Lecture 1: Metric and Ultrametric Embedding

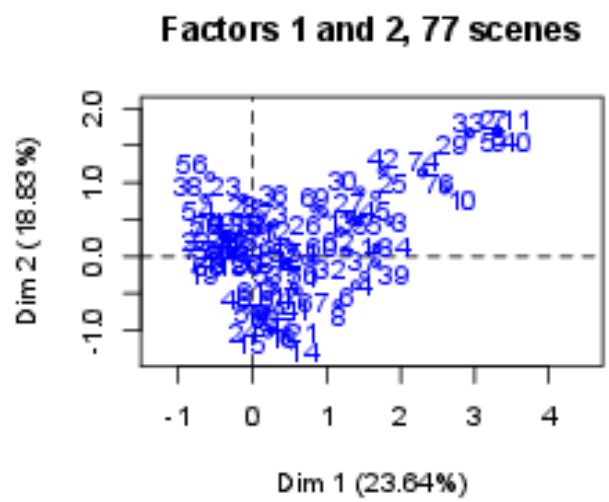
Themes of Lecture 1 are: metric projection and inducing a hierarchy, hence an ultrametric. Using Correspondence Analysis and agglomerative hierarchical clustering. Topics are as follows.

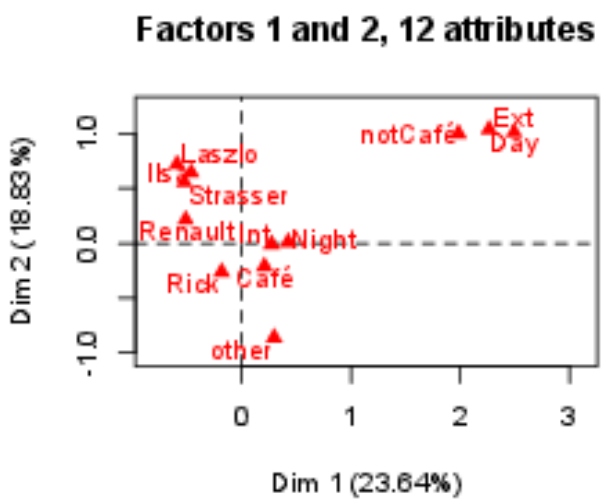
1. Examples, using Casablanca movie.
2. Metrics, clouds of points, masses, inertia.
3. Factors, decomposition of inertia, contributions, dual spaces.
4. Hierarchical agglomerative clustering
5. Minimum variance agglomerative hierarchical clustering criterion.

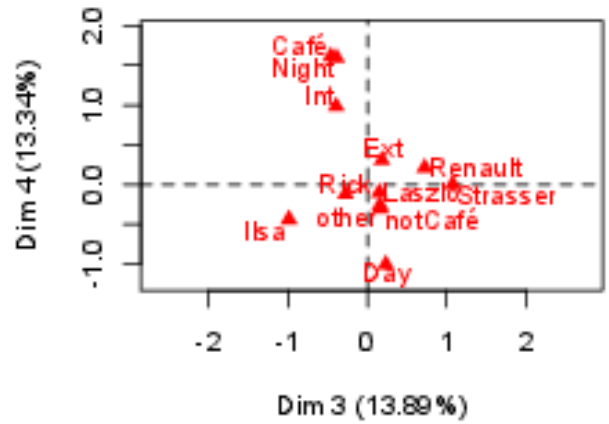
Casablanca Movie

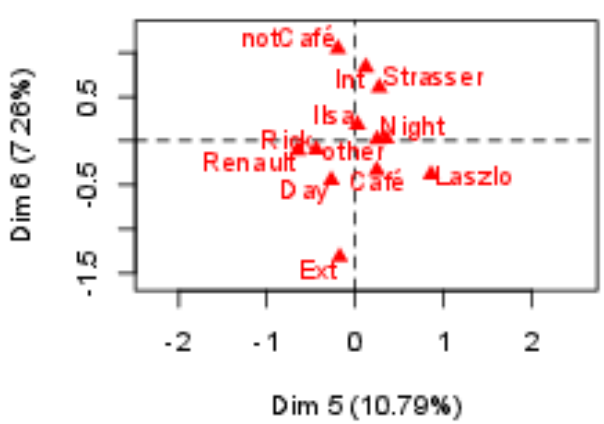
- 77 scenes in the movie. I use the filmscript. Here I use 12 attributes (metadata, characters, locations): Int(erior), Ext(erior), Day, Night, Ilsa, Rick, Renault, Strasser, Laszlo, Other (minor characters), Rick's Café, Elsewhere.
- Example of three scenes, 12–14, follows.

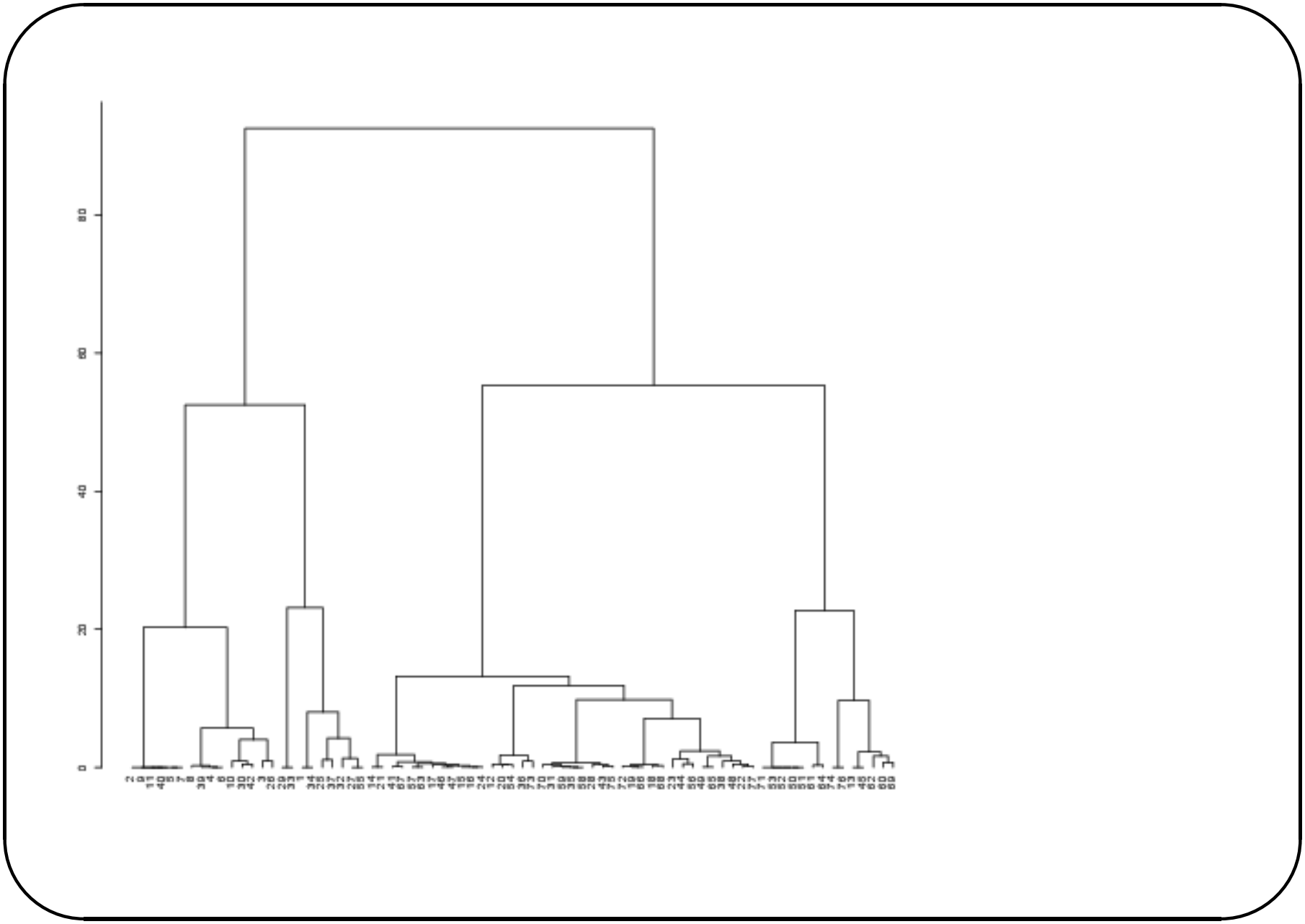
	INT	EXT	Day	Night	Ilsa	Rick	Renault	Strasser	Laszlo	Other	Cafe	Elsewhere
12	0	1	1	0	0	0	6	8	0	5	0	1
13	0	1	0	1	0	0	0	0	0	0	1	0
14	1	0	0	1	0	0	0	0	0	10	1	0

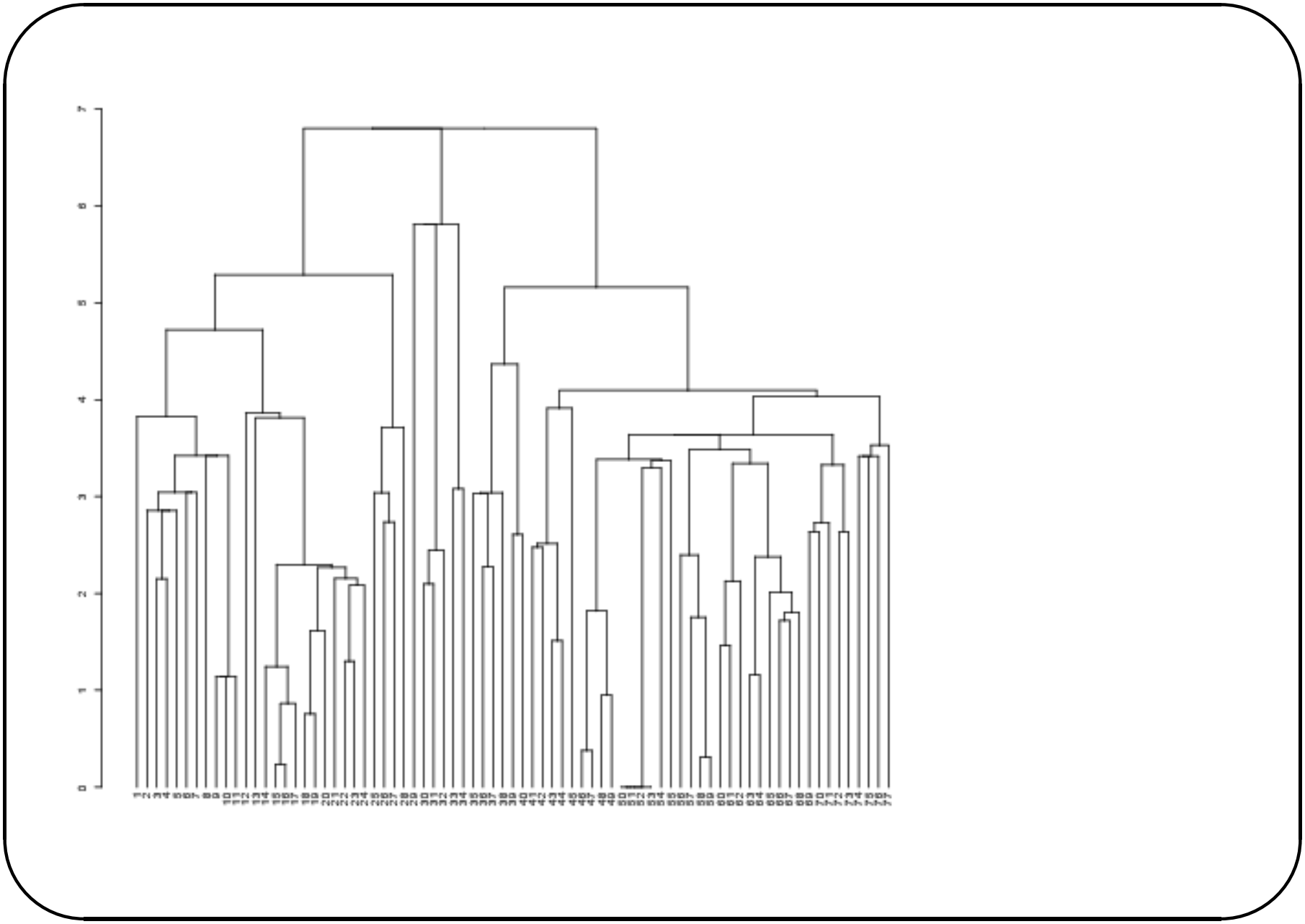

Factors 1 and 2, 77 scenes
Dim 2 (18.83%)
Dim 1 (23.64%)


Factors 1 and 2, 12 attributes
notCafé
Ext
Day
Laszlo
Ilsa
Strasser
Renault
Int
Night
Rick
Café
other
Dim 2 (18.83%)
Dim 1 (23.64%)


Factors 3 and 4, 12 attributes
Café
Night
Int
Ext
Renault
Rick
Laszlo
Strasser
Ilsa
other
notCafé
Day
Dim 4 (13.34%)
Dim 3 (13.89%)


Factors 5 and 6, 12 attributes
notCafé
Int
Strasser
Ilsa
Night
Rick
other
Renault
Day
Café
Laszlo
Ext
Dim 6 (7.26%)
Dim 5 (10.79%)


0
20
40
60
80


0
1
2
3
4
5
6
7

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. (Principal components analysis, multidimensional scaling; Latent Semantic Indexing.)

Data

- Matrix X defines a set of n vectors in m-dimensional space: $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ for $1 \leq i \leq n$.
- We have: $x_i \in \mathbb{R}^m$
- Matrix X also defines a set of m column vectors in n-dimensional space: $x_j = \{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ for $1 \leq j \leq m$.
- We have: $x_j \in \mathbb{R}^n$
- By convention we usually take the space of row points, i.e. $\mathbb{R}^m$, as X; and the space of column points, i.e. $\mathbb{R}^n$, as the transpose of X, i.e. X' or X^t.
- The row points define a cloud of n points in $\mathbb{R}^m$.
- The column points define a cloud of m points in $\mathbb{R}^n$.

Next topic: Metrics in Data Analysis. Euclidean, Chi Squared.

Metrics

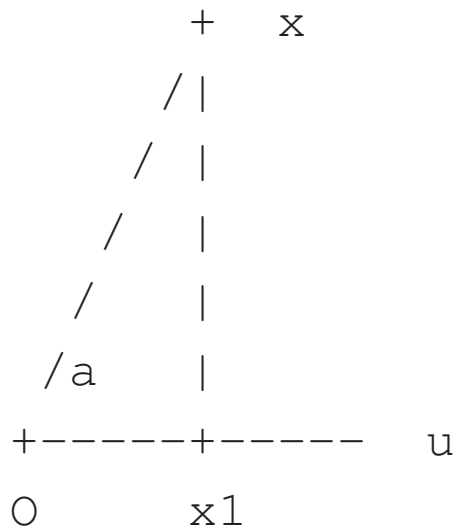
- The notion of distance is crucial, since we want to investigate relationships between observations and/or variables.
- Recall: $x = \{3, 4, 1, 2\}, y = \{1, 3, 0, 1\}$, then: scalar product $\langle x, y \rangle = \langle y, x \rangle = x'y = xy' = 3 \times 1 + 4 \times 3 + 1 \times 0 + 2 \times 1$.
- 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: $(3-1)^2 + (4-3)^2 + (1-0)^2 + (2-1)^2$
- 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 $\langle x, y \rangle_M = x'My$, norm is $\|x\|^2 = x'Mx$, 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.



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: $x' M_{p_I} x = \sum_{i \in I} p_i x_i^2 = \mathrm{var}(x)$; and $x' M_{p_I} y = \sum_{i \in I} p_i x_i y_i = \mathrm{cov}(x, y)$.

Objectives of PCA, Principal Components Analysis

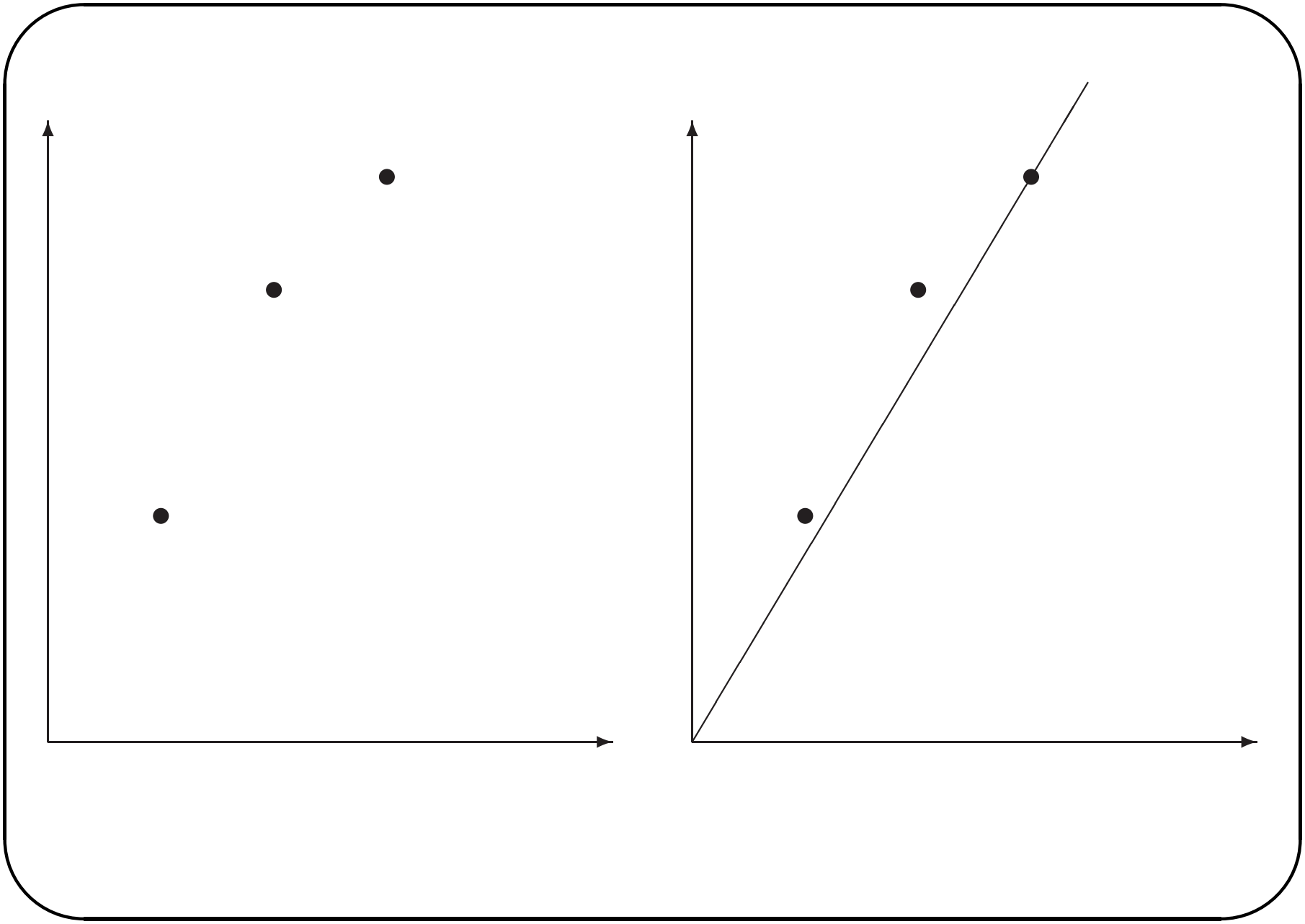
- dimensionality reduction;
- the determining of linear combinations of variables;
- feature selection: the choosing of the most useful variables;
- visualization of multidimensional data;
- identification of underlying variables;
- identification of groups of objects or of outliers.

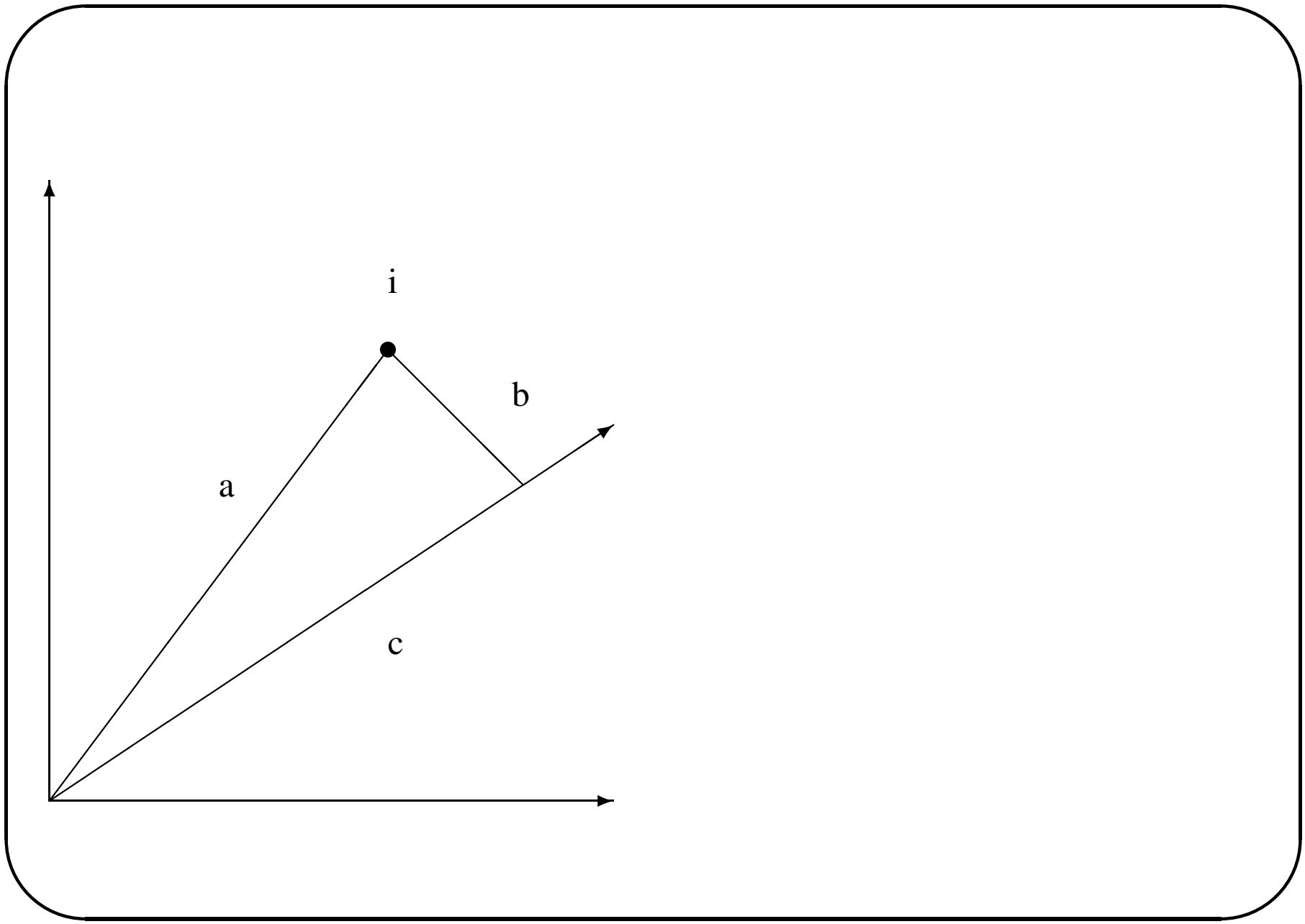
Least Squares Optimal Projection of Points

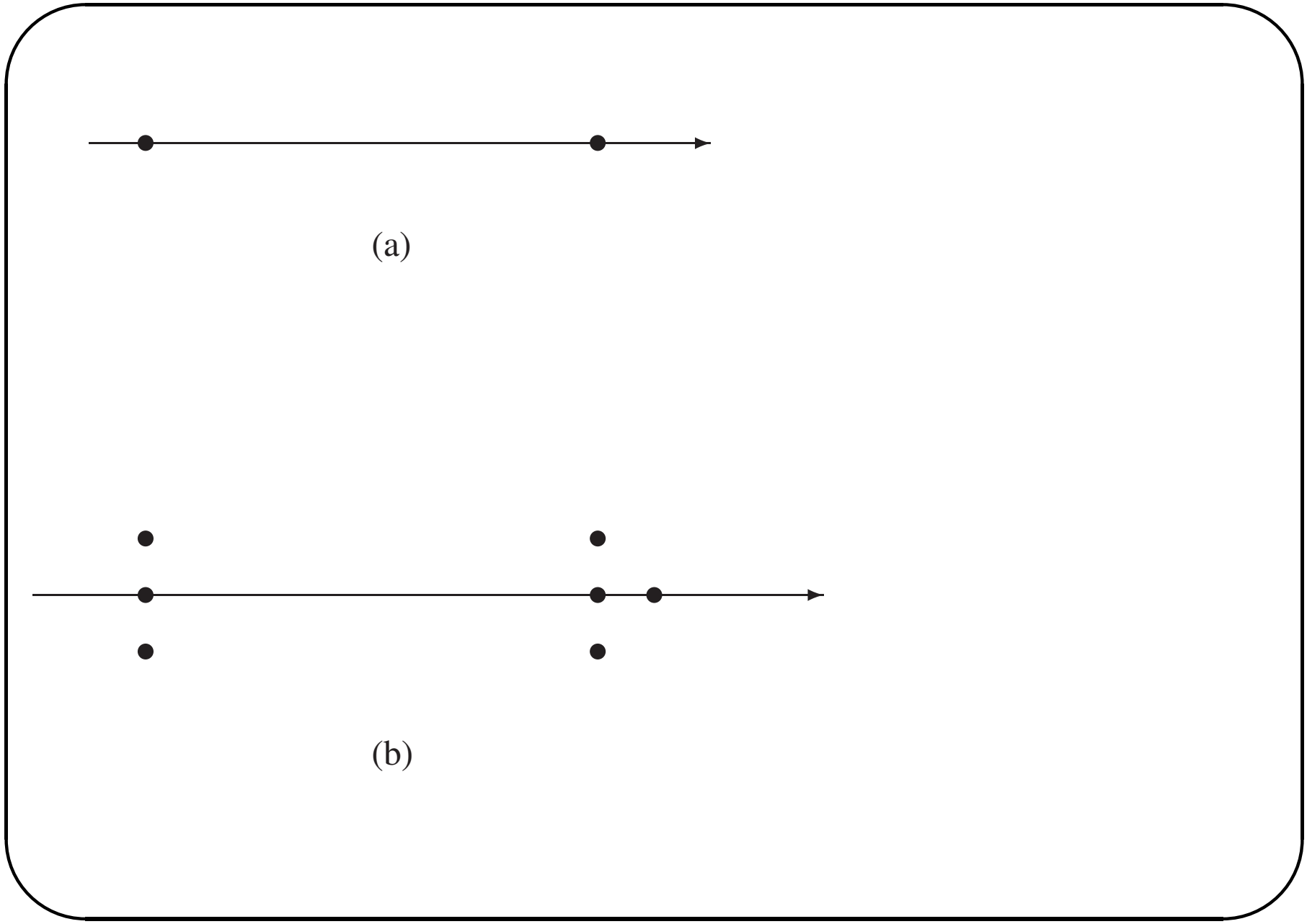
- Plot of 3 points in $\mathbb{R}^2$ (see following slides).
- PCA: determine best fitting axes.
- Examples follow.
- Note: optimization means either (i) closest axis to points, or (ii) maximum elongation of projections of points on the axis.
- This follows from Pythagoras's theorem: $x^2 + y^2 = z^2$. Call z the distance from the origin to a point. Let x be the distance of the projection of the point from the origin. Then y is the perpendicular distance from the axis to to the point.
- Minimizing y is the same as maximizing x (because z is fixed).

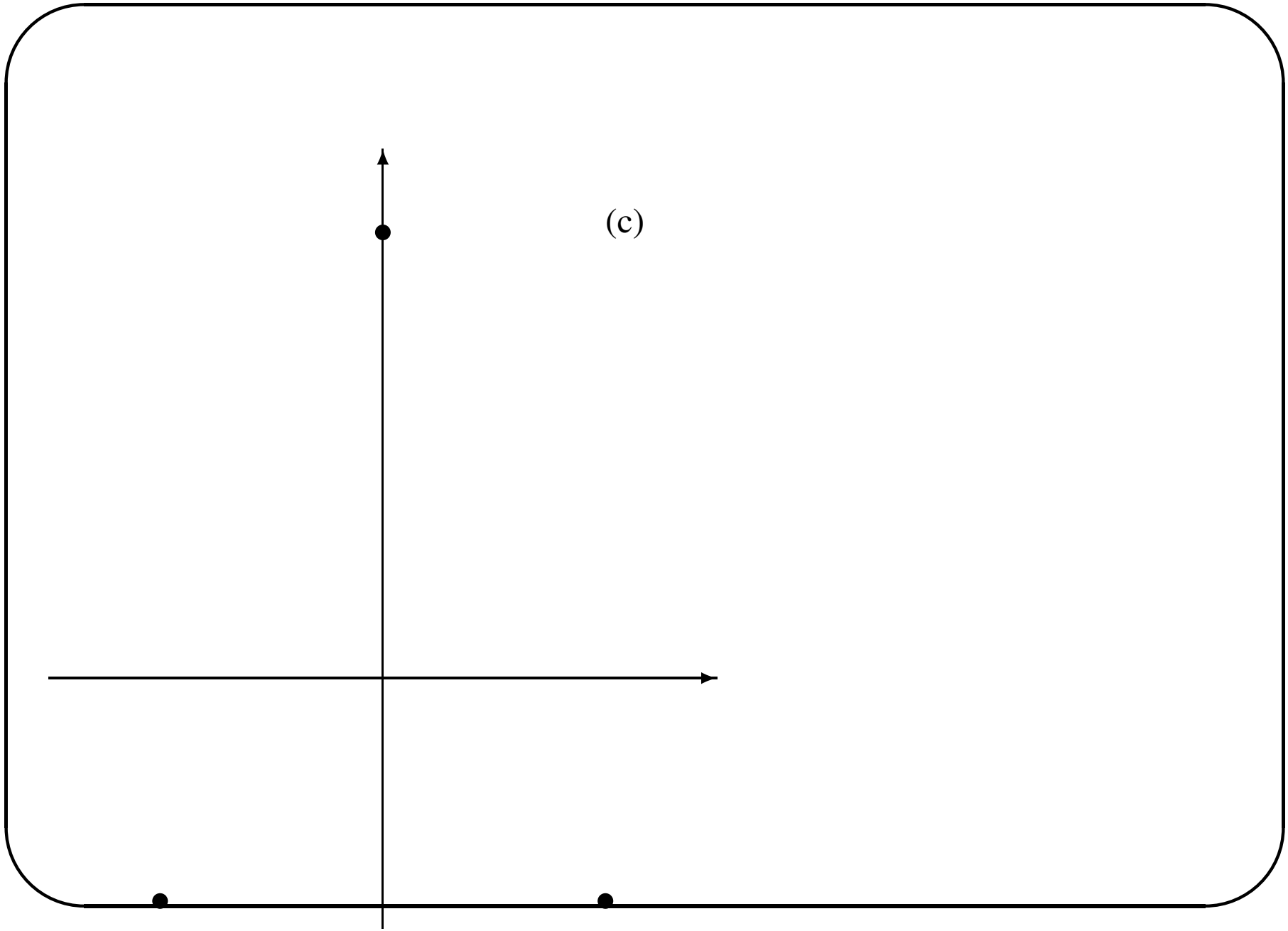
Examples of Optimal Projection

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 5 \end{pmatrix}$$




i
b
a
c


(a)
(b)


(c)

Questions We Will Now Address

- How is the PCA of an $n \times m$ matrix related to the PCA of the transposed $m \times n$ matrix?
- How may the new axes derived – the principal components – be said to be linear combinations of the original axes?
- How may PCA be understood as a series expansion?
- In what sense does PCA provide a lower-dimensional approximation to the original data?

PCA Algorithm

- The projection of vector $\mathbf{x}$ onto axis $\mathbf{u}$ is $\mathbf{y} = \frac{\mathbf{x}'M\mathbf{u}}{\|\mathbf{u}\|_M}\ \mathbf{u}$
- I.e. the coordinate of the projection on the axis is $\mathbf{x}'M\mathbf{u}/\|\mathbf{u}\|_M$.
- This becomes $\mathbf{x}'M\mathbf{u}$ when the vector $\mathbf{u}$ is of unit length.
- The cosine of the angle between vectors $\mathbf{x}$ and $\mathbf{y}$ in the usual Euclidean space is $\mathbf{x}'\mathbf{y}/\|\mathbf{x}\|\|\mathbf{y}\|$.
- That is to say, we make use of the triangle whose vertices are the origin, the projection of $\mathbf{x}$ onto $\mathbf{y}$, and vector $\mathbf{x}$.
- The cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$ is then the coordinate of the projection of $\mathbf{x}$ onto $\mathbf{y}$, divided by the – hypotenuse – length of $\mathbf{x}$.
- The correlation coefficient between two vectors is then simply the cosine of the angle between them, when the vectors have first been centred (i.e. $\mathbf{x} - \mathbf{g}$ and $\mathbf{y} - \mathbf{g}$ are used, where $\mathbf{g}$ is the overall centre of gravity).

PCA Algorithm (Cont'd.)

- $X = \{x_{ij}\}$
- In $\mathbb{R}^m$, the space of objects, PCA searches for the best-fitting set of orthogonal axes to replace the initially-given set of m axes in this space.
- An analogous procedure is simultaneously carried out for the dual space, $\mathbb{R}^n$.
- First, the axis which best fits the objects/points in $\mathbb{R}^m$ is determined.
- If $\mathbf{u}$ is this vector, and is of unit length, then the product $X\mathbf{u}$ of $n \times m$ matrix by $m \times 1$ vector gives the projections of the n objects onto this axis.
- The sum of squared projections of points on the new axis, for all points, is $(X\mathbf{u})'(X\mathbf{u})$.
- Such a quadratic form would increase indefinitely if $\mathbf{u}$ were arbitrarily large, so $\mathbf{u}$ is taken to be of unit length, i.e. $\mathbf{u}'\mathbf{u} = 1$.
- We seek a maximum of the quadratic form $\mathbf{u}'S\mathbf{u}$ (where $S = X'X$) subject to

the constraint that $\mathbf{u}'\mathbf{u} = 1$.

- This is done by setting the derivative of the Lagrangian equal to zero.
- Differentiation of $\mathbf{u}'S\mathbf{u} - \lambda(\mathbf{u}'\mathbf{u} - 1)$ where λ is a Lagrange multiplier gives $2S\mathbf{u} - 2\lambda\mathbf{u}$.
- The optimal value of $\mathbf{u}$ (let us call it $\mathbf{u}_1$) is the solution of $S\mathbf{u} = \lambda\mathbf{u}$.
- The solution of this equation is well-known: $\mathbf{u}$ is the eigenvector associated with the eigenvalue λ of matrix S.
- Therefore the eigenvector of $X'X$, $\mathbf{u}_1$, is the axis sought, and the corresponding largest eigenvalue, λ_1, is a figure of merit for the axis, – it indicates the amount of variance explained by the axis.
- The second axis is to be orthogonal to the first, i.e. $\mathbf{u}'\mathbf{u}_1 = 0$.
- The second axis satisfies the equation $\mathbf{u}'X'X\mathbf{u} - \lambda_2(\mathbf{u}'\mathbf{u} - 1) - \mu_2(\mathbf{u}'\mathbf{u}_1)$ where λ_2 and μ_2 are Lagrange multipliers.

- Differentiating gives $2S\mathbf{u} - 2\lambda_2\mathbf{u} - \mu_2\mathbf{u}_1$.
- This term is set equal to zero. Multiplying across by $\mathbf{u}_1'$ implies that μ_2 must equal 0.
- Therefore the optimal value of $\mathbf{u}$, $\mathbf{u}_2$, arises as another solution of $S\mathbf{u} = \lambda\mathbf{u}$.
- Thus λ_2 and $\mathbf{u}_2$ are the second largest eigenvalue and associated eigenvector of S.
- The eigenvectors of $S = X'X$, arranged in decreasing order of corresponding eigenvalues, give the line of best fit to the cloud of points, the plane of best fit, the three-dimensional hyperplane of best fit, and so on for higher-dimensional subspaces of best fit.
- $X'X$ is referred to as the *sums of squares and cross products* matrix.

Eigenvalues

- Eigenvalues are decreasing in value.
- $\lambda_i = \lambda_{i'}$? Then equally privileged directions of elongation have been found.
- $\lambda_i = 0$? Space is actually of dimensionality less than expected. Example: in 3D, points actually lie on a plane.
- Since PCA in $\mathbb{R}^n$ and in $\mathbb{R}^m$ lead respectively to the finding of n and of m eigenvalues, and since in addition it has been seen that these eigenvalues are identical, it follows that the number of *non-zero eigenvalues* obtained in either space is less than or equal to $\min(n, m)$.
- The eigenvectors associated with the p largest eigenvalues yield the best-fitting p-dimensional subspace of $\mathbb{R}^m$. A measure of the approximation is the percentage of variance explained by the subspace $\sum_{k \leq p} \lambda_k / \sum_{k=1}^{n} \lambda_k$ expressed as a percentage.

Dual Spaces

- In the dual space of attributes, $\mathbb{R}^n$, a PCA may equally well be carried out.
- For the line of best fit, $\mathbf{v}$, the following is maximized: $(X'\mathbf{v})'(X'\mathbf{v})$ subject to $\mathbf{v}'\mathbf{v} = \mathbf{1}$.
- In $\mathbb{R}^m$ we arrived at $X'X\mathbf{u}_1 = \lambda_1\mathbf{u}_1$.
- In $\mathbb{R}^n$, we have $XX'\mathbf{v}_1 = \mu_1\mathbf{v}_1$.
- Premultiplying the first of these relationships by X yields $(XX')(X\mathbf{u}_1) = \lambda_1(X\mathbf{u}_1)$.
- Hence $\lambda_1 = \mu_1$ because we have now arrived at two eigenvalue equations which are identical in form.
- Relationship between the eigenvectors in the two spaces: these must be of unit length.

- Find: $\mathbf{v}_1 = \frac{1}{\sqrt{\lambda_1}} X \mathbf{u}_1$.
- $\lambda > 0$ since if $\lambda = 0$ eigenvectors are not defined.
- For λ_k: $\mathbf{v}_k = \frac{1}{\sqrt{\lambda_k}} X \mathbf{u}_k$
- And: $\mathbf{u}_k = \frac{1}{\sqrt{\lambda_k}} X' \mathbf{v}_k$
- Taking $X \mathbf{u}_k = \sqrt{\lambda_k}\, \mathbf{v}_k$, postmultiplying by $\mathbf{u}_k'$, and summing gives: $X \sum_{k=1}^{n} \mathbf{u}_k \mathbf{u}_k' = \sum_{k=1}^{n} \sqrt{\lambda_k}\, \mathbf{v}_k \mathbf{u}_k'$.
- LHS gives the identity matrix (due to orthogonality of eigenvectors). Hence:
- $X = \sum_{k=1}^{n} \sqrt{\lambda_k}\, \mathbf{v}_k \mathbf{u}_k'$
- This is termed: Karhunen-Loève expansion or transform.
- We can approximate the data, X, by choosing some eigenvalues/vectors only.

Linear Combinations

- The variance of the projections on a given axis in $\mathbb{R}^m$ is given by $(X\mathbf{u})'(X\mathbf{u})$, which by the eigenvector equation, is seen to equal λ.
- In some software packages, the eigenvectors are rescaled so that $\sqrt{\lambda}\,\mathbf{u}$ and $\sqrt{\lambda}\,\mathbf{v}$ are used instead of $\mathbf{u}$ and $\mathbf{v}$. In this case, the *factor* $\sqrt{\lambda}\,\mathbf{u}$ gives the new, rescaled projections of the points in the space $\mathbb{R}^n$ (i.e. $\sqrt{\lambda}\,\mathbf{u} = X'\mathbf{v}$).
- The coordinates of the new axes can be written in terms of the old coordinate system. Since $\mathbf{u} = \frac{1}{\sqrt{\lambda}} X'\mathbf{v}$ each coordinate of the new vector $\mathbf{u}$ is defined as a linear combination of the initially-given vectors:
  $u_j = \sum_{i=1}^{n} \frac{1}{\sqrt{\lambda}} v_i x_{ij} = \sum_{i=1}^{n} c_i x_{ij}$ (where $i \leq j \leq m$ and x_{ij} is the $(i,j)^{th}$ element of matrix X).
- Thus the j^{th} coordinate of the new vector is a *synthetic* value formed from the j^{th} coordinates of the given vectors (i.e. x_{ij} for all $1 \leq i \leq n$).

Normalization or Standardization

- Let r_{ij} be the original measurements.
- Then define: $x_{ij} = \frac{r_{ij} - \overline{r}_j}{s_j \sqrt{n}}$
- $\overline{r}_j = \frac{1}{n} \sum_{i=1}^{n} r_{ij}$
- $s_j^2 = \frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)^2$
- Then the matrix to be diagonalized, $X'X$, is of $(j,k)^{th}$ term:
  $\rho_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} = \frac{1}{n} \sum_{i=1}^{n} (r_{ij} - \overline{r}_j)(r_{ik} - \overline{r}_k)/s_j s_k$
- This is the correlation coefficient between variables j and k.
- Have distance
  $d^2(j,k) = \sum_{i=1}^{n} (x_{ij} - x_{ik})^2 = \sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2 - 2 \sum_{i=1}^{n} x_{ij} x_{ik}$
- First two terms both yield 1. Hence:
- $d^2(j,k) = 2(1 - \rho_{jk})$

- Thus the distance between variables is directly proportional to the correlation between them.
- For row points (objects, observations):
  $d^2(i,h) = \sum_j (x_{ij} - x_{hj})^2 = \sum_j (\frac{r_{ij} - r_{hj}}{\sqrt{n} s_j})^2 = (\mathbf{r}_i - \mathbf{r}_h)' M (\mathbf{r}_i - \mathbf{r}_h)$
- $\mathbf{r}_i$ and $\mathbf{r}_h$ are column vectors (of dimensions $m \times 1$) and M is the $m \times m$ diagonal matrix of j^{th} element $1/ns_j^2$.
- Therefore d is a Euclidean distance associated with matrix M.
- Note that the row points are now centred but the column points are not: therefore the latter may well appear in one quadrant on output listings.

Implications of Standardization

- Analysis of the matrix of $(j,k)^{th}$ term ρ_{jk} as defined above is PCA on a *correlation* matrix.
- The row vectors are centred and reduced.
- Centring alone used, and not the rescaling of the variance: matrix of $(j,k)^{th}$ term $c_{jk} = \frac{1}{n}\sum_{i=1}^{n}(r_{ij} - \overline{r}_j)(r_{ik} - \overline{r}_k)$
- In this case we have PCA of the *variance-covariance* matrix.
- If we use no normalization, we have PCA of the *sums of squares and cross-products* matrix. That was what we used to begin with.
- Usually it is best to carry out analysis on correlations.

Next topic: Introducing the Chi Squared Distance

used for Correspondence Analysis

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.

Next Topic: Close Parallels between Data Analysis

and Classical Mechanics

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: $M^2(N_J(I)) = \sum_{i \in I} f_i \| f_J^i - f_J \|_{f_J}^2 = \sum_{i \in I} f_i \rho^2(i)$.
- 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

- $\sum_{i \in I} f_i F_\alpha(i) = 0; \quad \sum_{j \in J} f_j G_\alpha(j) = 0$
- $\sum_{i \in I} f_i F^2_\alpha(i) = \lambda_\alpha; \quad \sum_{j \in J} f_j G^2_\alpha(j) = \lambda_\alpha$
- $\sum_{i \in I} f_i F_\alpha(i) F_\beta(i) = \delta_{\alpha\beta}$
- $\sum_{j \in J} f_j G_\alpha(j) G_\beta(j) = \delta_{\alpha\beta}$
- 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).
- $\sum_{i \in I} f_i \phi_\alpha(i) = 0; \quad \sum_{j \in J} f_j \psi_\alpha(j) = 0$
- $\sum_{i \in I} f_i \phi^2_\alpha(i) = 1; \quad \sum_{j \in J} f_j \psi^2_\alpha(j) = 1$
- $\sum_{i \in I} f_i \phi_\alpha(i) \phi_\beta(i) = \delta_{\alpha\beta}; \quad \sum_{j \in J} f_j \psi_\alpha(j) \psi_\beta(j) = \delta_{\alpha\beta}$

- 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.
- $\| f_J^i - f_J^{i'} \|_{f_J}^2 = \sum_{j \in J} \left(f_j^i - f_j^{i'} \right)^2 / f_j = \sum_{\alpha=1..N} \left(F_\alpha(i) - F_\alpha(i') \right)^2$
- $\| f_I^j - f_I^{j'} \|_{f_I}^2 = \sum_{i \in I} \left(f_i^j - f_i^{j'} \right)^2 / f_i = \sum_{\alpha=1..N} \left(G_\alpha(j) - G_\alpha(j') \right)^2$
- 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.
- $\rho^2(i) = \| f_J^i - f_J \|_{f_J}^2 = \sum_{\alpha=1..N} F_\alpha^2(i)$
  $\rho^2(j) = \| f_I^j - f_I \|_{f_I}^2 = \sum_{\alpha=1..N} F_\alpha^2(j)$

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.
- $\rho^2(i) = \|f^i_J - f_J\|^2 = \sum_{j \in J} \left(f^i_j - f_j\right)^2 / f_j$
- 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.
- $\lambda_\alpha = \sum_{i \in I} f_i F^2_\alpha(i)$ and $\rho^2(i) = \sum_{\alpha=1..N} F^2_\alpha(i)$

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.
- $\mathrm{QLT}(i) = \sum_{\alpha=1..N'} \cos^2 a$, 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'.
- $\mathrm{INR}(I) = \rho^2(i)$ is the distance of element I from the centre of gravity of the cloud.
- $\mathrm{POID}(I) = f_i$ 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).

Data Input Coding in Correspondence Analysis

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.)

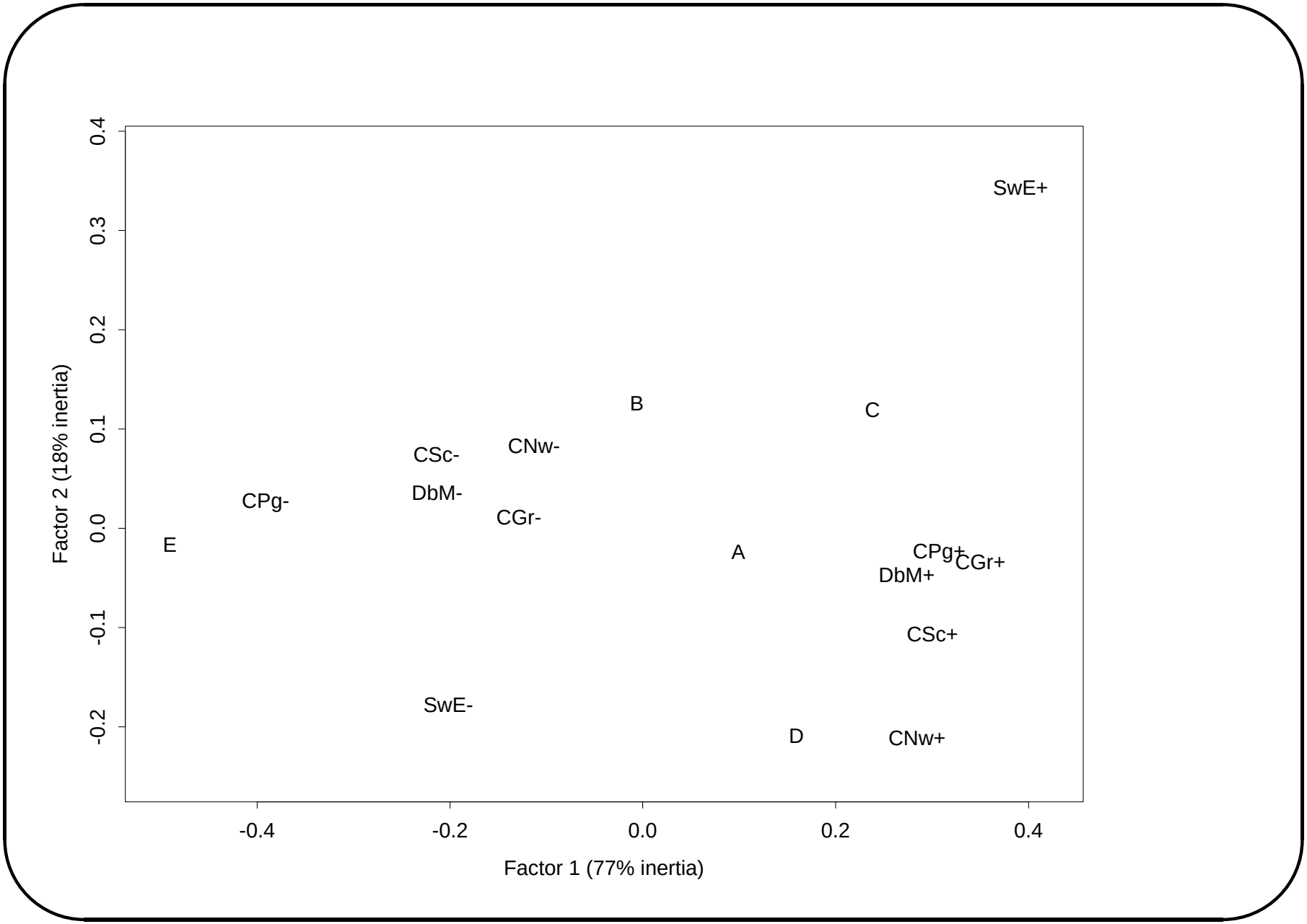
- Correspondence analysis highlights the similarities and the differences in the profiles.
- Note that all the scores of D and E are in the same proportion (E's scores are one-third those of D).
- Note also that E has the lowest scores both in absolute and relative terms in all the subjects.
- D and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

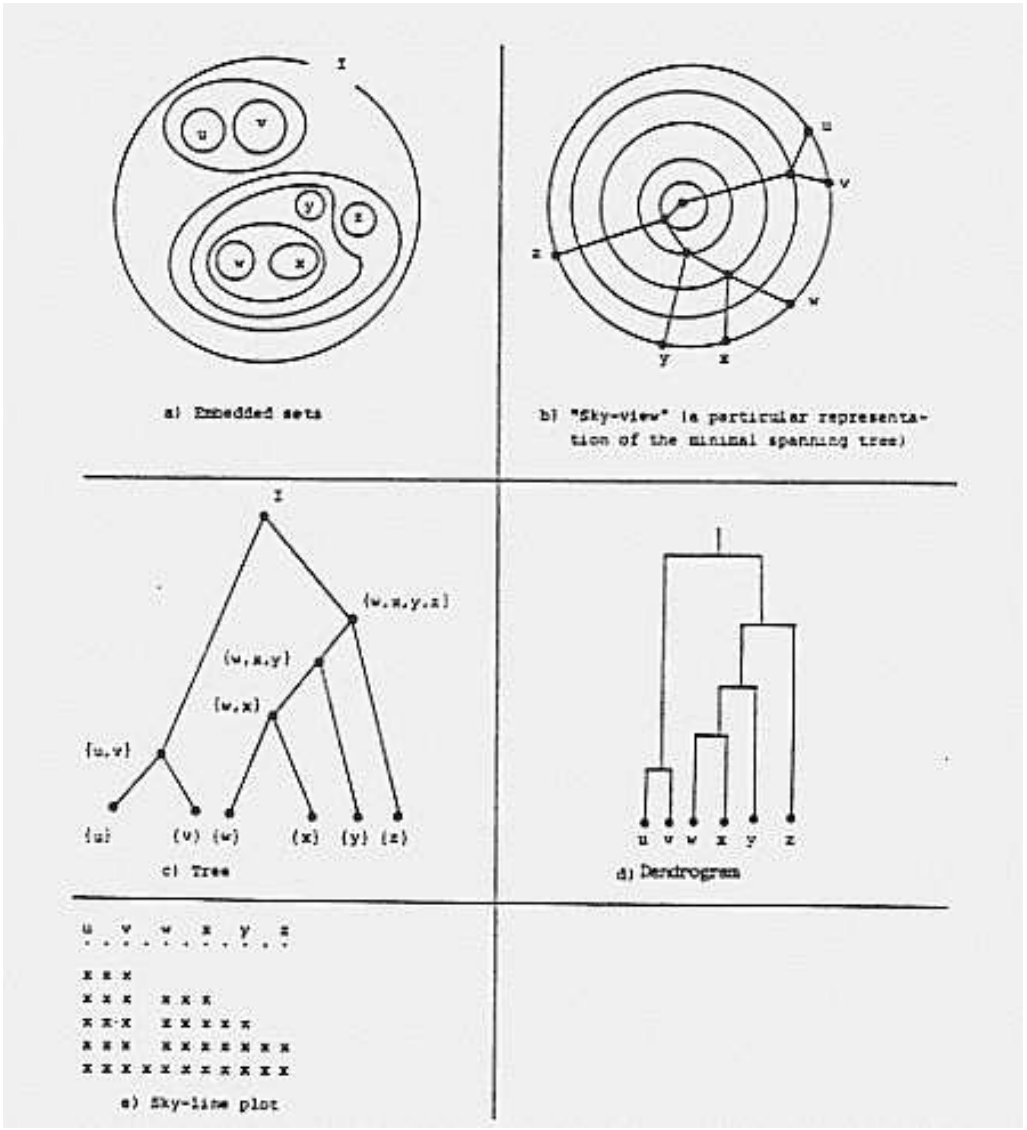
Data coding: Doubling

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

Next Topic: Hierarchical Clustering


I
u
v
y
z
w
x
a) Embedded sets
u
v
z
w
y
x
b) "Sky-view" (a particular representation of the minimal spanning tree)
I
{w,x,y,z}
{w,x,y}
{w,x}
{u,v}
{u}
{v}
{w}
{x}
{y}
{z}
c) Tree
u v w x y z
d) Dendrogram
u v w x y z
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 with small base, or equilateral. An ultrametric is a special distance associated with rooted trees. Ultrametric topology is 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.

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.

Hierarchical Cluster Analysis

Topics:

- Part 2: Hierarchical agglomerative cluster analysis. Using metric embedding.
- Example: globular cluster study (PCA and clustering).
- Metric and distance.
- Hierarchical agglomerative clustering.
- Single link, minimum variance criterion.
- Graph methods – minimal spanning tree, Voronoi diagram

Cluster Analysis

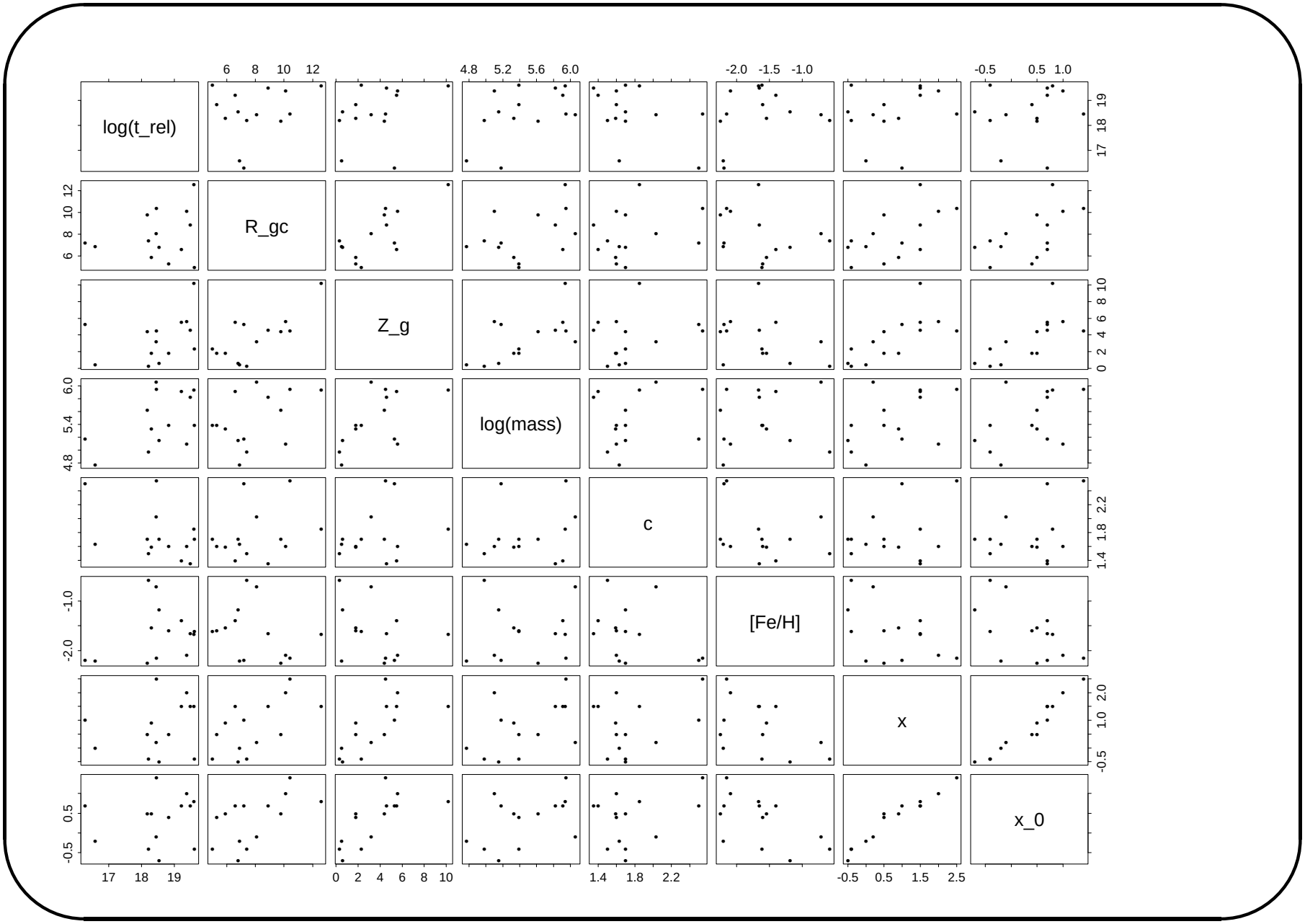
Some Terms

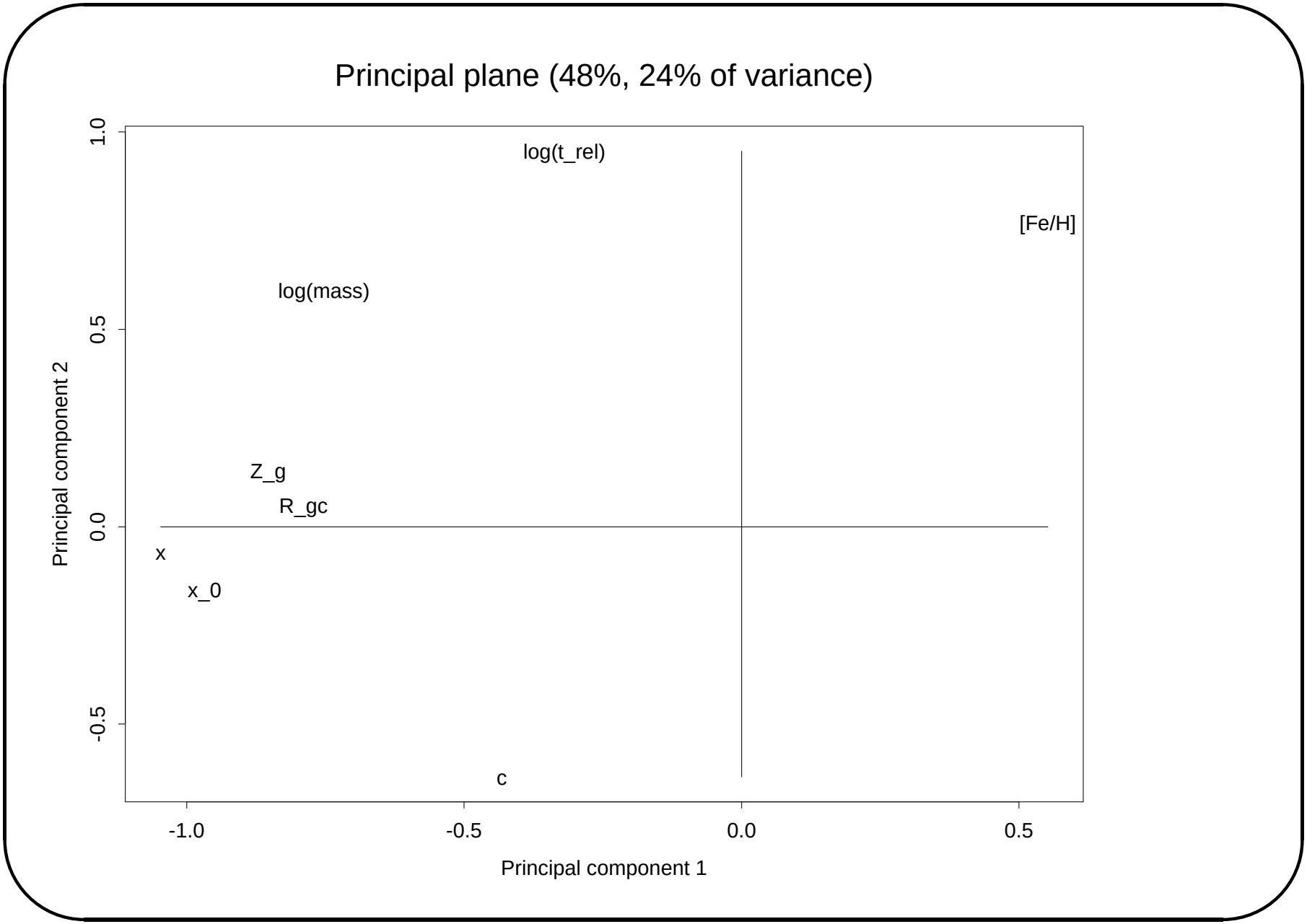
- Unsupervised classification, clustering, cluster analysis, automatic classification. Versus: Supervised classification, discrimant analysis, trainable classifier, machine learning.
- For clustering we will consider (i) partitioning methods, (ii) agglomerative hierarchical classification, (iii) graph methods, (iv) statistical methods, or distribution mixture models.
- For discrimination one can consider consider (i) multiple discriminant analysis (geometric), (ii) nearest neighbour discriminant analysis, (iii) neural networks – multilayer perceptron, (iv) machine learning methods, and (v) classification trees.
- Note that principal components analysis, correspondence analysis, or indeed visualization display methods, can be used as a basis for clustering.

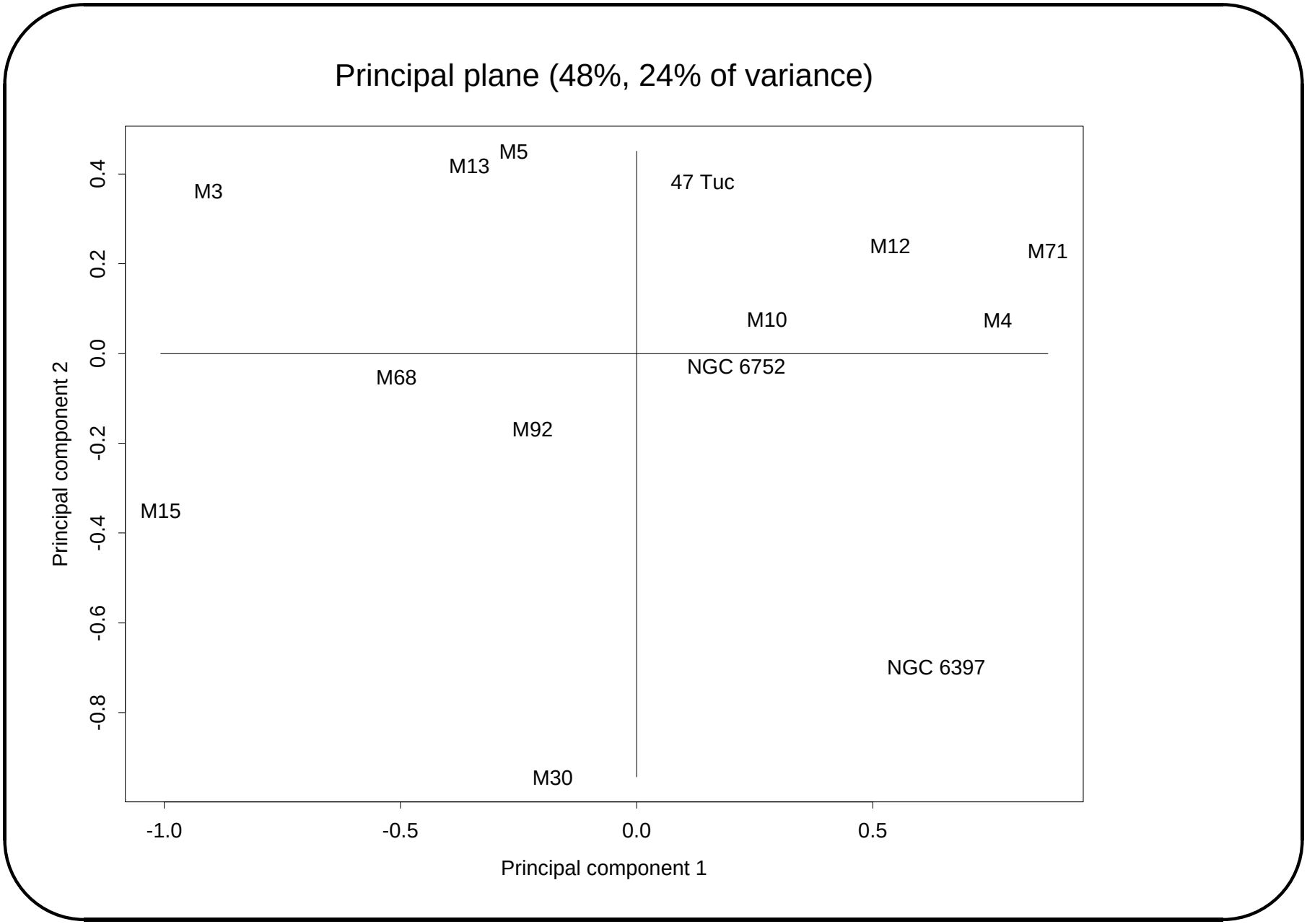
Example: analysis of globular clusters

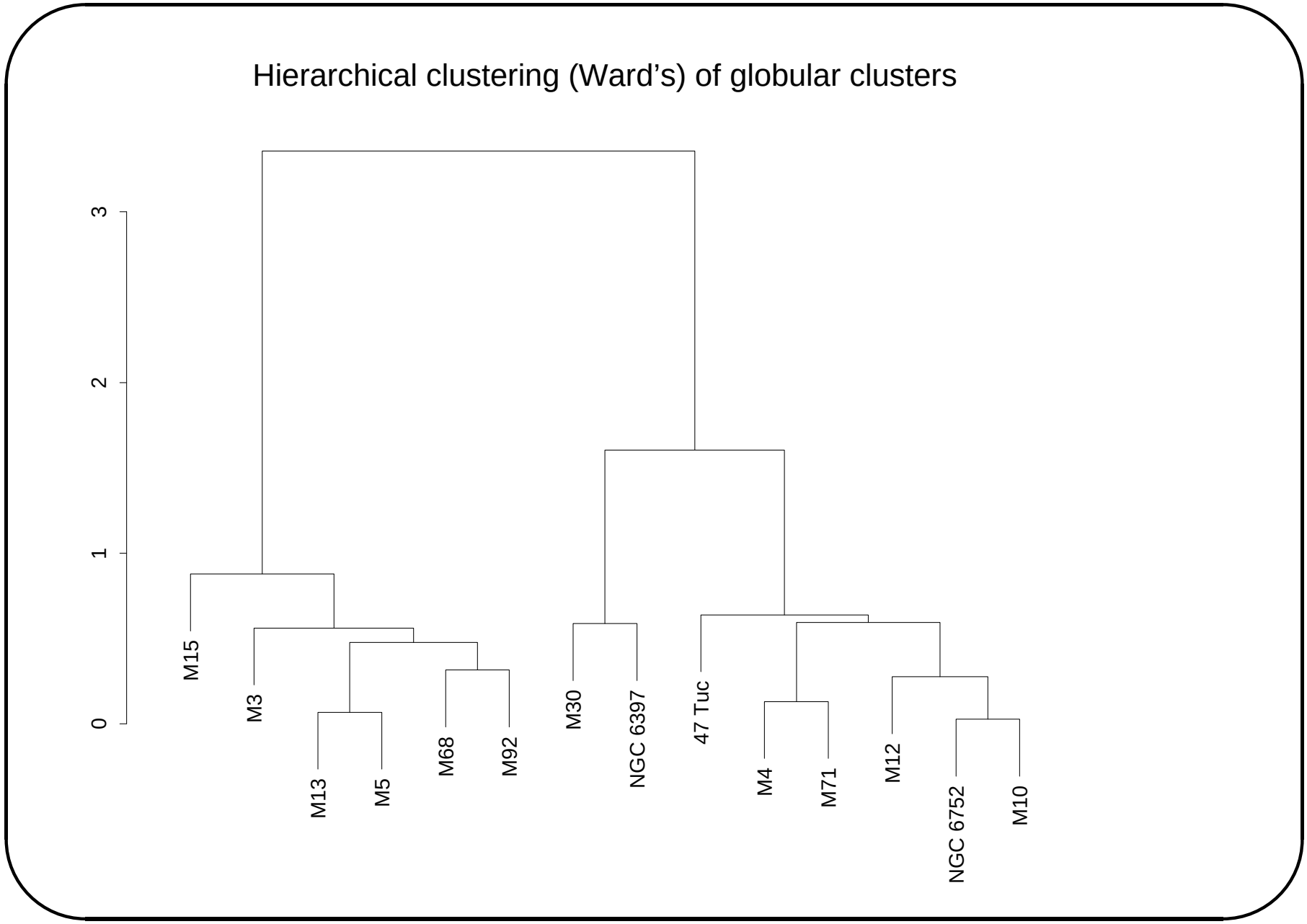
- M. Capaccioli, S. Ortolani and G. Piotto, "Empirical correlation between globular cluster parameters and mass function morphology", AA, 244, 298–302, 1991.
- 14 globular clusters, 8 measurement variables.
- Data collected in earlier CCD (digital detector) photometry studies.
- Pairwise plots of the variables.
- PCA of the variables.
- PCA of the objects (globular clusters).

Object	t_rlx years	Rgc Kpc	Zg Kpc	log(M/ M.)	c	[Fe/H]	x	x0
M15	1.03e+8	10.4	4.5	5.95	2.54	-2.15	2.5	1.4
M68	2.59e+8	10.1	5.6	5.1	1.6	-2.09	2.0	1.0
M13	2.91e+8	8.9	4.6	5.82	1.35	-1.65	1.5	0.7
M3	3.22e+8	12.6	10.2	5.94	1.85	-1.66	1.5	0.8
M5	2.21e+8	6.6	5.5	5.91	1.4	-1.4	1.5	0.7
M4	1.12e+8	6.8	0.6	5.15	1.7	-1.28	-0.5	-0.7
47 Tuc	1.02e+8	8.1	3.2	6.06	2.03	-0.71	0.2	-0.1
M30	1.18e+7	7.2	5.3	5.18	2.5	-2.19	1.0	0.7
NGC 6397	1.59e+7	6.9	0.5	4.77	1.63	-2.2	0.0	-0.2
M92	7.79e+7	9.8	4.4	5.62	1.7	-2.24	0.5	0.5
M12	3.26e+8	5.0	2.3	5.39	1.7	-1.61	-0.4	-0.4
NGC 6752	8.86e+7	5.9	1.8	5.33	1.59	-1.54	0.9	0.5
M10	1.50e+8	5.3	1.8	5.39	1.6	-1.6	0.5	0.4
M71	8.14e+7	7.4	0.3	4.98	1.5	-0.58	-0.4	-0.4


log(t_rel)
R_gc
Z_g
log(mass)
c
[Fe/H]
x
x_0


Principal plane (48%, 24% of variance)
log(t_rel)
[Fe/H]
log(mass)
Z_g
R_gc
x
x_0
c
Principal component 2
Principal component 1
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5


Principal plane (48%, 24% of variance)
M5
M13
M3
47 Tuc
M12
M71
M10
M4
NGC 6752
M68
M92
M15
NGC 6397
M30
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
Principal component 2
-1.0
-0.5
0.0
0.5
Principal component 1


Hierarchical clustering (Ward's) of globular clusters
0
1
2
3
M15
M3
M13
M5
M68
M92
M30
NGC 6397
47 Tuc
M4
M71
M12
NGC 6752
M10

Metric and Ultrametric

- Triangular inequality:

  Symmetry: $d(a,b) = d(b,a)$

  Positive semi-definiteness: $d(a,b) > 0,\ \text{if } a \neq b; d(a,b) = 0,\ \text{if } a = b$

  Triangular inequality: $d(a,b) \leq d(a,c) + d(c,b)$

- Ultrametric inequality: $d(a,b) \leq \max(d(a,c) + d(c,b))$
- Minkowski metric: $d_p(a,b) = \sqrt[p]{\sum_j \mid a_j - b_j \mid^p} \qquad p \geq 1.$
- Particular cases of the Minkowski metric: $p = 2$ gives Euclidean, $p = 1$ gives Hamming or city-block; and $= \infty$ gives $d_\infty(a,b) = max_j \mid a_j - b_j \mid$ which is the "maximum coordinate" or *Chebyshev* distance.
- Also termed L_2, L_1, and L_∞ distances.
- Question: show that squared Euclidean and Hamming distances are the same for binary data.

Single Linkage Hierarchical Clustering

Dissimilarity matrix defined for 5 objects

	1	2	3	4	5
1	0	4	9	5	8
2	4	0	6	3	6
3	9	6	0	6	3
4	5	3	6	0	5
5	8	6	3	5	0

Agglomerate 2 and 4 at dissimilarity 3

	1	2U4	3	5
1	0	4	9	8
2U4	4	0	6	5
3	9	6	0	3
5	8	5	3	0

Agglomerate 3 and 5 at dissimilarity 3

Single Linkage Hierarchical Clustering – 2

	1	2U4	3U5
1	0	4	8
2U4	4	0	5
3U5	8	5	0

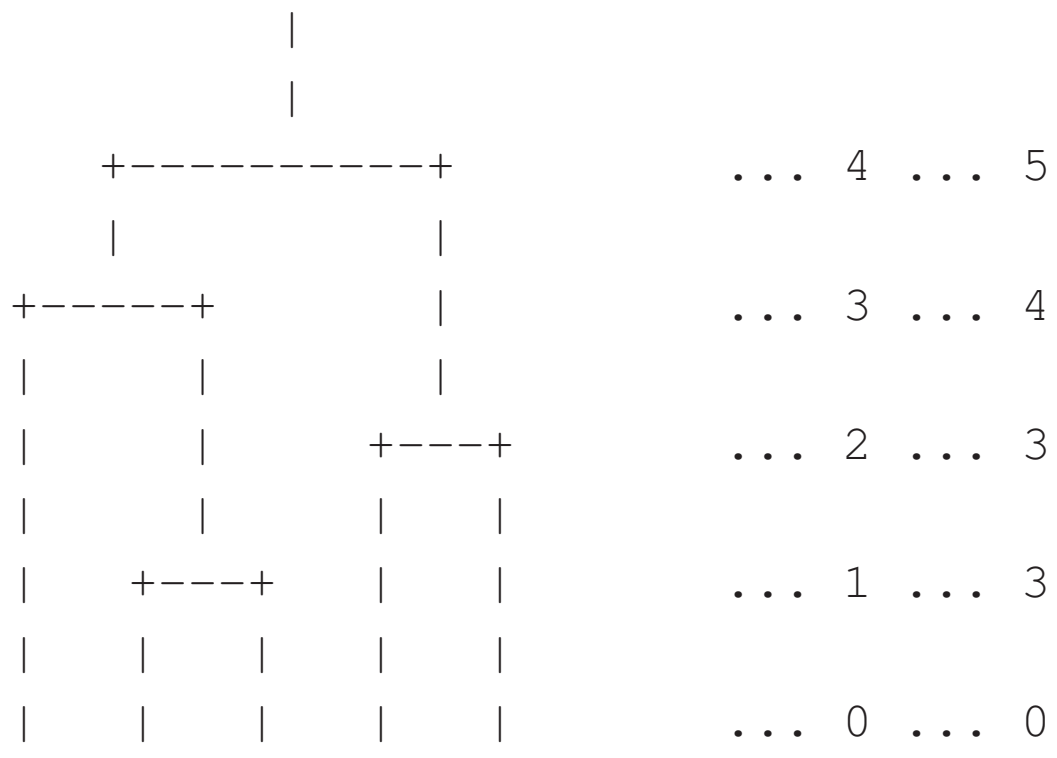
Agglomerate 1 and 2U4 at dissimilarity 4

	1U2U4	3U5
1U2U4	0	5
3U5	5	0

Finally agglomerate 1U2U4 and 3U5 at dissim. 5

Single Linkage Hierarchical Clustering – 3

Resulting dendrogram



r = ranks or levels. c = criterion values (linkage wts).

Single Linkage Hierarchical Clustering – 3

Input An $n(n-1)/2$ set of dissimilarities.

Step 1 Determine the smallest dissimilarity, d_{ik}.

Step 2 Agglomerate objects i and k: i.e. replace them with a new object, $i \cup k$; update dissimilarities such that, for all objects $j \neq i, k$:

$$d_{i \cup k, j} = \min \{d_{ij}, d_{kj}\}.$$

Delete dissimilarities d_{ij} and d_{kj}, for all j, as these are no longer used.

Step 3 While at least two objects remain, return to Step 1.

Single Linkage Hierarchical Clustering – 4

- Precisely $n - 1$ levels for n objects. Ties settled arbitrarily.
- Note single linkage criterion.
- Disadvantage: chaining. "Friends of friends" in the same cluster.
- Lance-Williams cluster update formula:
  $d(i \cup j, k) = \alpha_i d(i, k) + \alpha_j d(j, k) + \beta d(i, j) + \gamma \mid d(i, k) - d(j, k) \mid$ where coefficients α_i, α_j, β, and γ define the agglomerative criterion.
- For single link, $\alpha_i = 0.5$, $\beta = 0$ and $\gamma = -0.5$.
- These values always imply: $\min\{d_{ik}, d_{jk}\}$
- Ultrametric distance, δ, resulting from the single link method is such that $\delta(i, j) \leq d(i, j)$ always. It is also unique (with the exception of ties). So single link is also termed the subdominant ultrametric method.

Other Hierarchical Clustering Criteria

- Complete link: substitute max for min in single link.
- Complete link leads to compact clusters.
- Single link defines the cluster criterion from the closest object in the cluster. Complete link defines the cluster criterion from the furthest object in the cluster.
- Complete link yields a *minimal superior ultrametric*. Unfortunately this is not unique (as is the *maximal inferior ultrametric*, or *subdominant ultrametric*).
- Other criteria define $d(i \cup j, k)$ from the distance between k and something closer to the mean of i and j. These criteria include the median, centroid and minimum variance methods.
- A problem that can arise: inversions in the hierarchy. I.e. the cluster criterion value is not monotonically increasing. That leads to cross-overs in the dendrogram.

- Of the above agglomerative methods, the single link, complete link, and minimum variance methods can be shown to never allow inversions. They satisfy the *reducibility property*.

Hierarchical clustering methods (and aliases).	Lance and Williams dissimilarity update formula.	Coordinates of centre of cluster, which agglomerates clusters i and j.	Dissimilarity between cluster centres g_i and g_j.
Single link (nearest neighbour).	$\alpha_i = 0.5$ $\beta = 0$ $\gamma = -0.5$ (More simply: $min\{d_{ik}, d_{jk}\}$)		
Complete link (diameter).	$\alpha_i = 0.5$ $\beta = 0$ $\gamma = 0.5$ (More simply: $max\{d_{ik}, d_{jk}\}$)		
Group average (average link, UPGMA).	$\alpha_i = \frac{\lvert i \rvert}{\lvert i \rvert + \lvert j \rvert}$ $\beta = 0$ $\gamma = 0$		

Hierarchical clustering methods (and aliases).	Lance and Williams dissimilarity update formula.	Coordinates of centre of cluster, which agglomerates clusters i and j.	Dissimilarity between cluster centres g_i and g_j.
Median method (Gower's, WPGMC).	$\alpha_i = 0.5$ $\beta = -0.25$ $\gamma = 0$	$\mathbf{g} = \frac{\mathbf{g}_i + \mathbf{g}_j}{2}$	$\|\mathbf{g}_i - \mathbf{g}_j\|^2$
Centroid (UPGMC).	$\alpha_i = \frac{\|i\|}{\|i\|+\|j\|}$ $\beta = -\frac{\|i\|\|j\|}{(\|i\|+\|j\|)^2}$ $\gamma = 0$	$\mathbf{g} = \frac{\|i\|\mathbf{g}_i + \|j\|\mathbf{g}_j}{\|i\|+\|j\|}$	$\|\mathbf{g}_i - \mathbf{g}_j\|^2$
Ward's method (minimum variance, error sum of squares.	$\alpha_i = \frac{\|i\|+\|k\|}{\|i\|+\|j\|+\|k\|}$ $\beta = -\frac{\|k\|}{\|i\|+\|j\|+\|k\|}$ $\gamma = 0$	$\mathbf{g} = \frac{\|i\|\mathbf{g}_i + \|j\|\mathbf{g}_j}{\|i\|+\|j\|}$	$\frac{\|i\|\|j\|}{\|i\|+\|j\|}\|\mathbf{g}_i - \mathbf{g}_j\|^2$

Agglomerative Algorithm Based on Data

(i.e. directly on data, rather than directly on dissimilarities

Step 1 Examine all interpoint dissimilarities, and form cluster from two closest points.

Step 2 Replace two points clustered by representative point (centre of gravity) or by cluster fragment.

Step 3 Return to Step 1, treating clusters as well as remaining objects, until all objects are in one cluster.

Agglomerative Algorithm Based on Dissimilarities

Step 1 Form cluster from smallest dissimilarity.

Step 2 Define cluster; remove dissimilarity of agglomerated pair. Update dissimilarities from cluster to all other clusters/singletons.

Step 3 Return to Step 1, treating clusters as well as remaining objects, until all objects are in one cluster.

Example of Similarities

- Jaccard coefficient for binary vectors $\mathbf{a}$ and $\mathbf{b}$. N is counting operator: $s(a,b) = \frac{N_j(a_j{=}b_j{=}1)}{N_j(a_j{=}1)+N_j(b_j{=}1)-N_j(a_j{=}b_j{=}1)}$
- Jaccard similarity coefficient of vectors (10001001111) and (10101010111) is $5/(6+7-5) = 5/8$. In vector notation: $s(a,b) = \frac{\mathbf{a'b}}{\mathbf{a'a+b'b-a'b}}$.
- Note: max sim. value - sim. = dissim.
- Jaccard coefficient uses counts of presence/absences in cross-tabulation of binary presence/absence vectors:

	a/present	a/absent
b/present	n1	n2
b/absent	n3	n4

- A number of such measures have been used in information retrieival, or numerical taxonomy: Jaccard, Dice, Tanimoto, ...

- Another example based on coding of data:

Record x:	S1, 18.2, X
Record y:	S1, 6.7, —

Two records (x and y) with three variables (Seyfert type, magnitude, X-ray emission) showing disjunctive coding.

	Seyfert type spectrum				Integrated magnitude		X–ray data?
	S1	S2	S3	—	≤ 10	> 10	Yes
x	1	0	0	0	0	1	1
y	1	0	0	0	1	0	0

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.

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

- 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:

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

Graph Methods

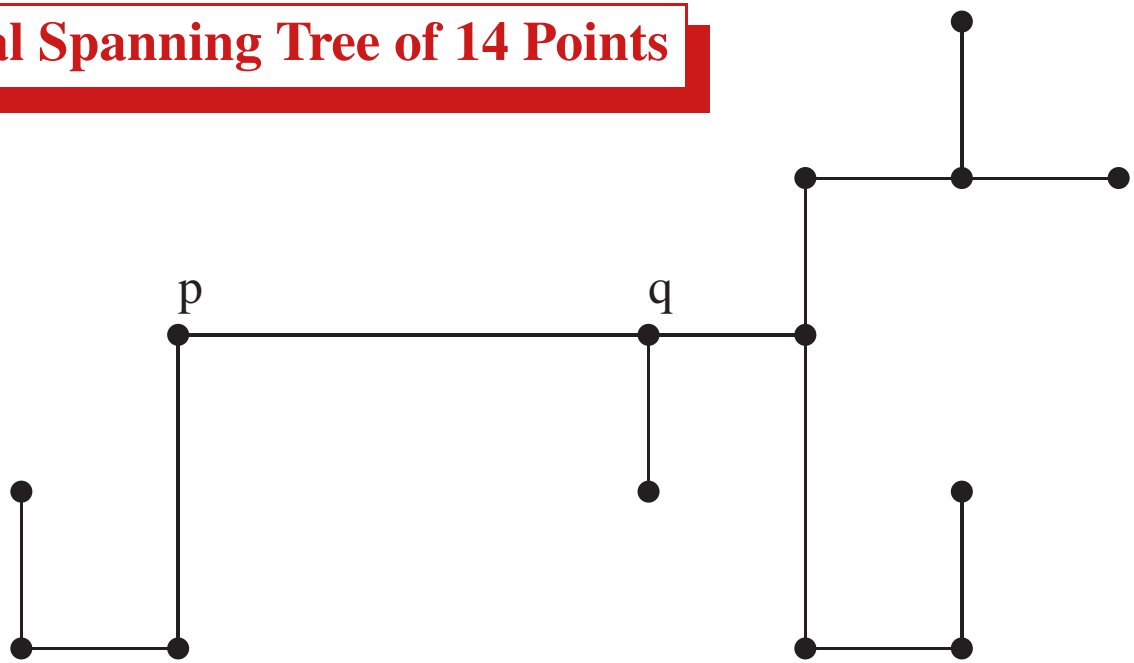
Minimal Spanning Tree

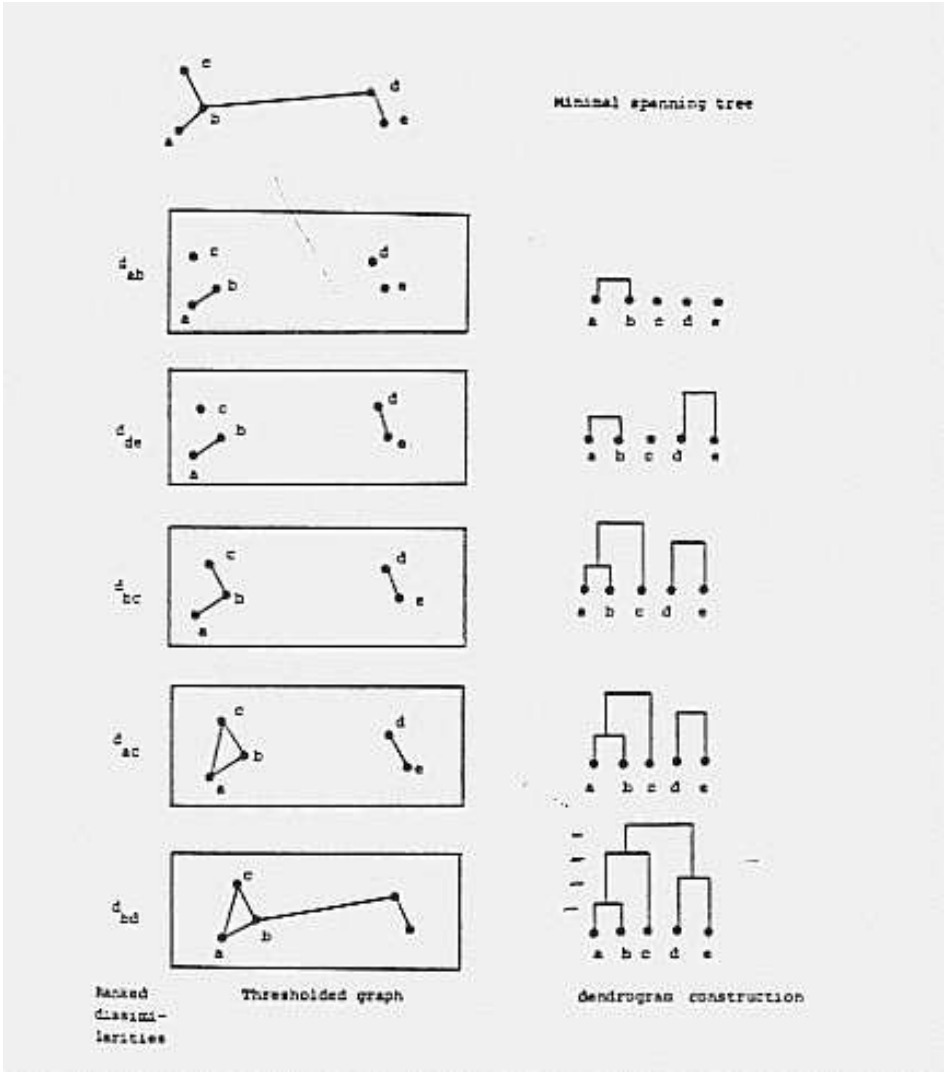
Step 1 Select an arbitrary point and connect it to the least dissimilar neighbour. These two points constitute a subgraph of the MST.

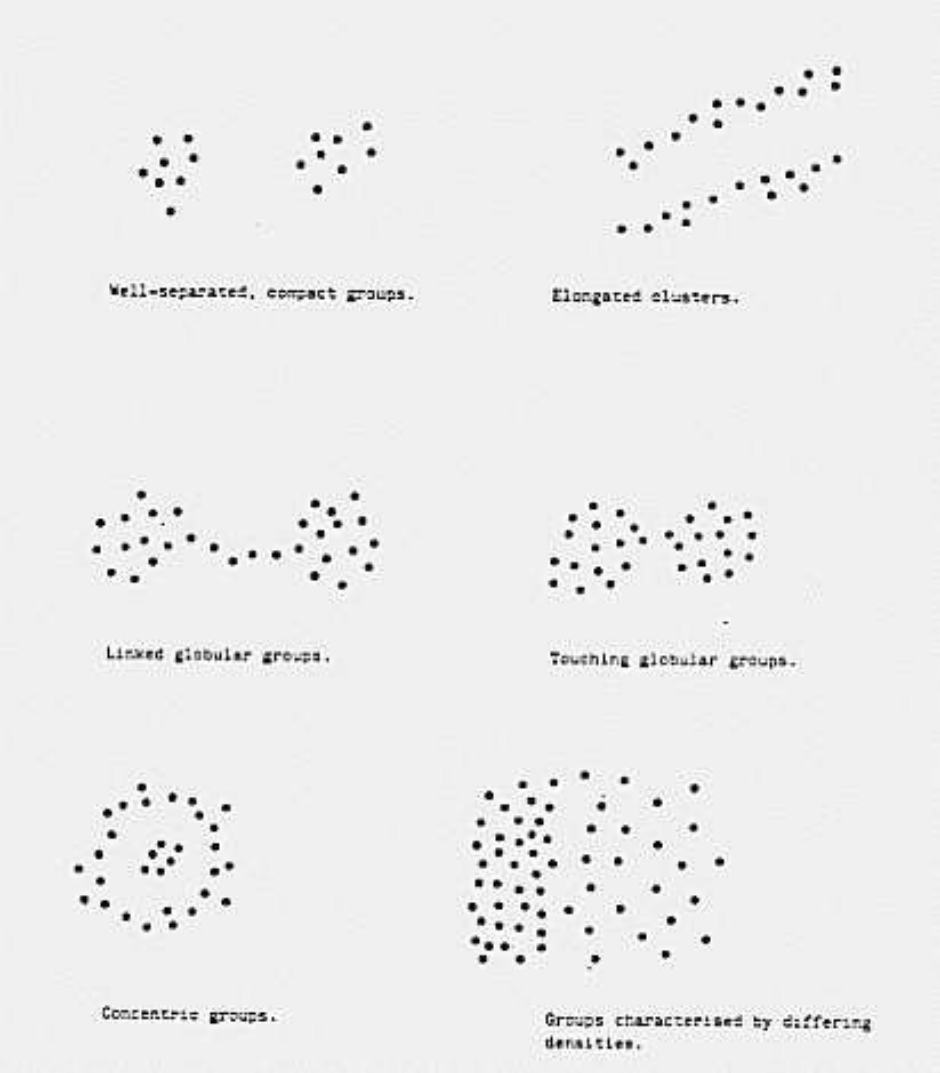
Step 2 Connect the current subgraph to the least dissimilar neighbour of any of the members of the subgraph.

Step 3 Loop on Step 2, until all points are in the one subgraph: this, then, is the MST.

Minimal Spanning Tree of 14 Points



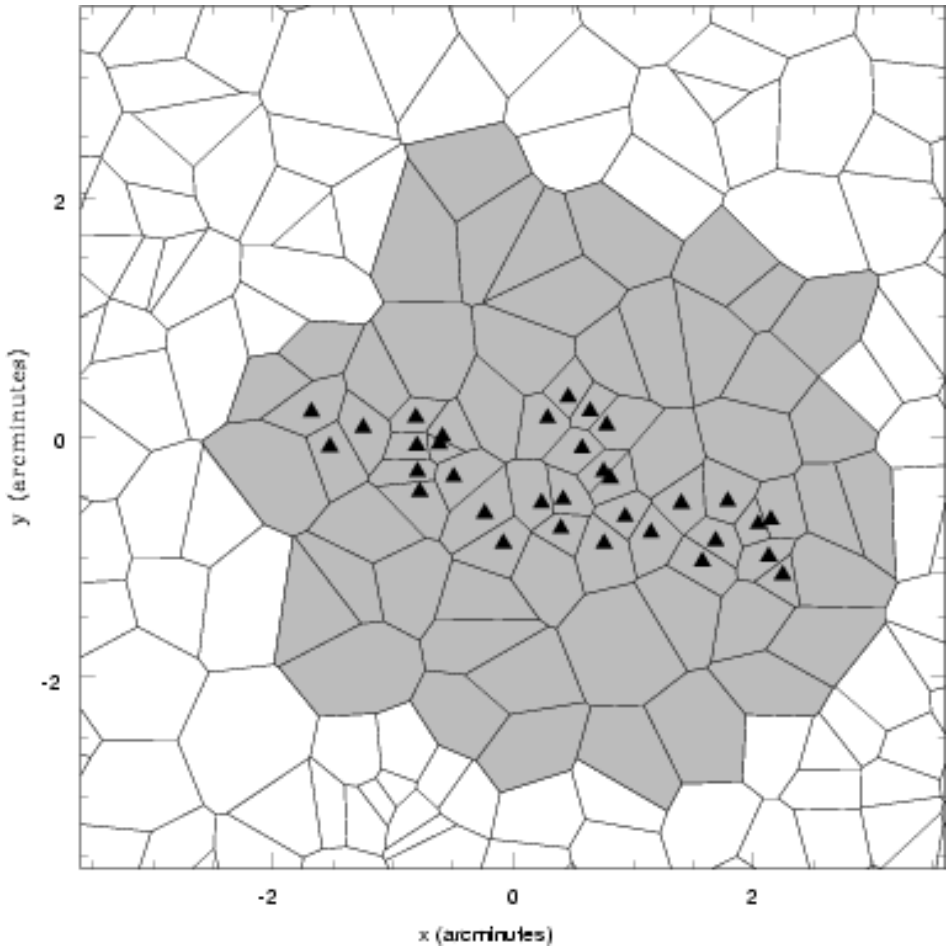

Minimal spanning tree
Ranked dissimilarities
Thresholded graph
dendrogram construction


Well-separated, compact groups.
Elongated clusters.
Linked globular groups.
Touching globular groups.
Concentric groups.
Groups characterised by differing densities.

Voronoi Diagram

- M. Ramella, W. Boschin, D. Fadda and M. Nonino, Finding galaxy clusters using Voronoi tessellations, A&A 368, 776-786 (2001)
- For lots on Voronoi diagrams: http://www.voronoi.com/cgi-bin/display.voronoi_applications.php?cat=Applications
- Voronoi diagram: for given points i, we define the Voronoi cell or region of i as $\{x | d(x,i) \leq d(x,i')\}\ \forall i'$.
- Delaunay triangulation: perpendicular bisectors of Voronoi boundaries.
- Theorem: MST $\subset$ Delaunay triangulation.

Voronoi Diagram



Some galaxies shown.

Partitioning

Iterative optimization algorithm for the variance criterion

Step 1 Arbitrarily define a set of k cluster centres.

Step 2 Assign each object to the cluster to which it is closest (using the Euclidean distance, $d^2(i,p) = \|\mathbf{i} - \mathbf{p}\|^2$).

Step 3 Redefine cluster centres on the basis of the current cluster memberships.

Step 4 If the totalled within class variances is better than at the previous iteration, then return to Step 2.

Partitioning – Properties

- Sub-optimal.
- Dependent on initial cluster centres.
- The two main steps define the EM algorithm. "Expectation": determine mean; "Maximization": assignment step.
- Widely used (since computational cost of hierarchical clustering is usually $O(n^2)$).