Input Data Coding in Correspondence Analysis

Topics:

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

Introduction

- Measurement scales introduced by S.S. Stevens in the 1940s for use in psychophysics: a measurement value was of scale type nominal, ordinal, interval or ratio. Appropriate analysis method depended on level of measurement. If data were not ratio level (and not real-valued), then a metric method like principal components analysis should not be used.
- But... Velleman and Wilkinson (1993) criticized this approach on the grounds of being irrelevant in practice.
- Correspondence analysis is open and flexible in regard to input data types. But input data coding is inextricably linked to the analysis. Cf. how the χ^2 distance between profiles becomes, when a particular data coding is used, the classical Euclidean distance.
- In other methods, standardization by dividing by data range is usual, or dividing centred data values by the standard deviation. In corr. analysis: data coding.

Data analysis = questionnaire analysis (1/3)

Homogeneity: the theme of the study delimits the domain from which one collects the data. The point of view of the study fixes the form of this data. That is to say, it fixes the level at which one describes reality: spatial dimensions, chemical composition, word counts, etc. However in practice it is often necessary to analyze heterogeneous sets of variables collected on different levels: qualities, integers; continuous quantities of different natures or orders of magnitudes. One tries in such cases to arrive at some measure of homogeneity using mathematical transformation or coding. 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 through analysis how a certain level of reality is ordered, in accordance with which axes it is necessary to have taken this level in its totality, or at least to have extracted a sample of uniform density. From this point of view, Louis Guttman considered every finite questionnaire as an extraction from a continuous universe of possible questions: hence the importance of continuous models. [...] We approximate an exhaustive description by a nomenclature which is more and more fine-grained. [...] The principle of distributional equivalence guarantees that cumulative rows or columns of neighbouring profiles in a table changes the results very little. What is more, if starting with a cloud $N(I)$ we form aggregated rows or columns arbitrarily (which therefore could include distant rows or columns), the cloud of centres of these aggregates has the same principal axes of inertia as $N(I)$, but with weaker moments of inertia.

Data analysis = questionnaire analysis (3/3)

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 multidimensional object $N(I)$. 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 analysis algorithms ...

Stability of results: ... in the same study different approaches seem to be possible. It is particularly satisfactory if all approaches point in the same direction, and give similar results. For this reason, it may be useful to consider a number of codings of the same set I, for a given cloud $N(I)$.

Scores 5 students in 6 subjects

	CSc	CPg	CGr	CNw	DbM	SwE
A	54	55	31	36	46	40
B	35	56	20	20	49	45
C	47	73	39	30	48	57
D	54	72	33	42	57	21
E	18	24	11	14	19	7

	CSc	CPg	CGr	CNw	DbM	SwE
mean profile:	.18	.24	.12	.12	.19	.15
profile of D:	.19	.26	.12	.15	.20	.08
profile of E:	.19	.26	.12	.15	.20	.08

Scores (out of 100) of 5 students, A–E, in 6 subjects. Subjects: CSc: Computer Science Proficiency, CPg: Computer Programming, CGr: Computer Graphics, CNw: Computer Networks, DbM: Database Management, SwE: Software Engineering.

Scores 5 students in 6 subjects (Cont'd.)

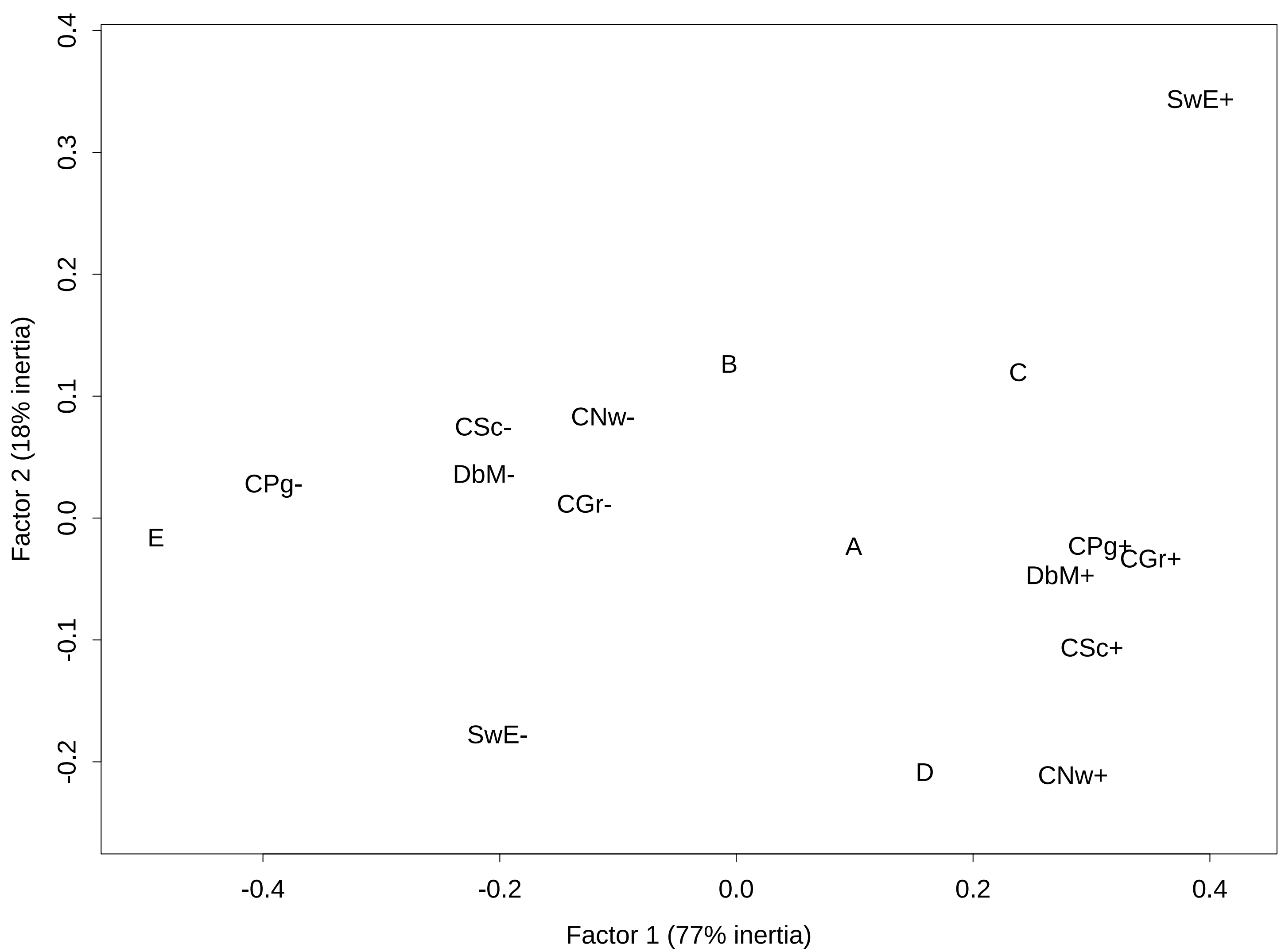
- Correspondence analysis highlights the similarities and the differences in the profiles.
- Note that all the scores of D and E are in the same proportion (E's scores are one-third those of D).
- Note also that E has the lowest scores both in absolute and relative terms in all the subjects.
- D and E have identical profiles: without data coding they would be located at the same location in the output display.
- 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`.

- 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", $k(i, j^+)$, is equal to the initial score. The "score not awarded", $k(i, j^-)$, is equal to its complement, i.e., $100 - k(i, j^+)$.
- Lever principle: a "$+$" variable and its corresponding "$-$" variable lie on the opposite sides of the origin and collinear with it.
- 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 in all the subjects.

Data coding: Doubling

	CSc+	CSc-	CPg+	CPg-	CGr+	CGr-	CNw+	CNw-	DbM+	DbM-	SwE+	SwE-
A	54	46	55	45	31	69	36	64	46	54	40	60
B	35	65	56	44	20	80	20	80	49	51	45	55
C	47	53	73	27	39	61	30	70	48	52	57	43
D	54	46	72	28	33	67	42	58	57	43	21	79
E	18	82	24	76	11	89	14	86	19	81	7	93

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


Factor 2 (18% inertia)
Factor 1 (77% inertia)
-0.4
-0.2
0.0
0.2
0.4
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
SwE+
B
C
CSc-
CNw-
CPg-
DbM-
CGr-
E
A
CPg+
CGr+
DbM+
CSc+
SwE-
D
CNw+

Coding: Terminology

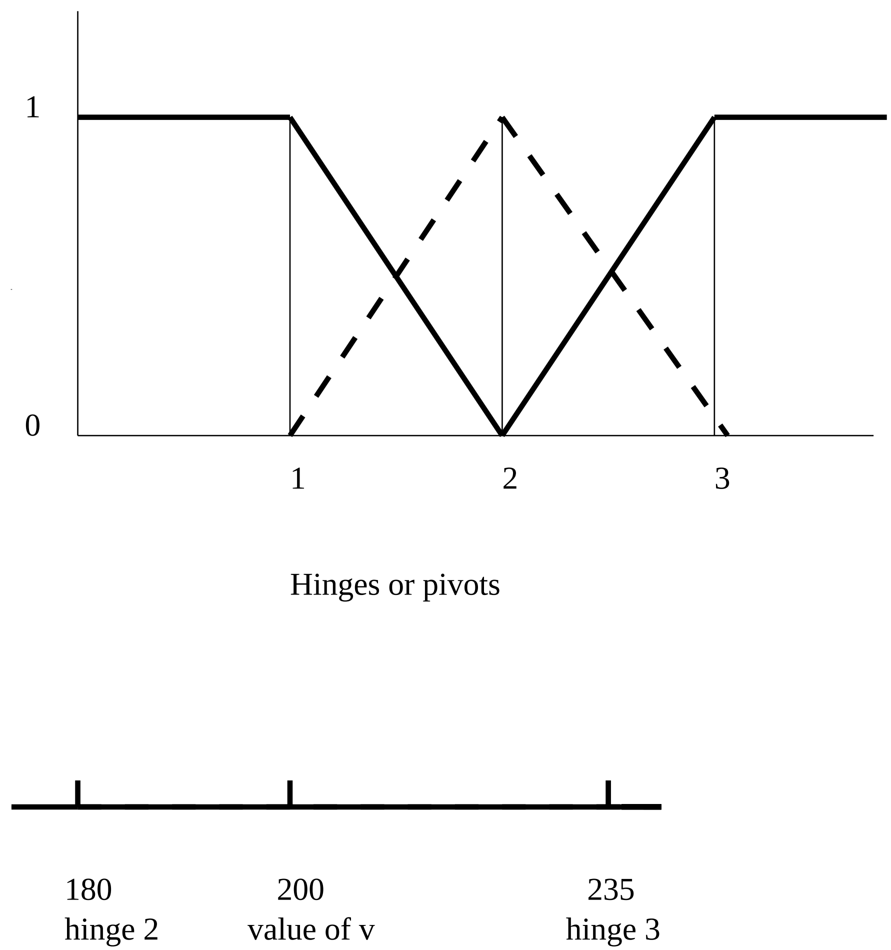
- Contingency table
- Description table
- Mixed qualitative and quantitative data
- Table of scores
- Doubling, lever principle
- Complete disjunctive form [looked at next...]
- Fuzzy, piecewise linear, or barycentric coding [Also looked at next...]
- Personal equation
- Double rescaling

Complete disjunctive form

- Responses of a set of subjects to a set of questions are coded as boolean (or logical) values.
- Let I be a set of subjects i, Q a set of questions q, J_q a set of the response categories corresponding to the question q; we suppose that the response of any subject to a question q falls under one of the categories J_q.
- J is the union of all the J_q, for q belonging to Q, i.e. J is the set of all the response categories pertaining to all the questions.
- k_{IJ} is the table of responses. With each individual, a row of the data table is associated.
- To each question q there corresponds a block J_q of columns. $k(i, j) = 1$ if the subject i chooses the category j, and zero otherwise. Hence in the row i in each block J_q there is a 1 in the column pertaining to the response category j chosen by the subject for the question q, and zeros elsewhere.

- The total of each row of the table k is therefore equal to the number of questions.
- Remark on Burt table. The analysis of a table $I \times J$ in complete disjunctive format furnishes for the set of categories J principal coordinates which (within a constant coefficient) are the same as those obtained by analyzing the Burt table ${k'}_{JJ}$. We have: $k'(j, j') =$ the number of individuals i of I belonging simultaneously to both the categories j and j'. The Burt table is a true contingency table.

Fuzzy coding (1/2)



Fuzzy coding (2/2)

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

Hinges: (125, 180, 235)

Shown above are hinges $v_2 = 180$ and $v_3 = 235$.

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

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

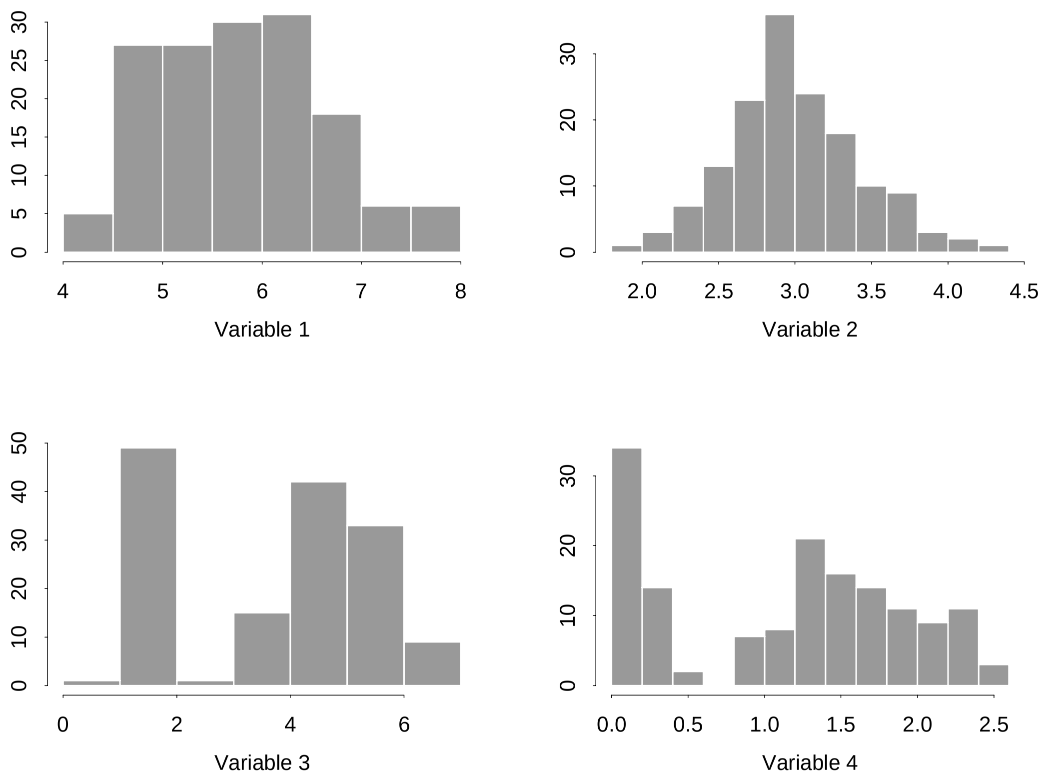
The value 200, lying between the middle and last hinges, can be considered as the barycentre (weighted average) of these two hinges with appropriate masses adding up to 1. The value 200 is at 20/55 units from the second hinge 180, and 35/55 units from the third hinge 235.

The value 200 is therefore coded as (0, 35/55, 20/55) = (0, .64, .36).

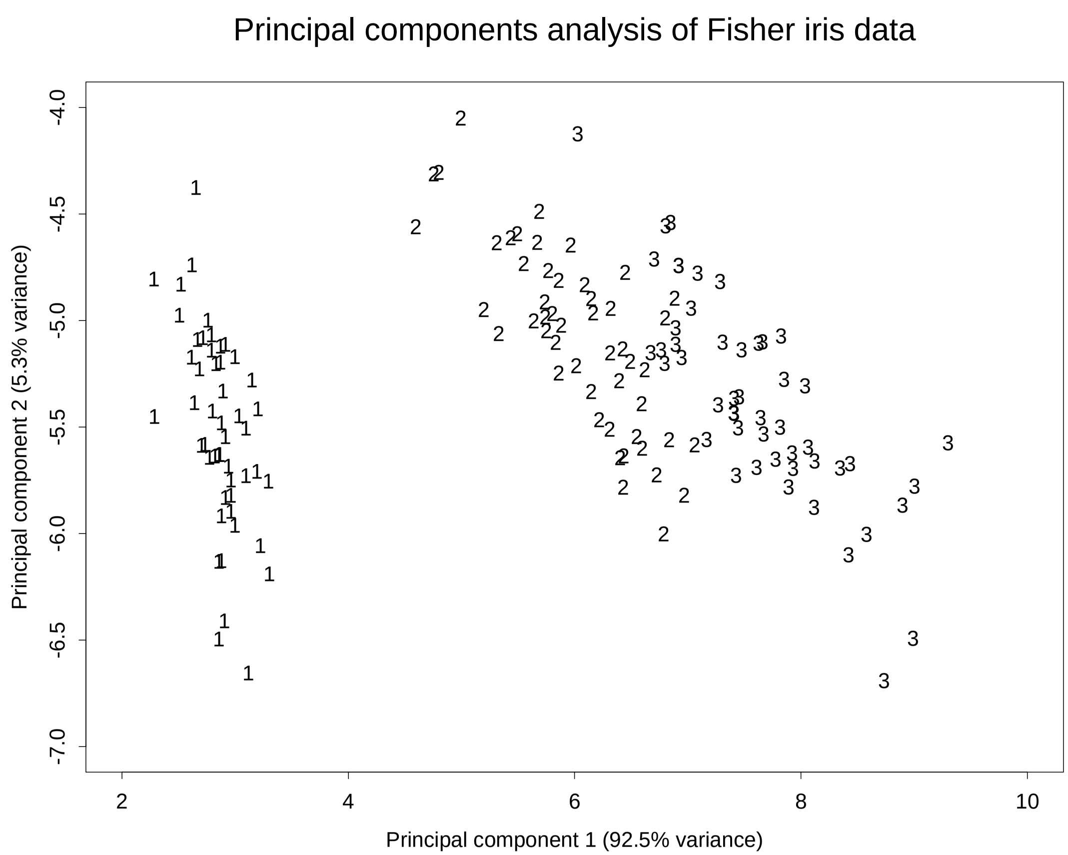
Test of fuzzy coding on Fisher iris data

- Fisher iris data (Fisher, 1936).
- 150 observations (iris flower), 4 measurements on each (sepal and petal width and length). Real values.
- Class 1 = obs. 1–50 well distinguished from others.
- Classes 2 and 3 = resp. obs. 51–100, and 101–150.
- We used principal components analysis on given data, with standardization to zero mean and unit variance for the variables.
- We also employed a fuzzy coding with two pivots at the 33rd and 66th percentiles. (Why this choice? To have equal weighting in each category.)
- One motivation for such fuzzy coding: multimodality in histograms of the variables.
- Figures to follow. We conclude: fuzzy coding is competitive...

Histograms of variables of iris data

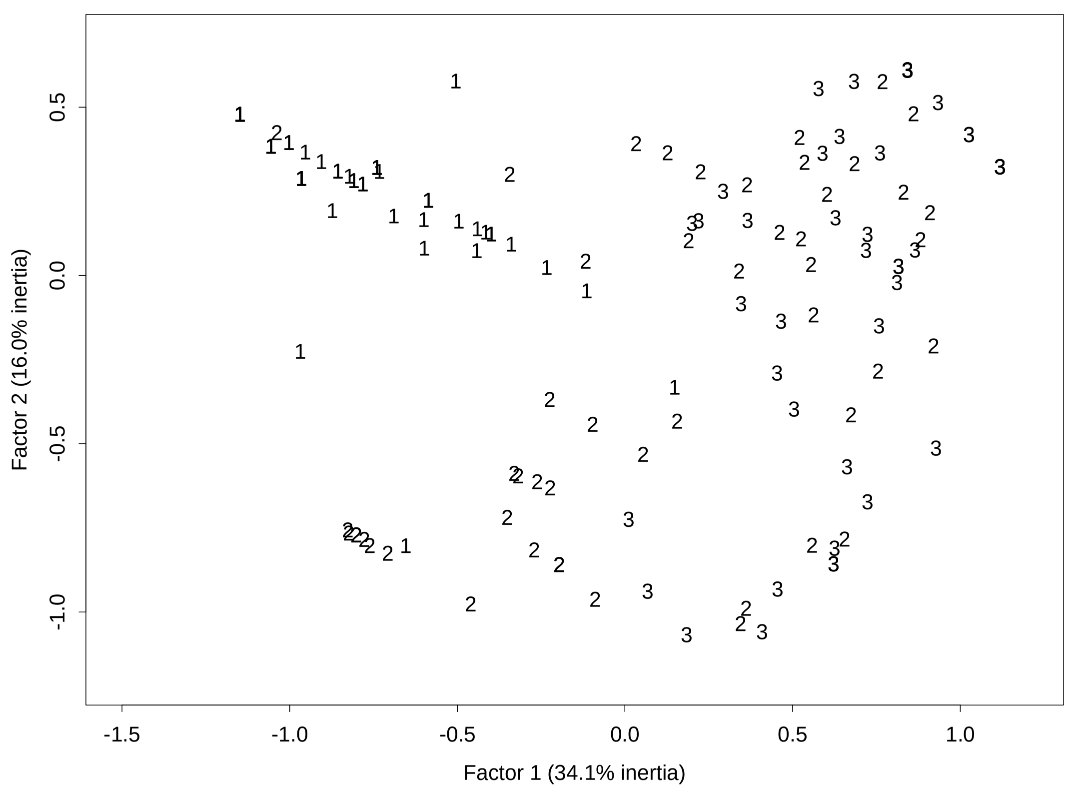


PCA principal plane of iris data

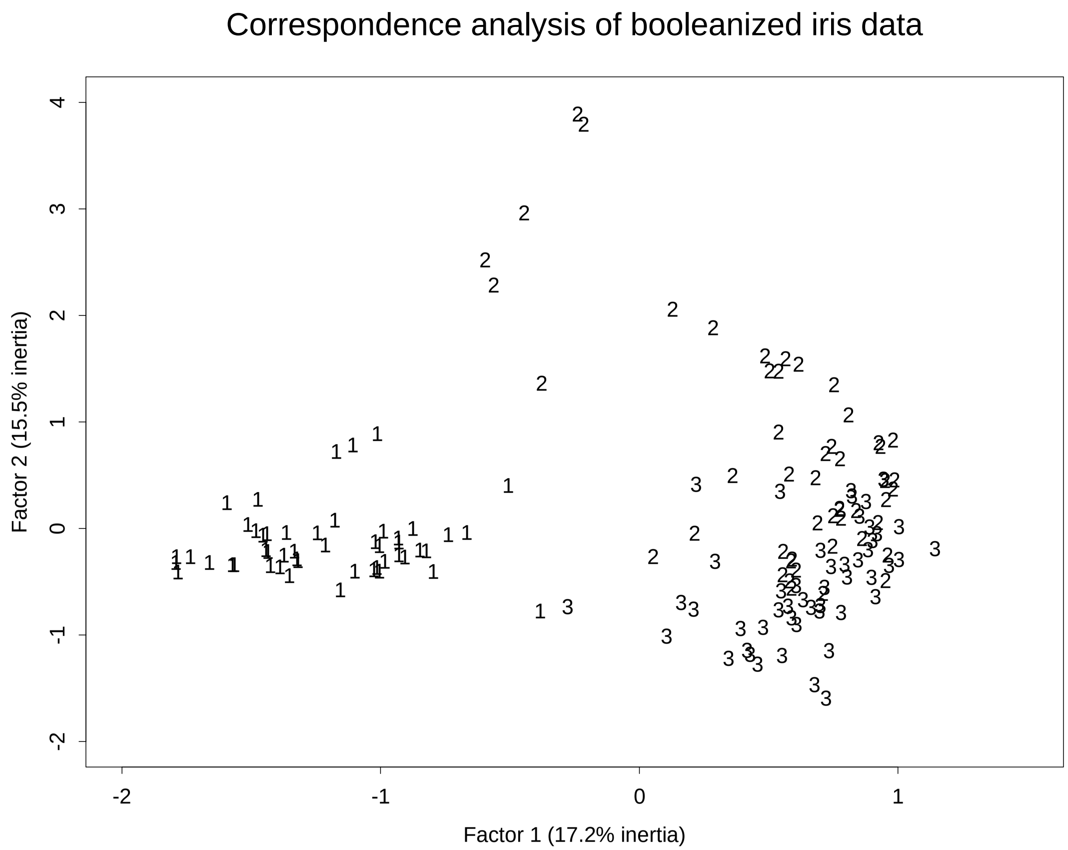


CA of fuzzily coded iris data, 3 pivots

Correspondence analysis of iris data, fuzzy coding, 3 pivots



CA of 123-dimensional booleanized iris data



Personal equation

- “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 -1 and $+1$” (Benzécri, 1989c). This is done using a formula known as personal equation, particular to each subject.
- For each subject i, the rescaling between -1 and $+1$ of all the scores attributed by him or her is done by computing their mean ($\texttt{ave}$), maximum ($\texttt{max}$) and minimum ($\texttt{min}$). The scores are first centred by subtracting the mean from them. Then all the positive scores are divided by ($\texttt{max} - \texttt{ave}$); all the negative scores are divided by ($\texttt{ave} - \texttt{min}$); thus the scores given by the subject i vary from -1 to $+1$.
- Now let $k(i, j)$ be a rescaled score; we code it across three categories by applying the formula:

```
if k(i,j) <= 0 then
        k(i,j+) = 0
        k(i,j=) = 1+k(i,j)
        k(i,j-) = k(i,j)
else
        k(i,j+) = k(i,j)
        k(i,j=) = 1-k(i,j)
        k(i,j-) = 0
endif
```

- 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 i as the hinges for barycentrically coding all the scores in that row.

Double scaling

- Use of the personal equation on both I and J.
- Here, too, it is a barycentric coding.
- “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 coherence of the results after each transformation.”

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.
- An appropriate coding granularity, or scale of problem representation, should be sought.
- In the case of a time-varying data signal (which also holds for spatial data, *mutatis mutandis*) non-respect of stationarity should be checked for: the consistency of our results will inform us about stationarity present in our data.
- Structures (or models or associations or relationships) found in training data are validated on unseen test data. But if a data set consistently supports or respects these structures then *a fortiori* leaving-k-out cross-validation is achieved.

- Departure from average behavior is make easy in the analysis framework adopted. This amounts to fingerprinting the data, i.e. determining patterns in the data that are characteristic of it.

Efficient market hypothesis and geometric Brownian motion

- Efficient market hypothesis (Samuelson, 1965): if y_i is the value of a financial asset, then the expected value at time $t+1$ is related to previous values as follows.

$$E\{y_{t+1} \mid y_0, y_1, \ldots y_t\} = y_t$$

- When stochastic processes satisfy this conditional probability, they are termed martingales (Doob, 1953).
- An implication of the efficient market hypothesis is that price changes are not predictable from a historical time series of these prices.
- Differenced values of the time series with constant time steps are studied through Brownian motion: for $0 \leq i < \infty$, the variable $y_{t+1} - y_t$ is independent of all $y_i, i < t$, and follows a Gaussian distribution.
- As in the efficient market hypothesis, in Brownian motion a future price

depends only on the present price, and not at all on the past prices. Furthermore in Brownian motion, price difference is Gaussian.

- These difficulties with Brownian motion in financial time series are avoided with geometric Brownian motion. In geometric Brownian motion, the variable y_{t+1}/y_t is not dependent on any $y_i, i < t$, and $\log(y_{t+1}/y_t)$ is Gaussian. Therefore the ratio of price y_{t+1} to present price y_t follows a lognormal distribution, and is independent of all past prices. With drift μ and volatility σ, geometric Brownian motion satisfies $E\{y_t\} = y_0 \exp t(\mu + \sigma^2/2)$.
- Using crude oil data, Ross (2003) shows how structure can be found in apparently geometric Brownian motion, through data recoding.
- Considering monthly oil price values, $P(i)$, and then $L(i) = \log(P(i))$, and finally $D(i) = L(i) - L(i-1)$, a histogram of $D(i)$ for all i should approximate a Gaussian.
- The following recoding, though, gives rise to a somewhat different picture: response categories or states 1, 2, 3, 4 are used for values of $D(i)$ less than or equal to -0.01, between the latter and 0, from 0 to 0.01, and greater than the

latter.

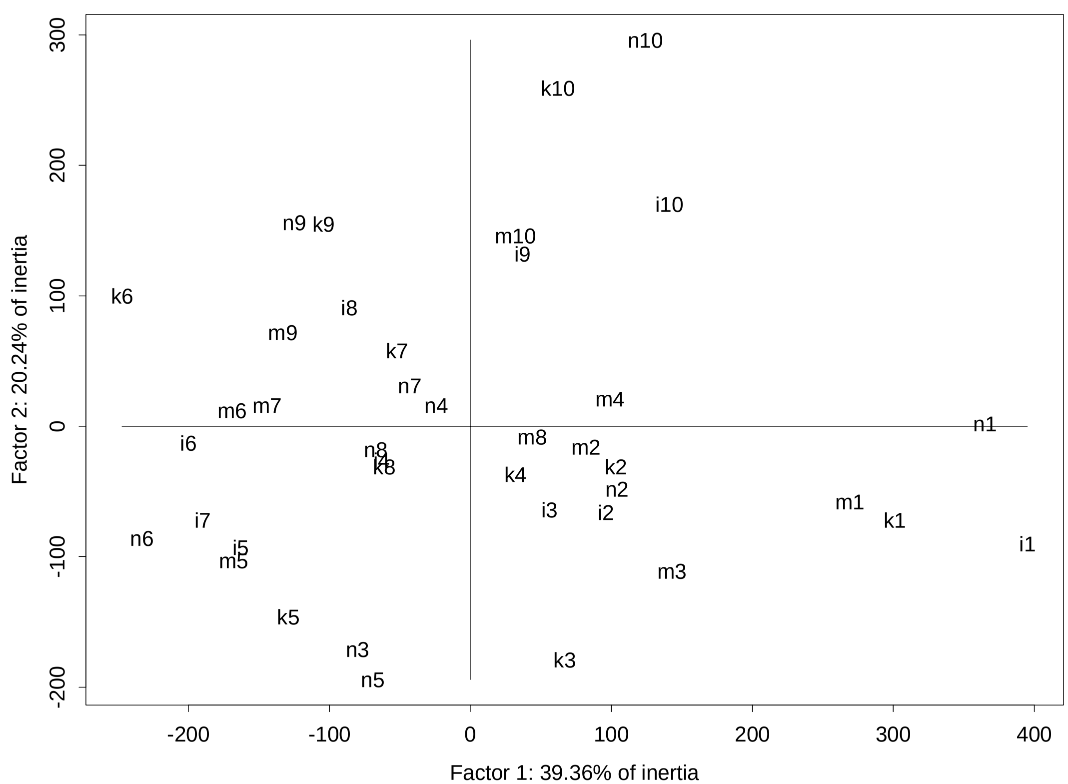
- Then a cross-tabulation of states 1 through 4 for y_{t+1}, against states 1 through 4 for y_t, is determined. The cross-tabulation can be expressed as a percentage. Under geometric Brownian motion, one would expect constant percentages. This is not what is found. Instead there is appreciable structure in the contingency table.
- 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 include in this exploration the possible aggregation of the fine-grained states.

Use of correspondence analysis

- We use quantile coding motivated (i) by the desire on our part to find structure in Brownian motion signals, and (ii) by the fact that it lends itself well (in that it furnishes a uniform mass density) to the analysis and display properties of correspondence analysis.
- We use an overly fine-grained set of coding categories, so that a satisfactory outcome (a *satisficing* solution in scheduling terminology) is obtained by aggregating these categories.
- To aggregate the fine-resolution coding categories used, we need strongly associated coding categories.
- Less influential coding categories are sought in order, possibly, to bypass them later in practical application.
- In addition we will take into account possible non-stationarity over the time period of the data stream.

- Generalizing the leaving-k-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)



VACOR for atypical price movements

Table crossing clusters (on I) and coordinates (J), giving correlations and contributions (as thousandths). Clusters retained here: 65, 68, 69, 70, 71, 72, 73. Coordinates: j1, j2, ... j10.

```
Top of hierarchy agglomerations:
( ( 65 ( 73 ( 69 71 ) ) ) ( 70 ( 68 72 ) ) )

Cluster 65: k9 n9 k7 n7 i4 m9              Predominant: 9
Cluster 68: i3 k3 m3 m4 i2 m2 k2 n2        Predominant: 2, 3
Cluster 69: n6 i8 m7                       Predominant: none
Cluster 70: i10 m10 i9 k10 n10             Predominant: 10
Cluster 71: i6 k4 n4 m8 k8 n8              Predominant: 8
Cluster 72: i1 m1 k1 n1                    Predominant: 1
Cluster 73: i5 m5 n3 k5 n5 k6 i7 m6        Predominant: 5

Clusters 65 through 73 represent the input coding categories.
Coordinates j1 through j10 represent the output coding categories.
```

```
Top of hierarchy agglomerations for output coding categories:
( 1 ( 16 ( 14 15) ) )

          j1         j14        j15        j16

          j1      j2,j3,j7  j8,j9,j10  j4,j5,j6

       very low     low,     high/       middle
                  spoiled   very high

        COR CTR    COR CTR    COR CTR    COR CTR
   65|  734 131|     4   4|  260 114|      2   0|
   68|  201  52|   399 583|  264 169|    137  52|
   69|  210  36|   261 250|    2   1|    527 132|
   70|    4   2|    26  57|  568 543|    402 229|
   71|  299   7|   239  32|   17   1|    445  15|
   72|  784 661|     8  37|   29  60|    179 221|
   73|  277 112|    17  38|  114 113|    592 349|
```

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 χ^2 and Euclidean metrics, and ultrametric.
- VACOR is a way to study clusters of observations, and clusters of variables.
- 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?
- Corr. analysis provides a multi-modal, multi-facteted analysis toolbox.