Title

stata.com

pca — Principal component analysis

Description

`pca` and `pcamat` display the eigenvalues and eigenvectors from the principal component analysis (PCA) eigen decomposition. The eigenvectors are returned in orthonormal form, that is, uncorrelated and normalized.

`pca` can be used to reduce the number of variables or to learn about the underlying structure of the data. `pcamat` provides the correlation or covariance matrix directly. For `pca`, the correlation or covariance matrix is computed from the variables in *varlist*.

Quick start

Principal component analysis of data

Principal component analysis of `v1`, `v2`, `v3`, and `v4`

```
pca v1 v2 v3 v4
```

As above, but retain only 2 components

```
pca v1 v2 v3 v4, components(2)
```

As above, but retain only those components with eigenvalues greater than or equal to 0.5

```
pca v1 v2 v3 v4, mineigen(.5)
```

Principal component analysis of covariance matrix instead of correlation matrix

```
pca v1 v2 v3 v4, covariance
```

Principal component analysis of a correlation matrix

Principal component analysis of matrix `C` representing the correlations from 1,000 observations

```
pcamat C, n(1000)
```

As above, but retain only 4 components

```
pcamat C, n(1000) components(4)
```

Menu

pca

Statistics > Multivariate analysis > Factor and principal component analysis > Principal component analysis (PCA)

pcamat

Statistics > Multivariate analysis > Factor and principal component analysis > PCA of a correlation or covariance matrix

Syntax

Principal component analysis of data

```
pca varlist [if] [in] [weight] [, options]
```

Principal component analysis of a correlation or covariance matrix

```
pcamat matname, n(#) [options pcamat_options]
```

matname is a $k \times k$ symmetric matrix or a $k(k+1)/2$ long row or column vector containing the upper or lower triangle of the correlation or covariance matrix.

options	Description
Model 2	
components(#)	retain maximum of # principal components; factors() is a synonym
mineigen(#)	retain eigenvalues larger than #; default is 1e-5
correlation	perform PCA of the correlation matrix; the default
covariance	perform PCA of the covariance matrix
vce(none)	do not compute VCE of the eigenvalues and vectors; the default
vce(normal)	compute VCE of the eigenvalues and vectors assuming multivariate normality
Reporting	
level(#)	set confidence level; default is level(95)
blanks(#)	display loadings as blanks when |loadings| < #
novce	suppress display of SEs even though calculated
* means	display summary statistics of variables
Advanced	
tol(#)	advanced option; see *Options* for details
ignore	advanced option; see *Options* for details
norotated	display unrotated results, even if rotated results are available (replay only)

* means is not allowed with pcamat.
norotated is not shown in the dialog box.

pcamat_options	Description
Model	
shape(full)	*matname* is a square symmetric matrix; the default
shape(lower)	*matname* is a vector with the rowwise lower triangle (with diagonal)
shape(upper)	*matname* is a vector with the rowwise upper triangle (with diagonal)
names(*namelist*)	variable names; required if *matname* is triangular
forcepsd	modifies *matname* to be positive semidefinite
* n(#)	number of observations
sds(*matname*$_2$)	vector with standard deviations of variables
means(*matname*$_3$)	vector with means of variables

* n() is required for pcamat.

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

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

aweights are not allowed with the jackknife prefix; see [R] jackknife.

aweights and fweights are allowed with pca; see [U] 11.1.6 weight.

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

Options

Model 2

components(#) and mineigen(#) specify the maximum number of components (eigenvectors or factors) to be retained. components() specifies the number directly, and mineigen() specifies it indirectly, keeping all components with eigenvalues greater than the indicated value. The options can be specified individually, together, or not at all. factors() is a synonym for components().

components(#) sets the maximum number of components (factors) to be retained. pca and pcamat always display the full set of eigenvalues but display eigenvectors only for retained components. Specifying a number larger than the number of variables in *varlist* is equivalent to specifying the number of variables in *varlist* and is the default.

mineigen(#) sets the minimum value of eigenvalues to be retained. The default is 1e-5 or the value of tol() if specified.

Specifying components() and mineigen() affects only the number of components to be displayed and stored in e(); it does not enforce the assumption that the other eigenvalues are 0. In particular, the standard errors reported when vce(normal) is specified do not depend on the number of retained components.

correlation and covariance specify that principal components be calculated for the correlation matrix and covariance matrix, respectively. The default is correlation. Unlike factor analysis, PCA is not scale invariant; the eigenvalues and eigenvectors of a covariance matrix differ from those of the associated correlation matrix. Usually, a PCA of a covariance matrix is meaningful only if the variables are expressed in the same units.

For pcamat, do not confuse the type of the matrix to be analyzed with the type of *matname*. Obviously, if *matname* is a correlation matrix and the option sds() is not specified, it is not possible to perform a PCA of the covariance matrix.

vce(none | normal) specifies whether standard errors are to be computed for the eigenvalues, the eigenvectors, and the (cumulative) percentage of explained variance (confirmatory PCA). These standard errors are obtained assuming multivariate normality of the data and are valid only for a PCA of a covariance matrix. Be cautious if applying these to correlation matrices.

Reporting

level(#) specifies the confidence level, as a percentage, for confidence intervals. The default is level(95) or as set by set level; see [U] 20.8 Specifying the width of confidence intervals. level() is allowed only with vce(normal).

blanks(#) shows blanks for loadings with absolute value smaller than #. This option is ignored when specified with vce(normal).

novce suppresses the display of standard errors, even though they are computed, and displays the PCA results in a matrix/table style. You can specify novce during estimation in combination with vce(normal). More likely, you will want to use novce during replay.

means displays summary statistics of the variables over the estimation sample. This option is not available with pcamat.

Advanced

tol(#) is an advanced, rarely used option and is available only with vce(normal). An eigenvalue, ev_i, is classified as being close to zero if $ev_i < \text{tol} \times \max(ev)$. Two eigenvalues, ev_1 and ev_2, are "close" if $\text{abs}(ev_1 - ev_2) < \text{tol} \times \max(ev)$. The default is tol(1e-5). See option ignore below and the technical note later in this entry.

ignore is an advanced, rarely used option and is available only with vce(normal). It continues the computation of standard errors and tests, even if some eigenvalues are suspiciously close to zero or suspiciously close to other eigenvalues, violating crucial assumptions of the asymptotic theory used to estimate standard errors and tests. See the technical note later in this entry.

The following option is available with pca and pcamat but is not shown in the dialog box:

norotated displays the unrotated principal components, even if rotated components are available. This option may be specified only when replaying results.

Options unique to pcamat

Model

shape(*shape_arg*) specifies the shape (storage mode) for the covariance or correlation matrix *matname*. The following shapes are supported:

full specifies that the correlation or covariance structure of k variables is stored as a symmetric $k \times k$ matrix. Specifying shape(full) is optional in this case.

lower specifies that the correlation or covariance structure of k variables is stored as a vector with $k(k+1)/2$ elements in rowwise lower-triangular order:

$$C_{11}\ C_{21}\ C_{22}\ C_{31}\ C_{32}\ C_{33}\ \ldots\ C_{k1}\ C_{k2}\ \ldots\ C_{kk}$$

upper specifies that the correlation or covariance structure of k variables is stored as a vector with $k(k+1)/2$ elements in rowwise upper-triangular order:

$$C_{11}\ C_{12}\ C_{13}\ \ldots\ C_{1k}\ C_{22}\ C_{23}\ \ldots C_{2k}\ \ldots\ C_{(k-1k-1)}\ C_{(k-1k)}\ C_{kk}$$

names(*namelist*) specifies a list of k different names, which are used to document output and to label estimation results and are used as variable names by predict. By default, pcamat verifies that the row and column names of *matname* and the column or row names of $matname_2$ and $matname_3$ from the sds() and means() options are in agreement. Using the names() option turns off this check.

forcepsd modifies the matrix *matname* to be positive semidefinite (psd) and so to be a proper covariance matrix. If *matname* is not positive semidefinite, it will have negative eigenvalues. By setting negative eigenvalues to 0 and reconstructing, we obtain the least-squares positive-semidefinite approximation to *matname*. This approximation is a singular covariance matrix.

n(#) is required and specifies the number of observations.

sds(*matname*$_2$) specifies a $k \times 1$ or $1 \times k$ matrix with the standard deviations of the variables. The row or column names should match the variable names, unless the names() option is specified. sds() may be specified only if *matname* is a correlation matrix.

means(*matname*$_3$) specifies a $k\times1$ or $1\times k$ matrix with the means of the variables. The row or column names should match the variable names, unless the names() option is specified. Specify means() if you have variables in your dataset and want to use predict after pcamat.

Remarks and examples

stata.com

Principal component analysis (PCA) is commonly thought of as a statistical technique for data reduction. It helps you reduce the number of variables in an analysis by describing a series of uncorrelated linear combinations of the variables that contain most of the variance. In addition to data reduction, the eigenvectors from a PCA are often inspected to learn more about the underlying structure of the data.

PCA originated with the work of Pearson (1901) and Hotelling (1933). For an introduction, see Rabe-Hesketh and Everitt (2007, chap. 14), van Belle, Fisher, Heagerty, and Lumley (2004), or Afifi, May, and Clark (2012). More advanced treatments are Mardia, Kent, and Bibby (1979, chap. 8), and Rencher and Christensen (2012, chap. 12). For monograph-sized treatments, including extensive discussions of the relationship between PCA and related approaches, see Jackson (2003) and Jolliffe (2002).

The objective of PCA is to find unit-length ($\mathbf{L}'\mathbf{L}=\mathbf{I}$) linear combinations of the variables with the greatest variance. The first principal component has maximal overall variance. The second principal component has maximal variance among all unit-length linear combinations that are uncorrelated to the first principal component, etc. The last principal component has the smallest variance among all unit-length linear combinations of the variables. All principal components combined contain the same information as the original variables, but the important information is partitioned over the components in a particular way: the components are orthogonal, and earlier components contain more information than later components. PCA thus conceived is just a linear transformation of the data. It does not assume that the data satisfy a specific statistical model, though it does require that the data be interval-level data—otherwise taking linear combinations is meaningless.

PCA is scale dependent. The principal components of a covariance matrix and those of a correlation matrix are different. In applied research, PCA of a covariance matrix is useful only if the variables are expressed in commensurable units.

❑ Technical note

Principal components have several useful properties. Some of these are geometric. Both the principal components and the principal scores are uncorrelated (orthogonal) among each other. The f leading principal components have maximal generalized variance among all f unit-length linear combinations.

It is also possible to interpret PCA as a fixed-effects factor analysis with homoskedastic residuals

$$y_{ij}=\mathbf{a}_i'\mathbf{b}_j+e_{ij}\qquad i=1,\ldots,n\qquad j=1,\ldots,p$$

where y_{ij} are the elements of the matrix $\mathbf{Y}$, $\mathbf{a}_i$ (scores) and $\mathbf{b}_j$ (loadings) are f-vectors of parameters, and e_{ij} are independent homoskedastic residuals. (In factor analysis, the scores $\mathbf{a}_i$ are random rather than fixed, and the residuals are allowed to be heteroskedastic in j.) It follows that $E(\mathbf{Y})$ is a matrix of rank f, with f typically substantially less than n or p. Thus we may think of PCA as a regression model with a restricted number but unknown independent variables. We may also say that the expected values of the rows (or columns) of $\mathbf{Y}$ are in some unknown f-dimensional space.

For more information on these properties and for other characterizations of PCA, see Jackson (2003) and Jolliffe (2002).

❑

Example 1: Principal component analysis of audiometric data

We consider a dataset of audiometric measurements on 100 males, age 39. The measurements are minimal discernible intensities at four different frequencies with the left and right ear (see Jackson 2003, 106). The variable `lft1000` refers to the left ear at 1,000 Hz.

```
. use http://www.stata-press.com/data/r15/audiometric
(Audiometric measures)
. correlate lft* rght*
(obs=100)
```

	lft500	lft1000	lft2000	lft4000	rght500	rght1000	rght2000
lft500	1.0000						
lft1000	0.7775	1.0000					
lft2000	0.4012	0.5366	1.0000				
lft4000	0.2554	0.2749	0.4250	1.0000			
rght500	0.6963	0.5515	0.2391	0.1790	1.0000		
rght1000	0.6416	0.7070	0.4460	0.2632	0.6634	1.0000	
rght2000	0.2372	0.3597	0.7011	0.3165	0.1589	0.4142	1.0000
rght4000	0.2041	0.2169	0.3262	0.7097	0.1321	0.2201	0.3746

	rght4000
rght4000	1.0000

As you may have expected, measurements on the same ear are more highly correlated than measurements on different ears. Also, measurements on different ears at the same frequency are more highly correlated than at different frequencies. Because the variables are in commensurable units, it would make theoretical sense to analyze the covariance matrix of these variables. However, the variances of the measures differ widely:

```
. summarize lft* rght*, sep(4)
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lft500	100	-2.8	6.408643	-10	15
lft1000	100	-.5	7.571211	-10	20
lft2000	100	2	10.94061	-10	45
lft4000	100	21.35	19.61569	-10	70
rght500	100	-2.6	7.123726	-10	25
rght1000	100	-.7	6.396811	-10	20
rght2000	100	1.6	9.289942	-10	35
rght4000	100	21.35	19.33039	-10	75

In an analysis of the covariances, the higher frequency measures would dominate the results. There is no clinical reason for such an effect (see also Jackson [2003]). Therefore, we will analyze the correlation matrix.

```
. pca lft* rght*
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          8
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6
lft500	0.4011	-0.3170	0.1582	-0.3278	0.0231	0.4459
lft1000	0.4210	-0.2255	-0.0520	-0.4816	-0.3792	-0.0675
lft2000	0.3664	0.2386	-0.4703	-0.2824	0.4392	-0.0638
lft4000	0.2809	0.4742	0.4295	-0.1611	0.3503	-0.4169
rght500	0.3433	-0.3860	0.2593	0.4876	0.4975	0.1948
rght1000	0.4114	-0.2318	-0.0289	0.3723	-0.3513	-0.6136
rght2000	0.3115	0.3171	-0.5629	0.3914	-0.1108	0.2650
rght4000	0.2542	0.5135	0.4262	0.1591	-0.3960	0.3660

Variable	Comp7	Comp8	Unexplained
lft500	0.3293	-0.5463	0
lft1000	-0.0331	0.6227	0
lft2000	-0.5255	-0.1863	0
lft4000	0.4269	0.0839	0
rght500	-0.1594	0.3425	0
rght1000	-0.0837	-0.3614	0
rght2000	0.4778	0.1466	0
rght4000	-0.4139	-0.0508	0

`pca` shows two panels. The first panel lists the eigenvalues of the correlation matrix, ordered from largest to smallest. The corresponding eigenvectors are listed in the second panel. These are the principal components and have unit length; the columnwise sum of the squares of the loadings is 1 ($0.4011^2 + 0.4210^2 + \cdots + 0.2542^2 = 1$).

Remark: Literature and software that treat principal components in combination with factor analysis tend to display principal components normed to the associated eigenvalues rather than to 1. This normalization is available in the postestimation command `estat loadings`; see [MV] **pca postestimation**.

The eigenvalues add up to the sum of the variances of the variables in the analysis—the "total variance" of the variables. Because we are analyzing a correlation matrix, the variables are standardized to have unit variance, so the total variance is 8. The eigenvalues are the variances of the principal components. The first principal component has variance 3.93, explaining 49% (3.93/8) of the total variance. The second principal component has variance 1.62 or 20% (1.62/8) of the total variance. Principal components are uncorrelated. You may want to verify that; for instance,

$$0.4011(-0.3170) + 0.4210(-0.2255) + \cdots + 0.2542(0.5135) = 0$$

As a consequence, we may also say that the first two principal components explain the sum of the variances of the individual components, or $49 + 20 = 69\%$ of the total variance. Had the components been correlated, they would have partly represented the same information, so the information contained in the combination would not have been equal to the sum of the information of the components. All eight principal components combined explain all variance in all variables; therefore, the unexplained variances listed in the second panel are all zero, and `Rho` $= 1.00$ as shown above the first panel.

More than 85% of the variance is contained in the first four principal components. We can list just these components with the option `components(4)`.

```
. pca lft* rght*, components(4)
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          4
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     0.8737
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
lft500	0.4011	-0.3170	0.1582	-0.3278	.1308
lft1000	0.4210	-0.2255	-0.0520	-0.4816	.1105
lft2000	0.3664	0.2386	-0.4703	-0.2824	.1275
lft4000	0.2809	0.4742	0.4295	-0.1611	.1342
rght500	0.3433	-0.3860	0.2593	0.4876	.1194
rght1000	0.4114	-0.2318	-0.0289	0.3723	.1825
rght2000	0.3115	0.3171	-0.5629	0.3914	.07537
rght4000	0.2542	0.5135	0.4262	0.1591	.1303

The first panel is not affected. The second panel now lists the first four principal components. These four components do not contain all information in the data, and therefore some of the variances in the variables are unaccounted for or unexplained. These equal the sums of squares of the loadings in the deleted components, weighted by the associated eigenvalues. The unexplained variances in all variables are of similar order. The average unexplained variance is equal to the overall unexplained variance of 13% $(1 - 0.87)$.

Look more closely at the principal components. The first component has positive loadings of roughly equal size on all variables. It can be interpreted as overall sensitivity of a person's ears. The second principal component has positive loadings on the higher frequencies with both ears and negative loadings for the lower frequencies. Thus the second principal component distinguishes sensitivity for higher frequencies versus lower frequencies. The third principal component similarly differentiates sensitivity at medium frequencies from sensitivity at other frequencies. Finally, the fourth principal component has negative loadings on the left ear and positive loadings on the right ear; it differentiates the left and right ear.

We stated earlier that the first principal component had similar loadings on all eight variables. This can be tested if we are willing to assume that the data are multivariate normal distributed. For this case, pca can estimate the standard errors and related statistics. To conserve paper, we request only the results of the first two principal components and specify the option vce(normal).

```
. pca l* r*, comp(2) vce(normal)
(with PCA/correlation, SEs and tests are approximate)
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          2
                                                  Trace            =          8
                                                  Rho              =     0.6934
SEs assume multivariate normality                 SE(Rho)          =     0.0273
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Eigenvalues						
Comp1	3.929005	.5556453	7.07	0.000	2.839961	5.01805
Comp2	1.618322	.2288653	7.07	0.000	1.169754	2.066889
Comp1						
lft500	.4010948	.0429963	9.33	0.000	.3168236	.485366
lft1000	.4209908	.0359372	11.71	0.000	.3505551	.4914264
lft2000	.3663748	.0463297	7.91	0.000	.2755702	.4571794
lft4000	.2808559	.0626577	4.48	0.000	.1580491	.4036628
rght500	.343251	.0528285	6.50	0.000	.2397091	.446793
rght1000	.4114209	.0374312	10.99	0.000	.3380571	.4847846
rght2000	.3115483	.0551475	5.65	0.000	.2034612	.4196354
rght4000	.2542212	.066068	3.85	0.000	.1247303	.3837121
Comp2						
lft500	-.3169638	.067871	-4.67	0.000	-.4499885	-.1839391
lft1000	-.225464	.0669887	-3.37	0.001	-.3567595	-.0941686
lft2000	.2385933	.1079073	2.21	0.027	.0270989	.4500877
lft4000	.4741545	.0967918	4.90	0.000	.284446	.6638629
rght500	-.3860197	.0803155	-4.81	0.000	-.5434352	-.2286042
rght1000	-.2317725	.0674639	-3.44	0.001	-.3639994	-.0995456
rght2000	.317059	.1215412	2.61	0.009	.0788427	.5552752
rght4000	.5135121	.0951842	5.39	0.000	.3269544	.7000697

```
LR test for independence:       chi2(28)  =     448.21   Prob > chi2 =  0.0000
LR test for   sphericity:       chi2(35)  =     451.11   Prob > chi2 =  0.0000
Explained variance by components
```

Components	Eigenvalue	Proportion	SE_Prop	Cumulative	SE_Cum	Bias
Comp1	3.929005	0.4911	0.0394	0.4911	0.0394	.056663
Comp2	1.618322	0.2023	0.0271	0.6934	0.0273	.015812
Comp3	.9753248	0.1219	0.0178	0.8153	0.0175	-.014322
Comp4	.4667822	0.0583	0.0090	0.8737	0.0127	.007304
Comp5	.34009	0.0425	0.0066	0.9162	0.0092	.026307
Comp6	.3158912	0.0395	0.0062	0.9557	0.0055	-.057717
Comp7	.2001111	0.0250	0.0040	0.9807	0.0031	-.013961
Comp8	.1544736	0.0193	0.0031	1.0000	0.0000	-.020087

Here pca acts like an estimation command. The output is organized in different equations. The first equation contains the eigenvalues. The second equation named, Comp1, is the first principal component, etc. pca reports, for instance, standard errors of the eigenvalues. Although testing the values of eigenvalues may, up to now, be rare in applied research, interpretation of results should take stability into consideration. It makes little sense to report the first eigenvalue as 3.929 if you see that the standard error is 0.56.

pca has also reported the standard errors of the principal components. It has also estimated the covariances.

```
. estat vce
  (output omitted)
```

Showing the large amount of information contained in the VCE matrix is not useful by itself. The fact that it has been estimated, however, enables us to test properties of the principal components. Does it make good sense to talk about the loadings of the first principal component being of the same size? We use testparm with two options; see [R] **test**. eq(Comp1) specifies that we are testing coefficients for equation Comp1, that is, the first principal component. equal specifies that instead of testing that the coefficients are zero, we want to test that the coefficients are equal to each other—a more sensible hypothesis because principal components are normalized to 1.

```
. testparm lft* rght*, equal eq(Comp1)
 ( 1)  - [Comp1]lft500 + [Comp1]lft1000 = 0
 ( 2)  - [Comp1]lft500 + [Comp1]lft2000 = 0
 ( 3)  - [Comp1]lft500 + [Comp1]lft4000 = 0
 ( 4)  - [Comp1]lft500 + [Comp1]rght500 = 0
 ( 5)  - [Comp1]lft500 + [Comp1]rght1000 = 0
 ( 6)  - [Comp1]lft500 + [Comp1]rght2000 = 0
 ( 7)  - [Comp1]lft500 + [Comp1]rght4000 = 0
           chi2(  7) =    7.56
         Prob > chi2 =    0.3729
```

We cannot reject the null hypothesis of equal loadings, so our interpretation of the first component does not seem to conflict with the data.

pca also displays standard errors of the proportions of variance explained by the leading principal components. Again this information is useful primarily to indicate the strength of formulations of results rather than to test hypotheses about these statistics. The information is also useful to compare studies: if in one study the leading two principal components explain 70% of variance, whereas in a replicating study they explain 80%, are these differences significant given the sampling variation?

Because pca is an estimation command just like regress or xtlogit, you may replay the output by typing just pca. If you have used pca with the vce(normal) option, you may use the option novce at estimation or during replay to display the standard PCA output.

```
. pca, novce
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          2
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     0.6934
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

```
Principal components (eigenvectors)
```

Variable	Comp1	Comp2	Unexplained
lft500	0.4011	-0.3170	.2053
lft1000	0.4210	-0.2255	.2214
lft2000	0.3664	0.2386	.3805
lft4000	0.2809	0.4742	.3262
rght500	0.3433	-0.3860	.2959
rght1000	0.4114	-0.2318	.248
rght2000	0.3115	0.3171	.456
rght4000	0.2542	0.5135	.3193

◁

❑ Technical note

Inference on the eigenvalues and eigenvectors of a covariance matrix is based on a series of assumptions:

(A1) The variables are multivariate normal distributed.

(A2) The variance–covariance matrix of the observations has all distinct and strictly positive eigenvalues.

Under assumptions A1 and A2, the eigenvalues and eigenvectors of the sample covariance matrix can be seen as maximum likelihood estimates for the population analogues that are asymptotically (multivariate) normally distributed (Anderson 1963; Jackson 2003). See Tyler (1981) for related results for elliptic distributions. Be cautious when interpreting because the asymptotic variances are rather sensitive to violations of assumption A1 (and A2). Wald tests of hypotheses that are in conflict with assumption A2 (for example, testing that the first and second eigenvalues are the same) produce incorrect p-values.

Because the statistical theory for a PCA of a correlation matrix is much more complicated, `pca` and `pcamat` compute standard errors and tests of a correlation matrix as if it were a covariance matrix. This practice is in line with the application of asymptotic theory in Jackson (2003). This will usually lead to some underestimation of standard errors, but we believe that this problem is smaller than the consequences of deviations from normality.

You may conduct tests for multivariate normality using the `mvtest normality` command (see [MV] **mvtest normality**):

```
. mvtest normality lft* rght*, stats(all)
Test for multivariate normality
    Mardia mSkewness =  14.52785   chi2(120) =  251.052   Prob>chi2 =  0.0000
    Mardia mKurtosis =  94.53331     chi2(1) =   33.003   Prob>chi2 =  0.0000
    Henze-Zirkler    =  1.272529     chi2(1) =  118.563   Prob>chi2 =  0.0000
    Doornik-Hansen                  chi2(16) =   95.318   Prob>chi2 =  0.0000
```

These tests cast serious doubt on the multivariate normality of the variables. We advise caution in interpreting the inference results. Time permitting, you may want to turn to bootstrap methods for inference on the principal components and eigenvalues, but you should be aware of some serious identification problems in using the bootstrap here (Milan and Whittaker 1995).

❑

Example 2: Analyzing the covariance instead of the correlation

We remarked before that the principal components of a correlation matrix are generally different from the principal components of a covariance matrix. `pca` defaults to performing the PCA of the correlation matrix. To obtain a PCA of the covariance matrix, specify the `covariance` option.

```
. pca l* r*, comp(4) covariance
Principal components/covariance                 Number of obs    =        100
                                                Number of comp.  =          4
                                                Trace            =     1154.5
    Rotation: (unrotated = principal)           Rho              =     0.9396
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	706.795	527.076	0.6122	0.6122
Comp2	179.719	68.3524	0.1557	0.7679
Comp3	111.366	24.5162	0.0965	0.8643
Comp4	86.8501	57.4842	0.0752	0.9396
Comp5	29.366	9.53428	0.0254	0.9650
Comp6	19.8317	6.67383	0.0172	0.9822
Comp7	13.1578	5.74352	0.0114	0.9936
Comp8	7.41432	.	0.0064	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Unexplained
lft500	0.0835	0.2936	-0.0105	0.3837	7.85
lft1000	0.1091	0.3982	0.0111	0.3162	11.71
lft2000	0.2223	0.5578	0.0558	-0.4474	11.13
lft4000	0.6782	-0.1163	-0.7116	-0.0728	.4024
rght500	0.0662	0.2779	-0.0226	0.4951	12.42
rght1000	0.0891	0.3119	0.0268	0.2758	11.14
rght2000	0.1707	0.3745	0.2721	-0.4496	14.71
rght4000	0.6560	-0.3403	0.6441	0.1550	.4087

As expected, the results are less clear. The total variance to be analyzed is 1,154.5; this is the sum of the variances of the eight variables, that is, the trace of the covariance matrix. The leading principal components now account for a larger fraction of the variance; this is often the case with covariance matrices where the variables have widely different variances. The principal components are somewhat harder to interpret; mainly the loadings are no longer of roughly comparable size.

◁

Example 3: PCA directly from a correlation matrix

Sometimes you do not have the original data but have only the correlation or covariance matrix. `pcamat` performs a PCA for such a matrix. To simplify presentation, we use the data on the left ear.

```
. correlate lft*, cov
(obs=100)
```

	lft500	lft1000	lft2000	lft4000
lft500	41.0707			
lft1000	37.7273	57.3232		
lft2000	28.1313	44.4444	119.697	
lft4000	32.101	40.8333	91.2121	384.775

Suppose that we have the covariances of the variables but not the original data. `correlate` stores the covariances in `r(C)`, so we can use that matrix and invoke `pcamat` with the options `n(100)`, specifying the number of observations, and `names()`, providing the variable names.

```
. matrix Cfull = r(C)
. pcamat Cfull, comp(2) n(100) names(lft500 lft1000 lft2000 lft4000)
Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          2
                                                  Trace            =          4
    Rotation: (unrotated = principal)             Rho              =     0.8169
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	2.37181	1.47588	0.5930	0.5930
Comp2	.895925	.366238	0.2240	0.8169
Comp3	.529687	.327106	0.1324	0.9494
Comp4	.202581	.	0.0506	1.0000

Principal components (eigenvectors)

Variable	Comp1	Comp2	Unexplained
lft500	0.5384	-0.4319	.1453
lft1000	0.5730	-0.3499	.1116
lft2000	0.4958	0.2955	.3387
lft4000	0.3687	0.7770	.1367

If we had to type in the covariance matrix, to avoid excess typing `pcamat` allows you to provide the covariance (or correlation) matrix with just the upper or lower triangular elements including the diagonal. (Thus, for correlations, you have to enter the 1s for the diagonal.) For example, we could enter the lower triangle of our covariance matrix row by row up to and including the diagonal as a one-row Stata matrix.

```
. matrix Clow = (41.0707, 37.7273, 57.3232, 28.1313, 44.4444,
>                119.697, 32.101,  40.8333, 91.2121, 384.775)
```

The matrix `Clow` has one row and 10 columns. To make seeing the structure easier, we prefer to enter these numbers in the following way:

```
. matrix Clow = (41.0707,
>                37.7273, 57.3232,
>                28.1313, 44.4444, 119.697,
>                32.101,  40.8333, 91.2121, 384.775)
```

When using the lower or upper triangle stored in a row or column vector, it is not possible to define the variable names as row or column names of the matrix; the option `names()` is required. Moreover, we have to specify the option `shape(lower)` to inform `pcamat` that the vector contains the lower triangle, not the upper triangle.

```
. pcamat Clow, comp(2) shape(lower) n(100) names(lft500 lft1000 lft2000 lft4000)
  (output omitted)
```

◁

Stored results

pca and pcamat without the vce(normal) option store the following in e():

Scalars

e(N)	number of observations
e(f)	number of retained components
e(rho)	fraction of explained variance
e(trace)	trace of e(C)
e(lndet)	ln of the determinant of e(C)
e(cond)	condition number of e(C)

Macros

e(cmd)	pca (even for pcamat)
e(cmdline)	command as typed
e(Ctype)	correlation or covariance
e(wtype)	weight type
e(wexp)	weight expression
e(title)	title in output
e(properties)	nob noV eigen
e(rotate_cmd)	program used to implement rotate
e(estat_cmd)	program used to implement estat
e(predict)	program used to implement predict
e(marginsnotok)	predictions disallowed by margins

Matrices

e(C)	$p \times p$ correlation or covariance matrix
e(means)	$1 \times p$ matrix of means
e(sds)	$1 \times p$ matrix of standard deviations
e(Ev)	$1 \times p$ matrix of eigenvalues (sorted)
e(L)	$p \times f$ matrix of eigenvectors = components
e(Psi)	$1 \times p$ matrix of unexplained variance

Functions

e(sample)	marks estimation sample

pca and pcamat with the vce(normal) option store the above, as well as the following:

Scalars

e(v_rho)	variance of e(rho)
e(chi2_i)	χ^2 statistic for test of independence
e(df_i)	degrees of freedom for test of independence
e(p_i)	significance of test of independence
e(chi2_s)	χ^2 statistic for test of sphericity
e(df_s)	degrees of freedom for test of sphericity
e(p_s)	significance of test of sphericity
e(rank)	rank of e(V)

Macros

e(vce)	multivariate normality
e(properties)	b V

Matrices

e(b)	$1 \times p+fp$ coefficient vector (all eigenvalues and retained eigenvectors)
e(Ev_bias)	$1 \times p$ matrix: bias of eigenvalues
e(Ev_stats)	$p \times 5$ matrix with statistics on explained variance
e(V)	variance–covariance matrix of the estimates e(b)

Methods and formulas

Methods and formulas are presented under the following headings:

Notation

Let $\mathbf{C}$ be the $p \times p$ correlation or covariance matrix to be analyzed. The spectral or eigen decomposition of $\mathbf{C}$ is

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i'$$

$$\mathbf{v}_i'\mathbf{v}_j = \delta_{ij} \qquad \text{(that is, orthonormality)}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$$

The eigenvectors $\mathbf{v}_i$ are also known as the principal components. The direction (sign) of principal components is not defined. `pca` returns principal components signed so that $\mathbf{1}'\mathbf{v}_i > 0$. In PCA, "total variance" equals $\text{trace}(\mathbf{C}) = \sum \lambda_j$.

Inference on eigenvalues and eigenvectors

The asymptotic distribution of the eigenvectors $\widehat{\mathbf{v}}_i$ and eigenvalues $\widehat{\lambda}_i$ of a covariance matrix $\mathbf{S}$ for a sample from a multivariate normal distribution $N(\mu, \mathbf{\Sigma})$ was derived by Girshick (1939); for more results, see also Anderson (1963) and Jackson (2003). Higher-order expansions are discussed in Lawley (1956). See Tyler (1981) for related results for elliptic distributions. The theory of the exact distribution is rather complicated (Muirhead 1982, chap. 9) and hard to implement. If we assume that eigenvalues of $\mathbf{\Sigma}$ are distinct and strictly positive, the eigenvalues and eigenvectors of $\mathbf{S}$ are jointly asymptotically multivariate normal distributed with the following moments (up to order n^{-3}):

$$E(\widehat{\lambda}_i) = \lambda_i \left\{ 1 + \frac{1}{n} \sum_{j \neq i}^{k} \left(\frac{\lambda_j}{\lambda_i - \lambda_j} \right) \right\} + O(n^{-3})$$

$$\text{Var}(\widehat{\lambda}_i) = \frac{2\lambda_i^2}{n} \left\{ 1 - \frac{1}{n} \sum_{j \neq i}^{k} \left(\frac{\lambda_j}{\lambda_i - \lambda_j} \right)^2 \right\} + O(n^{-3})$$

$$\text{Cov}(\widehat{\lambda}_i, \widehat{\lambda}_j) = \frac{2}{n^2} \left(\frac{\lambda_i \lambda_j}{\lambda_i - \lambda_j} \right)^2 + O(n^{-3})$$

$$\text{Var}(\widehat{\mathbf{v}}_i) = \frac{1}{n} \sum_{j \neq i}^{k} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{v}_j \mathbf{v}_j'$$

$$\text{Cov}(\widehat{\mathbf{v}}_i, \widehat{\mathbf{v}}_j) = -\frac{1}{n} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{v}_i \mathbf{v}_j'$$

For the asymptotic theory of the cumulative proportion of variance explained, see Kshirsagar (1972, 454).

More general tests for multivariate normal distributions

The likelihood-ratio χ^2 test of independence (Basilevsky 1994, 187) is

$$\chi^2 = -\left(n - \frac{2p+5}{6}\right) \ln\{\det(\mathbf{C})\}$$

with $p(p-1)/2$ degrees of freedom.

The likelihood-ratio χ^2 test of sphericity (Basilevsky 1994, 192) is

$$\chi^2 = -\left(n - \frac{2p^2+p+2}{6p}\right)\left[\ln\{\det(\widetilde{\mathbf{\Lambda}})\} - p\ln\left\{\frac{\text{trace}(\widetilde{\mathbf{\Lambda}})}{p}\right\}\right]$$

with $(p+2)(p-1)/2$ degrees of freedom and with $\widetilde{\mathbf{\Lambda}}$ the eigenvalues of the correlation matrix.

References

Afifi, A. A., S. May, and V. A. Clark. 2012. *Practical Multivariate Analysis*. 5th ed. Boca Raton, FL: CRC Press.

Anderson, T. W. 1963. Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* 34: 122–148.

Basilevsky, A. T. 1994. *Statistical Factor Analysis and Related Methods: Theory and Applications*. New York: Wiley.

Bontempi, M. E., and I. Mammi. 2015. Implementing a strategy to reduce the instrument count in panel GMM. *Stata Journal* 15: 1075–1097.

Dinno, A. 2009. Implementing Horn's parallel analysis for principal component analysis and factor analysis. *Stata Journal* 9: 291–298.

Girshick, M. A. 1939. On the sampling theory of roots of determinantal equations. *Annals of Mathematical Statistics* 10: 203–224.

Gorst-Rasmussen, A. 2012. tt: Treelet transform with Stata. *Stata Journal* 12: 130–146.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson. 2007. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology* 27: 1119–1152.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417–441, 498–520.

Jackson, J. E. 2003. *A User's Guide to Principal Components*. New York: Wiley.

Jolliffe, I. T. 2002. *Principal Component Analysis*. 2nd ed. New York: Springer.

Kshirsagar, A. M. 1972. *Multivariate Analysis*. New York: Dekker.

Lawley, D. N. 1956. Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika* 43: 128–136.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

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

Muirhead, R. J. 1982. *Aspects of Multivariate Statistical Theory*. New York: Wiley.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine, Series 6* 2: 559–572.

Rabe-Hesketh, S., and B. S. Everitt. 2007. *A Handbook of Statistical Analyses Using Stata*. 4th ed. Boca Raton, FL: Chapman & Hall/CRC.

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

Tyler, D. E. 1981. Asymptotic inference for eigenvectors. *Annals of Statistics* 9: 725–736.

van Belle, G., L. D. Fisher, P. J. Heagerty, and T. S. Lumley. 2004. *Biostatistics: A Methodology for the Health Sciences*. 2nd ed. New York: Wiley.

Weesie, J. 1997. smv7: Inference on principal components. *Stata Technical Bulletin* 37: 22–23. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 229–231. College Station, TX: Stata Press.

Also see

[MV] **pca postestimation** — Postestimation tools for pca and pcamat

[MV] **alpha** — Compute interitem correlations (covariances) and Cronbach's alpha

[MV] **biplot** — Biplots

[MV] **canon** — Canonical correlations

[MV] **factor** — Factor analysis

[D] **corr2data** — Create dataset with specified correlation structure

[R] **tetrachoric** — Tetrachoric correlations for binary variables

[U] 20 Estimation and postestimation commands