Chapter 24
The CORRESP Procedure

Chapter Table of Contents

Chapter 24
The CORRESP Procedure

Overview

The CORRESP procedure performs simple and multiple correspondence analysis. You can use correspondence analysis to find a low-dimensional graphical representation of the rows and columns of a crosstabulation or contingency table. Each row and column is represented by a point in a plot determined from the cell frequencies. PROC CORRESP can also compute coordinates for supplementary rows and columns.

PROC CORRESP can read two kinds of input: raw categorical responses on two or more classification variables, and a two-way contingency table. The correspondence analysis results can be output and displayed with the %PLOTIT macro.

Background

Correspondence analysis is a popular data analysis method in France and Japan. In France, correspondence analysis was developed under the strong influence of Jean-Paul Benzécri; in Japan, it was developed under Chikio Hayashi. The name *correspondence analysis* is a translation of the French *analyse des correspondances*. The technique apparently has many independent beginnings (for example, Richardson and Kuder 1933; Hirshfeld 1935; Horst 1935; Fisher 1940; Guttman 1941; Burt 1950; Hayashi 1950). It has had many other names, including optimal scaling, reciprocal averaging, optimal scoring, and appropriate scoring in the United States; quantification method in Japan; homogeneity analysis in the Netherlands; dual scaling in Canada; and scalogram analysis in Israel.

Correspondence analysis is described in more detail in French in Benzécri (1973) and Lebart, Morineau, and Tabard (1977). In Japanese, the subject is described in Komazawa (1982), Nishisato (1982), and Kobayashi (1981). In English, correspondence analysis is described in Lebart, Morineau, and Warwick (1984), Greenacre (1984), Nishisato (1980), Tenenhaus and Young (1985); Gifi (1990); Greenacre and Hastie (1987); and many other sources. Hoffman and Franke (1986) offer a short, introductory treatment using examples from the field of market research.

Getting Started

Data are available containing the numbers of Ph.Ds awarded in the United States during the years 1973 through 1978 (U.S. Bureau of the Census 1979). The table has six rows, one for each of six academic disciplines, and six columns for the six years. The following DATA step reads the complete table into a SAS data set, and PROC CORRESP displays correspondence analysis results including the inertia decomposition and coordinates. The concept of *inertia* in correspondence analysis is analogous to the concept of variance in principal component analysis, and it is proportional to the chi-square information. The %PLOTIT macro creates a graphical scatterplot of the results. See Appendix B, "Using the %PLOTIT Macro," for more information on the %PLOTIT macro.

```
title "Number of Ph.D's Awarded from 1973 to 1978";
data PhD;
   input Science $ 1-19 y1973-y1978;
   label y1973 = '1973'
         y1974 = '1974'
         y1975 = '1975'
         y1976 = '1976'
         y1977 = '1977'
         y1978 = '1978';
   datalines;
Life Sciences       4489 4303 4402 4350 4266 4361
Physical Sciences   4101 3800 3749 3572 3410 3234
Social Sciences     3354 3286 3344 3278 3137 3008
Behavioral Sciences 2444 2587 2749 2878 2960 3049
Engineering         3338 3144 2959 2791 2641 2432
Mathematics         1222 1196 1149 1003  959  959
;

proc corresp data=PhD out=Results short;
   var y1973-y1978;
   id Science;
run;

%plotit(data=Results, datatype=corresp, plotvars=Dim1 Dim2)
```

Number of Ph.D's Awarded from 1973 to 1978

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	19 38 57 76 95
0.05845	0.00342	368.653	96.04	96.04	**********
0.00861	0.00007	7.995	2.08	98.12	*
0.00694	0.00005	5.197	1.35	99.48	
0.00414	0.00002	1.852	0.48	99.96	
0.00122	0.00000	0.160	0.04	100.00	
Total	0.00356	383.856	100.00		

Degrees of Freedom = 25

Figure 24.1. Inertia and Chi-Square Decomposition

The total chi-square statistic, which is a measure of the association between the rows and columns in the full five dimensions of the (centered) table, is 383.856. The maximum number of dimensions (or axes) is the minimum of the number of rows and columns, minus one. Over 96% of the total chi-square and inertia is explained by the first dimension, indicating that the association between the row and column categories is essentially one dimensional. The plot shows how the number of doctorates in the different areas changes over time. The plot shows that the number of doctorates in the behavioral sciences is associated with later years, and the number of doctorates in mathematics and engineering is associated with earlier years. This is consistent with the data which shows that number of doctorates in the behavioral sciences is increasing, the number of doctorates in every other discipline is decreasing, and the rate of decrease is greatest for mathematics and engineering.

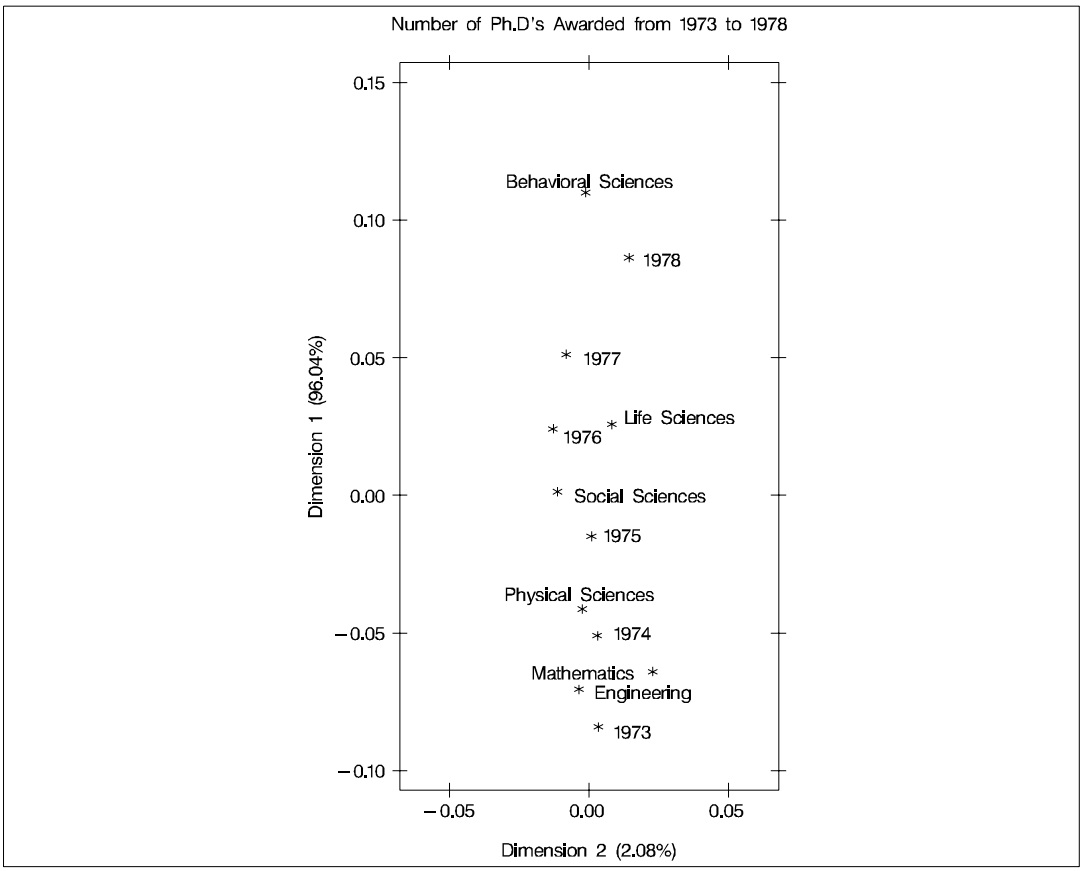


Figure 24.2. Plot of Dimension 1 versus Dimension 2 for Ph.D. Data

Syntax

The following statements are available in the CORRESP procedure.

PROC CORRESP < *options* > **;**
TABLES < *row-variables,* > *column-variables* **;**
VAR *variables* **;**
BY *variables* **;**
ID *variable* **;**
SUPPLEMENTARY *variables* **;**
WEIGHT *variable* **;**

There are two separate forms of input to PROC CORRESP. One form is specified in the TABLES statement, the other in the VAR statement. You must specify either the TABLES or the VAR statement, but not both, each time you run PROC CORRESP.

Specify the TABLES statement if you are using raw, categorical data, the levels of which define the rows and columns of a table.

Specify the VAR statement if your data are already in tabular form. PROC CORRESP is generally more efficient with VAR statement input than with TABLES statement input.

The other statements are optional. Each of the statements is explained in alphabetical order following the PROC CORRESP statement. All of the options in PROC CORRESP can be abbreviated to their first three letters, except for the OUTF= option. This is a special feature of PROC CORRESP and is not generally true of SAS/STAT procedures.

PROC CORRESP Statement

PROC CORRESP < *options* > **;**

The PROC CORRESP statement invokes the procedure. You can specify the following options in the PROC CORRESP statement. These options are described following Table 24.1.

Table 24.1. Summary of PROC CORRESP Statement Options

Task	Options
Specify data sets	
specify input SAS data set	DATA=
specify output coordinate SAS data set	OUTC=
specify output frequency SAS data set	OUTF=
Compute row and column coordinates	
specify the number of dimensions or axes	DIMENS=
perform multiple correspondence analysis	MCA
standardize the row and column coordinates	PROFILE=
Construct tables	
specify binary table	BINARY
specify cross levels of TABLES variables	CROSS=
specify input data in PROC FREQ output	FREQOUT
include observations with missing values	MISSING
Display output	
display all output	ALL
display inertias adjusted by Benzécri's method	BENZECRI
display cell contributions to chi-square	CELLCHI2
display column profile matrix	CP
display observed minus expected values	DEVIATION
display chi-square expected values	EXPECTED
display inertias adjusted by Greenacre's method	GREENACRE
suppress the display of column coordinates	NOCOLUMN=
suppress the display of all output	NOPRINT
suppress the display of row coordinates	NOROW=
display contingency table of observed frequencies	OBSERVED
display percentages or frequencies	PRINT=

Task	Options
display row profile matrix	RP
suppress all point and coordinate statistics	SHORT
display unadjusted inertias	UNADJUSTED
Other tasks	
specify rarely used column coordinate standardizations	COLUMN=
specify minimum inertia	MININERTIA=
specify number of classification variables	NVARS=
specify rarely used row coordinate standardizations	ROW=
specify effective zero	SINGULAR=
include level source in the OUTC= data set	SOURCE

The display options control the amount of displayed output. The CELLCHI2, EXPECTED, and DEVIATION options display additional chi-square information. See the "Details" section on page 961 for more information. The unit of the matrices displayed by the CELLCHI2, CP, DEVIATION, EXPECTED, OBSERVED, and RP options depends on the value of the PRINT= option. The table construction options control the construction of the contingency table; these options are valid only when you also specify a TABLES statement.

You can specify the following options in the PROC CORRESP statement. They are described in alphabetical order.

ALL
is equivalent to specifying the OBSERVED, RP, CP, CELLCHI2, EXPECTED, and DEVIATION options. Specifying the ALL option does not affect the PRINT= option. Therefore, only frequencies (not percentages) for these options are displayed unless you specify otherwise with the PRINT= option.

BENZECRI | BEN
displays adjusted inertias when performing multiple correspondence analysis. By default, unadjusted inertias, the usual inertias from multiple correspondence analysis, are displayed. However, adjusted inertias using a method proposed by Benzécri (1979) and described by Greenacre (1984, p. 145) can be displayed by specifying the BENZECRI option. Specify the UNADJUSTED option to output the usual table of unadjusted inertias as well. See the section "MCA Adjusted Inertias" on page 981 for more information.

BINARY
enables you to create binary tables easily. When you specify the BINARY option, specify only column variables in the TABLES statement. Each input data set observation forms a single row in the constructed table.

CELLCHI2 | CEL
displays the contribution to the total chi-square test statistic for each cell. See also the descriptions of the DEVIATION, EXPECTED, and OBSERVED options.

COLUMN=B | BD | DB | DBD | DBD1/2 | DBID1/2
COL=B | BD | DB | DBD | DBD1/2 | DBID1/2
provides other standardizations of the column coordinates. The COLUMN= option is rarely needed. Typically, you should use the PROFILE= option instead (see the section "The PROFILE=, ROW=, and COLUMN= Options" on page 978). By default, COLUMN=DBD.

CP
displays the column profile matrix. Column profiles contain the observed conditional probabilities of row membership given column membership. See also the RP option.

CROSS=BOTH | COLUMN | NONE | ROW
CRO=BOT | COL | NON | ROW
specifies the method of crossing (factorially combining) the levels of the TABLES variables. The default is CROSS=NONE.

- CROSS=NONE causes each level of every row variable to become a row label and each level of every column variable to become a column label.
- CROSS=ROW causes each combination of levels for all row variables to become a row label, whereas each level of every column variable becomes a column label.
- CROSS=COLUMN causes each combination of levels for all column variables to become a column label, whereas each level of every row variable becomes a row label.
- CROSS=BOTH causes each combination of levels for all row variables to become a row label and each combination of levels for all column variables to become a column label.

The "TABLES Statement" section on page 959 provides a more detailed description of this option.

DATA=*SAS-data-set*
specifies the SAS data set to be used by PROC CORRESP. If you do not specify the DATA= option, PROC CORRESP uses the most recently created SAS data set.

DEVIATION | DEV
displays the matrix of deviations between the observed frequency matrix and the product of its row marginals and column marginals divided by its grand frequency. For ordinary two-way contingency tables, these are the observed minus expected frequencies under the hypothesis of row and column independence and are components of the chi-square test statistic. See also the CELLCHI2, EXPECTED, and OBSERVED options.

DIMENS=*n*
DIM=*n*
specifies the number of dimensions or axes to use. The default is DIMENS=2. The maximum value of the DIMENS= option in an $(n_r \times n_c)$ table is $n_r - 1$ or $n_c - 1$, whichever is smaller. For example, in a table with 4 rows and 5 columns, the

maximum specification is DIMENS=3. If your table has 2 rows or 2 columns, specify DIMENS=1.

EXPECTED | EXP

displays the product of the row marginals and the column marginals divided by the grand frequency of the observed frequency table. For ordinary two-way contingency tables, these are the expected frequencies under the hypothesis of row and column independence and are components of the chi-square test statistic. In other situations, this interpretation is not strictly valid. See also the CELLCHI2, DEVIATION, and OBSERVED options.

FREQOUT | FRE

indicates that the PROC CORRESP input data set has the same form as an output data set from the FREQ procedure, even if it was not directly produced by PROC FREQ. The FREQOUT option enables PROC CORRESP to take shortcuts in constructing the contingency table.

When you specify the FREQOUT option, you must also specify a WEIGHT statement. The cell frequencies in a PROC FREQ output data set are contained in a variable called COUNT, so specify COUNT in a WEIGHT statement with PROC CORRESP. The FREQOUT option may produce unexpected results if the DATA= data set is structured incorrectly. Each of the two variable lists specified in the TABLES statement must consist of a single variable, and observations must be grouped by the levels of the row variable and then by the levels of the column variable. It is not required that the observations be sorted by the row variable and column variable, but they must be grouped consistently. There must be as many observations in the input data set (or BY group) as there are cells in the completed contingency table. Zero cells must be specified with zero weights. When you use PROC FREQ to create the PROC CORRESP input data set, you must specify the SPARSE option in the FREQ procedure's TABLES statement so that the zero cells are written to the output data set.

GREENACRE | GRE

displays adjusted inertias when performing multiple correspondence analysis. By default, unadjusted inertias, the usual inertias from multiple correspondence analysis, are displayed. However, adjusted inertias using a method proposed by Greenacre (1994, p. 156) can be displayed by specifying the GREENACRE option. Specify the UNADJUSTED option to output the usual table of unadjusted inertias as well. See the section "MCA Adjusted Inertias" on page 981 for more information.

MCA

requests a multiple correspondence analysis. This option requires that the input table be a Burt table, which is a symmetric matrix of crosstabulations among several categorical variables. If you specify the MCA option and a VAR statement, you must also specify the NVARS= option, which gives the number of categorical variables that were used to create the table. With raw categorical data, if you want results for the individuals as well as the categories, use the BINARY option instead.

MININERTIA=*n*
MIN=*n*

specifies the minimum inertia $(0 \leq n \leq 1)$ used to create the "best" tables—the indicator of which points best explain the inertia of each dimension. By default, MININERTIA=0.8. See the "Algorithm and Notation" section on page 976 for more information.

MISSING | MIS

specifies that observations with missing values for the TABLES statement variables are included in the analysis. Missing values are treated as a distinct level of each categorical variable. By default, observations with missing values are excluded from the analysis.

NOCOLUMN < = BOTH | DATA | PRINT >
NOC < = BOT | DAT | PRI >

suppresses the display of the column coordinates and statistics and omits them from the output coordinate data set.

BOTH	suppresses all column information from both the SAS listing and the output data set. The NOCOLUMN option is equivalent to the option NOCOLUMN=BOTH.
DATA	suppresses all column information from the output data set.
PRINT	suppresses all column information from the SAS listing.

NOPRINT | NOP

suppresses the display of all output. This option is useful when you need only an output data set. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 15, "Using the Output Delivery System."

NOROW < = BOTH | DATA | PRINT >
NOR < = BOT | DAT | PRI >

suppresses the display of the row coordinates and statistics and omits them from the output coordinate data set.

BOTH	suppresses all row information from both the SAS listing and the output data set. The NOROW option is equivalent to the option NOROW=BOTH.
DATA	suppresses all row information from the output data set.
PRINT	suppresses all row information from the SAS listing.

The NOROW option can be useful when the rows of the contingency table are replications.

NVARS=*n*
NVA=*n*
specifies the number of classification variables that were used to create the Burt table. For example, if the Burt table was originally created with the statement

```
tables a b c;
```

you must specify NVARS=3 to read the table with a VAR statement.

The NVARS= option is required when you specify both the MCA option and a VAR statement. (See the section "VAR Statement" on page 960 for an example.)

OBSERVED | OBS
displays the contingency table of observed frequencies and its row, column, and grand totals. If you do not specify the OBSERVED or ALL option, the contingency table is not displayed.

OUTC=*SAS-data-set*
OUT=*SAS-data-set*
creates an output coordinate SAS data set to contain the row, column, supplementary observation, and supplementary variable coordinates. This data set also contains the masses, squared cosines, quality of each point's representation in the DIMENS=n dimensional display, relative inertias, partial contributions to inertia, and best indicators.

OUTF=*SAS-data-set*
creates an output frequency SAS data set to contain the contingency table, row, and column profiles, the expected values, and the observed minus expected values and contributions to the chi-square statistic.

PRINT=BOTH | FREQ | PERCENT
PRI=BOT | FRE | PER
affects the OBSERVED, RP, CP, CELLCHI2, EXPECTED, and DEVIATION options. The default is PRINT=FREQ.

- The PRINT=FREQ option displays output in the appropriate raw or natural units. (That is, PROC CORRESP displays raw frequencies for the OBSERVED option, relative frequencies with row marginals of 1.0 for the RP option, and so on.)
- The PRINT=PERCENT option scales results to percentages for the display of the output. (All elements in the OBSERVED matrix sum to 100.0, the row marginals are 100.0 for the RP option, and so on.)
- The PRINT=BOTH option displays both percentages and frequencies.

PROFILE=BOTH | COLUMN | NONE | ROW
PRO=BOT | COL | NON | ROW
specifies the standardization for the row and column coordinates. The default is PROFILE=BOTH.

PROFILE=BOTH specifies a standard correspondence analysis, which jointly displays the principal row and column coordinates. Row coordinates are computed from the row profile matrix, and column coordinates are computed from the column profile matrix.

PROFILE=ROW specifies a correspondence analysis of the row profile matrix. The row coordinates are weighted centroids of the column coordinates.

PROFILE=COLUMN specifies a correspondence analysis of the column profile matrix. The column coordinates are weighted centroids of the row coordinates.

PROFILE=NONE is rarely needed. Row and column coordinates are the generalized singular vectors, without the customary standardizations.

ROW=A | AD | DA | DAD | DAD1/2 | DAID1/2

provides other standardizations of the row coordinates. The ROW= option is rarely needed. Typically, you should use the PROFILE= option instead (see the section "The PROFILE=, ROW=, and COLUMN= Options" on page 978). By default, ROW=DAD.

RP

displays the row profile matrix. Row profiles contain the observed conditional probabilities of column membership given row membership. See also the CP option.

SHORT | SHO

suppresses the display of all point and coordinate statistics except the coordinates. The following information is suppressed: each point's mass, relative contribution to the total inertia, and quality of representation in the DIMENS=n dimensional display; the squared cosines of the angles between each axis and a vector from the origin to the point; the partial contributions of each point to the inertia of each dimension; and the best indicators.

SINGULAR=*n*
SIN=*n*

specifies the largest value that is considered to be within rounding error of zero. The default value is 1E−8. This parameter is used when checking for zero rows and columns, when checking Burt table diagonal sums for equality, when checking denominators before dividing, and so on. Typically, you should not assign a value outside the range 1E−6 to 1E−12.

SOURCE | SOU

adds the variable _VAR_, which contains the name or label of the variable corresponding to the current level, to the OUTC= and OUTF= data sets.

UNADJUSTED | UNA

displays unadjusted inertias when performing multiple correspondence analysis. By default, unadjusted inertias, the usual inertias from multiple correspondence analysis, are displayed. However, if adjusted inertias are requested by either the GREENACRE option or the BENZECRI option, then the unadjusted inertia table is not displayed unless the UNADJUSTED option is specified. See the section "MCA Adjusted Inertias" on page 981 for more information.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC CORRESP to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the CORRESP procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide.*

ID Statement

ID *variable* **;**

You specify the ID statement only in conjunction with the VAR statement. You cannot specify the ID statement when you use the TABLES statement or the MCA option. When you specify an ID variable, PROC CORRESP labels the rows of the tables with the ID values and places the ID variable in the output data set.

SUPPLEMENTARY Statement

SUPPLEMENTARY *variables* **;**
SUP *variables* **;**

The SUPPLEMENTARY statement specifies variables that are to be represented as points in the joint row and column space but that are not used when determining the locations of the other, active row and column points of the contingency table. Supplementary observations on supplementary variables are ignored in simple correspondence analysis but are needed to compute the squared cosines for multiple corre-

spondence analysis. Variables that are specified in the SUPPLEMENTARY statement must also be specified in the TABLES or VAR statement.

When you specify a VAR statement, each SUPPLEMENTARY variable indicates one supplementary column of the table. Supplementary variables must be numeric with VAR statement input.

When you specify a TABLES statement, each SUPPLEMENTARY variable indicates a set of rows or columns of the table that is supplementary. Supplementary variables can be either character or numeric with TABLES statement input.

TABLES Statement

TABLES < *row-variables,* > *column-variables* **;**

The TABLES statement instructs PROC CORRESP to create a contingency table, Burt table, or binary table from the values of two or more categorical variables. The TABLES statement specifies classification variables that are used to construct the rows and columns of the contingency table. The variables can be either numeric or character. The variable lists in the TABLES statement and the CROSS= option together determine the row and column labels of the contingency table.

You can specify both row variables and column variables separated by a comma, or you can specify only column variables and no comma. If you do not specify row variables (that is, if you list variables but do not use the comma as a delimiter), then you should specify either the MCA or the BINARY option. With the MCA option, PROC CORRESP creates a Burt table, which is a crosstabulation of each variable with itself and every other variable. The Burt table is symmetric. With the BINARY option, PROC CORRESP creates a binary table, which consists of one row for each input data set observation and one column for each category of each TABLES statement variable. If the binary matrix is $\mathbf{Z}$, then the Burt table is $\mathbf{Z}'\mathbf{Z}$. Specifying the BINARY option with the NOROWS option produces the same results as specifying the MCA option (except for the chi-square statistics).

See Figure 24.3 for an example or see the section “The MCA Option” on page 980 for a detailed description of Burt tables.

You can use the WEIGHT statement with the TABLES statement to read category frequencies. Specify the SUPPLEMENTARY statement to name variables with categories that are supplementary rows or columns. You cannot specify the ID or VAR statement with the TABLES statement.

See the section “Using the TABLES Statement” on page 967 for an example.

VAR Statement

VAR *variables* **;**

You should specify the VAR statement when your data are in tabular form. The VAR variables must be numeric. The VAR statement instructs PROC CORRESP to read an existing contingency table, binary indicator matrix, fuzzy-coded indicator matrix, or Burt table, rather than raw data. See the "Algorithm and Notation" section on page 976 for a description of a binary indicator matrix and a fuzzy-coded indicator matrix.

You can specify the WEIGHT statement with the VAR statement to read category frequencies and designate supplementary rows. Specify the SUPPLEMENTARY statement to name supplementary variables. You cannot specify the TABLES statement with the VAR statement.

WEIGHT Statement

WEIGHT *variable* **;**

The WEIGHT statement specifies weights for each observation and indicates supplementary observations for simple correspondence analyses with VAR statement input. You can include only one WEIGHT statement, and the weight variable must be numeric.

If you omit the WEIGHT statement, each observation contributes a value of 1 to the frequency count for its category. That is, each observation represents one subject. When you specify a WEIGHT statement, each observation contributes the value of the weighting variable for that observation. For example, a weight of 3 means that the observation represents 3 subjects. Weight values are not required to be integers.

You can specify the WEIGHT statement with a TABLES statement to indicate category frequencies, as in the following example:

```
proc freq;
   tables a*b / out=outfreq sparse;
run;

proc corresp freqout;
   tables a, b;
   weight count;
run;
```

If you specify a VAR statement, you can specify the WEIGHT statement to indicate supplementary observations and to weight some rows of the table more heavily than others. When the value of the WEIGHT variable is negative, the observation is treated as supplementary, and the absolute value of the weight is used as the weighting value.

You cannot specify a WEIGHT statement with a VAR statement and the MCA option, because the table must be symmetric. Supplementary variables are indicated with the SUPPLEMENTARY statement, so differential weighting of rows is inappropriate.

Details

Input Data Set

PROC CORRESP can read two kinds of input:

- raw category responses on two or more classification variables with the TABLES statement
- a two-way contingency table with the VAR statement

You can use output from PROC FREQ as input for PROC CORRESP.

The classification variables referred to by the TABLES statement can be either numeric or character variables. Normally, all observations for a given variable that have the same formatted value are placed in the same level, and observations with different values are placed in different levels.

The variables in the VAR statement must be numeric. The values of the observations specify the cell frequencies. These values are not required to be integers, but only those observations with all nonnegative, nonmissing values are used in the correspondence analysis. Observations with one or more negative values are removed from the analysis.

The WEIGHT variable must be numeric. Observations with negative weights are treated as supplementary observations. The absolute values of the weights are used to weight the observations.

Types of Tables Used as Input

The following example explains correspondence analysis and illustrates some capabilities of PROC CORRESP.

```
data Neighbor;
   input Name $ 1-10 Age $ 12-18 Sex $ 19-25
         Height $ 26-30 Hair $ 32-37;
   datalines;
Jones      Old    Male   Short White
Smith      Young  Female Tall  Brown
Kasavitz   Old    Male   Short Brown
Ernst      Old    Female Tall  White
Zannoria   Old    Female Short Brown
Spangel    Young  Male   Tall  Blond
Myers      Young  Male   Tall  Brown
Kasinski   Old    Male   Short Blond
Colman     Young  Female Short Blond
Delafave   Old    Male   Tall  Brown
```

```
Singer     Young  Male   Tall  Brown
Igor       Old           Short
;
```

There are several types of tables, $\mathbf{N}$, that can be used as input to correspondence analysis —all tables can be defined using a binary matrix, $\mathbf{Z}$.

With the BINARY option, $\mathbf{N} = \mathbf{Z}$ is directly analyzed. The binary matrix has one column for each category and one row for each individual or case. A binary table constructed from m categorical variables has m partitions. The following table has four partitions, one for each of the four categorical variables. Each partition has a 1 in each row, and each row contains exactly four 1s since there are four categorical variables. More generally, the binary design matrix has exactly m 1s in each row. The 1s indicate the categories to which the observation applies.

Table 24.2. $\mathbf{Z}$, The Binary Coding of Neighbor Data Set

$\mathbf{Z}_{\mathbf{Hair}}$			$\mathbf{Z}_{\mathbf{Height}}$		$\mathbf{Z}_{\mathbf{Sex}}$		$\mathbf{Z}_{\mathbf{Age}}$	
Blond	Brown	White	Short	Tall	Female	Male	Old	Young
0	0	1	1	0	0	1	1	0
0	1	0	0	1	1	0	0	1
0	1	0	1	0	0	1	1	0
0	0	1	0	1	1	0	1	0
0	1	0	1	0	1	0	1	0
1	0	0	0	1	0	1	0	1
0	1	0	0	1	0	1	0	1
1	0	0	1	0	0	1	1	0
1	0	0	1	0	1	0	0	1
0	1	0	0	1	0	1	1	0
0	1	0	0	1	0	1	0	1

With the MCA option, the Burt table ($\mathbf{Z}'\mathbf{Z}$) is analyzed. A Burt table is a partitioned symmetric matrix containing all pairs of crosstabulations among a set of categorical variables. Each diagonal partition is a diagonal matrix containing marginal frequencies (a crosstabulation of a variable with itself). Each off-diagonal partition is an ordinary contingency table. Each contingency table above the diagonal has a transposed counterpart below the diagonal.

Table 24.3. $\mathbf{Z}'\mathbf{Z}$, The Burt Table

	Blond	Brown	White	Short	Tall	Female	Male	Old	Young
Blond	3	0	0	2	1	1	2	1	2
Brown	0	6	0	2	4	2	4	3	3
White	0	0	2	1	1	1	1	2	0
Short	2	2	1	5	0	2	3	4	1
Tall	1	4	1	0	6	2	4	2	4
Female	1	2	1	2	2	4	0	2	2
Male	2	4	1	3	4	0	7	4	3
Old	1	3	2	4	2	2	4	6	0
Young	2	3	0	1	4	2	3	0	5

This Burt table is composed of all pairs of crosstabulations among the variables Hair, Height, Sex, and Age. It is composed of sixteen individual subtables —the number of variables squared. Both the rows and the columns have the same nine categories (in this case Blond, Brown, White, Short, Tall, Female, Male, Old, and Young). The off-diagonal partitions are crosstabulations of each variable with every other variable. Below the diagonal are the following crosstabulations (from left to right, top to bottom): Height * Hair, Sex * Hair, Sex * Height, Age * Hair, Age * Height, and Age * Sex. Each crosstabulation below the diagonal has a transposed counterpart above the diagonal. Each diagonal partition contains a crosstabulation of a variable with itself (Hair * Hair, Height * Height, Sex * Sex, and Age * Age). The diagonal elements of the diagonal partitions contain marginal frequencies of the off-diagonal partitions.

For example, the table Hair * Height has three rows for Hair and two columns for Height. The values of the Hair * Height table, summed across rows, sum to the diagonal values of the Height * Height table, as displayed in the following table.

Table 24.4. $\mathbf{Z}_{\mathbf{Hair},\mathbf{Height}}{}'\mathbf{Z}_{\mathbf{Height}}$, The (Hair Height) $\times$ Height Crosstabulation

	Short	Tall
Blond	2	1
Brown	2	4
White	1	1
Short	5	0
Tall	0	6

A simple crosstabulation of Hair $\times$ Height is $\mathbf{N} = \mathbf{Z}_{\mathbf{Hair}}{}'\mathbf{Z}_{\mathbf{Height}}$. Crosstabulations such as this, involving only two variables, are the input to simple correspondence analysis.

Table 24.5. $\mathbf{Z}_{\mathbf{Hair}}{}'\mathbf{Z}_{\mathbf{Height}}$, The Hair $\times$ Height Crosstabulation

	Short	Tall
Blond	2	1
Brown	2	4
White	1	1

Tables such as the following ($\mathbf{N} = \mathbf{Z}_{\mathbf{Hair}}{}'\mathbf{Z}_{\mathbf{Height,Sex}}$), made up of several crosstabulations, can also be analyzed in simple correspondence analysis.

Table 24.6. $\mathbf{Z}_{\mathbf{Hair}}{}'\mathbf{Z}_{\mathbf{Height,Sex}}$, The Hair $\times$ (Height Sex) Crosstabulation

	Short	Tall	Female	Male
Blond	2	1	1	2
Brown	2	4	2	4
White	1	1	1	1

Coding, Fuzzy Coding, and Doubling

You can use an indicator matrix as input to PROC CORRESP using the VAR statement. An indicator matrix is composed of several submatrices, each of which is a design matrix with one column for each category of a categorical variable. In order to create an indicator matrix, you must code an indicator variable for each level of each categorical variable. For example, the categorical variable Sex, with two levels (Female and Male), would be coded using two indicator variables.

A binary indicator variable is coded 1 to indicate the presence of an attribute and 0 to indicate its absence. For the variable Sex, a male would be coded Female=0 and Male=1, and a female would be coded Female=1 and Male=0. The indicator variables representing a categorical variable must sum to 1.0. You can specify the BINARY option to create a binary table.

Sometimes binary data such as Yes/No data are available. For example, 1 means "Yes, I have bought this brand in the last month" and 0 means "No, I have not bought this brand in the last month".

```
title 'Doubling Yes/No Data';

proc format;
   value yn 0 = 'No '  1 = 'Yes';
   run;

data BrandChoice;
   input a b c;
   label a = 'Brand A' b = 'Brand B' c = 'Brand B';
   format a b c yn.;
   datalines;
0 0 1
1 1 0
0 1 1
0 1 0
1 0 0
;
```

Data such as these cannot be analyzed directly because the raw data do not consist of partitions, each with one column per level and exactly one 1 in each row. The data must be *doubled* so that both Yes and No are both represented by a column in

the data matrix. The TRANSREG procedure provides one way of doubling. In the following statements, the DESIGN option specifies that PROC TRANSREG is being used only for coding, not analysis. The option SEPARATORS=': ' specifies that labels for the coded columns are constructed from input variable labels, followed by a colon and space, followed by the formatted value. The variables are designated in the MODEL statement as CLASS variables, and the ZERO=NONE option creates binary variables for all levels. The OUTPUT statement specifies the output data set and drops the _NAME_, _TYPE_, and Intercept variables. PROC TRANSREG stores a list of coded variable names in a macro variable &_TRGIND, which in this case has the value "aNo aYes bNo bYes cNo cYes". This macro can be used directly in the VAR statement in PROC CORRESP.

```
proc transreg data=BrandChoice design separators=': ';
   model class(a b c / zero=none);
   output out=Doubled(drop=_: Intercept);
run;

proc print label;
run;

proc corresp data=Doubled norow short;
   var &_trgind;
run;
```

A fuzzy-coded indicator also sums to 1.0 across levels of the categorical variable, but it is coded with fractions rather than with 1 and 0. The fractions represent the distribution of the attribute across several levels of the categorical variable.

Ordinal variables, such as survey responses of 1 to 3 can be represented as two design variables.

Table 24.7. Coding an Ordinal Variable

Ordinal Values	Coding	
1	0.25	0.75
2	0.50	0.50
3	0.75	0.25

Values of the coding sum to one across the two coded variables.

This next example illustrates the use of binary and fuzzy-coded indicator variables. Fuzzy-coded indicators are used to represent missing data. Note that the missing values in the observation Igor are coded with equal proportions.

```
proc transreg data=Neighbor design cprefix=0;
   model class(Age Sex Height Hair / zero=none);
   output out=Neighbor2(drop=_: Intercept);
   id Name;
   run;

data Neighbor3;
   set Neighbor2;
```

```
   if Sex = ' ' then do;
      Female = 0.5;
      Male   = 0.5;
      end;
   if Hair = ' ' then do;
      White = 1/3;
      Brown = 1/3;
      Blond = 1/3;
      end;
run;

proc print label;
run;
```

Obs	Age Old	Age Young	Sex Female	Sex Male	Height Short	Height Tall	Hair Blond	Hair Brown	Hair White	Age	Sex	Height	Hair	Name
1	1	0	0.0	1.0	1	0	0.00000	0.00000	1.00000	Old	Male	Short	White	Jones
2	0	1	1.0	0.0	0	1	0.00000	1.00000	0.00000	Young	Female	Tall	Brown	Smith
3	1	0	0.0	1.0	1	0	0.00000	1.00000	0.00000	Old	Male	Short	Brown	Kasavitz
4	1	0	1.0	0.0	0	1	0.00000	0.00000	1.00000	Old	Female	Tall	White	Ernst
5	1	0	1.0	0.0	1	0	0.00000	1.00000	0.00000	Old	Female	Short	Brown	Zannoria
6	0	1	0.0	1.0	0	1	1.00000	0.00000	0.00000	Young	Male	Tall	Blond	Spangel
7	0	1	0.0	1.0	0	1	0.00000	1.00000	0.00000	Young	Male	Tall	Brown	Myers
8	1	0	0.0	1.0	1	0	1.00000	0.00000	0.00000	Old	Male	Short	Blond	Kasinski
9	0	1	1.0	0.0	1	0	1.00000	0.00000	0.00000	Young	Female	Short	Blond	Colman
10	1	0	0.0	1.0	0	1	0.00000	1.00000	0.00000	Old	Male	Tall	Brown	Delafave
11	0	1	0.0	1.0	0	1	0.00000	1.00000	0.00000	Young	Male	Tall	Brown	Singer
12	1	0	0.5	0.5	1	0	0.33333	0.33333	0.33333	Old		Short		Igor

Figure 24.3. Fuzzy Coding of Missing Values

There is one set of coded variables for each input categorical variable. If observation 12 is excluded, each set is a binary design matrix. Each design matrix has one column for each category and exactly one 1 in each row.

Fuzzy-coding is shown in the final observation, Igor. The observation Igor has missing values for the variables Sex and Hair. The design matrix variables are coded with fractions that sum to one within each categorical variable.

An alternative way to represent missing data is to treat missing values as an additional level of the categorical variable. This alternative is available with the MISSING option in the PROC statement. This approach yields coordinates for missing responses, allowing the comparison of "missing" along with the other levels of the categorical variables.

Greenacre and Hastie (1987) discuss additional coding schemes, including one for continuous variables. Continuous variables can be coded with PROC TRANSREG by specifying BSPLINE(*variables* / degree=1) in the MODEL statement.

Using the TABLES Statement

In the following TABLES statement, each variable list consists of a single variable:

```
proc corresp data=Neighbor dimens=1 observed short;
   ods select observed;
   tables Sex, Age;
run;
```

These statements create a contingency table with two rows (Female and Male) and two columns (Old and Young) and show the neighbors broken down by age and sex. The DIMENS=1 option overrides the default, which is DIMENS=2. The OBSERVED option displays the contingency table. The SHORT option limits the displayed output. Because it contains missing values, the observation where Name='Igor' is omitted from the analysis. Figure 24.4 displays the contingency table.

The CORRESP Procedure

Contingency Table

	Old	Young	Sum
Female	2	2	4
Male	4	3	7
Sum	6	5	11

Figure 24.4. Contingency Table for Sex, Age

The following statements create a table with six rows (**Blond*Short**, **Blond*Tall**, **Brown*Short**, **Brown*Tall**, **White*Short**, and **White*Tall**), and four columns (**Female**, **Male**, **Old**, and **Young**). The levels of the row variables are crossed, forming mutually exclusive categories, whereas the categories of the column variables overlap.

```
proc corresp data=Neighbor cross=row observed short;
   ods select observed;
   tables Hair Height, Sex Age;
run;
```

The CORRESP Procedure

Contingency Table

	Female	Male	Old	Young	Sum
Blond * Short	1	1	1	1	4
Blond * Tall	0	1	0	1	2
Brown * Short	1	1	2	0	4
Brown * Tall	1	3	1	3	8
White * Short	0	1	1	0	2
White * Tall	1	0	1	0	2
Sum	4	7	6	5	22

Figure 24.5. Contingency Table for Hair Height, Sex Age

You can enter supplementary variables with TABLES input by including a SUPPLEMENTARY statement. Variables named in the SUPPLEMENTARY statement indicate TABLES variables with categories that are supplementary. In other words, the categories of the variable Age are represented in the row and column space, but they are not used in determining the scores of the categories of the variables Hair, Height, and Sex. The variable used in the SUPPLEMENTARY statement must be listed in the TABLES statement as well. For example, the following statements create a Burt table with seven active rows and columns (**Blond**, **Brown**, **White**, **Short**, **Tall**, **Female**, **Male**) and two supplementary rows and columns (**Old** and **Young**).

```
proc corresp data=Neighbor observed short mca;
   ods select burt supcols;
   tables Hair Height Sex Age;
   supplementary Age;
run;
```

The CORRESP Procedure

Burt Table

	Blond	Brown	White	Short	Tall	Female	Male
Blond	3	0	0	2	1	1	2
Brown	0	6	0	2	4	2	4
White	0	0	2	1	1	1	1
Short	2	2	1	5	0	2	3
Tall	1	4	1	0	6	2	4
Female	1	2	1	2	2	4	0
Male	2	4	1	3	4	0	7

Supplementary Columns

	Old	Young
Blond	1	2
Brown	3	3
White	2	0
Short	4	1
Tall	2	4
Female	2	2
Male	4	3

Figure 24.6. Burt Table from PROC CORRESP

The following statements create a binary table with 7 active columns (**Blond**, **Brown**, **White**, **Short**, **Tall**, **Female**, **Male**), 2 supplementary columns (**Old** and **Young**), and 11 rows for the 11 observations with nonmissing values.

```
proc corresp data=Neighbor observed short binary;
   ods select binary supcols;
   tables Hair Height Sex Age;
   supplementary Age;
run;
```

The CORRESP Procedure

Binary Table

	Blond	Brown	White	Short	Tall	Female	Male
1	0	0	1	1	0	0	1
2	0	1	0	0	1	1	0
3	0	1	0	1	0	0	1
4	0	0	1	0	1	1	0
5	0	1	0	1	0	1	0
6	1	0	0	0	1	0	1
7	0	1	0	0	1	0	1
8	1	0	0	1	0	0	1
9	1	0	0	1	0	1	0
10	0	1	0	0	1	0	1
11	0	1	0	0	1	0	1

Supplementary Columns

	Old	Young
1	1	0
2	0	1
3	1	0
4	1	0
5	1	0
6	0	1
7	0	1
8	1	0
9	0	1
10	1	0
11	0	1

Figure 24.7. Binary Table from PROC CORRESP

Using the VAR Statement

With VAR statement input, the rows of the contingency table correspond to the observations of the input data set, and the columns correspond to the VAR statement variables. The values of the variables typically contain the table frequencies. The example displayed in Figure 24.4 could be run with VAR statement input using the following code:

```
data Ages;
   input Sex $ Old Young;
   datalines;
Female  2 2
Male    4 3
;

proc corresp data=Ages dimens=1 observed short;
   var Old Young;
   id Sex;
run;
```

Only nonnegative values are accepted. Negative values are treated as missing, causing the observation to be excluded from the analysis. The values are not required to be integers. Row labels for the table are specified with an ID variable. Column labels are constructed from the variable name or variable label if one is specified. When you specify multiple correspondence analysis (MCA), the row and column labels are the same and are constructed from the variable names or labels, so you cannot include an ID statement. With MCA, the VAR statement must list the variables in the order in which the rows occur. For example, the table displayed in Figure 24.6, which was created with the following TABLES statement,

```
tables Hair Height Sex Age;
```

is input as follows with the VAR statement:

```
proc corresp data=table nvars=4 mca;
   var Blond Brown White Short Tall Female Male Old Young;
run;
```

You must specify the NVARS= option to specify the number of original categorical variables with the MCA option. The option NVARS=n is needed to find boundaries between the subtables of the Burt table. If f is the sum of all elements in the Burt table $\mathbf{Z}'\mathbf{Z}$, then fn^{-2} is the number of rows in the binary matrix $\mathbf{Z}$. The sum of all elements in each diagonal subtable of the Burt table must be fn^{-2}.

To enter supplementary observations, include a WEIGHT statement with negative weights for those observations. Specify the SUPPLEMENTARY statement to include supplementary variables. You must list supplementary variables in both the VAR and SUPPLEMENTARY statements.

Missing and Invalid Data

With VAR statement input, observations with missing or negative frequencies are excluded from the analysis. Supplementary variables and supplementary observations with missing or negative frequencies are also excluded. Negative weights are valid with VAR statement input.

With TABLES statement input, observations with negative weights are excluded from the analysis. With this form of input, missing cell frequencies cannot occur. Observations with missing values on the categorical variables are excluded unless you specify the MISSING option. If you specify the MISSING option, ordinary missing values and special missing values are treated as additional levels of a categorical variable. In all cases, if any row or column of the constructed table contains only zeros, that row or column is excluded from the analysis.

Observations with missing weights are excluded from the analysis.

Creating a Data Set Containing the Crosstabulation

The CORRESP procedure can read or create a contingency or Burt table. PROC CORRESP is generally more efficient with VAR statement input than with TABLES statement input. TABLES statement input requires that the table be created from raw categorical variables, whereas the VAR statement is used to read an existing table. If PROC CORRESP runs out of memory, it may be possible to use some other method to create the table and then use VAR statement input with PROC CORRESP.

The following example uses the CORRESP, FREQ, and TRANSPOSE procedures to create rectangular tables from a SAS data set WORK.A that contains the categorical variables V1–V5. The Burt table examples assume that no categorical variable has a value found in any of the other categorical variables (that is, that each row and column label is unique).

You can use PROC CORRESP and ODS to create a rectangular two-way contingency table from two categorical variables.

```
proc corresp data=a observed short;
   ods listing close;
   ods output Observed=Obs(drop=Sum where=(Label ne 'Sum'));
   tables v1, v2;
run;

ods listing;
```

You can use PROC FREQ and PROC TRANSPOSE to create a rectangular two-way contingency table from two categorical variables.

```
proc freq data=a;
   tables v1 * v2 / sparse noprint out=freqs;
run;

proc transpose data=freqs out=rfreqs;
   id  v2;
   var count;
   by  v1;
run;
```

You can use PROC CORRESP and ODS to create a Burt table from five categorical variables.

```
proc corresp data=a observed short mca;
   ods listing close;
   ods output Burt=Obs;
   tables v1-v5;
run;

ods listing;
```

You can use a DATA step, PROC FREQ, and PROC TRANSPOSE to create a Burt table from five categorical variables.

```
data b;
   set a;
   array v[5] $ v1-v5;
   do i = 1 to 5;
      row = v[i];
      do j = 1 to 5;
         column = v[j];
         output;
         end;
      end;
   keep row column;
run;

proc freq data=b;
   tables row * column / sparse  noprint out=freqs;
run;

proc transpose data=freqs out=rfreqs;
   id  column;
   var count;
   by  row;
run;
```

Output Data Sets

The OUTC= Data Set

The OUTC= data set contains two or three character variables and $4n + 4$ numeric variables, where n is the number of axes from DIMENS=n (two by default). The OUTC= data set contains one observation for each row, column, supplementary row, and supplementary column point, and one observation for inertias.

The first variable is named _TYPE_ and identifies the type of observation. The values of _TYPE_ are as follows:

- The 'INERTIA' observation contains the total inertia in the INERTIA variable, and each dimension's inertia in the Contr1–Contr$n$ variables.
- The 'OBS' observations contain the coordinates and statistics for the rows of the table.
- The 'SUPOBS' observations contain the coordinates and statistics for the supplementary rows of the table.
- The 'VAR' observations contain the coordinates and statistics for the columns of the table.
- The 'SUPVAR' observations contain the coordinates and statistics for the supplementary columns of the table.

If you specify the SOURCE option, then the data set also contains a variable _VAR_ containing the name or label of the input variable from which that row originates. The name of the next variable is either _NAME_ or (if you specify an ID statement) the name of the ID variable.

For observations with a value of 'OBS' or 'SUPOBS' for the _TYPE_ variable, the values of the second variable are constructed as follows:

- When you use a VAR statement without an ID statement, the values are 'Row1', 'Row2', and so on.
- When you specify a VAR statement with an ID statement, the values are set equal to the values of the ID variable.
- When you specify a TABLES statement, the _NAME_ variable has values formed from the appropriate row variable values.

For observations with a value of 'VAR' or 'SUPVAR' for the _TYPE_ variable, the values of the second variable are equal to the names or labels of the VAR (or SUPPLEMENTARY) variables. When you specify a TABLES statement, the values are formed from the appropriate column variable values.

The third and subsequent variables contain the numerical results of the correspondence analysis.

- Quality contains the quality of each point's representation in the DIMENS=n dimensional display, which is the sum of squared cosines over the first n dimensions.
- Mass contains the masses or marginal sums of the relative frequency matrix.
- Inertia contains each point's relative contribution to the total inertia.
- Dim1–Dim$n$ contain the point coordinates.
- Contr1–Contr$n$ contain the partial contributions to inertia.
- SqCos1–SqCos$n$ contain the squared cosines.
- Best1–Best$n$ and Best contain the summaries of the partial contributions to inertia.

The OUTF= Data Set

The OUTF= data set contains frequencies and percentages. It is similar to a PROC FREQ output data set. The OUTF= data set begins with a variable called _TYPE_, which contains the observation type. If the SOURCE option is specified, the data set contains two variables _ROWVAR_ and _COLVAR_ that contain the names or labels of the row and column input variables from which each cell originates. The next two variables are classification variables that contain the row and column levels. If you use TABLES statement input and each variable list consists of a single variable, the names of the first two variables match the names of the input variables; otherwise,

these variables are named Row and Column. The next two variables are Count and Percent, which contain frequencies and percentages.

The _TYPE_ variable can have the following values:

- 'OBSERVED' observations contain the contingency table.
- 'SUPOBS' observations contain the supplementary rows.
- 'SUPVAR' observations contain the supplementary columns.
- 'EXPECTED' observations contain the product of the row marginals and the column marginals divided by the grand frequency of the observed frequency table. For ordinary two-way contingency tables, these are the expected frequency matrix under the hypothesis of row and column independence.
- 'DEVIATION' observations contain the matrix of deviations between the observed frequency matrix and the product of its row marginals and column marginals divided by its grand frequency. For ordinary two-way contingency tables, these are the observed minus expected frequencies under the hypothesis of row and column independence.
- 'CELLCHI2' observations contain contributions to the total chi-square test statistic.
- 'RP' observations contain the row profiles.
- 'SUPRP' observations contain supplementary row profiles.
- 'CP' observations contain the column profiles.
- 'SUPCP' observations contain supplementary column profiles.

Computational Resources

Let

$$
\begin{aligned}
n_r &= \text{number of rows in the table} \\
n_c &= \text{number of columns in the table} \\
n &= \text{number of observations} \\
v &= \text{number of VAR statement variables} \\
t &= \text{number of TABLES statement variables} \\
c &= \max(n_r, n_c) \\
d &= \min(n_r, n_c)
\end{aligned}
$$

For TABLES statement input, more than

$$32(t+1)+8(\max(2tn,(n_r+3)(n_c+3)))$$

bytes of array space are required.

For VAR statement input, more than

$$16(v+2)+8(n_r+3)(n_c+3)$$

bytes of array space are required.

Memory

The computational resources formulas are underestimates of the amounts of memory needed to handle most problems. If you use a utility data set, and if memory could be used with perfect efficiency, then roughly the stated amount of memory would be needed. In reality, most problems require at least two or three times the minimum.

PROC CORRESP tries to store the raw data (TABLES input) and the contingency table in memory. If there is not enough memory, a utility data set is used, potentially resulting in a large increase in execution time.

Time

The time required to perform the generalized singular value decomposition is roughly proportional to $2cd^2+5d^3$. Overall computation time increases with table size at a rate roughly proportional to $(n_r n_c)^{\frac{3}{2}}$.

Algorithm and Notation

This section is primarily based on the theory of correspondence analysis found in Greenacre (1984). If you are interested in other references, see the "Background" section on page 947.

Let $\mathbf{N}$ be the contingency table formed from those observations and variables that are not supplementary and from those observations that have no missing values and have a positive weight. This table is an $(n_r \times n_c)$ rank q matrix of nonnegative numbers with nonzero row and column sums. If $\mathbf{Z}_a$ is the binary coding for variable A, and $\mathbf{Z}_b$ is the binary coding for variable B, then $\mathbf{N}=\mathbf{Z}_a'\mathbf{Z}_b$ is a contingency table. Similarly, if $\mathbf{Z}_{b,c}$ contains the binary coding for both variables B and C, then $\mathbf{N}=\mathbf{Z}_a'\mathbf{Z}_{b,c}$ can also be input to a correspondence analysis. With the BINARY option, $\mathbf{N}=\mathbf{Z}$, and the analysis is based on a binary table. In multiple correspondence analysis, the analysis is based on a Burt table, $\mathbf{Z}'\mathbf{Z}$.

Let $\mathbf{1}$ be a vector of 1s of the appropriate order, let $\mathbf{I}$ be an identity matrix, and let $\mathrm{diag}(\cdot)$ be a matrix-valued function that creates a diagonal matrix from a vector. Let

$$\begin{aligned} f &= \mathbf{1}'\mathbf{N}\mathbf{1} \\ \mathbf{P} &= \frac{1}{f}\mathbf{N} \\ \mathbf{r} &= \mathbf{P}\mathbf{1} \end{aligned}$$

$$
\begin{aligned}
\mathbf{c} &= \mathbf{P}'\mathbf{1} \\
\mathbf{D_r} &= \text{diag}(\mathbf{r}) \\
\mathbf{D_c} &= \text{diag}(\mathbf{c}) \\
\mathbf{R} &= \mathbf{D_r^{-1}P} \\
\mathbf{C}' &= \mathbf{D_c^{-1}P}'
\end{aligned}
$$

The scalar f is the sum of all elements in $\mathbf{N}$. The matrix $\mathbf{P}$ is a matrix of relative frequencies. The vector $\mathbf{r}$ contains row marginal proportions or row "masses." The vector $\mathbf{c}$ contains column marginal proportions or column masses. The matrices $\mathbf{D_r}$ and $\mathbf{D_c}$ are diagonal matrices of marginals.

The rows of $\mathbf{R}$ contain the "row profiles." The elements of each row of $\mathbf{R}$ sum to one. Each (i, j) element of $\mathbf{R}$ contains the observed probability of being in column j given membership in row i. Similarly, the columns of $\mathbf{C}$ contain the column profiles. The coordinates in correspondence analysis are based on the generalized singular value decomposition of $\mathbf{P}$,

$$\mathbf{P} = \mathbf{A}\mathbf{D_u}\mathbf{B}'$$

where

$$\mathbf{A}'\mathbf{D_r^{-1}}\mathbf{A} = \mathbf{B}'\mathbf{D_c^{-1}}\mathbf{B} = \mathbf{I}$$

In multiple correspondence analysis,

$$\mathbf{P} = \mathbf{B}\mathbf{D_u^2}\mathbf{B}'$$

The matrix $\mathbf{A}$, which is the rectangular matrix of left generalized singular vectors, has n_r rows and q columns; the matrix $\mathbf{D_u}$, which is a diagonal matrix of singular values, has q rows and columns; and the matrix $\mathbf{B}$, which is the rectangular matrix of right generalized singular vectors, has n_c rows and q columns. The columns of $\mathbf{A}$ and $\mathbf{B}$ define the principal axes of the column and row point clouds, respectively.

The generalized singular value decomposition of $\mathbf{P} - \mathbf{rc}'$, discarding the last singular value (which is zero) and the last left and right singular vectors, is exactly the same as a generalized singular value decomposition of $\mathbf{P}$, discarding the first singular value (which is one), the first left singular vector, $\mathbf{r}$, and the first right singular vector, $\mathbf{c}$. The first (trivial) column of $\mathbf{A}$ and $\mathbf{B}$ and the first singular value in $\mathbf{D_u}$ are discarded before any results are displayed. You can obtain the generalized singular value decomposition of $\mathbf{P} - \mathbf{rc}'$ from the ordinary singular value decomposition of $\mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc}')\mathbf{D_c^{-1/2}}$.

$$\mathbf{D_r^{-1/2}}(\mathbf{P} - \mathbf{rc}')\mathbf{D_c^{-1/2}} = \mathbf{U}\mathbf{D_u}\mathbf{V}' = (\mathbf{D_r^{-1/2}}\mathbf{A})\mathbf{D_u}(\mathbf{D_c^{-1/2}}\mathbf{B})'$$

$$\mathbf{P} - \mathbf{rc}' = \mathbf{D_r^{1/2}}\mathbf{U}\mathbf{D_u}\mathbf{V}'\mathbf{D_c^{1/2}} = (\mathbf{D_r^{1/2}}\mathbf{U})\mathbf{D_u}(\mathbf{D_c^{1/2}}\mathbf{V})' = \mathbf{A}\mathbf{D_u}\mathbf{B}'$$

Hence, $\mathbf{A} = \mathbf{D_r^{1/2}U}$ and $\mathbf{B} = \mathbf{D_c^{1/2}V}$.

The default row coordinates are $\mathbf{D_r^{-1}AD_u}$, and the default column coordinates are $\mathbf{D_c^{-1}BD_u}$. Typically the first two columns of $\mathbf{D_r^{-1}AD_u}$ and $\mathbf{D_c^{-1}BD_u}$ are plotted to display graphically associations between the row and column categories. The plot consists of two overlaid plots, one for rows and one for columns. The row points are row profiles, rescaled so that distances between profiles can be displayed as ordinary Euclidean distances, then orthogonally rotated to a principal axes orientation. The column points are column profiles, rescaled so that distances between profiles can be displayed as ordinary Euclidean distances, then orthogonally rotated to a principal axes orientation. Distances between row points and other row points have meaning. Distances between column points and other column points have meaning. However, distances between column points and row points are not interpretable.

The PROFILE=, ROW=, and COLUMN= Options

The PROFILE=, ROW=, and COLUMN= options standardize the coordinates before they are displayed and placed in the output data set. The options PROFILE=BOTH, PROFILE=ROW, and PROFILE=COLUMN provide the standardizations that are typically used in correspondence analysis. There are six choices each for row and column coordinates. However, most of the combinations of the ROW= and COLUMN= options are not useful. The ROW= and COLUMN= options are provided for completeness, but they are not intended for general use.

ROW=	Matrix Formula
A	$\mathbf{A}$
AD	$\mathbf{AD_u}$
DA	$\mathbf{D_r^{-1}A}$
DAD	$\mathbf{D_r^{-1}AD_u}$
DAD1/2	$\mathbf{D_r^{-1}AD_u^{1/2}}$
DAID1/2	$\mathbf{D_r^{-1}A(I+D_u)^{1/2}}$
COLUMN=	**Matrix Formula**
B	$\mathbf{B}$
BD	$\mathbf{BD_u}$
DB	$\mathbf{D_c^{-1}B}$
DBD	$\mathbf{D_c^{-1}BD_u}$
DBD1/2	$\mathbf{D_c^{-1}BD_u^{1/2}}$
DBID1/2	$\mathbf{D_c^{-1}B(I+D_u)^{1/2}}$

When PROFILE=ROW (ROW=DAD and COLUMN=DB), the row coordinates $\mathbf{D_r^{-1}AD_u}$ and column coordinates $\mathbf{D_c^{-1}B}$ provide a correspondence analysis based on the row profile matrix. The row profile (conditional probability) matrix is defined as $\mathbf{R} = \mathbf{D_r^{-1}P} = \mathbf{D_r^{-1}AD_uB'}$. The elements of each row of $\mathbf{R}$ sum to one. Each (i, j) element of $\mathbf{R}$ contains the observed probability of being in column

j given membership in row i. The "principal" row coordinates $\mathbf{D_r^{-1}AD_u}$ and "standard" column coordinates $\mathbf{D_c^{-1}B}$ provide a decomposition of $\mathbf{D_r^{-1}AD_uB'D_c^{-1}} = \mathbf{D_r^{-1}PD_c^{-1}} = \mathbf{RD_c^{-1}}$. Since $\mathbf{D_r^{-1}AD_u} = \mathbf{RD_c^{-1}B}$, the row coordinates are weighted centroids of the column coordinates. Each column point, with coordinates scaled to standard coordinates, defines a vertex in $(n_c - 1)$-dimensional space. All of the principal row coordinates are located in the space defined by the standard column coordinates. Distances among row points have meaning, but distances among column points and distances between row and column points are not interpretable.

The option PROFILE=COLUMN can be described as applying the PROFILE=ROW formulas to the transpose of the contingency table. When PROFILE=COLUMN (ROW=DA and COLUMN=DBD), the principal column coordinates $\mathbf{D_c^{-1}BD_u}$ are weighted centroids of the standard row coordinates $\mathbf{D_r^{-1}A}$. Each row point, with coordinates scaled to standard coordinates, defines a vertex in $(n_r - 1)$-dimensional space. All of the principal column coordinates are located in the space defined by the standard row coordinates. Distances among column points have meaning, but distances among row points and distances between row and column points are not interpretable.

The usual sets of coordinates are given by the default PROFILE=BOTH (ROW=DAD and COLUMN=DBD). All of the summary statistics, such as the squared cosines and contributions to inertia, apply to these two sets of points. One advantage to using these coordinates is that both sets ($\mathbf{D_r^{-1}AD_u}$ and $\mathbf{D_c^{-1}BD_u}$) are postmultiplied by the diagonal matrix $\mathbf{D_u}$, which has diagonal values that are all less than or equal to one. When $\mathbf{D_u}$ is a part of the definition of only one set of coordinates, that set forms a tight cluster near the centroid whereas the other set of points is more widely dispersed. Including $\mathbf{D_u}$ in both sets makes a better graphical display. However, care must be taken in interpreting such a plot. No correct interpretation of distances between row points and column points can be made.

Another property of this choice of coordinates concerns the geometry of distances between points within each set. The default row coordinates can be decomposed into $\mathbf{D_r^{-1}AD_u} = \mathbf{D_r^{-1}AD_uB'D_c^{-1}B} = (\mathbf{D_r^{-1}P})(\mathbf{D_c^{-1/2}})(\mathbf{D_c^{-1/2}B})$. The row coordinates are row profiles ($\mathbf{D_r^{-1}P}$), rescaled by $\mathbf{D_c^{-1/2}}$ (rescaled so that distances between profiles are transformed from a chi-square metric to a Euclidean metric), then orthogonally rotated (with $\mathbf{D_c^{-1/2}B}$) to a principal axes orientation. Similarly, the column coordinates are column profiles rescaled to a Euclidean metric and orthogonally rotated to a principal axes orientation.

The rationale for computing distances between row profiles using the non-Euclidean chi-square metric is as follows. Each row of the contingency table can be viewed as a realization of a multinomial distribution conditional on its row marginal frequency. The null hypothesis of row and column independence is equivalent to the hypothesis of homogeneity of the row profiles. A significant chi-square statistic is geometrically interpreted as a significant deviation of the row profiles from their centroid, $\mathbf{c}'$. The chi-square metric is the Mahalanobis metric between row profiles based on their estimated covariance matrix under the homogeneity assumption (Greenacre and Hastie 1987). A parallel argument can be made for the column profiles.

When ROW=DAD1/2 and COLUMN=DBD1/2 (Gifi 1990; van der Heijden and de Leeuw 1985), the row coordinates $\mathbf{D_r^{-1}AD_u^{1/2}}$ and column coordinates $\mathbf{D_c^{-1}BD_u^{1/2}}$ are a decomposition of $\mathbf{D_r^{-1}PD_c^{-1}}$.

In all of the preceding pairs, distances between row and column points are not meaningful. This prompted Carroll, Green, and Schaffer (1986) to propose that row coordinates $\mathbf{D_r^{-1}A(I+D_u)^{1/2}}$ and column coordinates $\mathbf{D_c^{-1}B(I+D_u)^{1/2}}$ be used. These coordinates are (except for a constant scaling) the coordinates from a multiple correspondence analysis of a Burt table created from two categorical variables. This standardization is available with ROW=DAID1/2 and COLUMN=DBID1/2. However, this approach has been criticized on both theoretical and empirical grounds by Greenacre (1989). The Carroll, Green, and Schaffer standardization relies on the assumption that the chi-square metric is an appropriate metric for measuring the distance between the columns of a bivariate indicator matrix. See the section "Types of Tables Used as Input" on page 961 for a description of indicator matrices. Greenacre (1989) showed that this assumption cannot be justified.

The MCA Option

MCA= option) The MCA option performs a multiple correspondence analysis (MCA). This option requires a Burt table. You can specify the MCA option with a table created from a design matrix with fuzzy coding schemes as long as every row of every partition of the design matrix has the same marginal sum. For example, each row of each partition could contain the probabilities that the observation is a member of each level. Then the Burt table constructed from this matrix no longer contains all integers, and the diagonal partitions are no longer diagonal matrices, but MCA is still valid.

A TABLES statement with a single variable list creates a Burt table. Thus, you can always specify the MCA option with this type of input. If you use the MCA option when reading an existing table with a VAR statement, you must ensure that the table is a Burt table.

If you perform MCA on a table that is not a Burt table, the results of the analysis are invalid. If the table is not symmetric, or if the sums of all elements in each diagonal partition are not equal, PROC CORRESP displays an error message and quits.

A subset of the columns of a Burt table is not necessarily a Burt table, so in MCA it is not appropriate to designate arbitrary columns as supplementary. You can, however, designate all columns from one or more categorical variables as supplementary.

The results of a multiple correspondence analysis of a Burt table $\mathbf{Z'Z}$ are the same as the column results from a simple correspondence analysis of the binary (or fuzzy) matrix $\mathbf{Z}$. Multiple correspondence analysis is not a simple correspondence analysis of the Burt table. It is not appropriate to perform a simple correspondence analysis of a Burt table. The MCA option is based on $\mathbf{P = BD_u^2B'}$, whereas a simple correspondence analysis of the Burt table would be based on $\mathbf{P = BD_uB'}$.

Since the rows and columns of the Burt table are the same, no row information is displayed or written to the output data sets. The resulting inertias and the default (COLUMN=DBD) column coordinates are the appropriate inertias and coordinates for an MCA. The supplementary column coordinates, cosines, and quality of repre-

sentation formulas for MCA differ from the simple correspondence analysis formulas because the design matrix column profiles and left singular vectors are not available.

The following statements create a Burt table and perform a multiple correspondence analysis:

```
proc corresp data=Neighbor observed short mca;
   tables Hair Height Sex Age;
run;
```

Both the rows and the columns have the same nine categories (Blond, Brown, White, Short, Tall, Female, Male, Old, and Young).

MCA Adjusted Inertias

The usual principal inertias of a Burt Table constructed from m categorical variables in MCA are the eigenvalues u_k from $\mathbf{D_u^2}$. The problem with these inertias is that they provide a pessimistic indication of fit. Benzécri (1979) proposed the following inertia adjustment, which is also described by Greenacre (1984, p. 145):

$$\left(\frac{m}{m-1}\right)^2 \times \left(u_k - \frac{1}{m}\right)^2 \qquad \text{for } u_k > \frac{1}{m}$$

The Benzécri adjustment is available with the BENZECRI option.

Greenacre (1994, p. 156) argues that the Benzécri adjustment overestimates the quality of fit. Greenacre proposes instead the following inertia adjustment:

$$\left(\frac{m}{m-1}\right)^2 \times \left(\sqrt{u_k} - \frac{1}{m}\right)^2 \qquad \text{for } \sqrt{u_k} > \frac{1}{m}$$

The Greenacre adjustment is available with the GREENACRE option.

Ordinary unadjusted inertias are printed by default with MCA when neither the BENZECRI nor the GREENACRE option is specified. However, the unadjusted inertias are not printed by default when either the BENZECRI or the GREENACRE option is specified. To display both adjusted and unadjusted inertias, specify the UNADJUSTED option in addition to the relevant adjusted inertia option (BENZECRI, GREENACRE, or both).

Supplementary Rows and Columns

Supplementary rows and columns are represented as points in the joint row and column space, but they are not used when determining the locations of the other active rows and columns of the table. The formulas that are used to compute coordinates for the supplementary rows and columns depend on the PROFILE= option or on the ROW= and COLUMN= options. Let $\mathbf{S_o}$ be the matrix with rows that contain the supplementary observations and $\mathbf{S_v}$ be a matrix with rows that contain the supplementary variables. Note that $\mathbf{S_v}$ is defined to be the transpose of the supplementary variable partition of the table. Let $\mathbf{R_s} = \text{diag}(\mathbf{S_o 1})^{-1}\mathbf{S_o}$ be the supplementary observation profile matrix and $\mathbf{C_s} = \text{diag}(\mathbf{S_v 1})^{-1}\mathbf{S_v}$ be the supplementary variable profile matrix. Note that the notation $\text{diag}(\cdot)^{-1}$ means to convert the vector to a diagonal matrix, then invert the diagonal matrix. The coordinates for the supplementary observations and variables are as follows.

ROW=	Matrix Formula
A	$\frac{1}{f}\mathbf{S_o}\mathbf{D_c^{-1}}\mathbf{B}\mathbf{D_u^{-1}}$
AD	$\frac{1}{f}\mathbf{S_o}\mathbf{D_c^{-1}}\mathbf{B}$
DA	$\mathbf{R_s}\mathbf{D_c^{-1}}\mathbf{B}\mathbf{D_u^{-1}}$
DAD	$\mathbf{R_s}\mathbf{D_c^{-1}}\mathbf{B}$
DAD1/2	$\mathbf{R_s}\mathbf{D_c^{-1}}\mathbf{B}\mathbf{D_u^{-1/2}}$
DAID1/2	$\mathbf{R_s}\mathbf{D_c^{-1}}\mathbf{B}\mathbf{D_u^{-1}}(\mathbf{I}+\mathbf{D_u})^{\mathbf{1/2}}$

COLUMN=	Matrix Formula
B	$\frac{1}{f}\mathbf{S_v}\mathbf{D_r^{-1}}\mathbf{A}\mathbf{D_u^{-1}}$
BD	$\frac{1}{f}\mathbf{S_v}\mathbf{D_r^{-1}}\mathbf{A}$
DB	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{A}\mathbf{D_u^{-1}}$
DBD	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{A}$
DBD1/2	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{A}\mathbf{D_u^{-1/2}}$
DBID1/2	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{A}\mathbf{D_u^{-1}}(\mathbf{I}+\mathbf{D_u})^{\mathbf{1/2}}$

MCA COLUMN=	Matrix Formula
B	not allowed
BD	not allowed
DB	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{B}\mathbf{D_u^{-2}}$
DBD	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{B}\mathbf{D_u^{-1}}$
DBD1/2	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{B}\mathbf{D_u^{-3/2}}$
DBID1/2	$\mathbf{C_s}\mathbf{D_r^{-1}}\mathbf{B}\mathbf{D_u^{-2}}(\mathbf{I}+\mathbf{D_u})^{\mathbf{1/2}}$

Statistics that Aid Interpretation

The partial contributions to inertia, squared cosines, quality of representation, inertia, and mass provide additional information about the coordinates. These statistics are displayed by default. Include the SHORT or NOPRINT option in the PROC CORRESP statement to avoid having these statistics displayed.

These statistics pertain to the default PROFILE=BOTH coordinates, no matter what values you specify for the ROW=, COLUMN=, or PROFILE= option. Let $\mathrm{sq}(\cdot)$ be a matrix-valued function denoting elementwise squaring of the argument matrix. Let t be the total inertia (the sum of the elements in $\mathbf{D_u^2}$).

In MCA, let $\mathbf{D_s}$ be the Burt table partition containing the intersection of the supplementary columns and the supplementary rows. The matrix $\mathbf{D_s}$ is a diagonal matrix of marginal frequencies of the supplemental columns of the binary matrix $\mathbf{Z}$. Let p be the number of rows in this design matrix.

Statistic	Matrix Formula
Row partial contributions to inertia	$\mathbf{D_r^{-1}}\text{sq}(\mathbf{A})$
Column partial contributions to inertia	$\mathbf{D_c^{-1}}\text{sq}(\mathbf{B})$
Row squared cosines	$\text{diag}(\text{sq}(\mathbf{AD_u})\mathbf{1})^{-1}\text{sq}(\mathbf{AD_u})$
Column squared cosines	$\text{diag}(\text{sq}(\mathbf{BD_u})\mathbf{1})^{-1}\text{sq}(\mathbf{BD_u})$
Row mass	$\mathbf{r}$
Column mass	$\mathbf{c}$
Row inertia	$\frac{1}{t}\mathbf{D_r^{-1}}\ \text{sq}(\mathbf{AD_u})\mathbf{1}$
Column inertia	$\frac{1}{t}\mathbf{D_c^{-1}}\ \text{sq}(\mathbf{BD_u})\mathbf{1}$
Supplementary row squared cosines	$\text{diag}(\text{sq}(\mathbf{R_s} - \mathbf{1c}')\mathbf{D_c^{-1}}\mathbf{1})^{-1}\text{sq}(\mathbf{R_s D_c^{-1} B})$
Supplementary column squared cosines	$\text{diag}(\text{sq}(\mathbf{C_s} - \mathbf{1r}')\mathbf{D_r^{-1}}\mathbf{1})^{-1}\text{sq}(\mathbf{C_s D_r^{-1} A})$
MCA supplementary column squared cosines	$\mathbf{D_s}(p\mathbf{I} - \mathbf{D_s})^{-1}\ \text{sq}(\mathbf{C_s D_r^{-1} B D_u^{-1}})$

The quality of representation in the DIMENS=n dimensional display of any point is the sum of its squared cosines over only the n dimensions. Inertia and mass are not defined for supplementary points.

A table that summarizes the partial contributions to inertia table is also computed. The points that best explain the inertia of each dimension and the dimension to which each point contributes the most inertia are indicated. The output data set variable names for this table are Best1–Best$n$ (where DIMENS=n) and Best. The Best column contains the dimension number of the largest partial contribution to inertia for each point (the index of the maximum value in each row of $\mathbf{D_r^{-1}}\text{sq}(\mathbf{A})$ or $\mathbf{D_c^{-1}}\text{sq}(\mathbf{B})$).

For each row, the Best1–Best$n$ columns contain either the corresponding value of Best if the point is one of the biggest contributors to the dimension's inertia or 0 if it is not. Specifically, Best1 contains the value of Best for the point with the largest contribution to dimension one's inertia. A cumulative proportion sum is initialized to this point's partial contribution to the inertia of dimension one. If this sum is less than the value for the MININERTIA= option, then Best1 contains the value of Best for the point with the second largest contribution to dimension one's inertia. Otherwise, this point's Best1 is 0. This point's partial contribution to inertia is added to the sum. This process continues for the point with the third largest partial contribution, and so on, until adding a point's contribution to the sum increases the sum beyond the value of the MININERTIA= option. This same algorithm is then used for Best2, and so on.

For example, the following table contains contributions to inertia and the corresponding Best variables. The contribution to inertia variables are proportions that sum to 1 within each column. The first point makes its greatest contribution to the inertia

of dimension two, so Best for point one is set to 2 and Best1–Best3 for point one must all be 0 or 2. The second point also makes its greatest contribution to the inertia of dimension two, so Best for point two is set to 2 and Best1–Best3 for point two must all be 0 or 2, and so on.

Assume MININERTIA=0.8, the default. In dimension one, the largest contribution is 0.41302 for the fourth point, so Best1 is set to 1, the value of Best for the fourth point. Because this value is less than 0.8, the second largest value (0.36456 for point five) is found and its Best1 is set to its Best's value of 1. Because $0.41302 + 0.36456 = 0.77758$ is less than 0.8, the third point (0.0882 at point eight) is found and Best1 is set to 3 since the contribution to dimension 3 for that point is greater than the contribution to dimension 1. This increases the sum of the partial contributions to greater than 0.8, so the remaining Best1 values are all 0.

Contr1	Contr2	Contr3	Best1	Best2	Best3	Best
0.01593	0.32178	0.07565	0	2	2	2
0.03014	0.24826	0.07715	0	2	2	2
0.00592	0.02892	0.02698	0	0	0	2
0.41302	0.05191	0.05773	1	0	0	1
0.36456	0.00344	0.15565	1	0	1	1
0.03902	0.30966	0.11717	0	2	2	2
0.00019	0.01840	0.00734	0	0	0	2
0.08820	0.00527	0.16555	3	0	3	3
0.01447	0.00024	0.03851	0	0	0	3
0.02855	0.01213	0.27827	0	0	3	3

Displayed Output

The display options control the amount of displayed output. By default, the following information is displayed:

- an inertia and chi-square decomposition table including the total inertia, the principal inertias of each dimension (eigenvalues), the singular values (square roots of the eigenvalues), each dimension's percentage of inertia, a horizontal bar chart of the percentages, and the total chi-square with its degrees of freedom and decomposition. The chi-square statistics and degrees of freedom are valid only when the constructed table is an ordinary two-way contingency table.
- the coordinates of the rows and columns on the dimensions
- the mass, relative contribution to the total inertia, and quality of representation in the DIMENS=n dimensional display of each row and column
- the squared cosines of the angles between each axis and a vector from the origin to the point
- the partial contributions of each point to each dimension's inertia
- the Best table, indicators of which points best explain the inertia of each dimension

Specific display options and combinations of options display output as follows.

If you specify the OBSERVED or ALL option and you do not specify PRINT=PERCENT, PROC CORRESP displays

- the contingency table including the row and column marginal frequencies; or with BINARY, the binary table; or the Burt table in MCA
- the supplementary rows
- the supplementary columns

If you specify the OBSERVED or ALL option, with the PRINT=PERCENT or PRINT=BOTH option, PROC CORRESP displays

- the contingency table or Burt table in MCA, scaled to percentages, including the row and column marginal percentages
- the supplementary rows, scaled to percentages
- the supplementary columns, scaled to percentages

If you specify the EXPECTED or ALL option and you do not specify PRINT=PERCENT, PROC CORRESP displays the product of the row marginals and the column marginals divided by the grand frequency of the observed frequency table. For ordinary two-way contingency tables, these are the expected frequencies under the hypothesis of row and column independence.

If you specify the EXPECTED or ALL option with the PRINT=PERCENT or PRINT=BOTH option, PROC CORRESP displays the product of the row marginals and the column marginals divided by the grand frequency of the observed percentages table. For ordinary two-way contingency tables, these are the expected percentages under the hypothesis of row and column independence.

If you specify the DEVIATION or ALL option and you do not specify PRINT=PERCENT, PROC CORRESP displays the observed minus expected frequencies. For ordinary two-way contingency tables, these are the expected frequencies under the hypothesis of row and column independence.

If you specify the DEVIATION or ALL option with the PRINT=PERCENT or PRINT=BOTH option, PROC CORRESP displays the observed minus expected percentages. For ordinary two-way contingency tables, these are the expected percentages under the hypothesis of row and column independence.

If you specify the CELLCHI2 or ALL option and you do not specify PRINT=PERCENT, PROC CORRESP displays contributions to the total chi-square test statistic, including the row and column marginals. The intersection of the marginals contains the total chi-square statistic.

If you specify the CELLCHI2 or ALL option with the PRINT=PERCENT or the PRINT=BOTH option, PROC CORRESP displays contributions to the total chi-square, scaled to percentages, including the row and column marginals.

If you specify the RP or ALL option and you do not specify PRINT=PERCENT, PROC CORRESP displays the row profiles and the supplementary row profiles.

If you specify the RP or ALL option with the PRINT=PERCENT or the PRINT=BOTH option, PROC CORRESP displays the row profiles (scaled to percentages) and the supplementary row profiles (scaled to percentages).

If you specify the CP or ALL option and you do not specify PRINT=PERCENT, PROC CORRESP displays the column profiles and the supplementary column profiles.

If you specify the CP or ALL option with the PRINT=PERCENT or PRINT=BOTH option, PROC CORRESP displays the column profiles (scaled to percentages) and the supplementary column profiles (scaled to percentages).

If you do not specify the NOPRINT option, PROC CORRESP displays the inertia and chi-square decomposition table. This includes the nonzero singular values of the contingency table (or, in MCA, the binary matrix **Z** used to create the Burt table), the nonzero principal inertias (or eigenvalues) for each dimension, the total inertia, the total chi-square, the decomposition of chi-square, the chi-square degrees of freedom (appropriate only when the table is an ordinary two-way contingency table), the percent of the total chi-square and inertia for each dimension, and a bar chart of the percents.

If you specify the MCA option and you do not specify the NOPRINT option, PROC CORRESP displays the adjusted inertias. This includes the nonzero adjusted inertias, percents, cumulative percents, and a bar chart of the percents.

If you do not specify the NOROW, NOPRINT, or MCA option, PROC CORRESP displays the row coordinates and the supplementary row coordinates (displayed when there are supplementary row points).

If you do not specify the NOROW, NOPRINT, MCA, or SHORT option, PROC CORRESP displays

- the summary statistics for the row points including the quality of representation of the row points in the n-dimensional display, the mass, and the relative contributions to inertia
- the quality of representation of the supplementary row points in the n-dimensional display (displayed when there are supplementary row points)
- the partial contributions to inertia for the row points
- the row Best table, indicators of which row points best explain the inertia of each dimension
- the squared cosines for the row points
- the squared cosines for the supplementary row points (displayed when there are supplementary row points)

If you do not specify the NOCOLUMN or NOPRINT option, PROC CORRESP displays the column coordinates and the supplementary column coordinates (displayed when there are supplementary column points).

If you do not specify the NOCOLUMN, NOPRINT, or SHORT option, PROC CORRESP displays

- the summary statistics for the column points including the quality of representation of the column points in the n-dimensional display, the mass, and the relative contributions to inertia for the supplementary column points
- the quality of representation of the supplementary column points in the n-dimensional display (displayed when there are supplementary column points)
- the partial contributions to inertia for the column points
- the column Best table, indicators of which column points best explain the inertia of each dimension
- the squared cosines for the column points
- the squared cosines for the supplementary column points

ODS Table Names

PROC CORRESP assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 24.8. ODS Tables Produced in PROC CORRESP

ODS Table Name	Description	Option
AdjInGreenacre	Greenacre Inertia Adjustment	GREENACRE
AdjInBenzecri	Benzécri Inertia Adjustment	BENZECRI
Binary	Binary table	OBSERVED, BINARY
BinaryPct	Binary table percents	OBSERVED, BINARY *
Burt	Burt table	OBSERVED, MCA
BurtPct	Burt table percents	OBSERVED, MCA *
CellChiSq	Contributions to Chi Square	CELLCHI2
CellChiSqPct	Contributions, pcts	CELLCHI2 *
ColBest	Col best indicators	default
ColContr	Col contributions to inertia	default
ColCoors	Col coordinates	default
ColProfiles	Col profiles	CP
ColProfilesPct	Col profiles, pcts	CP *
ColQualMassIn	Col quality, mass, inertia	default
ColSqCos	Col squared cosines	default
DF	DF, Chi Square (not displayed)	default
Deviations	Observed - expected freqs	DEVIATIONS
DeviationsPct	Observed - expected pcts	DEVIATIONS *
Expected	Expected frequencies	EXPECTED
ExpectedPct	Expected percents	EXPECTED *
Inertias	Inertia decomposition table	default

Table 24.8. (continued)

ODS Table Name	Description	Option
Observed	Observed frequencies	OBSERVED
ObservedPct	Observed percents	OBSERVED *
RowBest	Row best indicators	default
RowContr	Row contributions to inertia	default
RowCoors	Row coordinates	default
RowProfiles	Row profiles	RP
RowProfilesPct	Row profiles, pcts	RP *
RowQualMassIn	Row quality, mass, inertia	default
RowSqCos	Row squared cosines	default
SupColCoors	Supp col coordinates	default
SupColProfiles	Supp col profiles	CP
SupColProfilesPct	Supp col profiles, pcts	CP *
SupColQuality	Supp col quality	default
SupCols	Supplementary col freq	OBSERVED
SupColsPct	Supplementary col pcts	OBSERVED *
SupColSqCos	Supp col squared cosines	default
SupRows	Supplementary row freqs	OBSERVED
SupRowCoors	Supp row coordinates	default
SupRowProfiles	Supp row profiles	RP
SupRowProfilesPct	Supp row profiles, pcts	RP *
SupRowQuality	Supp row quality	default
SupRowsPct	Supplementary row pcts	OBSERVED *
SupRowSqCos	Supp row squared cosines	default

*Percents are displayed when you specify the PRINT=PERCENT or PRINT=BOTH option.

Examples

Example 24.1. Simple Correspondence Analysis of Cars and Their Owners

In this example, PROC CORRESP creates a contingency table from categorical data and performs a simple correspondence analysis. The data are from a sample of individuals who were asked to provide information about themselves and their cars. The questions included origin of the car (American, Japanese, European) and family status (single, married, single and living with children, and married living with children). These data are used again in Example 24.2.

The first steps read the input data and assign formats. PROC CORRESP is used to perform the simple correspondence analysis. The ALL option displays all tables including the contingency table, chi-square information, profiles, and all results of the correspondence analysis. The OUTC= option creates an output coordinate data set. The TABLES statement specifies the row and column categorical variables. The %PLOTIT macro is used to plot the results.

Normally, you only need to tell the %PLOTIT macro the name of the input data set, DATA=Coor, and the type of analysis performed on the data, DATATYPE=CORRESP.

The following statements produce Output 24.1.1:

```
title 'Car Owners and Car Origin';

proc format;
   value Origin  1 = 'American' 2 = 'Japanese' 3 = 'European';
   value Size    1 = 'Small'    2 = 'Medium'   3 = 'Large';
   value Type    1 = 'Family'   2 = 'Sporty'   3 = 'Work';
   value Home    1 = 'Own'      2 = 'Rent';
   value Sex     1 = 'Male'     2 = 'Female';
   value Income  1 = '1 Income' 2 = '2 Incomes';
   value Marital 1 = 'Single with Kids' 2 = 'Married with Kids'
                 3 = 'Single'           4 = 'Married';
   run;

data Cars;
   missing a;
   input (Origin Size Type Home Income Marital Kids Sex) (1.) @@;
   * Check for End of Line;
   if n(of Origin -- Sex) eq 0 then do; input; return; end;
   marital = 2 * (kids le 0) + marital;
   format Origin Origin. Size Size. Type Type. Home Home.
          Sex Sex. Income Income. Marital Marital.;
   output;
   datalines;
13111221212111012111220113121101121122112211212113112212321122221221220 1
12112202312122123221110112212202212111012211210213111221112111011231110 1
21111211321122312112220222112211131112313121110232112222322122022122110 1
12212202212122021121220122112202112211013211220221311211133312261222211 01
1212110231AA22023112212113112112121220212212202112111022222110212121221
21121101221122221221110131311211312122012111221212111221221122222111211
22111101111222012221220113121101312122011311222213111201213111022111211
12111221221112112111220132112231131122111311221223211013121220221221101
13321101121222023331110221311102321112212131222221221211111122212111211
13311201121211221211221222122202213122222212110111112202221122011311221
21111201223222012122110221321101113122012121220121112211233122023331220
22212201211122021211220122112211221222022221221131112201211111011221221
11222201113111222121220232221102122211012122110133321101223211013221210
22322201311122011221110121121102211211021221110222112202111122011211211
11112202212111011331112232211112222121022221110121212202121122123211202
13311101131122112132220121312212111122122112202133122021212111212121.2212
121122.22121210233112212222121011311122121211102211122112121110121212101
31121202223122111211221121121131222122121311221222112202222211013121220 2
21312221131122121211222211312222122122021311122112121122121122122110210 2
13112221121122022122210122311201211122121211102223122111311222121111102
21211101211220221333122223111221213122121211102222112202111122011211211
11121202312111011111222121211101221122022132110122121112212122011211111 2
21221102211111012222110112111211212211012212223222112221221122122110210 2
21312211221111021212120113211012221110232111102212211012112220121212202
22111201121122012122110121121102221122111212110111111221212122111112212 01
21112212212211121211222111112231213211011312110112112222211122022212110 2
22121101212211022122110231211101212222012112110112112222211122221221102
21212202122212011311220212112221212111011311110112311221211122011311110 1
22111221132121013121221112121101122211012211222212312202312122311221220 2
```

```
3112110121311101312211021122110211312202131222012221110221212212213122O2
131.2252322111012221222113111241221122022112111213122022122220122122201
2121110113112202213122022211221232212101212222022231221212112212211111112
2111111212112212212201113122122131220222112222211122011311110112312211
2112220132212201212112113121222122201221122011112220112111101223111112
3121110212312201222121101211112112.2211022211221212112212221111012111210l
1212110132112222121112222321112112112110121321101113111012221220121312201
2132110122122202221211101321122121111220221121101122211021122110213112212
2121220112111221312211011212110222122202121211O1
;

*---Perform Simple Correspondence Analysis---;
proc corresp all data=Cars outc=Coor;
   tables Marital, Origin;
   run;

*---Plot the Simple Correspondence Analysis Results---;
%plotit(data=Coor, datatype=corresp)
```

Correspondence analysis locates all the categories in a Euclidean space. The first two dimensions of this space are plotted to examine the associations among the categories. Since the smallest dimension of this table is three, there is no loss of information when only two dimensions are plotted. The plot should be thought of as two different overlaid plots, one for each categorical variable. Distances between points within a variable have meaning, but distances between points from different variables do not.

Output 24.1.1. Simple Correspondence Analysis of a Contingency Table

Car Owners and Car Origin

The CORRESP Procedure

Contingency Table

	American	European	Japanese	Sum
Married	37	14	51	102
Married with Kids	52	15	44	111
Single	33	15	63	111
Single with Kids	6	1	8	15
Sum	128	45	166	339

Chi-Square Statistic Expected Values

	American	European	Japanese
Married	38.5133	13.5398	49.9469
Married with Kids	41.9115	14.7345	54.3540
Single	41.9115	14.7345	54.3540
Single with Kids	5.6637	1.9912	7.3451

Observed Minus Expected Values

	American	European	Japanese
Married	-1.5133	0.4602	1.0531
Married with Kids	10.0885	0.2655	-10.3540
Single	-8.9115	0.2655	8.6460
Single with Kids	0.3363	-0.9912	0.6549

Contributions to the Total Chi-Square Statistic

	American	European	Japanese	Sum
Married	0.05946	0.01564	0.02220	0.09730
Married with Kids	2.42840	0.00478	1.97235	4.40553
Single	1.89482	0.00478	1.37531	3.27492
Single with Kids	0.01997	0.49337	0.05839	0.57173
Sum	4.40265	0.51858	3.42825	8.34947

Car Owners and Car Origin

The CORRESP Procedure

Row Profiles

	American	European	Japanese
Married	0.362745	0.137255	0.500000
Married with Kids	0.468468	0.135135	0.396396
Single	0.297297	0.135135	0.567568
Single with Kids	0.400000	0.066667	0.533333

Column Profiles

	American	European	Japanese
Married	0.289063	0.311111	0.307229
Married with Kids	0.406250	0.333333	0.265060
Single	0.257813	0.333333	0.379518
Single with Kids	0.046875	0.022222	0.048193

Car Owners and Car Origin

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	19 38 57 76 95
0.15122	0.02287	7.75160	92.84	92.84	**********
0.04200	0.00176	0.59787	7.16	100.00	**
Total	0.02463	8.34947	100.00		

Degrees of Freedom = 6

Row Coordinates

	Dim1	Dim2
Married	-0.0278	0.0134
Married with Kids	0.1991	0.0064
Single	-0.1716	0.0076
Single with Kids	-0.0144	-0.1947

Summary Statistics for the Row Points

	Quality	Mass	Inertia
Married	1.0000	0.3009	0.0117
Married with Kids	1.0000	0.3274	0.5276
Single	1.0000	0.3274	0.3922
Single with Kids	1.0000	0.0442	0.0685

Car Owners and Car Origin

The CORRESP Procedure

Partial Contributions to Inertia for the Row Points

	Dim1	Dim2
Married	0.0102	0.0306
Married with Kids	0.5678	0.0076
Single	0.4217	0.0108
Single with Kids	0.0004	0.9511

Indices of the Coordinates that Contribute Most to Inertia for the Row Points

	Dim1	Dim2	Best
Married	0	0	2
Married with Kids	1	0	1
Single	1	0	1
Single with Kids	0	2	2

Squared Cosines for the Row Points

	Dim1	Dim2
Married	0.8121	0.1879
Married with Kids	0.9990	0.0010
Single	0.9980	0.0020
Single with Kids	0.0054	0.9946

Car Owners and Car Origin

The CORRESP Procedure

Column Coordinates

	Dim1	Dim2
American	0.1847	-0.0166
European	0.0013	0.1073
Japanese	-0.1428	-0.0163

Summary Statistics for the Column Points

	Quality	Mass	Inertia
American	1.0000	0.3776	0.5273
European	1.0000	0.1327	0.0621
Japanese	1.0000	0.4897	0.4106

Car Owners and Car Origin

The CORRESP Procedure

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
American	0.5634	0.0590
European	0.0000	0.8672
Japanese	0.4366	0.0737

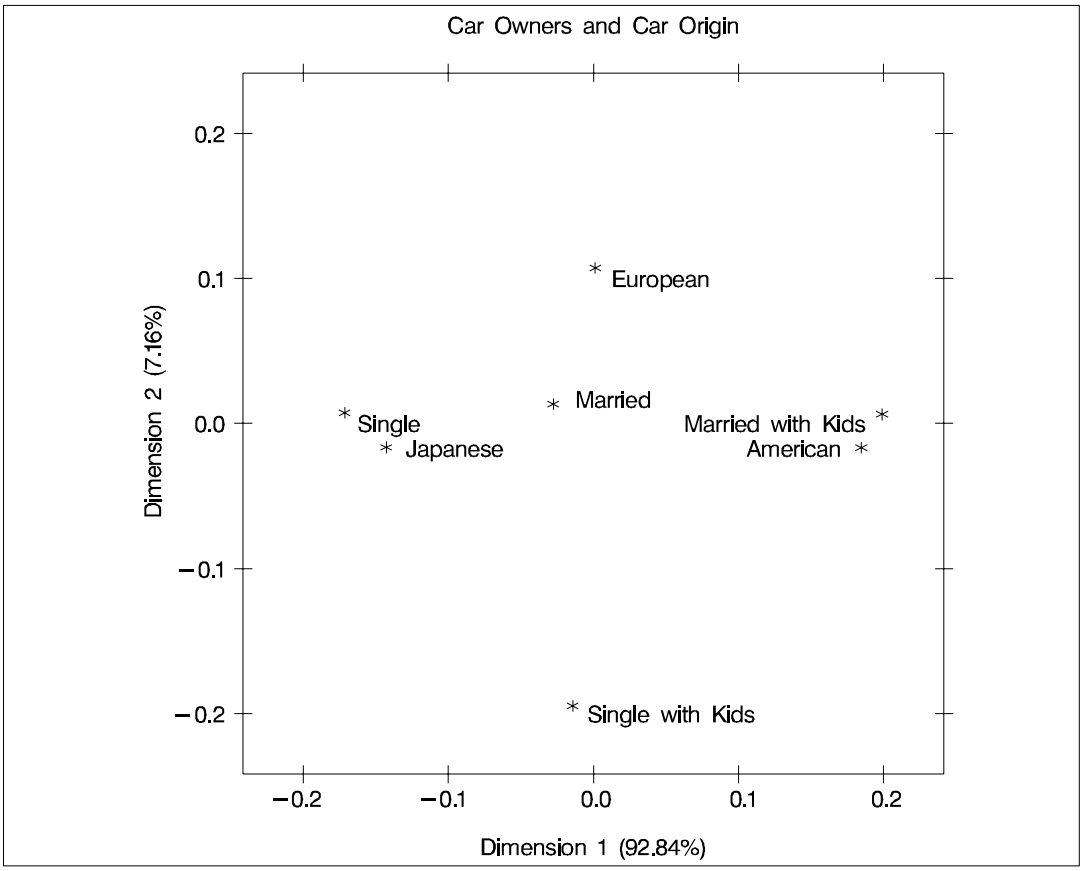
Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
American	1	0	1
European	0	2	2
Japanese	1	0	1

Squared Cosines for the Column Points

	Dim1	Dim2
American	0.9920	0.0080
European	0.0001	0.9999
Japanese	0.9871	0.0129

Output 24.1.2. Plot of Simple Correspondence Analysis of a Contingency Table



To interpret the plot, start by interpreting the row points separately from the column points. The European point is near and to the left of the centroid, so it makes a relatively small contribution to the chi-square statistic (because it is near the centroid), it contributes almost nothing to the inertia of dimension one (since its coordinate on dimension one has a small absolute value relative to the other column points), and it makes a relatively large contribution to the inertia of dimension two (since its coordinate on dimension two has a large absolute value relative to the other column points). Its squared cosines for dimension one and two, approximately 0 and 1, respectively, indicate that its position is almost completely determined by its location on dimension two. Its quality of display is 1.0, indicating perfect quality, since the table is two-dimensional after the centering. The American and Japanese points are far from the centroid, and they lie along dimension one. They make relatively large contributions to the chi-square statistic and the inertia of dimension one. The horizontal dimension seems to be largely determined by Japanese versus American car ownership.

In the row points, the Married point is near the centroid, and the Single with Kids point has a small coordinate on dimension one that is near zero. The horizontal dimension seems to be largely determined by the Single versus the Married with Kids points. The two interpretations of dimension one show the association with being Married with Kids and owning an American car, and being single and owning

a Japanese car. The fact that the Married with Kids point is close to the American point and the fact that the Japanese point is near the Single point should be ignored. Distances between row and column points are not defined. The plot shows that more people who are married with kids than you would expect if the rows and columns were independent drive an American car, and more people who are single than you would expect if the rows and columns were independent drive a Japanese car.

Example 24.2. Multiple Correspondence Analysis of Cars and Their Owners

In this example, PROC CORRESP creates a Burt table from categorical data and performs a multiple correspondence analysis. The data are from a sample of individuals who were asked to provide information about themselves and their cars. The questions included origin of the car (American, Japanese, European), size of car (Small, Medium, Large), type of car (Family, Sporty, Work Vehicle), home ownership (Owns, Rents), marital/family status (single, married, single and living with children, and married living with children), and sex (Male, Female).

The data are read and formats assigned in a previous step, displayed in Example 24.1. The variables used in this example are Origin, Size, Type, Income, Home, Marital, and Sex. MCA specifies multiple correspondence analysis, OBSERVED displays the Burt table, and the OUTC= option creates an output coordinate data set. The TABLES statement with only a single variable list and no comma creates the Burt table. The %PLOTIT macro is used to plot the results with vertical and horizontal reference lines.

The data used to produce Output 24.2.1 and Output 24.2.2 can be found in Example 24.1.

```
title 'MCA of Car Owners and Car Attributes';

*---Perform Multiple Correspondence Analysis---;
proc corresp mca observed data=Cars outc=Coor;
   tables Origin Size Type Income Home Marital Sex;
run;

*---Plot the Multiple Correspondence Analysis Results---;
%plotit(data=Coor, datatype=corresp, href=0, vref=0)
```

Output 24.2.1. Multiple Correspondence Analysis of a Burt Table

MCA of Car Owners and Car Attributes

The CORRESP Procedure

Burt Table

	American	European	Japanese	Large	Medium	Small	Family	Sporty	Work	1 Income
American	125	0	0	36	60	29	81	24	20	58
European	0	44	0	4	20	20	17	23	4	18
Japanese	0	0	165	2	61	102	76	59	30	74
Large	36	4	2	42	0	0	30	1	11	20
Medium	60	20	61	0	141	0	89	39	13	57
Small	29	20	102	0	0	151	55	66	30	73
Family	81	17	76	30	89	55	174	0	0	69
Sporty	24	23	59	1	39	66	0	106	0	55
Work	20	4	30	11	13	30	0	0	54	26
1 Income	58	18	74	20	57	73	69	55	26	150
2 Incomes	67	26	91	22	84	78	105	51	28	0
Own	93	38	111	35	106	101	130	71	41	80
Rent	32	6	54	7	35	50	44	35	13	70
Married	37	13	51	9	42	50	50	35	16	10
Married with Kids	50	15	44	21	51	37	79	12	18	27
Single	32	15	62	11	40	58	35	57	17	99
Single with Kids	6	1	8	1	8	6	10	2	3	14
Female	58	21	70	17	70	62	83	44	22	47
Male	67	23	95	25	71	89	91	62	32	103

Burt Table

	2 Incomes	Own	Rent	Married	Married with Kids	Single	Single with Kids	Female	Male
American	67	93	32	37	50	32	6	58	67
European	26	38	6	13	15	15	1	21	23
Japanese	91	111	54	51	44	62	8	70	95
Large	22	35	7	9	21	11	1	17	25
Medium	84	106	35	42	51	40	8	70	71
Small	78	101	50	50	37	58	6	62	89
Family	105	130	44	50	79	35	10	83	91
Sporty	51	71	35	35	12	57	2	44	62
Work	28	41	13	16	18	17	3	22	32
1 Income	0	80	70	10	27	99	14	47	103
2 Incomes	184	162	22	91	82	10	1	102	82
Own	162	242	0	76	106	52	8	114	128
Rent	22	0	92	25	3	57	7	35	57
Married	91	76	25	101	0	0	0	53	48
Married with Kids	82	106	3	0	109	0	0	48	61
Single	10	52	57	0	0	109	0	35	74
Single with Kids	1	8	7	0	0	0	15	13	2
Female	102	114	35	53	48	35	13	149	0
Male	82	128	57	48	61	74	2	0	185

MCA of Car Owners and Car Attributes

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	4 8 12 16 20
0.56934	0.32415	970.77	18.91	18.91	**********
0.48352	0.23380	700.17	13.64	32.55	**********
0.42716	0.18247	546.45	10.64	43.19	**********
0.41215	0.16987	508.73	9.91	53.10	**********
0.38773	0.15033	450.22	8.77	61.87	**********
0.38520	0.14838	444.35	8.66	70.52	**********
0.34066	0.11605	347.55	6.77	77.29	********
0.32983	0.10879	325.79	6.35	83.64	********
0.31517	0.09933	297.47	5.79	89.43	*******
0.28069	0.07879	235.95	4.60	94.03	******
0.26115	0.06820	204.24	3.98	98.01	*****
0.18477	0.03414	102.24	1.99	100.00	**
Total	1.71429	5133.92	100.00		

Degrees of Freedom = 324

MCA of Car Owners and Car Attributes

The CORRESP Procedure

Column Coordinates

	Dim1	Dim2
American	-0.4035	0.8129
European	-0.0568	-0.5552
Japanese	0.3208	-0.4678
Large	-0.6949	1.5666
Medium	-0.2562	0.0965
Small	0.4326	-0.5258
Family	-0.4201	0.3602
Sporty	0.6604	-0.6696
Work	0.0575	0.1539
1 Income	0.8251	0.5472
2 Incomes	-0.6727	-0.4461
Own	-0.3887	-0.0943
Rent	1.0225	0.2480
Married	-0.4169	-0.7954
Married with Kids	-0.8200	0.3237
Single	1.1461	0.2930
Single with Kids	0.4373	0.8736
Female	-0.3365	-0.2057
Male	0.2710	0.1656

Summary Statistics for the Column Points

	Quality	Mass	Inertia
American	0.4925	0.0535	0.0521
European	0.0473	0.0188	0.0724
Japanese	0.3141	0.0706	0.0422
Large	0.4224	0.0180	0.0729
Medium	0.0548	0.0603	0.0482
Small	0.3825	0.0646	0.0457
Family	0.3330	0.0744	0.0399
Sporty	0.4112	0.0453	0.0569
Work	0.0052	0.0231	0.0699
1 Income	0.7991	0.0642	0.0459
2 Incomes	0.7991	0.0787	0.0374
Own	0.4208	0.1035	0.0230
Rent	0.4208	0.0393	0.0604
Married	0.3496	0.0432	0.0581
Married with Kids	0.3765	0.0466	0.0561
Single	0.6780	0.0466	0.0561
Single with Kids	0.0449	0.0064	0.0796
Female	0.1253	0.0637	0.0462
Male	0.1253	0.0791	0.0372

MCA of Car Owners and Car Attributes

The CORRESP Procedure

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2
American	0.0268	0.1511
European	0.0002	0.0248
Japanese	0.0224	0.0660
Large	0.0268	0.1886
Medium	0.0122	0.0024
Small	0.0373	0.0764
Family	0.0405	0.0413
Sporty	0.0610	0.0870
Work	0.0002	0.0023
1 Income	0.1348	0.0822
2 Incomes	0.1099	0.0670
Own	0.0482	0.0039
Rent	0.1269	0.0103
Married	0.0232	0.1169
Married with Kids	0.0967	0.0209
Single	0.1889	0.0171
Single with Kids	0.0038	0.0209
Female	0.0223	0.0115
Male	0.0179	0.0093

MCA of Car Owners and Car Attributes

The CORRESP Procedure

Indices of the Coordinates that Contribute Most to Inertia for the Column Points

	Dim1	Dim2	Best
American	0	2	2
European	0	0	2
Japanese	0	2	2
Large	0	2	2
Medium	0	0	1
Small	0	2	2
Family	2	0	2
Sporty	2	2	2
Work	0	0	2
1 Income	1	1	1
2 Incomes	1	1	1
Own	1	0	1
Rent	1	0	1
Married	0	2	2
Married with Kids	1	0	1
Single	1	0	1
Single with Kids	0	0	2
Female	0	0	1
Male	0	0	1

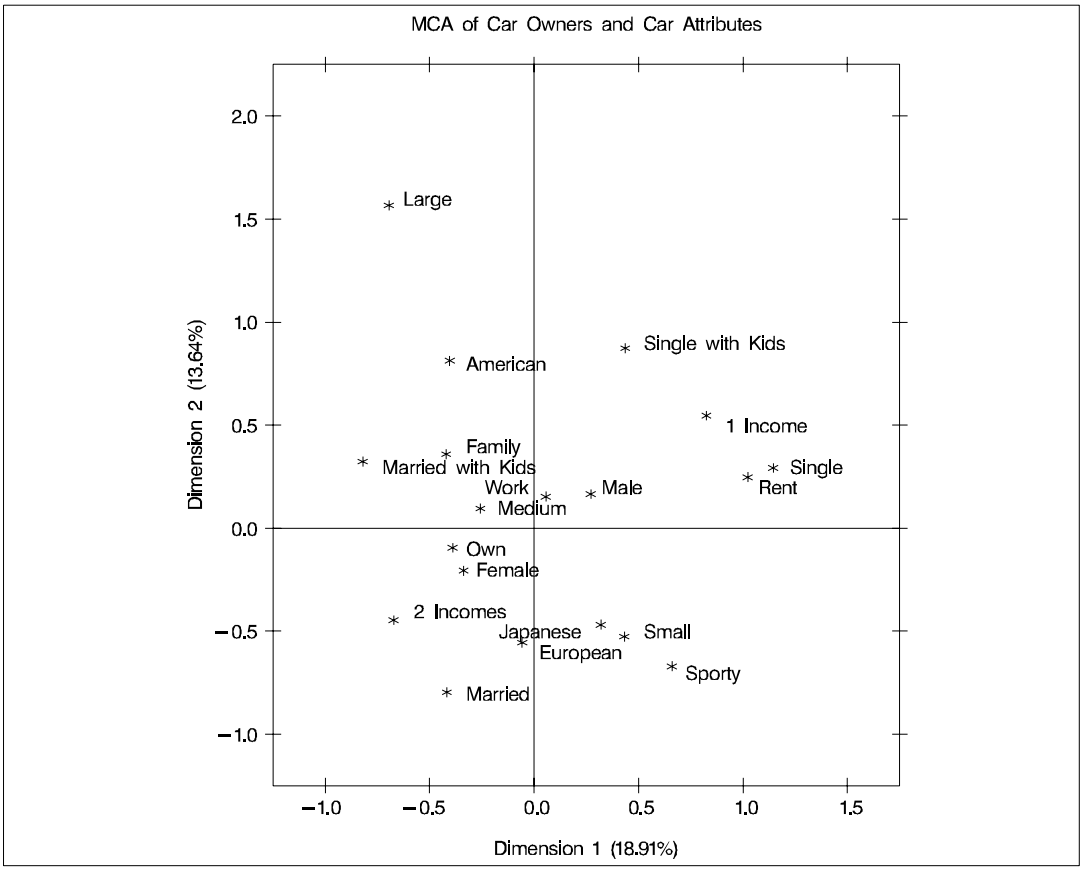
Squared Cosines for the Column Points

	Dim1	Dim2
American	0.0974	0.3952
European	0.0005	0.0468
Japanese	0.1005	0.2136
Large	0.0695	0.3530
Medium	0.0480	0.0068
Small	0.1544	0.2281
Family	0.1919	0.1411
Sporty	0.2027	0.2085
Work	0.0006	0.0046
1 Income	0.5550	0.2441
2 Incomes	0.5550	0.2441
Own	0.3975	0.0234
Rent	0.3975	0.0234
Married	0.0753	0.2742
Married with Kids	0.3258	0.0508
Single	0.6364	0.0416
Single with Kids	0.0090	0.0359
Female	0.0912	0.0341
Male	0.0912	0.0341

Multiple correspondence analysis locates all the categories in a Euclidean space. The first two dimensions of this space are plotted to examine the associations among the categories. The top-right quadrant of the plot shows that the categories single, single with kids, 1 income, and renting a home are associated. Proceeding clockwise, the categories sporty, small, and Japanese are associated. The bottom-left quadrant shows the association between being married, owning your own home, and having two incomes. Having children is associated with owning a large American family car. Such information could be used in market research to identify target audiences for advertisements.

This interpretation is based on points found in approximately the same direction from the origin and in approximately the same region of the space. Distances between points do not have a straightforward interpretation in multiple correspondence analysis. The geometry of multiple correspondence analysis is not a simple generalization of the geometry of simple correspondence analysis (Greenacre and Hastie 1987; Greenacre 1988).

Output 24.2.2. Plot of Multiple Correspondence Analysis of a Burt Table



If you want to perform a multiple correspondence analysis and get scores for the individuals, you can specify the BINARY option to analyze the binary table. In the interest of space, only the first ten rows of coordinates are printed.

```
title 'Car Owners and Car Attributes';
title2 'Binary Table';

*---Perform Multiple Correspondence Analysis---;
proc corresp data=Cars binary;
   ods select RowCoors;
   tables Origin Size Type Income Home Marital Sex;
run;
```

Output 24.2.3. Correspondence Analysis of a Binary Table

Car Owners and Car Attributes
Binary Table

The CORRESP Procedure

Row Coordinates

	Dim1	Dim2
1	-0.4093	1.0878
2	0.8198	-0.2221
3	-0.2193	-0.5328
4	0.4382	1.1799
5	-0.6750	0.3600
6	-0.1778	0.1441
7	-0.9375	0.6846
8	-0.7405	-0.1539
9	-0.3027	-0.2749
10	-0.7263	-0.0803

Example 24.3. Simple Correspondence Analysis of U.S. Population

In this example, PROC CORRESP reads an existing contingency table with supplementary observations and performs a simple correspondence analysis. The data are populations of the fifty states, grouped into regions, for each of the census years from 1920 to 1970 (U.S. Bureau of the Census 1979). Alaska and Hawaii are treated as supplementary regions. They were not states during this entire period and they are not physically connected to the other 48 states. Consequently, it is reasonable to expect that population changes in these two states operate differently from population changes in the other states. The correspondence analysis is performed giving the supplementary points negative weight, then the coordinates for the supplementary points are computed in the solution defined by the other points.

The initial DATA step reads the table, provides labels for the years, flags the supplementary rows with negative weights, and specifies absolute weights of 1000 for all observations since the data were originally reported in units of 1000 people.

In the PROC CORRESP statement, PRINT=PERCENT and the display options display the table of cell percentages (OBSERVED), cell contributions to the total chi-square scaled to sum to 100 (CELLCHI2), row profile rows that sum to 100 (RP), and column profile columns that sum to 100 (CP). The SHORT option specifies that the correspondence analysis summary statistics, contributions to inertia, and squared cosines should not be displayed. The option OUTC=COOR creates the output coordinate data set. Since the data are already in table form, a VAR statement is used to read the table. Row labels are specified with the ID statement, and column labels come from the variable labels. The WEIGHT statement flags the supplementary observations and restores the table values to populations.

The %PLOTIT macro is used to plot the results. Normally, you only need to tell the %PLOTIT macro the name of the input data set, DATA=Coor, and the type of analysis performed on the data, DATATYPE=CORRESP. In this case, PLOTVARS=Dim1

Dim2 is also specified to indicate that Dim1 is the vertical axis variable, as opposed to the default PLOTVARS=Dim2 Dim1.

For an essentially one-dimensional plot such as this, specifying PLOTVARS=Dim1 Dim2 improves the graphical display.

The following statements produce Output 24.3.1 and Output 24.3.2:

```
title 'United States Population';

data USPop;

   * Regions:
   * New England     - ME, NH, VT, MA, RI, CT.
   * Great Lake      - OH, IN, IL, MI, WI.
   * South Atlantic  - DE, MD, DC, VA, WV, NC, SC, GA, FL.
   * Mountain        - MT, ID, WY, CO, NM, AZ, UT, NV.
   * Pacific         - WA, OR, CA.
   *
   * Note: Multiply data values by 1000 to get populations.;

   input Region $14. y1920 y1930 y1940 y1950 y1960 y1970;

   label y1920 = '1920'    y1930 = '1930'    y1940 = '1940'
         y1950 = '1950'    y1960 = '1960'    y1970 = '1970';

   if region = 'Hawaii' or region = 'Alaska'
      then w = -1000;       /* Flag Supplementary Observations */
      else w =  1000;

   datalines;
New England        7401  8166  8437  9314 10509 11842
NY, NJ, PA        22261 26261 27539 30146 34168 37199
Great Lake        21476 25297 26626 30399 36225 40252
Midwest           12544 13297 13517 14061 15394 16319
South Atlantic    13990 15794 17823 21182 25972 30671
KY, TN, AL, MS     8893  9887 10778 11447 12050 12803
AR, LA, OK, TX    10242 12177 13065 14538 16951 19321
Mountain           3336  3702  4150  5075  6855  8282
Pacific            5567  8195  9733 14486 20339 25454
Alaska               55    59    73   129   226   300
Hawaii              256   368   423   500   633   769
;

*---Perform Simple Correspondence Analysis---;
proc corresp print=percent observed cellchi2 rp cp
     short outc=Coor;
   var y1920 -- y1970;
   id Region;
   weight w;
   run;

*---Plot the Simple Correspondence Analysis Results---;
%plotit(data=Coor, datatype=corresp, plotvars=Dim1 Dim2)
```

The contingency table shows that the population of all regions increased over this time period. The row profiles show that population is increasing at a different rate for the different regions. There is a small increase in population in the Midwest, for example, but the population has more than quadrupled in the Pacific region over the same period. The column profiles show that in 1920, the US population was concentrated in the NY, NJ, PA, Great Lakes, Midwest, and South Atlantic regions. With time, the population is shifting more to the South Atlantic, Mountain, and Pacific regions. This is also clear from the correspondence analysis. The inertia and chi-square decomposition table shows that there are five nontrivial dimensions in the table, but the association between the rows and columns is almost entirely one-dimensional.

Output 24.3.1. Simple Correspondence Analysis of a Contingency Table with Supplementary Observations

United States Population

The CORRESP Procedure

Contingency Table

Percents	1920	1930	1940	1950	1960	1970	Sum
New England	0.830	0.916	0.946	1.045	1.179	1.328	6.245
NY, NJ, PA	2.497	2.946	3.089	3.382	3.833	4.173	19.921
Great Lake	2.409	2.838	2.987	3.410	4.064	4.516	20.224
Midwest	1.407	1.492	1.516	1.577	1.727	1.831	9.550
South Atlantic	1.569	1.772	1.999	2.376	2.914	3.441	14.071
KY, TN, AL, MS	0.998	1.109	1.209	1.284	1.352	1.436	7.388
AR, LA, OK, TX	1.149	1.366	1.466	1.631	1.902	2.167	9.681
Mountain	0.374	0.415	0.466	0.569	0.769	0.929	3.523
Pacific	0.625	0.919	1.092	1.625	2.282	2.855	9.398
Sum	11.859	13.773	14.771	16.900	20.020	22.677	100.000

Supplementary Rows

Percents	1920	1930	1940	1950	1960	1970
Alaska	0.006170	0.006619	0.008189	0.014471	0.025353	0.033655
Hawaii	0.028719	0.041283	0.047453	0.056091	0.071011	0.086268

Contributions to the Total Chi-Square Statistic

Percents	1920	1930	1940	1950	1960	1970	Sum
New England	0.937	0.314	0.054	0.009	0.352	0.469	2.135
NY, NJ, PA	0.665	1.287	0.633	0.006	0.521	2.265	5.378
Great Lake	0.004	0.085	0.000	0.001	0.005	0.094	0.189
Midwest	5.749	2.039	0.684	0.072	1.546	4.472	14.563
South Atlantic	0.509	1.231	0.259	0.000	0.285	1.688	3.973
KY, TN, AL, MS	1.454	0.711	1.098	0.087	0.946	2.945	7.242
AR, LA, OK, TX	0.000	0.069	0.077	0.001	0.059	0.030	0.238
Mountain	0.391	0.868	0.497	0.098	0.498	1.834	4.187
Pacific	18.591	9.380	5.458	0.074	7.346	21.248	62.096
Sum	28.302	15.986	8.761	0.349	11.558	35.046	100.000

United States Population

The CORRESP Procedure

Row Profiles

Percents	1920	1930	1940	1950	1960	1970
New England	13.2947	14.6688	15.1557	16.7310	18.8777	21.2722
NY, NJ, PA	12.5362	14.7888	15.5085	16.9766	19.2416	20.9484
Great Lake	11.9129	14.0325	14.7697	16.8626	20.0943	22.3281
Midwest	14.7348	15.6193	15.8777	16.5167	18.0825	19.1691
South Atlantic	11.1535	12.5917	14.2093	16.8872	20.7060	24.4523
KY, TN, AL, MS	13.5033	15.0126	16.3655	17.3813	18.2969	19.4403
AR, LA, OK, TX	11.8687	14.1111	15.1401	16.8471	19.6433	22.3897
Mountain	10.6242	11.7898	13.2166	16.1624	21.8312	26.3758
Pacific	6.6453	9.7823	11.6182	17.2918	24.2784	30.3841

Supplementary Row Profiles

Percents	1920	1930	1940	1950	1960	1970
Alaska	6.5321	7.0071	8.6698	15.3207	26.8409	35.6295
Hawaii	8.6809	12.4788	14.3438	16.9549	21.4649	26.0766

Column Profiles

Percents	1920	1930	1940	1950	1960	1970
New England	7.0012	6.6511	6.4078	6.1826	5.8886	5.8582
NY, NJ, PA	21.0586	21.3894	20.9155	20.0109	19.1457	18.4023
Great Lake	20.3160	20.6042	20.2221	20.1788	20.2983	19.9126
Midwest	11.8664	10.8303	10.2660	9.3337	8.6259	8.0730
South Atlantic	13.2343	12.8641	13.5363	14.0606	14.5532	15.1729
KY, TN, AL, MS	8.4126	8.0529	8.1857	7.5985	6.7521	6.3336
AR, LA, OK, TX	9.6888	9.9181	9.9227	9.6503	9.4983	9.5581
Mountain	3.1558	3.0152	3.1519	3.3688	3.8411	4.0971
Pacific	5.2663	6.6748	7.3921	9.6158	11.3968	12.5921

United States Population

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	20 40 60 80 100
0.10664	0.01137	1.014E7	98.16	98.16	**********
0.01238	0.00015	136586	1.32	99.48	
0.00658	0.00004	38540	0.37	99.85	
0.00333	0.00001	9896.6	0.10	99.95	
0.00244	0.00001	5309.9	0.05	100.00	
Total	0.01159	1.033E7	100.00		

Degrees of Freedom = 40

Row Coordinates

	Dim1	Dim2
New England	0.0611	0.0132
NY, NJ, PA	0.0546	-0.0117
Great Lake	0.0074	-0.0028
Midwest	0.1315	0.0186
South Atlantic	-0.0553	0.0105
KY, TN, AL, MS	0.1044	-0.0144
AR, LA, OK, TX	0.0131	-0.0067
Mountain	-0.1121	0.0338
Pacific	-0.2766	-0.0070

Supplementary Row Coordinates

	Dim1	Dim2
Alaska	-0.4152	0.0912
Hawaii	-0.1198	-0.0321

Column Coordinates

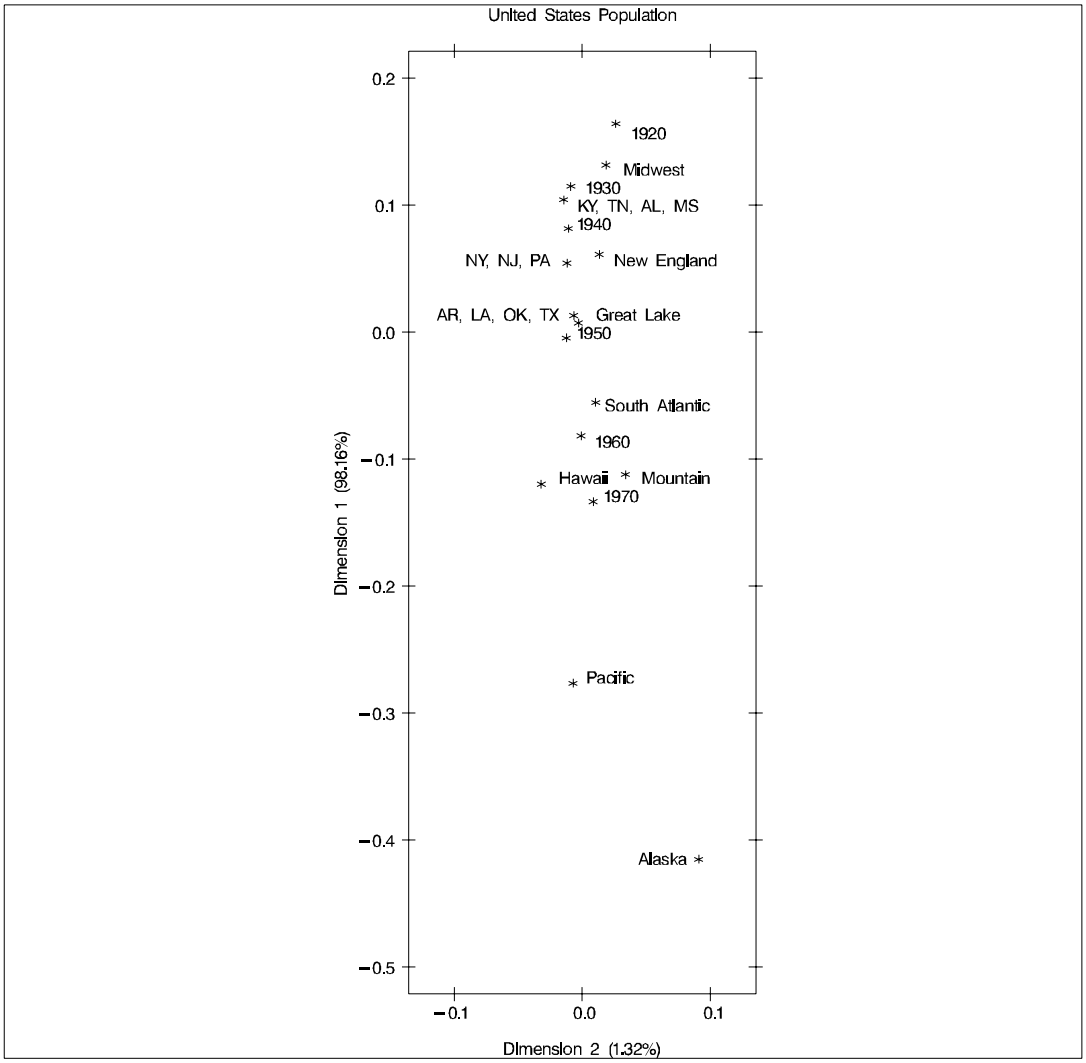
	Dim1	Dim2
1920	0.1642	0.0263
1930	0.1149	-0.0089
1940	0.0816	-0.0108
1950	-0.0046	-0.0125
1960	-0.0815	-0.0007
1970	-0.1335	0.0086

The plot shows that the first dimension correctly orders the years. There is nothing in the correspondence analysis that forces this to happen; PROC CORRESP knows nothing about the inherent ordering of the column categories. The ordering of the regions and the ordering of the years reflect the shift over time of the US population from the Northeast quadrant of the country to the South and to the West. The results show that the West and Southeast are growing faster than the rest of the contiguous 48 states.

The plot also shows that the growth pattern for Hawaii is similar to the growth pattern for the mountain states and that Alaska's growth is even more extreme than the Pacific states' growth. The row profiles confirm this interpretation.

The Pacific region is farther from the origin than all other active points. The Midwest is the extreme region in the other direction. The table of contributions to the total chi-square shows that 62% of the total chi-square statistic is contributed by the Pacific region, which is followed by the Midwest at over 14%. Similarly the two extreme years, 1920 and 1970, together contribute over 63% to the total chi-square, whereas the years nearer the origin of the plot contribute less.

Output 24.3.2. Plot of Simple Correspondence Analysis of a Contingency Table with Supplementary Observations



References

Benzécri, J.P. (1973), *L'Analyse des Données: T. 2, l'Analyse des Correspondances*, Paris: Dunod.

Benzécri, J.P. (1979), *Sur le Calcul des taux d'inertie dans l'analyse d'un questionaire*, Addendum et erratum á [BIN.MULT.]. Cahiers de l'Analyse des Données 4, 377–378.

Burt, C. (1950), "The Factorial Analysis of Qualitative Data," *British Journal of Psychology*, 3, 166–185.

Carroll, J.D, Green, P.E., and Schaffer, C.M. (1986), "Interpoint Distance Comparisons in Correspondence Analysis," *Journal of Marketing Research*, 23, 271–280.

Fisher, R.A. (1940), "The Precision of Discriminant Functions," *Annals of Eugenics*, 10, 422–429.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, New York: John Wiley & Sons, Inc.

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

Greenacre, M.J. (1988), "Correspondence Analysis of Multivariate Categorical Data by Weighted Least-Squares," *Biometrika*, 75, 457–467.

Greenacre, M.J. (1989), "The Carroll-Green-Schaffer Scaling in Correspondence Analysis: A Theoretical and Empirical Appraisal," *Journal of Market Research*,26, 358–365.

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

Greenacre, M.J. and Hastie, T. (1987), "The Geometric Interpretation of Correspondence Analysis," *Journal of the American Statistical Association*, 82, 437–447.

Guttman, L. (1941), "The Quantification of a Class of Attributes: A Theory and Method of Scale Construction," in P. Horst, et al. (ed)., *The Prediction of Personal Adjustment*, New York: Social Science Research Council.

Hayashi, C. (1950), "On the Quantification of Qualitative Data from the Mathematico-Statistical Point of View," *Annals of the Institute of Statistical Mathematics*, 2 (1), 35–47.

van der Heijden, P.G.M, and de Leeuw, J. (1985), "Correspondence Analysis Used Complementary to Loglinear Analysis," *Psychometrika*, 50, 429–447.

Hirshfield, H.O. (1935), "A Connection Between Correlation and Contingency," *Cambridge Philosophical Society Proceedings*, 31, 520–524.

Hoffman, D.L. and Franke, G.R. (1986), "Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research," *Journal of Marketing Research*, 23, 213–227.

Horst, P. (1935), "Measuring Complex Attitudes," *Journal of Social Psychology*, 6, 369–374.

Kobayashi, R. (1981), *An Introduction to Quantification Theory*, Tokyo: Japan Union of Scientists and Engineers.

Komazawa, T. (1982), *Quantification Theory and Data Processing*, Tokyo: Asakura-shoten.

Lebart, L., Morineau, A., and Tabard, N. (1977), *Techniques de la Description Statistique*, Paris: Dunod.

Lebart, L., Morineau, A., and Warwick, K.M. (1984), *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, New York: John Wiley & Sons, Inc.

Nishisato, S. (1980), *Analysis of Categorical Data: Dual Scaling and Its Applications*, Toronto: University of Toronto Press.

Nishisato, S. (1982), *Quantification of Qualitative Data - Dual Scaling and Its Applications*, Tokyo: Asakura-shoten.

Richardson, M., and Kuder, G.F. (1933), "Making a Rating Scale that Measures," *Personnel Journal*, 12, 36–40.

Tenenhaus, M. and Young, F.W. (1985), "An Analysis and Synthesis of Multiple Correspondence Analysis, Optimal Scaling, Dual Scaling, Homogeneity Analysis, and Other Methods of Quantifying Categorical Multivariate Data," *Psychometrika*, 50, 91–119.

U.S. Bureau of the Census (1979), *Statistical Abstract of the United States*, (100th Edition), Washington DC.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.

ISBN 1–58025–494–2

All rights reserved. Produced in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The Institute is a private company devoted to the support and further development of its software and related services.