Title

ca — Simple correspondence analysis

Syntax

Simple correspondence analysis of two categorical variables

`ca` *rowvar colvar* [*if*] [*in*] [*weight*] [, *options*]

Simple correspondence analysis with crossed (stacked) variables

`ca` *row_spec col_spec* [*if*] [*in*] [*weight*] [, *options*]

Simple correspondence analysis of an $n_r \times n_c$ *matrix*

`camat` *matname* [, *options*]

where *spec* = *varname* | (*newvar* : *varlist*)

options	Description
Model 2	
`dimensions(#)`	number of dimensions (factors, axes); default is `dim(2)`
`normalize(`*nopts*`)`	normalization of row and column coordinates
`rowsupp(`$matname_r$`)`	matrix of supplementary rows
`colsupp(`$matname_c$`)`	matrix of supplementary columns
`rowname(`*string*`)`	label for rows
`colname(`*string*`)`	label for columns
`missing`	treat missing values as ordinary values (`ca` only)
Codes (`ca` only)	
`report(variables)`	report coding of crossing variables
`report(crossed)`	report coding of crossed variables
`report(all)`	report coding of crossing and crossed variables
`length(min)`	use minimal length unique codes of crossing variables
`length(#)`	use # as coding length of crossing variables
Reporting	
`ddimensions(#)`	number of singular values to be displayed; default is `ddim(.)`
`norowpoints`	suppress table with row category statistics
`nocolpoints`	suppress table with column category statistics
`compact`	display tables in a compact format
`plot`	plot the row and column coordinates
`maxlength(#)`	maximum number of characters for labels; default is `maxlength(12)`

nopts	Description
`symmetric`	symmetric coordinates (`canonical`); the default
`standard`	row and column standard coordinates
`row`	row principal, column standard coordinates
`column`	column principal, row standard coordinates
`principal`	row and column principal coordinates
#	power $0 \leq \# \leq 1$ for row coordinates; seldom used

`bootstrap`, `by`, `jackknife`, `rolling`, and `statsby` are allowed with `ca`; see **[U] 11.1.10 Prefix commands**. However, `bootstrap` and `jackknife` results should be interpreted with caution; identification of the `ca` parameters involves data-dependent restrictions, possibly leading to badly biased and overdispersed estimates (Milan and Whittaker 1995).

Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.

`aweight`s are not allowed with the `jackknife` prefix; see [R] **jackknife**.

`fweight`s, `aweight`s, and `iweight`s are allowed with `ca`; see [U] **11.1.6 weight**.

See **[U] 20 Estimation and postestimation commands** for more capabilities of estimation commands.

Menu

ca

Statistics > Multivariate analysis > Correspondence analysis > Two-way correspondence analysis (CA)

camat

Statistics > Multivariate analysis > Correspondence analysis > Two-way correspondence analysis of a matrix

Description

`ca` performs a simple correspondence analysis (CA) of the cross-tabulation of the integer-valued variables *rowvar* and *colvar* with n_r and n_c categories with $n_r, n_c \geq 2$. CA is formally equivalent to various other geometric approaches, including dual scaling, reciprocal averaging, and canonical correlation analysis of contingency tables (Greenacre 1984, chap. 4).

`camat` performs a simple CA of an $n_r \times n_c$ matrix *matname* having nonnegative entries and strictly positive margins. The correspondence table need not contain frequencies. The labels for the row and column categories are obtained from the matrix row and column names.

Optionally, a CA biplot may be produced. The biplot displays the row and column coordinates within the same two-dimensional graph.

Results may be replayed using `ca` or `camat`; there is no difference.

Options

Model 2

`dimensions(`*#*`)` specifies the number of dimensions (= factors = axes) to be extracted. The default is `dimensions(2)`. If you may specify `dimensions(1)`, the row and column categories are placed on one dimension. *#* should be strictly smaller than the number of rows and the number of columns, counting only the active rows and columns, excluding supplementary rows and columns (see options `rowsupp()` and `colsupp()`).

CA is a hierarchical method, so extracting more dimensions does not affect the coordinates and decomposition of inertia of dimensions already included. The percentages of inertia accounting for the dimensions are in decreasing order as indicated by singular values. The first dimension accounts for the most inertia, followed by the second dimension, and then the third dimension, etc.

normalize(*nopt*) specifies the normalization method, that is, how the row and column coordinates are obtained from the singular vectors and singular values of the matrix of standardized residuals. See *Normalization and interpretation of correspondence analysis* in *Remarks and examples* for a discussion of these different normalization methods.

symmetric, the default, distributes the inertia equally over rows and columns, treating the rows and columns symmetrically. The symmetric normalization is also known as the standard, or canonical, normalization. This is the most common normalization when making a biplot. normalize(symmetric) is equivalent to normalize(0.5). canonical is a synonym for symmetric.

standard specifies that row and column coordinates should be in standard form (singular vectors divided by the square root of mass). This normalization method is not equivalent to normalize(*#*) for any *#*.

row specifies principal row coordinates and standard column coordinates. This option should be chosen if you want to compare row categories. Similarity of column categories should not be interpreted. The biplot interpretation of the relationship between row and column categories is appropriate. normalize(row) is equivalent to normalize(1).

column specifies principal column coordinates and standard row coordinates. This option should be chosen if you want to compare column categories. Similarity of row categories should not be interpreted. The biplot interpretation of the relationship between row and column categories is appropriate. normalize(column) is equivalent to normalize(0).

principal is the normalization to choose if you want to make comparisons among the row categories and among the column categories. In this normalization, comparing row and column points is not appropriate. Thus a biplot in this normalization is best avoided. In the principal normalization, the row and column coordinates are obtained from the left and right singular vectors, multiplied by the singular values. This normalization method is not equivalent to normalize(*#*) for any *#*.

#, $0 \leq \# \leq 1$, is seldom used; it specifies that the row coordinates are obtained as the left singular vectors multiplied by the singular values to the power *#*, whereas the column coordinates equal the right singular vectors multiplied by the singular values to the power $1 - \#$.

rowsupp(*matname*$_r$) specifies a matrix of supplementary rows. *matname*$_r$ should have n_c columns. The row names of *matname*$_r$ are used for labeling. Supplementary rows do not affect the computation of the dimensions and the decomposition of inertia. They are, however, included in the plots and in the table with statistics of the row points. Because supplementary points do not contribute to the dimensions, their entries under the column labeled contrib are left blank.

colsupp(*matname*$_c$) specifies a matrix of supplementary columns. *matname*$_c$ should have n_r rows. The column names of *matname*$_c$ are used for labeling. Supplementary columns do not affect the computation of the dimensions and the decomposition of inertia. They are, however, included in the plots and in the table with statistics of the column points. Because supplementary points do not contribute to the dimensions, their entries under the column labeled contrib are left blank.

rowname(*string*) specifies a label to refer to the rows of the matrix. The default is rowname(rowvar) for ca and rowname(rows) for camat.

colname(*string*) specifies a label to refer to the columns of the matrix. The default is colname(colvar) for ca and colname(columns) for camat.

`missing`, allowed only with `ca`, treats missing values of *rowvar* and *colvar* as ordinary categories to be included in the analysis. Observations with missing values are omitted from the analysis by default.

Codes

`report(`*opt*`)` displays coding information for the crossing variables, crossed variables, or both. `report()` is ignored if you do not specify at least one crossed variable.

`report(variables)` displays the coding schemes of the crossing variables, that is, the variables used to define the crossed variables.

`report(crossed)` displays a table explaining the value labels of the crossed variables.

`report(all)` displays the codings of the crossing and crossed variables.

`length(`*opt*`)` specifies the coding length of crossing variables.

`length(min)` specifies that the minimal-length unique codes of crossing variables be used.

`length(`*#*`)` specifies that the coding length # of crossing variables be used, where # must be between 4 and 32.

Reporting

`ddimensions(`*#*`)` specifies the number of singular values to be displayed. The default is `ddimensions(.)`, meaning all.

`norowpoints` suppresses the table with row point (category) statistics.

`nocolpoints` suppresses the table with column point (category) statistics.

`compact` specifies that the table with point statistics be displayed multiplied by 1,000 as proposed by Greenacre (2007), enabling the display of more columns without wrapping output. The compact tables can be displayed without wrapping for models with two dimensions at line size 79 and with three dimensions at line size 99.

`plot` displays a plot of the row and column coordinates in two dimensions. With row principal normalization, only the row points are plotted. With column principal normalization, only the column points are plotted. In the other normalizations, both row and column points are plotted. You can use `cabiplot` directly if you need another selection of points to be plotted or if you want to otherwise refine the plot; see [MV] **ca postestimation plots**.

`maxlength(`*#*`)` specifies the maximum number of characters for row and column labels in plots. The default is `maxlength(12)`.

Note: the reporting options may be specified during estimation or replay.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Introduction

Correspondence analysis (CA) offers a geometric representation of the rows and columns of a two-way frequency table that is helpful in understanding the similarities between the categories of variables and the association between the variables. For an informal introduction to CA and related metric approaches, see Weller and Romney (1990). Greenacre (2007) provides a much more thorough introduction with few mathematical prerequisites. More advanced treatments are given by Greenacre (1984) and Gower and Hand (1996).

In some respects, CA can be thought of as an analogue to principal components for nominal variables. It is also possible to interpret CA in reciprocal averaging (Greenacre 1984, 96–102; Cox and Cox 2001, 193–200), in optimal scaling (Greenacre 1984, 102–108), and in canonical correlations (Greenacre 1984, 108–116; Gower and Hand 1996, 183–185). Scaling refers to the assignment of scores to the categories of the row and column variables. Different criteria for the assignment of scores have been proposed, generally with different solutions. If the aim is to maximize the correlation between the scored row and column, the problem can be formulated in terms of CA. The optimal scores are the coordinates on the first dimension. The coordinates on the second and subsequent dimensions maximize the correlation between row and column scores subject to orthogonality constraints. See also [MV] **ca postestimation**.

A first example

▷ Example 1: A well-known correspondence analysis example

We illustrate CA with an example of smoking behavior by different ranks of personnel. This example is often used in the CA literature (for example, Greenacre 1984, 55; Greenacre 2007, 66), so you have probably encountered these (artificial) data before. By using these familiar data, we make it easier to relate the literature on CA to the output of the ca command.

```
. use http://www.stata-press.com/data/r13/ca_smoking

. tabulate rank smoking

                           smoking intensity
        rank |      none      light     medium      heavy |     Total
-------------+--------------------------------------------+----------
 senior_mngr |         4          2          3          2 |        11 
 junior_mngr |         4          3          7          4 |        18 
 senior_empl |        25         10         12          4 |        51 
 junior_empl |        18         24         33         13 |        88 
   secretary |        10          6          7          2 |        25 
-------------+--------------------------------------------+----------
       Total |        61         45         62         25 |       193 
```

ca displays the results of a CA on two categorical variables in a multipanel format.

```
. ca rank smoking

Correspondence analysis                          Number of obs      =        193
                                                 Pearson chi2(12)   =      16.44
                                                 Prob > chi2        =     0.1718
                                                 Total inertia      =     0.0852
    5 active rows                                Number of dim.     =          2
    4 active columns                             Expl. inertia (%)  =      99.51

                |   singular    principal                                cumul
      Dimension |    value       inertia          chi2    percent   percent
    ------------+-------------------------------------------------------------
          dim 1 |   .2734211    .0747591         14.43      87.76      87.76
          dim 2 |   .1000859    .0100172          1.93      11.76      99.51
          dim 3 |   .0203365    .0004136          0.08       0.49     100.00
    ------------+-------------------------------------------------------------
          total |               .0851899         16.44     100

Statistics for row and column categories in symmetric normalization

                 |          overall          |        dimension_1
      Categories |    mass  quality   %inert |   coord   sqcorr  contrib
    -------------+---------------------------+----------------------------
    rank         |                           |
     senior mngr |   0.057    0.893    0.031 |   0.126    0.092    0.003
     junior mngr |   0.093    0.991    0.139 |  -0.495    0.526    0.084
     senior empl |   0.264    1.000    0.450 |   0.728    0.999    0.512
     junior empl |   0.456    1.000    0.308 |  -0.446    0.942    0.331
       secretary |   0.130    0.999    0.071 |   0.385    0.865    0.070
    -------------+---------------------------+----------------------------
    smoking      |                           |
            none |   0.316    1.000    0.577 |   0.752    0.994    0.654
           light |   0.233    0.984    0.083 |  -0.190    0.327    0.031
          medium |   0.321    0.983    0.148 |  -0.375    0.982    0.166
           heavy |   0.130    0.995    0.192 |  -0.562    0.684    0.150
    -------------+---------------------------+----------------------------

                 |      dimension_2
      Categories |   coord   sqcorr  contrib
    -------------+----------------------------
    rank         |
     senior mngr |   0.612    0.800    0.214
     junior mngr |   0.769    0.465    0.551
     senior empl |   0.034    0.001    0.003
     junior empl |  -0.183    0.058    0.152
       secretary |  -0.249    0.133    0.081
    -------------+----------------------------
    smoking      |
            none |   0.096    0.006    0.029
           light |  -0.446    0.657    0.463
          medium |  -0.023    0.001    0.002
           heavy |   0.625    0.310    0.506
    -------------+----------------------------
```

The order in which we specify the variables is mostly immaterial. The first variable (rank) is also called the row variable, and the second (smoking) is the column variable. This ordering is important only as far as the interpretation of some options and some labeling of output are concerned. For instance, the option norowpoints suppresses the table with row points, that is, the categories of rank. ca requires two integer-valued variables. The rankings of the categories and the actual values used to code categories are not important. Thus, rank may be coded 1, 2, 3, 4, 5, or 0, 1, 4, 9, 16, or −2, −1, 0, 1, 2; it does not matter. We do suggest assigning value labels to the variables to improve the interpretability of tables and plots.

Correspondence analysis seeks to offer a low-dimensional representation describing how the row and column categories contribute to the inertia in a table. ca reports Pearson's test of independence, just like tabulate with the chi2 option. Inertia is Pearson's chi-squared statistic divided by the sample size, $16.44/193 = 0.0852$. Pearson's chi-squared test has significance level $p = 0.1718$, casting doubt on any association between rows and columns. Still, given the prominence of this example in the CA literature, we will continue.

The first panel produced by ca displays the decomposition of total inertia in orthogonal dimensions—analogous to the decomposition of the total variance in principal component analysis (see [MV] **pca**). The first dimension accounts for 87.76% of the inertia; the second dimension accounts for 11.76% of the inertia. Because the dimensions are orthogonal, we may add the contributions of the two dimensions and say that the two leading dimensions account for $87.76\% + 11.76\% = 99.52\%$ of the total inertia. A two-dimensional representation seems in order. The remaining output is discussed later.

◁

How many dimensions?

▷ Example 2: Specifying the number of dimensions

In the first example with the smoking data, we displayed coordinates and statistics for a two-dimensional approximation of the rows and columns. This is the default. We can specify more or fewer dimensions with the option dimensions(). The maximum number is $\min(n_r - 1, n_c - 1)$. At this maximum, the chi-squared distances between the rows and columns are exactly represented by CA; 100% of the inertia is accounted for. This is called the saturated model; the fitted values of the CA model equal the observed correspondence table.

The minimum number of dimensions is one; the model with zero dimensions would be a model of independence of the rows and columns. With one dimension, the rows and columns of the table are identified by points on a line, with distance on the line approximating the chi-squared distance in the table, and a biplot is no longer feasible.

```
. ca rank smoking, dim(1)
Correspondence analysis                          Number of obs      =       193
                                                 Pearson chi2(12)   =     16.44
                                                 Prob > chi2        =    0.1718
                                                 Total inertia      =    0.0852
    5 active rows                                Number of dim.     =         1
    4 active columns                             Expl. inertia (%)  =     87.76

                 |   singular    principal                             cumul
      Dimension  |    value       inertia        chi2     percent    percent
    -------------+--------------------------------------------------------------
          dim 1  |   .2734211    .0747591       14.43       87.76      87.76
          dim 2  |   .1000859    .0100172        1.93       11.76      99.51
          dim 3  |   .0203365    .0004136        0.08        0.49     100.00
    -------------+--------------------------------------------------------------
          total  |               .0851899       16.44      100
```

```
     Statistics for row and column categories in symmetric normalization
         ─────────────┬─────────────────────────┬────────────────────────
                      │          overall        │       dimension_1
           Categories │  mass  quality  %inert  │  coord   sqcorr  contrib
         ─────────────┼─────────────────────────┼────────────────────────
         rank         │                         │
          senior mngr │ 0.057    0.092   0.031  │  0.126    0.092    0.003
          junior mngr │ 0.093    0.526   0.139  │ -0.495    0.526    0.084
          senior empl │ 0.264    0.999   0.450  │  0.728    0.999    0.512
          junior empl │ 0.456    0.942   0.308  │ -0.446    0.942    0.331
            secretary │ 0.130    0.865   0.071  │  0.385    0.865    0.070
         ─────────────┼─────────────────────────┼────────────────────────
         smoking      │                         │
                 none │ 0.316    0.994   0.577  │  0.752    0.994    0.654
                light │ 0.233    0.327   0.083  │ -0.190    0.327    0.031
               medium │ 0.321    0.982   0.148  │ -0.375    0.982    0.166
                heavy │ 0.130    0.684   0.192  │ -0.562    0.684    0.150
         ─────────────┴─────────────────────────┴────────────────────────
```

The first panel produced by `ca` does not depend on the number of dimensions extracted; thus, we will always see all singular values and the percentage of inertia explained by the associated dimensions. In the second panel, the only thing that depends on the number of dimensions is the overall quality of the approximation. The overall quality is the sum of the quality scores on the extracted dimensions and so increases with the number of extracted dimensions. The higher the quality, the better the chi-squared distances with other rows (columns) are represented by the extracted number of dimensions. In a saturated model, the overall quality is 1 for each row and column category.

So, how many dimensions should we retain? It is common for researchers to extract the minimum number of dimensions in a CA to explain at least 90% of the inertia, analogous to similar heuristic rules on the number of components in principal component analysis. We could probably also search for a scree, the number of dimensions where the singular values flatten out (see [MV] **screeplot**). A `screeplot` of the singular values can be obtained by typing

```
. screeplot e(Sv)
  (output omitted )
```

where `e(Sv)` is the name where `ca` has stored the singular values.

◁

Statistics on the points

▷ Example 3: A more compact table of row and column statistics

We now turn our attention to the second panel. The overall section of the panel lists the following statistics:

- The mass of the category, that is, the proportion in the marginal distribution. The masses of all categories of a variable add up to 1.
- The quality of the approximation for a category, expressed as a number between 0 (very bad) and 1 (perfect). In a saturated model, quality is 1.
- The percentage of inertia contained in the category. Categories are divided through by the total inertia; the inertias of the categories of a variable add up to 100%.

For each of the dimensions, the panel lists the following:

- The coordinate of the category.

- The squared residuals between the profile and the categories. The sum of the squared residuals over the dimensions adds up to the quality of the approximation for the category.
- The contribution made by the categories to the dimensions. These add up to 1 over all categories of a variable.

The table with point statistics becomes pretty large, especially with more than two dimensions. `ca` can also list the second panel in a more compact form, saving space by multiplying all entries by 1,000; see Greenacre (2007).

```
. ca rank smoking, dim(2) compact
Correspondence analysis                         Number of obs     =       193
                                                Pearson chi2(12)  =     16.44
                                                Prob > chi2       =    0.1718
                                                Total inertia     =    0.0852
    5 active rows                               Number of dim.    =         2
    4 active columns                            Expl. inertia (%) =     99.51

             |   singular    principal                                cumul
   Dimension |     value      inertia         chi2     percent     percent
  -----------+------------------------------------------------------------
       dim 1 |   .2734211    .0747591        14.43       87.76       87.76
       dim 2 |   .1000859    .0100172         1.93       11.76       99.51
       dim 3 |   .0203365    .0004136         0.08        0.49      100.00
  -----------+------------------------------------------------------------
       total |               .0851899        16.44         100
Statistics for row and column categories in symmetric norm. (x 1000)

  ------------------- overall ------------ dimension 1 ------- dimension 2 --
  Categories | mass qualt %inert | coord sqcor contr | coord sqcor contr
  -----------+-------------------+-------------------+------------------
  rank       |                   |                   |
   senior mngr |   57   893     31 |   126    92     3 |   612   800   214
   junior mngr |   93   991    139 |  -495   526    84 |   769   465   551
   senior empl |  264  1000    450 |   728   999   512 |    34     1     3
   junior empl |  456  1000    308 |  -446   942   331 |  -183    58   152
     secretary |  130   999     71 |   385   865    70 |  -249   133    81
  -----------+-------------------+-------------------+------------------
  smoking    |                   |                   |
        none |  316  1000    577 |   752   994   654 |    96     6    29
       light |  233   984     83 |  -190   327    31 |  -446   657   463
      medium |  321   983    148 |  -375   982   166 |   -23     1     2
       heavy |  130   995    192 |  -562   684   150 |   625   310   506
  ------------------------------------------------------------------------
```

◁

Normalization and interpretation of correspondence analysis

The normalization method used in CA determines whether and how the similarity of the row categories, the similarity of the column categories, and the relationship (association) between the row and column variables can be interpreted in terms of the row and column coordinates and the origin of the plot.

How does one compare row points—provided that the normalization method allows such a comparison? Formally, the Euclidean distance between the row points approximates the chi-squared distances between the corresponding row profiles. Thus in the biplot, row categories mapped close together have similar row profiles; that is, the distributions on the column variable are similar. Row categories mapped widely apart have dissimilar row profiles. Moreover, the Euclidean distance between a row point and the origin approximates the chi-squared distance from the row profile and the row centroid, so it indicates how different a category is from the population.

An analogous interpretation applies to column points.

For the association between the row and column variables: in the CA biplot, you should not interpret the distance between a row point r and a column point c as the relationship of r and c. Instead, think in terms of the vectors origin to r (OR) and origin to c (OC). Remember that CA decomposes scaled deviations $d(r,c)$ from independence and $d(r,c)$ is approximated by the inner product of OR and OC. The larger the absolute value of $d(r,c)$, the stronger the association between r and c. In geometric terms, $d(r,c)$ can be written as the product of the length of OR, the length of OC, and the cosine of the angle between OR and OC.

What does this mean? First, consider the effects of the angle. The association in (r,c) is strongly positive if OR and OC point in roughly the same direction; the frequency of (r,c) is much higher than expected under independence, so r tends to flock together with c—if the points r and c are close together. Similarly, the association is strongly negative if OR and OC point in opposite directions. Here the frequency of (r,c) is much lower than expected under independence, so r and c are unlikely to occur simultaneously. Finally, if OR and OC are roughly orthogonal (angle = ±90), the deviation from independence is small.

Second, the association of r and c increases with the lengths of OR and OC. Points far from the origin tend to have large associations. If a category is mapped close to the origin, all its associations with categories of the other variable are small: its distribution resembles the marginal distribution.

Here are the interpretations enabled by the main normalization methods as specified in the `normalize()` option.

Normalization method	Similarity row cat.	Similarity column cat.	Association row vs. column
`symmetric`	No	No	Yes
`principal`	Yes	Yes	No
`row`	Yes	No	Yes
`column`	No	Yes	Yes

If we say that a comparison between row categories or between column categories is not possible, we really mean that the chi-squared distance between row profiles or column profiles is actually approximated by a weighted Euclidean distance between the respective plots in which the weights depend on the inertia of the dimensions rather than on the standard Euclidean distance.

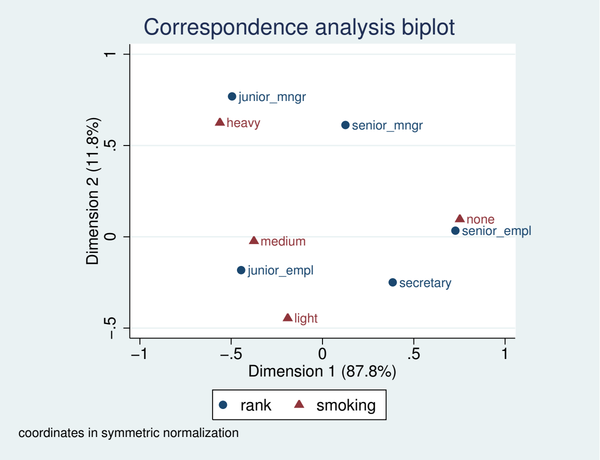
You may want to do a CA in principal normalization to study the relationship between the categories of a variable and do a CA in symmetric normalization to study the association of the row and column categories.

Plotting the points

▷ Example 4: A correspondence biplot

In our discussion of normalizations, we stated that CA offers simple geometric interpretations to the similarity of categories and the association of the variables. We may specify the option `plot` with `ca` during estimation or during replay.

```
. ca, norowpoint nocolpoint plot
Correspondence analysis                        Number of obs     =       193
                                               Pearson chi2(12)  =     16.44
                                               Prob > chi2       =    0.1718
                                               Total inertia     =    0.0852
    5 active rows                              Number of dim.    =         2
    4 active columns                           Expl. inertia (%) =     99.51

             |   singular    principal                             cumul
   Dimension |     value      inertia        chi2     percent     percent
  -----------+-----------------------------------------------------------
       dim 1 |   .2734211    .0747591       14.43       87.76       87.76
       dim 2 |   .1000859    .0100172        1.93       11.76       99.51
       dim 3 |   .0203365    .0004136        0.08        0.49      100.00
  -----------+-----------------------------------------------------------
       total |               .0851899       16.44      100
```



The options `norowpoint` and `nocolpoint` suppress the large tables of statistics for the rows and columns. If we did not request the plot during estimation, we can still obtain it with the `cabiplot` postestimation command. Unlike requesting the plot at estimation time, `cabiplot` allows us to fine-tune the plot; see [MV] **ca postestimation plots**.

The horizontal dimension seems to distinguish smokers from nonsmokers, whereas the vertical dimensions can be interpreted as intensity of smoking. Because the orientations from the origin to `none` and from the origin to `senior_empl` are so close, we conclude that senior employees tend not to smoke. Similarly, junior managers tend to be heavy smokers, and junior employees tend to be medium smokers.

◁

Supplementary points

A useful feature of CA is the ability to locate supplementary rows and columns in the space generated by the "active" rows and columns (see Greenacre [1984, 70–74]; Greenacre [2007, chap. 12], for an extensive discussion). Think of supplementary rows and columns as having mass 0; therefore, supplementary points do not influence the approximating space—their contribution values are zero.

▷ Example 5: Supplementary rows and columns

In our example, we want to include the national distribution of smoking intensity as a supplementary row.

ca requires that we define the supplementary row distributions as rows of a matrix. In this example, we have only one supplementary row, with the percentages of the smoking categories in a national sample. The matrix should have one row per supplementary row category and as many columns as there are active columns. We define the row name to obtain appropriately labeled output.

```
. matrix S_row = (42, 29, 20, 9)
. matrix rowname S_row = national
```

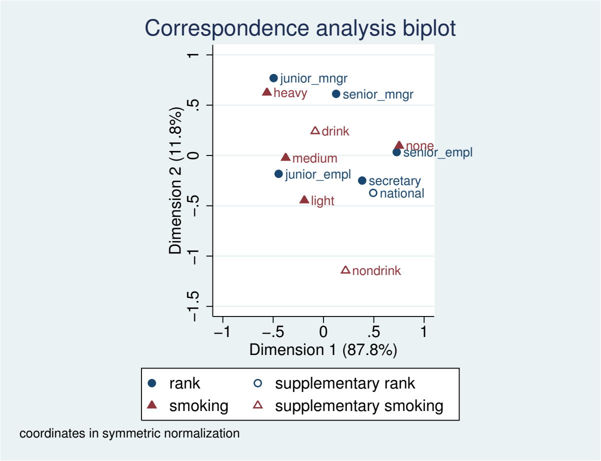
Before we show the CA analysis with the supplementary row, we also include two supplementary columns for the rank distribution of alcoholic beverage drinkers and nondrinkers. It will be interesting to see where smoking is located relative to drinking and nondrinking.

```
. matrix S_col = ( 0, 11 \
>                  1, 19 \
>                  5, 44 \
>                 10, 78 \
>                  7, 18)
. matrix colnames S_col = nondrink drink
```

We now invoke ca, specifying the names of the matrices with supplementary rows and columns with the options rowsupp() and colsupp().

```
. ca rank smoking, rowsupp(S_row) colsupp(S_col) plot
Correspondence analysis                          Number of obs     =       193
                                                 Pearson chi2(12)  =     16.44
                                                 Prob > chi2       =    0.1718
                                                 Total inertia     =    0.0852
    5 active + 1 supplementary rows              Number of dim.    =         2
    4 active + 2 supplementary columns           Expl. inertia (%) =     99.51

                 singular   principal                                cumul
    Dimension      value     inertia          chi2     percent     percent

        dim 1   .2734211    .0747591         14.43       87.76       87.76
        dim 2   .1000859    .0100172          1.93       11.76       99.51
        dim 3   .0203365    .0004136          0.08        0.49      100.00

        total               .0851899         16.44      100
```

```
Statistics for row and column categories in symmetric normalization

                 |          overall          |        dimension_1
      Categories |    mass  quality   %inert |   coord   sqcorr  contrib
    -------------+---------------------------+----------------------------
    rank         |                           |
     senior mngr |   0.057    0.893    0.031 |   0.126    0.092    0.003
     junior mngr |   0.093    0.991    0.139 |  -0.495    0.526    0.084
     senior empl |   0.264    1.000    0.450 |   0.728    0.999    0.512
     junior empl |   0.456    1.000    0.308 |  -0.446    0.942    0.331
       secretary |   0.130    0.999    0.071 |   0.385    0.865    0.070
    -------------+---------------------------+----------------------------
    suppl_rows   |                           |
        national |   0.518    0.761    0.644 |   0.494    0.631
    -------------+---------------------------+----------------------------
    smoking      |                           |
            none |   0.316    1.000    0.577 |   0.752    0.994    0.654
           light |   0.233    0.984    0.083 |  -0.190    0.327    0.031
          medium |   0.321    0.983    0.148 |  -0.375    0.982    0.166
           heavy |   0.130    0.995    0.192 |  -0.562    0.684    0.150
    -------------+---------------------------+----------------------------
    suppl_cols   |                           |
        nondrink |   0.119    0.439    0.460 |   0.220    0.040
           drink |   0.881    0.838    0.095 |  -0.082    0.202
    -------------+---------------------------+----------------------------

                 |        dimension_2
      Categories |   coord   sqcorr  contrib
    -------------+---------------------------
    rank         |
     senior mngr |   0.612    0.800    0.214
     junior mngr |   0.769    0.465    0.551
     senior empl |   0.034    0.001    0.003
     junior empl |  -0.183    0.058    0.152
       secretary |  -0.249    0.133    0.081
    -------------+---------------------------
    suppl_rows   |
        national |  -0.372    0.131
    -------------+---------------------------
    smoking      |
            none |   0.096    0.006    0.029
           light |  -0.446    0.657    0.463
          medium |  -0.023    0.001    0.002
           heavy |   0.625    0.310    0.506
    -------------+---------------------------
    suppl_cols   |
        nondrink |  -1.144    0.398
           drink |   0.241    0.636
    -------------+---------------------------
```



The first panel and the information about the five active rows and the four active columns have not changed—the approximating space is fully determined by the active rows and columns and is independent of the location of the supplementary rows and columns.

The table with statistics for the row and column categories now also contains entries for the supplementary rows and columns. The `contrib` entries for the supplementary points are blank. Supplementary points do not "contribute to" the location of the dimensions—their contribution is 0.000, but displaying blanks makes the point more clearly. All other columns for the supplementary points are informative. The inertia of supplementary points is the chi-squared distance to the respective centroid. The coordinates of supplementary points are obtained by applying the transition equations of the CA. Correlations of the supplementary profiles with the dimensions are also well defined. Finally, we may consider the quality of the two-dimensional approximation for the supplementary points. These are lower than for the active points, which will be the case in most applications—the active points exercise influence on the dimensions to improve their quality, whereas the supplementary points simply have to accept the dimensions as determined by the active points.

If we look at the biplot, the supplementary points are shown along with the active points. We may interpret the supplementary points just like the active points. Secretaries are close to the national sample in terms of smoking. Drinking alcohol is closer to the smoking categories than to nonsmoking, indicating that alcohol consumption and smoking are similar behaviors—but concluding that the *same* people smoke and drink is not possible because we do not have three-way data.

◁

Matrix input

▷ Example 6: Correspondence analysis of a frequency table

If we want to do a CA of a published two-way frequency table, we typically do not have immediate access to the data in the form of a dataset. We could enter the data with frequency weights.

```
. input rank smoking freq
  1.       1       1    4
  2.       1       2    2
  3.       1       3    3
 (output omitted )
 19.       5       3    7
 20.       5       4    2
 21.   end
. label define vl_rank  1  "senior_mngr" ...
. label value rank vl_rank
. label define vl_smoke 1  "none" ...
. label value smoke vl_smoke
. ca rank smoking [fw=freq]
 (output omitted )
```

Or we may enter the data as a matrix and use `camat`. First, we enter the frequency matrix with proper column and row names and then list the matrix for verification.

```
. matrix F = (4,2,3,2 \ 4,3,7,4 \ 25,10,12,4 \ 18,24,33,13 \ 10,6,7,2)
. matrix colnames F = none light medium heavy
. matrix rownames F = senior_mngr junior_mngr senior_empl junior_empl secretary
. matlist F, border
```

```
  ----------------------------------------------------------
               |      none      light     medium      heavy
  -------------+--------------------------------------------
   senior_mngr |         4          2          3          2
   junior_mngr |         4          3          7          4
   senior_empl |        25         10         12          4
   junior_empl |        18         24         33         13
     secretary |        10          6          7          2
  ----------------------------------------------------------
```

We can use `camat` on `F` to obtain the same results as from the raw data. We use the `compact` option for a more compact table.

```
. camat F, compact

Correspondence analysis                         Number of obs     =       193
                                                Pearson chi2(12)  =     16.44
                                                Prob > chi2       =    0.1718
                                                Total inertia     =    0.0852
    5 active rows                               Number of dim.    =         2
    4 active columns                            Expl. inertia (%) =     99.51

                |   singular    principal                               cumul
      Dimension |     value      inertia         chi2      percent    percent
    ------------+-----------------------------------------------------------
          dim 1 |   .2734211    .0747591        14.43        87.76      87.76
          dim 2 |   .1000859    .0100172         1.93        11.76      99.51
          dim 3 |   .0203365    .0004136         0.08         0.49     100.00
    ------------+-----------------------------------------------------------
          total |               .0851899        16.44       100

Statistics for row and column categories in symmetric norm. (x 1000)

    ------------- overall ---------- dimension 1 ------ dimension 2 ---
      Categories | mass  qualt %inert | coord sqcor contr | coord sqcor contr
    -------------+--------------------+-------------------+-----------------
    rows         |                    |                   |
     senior mngr |   57   893     31  |   126    92     3 |   612   800   214
     junior mngr |   93   991    139  |  -495   526    84 |   769   465   551
     senior empl |  264  1000    450  |   728   999   512 |    34     1     3
     junior empl |  456  1000    308  |  -446   942   331 |  -183    58   152
       secretary |  130   999     71  |   385   865    70 |  -249   133    81
    -------------+--------------------+-------------------+-----------------
    columns      |                    |                   |
            none |  316  1000    577  |   752   994   654 |    96     6    29
           light |  233   984     83  |  -190   327    31 |  -446   657   463
          medium |  321   983    148  |  -375   982   166 |   -23     1     2
           heavy |  130   995    192  |  -562   684   150 |   625   310   506
    ------------------------------------------------------------------------
```

◁

▷ Example 7: Correspondence analysis of nonfrequency data

The command `camat` may also be used for a CA of nonfrequency data. The data should be nonnegative, with strictly positive margins. An example are the compositional data on the distribution of government R&D funds over 11 areas in five European countries in 1989; the data are listed in Greenacre (1993, 82). The expenditures are scaled to 1,000 within country, to focus the analysis on the intranational distribution policies. Moreover, with absolute expenditures, small countries, such as The Netherlands, would have been negligible in the analysis.

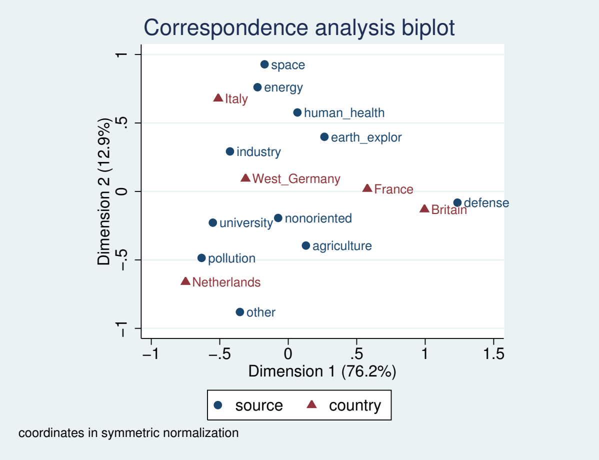
We enter the data as a Stata matrix. The command `matrix input` (see [P] **matrix define**) allows us to input row entries separated by blanks, rather than by commas; rows are separated by the backward slash (\).

```
. matrix input RandD = (
>    18  19  14  14   6 \
>    12  34   4  15  31 \
>    44  33  36  58  25 \
>    37  88  67 101  40 \
>    42  20  36  28  43 \
>    90 156 107 224 176 \
>    28  50  59  88  28 \
>   165 299 120 303 407 \
>    48 128 147  62 103 \
>   484 127 342  70  28 \
>    32  46  68  37 113)
```

```
. matrix colnames RandD = Britain West_Germany France Italy Netherlands

. matrix rownames RandD = earth_exploration pollution human_health
>                         energy agriculture industry space university
>                         nonoriented defense other
```

We perform a CA, suppressing the voluminous row- and column-point statistics. We want to show a biplot, and therefore we select symmetric normalization.

```
. camat RandD, dim(2) norm(symm) rowname(source) colname(country) norowpoints
> nocolpoints plot

Correspondence analysis                         Number of obs     =      5000
                                                Pearson chi2(40)  =   1321.55
                                                Prob > chi2       =    0.0000
                                                Total inertia     =    0.2643
    11 active rows                              Number of dim.    =         2
     5 active columns                           Expl. inertia (%) =     89.08

                 |   singular    principal                           cumul
       Dimension |     value      inertia         chi2    percent   percent
    -------------+-----------------------------------------------------------
           dim 1 |    .448735    .2013631      1006.82      76.18     76.18
           dim 2 |   .1846219    .0340852       170.43      12.90     89.08
           dim 3 |   .1448003    .0209671       104.84       7.93     97.01
           dim 4 |   .0888532    .0078949        39.47       2.99    100.00
    -------------+-----------------------------------------------------------
           total |               .2643103      1321.55       100
```



coordinates in symmetric normalization

The two dimensions account for 89% of the inertia in this example, justifying an interpretation of the biplot. Let us focus on the position of The Netherlands. The orientation of The Netherlands from the origin is in the same direction as the orientation of pollution and university from the origin, indicating that The Netherlands spends more on academic research and on research to reduce environmental pollution than the average country. Earth exploration and human health are in the opposite direction, indicating investments much lower than average in these areas. Industry and agriculture are approximately orthogonal to the orientation of The Netherlands, indicating average investments by The Netherlands in these areas. Britain and France have big military investments, whereas Germany and Italy have more of an industrial orientation.

◁

Technical note

The interpretation of the biplot is not fully in line with what we easily see in the row and column profiles—surprisingly, Greenacre does not seem to feel the need to comment on this. Why is this the case? The clue is in the statistics we did not show. Although the two dimensions account for 90% of the total inertia, this does not mean that all rows and columns are approximated to this extent. There are some row and column categories that are not well described in two dimensions. For instance, the quality of the `source` categories `nonoriented`, `agriculture`, and `earth_exploration` are only 0.063, 0.545, and 0.584, respectively, indicating that these rows are poorly represented in a two-dimensional space. The quality of `West_Germany` is also rather low at 0.577. Adding a third dimension improves the quality of the category `nonoriented` but hardly affects the other two problematic categories. This effect can be seen only from the squared correlations between the third dimension and the profiles of the row and column categories—these correlations are small for all categories but `nonoriented`. Thus, `nonoriented` does not seem to really belong with the other categories and should probably be omitted from the analysis.

Crossed variables

`ca` can include interactions between variables in the analysis; variables that contain interactions are called crossed or stacked variables, whereas the variables that make them up are the crossing or stacking variables.

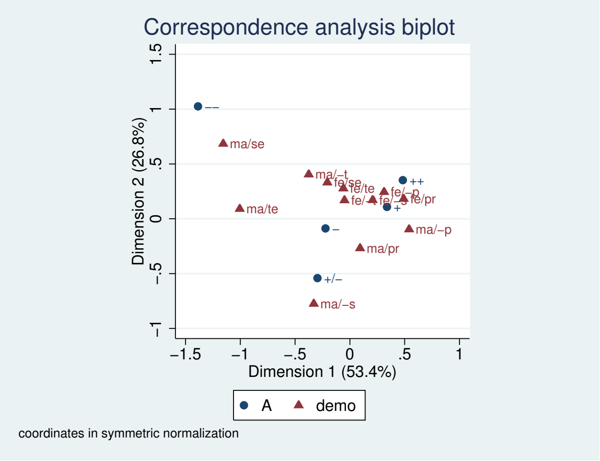
Example 8: Correspondence analysis with crossed variables

We illustrate crossed variables with `ca` by using the ISSP (1993) data from [MV] **mca**, which explores attitudes toward science and the environment. We are interested in whether responses to item A differ with education and gender. The item asks for a response to the statement “We believe too often in science, and not enough in feelings or faith,” with a 1 indicating strong agreement and a 5 indicating strong disagreement. We are interested in how education and gender influence response. We cross the variables `sex` and `edu` into one demographic variable labeled `demo` to explore this question.

```
. use http://www.stata-press.com/data/r13/issp93
(Selection from ISSP (1993))
. tabulate A edu
too much science, not  |        education (6 categories)
enough feelings&faith  | primary i  primary c  secondary  secondary |     Total
-----------------------+--------------------------------------------+----------
       agree strongly  |         7         59         29         11 |       119
                agree  |        15        155         84         27 |       322
neither agree nor dis  |         7         84         65         18 |       204
             disagree  |         8         68         54         26 |       178
    disagree strongly  |         1         12         10         12 |        48
-----------------------+--------------------------------------------+----------
                Total  |        38        378        242         94 |       871

                       |     education (6
too much science, not  |      categories)
enough feelings&faith  |  tertiary   tertiary |     Total
-----------------------+----------------------+----------
       agree strongly  |         5          8 |       119
                agree  |        20         21 |       322
neither agree nor dis  |        11         19 |       204
             disagree  |         8         14 |       178
    disagree strongly  |         5          8 |        48
-----------------------+----------------------+----------
                Total  |        49         70 |       871
```

We notice immediately the long labels for variable A and on edu. We replace these labels with short labels that can be abbreviated, so that in our analysis we will easily be able to identify categories. We use the length(2) option to ca to ensure that labels from each of the crossing variables are restricted to two characters.

```
. label define response 1 "++" 2 "+" 3 "+/-" 4 "-" 5 "--"
. label values A response
. label define education 1 "-pri" 2 "pri" 3 "-sec" 4 "sec" 5 "-ter" 6 "ter"
. label values edu education
. ca A (demo: sex edu), norowpoints nocolpoints length(2) plot norm(symmetric)
Correspondence analysis                        Number of obs     =       871
                                               Pearson chi2(44)  =     72.52
                                               Prob > chi2       =    0.0043
                                               Total inertia     =    0.0833
    5 active rows                              Number of dim.    =         2
    12 active columns                          Expl. inertia (%) =     80.17

                 |   singular    principal                            cumul
       Dimension |     value      inertia       chi2     percent    percent
    -------------+----------------------------------------------------------
           dim 1 |   .2108455     .0444558     38.72       53.39      53.39
           dim 2 |     .14932     .0222965     19.42       26.78      80.17
           dim 3 |   .1009876     .0101985      8.88       12.25      92.42
           dim 4 |   .0794696     .0063154      5.50        7.58     100.00
    -------------+----------------------------------------------------------
           total |                .0832662     72.52      100
```



We see clearly that the responses of the males vary more widely than those of the females. Strong agreement with item A is most closely associated with females with little education, and strong disagreement is most closely associated with males with a secondary or tertiary education. Educated males are more closely associated with a negative response than educated females are, and females with little education are more closely associated with a positive response than males with little education are.

◁

Stored results

Let r be the number of rows, c be the number of columns, and f be the number of retained dimensions. `ca` and `camat` store the following in `e()`:

Scalars

`e(N)`	number of observations
`e(f)`	number of dimensions (factors, axes); maximum of min($r-1$,$c-1$)
`e(inertia)`	total inertia = `e(X2)/e(N)`
`e(pinertia)`	inertia explained by `e(f)` dimensions
`e(X2)`	χ^2 statistic
`e(X2_df)`	degrees of freedom $(r-1)(c-1)$
`e(X2_p)`	p-value for `e(X2)`

Macros

`e(cmd)`	`ca` (even for `camat`)
`e(cmdline)`	command as typed
`e(Rcrossvars)`	row crossing variable names (`ca` only)
`e(Ccrossvars)`	column crossing variable names (`ca` only)
`e(varlist)`	the row and column variable names (`ca` only)
`e(wtype)`	weight type (`ca` only)
`e(wexp)`	weight expression (`ca` only)
`e(title)`	title in estimation output
`e(ca_data)`	`variables` or `crossed`
`e(Cname)`	name for columns
`e(Rname)`	name for rows
`e(norm)`	normalization method
`e(sv_unique)`	`1` if the singular values are unique, `0` otherwise
`e(properties)`	`nob noV eigen`
`e(estat_cmd)`	program used to implement `estat`
`e(predict)`	program used to implement `predict`
`e(marginsnotok)`	predictions disallowed by `margins`

Matrices

`e(Ccoding)`	column categories ($1\times c$) (`ca` only)
`e(Rcoding)`	row categories ($1\times r$) (`ca` only)
`e(GSC)`	column statistics ($c\times 3(1+f)$)
`e(GSR)`	row statistics ($r\times 3(1+f)$)
`e(TC)`	normalized column coordinates ($c\times f$)
`e(TR)`	normalized row coordinates ($r\times f$)
`e(Sv)`	singular values ($1\times f$)
`e(C)`	column coordinates ($c\times f$)
`e(R)`	row coordinates ($r\times f$)
`e(c)`	column mass (margin) ($c\times 1$)
`e(r)`	row mass (margin) ($r\times 1$)
`e(P)`	analyzed matrix ($r\times c$)
`e(GSC_supp)`	supplementary column statistics
`e(GSR_supp)`	supplementary row statistics
`e(PC_supp)`	principal coordinates supplementary column points
`e(PR_supp)`	principal coordinates supplementary row points
`e(TC_supp)`	normalized coordinates supplementary column points
`e(TR_supp)`	normalized coordinates supplementary row points

Functions

`e(sample)`	marks estimation sample (`ca` only)

Methods and formulas

Our presentation of simple CA follows that of Greenacre (1984, 83–125); see also Blasius and Greenacre (1994) and Rencher and Christensen (2012, 565–580). See Greenacre and Blasius (1994) for a concise presentation of CA from a computational perspective. Simple CA seeks a geometric representation of the rows and column of a (two mode) matrix with nonnegative entries in a common low-dimensional space so that chi-squared distances between the rows and between the columns are well approximated by the Euclidean distances in the common space.

Let $\mathbf{N}$ be an $I \times J$ matrix with nonnegative entries and strictly positive margins. $\mathbf{N}$ may be frequencies of a two-way cross-tabulation, but this is not assumed in most of CA. Let $n = N_{++}$ be the overall sum of N_{ij} ("number of observations"). Define the *correspondence table* as the matrix $\mathbf{P}$ where $P_{ij} = N_{ij}/n$, so the overall sum of P_{ij} is $P_{++} = 1$. Let $\mathbf{r} = \mathbf{P}\,\mathbf{1}$ be the row margins, also known as the *row masses*, with elements $r_i > 0$. Similarly, $\mathbf{c} = \mathbf{P}'\mathbf{1}$ contains the column margins, or *column masses*, with elements $c_j > 0$.

CA is defined in terms of the generalized singular value decomposition (GSVD) of $\mathbf{P} - \mathbf{r}\mathbf{c}'$ with respect to the inner products normed by $\mathbf{D}_r^{-1}$ and $\mathbf{D}_c^{-1}$, where $\mathbf{D}_r = \text{diag}(\mathbf{r})$ and $\mathbf{D}_c = \text{diag}(\mathbf{c})$. The GSVD can be expressed in terms of the orthonormal (or standard) SVD of the standardized residuals

$$\mathbf{Z} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}')\mathbf{D}_c^{-\frac{1}{2}} \quad \text{with elements} \quad Z_{ij} = \frac{P_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

Denote by $\mathbf{Z} = \mathbf{R}\boldsymbol{\Lambda}\mathbf{C}'$ the SVD of $\mathbf{Z}$ with $\mathbf{R}'\mathbf{R} = \mathbf{C}'\mathbf{C} = \mathbf{I}$ and $\boldsymbol{\Lambda}$ a diagonal matrix with singular values in decreasing order. `ca` displays a warning message if $\mathbf{Z}$ has common singular values.

The *total principal inertia* of the correspondence table $\mathbf{P}$ is defined as $\chi^2/n = \sum_{i,j} Z_{ij}^2$, where χ^2 is Pearson's chi-squared statistic. We can express the inertia of $\mathbf{P}$ in terms of the singular values of $\mathbf{Z}$:

$$\text{inertia} = \frac{1}{n}\chi^2 = \sum_{k=1}^{\min(I-1,J-1)} \lambda_k^2$$

The inertia accounted for by d dimensions is $\sum_{k=1}^{d} \lambda_k^2$. The fraction of inertia accounted for (explained) by the d dimensions is defined as

$$\text{explained inertia} = \frac{\sum_{k=1}^{d} \lambda_k^2}{\sum_{k=1}^{\min(I-1,J-1)} \lambda_k^2}$$

Principal row ($\widetilde{R}_{ik}$) and principal column ($\widetilde{C}_{jk}$) coordinates are defined as

$$\widetilde{R}_{ik} = \frac{R_{ik}\lambda_k}{\sqrt{r_i}} = (\mathbf{D}_r^{-\frac{1}{2}}\mathbf{R}\boldsymbol{\Lambda})_{ik} \qquad \widetilde{C}_{jk} = \frac{C_{jk}\lambda_k}{\sqrt{c_j}} = (\mathbf{D}_c^{-\frac{1}{2}}\mathbf{C}\boldsymbol{\Lambda})_{jk}$$

The α-normalized row and column coordinates are defined as

$$R_{ik}^{(\alpha)} = \frac{R_{ik}\lambda_k^{\alpha}}{\sqrt{r_i}} \qquad C_{jk}^{(\alpha)} = \frac{C_{jk}\lambda_k^{1-\alpha}}{\sqrt{c_j}}$$

The row principal coordinates are obtained with $\alpha = 1$. The column principal coordinates are obtained with $\alpha = 0$. The symmetric coordinates are obtained with $\alpha = 1/2$.

Decomposition of inertia by rows ($\text{In}^{(r)}$) and by columns ($\text{In}^{(c)}$) is defined as

$$\text{In}_i^{(r)} = \sum_{j=1}^{J} Z_{ij}^2 \qquad \text{In}_j^{(c)} = \sum_{i=1}^{I} Z_{ij}^2$$

Quality of subspace approximations for the row and column categories are defined as

$$Q_i^{(r)} = \frac{r_i}{\mathrm{In}_i^{(r)}} \sum_{k=1}^{d} \widetilde{R}_{ik}^2 \qquad Q_j^{(c)} = \frac{c_j}{\mathrm{In}_j^{(c)}} \sum_{k=1}^{d} \widetilde{C}_{jk}^2$$

If $d = \min(I-1, J-1)$, the quality index satisfies $Q_i^{(r)} = Q_j^{(c)} = 1$.

CA provides several diagnostics for a more detailed analysis of inertia: what do the categories contribute to the inertia explained by the dimensions, and what do the dimensions contribute to the inertia explained for the categories?

The relative contributions of row i ($G_{ik}^{(r)}$) and of column j ($G_{jk}^{(c)}$) to the inertia of principal dimension k are defined as

$$G_{ik}^{(r)} = \frac{r_i \widetilde{R}_{ik}^2}{\lambda_k^2} \qquad G_{jk}^{(c)} = \frac{c_j \widetilde{C}_{jk}^2}{\lambda_k^2}$$

$G_{+k}^{(r)} = G_{+k}^{(c)} = 1$.

The correlations $H_{ik}^{(r)}$ of the ith row profile and kth principal row dimension and, analogously, $H_{jk}^{(c)}$ for columns are

$$H_{ik}^{(r)} = \frac{r_i}{\mathrm{In}_i^{(r)}} \widetilde{R}_{ik}^2 \qquad H_{jk}^{(c)} = \frac{c_j}{\mathrm{In}_j^{(c)}} \widetilde{C}_{jk}^2$$

We now define the quantities returned by the `estat` subcommands after `ca`. The row profiles are $\mathbf{U} = \mathbf{D}_r^{-1}\mathbf{P}$. The chi-squared distance between rows i_1 and i_2 of $\mathbf{P}$ is defined as the Mahalanobis distance between the respective row profiles $\mathbf{U}_{i_1}$ and $\mathbf{U}_{i_2}$ with respect to $\mathbf{D}_c$,

$$(\mathbf{U}_{i_1} - \mathbf{U}_{i_2})\mathbf{D}_c^{-1}(\mathbf{U}_{i_1} - \mathbf{U}_{i_2})'$$

The column profiles and the chi-squared distances between columns are defined analogously. The chi-squared distances for the approximated correspondence table are defined analogously in terms of $\widehat{\mathbf{P}}$.

The fitted or reconstructed values $\widehat{P}_{ij}$ are

$$\widehat{P}_{ij} = r_i c_j \left(1 + \lambda_k^{-1} \sum_{k=1}^{d} \widetilde{R}_{ik} \widetilde{C}_{jk}\right)$$

References

Blasius, J., and M. J. Greenacre. 1994. Computation of correspondence analysis. In *Correspondence Analysis in the Social Sciences*, ed. M. J. Greenacre and J. Blasius. London: Academic Press.

Cox, T. F., and M. A. A. Cox. 2001. *Multidimensional Scaling*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Gower, J. C., and D. J. Hand. 1996. *Biplots*. London: Chapman & Hall.

Greenacre, M. J. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press.

——. 1993. *Correspondence Analysis in Practice*. London: Academic Press.

——. 2007. *Correspondence Analysis in Practice*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Greenacre, M. J., and J. Blasius, ed. 1994. *Correspondence Analysis in the Social Sciences*. London: Academic Press.

ISSP. 1993. International Social Survey Programme: Environment. http://www.issp.org.

Milan, L., and J. C. Whittaker. 1995. Application of the parametric bootstrap to models that incorporate a singular value decomposition. *Applied Statistics* 44: 31–49.

Rencher, A. C., and W. F. Christensen. 2012. *Methods of Multivariate Analysis*. 3rd ed. Hoboken, NJ: Wiley.

Van Kerm, P. 1998. sg78: Simple and multiple correspondence analysis in Stata. *Stata Technical Bulletin* 42: 32–37. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 210–217. College Station, TX: Stata Press.

Weller, S. C., and A. K. Romney. 1990. *Metric Scaling: Correspondence Analysis*. Newbury Park, CA: Sage.

Also see

[MV] **ca postestimation** — Postestimation tools for ca and camat

[MV] **ca postestimation plots** — Postestimation plots for ca and camat

[MV] **mca** — Multiple and joint correspondence analysis

[R] **tabulate twoway** — Two-way table of frequencies

[U] **20 Estimation and postestimation commands**