clustermat — Introduction to clustermat commands

Description Syntax Remarks and examples References Also see

Description

`clustermat` performs hierarchical cluster analysis on the dissimilarity matrix *matname*. `clustermat` is part of the `cluster` suite of commands; see [MV] **cluster**. All Stata hierarchical clustering methods are allowed with `clustermat`. The partition-clustering methods (`kmeans` and `kmedians`) are not allowed because they require the data.

See [MV] **cluster** for a listing of all the `cluster` and `clustermat` commands. The `cluster dendrogram` command (see [MV] **cluster dendrogram**) will display the resulting dendrogram, the `clustermat stop` command (see [MV] **cluster stop**) will help in determining the number of groups, and the `cluster generate` command (see [MV] **cluster generate**) will produce grouping variables. Other useful `cluster` subcommands include `notes`, `dir`, `list`, `drop`, `use`, `rename`, and `renamevar`; see [MV] **cluster notes** and [MV] **cluster utility**.

Syntax

`clustermat` *linkage matname* . . .

linkage	Description
`singlelinkage`	single-linkage cluster analysis
`averagelinkage`	average-linkage cluster analysis
`completelinkage`	complete-linkage cluster analysis
`waveragelinkage`	weighted-average linkage cluster analysis
`medianlinkage`	median-linkage cluster analysis
`centroidlinkage`	centroid-linkage cluster analysis
`wardslinkage`	Ward's linkage cluster analysis

See [MV] **cluster linkage**.

`clustermat stop` has similar syntax to that of `cluster stop`; see [MV] **cluster stop**. For the remaining postclustering subcommands and user utilities, you may specify either `cluster` or `clustermat`—it does not matter which.

Remarks and examples

If you are clustering observations by using one of the similarity or dissimilarity measures provided by Stata, the `cluster` command is what you need. If, however, you already have a dissimilarity matrix or can produce one for a dissimilarity measure that Stata does not provide, or if you want to cluster variables instead of observations, the `clustermat` command is what you need.

▷ Example 1

Table 6 of Kaufman and Rousseeuw (1990) provides a subjective dissimilarity matrix among 11 sciences. Fourteen postgraduate economics students from different parts of the world gave subjective dissimilarities among these 11 sciences on a scale from 0 (identical) to 10 (very different). The final dissimilarity matrix was obtained by averaging the results from the students.

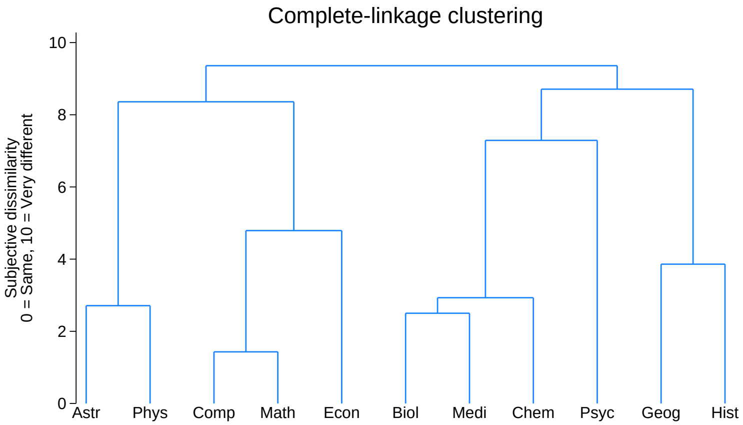
We begin by creating a label variable and a shorter version of the label variable corresponding to the 11 sciences. Then we create a row vector containing the lower triangle of the dissimilarity matrix.

```
. input str13 science

           science
  1. Astronomy
  2. Biology
  3. Chemistry
  4. Computer sci.
  5. Economics
  6. Geography
  7. History
  8. Mathematics
  9. Medicine
 10. Physics
 11. Psychology
 12. end

. generate str4 shortsci = substr(science,1,4)

. matrix input D = (
> 0.00
> 7.86 0.00
> 6.50 2.93 0.00
> 5.00 6.86 6.50 0.00
> 8.00 8.14 8.21 4.79 0.00
> 4.29 7.00 7.64 7.71 5.93 0.00
> 8.07 8.14 8.71 8.57 5.86 3.86 0.00
> 3.64 7.14 4.43 1.43 3.57 7.07 9.07 0.00
> 8.21 2.50 2.93 6.36 8.43 7.86 8.43 6.29 0.00
> 2.71 5.21 4.57 4.21 8.36 7.29 8.64 2.21 5.07 0.00
> 9.36 5.57 7.29 7.21 6.86 8.29 7.64 8.71 3.79 8.64 0.00 )
```

There are several ways that we could have stored the dissimilarity information in a matrix. To avoid entering both the upper and lower triangle of the matrix, we entered the dissimilarities as a row vector containing the lower triangular entries of the dissimilarity matrix, including the diagonal of zeros (although there are options that would allow us to omit the diagonal of zeros). We typed `matrix input D =` . . . instead of `matrix D =` . . . so that we could omit the commas between entries; see [P] **matrix define**.

We now perform a complete-linkage cluster analysis on these dissimilarities. The `name()` option names the cluster analysis. We will name it `complink`. The `shape(lower)` option is what signals that the dissimilarity matrix is stored as a row vector containing the lower triangle of the dissimilarity matrix, including the diagonal of zeros. The `add` option indicates that the resulting cluster information should be added to the existing dataset. Here the existing dataset consists of the `science` label variable and the shortened version `shortsci`. See [MV] **cluster linkage** for details concerning these options. The short labels are passed to `cluster dendrogram` so that we can see which subjects were most closely related when viewing the dendrogram; see [MV] **cluster dendrogram**.

```
. clustermat completelinkage D, shape(lower) add name(complink)

. cluster dendrogram complink, labels(shortsci)
> title(Complete-linkage clustering)
> ytitle("Subjective dissimilarity" "0 = Same, 10 = Very different")
```



From the dendrogram, we see that mathematics and computer science were deemed most similar and that the economists most closely related their field of study to those two disciplines.

◁

▷ Example 2

Stata does not provide the Bray and Curtis (1957) dissimilarity measure first described by Odum (1950). Using the same notation as that found in [MV] ***measure_option***, we find that the Bray–Curtis dissimilarity between observations i and j is

$$\frac{\sum_{a=1}^{p} |x_{ia} - x_{ja}|}{\sum_{a=1}^{p} (x_{ia} + x_{ja})}$$

Stata does not provide this measure because of the many cases where the measure is undefined (because of dividing by zero). However, when the data are positive the Bray–Curtis dissimilarity is well behaved.

Even though Stata does not automatically provide this measure, it is easy to obtain it and then use it with `clustermat` to perform hierarchical clustering. The numerator of the Bray–Curtis dissimilarity measure is the `L1` (absolute value) distance. We use the `matrix dissimilarity` command (see [MV] **matrix dissimilarity**) to obtain the `L1` dissimilarity matrix and then divide the elements of that matrix by the appropriate values to obtain the Bray–Curtis dissimilarity.

Fisher (1936) presented data, originally from Anderson (1935), on three species of iris. Measurements of the length and width of the sepal and petal were obtained for 50 samples of each of the three iris species. We obtained the data from Morrison (2005). Here we demonstrate average-linkage clustering of these 150 observations.

```
. use https://www.stata-press.com/data/r19/iris, clear
(Iris data)

. summarize seplen sepwid petlen petwid

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      seplen |        150    5.843333    .8280661        4.3        7.9
      sepwid |        150    3.057333    .4358663          2        4.4
      petlen |        150       3.758    1.765298          1        6.9
      petwid |        150    1.199333    .7622377         .1        2.5

. matrix dissimilarity irisD = seplen sepwid petlen petwid, L1

. egen rtot = rowtotal(seplen sepwid petlen petwid)

. forvalues a = 1/150 {
  2.          forvalues b = 1/150 {
  3.                  mat irisD[`a',`b'] = irisD[`a',`b']/(rtot[`a']+rtot[`b'])
  4.          }
  5. }

. matlist irisD[1..5,1..5]

             |      obs1       obs2       obs3       obs4       obs5
-------------+-------------------------------------------------------
        obs1 |         0
        obs2 |   .035533          0
        obs3 |  .0408163    .026455          0
        obs4 |  .0510204    .026455   .0212766          0
        obs5 |  .0098039    .035533   .0408163   .0510204          0
```

The egen `rowtotal()` function provided the row totals used in the denominator of the Bray–Curtis dissimilarity measure; see [D] **egen**. We listed the dissimilarities between the first 5 observations.

We now compute the average-linkage cluster analysis on these 150 observations (see [MV] **cluster linkage**) and examine the Caliński–Harabasz pseudo-*F* index and the Duda–Hart Je(2)/Je(1) index (cluster stopping rules; see [MV] **cluster stop**) to try to determine the number of clusters.

```
. clustermat averagelink irisD, name(iris) add

. clustermat stop, variables(seplen sepwid petlen petwid)

+---------------------------+
|             |  Calinski/  |
|  Number of  |  Harabasz   |
|  clusters   |  pseudo-F   |
|-------------+-------------|
|      2      |    502.82   |
|      3      |    299.96   |
|      4      |    201.58   |
|      5      |    332.89   |
|      6      |    288.61   |
|      7      |    244.61   |
|      8      |    252.39   |
|      9      |    223.28   |
|     10      |    268.47   |
|     11      |    241.51   |
|     12      |    232.61   |
|     13      |    233.46   |
|     14      |    255.84   |
|     15      |    273.96   |
+---------------------------+
. clustermat stop, variables(seplen sepwid petlen petwid) rule(duda)

+---------------------------------------+
|             |        Duda/Hart        |
|  Number of  |             |  pseudo   |
|  clusters   | Je(2)/Je(1) | T-squared |
|-------------+-------------+-----------|
|      1      |   0.2274    |   502.82  |
|      2      |   0.8509    |    17.18  |
|      3      |   0.8951    |     5.63  |
|      4      |   0.4472    |   116.22  |
|      5      |   0.6248    |    28.23  |
|      6      |   0.9579    |     2.55  |
|      7      |   0.5438    |    28.52  |
|      8      |   0.8843    |     5.10  |
|      9      |   0.5854    |    40.37  |
|     10      |   0.0000    |        .  |
|     11      |   0.8434    |     6.68  |
|     12      |   0.4981    |    37.28  |
|     13      |   0.5526    |    25.91  |
|     14      |   0.6342    |    16.15  |
|     15      |   0.6503    |     3.23  |
+---------------------------------------+
```

The stopping rules are not conclusive here. From the Duda–Hart pseudo-T^2 (small values) you might best conclude that there are three, six, or eight natural clusters. The Caliński and Harabasz pseudo-F (large values) indicates that there might be two, three, or five groups.

With the iris data, we know the three species. Let's compare the average-linkage hierarchical cluster solutions with the actual species. The `cluster generate` command (see [MV] **cluster generate**) will generate grouping variables for our hierarchical cluster analysis.

```
. cluster generate g = groups(2/6)
. tabulate g2 iris

           |           Iris species
        g2 |    Setosa  Versicolo  Virginica |     Total
-----------+---------------------------------+----------
         1 |        50          0          0 |        50
         2 |         0         50         50 |       100
-----------+---------------------------------+----------
     Total |        50         50         50 |       150
. tabulate g3 iris

           |           Iris species
        g3 |    Setosa  Versicolo  Virginica |     Total
-----------+---------------------------------+----------
         1 |        50          0          0 |        50
         2 |         0         46         50 |        96
         3 |         0          4          0 |         4
-----------+---------------------------------+----------
     Total |        50         50         50 |       150
```

```
. tabulate g4 iris

           |          Iris species
        g4 |    Setosa  Versicolo  Virginica |     Total
-----------+---------------------------------+----------
         1 |        49          0          0 |        49 
         2 |         1          0          0 |         1 
         3 |         0         46         50 |        96 
         4 |         0          4          0 |         4 
-----------+---------------------------------+----------
     Total |        50         50         50 |       150 
. tabulate g5 iris

           |          Iris species
        g5 |    Setosa  Versicolo  Virginica |     Total
-----------+---------------------------------+----------
         1 |        49          0          0 |        49 
         2 |         1          0          0 |         1 
         3 |         0         45         15 |        60 
         4 |         0          1         35 |        36 
         5 |         0          4          0 |         4 
-----------+---------------------------------+----------
     Total |        50         50         50 |       150 
. tabulate g6 iris

           |          Iris species
        g6 |    Setosa  Versicolo  Virginica |     Total
-----------+---------------------------------+----------
         1 |        41          0          0 |        41 
         2 |         8          0          0 |         8 
         3 |         1          0          0 |         1 
         4 |         0         45         15 |        60 
         5 |         0          1         35 |        36 
         6 |         0          4          0 |         4 
-----------+---------------------------------+----------
     Total |        50         50         50 |       150 
```

The two-group cluster solution splits *Iris setosa* from *Iris versicolor* and *Iris virginica*. The three- and four-group cluster solutions appear to split off some outlying observations from the two main groups. The five-group solution finally splits most of *Iris virginica* from the *Iris versicolor* but leaves some overlap.

Though this is not shown here, cluster solutions that better match the known species can be found by using dissimilarity measures other than Bray–Curtis.

◁

▷ Example 3

The `cluster` command clusters observations. If you want to cluster variables, you have two choices. You can use `xpose` (see [D] **xpose**) to transpose the variables and observations, or you can use `matrix dissimilarity` with the `variables` option (see [MV] **matrix dissimilarity**) and then use `clustermat`.

In example 2 of [MV] **cluster kmeans and kmedians**, we introduce the women's club data. Thirty women were asked 35 yes–no questions. In [MV] **cluster kmeans and kmedians**, our interest was in clustering the 30 women for placement at luncheon tables. Here our interest is in understanding the relationship among the 35 variables. Which questions produced similar response patterns from the 30 women?

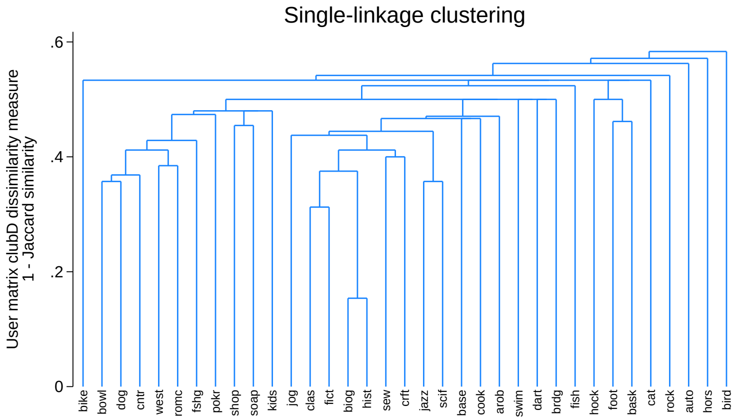
```
. use https://www.stata-press.com/data/r19/wclub, clear

. describe

Contains data from https://www.stata-press.com/data/r19/wclub.dta
 Observations:            30                  
    Variables:            35                  1 May 2024 16:56
--------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------
bike            byte    %8.0g                 Enjoy bicycle riding Y/N
bowl            byte    %8.0g                 Enjoy bowling Y/N
swim            byte    %8.0g                 Enjoy swimming Y/N
jog             byte    %8.0g                 Enjoy jogging Y/N
hock            byte    %8.0g                 Enjoy watching hockey Y/N
foot            byte    %8.0g                 Enjoy watching football Y/N
base            byte    %8.0g                 Enjoy baseball Y/N
bask            byte    %8.0g                 Enjoy basketball Y/N
arob            byte    %8.0g                 Participate in aerobics Y/N
fshg            byte    %8.0g                 Enjoy fishing Y/N
dart            byte    %8.0g                 Enjoy playing darts Y/N
clas            byte    %8.0g                 Enjoy classical music Y/N
cntr            byte    %8.0g                 Enjoy country music Y/N
jazz            byte    %8.0g                 Enjoy jazz music Y/N
rock            byte    %8.0g                 Enjoy rock and roll music Y/N
west            byte    %8.0g                 Enjoy reading western novels Y/N
romc            byte    %8.0g                 Enjoy reading romance novels Y/N
scif            byte    %8.0g                 Enjoy reading sci. fiction Y/N
biog            byte    %8.0g                 Enjoy reading biographies Y/N
fict            byte    %8.0g                 Enjoy reading fiction Y/N
hist            byte    %8.0g                 Enjoy reading history Y/N
cook            byte    %8.0g                 Enjoy cooking Y/N
shop            byte    %8.0g                 Enjoy shopping Y/N
soap            byte    %8.0g                 Enjoy watching soap operas Y/N
sew             byte    %8.0g                 Enjoy sewing Y/N
crft            byte    %8.0g                 Enjoy craft activities Y/N
auto            byte    %8.0g                 Enjoy automobile mechanics Y/N
pokr            byte    %8.0g                 Enjoy playing poker Y/N
brdg            byte    %8.0g                 Enjoy playing bridge Y/N
kids            byte    %8.0g                 Have children Y/N
hors            byte    %8.0g                 Have a horse Y/N
cat             byte    %8.0g                 Have a cat Y/N
dog             byte    %8.0g                 Have a dog Y/N
bird            byte    %8.0g                 Have a bird Y/N
fish            byte    %8.0g                 Have a fish Y/N
--------------------------------------------------------------------------------
Sorted by:
```

The `matrix dissimilarity` command allows us to compute the Jaccard similarity measure (the `Jaccard` option), comparing variables (the `variables` option) instead of observations, saving one minus the Jaccard measure (the `dissim(oneminus)` option) as a dissimilarity matrix.

```
. matrix dissimilarity clubD = , variables Jaccard dissim(oneminus)

. matlist clubD[1..5,1..5]

             |      bike       bowl       swim        jog       hock
-------------+-------------------------------------------------------
        bike |         0
        bowl |  .7333333          0
        swim |     .5625       .625          0
         jog |        .6   .8235294   .5882353          0
        hock |  .8461538         .6         .8   .8571429          0
```

We pass the clubD matrix to clustermat and ask for a single-linkage cluster analysis. We need to specify the clear option to replace the 30 observations currently in memory with the 35 observations containing the cluster results. Using the labelvar() option, we also ask for a label variable, question, to be created from the clubD matrix row names. To see the resulting cluster analysis, we call cluster dendrogram; see [MV] **cluster dendrogram**.

```
. clustermat singlelink clubD, name(club) clear labelvar(question)
Number of observations (_N) was 0, now 35.

. describe

Contains data
 Observations:            35
    Variables:             4
────────────────────────────────────────────────────────────────────────────────
Variable      Storage   Display    Value
    name         type    format    label      Variable label
────────────────────────────────────────────────────────────────────────────────
club_id         byte    %8.0g
club_ord        byte    %8.0g
club_hgt        double  %10.0g
question        str4    %9s
────────────────────────────────────────────────────────────────────────────────
Sorted by:
     Note: Dataset has changed since last saved.

. cluster dendrogram club, labels(question)
>                 xlabel(, angle(90) labsize(*.75))
>                 title(Single-linkage clustering)
>                 ytitle(1 - Jaccard similarity, suffix)
```



From these 30 women, we see that the biog (enjoy reading biographies) and hist (enjoy reading history) questions were most closely related. auto (enjoy automobile mechanics), hors (have a horse), and bird (have a bird) seem to be the least related to the other variables. These three variables, in turn, merge last into the supergroup containing the remaining variables.

◁

References

Anderson, E. 1935. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society* 59: 2–5.

Bray, R. J., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* 27: 325–349. https://doi.org/10.2307/1942268.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

Kaufman, L., and P. J. Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley. https://doi.org/10.1002/9780470316801.

Morrison, D. F. 2005. *Multivariate Statistical Methods*. 4th ed. Belmont, CA: Duxbury.

Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. *Ecology* 31: 587–605. https://doi.org/10.2307/1931577.

Also see

[MV] **cluster programming subroutines** — Add cluster-analysis routines

[MV] **cluster programming utilities** — Cluster-analysis programming utilities

[MV] **cluster** — Introduction to cluster-analysis commands

For suggested citations, see the FAQ on citing Stata documentation.