Correspondence Analysis

Topics:

- Basics, and preliminary example (student exam scores)
- Metrics, clouds of points, masses, inertia
- Factors, decomposition of inertia, contributions, dual spaces
- Hierarchical agglomerative clustering
- Minimum variance criterion
- Examples in depth (ppt file)
- Java application: http://astro.u-strasbg.fr/~fmurtagh/mda-sw

Basics

- Observations $\times$ variables matrix.
- Through display and through quantitative measures, investigate relationships between observations, and between variables.
- Similar in these objectives to principal components analysis, multidimensional scaling, Kohonen self-organizing feature map, and others.
- Correspondence analysis is often used in conjunction with clustering.
- Input data, and input data coding, are the major issues which distinguish correspondence analysis from other algorithmically-similar (or alternative algorithmic) methods.

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.)

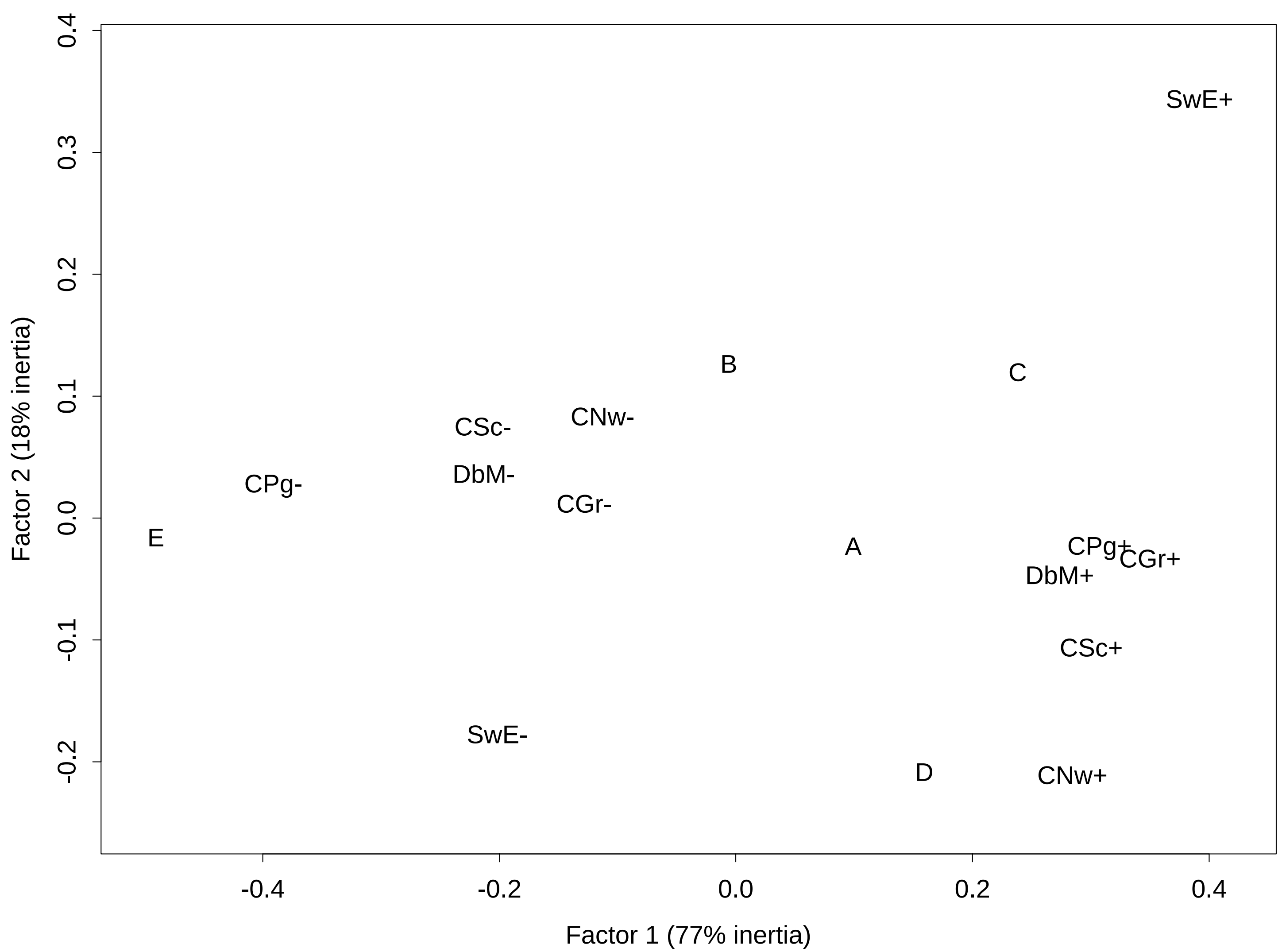
- Correspondence analysis highlights the similarities and the differences in the profiles.
- Note that all the scores of D and E are in the same proportion (E's scores are one-third those of D).
- Note also that E has the lowest scores both in absolute and relative terms in all the subjects.
- D and 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.


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

Metrics

- The notion of distance is crucial, since we want to investigate relationships between observations and/or variables.
- Recall: $x = \{3, 4, 1, 2\}, y = \{1, 3, 0, 1\}$, then: scalar product $\langle x, y \rangle = \langle y, x \rangle = x'y = xy' = 3 \times 1 + 4 \times 3 + 1 \times 0 + 2 \times 1$.
- Euclidean norm: $\|x\|^2 = 3 \times 3 + 4 \times 4 + 1 \times 1 + 2 \times 2$.
- Euclidean distance: $d(x, y) = \|x - y\|$. The squared Euclidean distance is: $3 - 1 + 4 - 3 + 1 - 0 + 2 - 1$
- Orthogonality: x is orthogonal to y if $\langle x, y \rangle = 0$.
- Distance is symmetric ($d(x, y) = d(y, x)$), positive ($d(x, y) \geq 0$), and definite ($d(x, y) = 0 \Longrightarrow x = y$).

Metrics (cont'd.)

- Any symmetric, positive, definite matrix M defines a generalized Euclidean space. Scalar product is $\langle x, y \rangle_M = x' M y$, norm is $\|x\|^2 = x' M x$, and Euclidean distance is $d(x, y) = \|x - y\|_M$.
- Classical case: $M = I_n$, the identity matrix.
- Normalization to unit variance: M is diagonal matrix with ith diagonal term $1/\sigma_i^2$.
- Mahalanobis distance: M is inverse variance-covariance matrix.
- Next topic: Scalar product defines orthogonal projection.

Metrics (cont'd.)

- Projected value, projection, coordinate: $x_1 = (x'Mu/u'Mu)u$. Here x_1 and u are both vectors.
- Norm of vector $x_1 = (x'Mu/u'Mu)\|u\| = (x'Mu)/\|u\|$.
- The quantity $(x'Mu)/(\|x\|\|u\|$ can be interpreted as the cosine of the angle a between vectors x and u.

```
      +  x
     /|
    / |
   /  |
  /   |
 /a   |
+-----+-----  u
O     x1
```

Metrics (cont'd.)

- Consider the case of centred n-valued coordinates or variables, x_i.
- The sum of variable vectors is a constant, proportional to the mean variable.
- Therefore the centred vectors lie on a hyperplane H, or a sub-space, of dimension $n - 1$.
- Consider a probability distribution p defined on I, i.e. for all i we have $p_i > 0$ (note: > 0 to avoid inconvenience of lower dim. subspace) and $\sum_{i \in I} p_i = 1$.
- Covariance matrix: M_{p_I}, diagonal matrix with diagonal elements consisting of the p terms.
- Have: $x' M_{p_I} x = \sum_{i \in I} p_i x_i^2 = \text{var}(x)$; and $x' M_{p_I} y = \sum_{i \in I} p_i x_i y_i = \text{cov}(x, y)$.

Metrics (cont'd.)

- Use of metric M_{p_I} on I is associated with the following χ^2 distance relative to centre p_I.
- This new distance is a generalized Euclidean M_{1/p_I} metric.
- Let both p_I and r_I be probability densities.
- Then: $\|p_{IJ} - q_{IJ}\|^2_{q_{IJ}} = \sum_{(i,j) \in I \times J} (p_{ij} - p_i p_j)^2 / p_i p_j$.
- Link with χ^2 statistic: let p_{IJ} be a data table of probabilities derived from frequencies or counts. $p_{IJ} = \{p_{ij} | i \in I, j \in J\}$.
- Marginals of this table are p_I and p_J. Consider independence of effects where the data table is $q_{IJ} = p_I p_J$.
- Then the χ^2 distance of centre q_{IJ} between the densities p_{IJ} and q_{IJ} is $\|p_{IJ} - q_{IJ}\|^2_{q_{IJ}} = \sum_{(i,j) \in I \times J} (p_{ij} - p_i p_j)^2 / p_i p_j$.

- With the coefficient $\sqrt{n}$, this is the quantity which can be assessed with a χ^2 test with $n - 1$ degrees of freedom.
- The χ^2 distance is used in correspondence analysis.
- Clearly, under appropriate circumstances (when $p_I = p_J =$ constant) then it becomes a classical Euclidean distance.

Input data table, marginals, and masses

- The given contingency table data are denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.
- We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$.
- From frequencies to probabilities: $f_{IJ} = \{f_{ij} = k(i,j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly f_I is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and f_J analogously.
- The conditional distribution of f_J knowing $i \in I$, also termed the jth profile with coordinates indexed by the elements of I, is $f_J^i = \{f_j^i = f_{ij}/f_i = (k_{ij}/k)/(k_i/k); f_i \neq 0; j \in J\}$ and likewise for f_I^j.

Clouds of points, masses, and inertia

- Moment of inertia of a cloud of points in a Euclidean space, with both distances and masses defined: $M^2(N_J(I)) = \sum_{i \in I} f_i \| f_J^i - f_J \|^2_{f_J} = \sum_{i \in I} f_i \rho^2(i)$.
- Here: ρ is the Euclidean distance from the cloud centre, and f_i is the mass of element i.
- The mass is the marginal distribution of the input data table.
- Correspondence analysis is, as will be seen, a decomposition of the inertia of a cloud of points, endowed with masses.

Inertia and Distributional Equivalence

- Another expression for inertia: $M^2(N_J(I)) = M^2(N_I(J)) = \|f_{IJ} - f_I f_J\|^2_{f_I f_J} = \sum_{i \in I, j \in J} (f_{ij} - f_i f_j)^2 / f_i f_j$.
- The term $\|f_{IJ} - f_I f_J\|^2_{f_I f_J}$ is the χ^2 metric between the probability distribution f_{IJ} and the product of marginal distributions $f_I f_J$, with as centre of the metric the product $f_I f_J$.
- *Principle of distributional equivalence:* Consider two elements j_1 and j_2 of J with identical profiles: i.e. $f_I^{j_1} = f_I^{j_2}$. Consider now that elements (or columns) j_1 and j_2 are replaced with a new element j_s such that the new coordinates are aggregated profiles, $f_{ij_s} = f_{ij_1} + f_{ij_2}$, and the new masses are similarly aggregated: $f_{ij_s} = f_{ij_1} + f_{ij_2}$. Then there is *no effect* on the distribution of distances between elements of I. The distance between elements of J, other than j_1 and j_2 is naturally not modified.

Inertia and Distributional Equivalence (Cont'd.)

- The principle of distributional equivalence leads to representational self-similarity: aggregation of rows or columns, as defined above, leads to the same analysis. Therefore it is very appropriate to analyze a contingency table with fine granularity, and seek in the analysis to merge rows or columns, through aggregation.

Factors

- Correspondence Analysis produces an ordered sequence of pairs, called factors, (F_α, G_α) associated with real numbers called eigenvalues $0 \leq \lambda_\alpha \leq 1$.
- We denote $F_\alpha(I)$ the value of the factor of rank α for element i of I; and similarly $G_\alpha(J)$ is the value of the factor of rank α for element j of J.
- We see that F is a function on I, and G is a function on J.
- The number of eigenvalues and associated factor couples is: $\alpha = 1, 2, \ldots, N = \inf(\mid I \mid -1, \mid J \mid -1)$, where $\mid . \mid$ denotes set cardinality.

Properties of factors

- $\sum_{i \in I} f_i F_\alpha(i) = 0; \quad \sum_{j \in J} f_j G_\alpha(j) = 0$
- $\sum_{i \in I} f_i F_\alpha^2(i) = \lambda_\alpha; \quad \sum_{j \in J} f_j G_\alpha^2(j) = \lambda_\alpha$
- $\sum_{i \in I} f_i F_\alpha(i) F_\beta(i) = \delta_{\alpha\beta}$
- $\sum_{j \in J} f_j G_\alpha(j) G_\beta(j) = \delta_{\alpha\beta}$
- Notation: $\delta_{\alpha\beta} = 0$ if $\alpha \neq \beta$ and $= 1$ if $\alpha = \beta$.
- Normalized factors: on the sets I and J, we next define the functions ϕ^I and ψ^J of zero mean, of unit variance, pairwise uncorrelated on I (resp. J), and associated with masses f_J (resp. f_I).
- $\sum_{i \in I} f_i \phi_\alpha(i) = 0; \quad \sum_{j \in J} f_j \psi_\alpha(j) = 0$
- $\sum_{i \in I} f_i \phi_\alpha^2(i) = 1; \quad \sum_{j \in J} f_j \psi_\alpha^2(j) = 1$
- $\sum_{i \in I} f_i \phi_\alpha(i) \phi_\beta(i) = \delta_{\alpha\beta}; \quad \sum_{j \in J} f_j \psi_\alpha(j) \psi_\beta(j) = \delta_{\alpha\beta}$

- Between unnormalized and normalized factors, we have the following relations.
- $\phi_\alpha(i) = \lambda_\alpha^{-\frac{1}{2}} F_\alpha(i) \ \forall i \in I, \ \forall \alpha = 1, 2, \ldots N$
- $\psi_\alpha(j) = \lambda_\alpha^{-\frac{1}{2}} G_\alpha(j) \ \forall j \in J, \ \forall \alpha = 1, 2, \ldots N$
- The moment of inertia of the clouds $N_J(I)$ and $N_I(J)$ in the direction of the α axis is λ_α.

Forward transform

- Have that the χ^2 metric is defined in direct space, i.e. space of profiles.
- The Euclidean metric is defined for the factors.
- We can characterize correspondence analysis as the mapping of a cloud in χ^2 space to Euclidean space.
- Distances between profiles are as follows.
- $\| f_J^i - f_J^{i'} \|^2_{f_J} = \sum_{j \in J} \left(f_j^i - f_j^{i'} \right)^2 / f_j = \sum_{\alpha = 1..N} \left(F_\alpha(i) - F_\alpha(i') \right)^2$
- $\| f_I^j - f_I^{j'} \|^2_{f_I} = \sum_{i \in I} \left(f_i^j - f_i^{j'} \right)^2 / f_i = \sum_{\alpha = 1..N} \left(G_\alpha(j) - G_\alpha(j') \right)^2$
- Norm, or distance of a point $i \in N_J(I)$ from the origin or centre of gravity of the cloud $N_J(I)$, is as follows.
- $\rho^2(i) = \| f_J^i - f_J \|^2_{f_J} = \sum_{\alpha = 1..N} F_\alpha^2(i)$
$\rho^2(j) = \| f_I^j - f_I \|^2_{f_I} = \sum_{\alpha = 1..N} F_\alpha^2(j)$

Inverse transform

- The correspondence analysis transform, taking profiles into a factor space, is reversed with no loss of information as follows $\forall (i, j) \in I \times J$.
- $f_{ij} = f_i f_j \left(1 + \sum_{\alpha=1..N} \lambda_\alpha^{-\frac{1}{2}} F_\alpha(i) G_\alpha(j)\right)$
- For profiles we have the following.
- $f_i^j = f_i \left(1 + \sum_\alpha \lambda_\alpha^{-\frac{1}{2}} F_\alpha(i) G_\alpha(j)\right)$
- $f_j^i = f_j \left(1 + \sum_\alpha \lambda_\alpha^{-\frac{1}{2}} F_\alpha(i) G_\alpha(j)\right)$

Decomposition of inertia

- The distance of a point from the centre of gravity of the cloud is as follows.
- $\rho^2(i) = \| f_J^i - f_J \|^2 = \sum_{j \in J} \left(f_j^i - f_j \right)^2 / f_j$
- Decomposition of the cloud's inertia is as follows.
- $M^2(N_J(I)) = \sum_{\alpha = 1..N} \lambda_\alpha = \sum_{i \in I} f_i \rho^2(i)$
- In greater detail, we have the following for this decomposition.
- $\lambda_\alpha = \sum_{i \in I} f_i F_\alpha^2(i)$ and $\rho^2(i) = \sum_{\alpha = 1..N} F_\alpha^2(i)$

Relative and absolute contributions

- $f_i \rho^{(i)}$ is the absolute contribution of point i to the inertia of the cloud, $M^2(N_J(I))$, or the variance of point i.
- $f_i F_\alpha^2(i)$ is the absolute contribution of point i to the moment of inertia λ_α.
- $f_i F_\alpha^2(i)/\lambda_\alpha$ is the relative contribution of point i to the moment of inertia λ_α. (Often denoted CTR.)
- $F_\alpha^2(i)$ is the contribution of point I to the χ^2 distance between i and the centre of the cloud $N_J(I)$.
- $\cos^2 a = F_\alpha^2(i)/\rho^2(i)$ is the relative contribution of the factor α to point i. (Often denoted COR.)
- Based on the latter term, we have: $\sum_{\alpha=1..N} F_\alpha^2(i)/\rho^2(i) = 1$.
- Analogous formulas hold for the points j in the cloud $N_I(J)$.

Reduction of dimensionality

- Interpretation is usually limited to the first few factors.
- Decomposition of inertia is usually far less decisive than (cumulative) percentage variance explained in principal components analysis. One reason for this: in CA, often recoding tends to bring input data coordinates closer to vertices of hypercube.
- QLT$(i) = \sum_{\alpha=1..N'} \cos^2 a$, where angle a has been defined above (previous section) and where $N' < N$ is the quality of representation of element i in the factor space of dimension N'.
- INR$(I) = \rho^2(i)$ is the distance of element I from the centre of gravity of the cloud.
- POID$(I) = f_i$ is the mass or marginal frequency of the element i.

Interpretation of results

1. Projections onto factors 1 and 2, 2 and 3, 1 and 3, etc. of set I, set J, or both sets simultaneously.
2. Spectrum of non-increasing values of eigenvalues.
3. Interpretation of axes. We can distinguish between the general (latent semantic, conceptual) meaning of axes, and axes which have something specific to say about groups of elements. Usually contrast is important: what is found to be analogous at one extremity versus the other extremity; or oppositions or polarities.
4. Factors are determined by how much the elements contribute to their dispersion. Therefore the values of CTR are examined in order to identify or to name the factors (for example, with higher order concepts). (Informally, CTR allows us to work from the elements towards the factors.)
5. The values of COR are squared cosines, which can be considered as being like

correlation coefficients. If COR(i, α) is large (say, around 0.8) then we can say that that element is well explained by the axis of rank α. (Informally, COR allows us to work from the factors towards the elements.)

Analysis of the dual spaces

- We have the following.
- $F_\alpha(i) = \lambda_\alpha^{-\frac{1}{2}} \sum_{j \in J} f_j^i G_\alpha(j)$ for $\alpha = 1, 2, \ldots N; i \in I$
- $G_\alpha(j) = \lambda_\alpha^{-\frac{1}{2}} \sum_{i \in I} f_i^j F_\alpha(i)$ for $\alpha = 1, 2, \ldots N; j \in J$
- These are termed the *transition formulas*. The coordinate of element $i \in I$ is the barycentre of the coordinates of the elements $j \in J$, with associated masses of value given by the coordinates of f_j^i of the profile f_J^i. This is all to within the $\lambda_\alpha^{-\frac{1}{2}}$ constant.

Analysis of the dual spaces (cont'd.)

- We also have the following.
- $\phi_\alpha(i) = \lambda_\alpha^{-\frac{1}{2}} \sum_{j \in J} f_j^i \psi_\alpha(j)$
- $\psi_\alpha(j) = \lambda_\alpha^{-\frac{1}{2}} \sum_{i \in I} f_i^j \phi_\alpha(i)$
- This implies that we can pass easily from one space to the other. I.e. we carry out the diagonalization, or eigen-reduction, in the more computationally favourable space which is usually $\mathbb{R}^J$. In the output display, the barycentric principle comes into play: this allows us to simultaneously view and interpret observations and attributes.

Supplementary elements

- Overly-preponderant elements (i.e. row or column profiles), or exceptional elements (e.g. a sex attribute, given other performance or behavioural attributes) may be placed as supplementary elements.
- This means that they are given zero mass in the analysis, and their projections are determined using the transition formulas.
- This amounts to carrying out a correspondence analysis first, without these elements, and then projecting them into the factor space following the determination of all properties of this space.

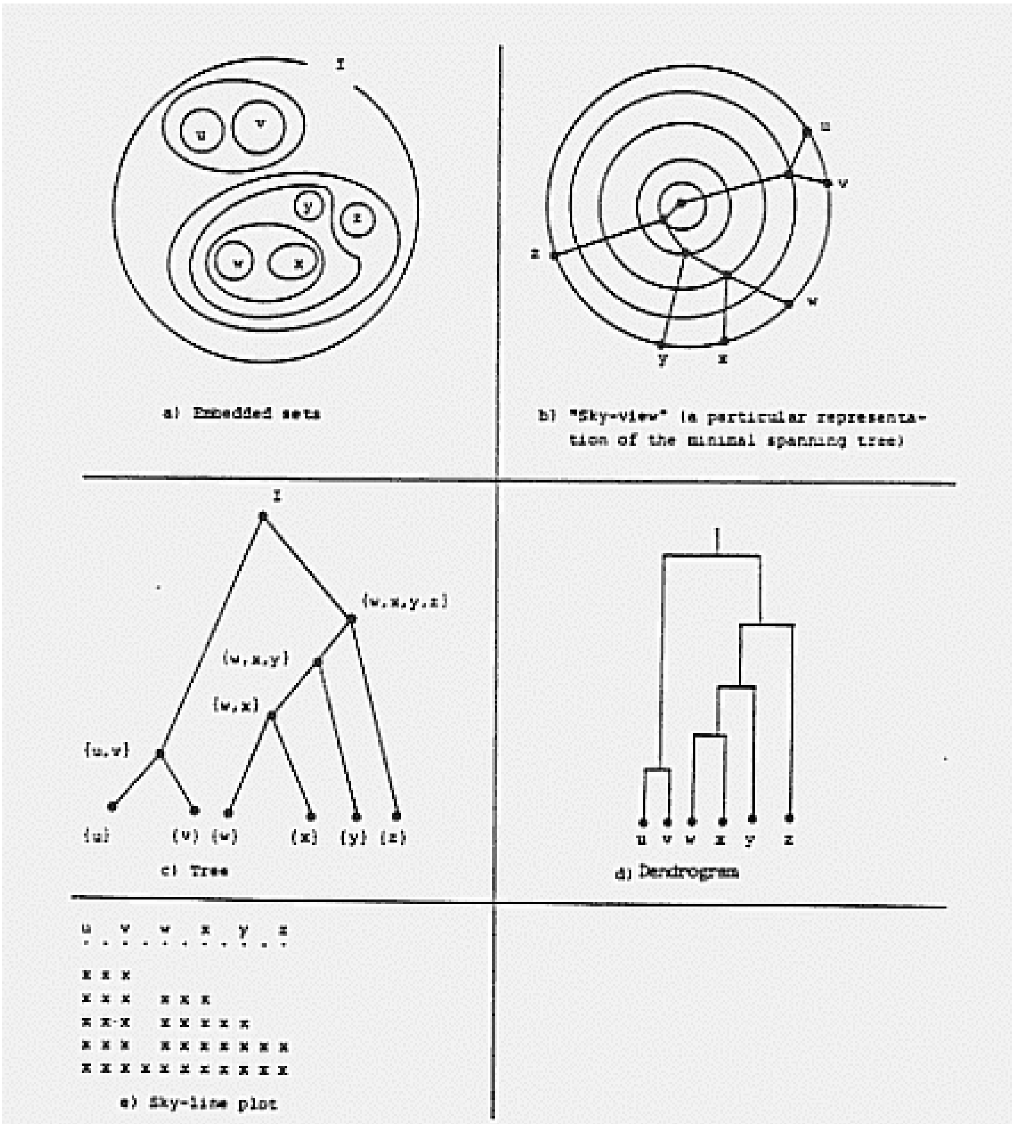
Summary

Space $\mathbb{R}^m$:

1. n row points, each of m coordinates.
2. The j^{th} coordinate is x_{ij}/x_i.
3. The mass of point i is x_i.
4. The χ^2 distance between row points i and k is: $d^2(i,k) = \sum_j \frac{1}{x_j}(\frac{x_{ij}}{x_i} - \frac{x_{kj}}{x_k})^2$.
   Hence this is a Euclidean distance, with respect to the weighting $1/x_j$ (for all j), between *profile* values x_{ij}/x_i etc.
5. The criterion to be optimized: the weighted sum of squares of projections, where the weighting is given by x_i (for all i).

Space $\mathbb{R}^n$:

1. m column points, each of n coordinates.
2. The i^{th} coordinate is x_{ij}/x_j.
3. The mass of point j is x_j.
4. The χ^2 distance between column points g and j is:
   $d^2(g,j) = \sum_i \frac{1}{x_i}(\frac{x_{ig}}{x_g} - \frac{x_{ij}}{x_j})^2$.
   Hence this is a Euclidean distance, with respect to the weighting $1/x_i$ (for all i), between *profile* values x_{ig}/x_g etc.
5. The criterion to be optimized: the weighted sum of squares of projections, where the weighting is given by x_j (for all j).


I
a) Embedded sets
b) "Sky-view" (a particular representation of the minimal spanning tree)
c) Tree
d) Dendrogram
e) Sky-line plot

Hierarchical clustering

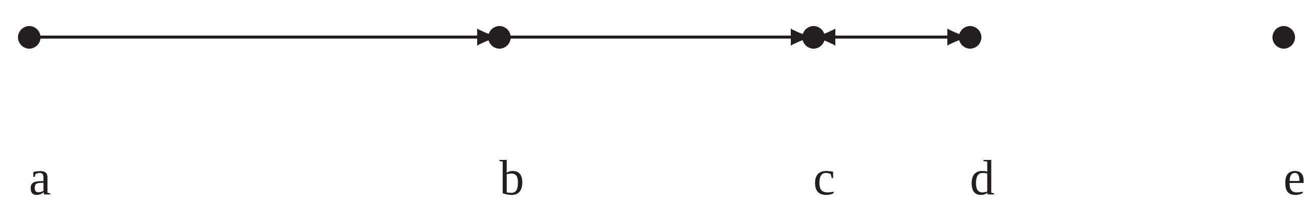
- Hierarchical agglomeration on n observation vectors, $i \in I$, involves a series of $1, 2, \ldots, n-1$ pairwise agglomerations of observations or clusters, with the following properties.
- A hierarchy $H = \{q | q \in 2^I\}$ such that:
  1. $I \in H$
  2. $i \in H \; \forall i$
  3. for each $q \in H, q' \in H : q \cap q' \neq \emptyset \Longrightarrow q \subset q'$ or $q' \subset q$
- An indexed hierarchy is the pair (H, ν) where the positive function defined on H, i.e., $\nu : H \rightarrow \mathbb{R}^+$, satisfies:
  1. $\nu(i) = 0$ if $i \in H$ is a singleton
  2. $q \subset q' \Longrightarrow \nu(q) < \nu(q')$
- Function ν is the agglomeration level.

- Take $q \subset q'$, let $q \subset q''$ and $q' \subset q''$, and let q'' be the lowest level cluster for which this is true. Then if we define $D(q, q') = \nu(q'')$, D is an ultrametric.
- Recall: Distances satisfy the triangle inequality $d(x, z) \leq d(x, y) + d(y, z)$. An ultrametric satisfies $d(x, z) \leq \max(d(x, y), d(y, z))$. In an ultrametric space triangles formed by any three points are isosceles. An ultrametric is a special distance associated with rooted trees. Ultrametrics are used in other fields also – in quantum mechanics, numerical optimization, number theory, and algorithmic logic.
- In practice, we start with a Euclidean distance or other dissimilarity, use some criterion such as minimizing the change in variance resulting from the agglomerations, and then define $\nu(q)$ as the dissimilarity associated with the agglomeration carried out.

Minimum variance agglomeration

- For Euclidean distance inputs, the following definitions hold for the minimum variance or Ward error sum of squares agglomerative criterion.
- Coordinates of the new cluster center, following agglomeration of q and q', where m_q is the mass of cluster q defined as cluster cardinality, and (vector) q denotes using overloaded notation the center of (set) cluster q:
  $q'' = (m_q q + m_{q'} q')/(m_q + m_{q'})$.
- Following the agglomeration of q and q', we define the following dissimilarity:
  $(m_q m_{q'})/(m_q + m_{q'}) \|q - q'\|^2$.
- Hierarchical clustering is usually based on factor projections, if desired using a limited number of factors (e.g. 7) in order to filter out the most useful information in our data.
- In such a case, hierarchical clustering can be seen to be a mapping of Euclidean distances into ultrametric distances.

Efficient NN chain algorithm



- A *NN*-chain (nearest neighbour chain)

Efficient NN chain algorithm (cont'd.)

- An *NN*-chain consists of an arbitrary point followed by its *NN*; followed by the *NN* from among the remaining points of this second point; and so on until we necessarily have some pair of points which can be termed reciprocal or mutual *NN*s. (Such a pair of *RNN*s may be the first two points in the chain; and we have assumed that no two dissimilarities are equal.)
- In constructing a *NN*-chain, irrespective of the starting point, we may agglomerate a pair of *RNN*s as soon as they are found.
- Exactness of the resulting hierarchy is guaranteed when the cluster agglomeration criterion respects the *reducibility property*.
- Inversion impossible if: $d(i,j) < d(i,k)$ or $d(j,k) \Rightarrow d(i,j) < d(i \cup j, k)$

Minimum variance method: properties

- We seek to agglomerate two clusters, c_1 and c_2, into cluster c such that the within-class variance of the partition thereby obtained is minimum.
- Alternatively, the between-class variance of the partition obtained is to be maximized.
- Let P and Q be the partitions prior to, and subsequent to, the agglomeration; let $p_1, p_2, \ldots$ be classes of the partitions.

$$\begin{aligned} P &= \{p_1, p_2, \ldots, p_k, c_1, c_2\} \\ Q &= \{p_1, p_2, \ldots, p_k, c\}. \end{aligned}$$

- Total variance of the cloud of objects in m-dimensional space is decomposed into the sum of within-class variance and between-class variance. This is Huyghen's theorem in classical mechanics.
- Total variance, between-class variance, and within-class variance are as follows:

$V(I) = \frac{1}{n} \sum_{i \in I} (i - g)^2$, $V(P) = \sum_{p \in P} \frac{|p|}{n} (p - g)^2$; and $\frac{1}{n} \sum_{p \in P} \sum_{i \in p} (i - p)^2$.

- For two partitions, before and after an agglomeration, we have respectively:

$$V(I) = V(P) + \sum_{p \in P} V(p)$$

$$V(I) = V(Q) + \sum_{p \in Q} V(p)$$

- From this, it can be shown that the criterion to be optimized in agglomerating c_1 and c_2 into new class c is:

$$\begin{aligned} V(P) - V(Q) &= V(c) - V(c_1) - V(c_2) \\ &= \frac{|c_1|\ |c_2|}{|c_1| + |c_2|} \|\mathbf{c_1} - \mathbf{c_2}\|^2 , \end{aligned}$$

FACOR and VACOR: Analysis of clusters

- The barycentric principle allows both row points and column points to be displayed simultaneously as projections.
- We therefore can consider:
  - simultaneous display of I and J
  - tree on I
  - tree on J
- To help analyze these outputs we can explore the representation of clusters (derived from the hierarchical trees) in factor space, leading to programs traditionally called FACOR.
- And the representation of clusters in the profile coordinate space, leading to programs traditionally called VACOR.

- In the case of FACOR, for every couple q, q' of a partition of I, we calculate

$$\frac{(f_q f_q')}{(f_q + f_q')} \|q - q'\|^2$$

This can be decomposed using the axes of $\mathbb{R}_J$, as well as using the factorial axes.

- In the case of VACOR, we can explore the cluster dipoles which takes account of the “elder” and “younger” cluster components:

```
           n
          / \
         /   \
        /     \
     a(n)    b(n)
```

- We have $F_\alpha(a) = \sum_{i \in q} (f_i / f_q) F_\alpha(i)$. We consider the vectors defining the dipole: $[q, a(q)]$ and $[q, b(q)]$.

- We then study the squared cosine of the angle between vector $[a(q), b(q)]$ and the factorial axis of rank α.

- This squared cosine defines the relative contribution of the pair q, α to the level index $\nu(q)$ of the class q.

Summary

- Correspondence analysis displays observation profiles in a low-dimensional factorial space.
- Profiles are points endowed with χ^2 distance.
- Under appropriate circumstances, the χ^2 distance reduces to a Euclidean distance.
- A factorial space is nearly always Euclidean.
- Simultaneously a hierarchical clustering is built using the observation profiles.
- Usually one or a small number of partitions are derived from the hierarchical clustering.
- A hierarchical clustering defines an ultrametric distance.
- Input for the hierarchical clustering is usually factor projections.

- In summary, correspondence analysis involves mapping a χ^2 distance into a particular Euclidean distance; and mapping this Euclidean distance into an ultrametric distance.
- The aim is to have different but complementary analytic tools to facilitate interpretation of our data.

To read further

- Ch. Bastin, J.P. Benzécri, Ch. Bourgarit and P. Cazes, *Pratique de l'Analyse des Données, Tome 2*, Dunod, Paris, 1980.
- J.P. Benzécri and F. Benzécri, F. *Pratique de l'Analyse des Données, Vol. 1: Analyse des Correspondances. Exposé Élémentaire*, Dunod, Paris, 1980.
- J.P. Benzécri, *L'Analyse des Données. Tome 1. La Taxinomie*, 2nd ed., Dunod, Paris, 1976.
- J.P. Benzécri, *L'Analyse des Données. Tome 2. L'Analyse des Correspondances*, 2nd ed., Dunod, Paris, 1976.
- J.P. Benzécri, *Correspondence Analysis Handbook*, Marcel Dekker, Basel, 1992.
- M. Jambu, *Classification Automatique pour l'Analyse des Données. 1. Méthodes et Algorithmes*, Dunod, Paris, 1978.

- L. Lebart, A. Morineau and K.M. Warwick, *Multivariate Descriptive Statistical Analysis*, Wiley, New York, 1984.
- F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms", *The Computer Journal*, 26, 354-359, 1983.
- F. Murtagh, *Multidimensional Clustering Algorithms*, COMPSTAT Lectures Volume 4, Physica-Verlag, Vienna, 1985.
- F. Murtagh and A. Heck, *Multivariate Data Analysis*, Kluwer, 1987.
- H. Rouanet and B. Le Roux, *Analyse des Données Multidimensionnelles*, Dunod, Paris, 1993.
- M. Volle, *Analyse des Données*, 2nd Edition, Economica, Paris, 1980.