58</td>
<td>57</td>
</tr>
</tbody>
</table>

Doubled table of scores derived from previous table. Note: all rows now have the same total.
Factor 1 (77% inertia)

Factor 2 (18% inertia)

A
B
C
D
E

CSc+
CSc-
CPg+
CPg-
CGr+
CGr-
CNw+
CNw-
DbM+
DbM-
SwE+
SwE-
Coding

Terminology

- Double rescaling
- Personal equation
- Fuzzy, piecewise linear, or barycentric coding
- Complete disjunctive form
- Doubling, lever principle
- Table of scores
- Mixed qualitative and quantitative data
- Description table
- Contingency table
Responses of a set of subjects to a set of questions are coded as boolean (or logical) values. Let $C_1$ be a set of subjects, $C_9$ a set of questions, $D_5$ a set of the response categories corresponding to the question $D_5$; we suppose that the response of any subject to a question falls under one of the categories $D_5$. Let $I$ be a set of subjects, $\tilde{I}$ a set of questions, $b$ a set of the response categories corresponding to the question $b$; we suppose that the response of any subject $i$ to the question $b$ falls under one of the categories $C_2$.

To each question $b$ corresponds a block $b$ of columns of the data table. With each individual, a row of the data table is associated. Each $\tilde{I}$ is the table of responses. $I$ is the union of all the $b$ for $b$ belonging to $\tilde{I}$, i.e. $I = \bigcup_{b \in \tilde{I}} b$. $C_9$ is the set of all the response categories pertaining to all the questions. $C_2$ is the union of all the $C_2$ for $b \in \tilde{I}$. $C_1 \cap C_2$ is the table of responses. With each individual, a row of the data table is associated. $C_1 \cap C_2$ is the table of responses. With each individual, a row of the data table is associated.

Complete disjunctive form
Remark on Burt table. The analysis of a table $f \times I$ in complete disjunctive form furnishes for the set of categories $f$ principal coordinates which (within a constant coefficient) are the same as those obtained by analyzing the Burt table $k_i f_i j$. We have: $f_i j = (f_i k_j, y_i k_j)$. The number of individuals $I$ belonging simultaneously to both the categories $k_i$ and $k_j$ of the table $k_i f_i j$ is therefore equal to the number of questions.

The total of each row of the table $f_i$ is therefore equal to the number of
Fuzzy coding (1/2)

Hinges or pivots

Value of v

Hinge 2

Hinge 3

180

200

235
The value 200 is therefore coded as (0, 35/55, 20/55) = (0, 0.64, 0.36).

From the third hinge 235 up to 1, the value 200 is at 20/55 units from the second hinge 180, and 35/55 units from the third hinge 235.

Hinges: (125, 180, 235)

Shown above are hinges $v_1 = 180$ and $v_2 = 235$.

Hinges: (125, 180, 235)

How will the value $v = 200$ be coded?

Pieces between the second and the third hinges can be considered as the category, $v_1$, is zero.

The value 200 lies between the second and the third hinges, therefore the first category, $v_1$, is zero.

The value 200 lies between the second and the third hinges, therefore the first category, $v_1$, is zero.

The value 200 lies between the second and the third hinges, therefore the first category, $v_1$, is zero.

The value 200 lies between the second and the third hinges, therefore the first category, $v_1$, is zero.

Hinges used in piecewise linear (or fuzzy, or barycentric) coding.

Fuzzy coding (2/2)
Test of fuzzy coding on Fisher iris data
Histograms of variables of iris data.
Principal components analysis of Fisher Iris data

Principal component 1 (92.5% variance)

Principal component 2 (5.3% variance)
Correspondence analysis of iris data, fuzzy coding, 3 pivots

Factor 1 (34.1% inertia)

Factor 2 (16.0% inertia)

CA of fuzzy coded iris data, 3 pivots
Correspondence analysis of booleanized iris data
The practice of correspondence analysis has however established that we gain by considering the mean of the scores attributed by a given subject as the zero-point of the scale adopted by him, in order to use this zero for rescaling the scores between the minimum (min) and maximum (max) of all the scores attributed by him or her.

For each subject, the rescaling between the scores attributed by a given subject as the zero-point of the scale adopted by him, in order to use this zero for rescaling the scores between the minimum (min) and maximum (max) of all the scores attributed by him or her is done by computing their mean (ave), maximum (max) and minimum (min). The scores are first centred by subtracting the mean from them. Then all the positive scores are divided by (max - ave) and all the negative scores are divided by (ave - min). The rescaled scores are then given by the formula:

\[ f(i, j) = \frac{A(i) - \bar{A}}{A(i) - \bar{A}} \]

where \( f(i, j) \) is the rescaled score for the subject \( j \), \( A(i) \) is the score attributed by the subject \( i \), and \( \bar{A} \) is the mean of all scores attributed by the subject \( i \).

Now let \( f(i, j) \) be a rescaled score; we code it across three categories by applying the formula:

\[ \text{Personal equation:} \]

from \(-1\) to \(+1\).

For each subject, the rescaling between the scores attributed by a given subject as the zero-point of the scale adopted by him, in order to use this zero for rescaling the scores between the minimum (min) and maximum (max) of all the scores attributed by him or her is done by computing their mean (ave), maximum (max) and minimum (min). The scores are first centred by subtracting the mean from them. Then all the positive scores are divided by (max - ave) and all the negative scores are divided by (ave - min). The scores given by the subject vary from \(-1\) to \(+1\).
If we used the min, ave, max of each row as the hinges for barycentrically coding all the scores in that row, it is easy to recognize a barycentric principle in this coding since the same result is achieved if we used the min, ave, max of each row.

```
end if

0 = (-\(j', t\)) \(k\)

(\(j', t\)) \(k - t\) = (\(= j', t\)) \(k\)

(\(j', t\)) \(k + t\) = (\(+ j', t\)) \(k\)
```
Double scaling

coherence of the results after each transformation. It should however be borne in mind that the larger the number of transformations effected on the data, the more circumspect one should be. One of the ways of ensuring that the coding does not distort the data is to check the use of the personal equation on both I and J.

Here, too, it is a barycentric coding.
Some conclusions for the financial case study to follow

Using categorical or qualitative coding may allow structure, imperceptible with quantitative data, to be discovered.

Quantile-based categorical coding (i.e., the uniform prior case) has beneficial properties.

In the case of time-varying data signals (which also holds for spatial data), quantile-based categorical coding may allow structure, imperceptible with quantitative data, to be discovered.

An appropriate coding granularity, or scale of problem representation, should be sought.

Structures (or models of associations or relationships) found in training data are validated on unseen test data, but if a data set consistently supports or respects non-respect of stationarity should be checked for: the mutatis mutandis (mutatis mutandis) consistency of our results will inform us about stationarity present in our data.

These structures then a fortiori leaving-k-out cross-validation is achieved.

An appropriate coding granularity, or scale of problem representation, should be sought.

Quantile-based categorical coding (i.e., the uniform prior case) has beneficial properties.
Departure from average behavior is made easy in the analysis framework adopted. This amounts to fingerprinting the data, i.e., determining patterns in the data that are characteristic of it.
As in the efficient market hypothesis, in Brownian motion a future price
independent of all \( y_i \) for \( t \geq t + 1 \) and follows a Gaussian distribution.

Brownian motion: for \( 0 \leq t \leq \infty \), the variable \( y_{t+1} \) is

Differenced values of the time series with constant time steps are studied
predictable from a historical time series of these prices.

An implication of the efficient market hypothesis is that price changes are not
martingales (Doob, 1953).

An implication of the efficient market hypothesis is that price changes are not

When stochastic processes satisfy this conditional probability, they are termed

\[ y_t = \{ y_0, \ldots, y_t \} \]

\[ y_{t+1} \]

follows.

Efficient market hypothesis (Samuelson, 1965): if \( y_t \) is the value of a financial
asset, then the expected value at time \( t + 1 \) is related to previous values as
In input data coding in correspondence analysis – F Murtagh

In Brownian motion, price difference is Gaussian. Furthermore, in geometric Brownian motion, the variable is not dependent on any past price, and hence the price difference is Gaussian.

For all \( t \), a histogram of \( \tau T \) is \( \tau T = (\tau)D \), and then

\[
\frac{(\tau)D}{\log} = (\tau)T \quad \text{and then } \quad \log \quad \text{and}
\]

Considering monthly oil price values, price values, \( P \), and then

Therefore the ratio of price \( P \) to present price \( P \) follows a lognormal distribution, and is independent of all past prices. With drift \( \mu \) and volatility \( \sigma \), geometric Brownian motion satisfies Gaussian.

These difficulties with Brownian motion in financial time series are avoided.
To address the issue of number of coding states to use, in order to search for latent structure in such data, one approach that seems very reasonable is to explore the dependencies and associations based on fine-grained structure; and to include in this exploration the possible aggregation of the fine-grained states.

Under geometric Brownian motion, one would expect constant percentages. This is not what is found. Instead there is appreciable structure in the contingency table. Instead of being expressed as a percentage, for $y_i$ is determined, the cross-tabulation can be expressed as a percentage. Then a cross-tabulation of states 1 through 4, for $y_i^{t+1}$, against states 1 through 4, latter.
Use of Correspondence Analysis

Period of the data stream:

• In addition we will take into account possible non-stationarity over the time

Later in practical application:

• Less influential coding categories are sought in order, possibly, to bypass them

• Less influential coding categories are sought in order, possibly, to bypass them

• Less influential coding categories are sought in order, possibly, to bypass them

• Less influential coding categories are sought in order, possibly, to bypass them

• Less influential coding categories are sought in order, possibly, to bypass them

To aggregate the fine-resolution coding categories used, we need strongly

• To aggregate the fine-resolution coding categories used, we need strongly

• To aggregate the fine-resolution coding categories used, we need strongly

• To aggregate the fine-resolution coding categories used, we need strongly

• To aggregate the fine-resolution coding categories used, we need strongly

We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.

• We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.

• We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.

• We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.

• We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.

• We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.

• We use an overly fine-grained set of coding categories, so that a satisfactory

correspondence analysis.
Generalizing the leave-one-out approach to validation, we will seek consistency of results obtained for sub-intervals. If we can experimentally show that all possible sufficiently-sized sub-intervals of the time series manifest the same results, then a fortiori we are exemplifying how unseen data will behave.
Consistency for 4 different time series intervals (i,m,k,n)
Table crossing clusters (on $I$) and coordinates and contributions (as thousandths).

Clusters retained here: 65, 68, 69, 70, 71, 72, 73.

Clusters 65 through 73 represent the input coding categories.

Clusters 6 through 72 represent the output coding categories.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Coordinates</th>
<th>Predominant:</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>j1, j2, ... j10.</td>
<td>none</td>
</tr>
<tr>
<td>66</td>
<td>k9, m7, n7, i4, m9</td>
<td>none</td>
</tr>
<tr>
<td>67</td>
<td>k6, m6, n6, i4, m6</td>
<td>none</td>
</tr>
<tr>
<td>68</td>
<td>k3, m3, n2, i2, m2</td>
<td>3, 2</td>
</tr>
<tr>
<td>69</td>
<td>i8, m7, n5, i2, m2</td>
<td>none</td>
</tr>
<tr>
<td>70</td>
<td>i10, m10, i9, k10, m10</td>
<td>8</td>
</tr>
<tr>
<td>71</td>
<td>i1, m1, k1, n1, m1</td>
<td>1</td>
</tr>
<tr>
<td>72</td>
<td>i5, m5, n3, i5, m5</td>
<td>5</td>
</tr>
</tbody>
</table>

Clusters 65 through 73 represent the input coding categories.
Top of hierarchy agglomerations for output coding categories:

( ( ( 14 15 ) 16 ) 14 )

Top output data coding in Correspondence Analysis – F. Murtagh
Some conclusions from the financial case study

Coding allows us to find structure (patterns) in data which would not otherwise be found.

How can this work? We are adding semantic information to the data. Cf. earlier quotation from Benzécri: to say that a patient has a temperature of 36.9 degrees is really only meaningful in relation to what is expected or normal. Additionally, an interpretation leading to a decision is based on additional semantics.

We have again the multiple perspectives provided by the $\chi^2$ and Euclidean metrics, and ultrametric.

Cor. analysis provides a multi-modal, multi-faceted analysis toolbox.

Studying clusters on $I$ and $J$ is one way to address the question: What is the most appropriate resolution scale for analyzing the given problem?

VACOR is a way to study clusters of observations, and clusters of variables.

We have again the multiple perspectives provided by the $\chi^2$ and Euclidean metrics, and ultrametric.

An interpretation leading to a decision is based on additional semantics. Additionally, it really only meaningful in relation to what is expected or normal. Additionally, quotation from Benzécri: to say that a patient has a temperature of 36.9 degrees...