

## Title

**multivariate** — Introduction to multivariate commands

## Description

The *Multivariate Reference Manual* organizes the commands alphabetically, which makes it easy to find individual command entries if you know the name of the command. This overview organizes and presents the commands conceptually, that is, according to the similarities in the functions that they perform.

The commands listed under the heading **Cluster analysis** perform cluster analysis on variables or the similarity or dissimilarity values within a matrix. An introduction to cluster analysis and a description of the `cluster` and `clustermat` subcommands is provided in [MV] **cluster**.

The commands listed under the heading **Factor analysis and principal component analysis** provide factor analysis of a correlation matrix and principal component analysis (PCA) of a correlation or covariance matrix. The correlation or covariance matrix can be provided directly or computed from variables.

The commands listed under the heading **Rotation** provide methods for rotating a factor or PCA solution or for rotating a matrix. Also provided is Procrustean rotation analysis for rotating a set of variables to best match another set of variables.

The commands listed under **Multivariate analysis of variance and related techniques** provide canonical correlation analysis, multivariate regression, multivariate analysis of variance (MANOVA), and comparison of multivariate means.

The commands listed under **Multidimensional scaling and biplots** provide classic MDS and two-dimensional biplots. MDS can be performed on the variables or on proximity data in a matrix or as proximity data in long format.

The commands listed under **Correspondence analysis** provide simple correspondence analysis (CA) on the cross tabulation of two categorical variables or on a matrix.

### Cluster analysis

<code>cluster</code>	Introduction to cluster-analysis commands
<code>cluster ...</code>	(See [MV] <b>cluster</b> for details)
<code>clustermat</code>	Cluster analysis of dissimilarity matrix
<code>matrix dissimilarity</code>	Compute similarity or dissimilarity measures; may be used by <code>clustermat</code>

### Factor analysis and principal component analysis

<code>factor</code>	Factor analysis
<code>factor postestimation</code>	Postestimation tools for factor and <code>factormat</code>
<code>pca</code>	Principal component analysis
<code>pca postestimation</code>	Postestimation tools for <code>pca</code> and <code>pcamat</code>
<code>rotate</code>	Orthogonal and oblique rotations after factor and <code>pca</code>
<code>screeplot</code>	Scree plot of eigenvalues
<code>scoreplot</code>	Score and loading plots after factor and <code>pca</code>

**Rotation**

<code>rotate</code>	Orthogonal and oblique rotations after factor and pca
<code>rotatemat</code>	Orthogonal and oblique rotations
<code>procrustes</code>	Procrustes transformation
<code>procrustes postestimation</code>	Postestimation tools for procrustes

**Multivariate analysis of variance and related techniques**

<code>canon</code>	Canonical correlations
<code>canon postestimation</code>	Postestimation tools for canon
<code>mvreg</code>	See [R] <b>mvreg</b> ; Multivariate regression
<code>mvreg postestimation</code>	See [R] <b>mvreg postestimation</b> ; Postestimation tools for mvreg
<code>manova</code>	Multivariate analysis of variance and covariance
<code>manova postestimation</code>	Postestimation tools for manova
<code>hotelling</code>	Hotelling's <i>T</i> -squared generalized means test

**Multidimensional scaling and biplots**

<code>mds</code>	Multidimensional scaling for two way data
<code>mds postestimation</code>	Postestimation tools for mds, mdsmat, and mdslong
<code>mdslong</code>	Multidimensional scaling of proximity data in long format
<code>mdsmat</code>	Multidimensional scaling of proximity data in a matrix
<code>biplot</code>	Biplots

**Correspondence analysis**

<code>ca</code>	Simple correspondence analysis
<code>ca postestimation</code>	Postestimation tools for ca

**Remarks**

Remarks are presented under the headings

*Cluster analysis*  
*Factor analysis and principal component analysis*  
*Rotation*  
*Multivariate analysis of variance and related techniques*  
*Multidimensional scaling and biplots*  
*Correspondence analysis*

**Cluster analysis**

Cluster analysis is concerned with finding natural groupings or clusters. Stata's cluster-analysis commands provide several hierarchical and partition clustering methods, postclustering summarization methods, and cluster-management tools. The hierarchical clustering methods may be applied to the data using the `cluster` command or to a user-supplied dissimilarity matrix using the `clustermat` command. See [MV] **cluster** for an introduction to cluster analysis and the `cluster` and `clustermat` suite of commands.

A wide variety of similarity and dissimilarity measures are available for comparing observations; see [MV] *measure\_option*. Dissimilarity matrices, for use with `clusteramat`, are easily obtained using the `matrix dissimilarity` command; see [MV] **matrix dissimilarity**. This provides the building blocks necessary for clustering variables instead of observations, or for clustering using a dissimilarity not automatically provided by Stata; [MV] **clusteramat** provides examples.

## Factor analysis and principal component analysis

Factor analysis and principal component analysis (PCA) have dual uses. They may be used as a dimension-reduction technique, and they may be used in describing the underlying data.

In PCA, the leading eigenvectors from the eigen decomposition of the correlation or covariance matrix of the variables describe a series of uncorrelated linear combinations of the variables that contain most of the variance. For data reduction, a small number of these leading components are retained. For describing the underlying structure of the data, the magnitudes and signs of the eigenvector elements are interpreted in relation to the original variables (rows of the eigenvector).

`pca` uses the correlation or covariance matrix computed from the dataset. `pcamat` allows the correlation or covariance matrix to be directly provided. The `vce(normal)` option provides standard errors for the eigenvalues and eigenvectors which aids in their interpretation. See [MV] **pca** for details.

Factor analysis finds a small number of common factors that linearly reconstruct the original variables. Reconstruction is defined in terms of prediction of the correlation matrix of the original variables, unlike PCA where reconstruction means minimum residual variance summed across all variables. Factor loadings are examined for interpretation of the structure of the data.

`factor` computes the correlation from the dataset, whereas `factormat` is supplied the matrix directly. They both display the eigenvalues of the correlation matrix, the factor loadings, and the “uniqueness” of the variables. See [MV] **factor** for details.

To perform factor analysis or PCA on binary data, compute the tetrachoric correlations and use these with `factormat` or `pcamat`. Tetrachoric correlations are available with the `tetrachoric` command; see [R] **tetrachoric**.

After factor analysis and PCA a suite of commands are available that provide for rotation of the loadings; generation of score variables; graphing of scree plots, loading plots, and score plots; display of matrices and scalars of interest such as anti-image matrices, residual matrices, Kaiser–Meyer–Olkin measures of sampling adequacy, squared multiple correlations, and more. See [MV] **factor postestimation**, [MV] **pca postestimation**, [MV] **rotate**, [MV] **screepplot**, and [MV] **scoreplot** for details.

## Rotation

Rotation provides a modified solution that is rotated from an original multivariate solution in such a way that interpretation is enhanced. Rotation is provided through three commands: `rotate`, `rotatemat`, and `procrustes`.

`rotate` works directly after `pca`, `pcamat`, `factor`, and `factormat`. It knows where to obtain the component- or factor-loading matrix for rotation, and after rotating the loading matrix, it places the rotated results in `e()` so that all the postestimation tools available after `pca` and `factor` may be applied to the rotated results. See [MV] **rotate** for details.

Perhaps you have the component or factor loadings from a published source and want to investigate various rotations, or perhaps you wish to rotate a loading matrix from some other multivariate command. `rotatemat` provides rotations for a specified matrix. See [MV] **rotatemat** for details.

A large selection of orthogonal and oblique rotations are provided for `rotate` and `rotatemat`. These include varimax, quartimax, equamax, parsimax, minimum entropy, Comrey's tandem 1 and 2 principles, promax family, oblimin family, Crawford–Ferguson family, Bentler's invariant pattern simplicity, oblimax, quartimin, target, and weighted target rotations. Horst normalization is also available.

The `procrustes` command provides Procrustean analysis. The goal is to transform a set of source variables to be as close as possible to a set of target variables. The permitted transformations are any combination of dilation (uniform scaling), rotation and reflection (orthogonal and oblique transformations), and translation. Closeness is measured by the residual sum of squares. See [MV] **procrustes** for details.

A set of postestimation commands are available after `procrustes` for generating fitted values and residuals; providing fit statistics for orthogonal, oblique, and unrestricted transformations; and providing a Procrustes overlay graph. See [MV] **procrustes postestimation** for details.

## Multivariate analysis of variance and related techniques

The first canonical correlation is the maximum correlation that can be obtained between a linear combination of one set of variables and a linear combination of another set of variables. The second canonical correlation is the maximum correlation that can be obtained between linear combinations of the two sets of variables subject to the constraint that these second linear combinations are orthogonal to the first linear combinations, and so on.

`canon` estimates these canonical correlations and provides the loadings that describe the linear combinations of the two sets of variables that produce the correlations. Standard errors of the loadings are provided, and tests of the significance of the canonical correlations are available. See [MV] **canon** for details.

Postestimation tools are available after `canon` for generating the variables corresponding to the linear combinations underlying the canonical correlations. Various matrices and correlations of interest may also be displayed. See [MV] **canon postestimation** for details.

In canonical correlation there is no real distinction between the two sets of original variables. In multivariate regression, however, the two sets of variables take on the roles of dependent and independent variables. `multivariate regression` is an extension of regression that allows for multiple dependent variables. See [R] **mvreg** for multivariate regression, and see [R] **mvreg postestimation** for the postestimation tools available after multivariate regression.

Just as analysis of variance (ANOVA) can be formulated in terms of regression where the categorical independent variables are represented by indicator (sometimes called dummy) variables, likewise multivariate analysis-of-variance (MANOVA), a generalization of ANOVA that allows for multiple dependent variables, can be formulated in terms of multivariate regression where the categorical independent variables are represented by indicator variables. Multivariate analysis-of-covariance (MANCOVA) allows for both continuous and categorical independent variables.

The `manova` command fits MANOVA and MANCOVA models for balanced and unbalanced designs, including designs with missing cells, and for factorial, nested, or mixed designs, or designs involving repeated measures. Four multivariate test statistics—Wilks' lambda, Pillai's trace, Lawley–Hotelling trace, and Roy's largest root—are computed for each term in the model. See [MV] **manova** for details.

Postestimation tools are available after `manova` that provide for univariate Wald tests of expressions involving the coefficients of the underlying regression model and that provide for multivariate tests involving terms or linear combinations of the underlying design matrix. Linear combinations of the dependent variables is also supported. See [MV] **manova postestimation** for details.

Related to MANOVA is Hotelling's  $T$ -squared test of whether a set of means is zero, or two sets of means are equal. It is a multivariate test that reduces to a standard  $t$  test if only one variable is involved. The `hotelling` command provides Hotelling's  $T$ -squared test; see [MV] `hotelling`.

## Multidimensional scaling and biplots

Multidimensional scaling (MDS) is a dimension-reduction and visualization technique. Dissimilarities (for instance, Euclidean distances) between observations in a high-dimensional space are represented in a lower-dimensional space (typically two dimensions) so that the Euclidean distance in the lower-dimensional space approximates the dissimilarities in the higher-dimensional space.

The `mds` command provides classical metric MDS for dissimilarities between observations with respect to the variables; see [MV] `mds`. A wide variety of similarity and dissimilarity measures are allowed (the same ones available for the `cluster` command); see [MV] `measure_option`.

`mdslong` and `mdsmat` provide classical metric MDS directly on the dissimilarities recorded either as data in long format (`mdslong`) or as a dissimilarity matrix (`mdsmat`); see [MV] `mdslong` and [MV] `mdsmat`.

Postestimation tools available after `mds`, `mdslong`, and `mdsmat` provide MDS configuration plots and Shepard diagrams; generation of the approximating configuration or the dissimilarities, distances, and raw residuals; and various matrices and scalars of interest such as Kruskal stress (loss), quantiles of the residuals per object, and correlations between dissimilarities and approximating distances. See [MV] `mds postestimation` for details.

Biplots are two-dimensional representations of data. Both the observations and the variables are represented. The observations are represented by marker symbols, and the variables are represented by arrows from the origin. Observations are projected to two dimensions so that the distance between the observations is approximately preserved. The cosine of the angle between arrows approximates the correlation between the variables. A biplot aids in understanding the relationship between the variables, the observations, and the observations and variables jointly. The `biplot` command produces biplots; see [MV] `biplot`.

## Correspondence analysis

Simple correspondence analysis (CA) is a technique for jointly exploring the relationship between rows and columns in a cross tabulation. It is known by many names including dual scaling, reciprocal averaging, and canonical correlation analysis of contingency tables.

`ca` performs CA on the cross-tabulation of two integer-valued variables. `camat` performs CA on a matrix with non-negative entries—perhaps from a published table. See [MV] `ca` for details.

A suite of commands are available following `ca` and `camat`. These include commands for producing CA biplots and dimensional projection plots; generating fitted values, row coordinates, and column coordinates; and displaying distances between row and column profiles, individual cell inertia contributions,  $\chi^2$  distances between row and column profiles, and the fitted correspondence table. See [MV] `ca postestimation` for details.

## Also See

**Complementary:** [U] 1.3 What's new

**Background:** [R] intro