Chapter 11

Correspondence Analysis

Software and Documentation by:
Bee-Leng Lee

This chapter describes ViSta-Corresp, the ViSta procedure for performing simple correspondence analysis, a way of analyzing a two-way table of data. ViSta-Corresp can analyze row or column profiles, or both. Its visualization includes a spinplot and scatterplot of the row and column points, a residual plot and a fit plot. ViSta-Corresp supports model re-vision. ViSta-Corresp does not perform multiple Correspondence Analysis of multi-way data.

11.1 Introduction to Correspondence Analysis

Correspondence analysis is a statistical visualization method for picturing the associations between the levels of a two-way contingency table. The name is a translation of the French *Analyses des Correspondances*, where the term *correspondance* denotes a "system of associations" between the elements of two sets.

In a two-way contingency table, the observed association of two traits is summarized by the cell frequencies, and a typical inferential aspect is the study of whether certain levels of one charactertistic are associated with some levels of another. Correspondence analysis is a geometric technique for displaying the rows and columns of a two-way contingency table as points in a low-dimensional space, such that the positions of the row and column points are consistent with their associations in the table. The goal is to have a global view of the data that is useful for interpretation.

To illustrate correspondence analysis, consider the multidimensional time series on the number of science doctorates conferred in the USA from 1960 to 1975 that is shown in Table 1 (Greenacre, 1984). Correspondence analysis of these data yields

Table 1: Science Doctorates in the USA, 1960-1975

Discipline/Year	1960	1965	1970	1971	1972	1973	1974	1975
Engineering	794	2073	3432	3495	3475	3338	3144	2959
Mathematics	291	685	1222	1236	1281	1222	1196	1149
Physics	530	1046	1655	1740	1635	1590	134	1293
Chemistry	1078	1444	2234	2204	2011	1849	1792	1762
Earth Sciences	253	375	511	550	580	577	570	556
Biology	1245	1963	3360	3633	3580	3636	3473	3498
Agriculture	414	576	803	900	855	853	830	904
Psychology	772	954	1888	2116	2262	2444	2587	2749
Sociology	162	239	504	583	638	599	645	680
Economics	341	538	826	791	863	907	833	867
Anthropology	69	82	217	240	260	324	381	385
Others	314	502	1079	1392	1500	1609	1531	1550

the graphical display shown in Figure 1. It has two sets of points, as indicated by the two types of point symbols. The points are *row* points for the rows of the data and *column* points for the columns. In Figure 1, there are row points for the disciplines, and column points for the years.

The distance between the row points is a measure of similarity between the row-frequency profiles - the anthropology degree and the engineering degree are far from

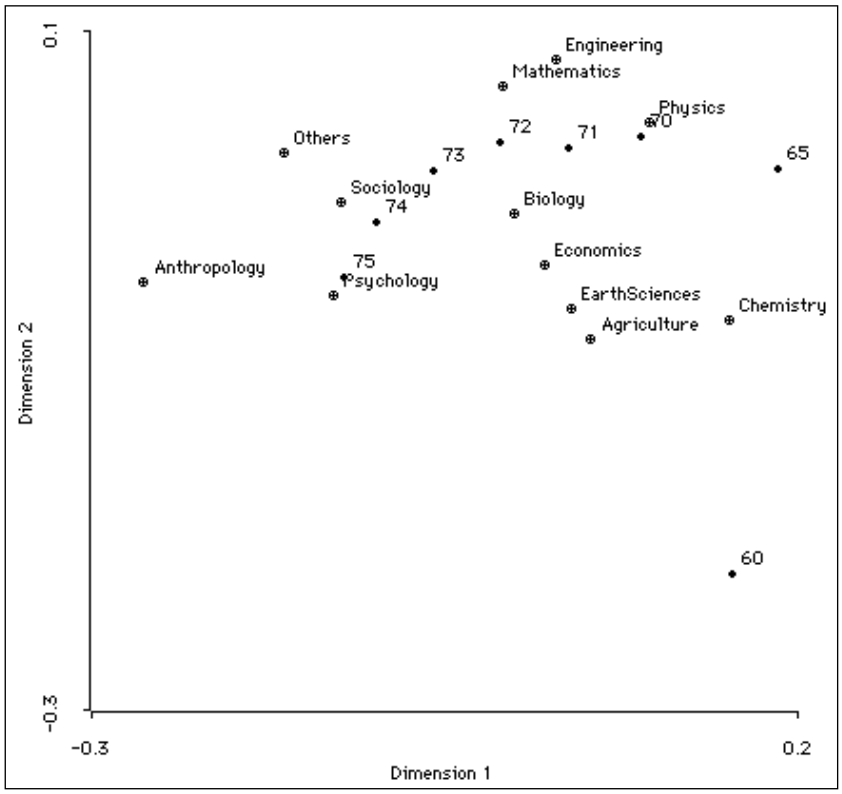

Figure 1: Correspondence Analysis of Doctorate Data

each other because their profiles are different, whereas the mathematics degree is near the engineering degree because their profiles are similar. Distances between the points representing years are interpreted in the same way – each year point represents the profile of that year across the various disciplines.

Note that the positions of two sets of points with respect to each other are not directly comparable and should be interpreted with caution. The interpretation given by Greenacre (1984) for this example is that each discipline point will lie in the neighborhood of the year in which the discipline's profile is prominent. Thus, there are relatively more agriculture, earth science and chemistry degrees in 1960, while the trend from 1965 to 1975 appears to be away from the physical sciences towards the social sciences. Furthermore, notice that points such as earth sciences and economics lie within the parabolic configuration of the years points; this implies that the profiles of these disciplines are higher than average in the early and later years. This example illustrates how a low-dimensional graphical representation of what is basically a deterministic trend supports a rich description of the data.

11.2 Using the Correspondence Analysis Procedure

ViSta's correspondence analysis procedure incorporates the traditional analysis methods of correspondence analysis with dynamic graphical procedures (Young, 1994). The subsequent sections provide a guide to the use of this procedure.

11.2.1 Input Datafile

The data suitable for correspondence analysis is a two-way contingency table, which is represented in ViSta by a multivariate (not a table) data object. The values of the observations specify the cell frequencies and hence must be nonnegative — negative or missing values cause an error. The data object can be created by typing data into the datasheet or by using the `data` function. The `data` function may be typed in the listener window or entered by an editor and saved as a datafile. The doctorates datafile, and its data function, is shown Figure 2. The datasheet resulting from this is shown in Figure 3. We use these data in this chapter.

The first argument of the `data` function must be a string that names the data object. In the example the data are named `"Doctorates"`. The `:title` keyword, which is optional, has a character string argument that specifies the title of the data. The `:variables` keyword, which is required, is used to specify the names of the variables, while the `:labels` keyword is used to specify the label for each observation. For a two-way contingency table, these keywords are used to specify the levels of each way of the table. Data values are specified using the `:data` key-

```
(data "Doctorates"
:title "Science Doctorates Conferred in the USA, 1960-1975"
:variables '("60" "65" "70" "71" "72" "73" "74" "75")
:labels '("Engineering" "Mathematics" "Physics" "Chemistry"
          "EarthSciences" "Biology" "Agriculture" "Psychology"
          "Sociology" "Economics" "Anthropology" "Others")
:data '(
 794 2073 3432 3495 3475 3338 3144 2959
 291  685 1222 1236 1281 1222 1196 1149
 530 1046 1655 1740 1635 1590 1334 1293
1078 1444 2234 2204 2011 1849 1792 1762
 253  375  511  550  580  577  570  556
1245 1963 3360 3633 3580 3636 3473 3498
 414  576  803  900  855  853  830  904
 772  954 1888 2116 2262 2444 2587 2749
 162  239  504  583  638  599  645  680
 341  538  826  791  863  907  833  867
  69   82  217  240  260  324  381  385
 314  502 1079 1392 1500 1609 1531 1550))
```

Figure 2: Doctorates datafile

8 Vars / 12 Obs	1960	1965	1970	1971	1972	1973	1974	1975
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
Engineering	794.00	2073.00	3432.00	3495.00	3475.00	3338.00	3144.00	2959.00
Mathematics	291.00	685.00	1222.00	1236.00	1281.00	1222.00	1196.00	1149.00
Physics	530.00	1046.00	1655.00	1740.00	1635.00	1590.00	1334.00	1293.00
Chemistry	1078.00	1444.00	2234.00	2204.00	2011.00	1849.00	1792.00	1762.00
EarthSciences	253.00	375.00	511.00	550.00	580.00	577.00	570.00	556.00
Biology	1245.00	1963.00	3360.00	3633.00	3580.00	3636.00	3473.00	3498.00
Agriculture	414.00	576.00	803.00	900.00	855.00	853.00	830.00	904.00
Psychology	772.00	954.00	1888.00	2116.00	2262.00	2444.00	2587.00	2749.00
Sociology	162.00	239.00	504.00	583.00	638.00	599.00	645.00	680.00
Economics	341.00	538.00	826.00	791.00	863.00	907.00	833.00	867.00
Anthropology	69.00	82.00	217.00	240.00	260.00	324.00	381.00	385.00
Others	314.00	502.00	1079.00	1392.00	1500.00	1609.00	1531.00	1550.00

Figure 3: Doctorates Datasheet

word, where each row corresponds to a label specified by the `:labels` keyword, and each column corresponds to a label specified by the `:variables` keyword. Note that the data are organized such that the columns of the data correspond to variables and the rows of the data correspond to observations. Since correspondence analysis treats the rows and columns in the same manner, the data could be transposed and analyzed with the same result. In this case, the names for the levels of the two classifications specified by `:variables` and `:labels` would be switched.

11.2.2 Analysis Options

You can perform a correspondence analysis by selecting the **Correspondence Analysis** item from the **Analyze** menu; by using the Coresp button on the toolbar, or by typing the `(correspondence-analysis)` function in the listener window. In any case, you will see the options dialog box shown in Figure 4 (when

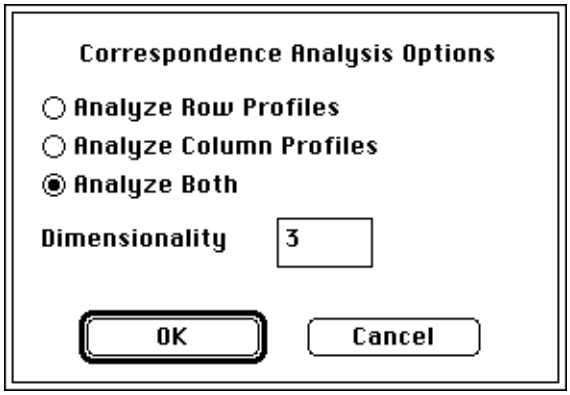

Figure 4: Options Dialog Box

typing, you only see the dialog when you use the `:dialog t` argument). The first set of options concern three different methods for computing the row and column points. The default is **Analyze Both,** as shown in the figure. More information on these methods is given below. The options are:

Analyze Row Profile: Row and column coordinates are computed from the row profiles. Row coordinates are weighted centroids of column coordinates.

Analyze Column Profile: Row and column coordinates are computed from column profiles. Column coordinates are weighted centroids of row coordinates.

Analyze Both: Row coordinates are computed from the row profiles, and column coordinates are computed from the column profiles.

The second option concerns the number of dimensions to use. The default value for this option is 3, as shown in the figure. The minimum is also 3. The maximum dimensionality for an (*n* x *m*) table is the minimum of *n - 1* and *m* - 1. Thus, the maximum dimensionality for the `Doctorates` data is 7.

As mentioned above, you can type the `(correspondence-analysis)` function in the listener window. The entire correspondence analysis method function with its keyword parameters and their default arguments is:

```
(correspondence-analysis
   :data current-data
   :title "Correspondence Analysis"
   :name (strcat "CRS-" (send current-data :name)
   :dialog nil
   :profile "both"
   :dimensions 3
   :precision 4)
```

To perform a correspondence analysis on the current data using the default arguments, the user types `(correspondence-analysis).` Otherwise, any of the following keyword arguments may be supplied:

`:data`	The argument to the `:data` keyword is the name of the data to be analyzed, without quotes. By default, the current data are analyzed.
`:title`	The argument of the `:title` keyword allows the user to provide an alternative title for the report. The title text is enclosed in quotes.
`:name`	The argument of the `:name` keyword allows the user to name the correspondence-analysis model object that is created. The default name is created from the name of the data object being

analyzed, preceded by the method's abbreviation `CRS.` The name must be enclosed in quotes.

`:dialog` If the argument to the `:dialog` keyword is `t`, the dialog box shown in Figure 4 is presented. Normally, when the user is typing directly in the listener window, the dialog box is not needed. Hence the default argument `nil` is appropriate.

`:profile` The `:profile` keyword is used to specify the method for computing row and column coordinates. The valid arguments are `"row", "column",` or `"both",` which are equivalent to the first set of options in the dialog box. The default argument is `"both"`.

`:dimensions` The desired dimensionality is specified by the `:dimensions` keyword followed by an integer argument in the range specified above. The default is 3.

`:precision` The `:precision` keyword controls the number of decimal places printed in the report (the default is 4).

For example, to perform a three-dimensional correspondence analysis on the `Doctorates` data using the default normalization method and precision of report, type:

```
(correspondence-analysis
   :data Doctorates
   :title "Correspondence Analysis of Science
Doctorates Conferred in USA")
```

Note that the keywords `:data, :title, :name, :dimensions,` and `:precision` are also the message selectors for the model object, once it has been created. For instance, typing the statement `(send CRS-Doctorates :precision)` returns the precision of report (which is 4 in this case), whereas typing `(send CRS-Doctorates :precision 6)` changes the precision to 6. You could also type `(send CRS-Doctorates :name "My-Model-Object")` to rename the model object `"My-Model-Object".`

11.2.3 Report

A convenient way of obtaining a report of the analysis is to select the **Report Model** item from the **Model** menu. The analyst is presented with the dialog box (Figure 5) which is used to control the type and amount of output reported. Briefly, the options refer to the following information: The contingency table is the table of relative frequencies, scaled to percentages; The row and column profiles are the row and column frequencies, scaled to percentages; Chi-Square statistics are defined for each cell of the table as the observed frequency minus the expected frequency under the hypothesis of row and column independence; the row and column coordinates

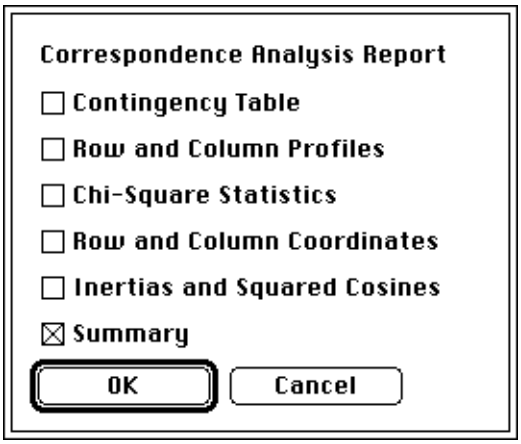

Figure 5: Report Dialog Box

are for the first *k* dimensions, where *k* is the dimensionality specified under the analysis options; the inertias and squared cosines are the partial contribution to principal inertias for row points and column points and the squared cosines for row points and column points; and the summary statistics are summaries for row points and column points including the quality of representation of points in the *k*-dimensional display, the masses, inertias, and the points that contribute most to inertia.

Keyboard users can type `(report-model)` to obtain the dialog box, or can send the model object the `:report` message followed by the `:level` keyword. This keyword accepts one of the following arguments: `"Brief"`, to obtain summary statistics; `"Normal"`, to also obtain the contingency table and the row and column profiles and coordinates; and `"Detailed"`, to obtain the complete report. For example `(send CRS-Doctorates :report :level "Detailed")` produces a complete report.

11.2.4 Statistical Visualization

The user can obtain a visualization of the analysis results by selecting the **Model** menu's **Visualize Model** item, or by typing `(visualize-model)`. In either case, a spreadplot of six linking windows will spread out across the screen, as shown in Figure 6. The individual plots that form a spreadplot can be thought of as "cells" in the spreadplot that can communicate with other cells via equations that define their relationships.

The **Spinplot** is a plot of the row and column points in the first three of the dimensions selected in the **Dimensions** window. The mouse can be in one of three modes: *Spinning, Brushing* and *Selecting*. The default mouse mode is *Spinning*. In this mode the cursor looks like a hand. Holding the mouse button down and moving

the cursor around the plot causes the plot to rotate. If you first hold the shift key down, then the plot will continue to rotate when you let up on the mouse button. You can also make the plot rotate by using the Pitch, Roll and Yaw buttons at the bottom. When you place the mouse mode in *Brushing*, the cursor looks like a tiny paint brush with a rectangle attached to it. Moving the brush across the plot selects

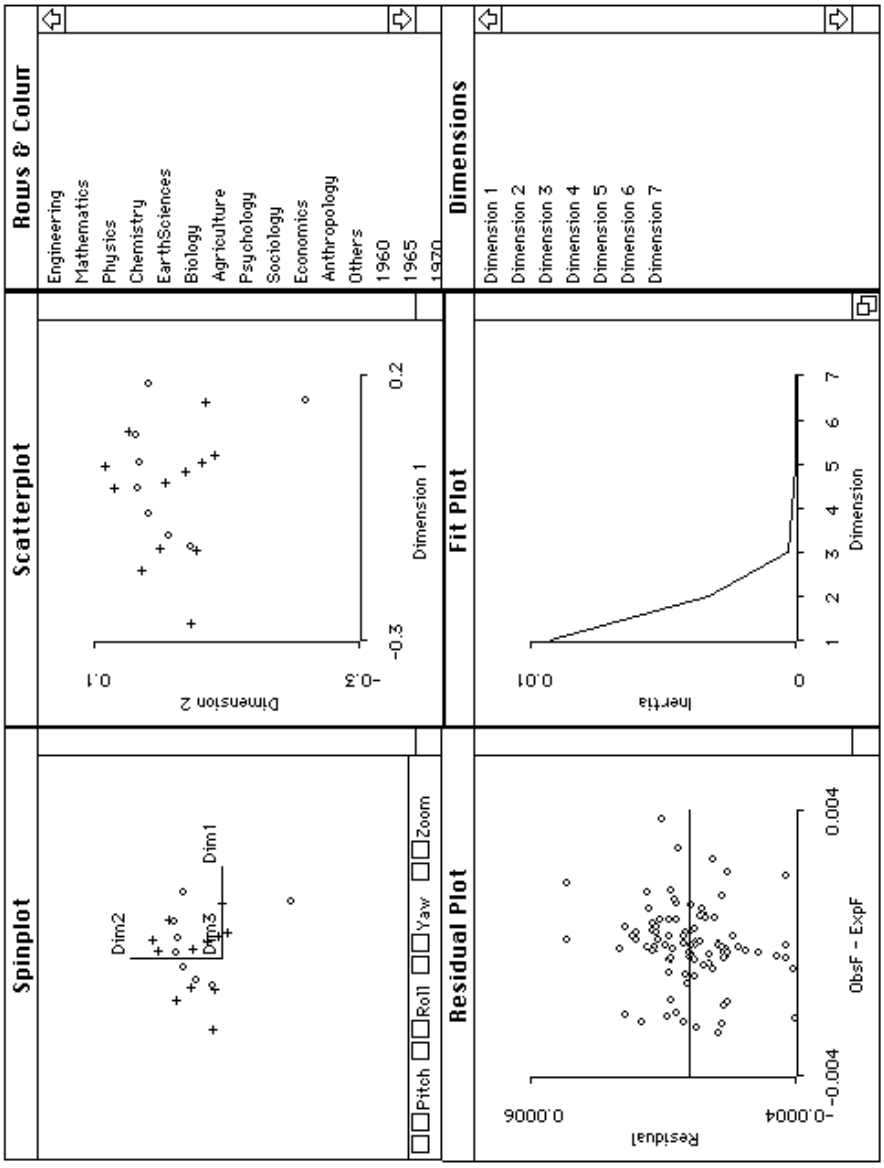

Figure 6: Correspondence Analysis Visualization

the points in the rectangle and highlights these points. When the mouse mode is changed to *Selecting* the cursor looks like an arrow and any points that are clicked on will be selected and highlighted. In addition, if the cursor is dragged across an area, any points inside the area are also selected and highlighted. Labels of selected points will be shown in whatever plots are linked to the spinplot and which have labeling turned on.

The **Scatterplot** plots the first two dimensions that are selected in the Dimensions window. This plot has two mouse modes — *Brushing* and *Selecting* — which are the same as those modes for the spinplot. The information in the scatterplot was displayed in Figure 1.

The **Rows & Columns** window, which contains the labels for the row and column points, is useful for locating or identifying points in the **Spinplot, Scatterplot** and **Residual Plot**. Since each cell frequency corresponds to the intersection of a row and a column in a contingency table, when more than two labels are selected or when the two labels belong to the same way of the table, the points in the **Residual Plot** will not respond to the selection and the previously highlighted residual point will return to the normal state.

The **Residual Plot** is a plot of the residuals versus the centered observed frequencies. Specifically, the residuals are obtained from the rank k least-squares approximation of the centered matrix of relative frequencies, where the value of k depends on the number of dimensions selected in the **Dimensions** window. Selecting a point in the **Residual Plot** causes the two points for the residual's row and column to be highlighted in the other three windows (only one residual may be selected at a time). The plot can be used for diagnostic checking as in a regression analysis.

The **Fit Plot** is a plot of the principal inertias against each dimension, showing the relative amount of fit for each dimension of the analysis. It serves the same purpose as the scree plot in principal component analysis.

The **Dimensions** window contains a list of dimensions. It serves as a control panel for the visualizations in the **Spinplot, Scatterplot** and **Residual Plot** windows. Selecting at least two dimensions will change the current display of the row and column points in the **Scatterplot** to that formed by the first two selected dimensions. For example, shift-clicking `Dimension 2, Dimension 3,` and `Dimension 5` produces a display of the points in the second and third dimensions. Selecting three or more dimensions will change the display in both the **Spinplot** and **Scatterplot**. In addition, selections in the **Dimensions** windows are tantamount to a re-specification of the dimensionality of analysis, which causes the residuals to be re-calculated and re-plotted in the **Residual Plot**.

11.3 Algorithm

Let X be an (n x m) matrix of observed frequencies of rank q such that the row and column sums are nonzero. Let 1 be a row vector of ones and I be an identity matrix, each of appropriate order. Denote a matrix-valued function that creates a diagonal matrix from a vector by diag `()`. Define

i. $s = 1'X1$ as the sum of all elements in X;

ii. $P = \frac{1}{s}X$ as the matrix of relative frequencies (the correspondence matrix);

iii. $r = P1$ as the vector of row marginal proportions (row masses);

iv. $c = P'1$ as the vector of column marginal proportions (column masses);

v. $D_r = diag(r)$ a diagonal matrix of row masses; and

vi. $D_c = diag(c)$ a diagonal matrix of column masses.

The generalized singular value decomposition (abbreviated SVD) of P provides the required solution to the point coordinates of correspondence analysis:

$$P = AD_uB',$$

where

i. A is an (n x q) matrix whose columns are the left generalized singular vectors;

ii. D_u is a (q x q) diagonal matrix of generalized singular values;

iii. B is an (m x q) matrix whose columns are the right generalized singular vectors; and where

iv. $A'D_r^{-1}A = B'D_c^{-1}B = I$.

There is a trivial part of the generalized SVD of P consisting of a singular value of 1 and associated left and right singular vectors, which is discarded before any results are displayed. The remaining left and right singular vectors define the orthogonal principal axes of the column and row points, respectively. In practice, the generalized SVD is computed indirectly by performing an ordinary SVD, where the ordinary SVD of any matrix Q is given by

$$Q = UD_\alpha V',$$

under the constraint $U'U = V'V = I$. Thus, to compute the generalized SVD of P, we perform the following steps:

i. Let $Q = D_r^{-1/2} P D_c^{-1/2}$.

ii. Obtain the ordinary SVD of **Q**, giving $Q = UD_\alpha V'$.

iii. Let $A = D_r^{1/2} U$, $B = D_c^{1/2} V$, and $D_u = D_\alpha$.

iv. Then $P = AD_u B'$ is the required generalized SVD.

v. The row coordinates F and column coordinates G are then computed according to the appropriate selection of the formulas in Table 2

Table 2: Formulas for Coordinates

Dialog Box Choice	Keyboard Argument	Formula for Row Coordinates	Formula for Column Coordinates
Analyze Row Profile	"Row"	$F = D_r^{-1} A D_u$	$G = D_c^{-1} B$
Analyze Column Profile	"Column"	$F = D_r^{-1} A$	$G = D_c^{-1} B D_u$
Analyze Both (default)	"Both"	$F = D_r^{-1} A D_u$	$G = D_c^{-1} B D_u$

For the various options given in Table 2, the following formulas apply:

• **Analyze Row Profile:** In this case, the row coordinates and column coordinates are a decomposition of $D_r^{-1} P D_c^{-1}$ into:

$$D_r^{-1} P D_c^{-1} = (D_r^{-1} P)(D_c^{-1} B) .$$

The row coordinates are weighted centroids of the column coordinates:

$$F = D_r^{-1} A D_u = (D_r^{-1} P)\,(D_c^{-1} B) .$$

The inertias D_u^2 refer only to the row points.

• **Analyze Column Profile:** In this case the row coordinates and column coordinates are a decomposition of $D_c^{-1} P D_r^{-1}$ into:

$$D_c^{-1} P D_r^{-1} = (D_c^{-1} B D_u)(D_r^{-1} A)' .$$

The column coordinates are weighted centroids of the row coordinates:

$$G = D_c^{-1} B D_u = (D_c^{-1} P')\,(D_r^{-1} A) .$$

The inertias D_u^2 refer only to the column points.

• **Analyze Both:** In this case the row and column coordinates are separate decompositions according to the following formulas:

$$F = D_r^{-1} A D_u = (D_r^{-1} P)\,(D_c^{-1/2})\,(D_c^{-1/2} B)\,, \text{ and}$$
$$G = D_c^{-1} B D_u = (D_c^{-1} P')\,(D_r^{-1/2})\,(D_r^{-1/2})\,(D_r^{-1/2} A)\,.$$

Neither the row nor column coordinates are weighted centroids of the other set of coordinates. The inertias D_u^2 refer to both sets of points, since:

$$D_u^2 = F' D_r F = G' D_c G\,.$$

Since $D_r^{-1} P$ is the matrix of row profiles, the default row coordinates can be interpreted as scaled row profiles (the scaling factor being $D_c^{-1/2}$) that are orthogonally rotated with $D_c^{-1/2} B$ to a principal axis orientation. Similarly, the default column coordinates are column profiles $D_c^{-1} P'$ scaled by the factor $D_r^{-1/2}$ and orthogonally rotated with $D_r^{-1/2} A$ to a principal axis orientation.

Plotted Values: Once the row and column coordinates have been calculated by the formulas given above, they are plotted in the spinplot and scatterplot. The spinplot shows the first 3 dimensions, whereas the scatterplot shows the first 2. The squared singular values (i.e., the square of the diagonal entries of the D_u matrix) are plotted versus each dimension index in the fit plot, to produce a plot of the goodness of fit. The residual plot displays residuals versus the centered frequencies. The centered data are calculated by the formula $P - rc'$. The residuals formula subtracts the expected relative frequencies from these centered frequencies, by the formula:

$$(P - rc') - A^k D_u^k B^{k\prime}$$

The superscripts *k* in this equation refer to the fact that only k of the dimensions are involved in the calculation. The specific columns of A^k and B^k and the specific diagonal elements of D_u^k that are involved correspond to the specific dimensions selected in the **Dimensions** window, which may not be the first *k* dimensions.

Several summary statistics which pertain to the default coordinates are available to aid the interpretation of the resulting graphical display. These statistics, which can be printed by the **Report** option, are summarized in Table 3. In these formulas, *T* denotes the total inertia given by the sum of squared (nontrivial) generalized singu-

lar values. The partial contribution of a point to the inertia of the *i*th principal axis (dimension) is a measure of the influence of the point on the final orientation of the principal axis. Normally, the principal axis tends more towards the higher mass points. The quality of representation of any point in a *k* dimensional display is given by the sum of its squared cosines over the *k* dimensions. The squared cosine of a point in the *j*th dimension is the contribution of the *j*th principal axis to the point's inertia. Usually, high contribution of a point to the inertia of an axis implies a high relative contribution of the axis to the inertia of the point, but not conversely.

Table 3: Formulas for Summary Statistics

Summary statistic	Formula
Row mass	r
Column mass	c
Row partial contribution to inertia	$D_c^{-1} sq(A)$
Column partial contribution to inertia	$D_c^{-1} sq(B)$
Row inertia	$(1/T)\ D_r^{-1} sq(AD_u)\ 1$
Column inertia	$(1/T)\ D_c^{-1} sq(BD_u)\ 1$
Row squared cosines	$\text{diag}\ [sq(AD_u)\quad 1]^{-1} sq(AD_u)$
Column squared cosines	$\text{diag}\ [sq(BD_u)\quad 1]^{-1} sq(BD_u)$

11.4 Statistical Revision

Statistical revision permits the user to move row or column points to new locations which may be more "interpretable", but which no longer satisfy all of the characteristics of the correspondence analysis equations. When the user moves a point, the software responds by adjusting the positions of the other points so that they approximate the correspondence analysis equations as well as possible, in a specific sense. In particular, when "row" points are moved by the user, the software calculates new positions for the "column" points, and conversely.

The calculations of the new positions of the "other" set of points is done so that the basic relationship $P = AD_uB'$ is maintained. This is done by noting that:

$$P = AD_uB' = D_rFG'D_c,$$

when the normalization is either "row" or "column". However, the relationship specified by the equation does not hold when the normalization is "both", which is why point-moving is not possible in that case.

If we move one or more "row" points we change F to, say, $\hat{F}$, and we must calculate a new $\hat{G}$ such that

$$P = D_r \hat{F} \hat{G}' D_c .$$

We note that

$$\hat{F}\hat{G}' = D_r^{-1} P D_c^{-1} ,$$

and we solve for $\hat{G}$ by the equation

$$\hat{G}' = [(\hat{F}'\hat{F})^{-1}\hat{F}']D_r^{-1} P D_c^{-1} .$$

While the basic relation $P = D_r \hat{F} \hat{G}' D_c$ is maintained, the orthogonality constraint of correspondence analysis is violated because the equations $\hat{F}' D_r \hat{F}$ and $\hat{G}' D_c \hat{G}$ are no longer diagonal. In fact, we even see that their equality is lost. That is:

$$\hat{F}' D_r \hat{F} \neq \hat{G}' D_c \hat{G} \ !$$

Thus, when a row point is moved to a new location in the scatterplot window, the new column coordinates $\hat{G}$ are calculated. Then, the new column coordinates are displayed in both the scatterplot and spinplot (the new location of the moved row point is also displayed in the spinplot). The residuals plot is then updated using the new values in $\hat{F}$ and $\hat{G}$.

To obtain an approximate measure of the goodness of fit after point moving, we solve for the new D_u matrix by the equation

$$D_u = (A'A)^{-1} A' D_r \hat{F} ,$$

and then plot the squared diagonal entries against each dimension as a dashed line in the fit plot window. Note that since the orthogonality constraint has been violated, the squared diagonal values will overestimate the true inertias.

Finally, note that the row points are plotted in their principal axes, while the column points are plotted in their standard axes. A consequence of this asymmetry is that the row points will often form a small cluster around the origin, making it difficult for you to move a point. In this case, you can use the **Hide Column Points** menu item in the **Scatterplot** menu to mask column points and rescale the row points.