Chapter 430

Correspondence Analysis

Introduction

Correspondence analysis (CA) is a technique for graphically displaying a two-way table by calculating *coordinates* representing its rows and columns. These coordinates are analogous to factors in a principal components analysis (used for continuous data), except that they partition the Chi-square value used in testing independence instead of the total variance.

For those of you who are new to CA, we suggest that you obtain Greenacre (1993). This is an excellent introduction to the subject, is very readable, and is suitable for self-study. If you want to understand the technique in detail, you should obtain this (paperback) book.

Discussion

We will explain CA using the following example. Suppose an aptitude survey consisting of eight yes or no questions is given to a group of tenth graders. The instructions on the survey allow the students to answer only those questions that they want to. The results of the survey are tabulated as follows.

Aptitude Survey Results – Counts

Question	Yes	No	Total
Q1	155	938	1093
Q2	19	63	82
Q3	395	542	937
Q4	61	64	125
Q5	1336	876	2212
Q6	22	14	36
Q7	864	354	1218
Q8	920	185	1105
Total	3772	3036	6808

Take a few moments to study this table and see what you can discover. The most obvious pattern is that many of the students did not answer all the questions. This makes response patterns between rows difficult to analyze.

To solve this problem of differential response rates, we create a table of row percents (or *row profiles* as they are called in CA).

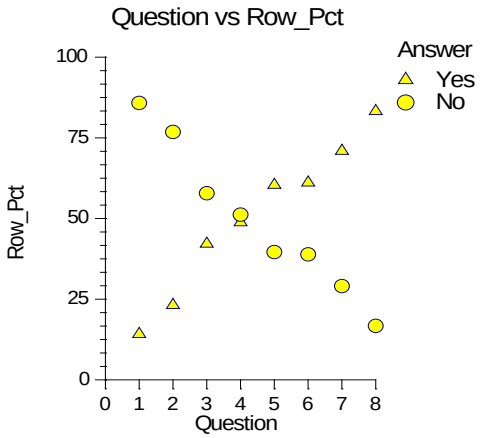
Aptitude Survey Results – Row Profiles

Question	Yes	No	Total
Q1	14.18	85.82	100.00
Q2	23.17	76.83	100.00
Q3	42.16	57.84	100.00
Q4	48.80	51.20	100.00
Q5	60.40	39.60	100.00
Q6	61.11	38.89	100.00
Q7	70.94	29.06	100.00
Q8	83.26	16.74	100.00
Total	55.41	44.59	100.00

This table allows us to see the underlying patterns within the table. We note that only 14% answered yes to question one while 83% answered yes to question eight.

Although we can inspect this table directly, a picture of the data will allow us to find patterns much more quickly. This is done in the following scatter plot. The plot shows the row profiles with the questions on the horizontal axis and the row percents on the vertical axis. Notice that there are two possible responses (yes or no) and two corresponding plotting symbols.

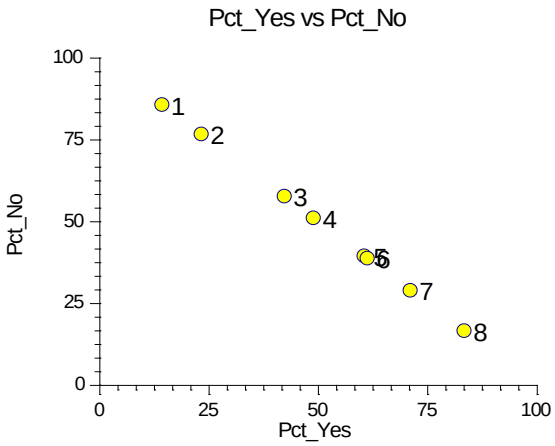
Aptitude Survey Results – Row Percents versus Questions



We notice a steady gain in the percent of students answering yes as we move from question one to question eight. Also, we can see the obvious relationship between the percent answering yes and the percent answering no. In fact, if you think about it for a moment you will realize that we really only need to plot the yes's or the no's, but not both since they both relate the same information.

Another way of plotting this data is to plot the percentage of each possible answer on a different axis. Since, in this example, we have two possible answers, we plot the yes percentage on one axis and the no percentage on the other axis.

Aptitude Survey Results – Scatter Plots



Spend some time analyzing this plot. Can you see the connection between the last plot and this plot? In the previous plot, each possible answer was a horizontal set of points. In this plot, each answer is an axis. Hence, if our survey was made up of multiple-choice questions each with three possible answers, this plot would need to have been three dimensional.

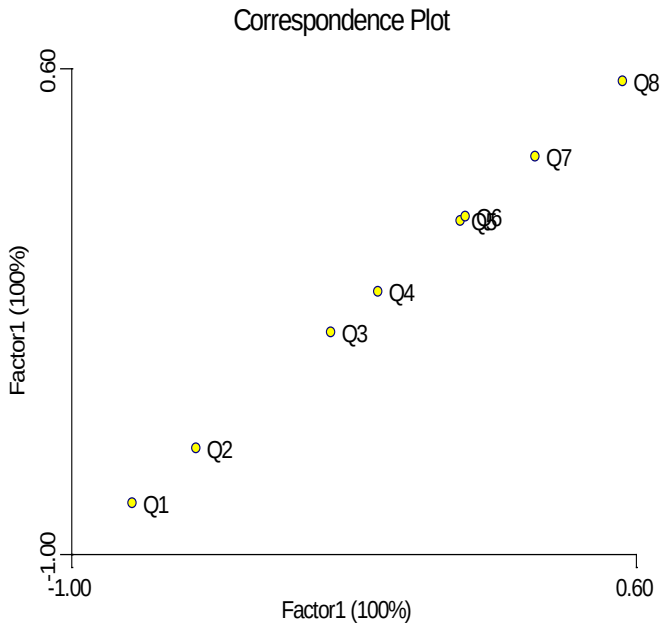
This plot is an example of a *correspondence map*, the primary output of CA. It is important to understand the features of this plot. Each axis of the plot represents a column and each point represents a row of the original table. If you were to draw a bar chart for this data, you would create bars representing the distances from each point to each axis. Since there are eight points and two axes, the bar chart would have sixteen bars.

Notice also how you interpret the plot. Each point represents a specific yes or no combination. For example, question eight had about 83% answering yes and 17% answering no. This high proportion of yes respondents positions the point very close to the Pct_Yes (horizontal) axis. Conversely, question one had a large percentage of no answers and is very near the Pct_No axis. We see that points relatively close to the end of a particular axis have a high percentage value on that axis.

Also notice that points that are close to each other have very similar patterns of yes or no answers. Look at the row profiles (percents) for questions five and six (two points that are almost in the same position). Notice that they differ by only one percentage point.

The following figure shows the CA plot of this data generated by the program. Although the orientation of the plot is different, the distances between the points are the same.

Aptitude Survey Results – CA Plot



This plot and the last plot appear the same because, in this example, the CA plot reproduces the plot of the row percents. This occurs because there were only two possible answers to each question. If the questions in our survey had been multiple choice so that there were four or five answers, we would have needed to create a four- or five-dimensional plot to reproduce the Row Profile Plot.

To summarize, then, a CA plot is a plot of the row profiles (percentages) constructed so that each column category becomes a different dimension. Since it is only possible to view two dimensions at a time, we must project this high dimensional space onto a two-dimensional subspace. This projection is constructed so as to maintain as much of the original information (or variation) as possible.

Technical Details

We will now present an outline of the computational methods used to perform the analysis. We will use standard matrix terminology to present the steps.

1. Read in the *n* (rows) by *m* (columns) data matrix, ***K***. Note that the elements of ***K*** must be non-negative and that none of the row or column totals is zero.
2. Compute the proportion matrix, **P**, by dividing the elements of **K** by the total of all numbers in **K**. Mathematically, we write

$$\mathbf{P} = \left\{ p_{ij} \right\} = \left\{ k_{ij} / k_{\cdot\cdot} \right\}$$

3. Compute the totals of the rows of **P** and the columns of **P**, putting the results in the vectors **r** and **c**. Using standard matrix notation, we write

$$\mathbf{r} = \mathbf{P1}$$

$$\mathbf{c} = \mathbf{P'1}$$

where **1** is an appropriately dimensioned vector of ones.

4. Change the square roots of the vectors **r** and **c** into diagonal matrices and take the inverse of the resulting square matrices.

$$\mathbf{D}_r = \left[diag(\mathbf{r})\right]^{-1/2}$$

$$\mathbf{D}_c = \left[diag(\mathbf{c})\right]^{-1/2}$$

5. Compute the scaled matrix, **A**.

$$\mathbf{A} = \mathbf{D}_r\mathbf{P}\mathbf{D}_c$$

6. Compute the Singular Value Decomposition (SVD) of **A**.

$$\langle \mathbf{B}, \mathbf{W}, \mathbf{C} \rangle = SVD(\mathbf{A})$$

7. Compute the coordinate matrices, **F** and **G**, as follows:

$$\mathbf{F} = \mathbf{D}_r\mathbf{B}\mathbf{W}$$

$$\mathbf{G} = \mathbf{D}_c\mathbf{C}\mathbf{W}'$$

8. Compute the eigenvalues, **V**.

$$\mathbf{V} = \mathbf{W}\mathbf{W}'$$

9. Compute the row distances, d_i, and the column distances, d_j.

$$d_i = \sum_j \left(\frac{1}{p_{.j}}\right)\left(\frac{p_{ij}}{p_{i.}} - p_{.j}\right)^2$$

$$d_j = \sum_i \left(\frac{1}{p_{i.}}\right)\left(\frac{p_{ij}}{p_{.j}} - p_{i.}\right)^2$$

10. Note that the weights, w_i and w_j, come from the vectors **r** and **c** that were formed in step 3.

$$w_i = \{r_i\}$$

$$w_j = \{c_j\}$$

11. Compute the reported statistics as follows:

Statistic	Formula
Mass	w_i
Inertia	$\frac{w_i d_i^2}{\sum_k w_k^2 d_k^2}$
Distance	d_i^2
Row Factor	f_{ij}
Column Factor	g_{ij}
Row COR	$\frac{f_{ij}^2}{d_i^2}$
Column COR	$\frac{g_{ij}^2}{d_j^2}$

Statistic	Formula
Row CTR	$\frac{w_i f_{ij}^2}{v_i}$
Column CTR	$\frac{w_j g_{ij}^2}{v_j}$
Angle	$ArcCos\left(\sqrt{COR_{ij}}\right)$

Data Structure

We will use a set of data from Greenacre (1993) in the tutorial that follows. The table below shows the results of a survey relating the smoking habits of the employees of a fictitious company to their position within the company. These data are contained in the Corres1 dataset.

The entries in the table are the counts of the number of employees falling into each cell.

Corres1 dataset

None	Light	Medium	Heavy	Staff
4	2	3	2	(SM) Senior Managers
4	3	7	4	(JM) Junior Managers
25	10	12	4	(SE) Senior Employees
18	24	33	13	(JE) Junior Employees
10	6	7	2	(SE) Secretaries

Example 1 – Correspondence Analysis

This section presents an example of how to run an analysis of the data presented in the table above. These data are contained in the Corres1 dataset.

Setup

To run this example, complete the following steps:

1 Open the Corres1 example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Corres1** and click **OK**.

2 Specify the Correspondence Analysis procedure options

- Find and open the **Correspondence Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables Tab	
Table Variables	**None-Heavy**
Row Label Variable	**Staff**
Plots Tab	
Correspondence Plots	**Both**

Raw Data Section

Raw Data Section

Staff	None	Light	Medium	Heavy	Total
SM	4	2	3	2	11
JM	4	3	7	4	18
SE	25	10	12	4	51
JE	18	24	33	13	88
SC	10	6	7	2	25
Total	61	45	62	25	193

This report displays the raw data. It documents the data that were used by the procedure so that you can check for data-entry errors.

Row Profiles Section

Row Profiles Section

Staff	None	Light	Medium	Heavy	Total
SM	36.36	18.18	27.27	18.18	100.00
JM	22.22	16.67	38.89	22.22	100.00
SE	49.02	19.61	23.53	7.84	100.00
JE	20.45	27.27	37.50	14.77	100.00
SC	40.00	24.00	28.00	8.00	100.00
Total	31.61	23.32	32.12	12.95	100.00

This report shows the row profiles (percentages). These are the values the will be plotted on the row oriented plot. Note that since there are five rows, these data would require five dimensions to be plotted in the standard fashion. CA investigates the differences between each individual row profile and the average row profile (the row labeled “Total”).

Column Profiles Section

Column Profiles Section

Staff	None	Light	Medium	Heavy	Total
SM	6.56	4.44	4.84	8.00	5.70
JM	6.56	6.67	11.29	16.00	9.33
SE	40.98	22.22	19.35	16.00	26.42
JE	29.51	53.33	53.23	52.00	45.60
SC	16.39	13.33	11.29	8.00	12.95
Total	100.00	100.00	100.00	100.00	100.00

This report shows the column profiles (percentages). These are the values the will be plotted in a column oriented CA plot. Note that since there are four columns, these data would require four dimensions to be plotted in the standard fashion. CA investigates the differences between each individual column profile and the average column profile (the column labeled “Total”).

Eigenvalue Section

Eigenvalue Section

Factor No.	Eigenvalue	Individual Percent	Cumulative Percent	Bar Chart
1	0.074759	87.76	87.76	||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2	0.010017	11.76	99.51	||||||||||
3	0.000414	0.49	100.00	|
Total	0.085190			

Since CA projects the row (or column) profiles onto a two-dimensional subspace, a critical issue is how well this projection works. The eigenvalues gives us important information regarding this. The Cumulative Percent column tells us how much of the total information is reproduced by each number of dimensions.

In this example, the CA plot using the first two factors accounts for 99.5% of the variation. In other words, the dimension reduction is only costing us a 0.5% loss in information. We can be confident that the patterns we see in the CA plot represent the patterns that we would see if we could peer into n-dimensional space.

Factor No.

This is the number of the factor (coordinate or dimension) that is reported about on this row of the report.

Eigenvalue

This is the eigenvalue associated with this dimension. It gives a relative size (importance) of this dimension.

Individual and Cumulative Percents

The first column gives the percentage of the total of the eigenvalues accounted for by this dimension. The second column is the cumulative total of the percentage.

In ideal situations, the first two dimensions will account for over 90% of the variation. If the cumulative percentage is less than 50%, CA is not appropriate.

Bar Chart

This is a rough bar plot of the eigenvalues. It enables you to quickly note the relative size of each eigenvalue.

Plot Detail Section

Plot Detail Section for Rows

				Axis1			Axis2		
Name	Quality	Mass	Inertia	Factor	COR	CTR	Factor	COR	CTR
1 SM	0.893	0.057	0.031	0.066	0.092	0.003	-0.194	0.800	0.214
2 JM	0.991	0.093	0.139	-0.259	0.526	0.084	-0.243	0.465	0.551
3 SE	1.000	0.264	0.450	0.381	0.999	0.512	-0.011	0.001	0.003
4 JE	1.000	0.456	0.308	-0.233	0.942	0.331	0.058	0.058	0.152
5 SC	0.999	0.130	0.071	0.201	0.865	0.070	0.079	0.133	0.081

This report provides the information you need to interpret a correspondence plot correctly. A similar report is generated for each CA plot.

This report is used as follows. First, for each axis, look down the CTR column to determine which profiles contribute highly to the axis. This is useful in finding possible interpretations of the axis. Next, look across the COR values to identify which of the axes represent the profile well. Finally, the Quality column shows how well the profile is reproduced in the subspace defined by the two axes.

Axis1, Axis 2

These are the two axes (coordinates or dimensions) that are reported on here.

Name

The name of the dimension (profile) being reported about on this line of the report.

Quality

This is the sum of the two COR values. It is the proportion of the variation in this profile that is reproduced by the two factors being reported on here.

In this example, we see that all of the profiles are above 89%. In fact, all but the SM profile are over 99%. We can feel confident that the points shown in this plot are not distorted by the projection process.

Mass

The mass (or weight) is the proportion of the whole table that is in the category represented by this row. It is the ratio of the row count to the total table count. You will find the masses also reported as percentages in the last column of the Column Profile Section.

Inertia

The inertia of the whole table is a function of the Chi-square statistic, χ^2. If

$$\chi^2 = \sum_{all\ i,j} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where O_{ij} is the count of row *i* and column *j* of the table, E_{ij} is the value expected under the assumption of row-by-column independence, and *N* is the total table count, then the total inertia of the table is given by

$$Total\ Inertia = \frac{\chi^2}{N}$$

The inertia value reported is the proportion of the total inertia that is due to this profile.

Another way to interpret the inertia is that it is the weighted average of the Chi-square distances between the row profiles and their average profile.

Factor

The coordinate of the profile along this axis. This is the value of the row profile projected onto the line defined by this axis. It is the value that is plotted.

COR

This is the correlation between this profile and the axis. It allows you to determine which of the axes represent the profile well. This is the proportion of the variance in a profile explained by the axis.

This is the contribution of this axis to the inertia of this profile. The formula used to compute this was given earlier.

CTR

The contribution of this profile to the inertia of this axis. This is the proportion of variance in the axis accounted for by this profile. The formula used to compute this was given earlier.

Principal Coordinate Section

Principal Coordinate Section for Rows - Axis 1

Name	Mass	Inertia	Distance	Factor	COR	CTR	Angle	Eigenvalue
1 SM	0.057	0.031	0.047	0.066	0.092	0.003	72.3	0.000247
2 JM	0.093	0.139	0.127	-0.259	0.526	0.084	43.5	0.006254
3 SE	0.264	0.450	0.145	0.381	0.999	0.512	1.8	0.038277
4 JE	0.456	0.308	0.058	-0.233	0.942	0.331	13.9	0.024743
5 SC	0.130	0.071	0.047	0.201	0.865	0.070	21.5	0.005238

Principal Coordinate Section for Rows - Axis 2

Name	Mass	Inertia	Distance	Factor	COR	CTR	Angle	Eigenvalue
1 SM	0.057	0.031	0.047	-0.194	0.800	0.214	26.5	0.002139
2 JM	0.093	0.139	0.127	-0.243	0.465	0.551	47.0	0.005521
3 SE	0.264	0.450	0.145	-0.011	0.001	0.003	88.4	0.000030
4 JE	0.456	0.308	0.058	0.058	0.058	0.152	76.1	0.001520
5 SC	0.130	0.071	0.047	0.079	0.133	0.081	68.6	0.000807

This report provides all information about each axis (dimension or factor). Much of the information is duplicated in the Plot Detail Section (see above) and will not be redefined here. We will present only those items that were not defined in the last report.

Distance

This is the weighted distance of the row profile from the average row profile. It is provided for completeness.

Angle

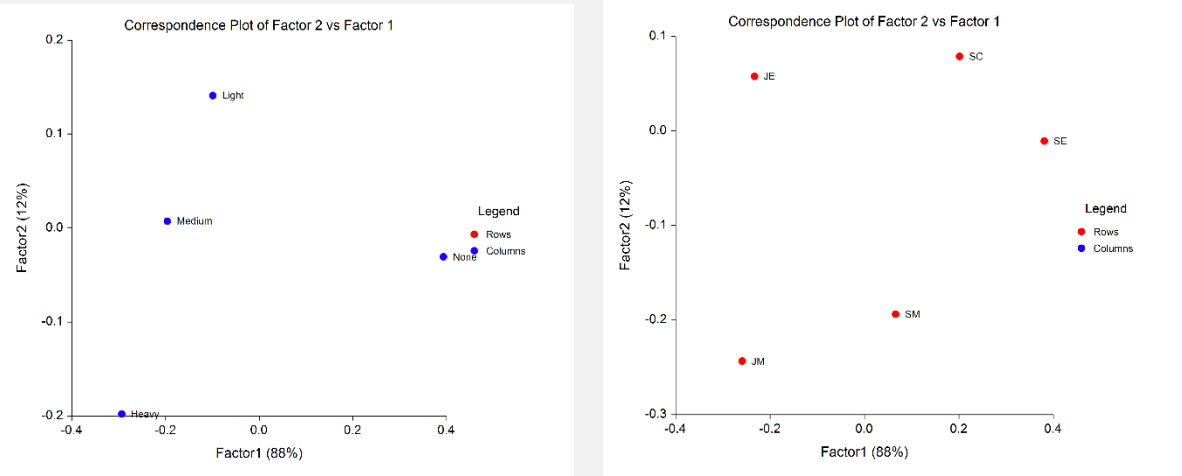
This is the angle between the axis and the profile.

Eigenvalue

If we partition the eigenvalue associated with this axis into separate parts for each profile, this is the absolution amount of the eigenvalue that is due to this profile. This value is provided more for completeness than interpretation.

Correspondence Plots

Correspondence Plots

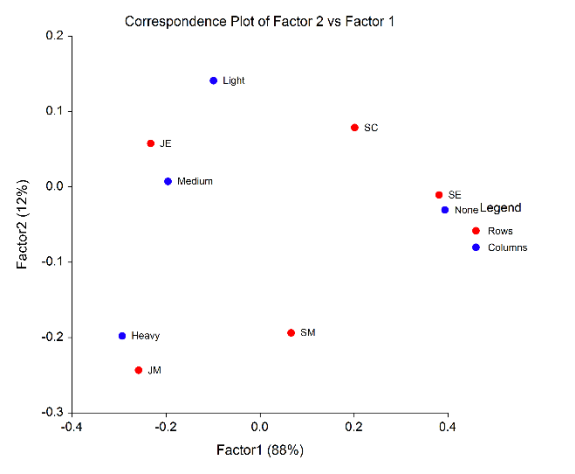


This plot is the main objective of a CA. The plot on the left shows the column profiles and the plot on the right shows the row profiles. It is important to remember that each point represents a profile projected onto the plane defined by the two axes.

Lets begin by discussing the left plot, the one presenting the four column profiles. These profiles represented the proportions belonging to each staff category. We can see from this plot that the first factor seems to separate those who smoke from those who do not. The second factor seems to separate the three types of smokers: light, medium, and heavy.

The right plot presents the five row profiles. The first axis appears to separate junior people from senior people. The second axis seems to separate managers (near the bottom) from non-managers (near the top).

Note that the distances between points on these plots are Chi-square distances between the profiles those of two points. Hence, the closer two points appear, the closer their profile patterns are to each other.



Finally, we come to the most popular CA plot in which we overlay the two plots shown above onto one plot. Extreme caution must be used when interpreting this plot. The critical point to remember is that this is a combination of two independent plots. Distances between the row profile points and column profile points are not defined. Hence, the distance between the categories SE and None (although this appear near each other on the plot) is not defined. Here's why: The point SE is a projection of the SE profile from the four dimensional space to the two- dimensional subspace defined by our axes. The point None is a projection of the None profile from the five-dimensional space to the two-dimensional subspace defined by the two axes. The original spaces are different. They represent different things. In the case of the row profiles, each of the four axes represents a smoking pattern (none, light, medium, and heavy). In the case of column profiles, each of the five axes represents a staff category. The point is that the meaning of the original spaces was completely different. Their axes have completely different definitions. This is the classical apples and oranges situation. As we view a subspace from each overlaid onto one plot, what is the connection between these two subspaces?

To understand why analysts like to plot the row and column profiles on one graph, we will create a new version of our row profile plot with the addition of supplementary rows. The following table presents the data being analyzed plus additional "supplementary" rows.

Corres1 dataset with supplementary rows (Corres1b)

None	Light	Medium	Heavy	Staff	RowType
4	2	3	2	(SM) Senior Managers	**1**
4	3	7	4	(JM) Junior Managers	**1**
25	10	12	4	(SE) Senior Employees	**1**
18	24	33	13	(JE) Junior Employees	**1**
10	6	7	2	(SE) Secretaries	**1**
100	**0**	**0**	**0**	**N**	**2**
0	**100**	**0**	**0**	**L**	**2**
0	**0**	**100**	**0**	**M**	**2**
0	**0**	**0**	**100**	**H**	**2**

Note the addition of the RowType variable with its 1's and 2's. This is used to indicate which rows contain data and which rows are supplementary.

Now, if you consider the four rows that have been added to the bottom, you will see that each has a value of 100 in one column and 0 in all the rest. Hence each row represents a particular type of smoker. N represents the None group, L represents the Light group, M represents the Medium group, and H represents the Heavy group. If we could peer into four-dimensional space, we would see that each of the points fall on the corresponding axis. That is, the four supplementary rows represent the four axes.

Incidentally, we could have entered “1” in each position instead of “100” since the program rescales these values.

Now let’s take a look at the row profile CA plot with these supplementary rows.

Row Profile Plot with Supplemental Axis Points

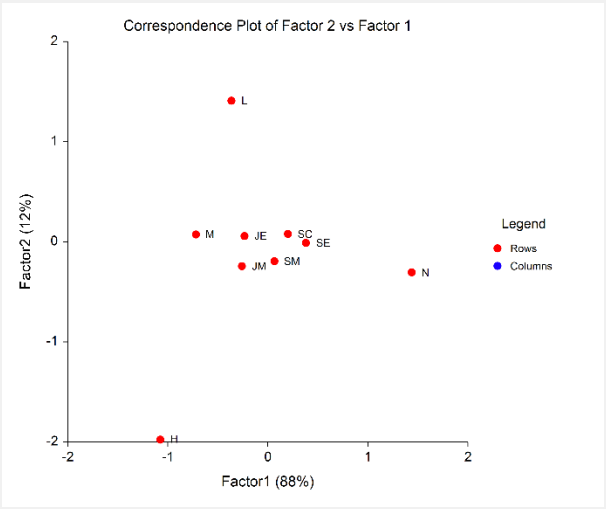
Studying this plot, we note that supplemental points (the N, L, M, and H) seem to surround the regular points. This is because each of these points defines an edge point, or vertex, to the four-dimensional space we are considering.

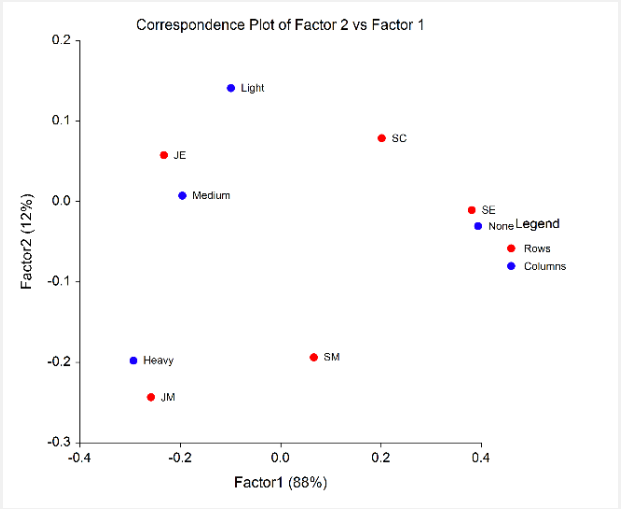
Now, a point near one of these supplemental points is one whose profile is similar to that point. For example, it appears that both JE and JM are near M (Medium). If you look at the Row Profile Section, you will see that they had 37.5% and 38.9% in the medium category, respectively. These are much larger than the other groups.

We see that points closer to one of our supplementary points tend to have higher than normal values for that category. However, since none of the row profiles had more than 50% in any one category, none of the profiles is right next to the vertex point (as defined by the supplementary row).

We are now ready to see why we can legitimately overlay the row profile and column profile plots. We will redisplay the last two plots side by side.

Row Profile Plot with Supplemental Axis Points and Overlaid CA Plot





The plot on the left is the regular row profile CA plot with supplementary points. Compare the relative positions of the L, M, H, and N with those of Light, Medium, Heavy, and None in the overlay plot on the right. You can see that these points maintain their relative position. They just shrink inward toward the center.

That this is the case can be shown mathematically. The message is clear. When the two plots are overlaid, the points from one space (row or column) represent the vertices of the other space, except they have been shrunken in towards the center. Hence, as we analyze the right plot from the row profile context, we must mentally move the column profile points out from the middle. That is, we must realize that each point represents the direction of the end point of that axis, but the point is not at the actual position of the end point.

Personally, I believe that interpretation is easier if you always construct two plots: one for the row profiles and another for the column profiles. The axes of each space are shown as supplementary rows (or columns). This avoids the temptation to see points from two spaces as being “near” each other.

This concludes our discussion of correspondence analysis. We again encourage you to obtain the workbook by Greenacre (1993) if you want to study this technique in more depth.