Cluster Analysis

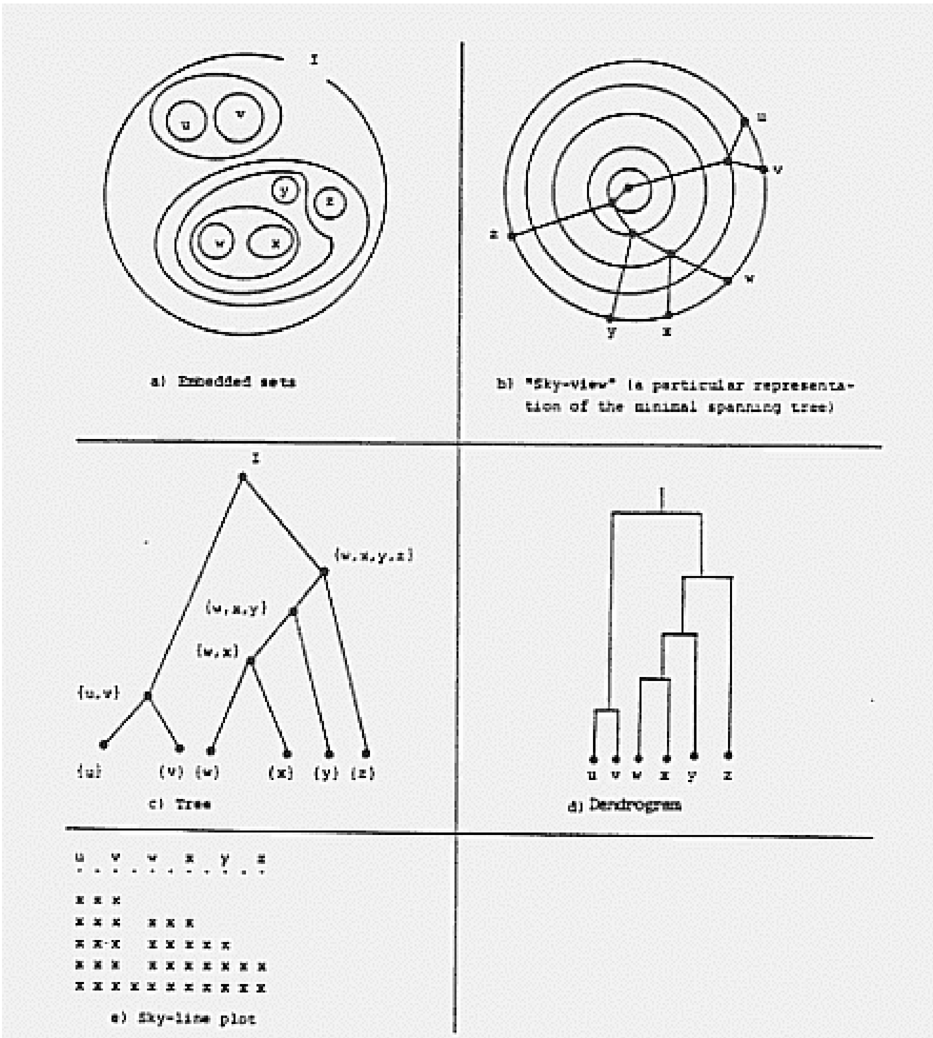
Topics:

- 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
- Distribution mixture modelling – Bayes factors
- Kohonen self-organizing maps
- Examples: BATSE gamma ray bursts – numbers of classes; interactive visual user interfaces.
- Software: http://astro.u-strasbg.fr/∼fmurtagh/mda-sw

Cluster Analysis

Some Terms

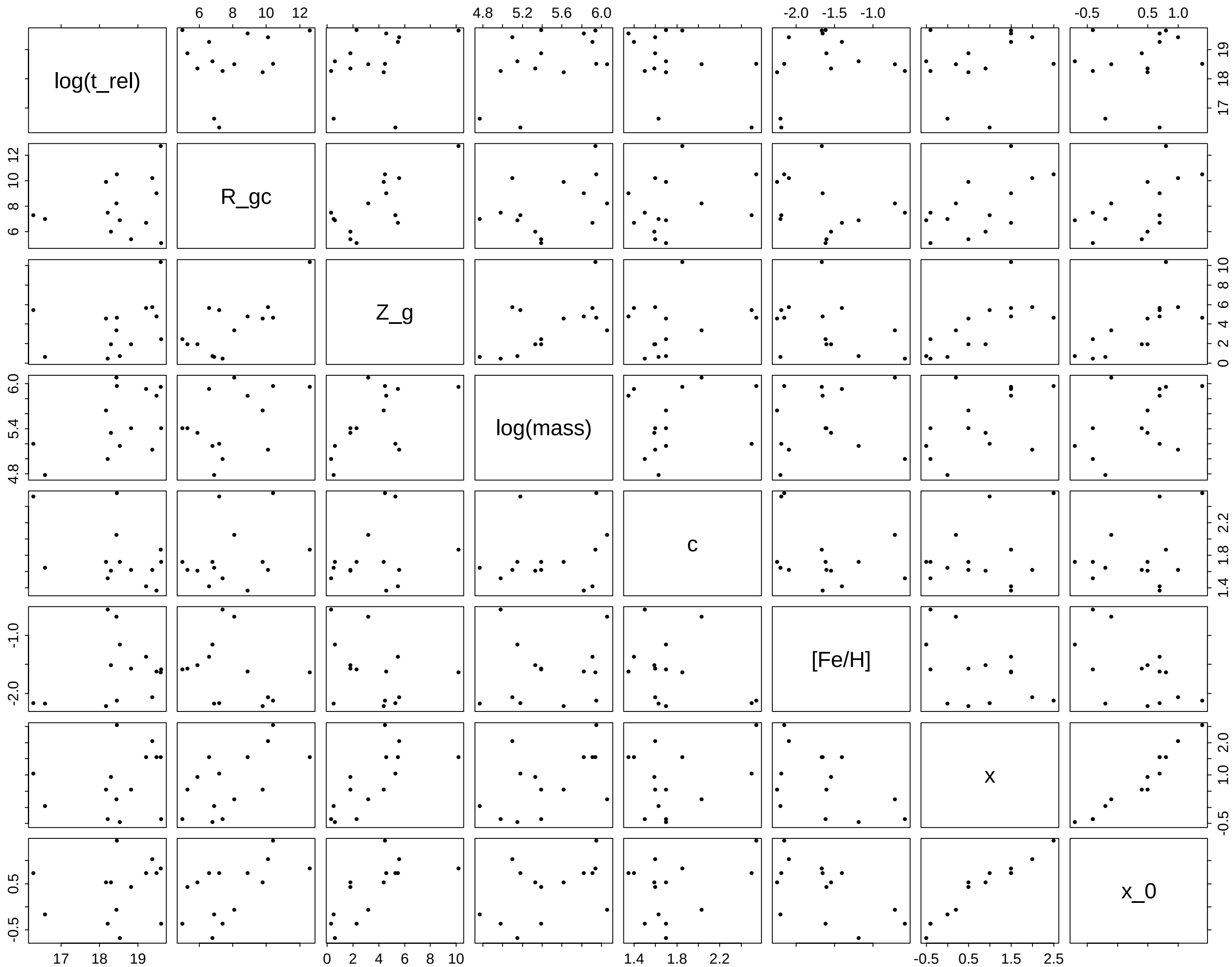
- Unsupervised classification, clustering, cluster analysis, automatic classification. Versus: Supervised classification, discrimant analysis, trainable classifier, machine learning.
- For clustering we will consider (i) partitioning methods, (ii) agglomerative hierarchical classification, (iii) graph methods, (iv) statistical methods, or distribution mixture models, (v) Kohonen self-organizing feature map.
- Later for discrimination we will 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 for clustering.

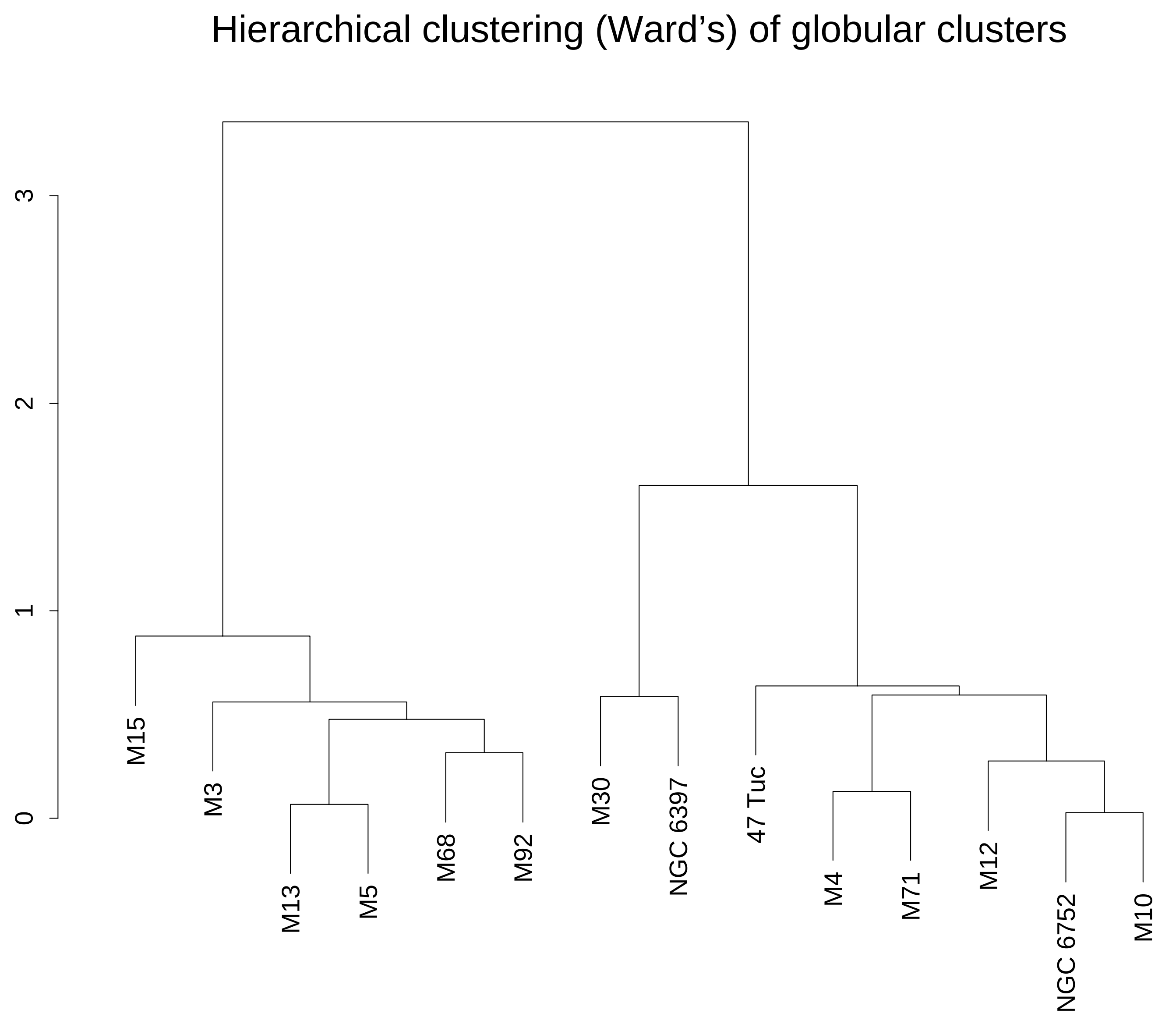

I
u
v
y
z
w
x
a) Embedded sets
u
v
z
w
y
x
b) "Sky-view" (a particular representation of the minimal spanning tree)
I
{w,x,y,z}
{w,x,y}
{w,x}
{u,v}
{u} {v} {w} {x} {y} {z}
c) Tree
u v w x y z
d) Dendrogram
u v w x y z
e) Sky-line plot

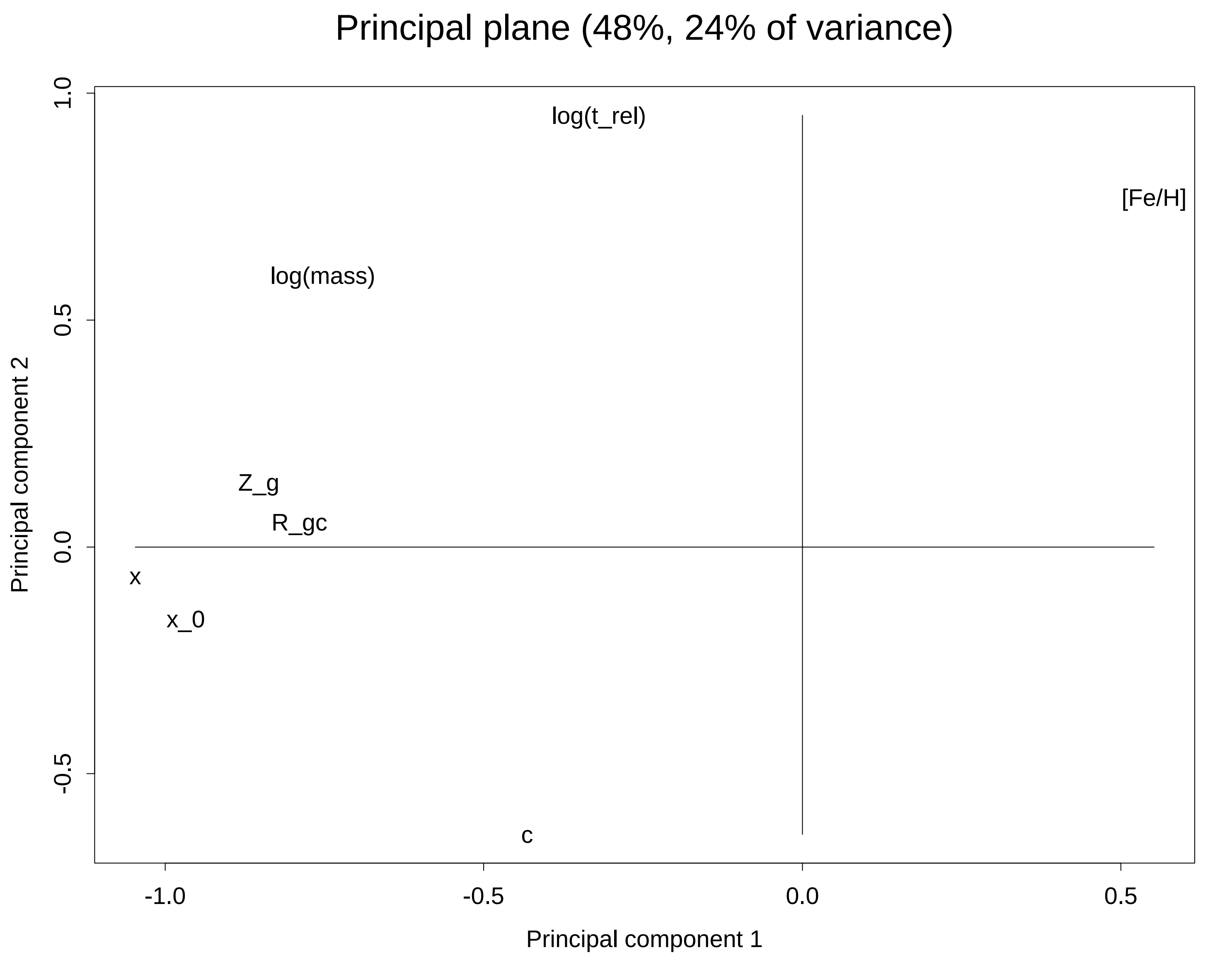
Example: analysis of globular clusters

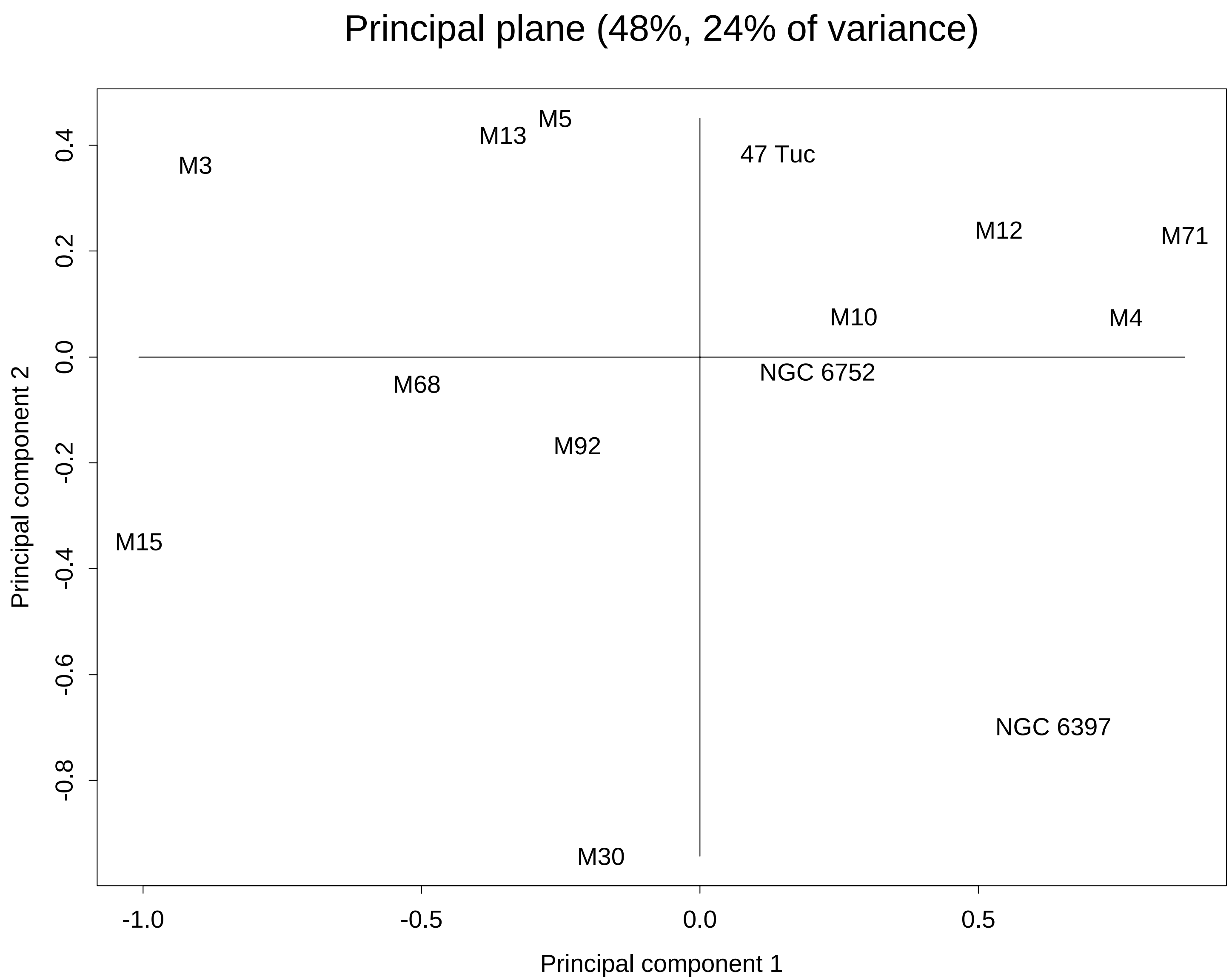
- M. Capaccioli, S. Ortolani and G. Piotto, "Empirical correlation between globular cluster parameters and mass function morphology", AA, 244, 298–302, 1991.
- 14 globular clusters, 8 measurement variables.
- Data collected in earlier CCD (digital detector) photometry studies.
- Pairwise plots of the 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
6 8 10 12
4.8 5.2 5.6 6.0
-2.0 -1.5 -1.0
-0.5 0.5 1.0
17 18 19
0 2 4 6 8 10
1.4 1.8 2.2
-0.5 0.5 1.5 2.5


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


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


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
-1.0
-0.5
0.0
0.5
Principal component 1
Principal component 2

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.

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: $d_p(a, b) = \sqrt[p]{\sum_j \mid a_j - b_j \mid^p} \quad p \geq 1.$
- 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     c
                              |
                              |
                        +----------+         ... 4 ... 5
                        |          |
                     +-----+       |         ... 3 ... 4
                     |     |       |
                     |     |     +---+       ... 2 ... 3
                     |     |     |   |
                     |   +---+   |   |       ... 1 ... 3
                     |   |   |   |   |
                     |   |   |   |   |       ... 0 ... 0

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).	$\alpha_i = \frac{\|i\|}{\|i\|+\|j\|}$ $\beta = 0$ $\gamma = 0$		

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

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: $s(a,b) = \frac{\mathbf{a}'\mathbf{b}}{\mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - \mathbf{a}'\mathbf{b}}$.
- 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.

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

- 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:

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

Graph Methods

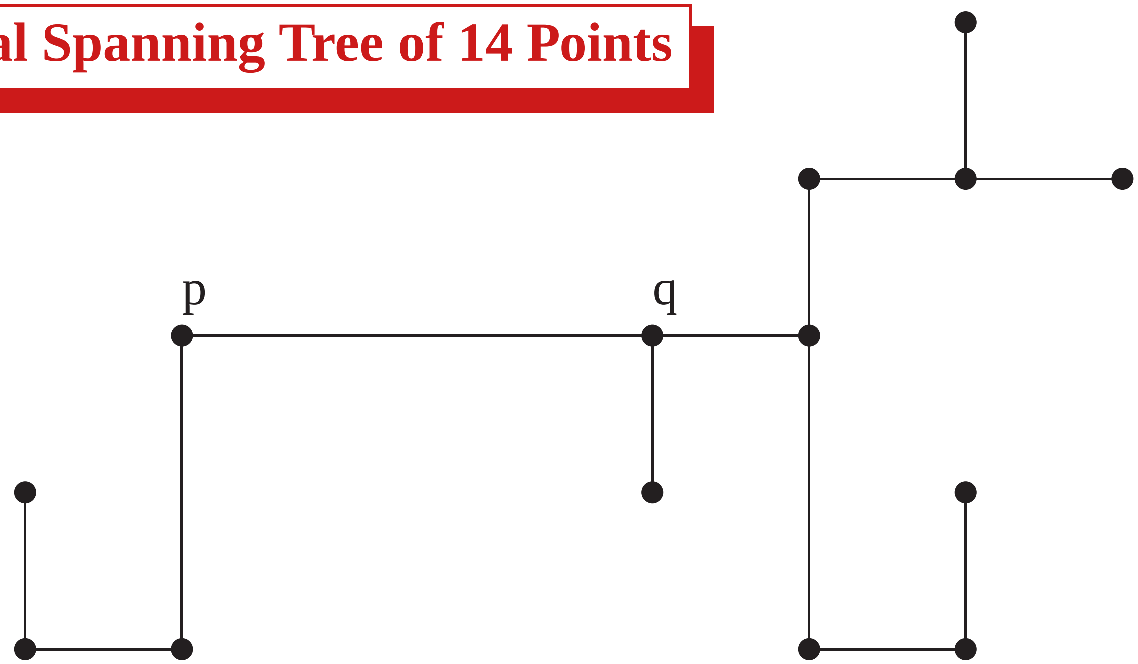
Minimal Spanning Tree

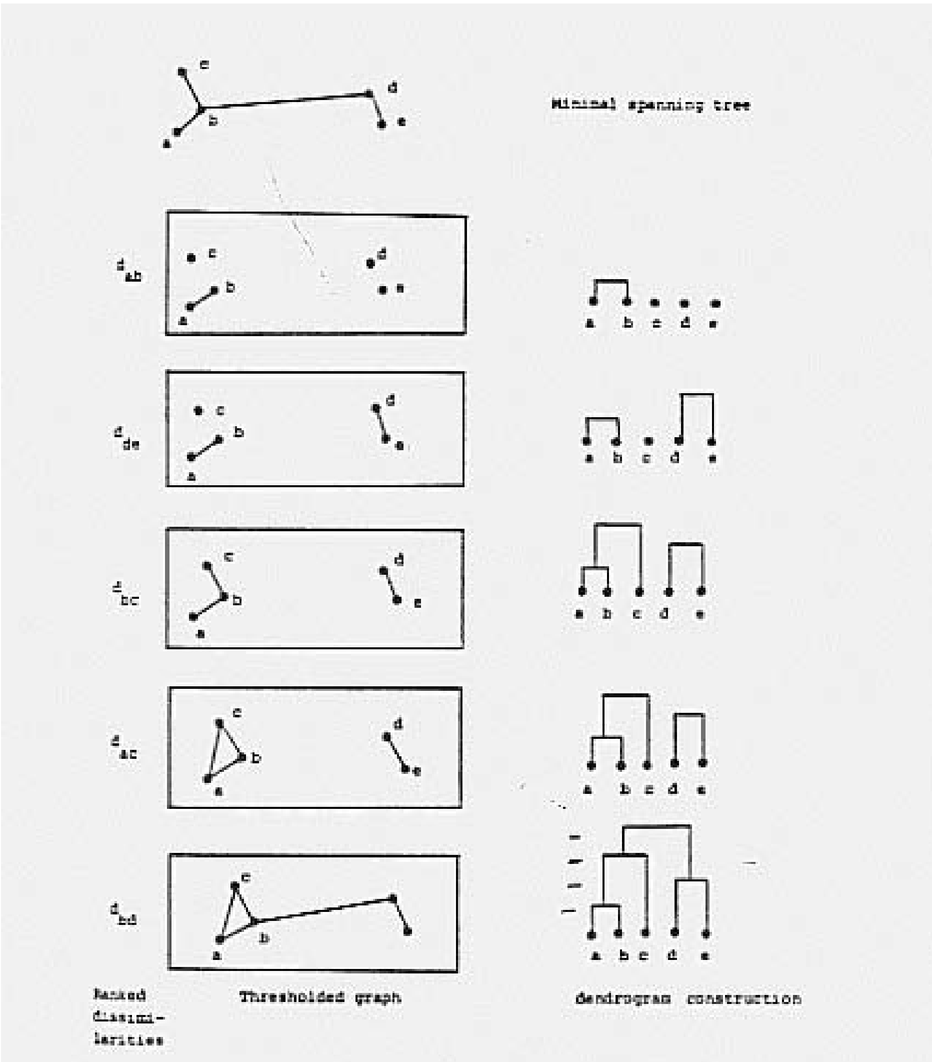
Step 1 Select an arbitrary point and connect it to the least dissimilar neighbour. These two points constitute a subgraph of the MST.

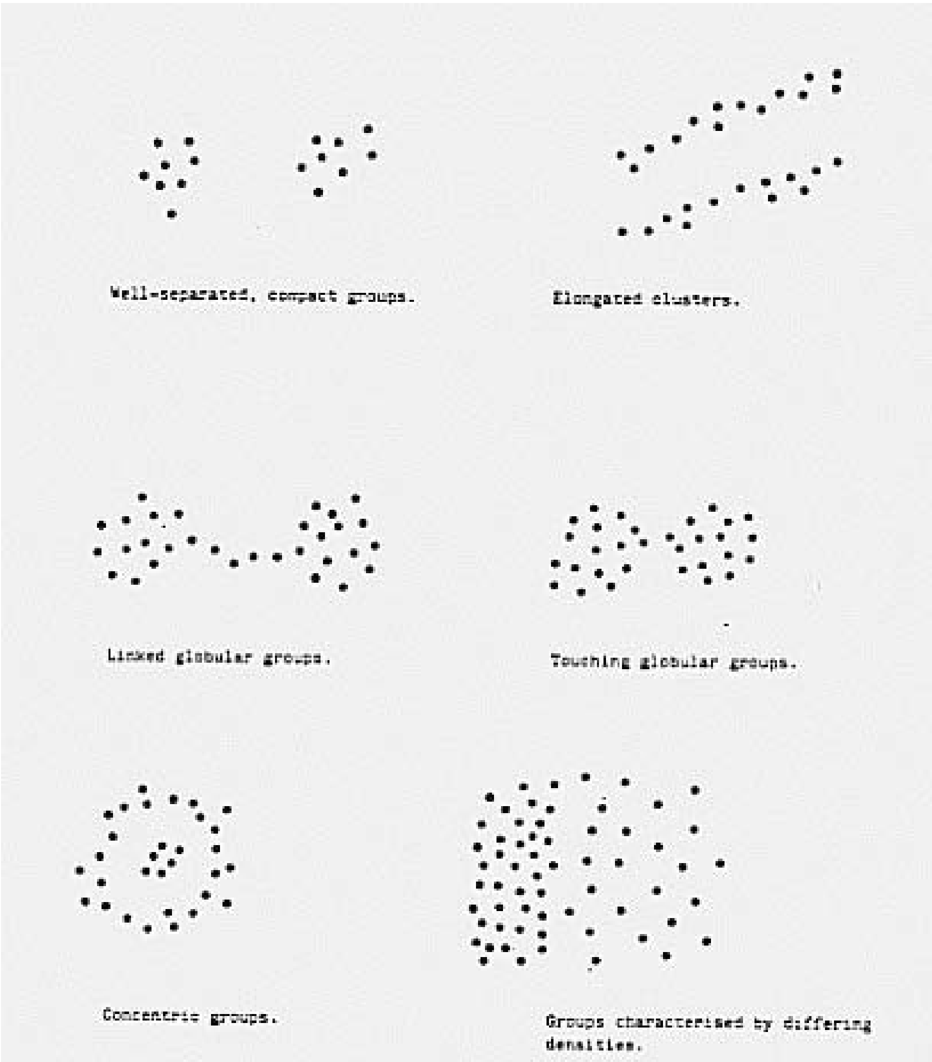
Step 2 Connect the current subgraph to the least dissimilar neighbour of any of the members of the subgraph.

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



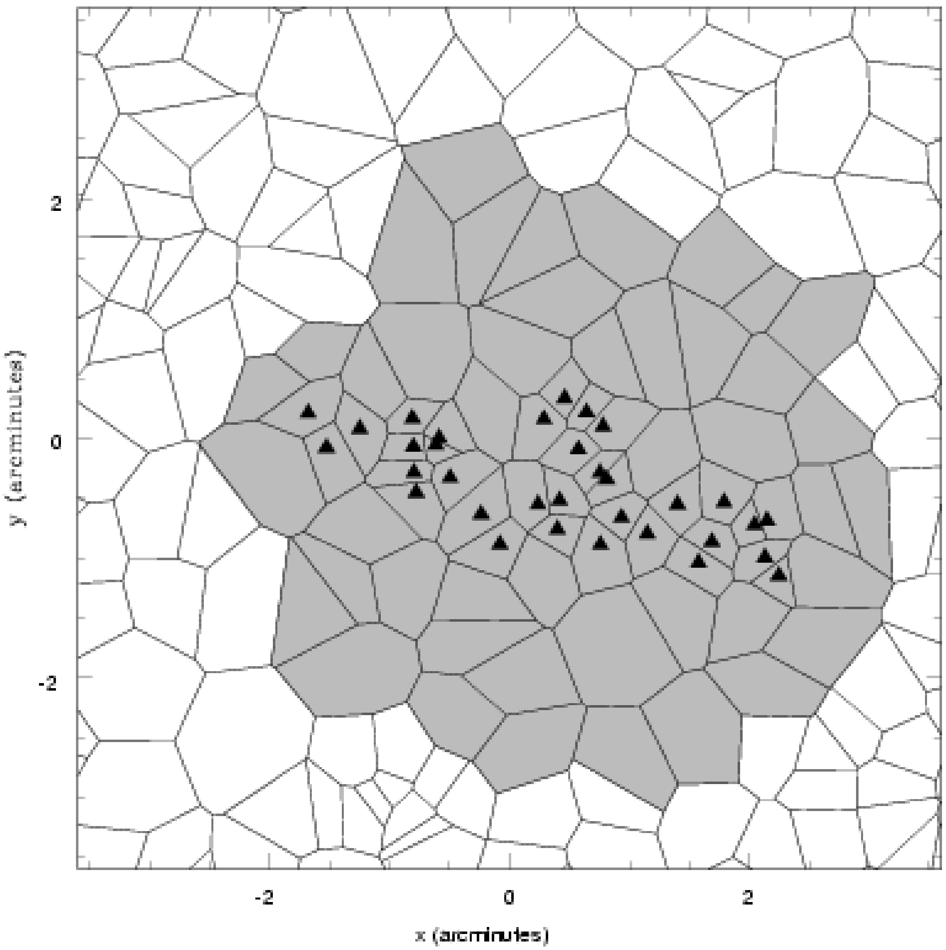

Minimal spanning tree
Ranked dissimi- larities
Thresholded graph
dendrogram construction


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.
- Demo: http://www.csie.ntu.edu.tw/~b5506061/voronoi/Voronoi.html
- 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 = mean; and Maximization = assignment step.
- Diday's *nuées dynamiques*.
- Widely used (since computational cost of hierarchical clustering is usually $O(n^2)$).

Partitioning: Späth's Exchange Algorithm

Exchange method for the minimum variance criterion

Step 1 Arbitrarily choose an initial partition.

Step 2 For each $i \in p$, see if the criterion is bettered by relocating i in another class q. If this is the case, we choose class q such that the criterion V is least; if it is not the case, we proceed to the next i.

Step 3 If the maximum possible number of iterations has not been reached, and if at least one relocation took place in Step 2, return again to Step 2.

Exchange Algorithm – Properties

- Clusters will not become empty.
- The change in variance brought about by relocating object i from class p to class q can be shown to be $\frac{|p|}{|p|-1} ||\mathbf{i} - \mathbf{p}||^2 - \frac{|q|}{|q|-1} ||\mathbf{i} - \mathbf{q}||^2$

Mixture Modelling

- Data is a mixture of G multivariate Gaussians:

$$f_k(x;\theta) \sim \text{MVN}(\mu_k, \Sigma_k) \qquad k = 1, \ldots, G$$

$$f(x;\theta) = \sum_{k=1}^{G} \pi_k f_k(x;\theta)$$

$$\text{Mixing or prior probabilities, } \sum_{k=1}^{G} \pi_k = 1$$

- Estimate parameters θ, π by maximizing the mixture likelihood:

$$L(\theta,\gamma) = \Pi_{i=1}^{n} f(x_i;\theta)$$

where x_i is the ith observation, and γ is a cluster assignment function.

Mixture Modelling – 2

- Implementation: hierarchical agglomerative; iterative relocation; EM; start with agglomerative and refine with EM.
- Choosing the number of clusters – the Bayes Information Criterion (BIC).

  Bayes factor, $B = p(x \mid M_2)/p(x \mid M_1)$

  $p(x \mid M_2)$ = integrated likelihood of the mixture model 2 obtained by integrating over parameter space.
- Approximate the Bayes factor by the BIC:

  Let $p(x \mid G)$ be the integrated likelihood of the data given that there are G clusters.

  Then:

  $2 \log p(x \mid G) \approx 2l(x; \hat{\theta}, G) - m_G \log n = BIC$

  $l(x; \hat{\theta}, G)$ is the maximized mixture log-likelihood with G clusters.

m_G is the number of independent parameters to be estimated in the G-cluster model.

The larger the value of BIC, the better the model.

Example: Gamma-Ray Bursts

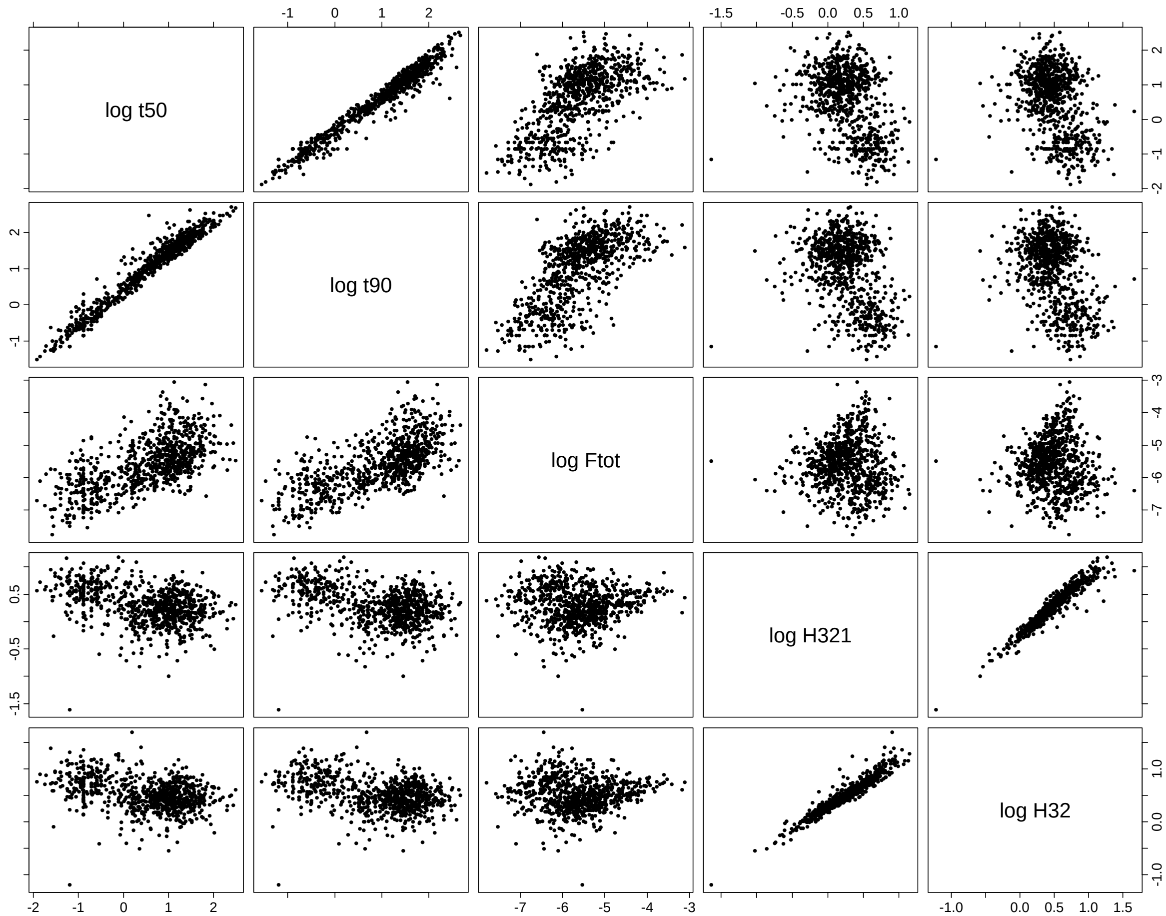
- Few gamma-ray burst (GRB) sources have astronomical counterparts at other wavebands. Hence empirical studies of GRBs have been largely restricted to the analysis of their gamma ray properties.
- Bulk properties such as fluence and spectral hardness are used.
- Studies fall into two categories: examination whether GRB bulk properties comprise a homogeneous population or are divided into distinct classes; and search for relationships between bulk properties.
- Generally accepted taxonomy of GRBs is division between short-hard and long-soft bursts.
- We use GRBs from the Third BATSE Catalog, from the Compton Gamma Ray Observatory. Data from 1996.
- There are roughly eleven variables of potential astrophysical interest: two

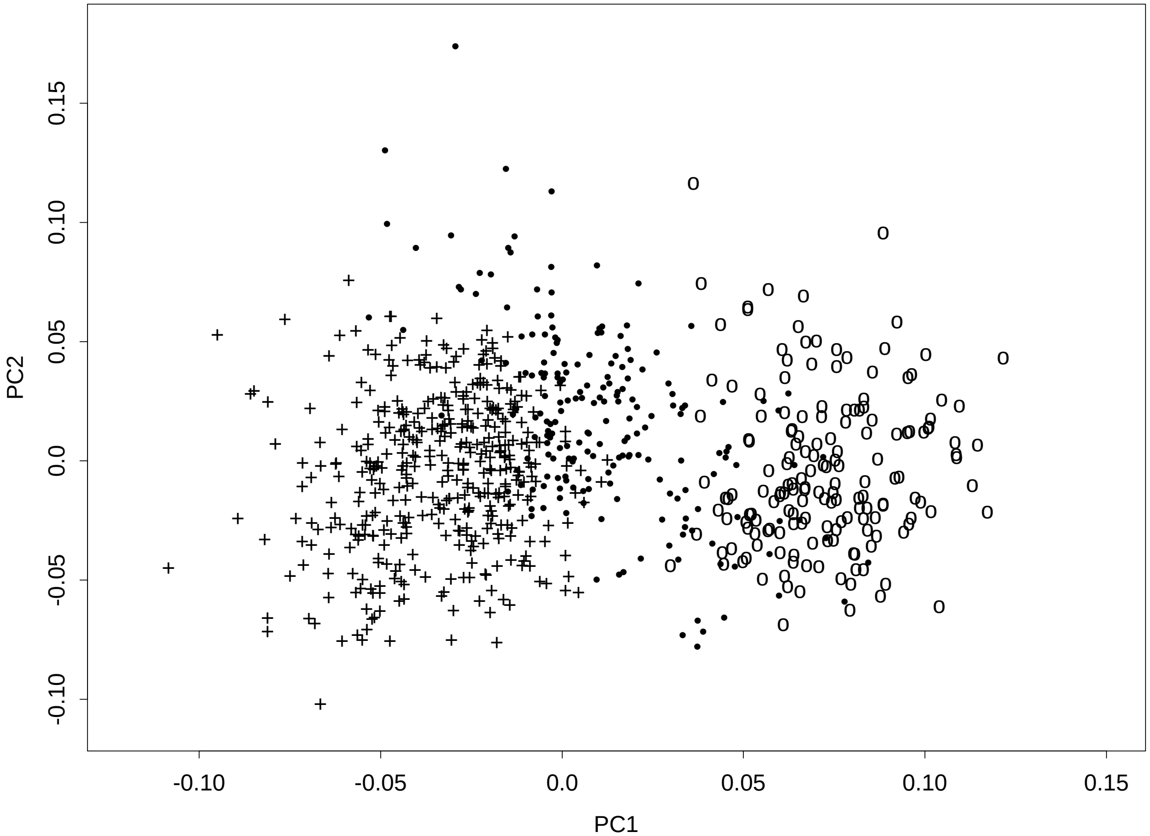
measures of location in Galactic coordinates, l and b; two measures of burst durations, the times within which 50% (T_{50}) and 90% (T_{90}) of the flux arrives; three peak fluxes P_{64}, P_{256} and P_{1024} measured in 64 ms, 256 ms and 1024 ms bins respectively; and four time-integrated fluences $F_1 - F_4$ in the 0-50 keV, 50-100 keV, 100-300 keV and $>$ 300 keV spectral channels respectively

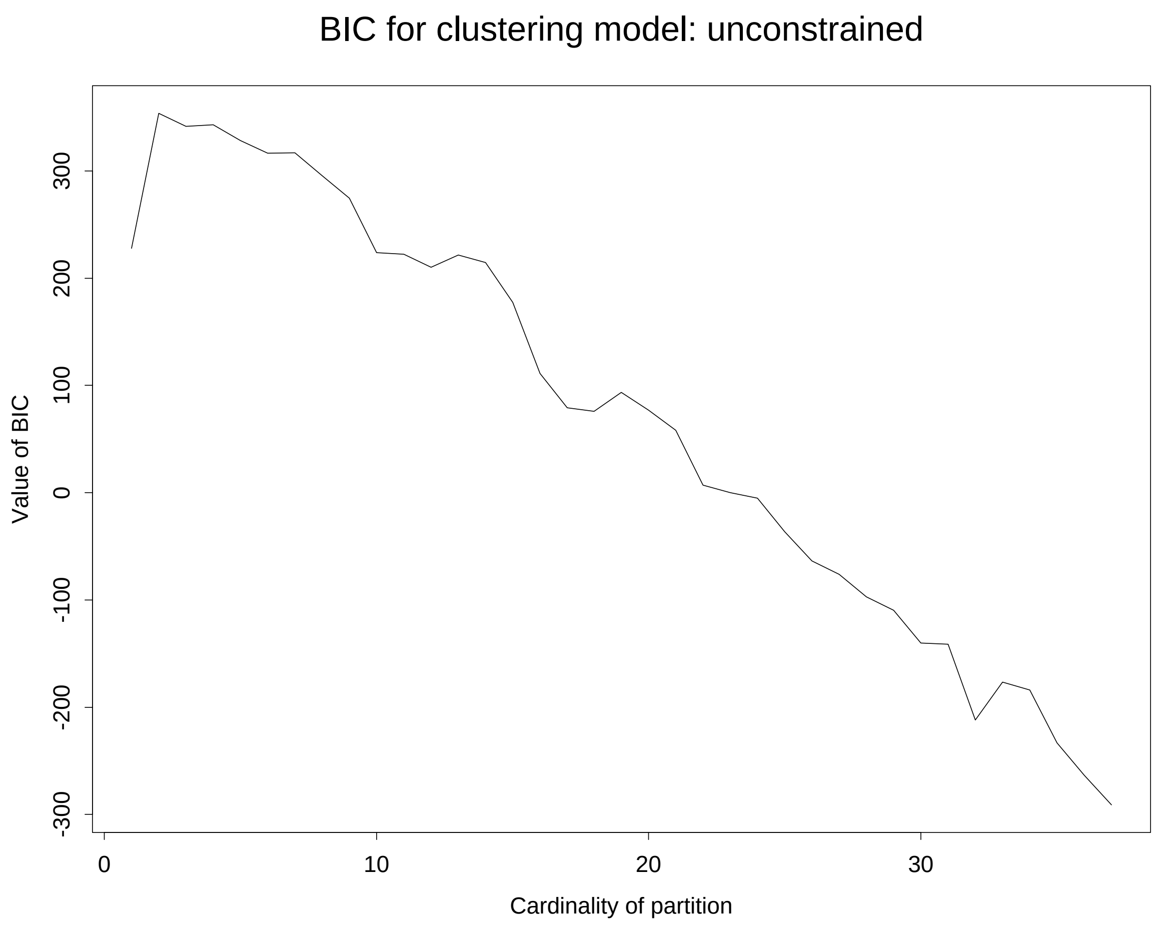
- Consider three composite variables: the total fluence, $F_T = F_1 + F_2 + F_3 + F_4$, and two measures of spectral hardness derived from the ratios of channel fluences, $H_{32} = F_3/F_2$ and $H_{321} = F_3/(F_1 + F_2)$. Of the 1122 listed bursts, 807 have data on all the variables described above.
- Our sample had 797 GRBs. For some analyses, we also used a subset of 644 bursts with 'debiased' durations, T_{90}^d. Here the durations are modified to account for the effect that brighter bursts will have signal above the noise for longer periods than fainter bursts with the same time profiles.
- We use log variables, rather than normalized or standardized variables.
- Our analysis was performed using log T_{50}, log T_{90}, log F_{tot}, log P_{256}, log H_{321} and log H_{32}.

Example: Gamma-Ray Bursts. Plots To Follow.

- Reference: S. Mukherjee, E.D. Feigelson, G.J. Babu, F. Murtagh, C. Fraley and A. Raftery, “Three types of gamma ray bursts”, The Astrophysical Journal, 508, 314-327, 1998.
- Pairwise plots of BATSE data showing strong correlation between variables 1 and 2, and 4 and 5.
- 3-cluster results on unconstrained model clustering (on variables 1, 3 and 4) in principal component space.
- Corresponding BIC values with maximum value corresponding to the 3-cluster solution.


log t50
log t90
log Ftot
log H321
log H32


PC2
PC1
0.15
0.10
0.05
0.0
-0.05
-0.10
-0.10
-0.05
0.0
0.05
0.10
0.15


BIC for clustering model: unconstrained
Value of BIC
300
200
100
0
-100
-200
-300
0
10
20
30
Cardinality of partition

Raftery's Cluster Modelling

- We will parametrize the standard spectral decomposition of Σ_k:

  $\Sigma_k = \lambda_k D_k A_k D_k^T$

  λ_k is largest eigenvalue of Σ:

  controls volume of cluster.

  D_k is matrix of eigenvectors:

  controls orientation of cluster.

  A_k is diag$\{1, \alpha_{2k} \ldots \alpha_{pk}\}$:

  controls shape of cluster.

- Example 1: set shape, different sizes and orientations:

  For p = 2 dimensional data,

  $A_k = \text{diag}\{1, \alpha\}, \alpha = \lambda_2/\lambda_1$

  $\alpha < 1 \Longrightarrow$ long and narrow cluster.

  Use: finding aligned sets of points.

Raftery's Cluster Modelling – 2

- Example 2: hyperspherical clusters, different sizes: $\Sigma_k = \lambda_k I$ (I = identity matrix).
- Example 3: hyperspherical, same size (Ward's method): $\Sigma_k = \lambda I$.
- Example 4: unconstrained Σ_k.

A.J. Scott and M.J. Symons, "Clustering methods based on likelihood ratio criteria", Biometrics, 27, 387–397, 1971.

W_k = SSCP matrix for cluster k,

x_k = mean of cluster k,

n_k = cardinality of cluster k,

$W_k = \sum_{i \in \text{ cluster}} (x_i - x_k)(x_i - x_k)^T$

W_k / n_k = MLE of Σ_k.

Maximize $\sum_{k=1}^{G} n_k \log \left| \frac{W_k}{n_k} \right|$ (| . | = det).

Kohonen Self-Organizing Feature Map

- Regular grid output representational or display space.
- Determine vectors w_k, such that inputs x_i are parsimoniously summarized (clustering objective); and in addition the vectors w_k are positioned in representational space so that similar vectors are close (low-dimensional projection objective) in *representation space*.
- **Clustering:** Associate each x_i with some one w_k such that

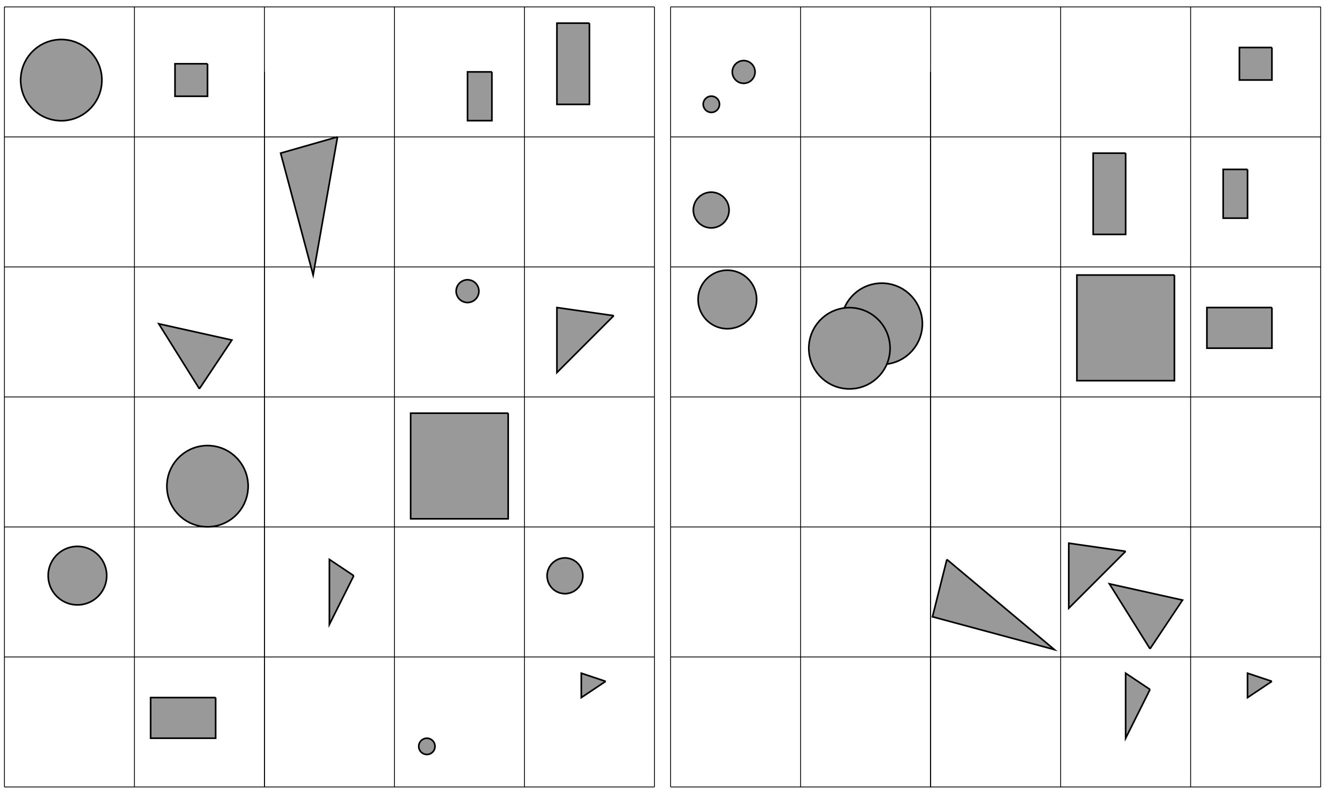
  $$k = argmin \parallel x_i - w_k \parallel$$

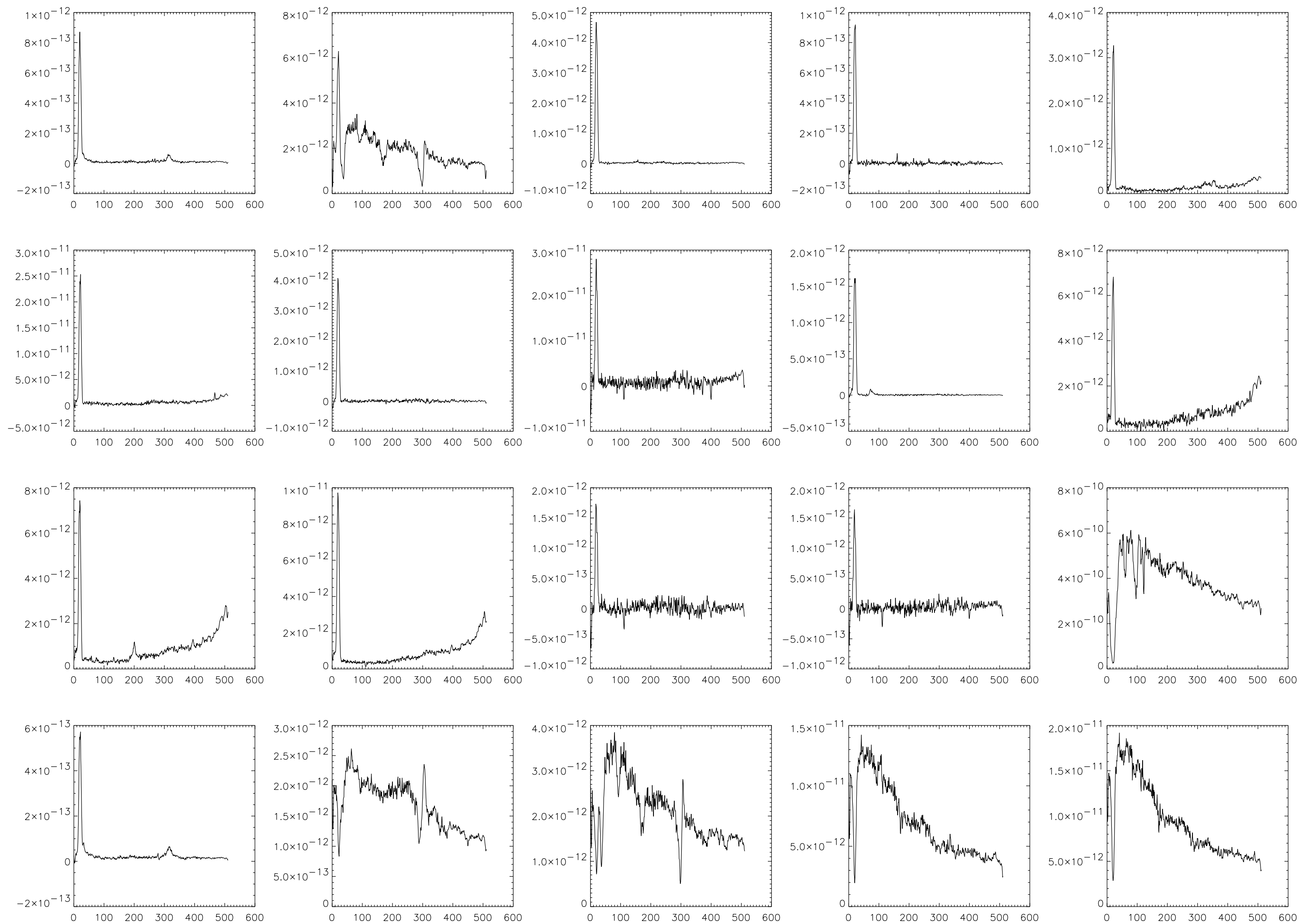
  Low-Dimensional projection:

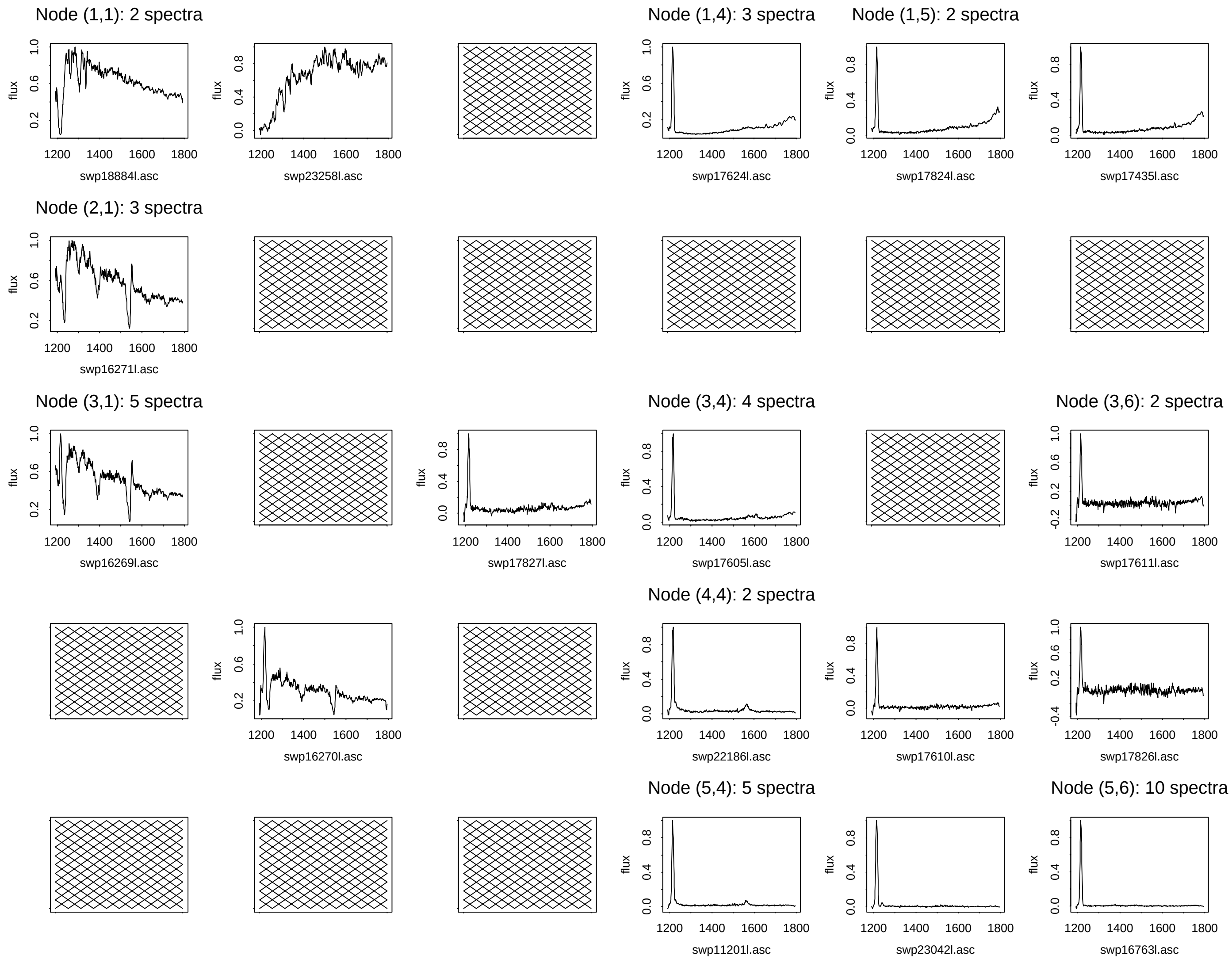
  $$\parallel w_k - w'_k \parallel < \parallel w_k - w''_k \parallel \Longrightarrow \parallel k - k' \parallel \leq \parallel k - k'' \parallel$$
- Initial random choice of values for w_k.
- Updated the set of w_k ($\forall k$) on the basis of presentation of input vectors, x_i.
- Processing one x_i is termed an iteration. Going through all x_i once is termed an

epoch.

- Update not just the so-called winner w_k, but also neighbors of w_k with respect to the representational space.
- The neighborhood is initially chosen to be quite large (e.g. a 4×4 zone) and as the epochs proceed, is reduced to 1×1 (i.e. no neighborhood).
- Example: set of 45 spectra of the complex AGN (active galactic nucleus) object, NGC 4151, taken with the IUE (International Ultraviolet Explorer) satellite.
- 45 spectra observed with the SWP spectral camera, with wavelengths from 1191.2 Å to approximately 1794.4 Å, with values at 512 interval steps.
- We will show sample of 20 spectra; and then Kohonen map of these.

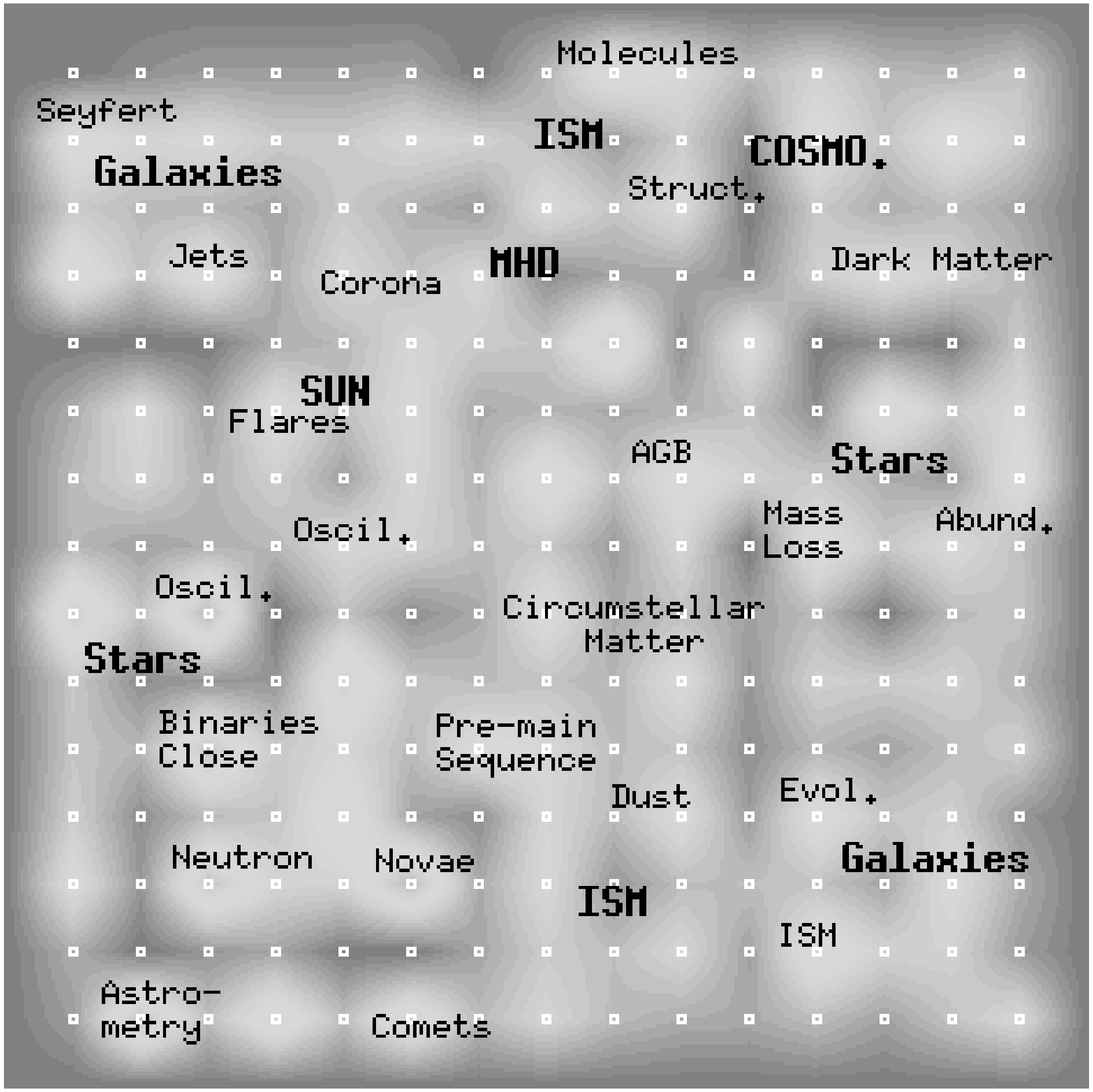


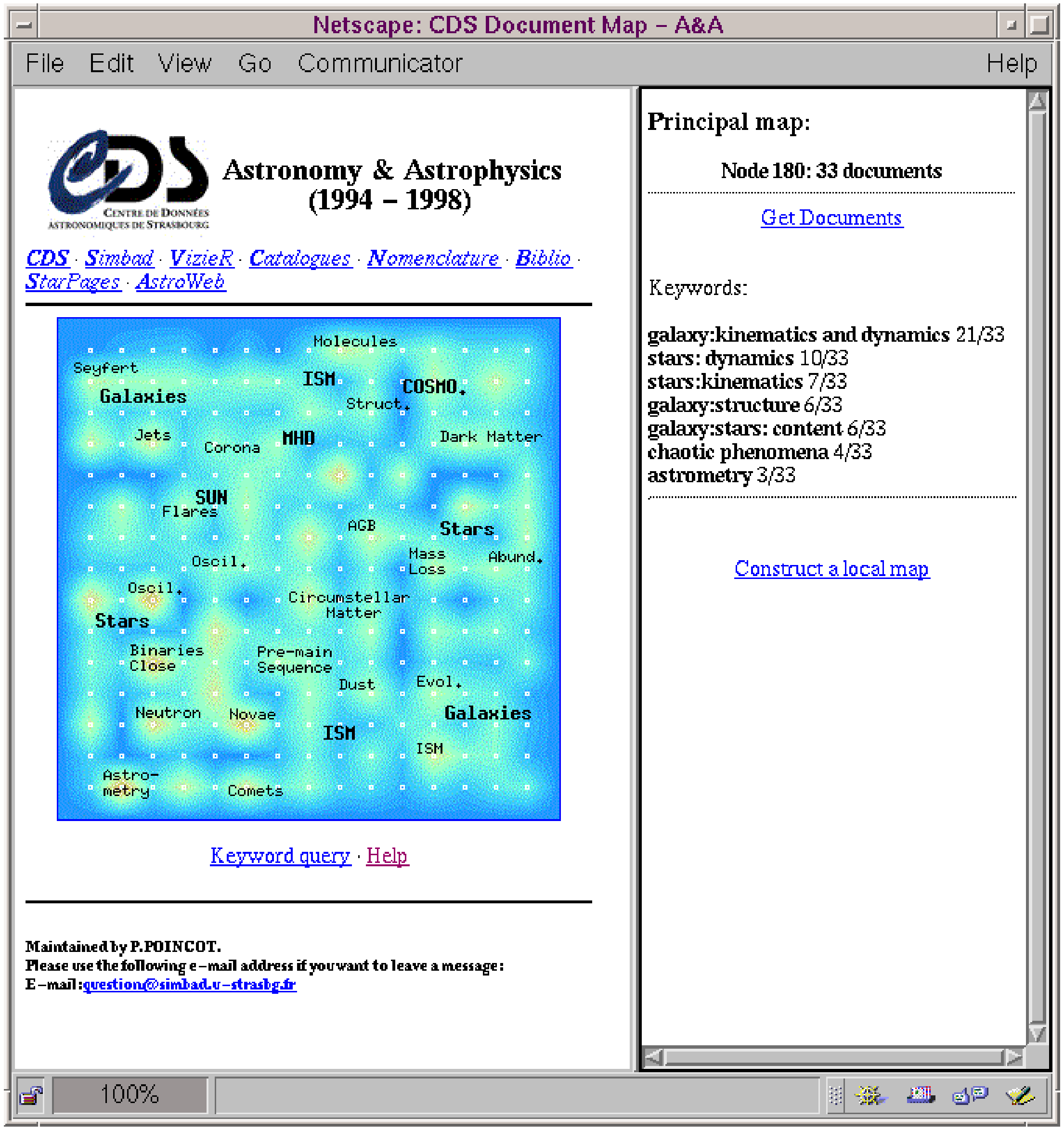


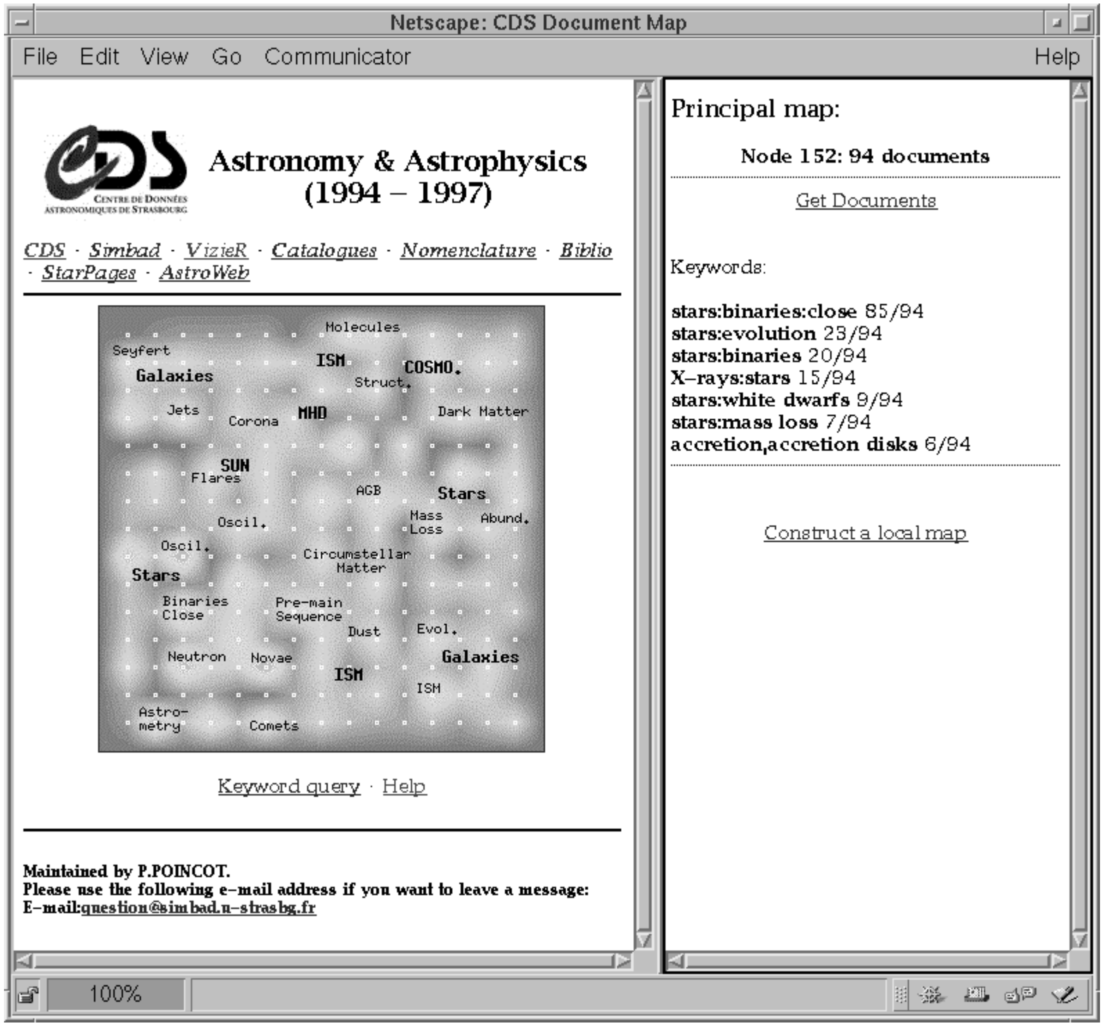

Node (1,1): 2 spectra
Node (1,4): 3 spectra
Node (1,5): 2 spectra
swp18884l.asc
swp23258l.asc
swp17624l.asc
swp17824l.asc
swp17435l.asc
Node (2,1): 3 spectra
swp16271l.asc
Node (3,1): 5 spectra
Node (3,4): 4 spectra
Node (3,6): 2 spectra
swp16269l.asc
swp17827l.asc
swp17605l.asc
swp17611l.asc
Node (4,4): 2 spectra
swp16270l.asc
swp22186l.asc
swp17610l.asc
swp17826l.asc
Node (5,4): 5 spectra
Node (5,6): 10 spectra
swp11201l.asc
swp23042l.asc
swp16763l.asc
flux
1200 1400 1600 1800

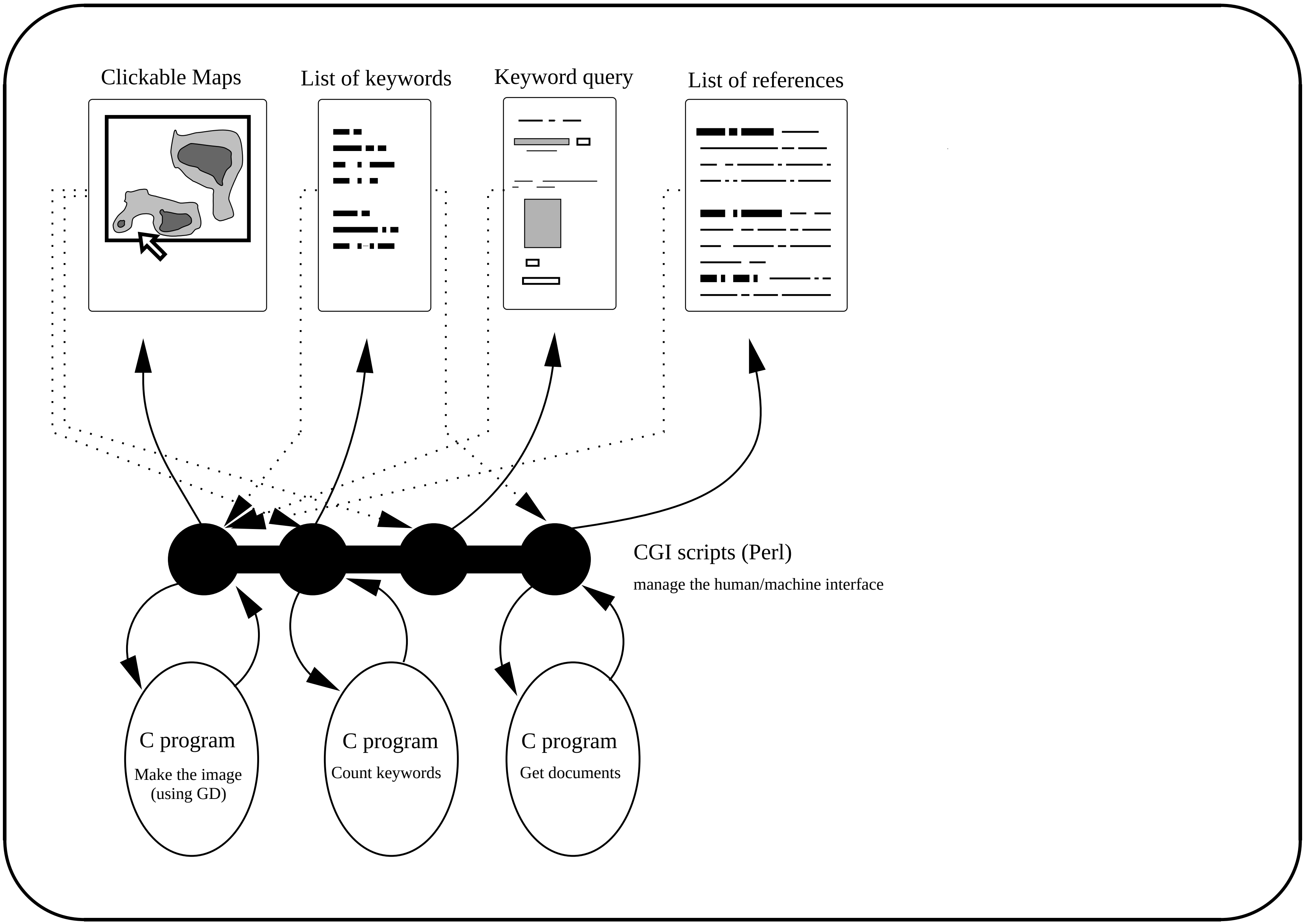
Kohonen Map: Interactive User Interface

- About 10,000 documents described by 269 keywords from articles published in A&A; also in ApJ.
- 15×15 grid was used for the principal map, and a 5×5 grid for detailed maps.
- User clicks on thematic area, or enters keywords.
- A detailed map is produced. Any document listed allows access to the full document through ADS.
- This system is server-side, based on imagemap and CGI scripts.


Molecules
Seyfert
ISM
COSMO.
Galaxies
Struct.
Jets
Corona
MHD
Dark Matter
SUN
Flares
AGB
Stars
Mass
Loss
Abund.
Oscil.
Oscil.
Circumstellar
Matter
Stars
Binaries
Close
Pre-main
Sequence
Dust
Evol.
Neutron
Novae
Galaxies
ISM
ISM
Astro-
metry
Comets


Netscape: CDS Document Map – A&A
File Edit View Go Communicator Help
Astronomy & Astrophysics (1994 – 1998)
CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG
CDS · Simbad · VizieR · Catalogues · Nomenclature · Biblio · StarPages · AstroWeb
Molecules
Seyfert
Galaxies
ISM
COSMO.
Struct.
Jets
Corona
MHD
Dark Matter
SUN
Flares
AGB
Stars
Mass Loss
Abund.
Oscil.
Oscil.
Circumstellar Matter
Stars
Binaries Close
Pre-main Sequence
Dust
Evol.
Neutron
Novae
Galaxies
ISM
ISM
Astro-metry
Comets
Keyword query · Help
Maintained by P.POINCOT.
Please use the following e-mail address if you want to leave a message:
E-mail: question@simbad.u-strasbg.fr
Principal map:
Node 180: 33 documents
Get Documents
Keywords:
galaxy:kinematics and dynamics 21/33
stars: dynamics 10/33
stars:kinematics 7/33
galaxy:structure 6/33
galaxy:stars: content 6/33
chaotic phenomena 4/33
astrometry 3/33
Construct a local map
100%


Netscape: CDS Document Map
File Edit View Go Communicator Help
Astronomy & Astrophysics
(1994 – 1997)
CDS · Simbad · VizieR · Catalogues · Nomenclature · Biblio · StarPages · AstroWeb
Molecules
Seyfert
Galaxies
ISM
COSMO.
Struct.
Jets
Corona
MHD
Dark Matter
SUN
Flares
AGB
Stars
Mass Loss
Abund.
Oscil.
Oscil.
Circumstellar Matter
Stars
Binaries Close
Pre-main Sequence
Dust
Evol.
Neutron
Novae
ISM
Galaxies
ISM
Astro-metry
Comets
Keyword query · Help
Maintained by P.POINCOT.
Please use the following e-mail address if you want to leave a message:
E-mail:question@simbad.u-strasbg.fr
Principal map:
Node 152: 94 documents
Get Documents
Keywords:
stars:binaries:close 85/94
stars:evolution 23/94
stars:binaries 20/94
X-rays:stars 15/94
stars:white dwarfs 9/94
stars:mass loss 7/94
accretion,accretion disks 6/94
Construct a local map
100%



Clickable Maps
List of keywords
Keyword query
List of references
CGI scripts (Perl)
manage the human/machine interface
C program
Make the image
(using GD)
C program
Count keywords
C program
Get documents

Some References

- J.D. Barrow, S.P. Bhavsar and D.H. Sonoda, "Minimal spanning trees filaments and galaxy clustering", *Monthly Notices of the Royal Astronomical Society*, **216**, 17–35, 1985.
- C.R. Cowley and R. Henry, "Numerical taxonomy of Ap and Am stars", *The Astrophysical Journal*, **233**, 633–643, 1979.
- J.K. Davies, N. Eaton, S.F. Green, R.S. McCheyne and A.J. Meadows, "The classification of asteroids", *Vistas in Astronomy*, **26**, 243–251, 1982.
- J.P. Huchra and M.J. Geller, "Groups of galaxies. I. Nearby groups", *The Astrophysical Journal*, **257**, 423–437, 1982
- J.F. Jarvis and J.A. Tyson, "FOCAS: faint object classification and analysis system", *The Astronomical Journal*, **86**, 476–495, 1981.
- M.O. Mennessier, "A classification of miras from their visual and near–infrared

light curves: an attempt to correlate them with their evolution", *Astronomy and Astrophysics*, **144**, 463–470, 1985.

- D.J. Tholen, "Asteroid taxonomy from cluster analysis of photometry", PhD Thesis, University of Arizona, 1984.
- P. Poinçot, S. Lesteven and F. Murtagh, "A spatial user interface to the astronomical literature", *Astronomy and Astrophysics Supplement Series*, **130**, 183–191, 1998.
- P. Poinçot, S. Lesteven and F. Murtagh, "Maps of information spaces: assessments from astronomy", *Journal of the American Society for Information Science*, **51**, 1081-1089, 2000.
- D. Egret, R.J. Hanisch and F. Murtagh, "Search and discovery tools for astronomical on-line resources and services", *Astronomy and Astrophysics Supplement*, **143**, 137-143, 2000