Principal Component Analysis

and

Factor Analysis

Principal Component Analysis and Factor Analysis

- Principal component analysis (PCA) and factor analysis are data reduction methods used to re-express multivariate data with fewer dimensions.
- The goal of these methods is to re-orient the data so that a multitude of original variables can be summarized with relatively few “factors” or “components” that capture the maximum possible information (variation) from the original variables.
- PCA is also useful in identifying patterns of association across variables.
- Factor analysis and principal component analysis are similar methods used for reduction of multivariate data; the difference between them is that factor analysis assumes the existence of a few common factors driving the variation in the data while principal component analysis does not make such an assumption.

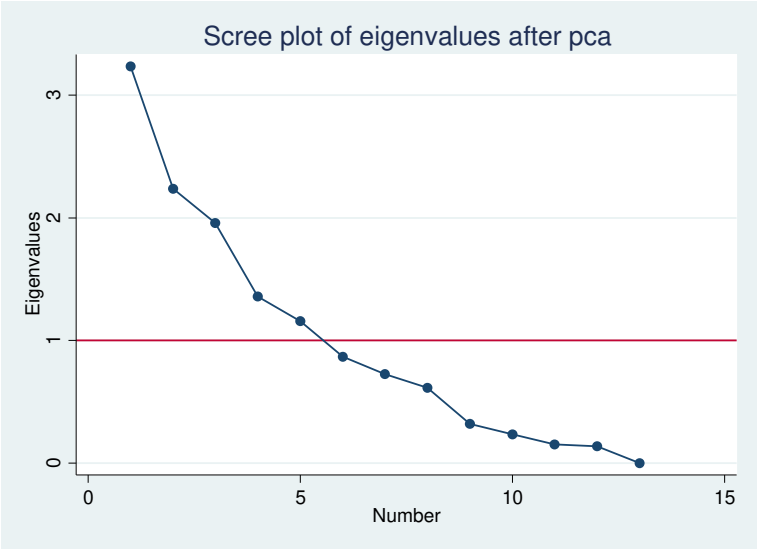
PCA Methodology

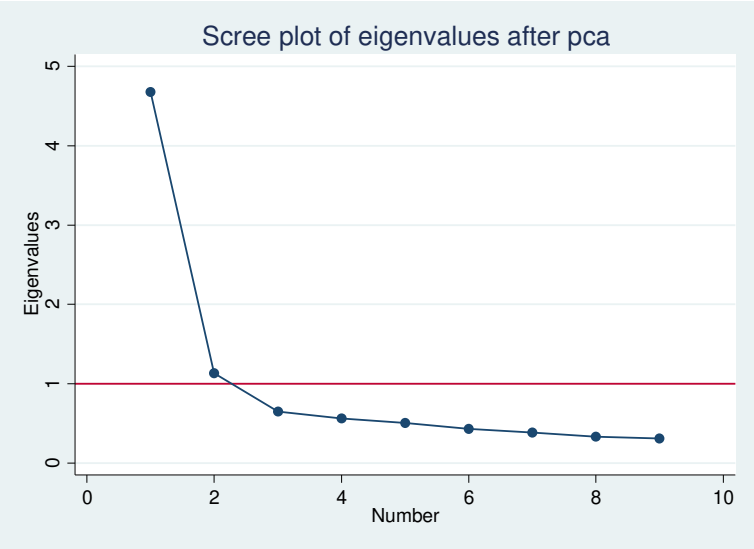
- The goal of PCA is to find components $z = [z_1, z_2, \ldots, z_p]$, which are a linear combination $u = [u_1, u_2, \ldots, u_p]'$ of the original variables $x = [x_1, x_2, \ldots, x_p]$ that achieve maximum variance.
- The first component z_1 is given by the linear combination of the original variables *x* and accounts for maximum possible variance. The second component captures most information not captured by the first component and is also uncorrelated with the first component.
- PCA seeks to maximize the variance so it is sensitive to scale differences in the variables. It is best to standardize the data and work with correlations rather than covariance among the original variables.
- PCA maximizes the variance of the elements of $z = xu$, such that $u'u = 1$.
- The solution is obtained by performing an eigenvalue decomposition of the correlation matrix, by finding the principal axes of the shape formed by the scatter plot of the data. The eigenvectors represent the direction of one of these principal axes.
- Solving the equation $(R - \lambda I)u = 0$, where *R* is the sample correlation matrix of the original variables *x*, λ is the eigenvalue, and *u* is the eigenvector.
- The eigenvalues λ are the variances of the associated components/factors *z*. The diagonal covariance matrix of the components is denoted as $D = diag(\lambda)$.

- The proportion of the variance in each original variable x_i accounted for by the first c factors is given by the sum of the squared factor loadings; that is, $\sum_{k=1}^{c} f_{ik}^2$. When $c=p$ (all components are retained), $\sum_{k=1}^{c} f_{ik}^2=1$ (all variation in the data are explained).
- Factor loadings are the correlations between the original variables x and the components/factors z, denoted as $F = cor(x, z) = uD^{1/2}$.
    - Because the factor loadings matrix shows the correlation between the factors and the original variables, typically the factors are named after the set of variables they are most correlated with.
    - The components can also be "rotated" to simplify the structure of the loadings matrix and the interpretations of the results.

Factor Retention

- Since principal components analysis and factor analysis are data reduction methods, there is a need to retain an appropriate number of factors based on the trade-off between simplicity (retaining as few as possible factors) and completeness (explaining most of the variation in the data).
- The Kaiser's rule recommends retaining only factors with eigenvalues λ exceeding unity. Intuitively, this rule means that any retained factor z should account for at least as much variation as any of the original variables x.
- In practice, the scree plot of the eigenvalues is examined to determine whether there is a "break" in the plot with the remaining factors explaining considerably less variation.





Factor Rotation

- The factor loadings matrix is usually "rotated" or re-oriented in order to make most factor loadings on any specific factor small while only a few factor loadings large in absolute value.
- This simple structure allows factors to be easily interpreted as the clusters of variables that are highly correlated with a particular factor. The goal is to find clusters of variables that to a large extent define only one factor.

Orthogonal rotation – preserves the perpendicularity of the axes (rotated components/factors remain uncorrelated)

- Varimax rotation –preserves simple structure by focusing on the columns of the factor loading matrix. The Kaiser's varimax rotation is an orthogonal rotation (preserving the independence of the factors) aiming to maximize the squared loadings variance across variables summed over all factors.
- Quartimax rotation – preserves simple structure by focusing on the rows of the factor loading matrix

Oblique rotation – allows for correlation between the rotated factors. The purpose is to align the factor axes as closely as possible to the groups of the original variables. The goal is to facilitate the interpretation of the results.

- Promax rotation

When to use Principal Component Analysis?

- Principal components analysis is undertaken in cases when there is a sufficient correlation among the original variables to warrant the factor/component representation.
- The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy takes values between 0 and 1, with small values indicating that overall the variables have little in common to warrant a principal components analysis and values above 0.5 are considered satisfactory for a principal components analysis.
- Bartlett's sphericity test examines whether the correlation matrix should be factored, i.e. the data are not independent. It is a chi-square test with a test statistic that is a function of the determinant of the correlation matrix of the variables.

Exploratory Factor Analysis

- Common factor model – observed variance in each measure is attributable to a relatively small number of common factors and a single specific factor (unrelated to other factors in the model).

$$X_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \cdots \lambda_{ic}\xi_c + \delta_i$$

  - The common factors ξ contribute to the variation in all variables *X*.
  - The specific factor δ can be thought of as the error term.

- Factor analysis is appropriate when there is a "latent trait" or "unobservable characteristics."
- The factor scores can be obtained from the analysis of dependence.
- Factor analysis is used with survey questions about attitudes – the goal is to identify common factors capturing the variance from these questions and which can also be used as factor scores.

- Assumptions to determine a solution to the common factor model:
  - The common factors are uncorrelated with each other.
  - The specific factors are uncorrelated with each other.
  - The common factors and specific factors are uncorrelated with each other.

- The communality is the proportion of variance in *X* attributable to the common factors

$$h_i^2 = \sum_k \lambda_{ik}^2 = 1 - \theta_{ii}^2$$

  where $\theta_{ii}^2 = var(\delta_i)$ is the factor uniqueness.
- The solution to the common factor model is determined by orienting the first factor so that it captures the greatest possible variance, then the second factor is obtained so that it captures the greatest possible variance but is uncorrelated with the first factor.

- The correlations between the X variables and the Ξ factors are called factor loadings Λ.
- The factor scores present the positions of the observations in the common factor space. The factor score coefficients are given by

$$B = R^{-1}\Lambda_c$$

  where R is the correlation matrix.
- The factor scores are calculated as:

$$\Xi = X_s B$$

- These factor scores are included in the data and can be used instead of the original variables.

Principal Component Analysis and Factor Analysis Example

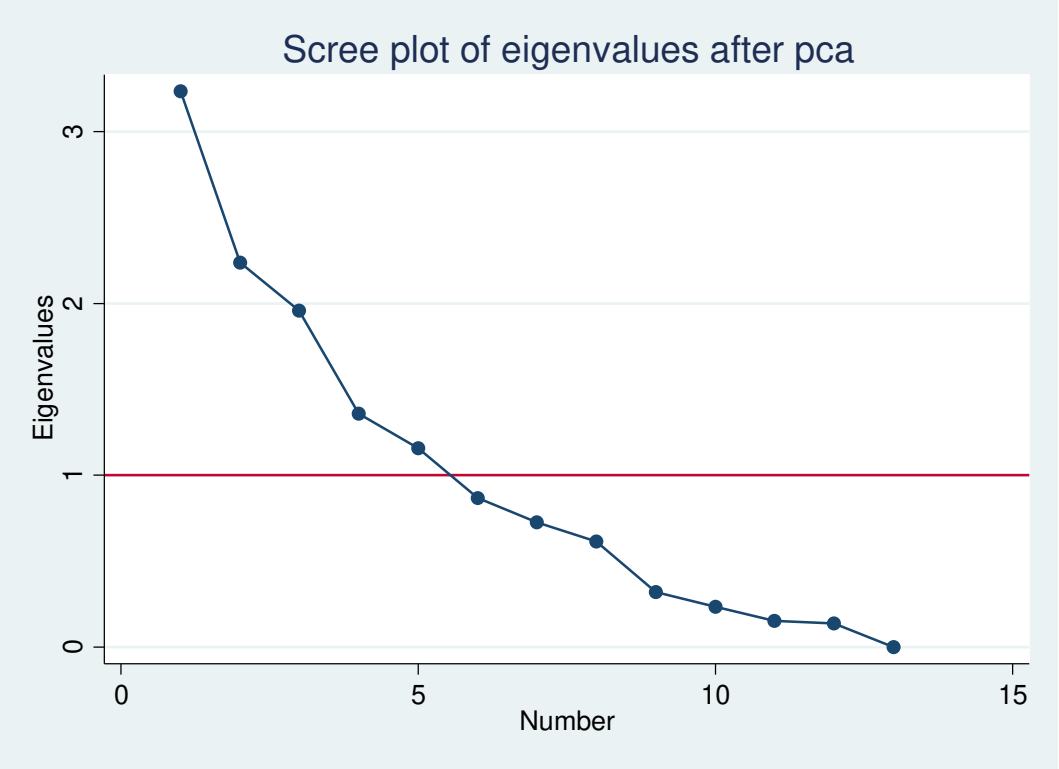
- We have data on gross state product from Lattin, Carroll, and Green (2003).
- Data are for 50 observations (U.S. states) and 13 categories (ag, mining, trade, etc.) for the gross state product expressed as shares.

Principal components, eigenvalues, and proportion of variance explained

Component	Eigenvalue	Difference b/n eigenvalues	Standard deviation (R)	Proportion of variance explained	Cumulative proportion of variance explained
Comp1	3.24	1.00	1.80	0.25	0.25
Comp2	2.24	0.28	1.50	0.17	0.42
Comp3	1.96	0.60	1.40	0.15	0.57
Comp4	1.36	0.20	1.17	0.10	0.68
Comp5	1.16	0.29	1.08	0.09	0.77
Comp6	0.87	0.14	0.93	0.07	0.83
Comp7	0.72	0.11	0.85	0.06	0.89
Comp8	0.62	0.30	0.78	0.05	0.94
Comp9	0.32	0.08	0.56	0.02	0.96
Comp10	0.24	0.08	0.49	0.02	0.98
Comp11	0.15	0.02	0.39	0.01	0.99
Comp12	0.14	0.14	0.37	0.01	1.00
Comp13	0.00	.	0.01	0.00	1.00

- Number of components equal to total number of variables (13).
- All 13 components explain the full variation in the data (1.00).
- The first 5 components have eigenvalues above 1 and explain 77% of variation.
- The first 3 components explain 57% of variation.

Scree plot



- First 5 components have eigenvalues above 1 (meaning that the component explains at least as much of the variation as the original variables).
- There is an "elbow" between components 3 and 5. We will use 3 components for the rest of the analysis (but using 5 components is also recommended).

Component loadings

	Component 1	Component 2	Component 3	Component 4	Component 5	Unexplained variation 5 components	Unexplained variation 3 components
Ag	0.13	-0.01	0.39	0.37	-0.41	0.27	0.65
Mining	0.47	0.00	-0.26	-0.07	-0.06	0.14	0.15
Constr	0.04	0.39	0.26	-0.35	-0.20	0.31	0.52
Manuf	-0.18	-0.38	0.38	-0.15	-0.11	0.26	0.30
Manuf_nd	-0.01	-0.46	0.04	0.05	0.47	0.27	0.53
Transp	0.42	0.15	0.01	0.37	-0.14	0.18	0.39
Comm	-0.15	0.32	-0.08	0.34	0.55	0.18	0.69
Energy	0.25	-0.14	0.07	-0.42	0.20	0.47	0.75
TradeW	-0.32	-0.03	0.29	0.44	0.01	0.25	0.51
TradeR	-0.09	0.26	0.51	-0.23	0.25	0.17	0.32
RE	-0.36	0.03	-0.45	0.01	-0.17	0.15	0.18
Services	-0.38	0.38	-0.13	-0.18	-0.13	0.11	0.17
Govt	0.29	0.37	0.09	0.08	0.29	0.30	0.41

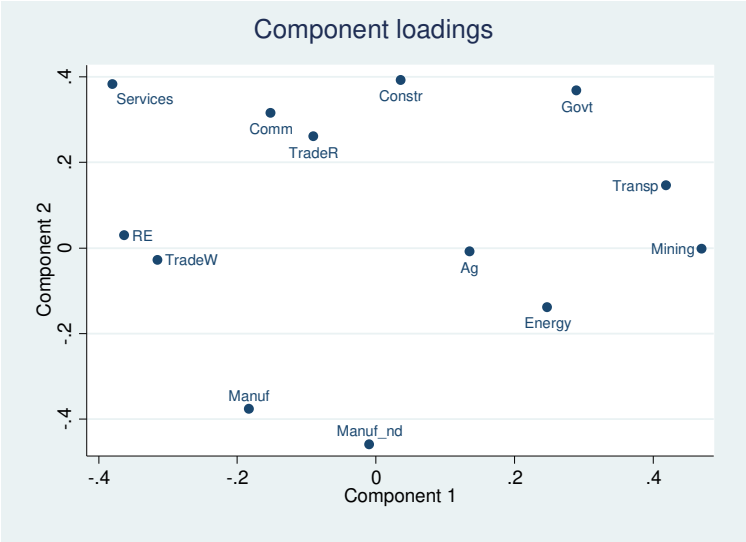
- The component loadings represent the correlation between the components and original variable.
- We concentrate on loadings above .3 or below -0.3.
- We can retain 3 or 5 components, also see the unexplained variation on ag, comm, and energy.

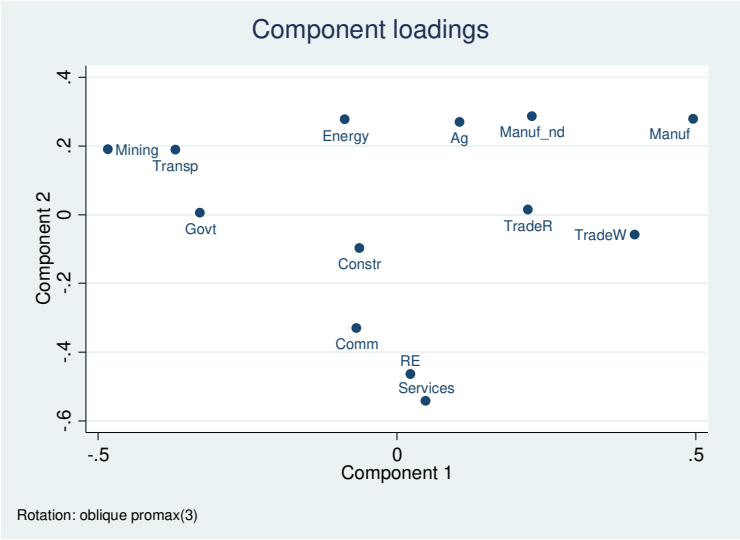
Component rotations (Stata)

	No rotation			Varimax rotation			Promax rotation		
	Comp 1	Comp 2	Comp3	Comp 1	Comp 2	Comp3	Comp 1	Comp 2	Comp3
Ag			0.39			0.31			0.33
Mining	0.47			-0.47			-0.48		
Constr		0.39				0.45			0.44
Manuf		-0.38	0.38	0.49			0.50		
Manuf_nd		-0.46							
Transp	0.42			-0.38			-0.37		
Comm		0.32			0.33			-0.33	
Energy									
TradeW	-0.32			0.39			0.40		
TradeR			0.51			0.55			0.56
RE	-0.36		-0.45		0.44	-0.37		-0.46	-0.39
Services	-0.38	0.38			0.54			-0.54	
Govt		0.37		-0.35		0.33	-0.33		0.31

- Principal components are only shown with loadings above 0.3 or below -0.3.
- The varimax and promax rotations give similar results – second component has reverse signs but still the same magnitude.
- Usually components/factors are “named” based on the highest loadings.

Component loadings rotations – no rotation, varimax, and promax





Factor loadings

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor uniqueness with 5 factors	Factor uniqueness with 3 factors	Factor commonality
Ag	0.24	-0.01	0.54	0.44	-0.44	0.26	0.65	0.73
Mining	0.85	0.00	-0.36	-0.08	-0.07	0.14	0.15	0.86
Constr	0.06	0.58	0.36	-0.40	-0.21	0.32	0.53	0.69
Manuf	-0.33	-0.56	0.52	-0.17	-0.12	0.25	0.30	0.74
Manuf_nd	-0.02	-0.69	0.05	0.05	0.50	0.27	0.52	0.73
Transp	0.75	0.22	0.01	0.43	-0.15	0.18	0.39	0.82
Comm	-0.27	0.47	-0.11	0.39	0.59	0.19	0.69	0.82
Energy	0.44	-0.21	0.09	-0.49	0.21	0.47	0.75	0.53
TradeW	-0.57	-0.04	0.41	0.51	0.01	0.25	0.51	0.75
TradeR	-0.16	0.39	0.71	-0.27	0.27	0.18	0.32	0.83
RE	-0.65	0.05	-0.63	0.02	-0.19	0.15	0.18	0.85
Services	-0.68	0.58	-0.18	-0.21	-0.14	0.11	0.17	0.89
Govt	0.52	0.55	0.12	0.09	0.32	0.30	0.41	0.70

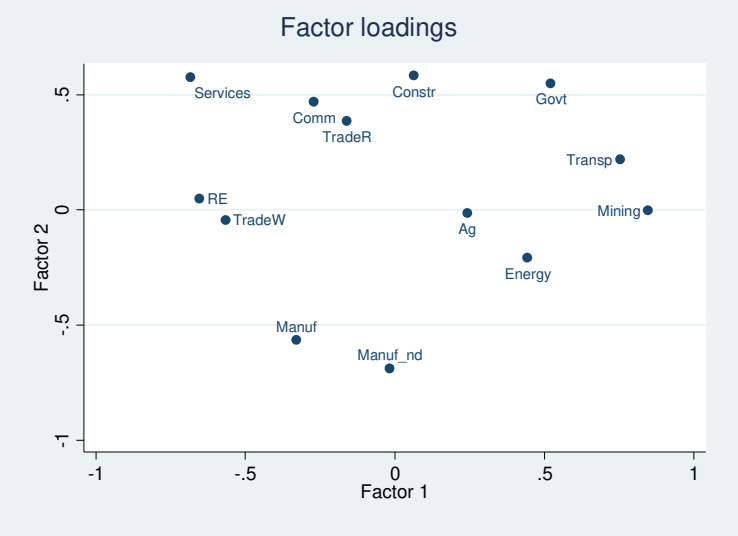
- Factor uniqueness is 1 minus the factor commonality.

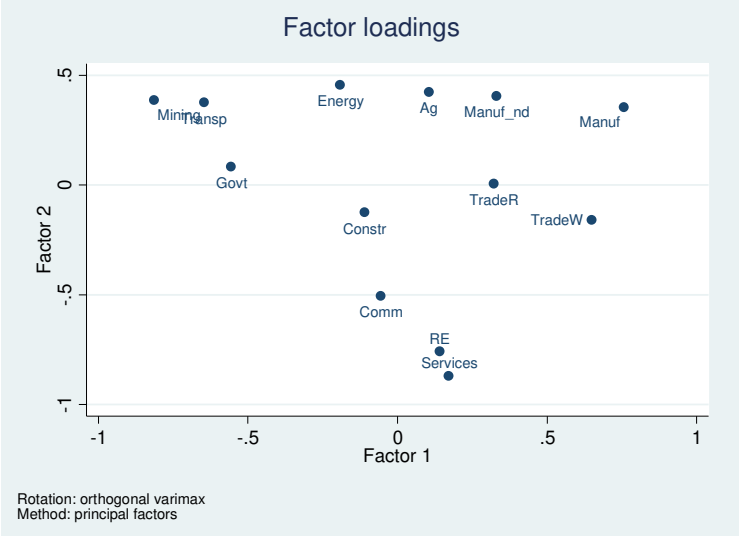
Factor rotations

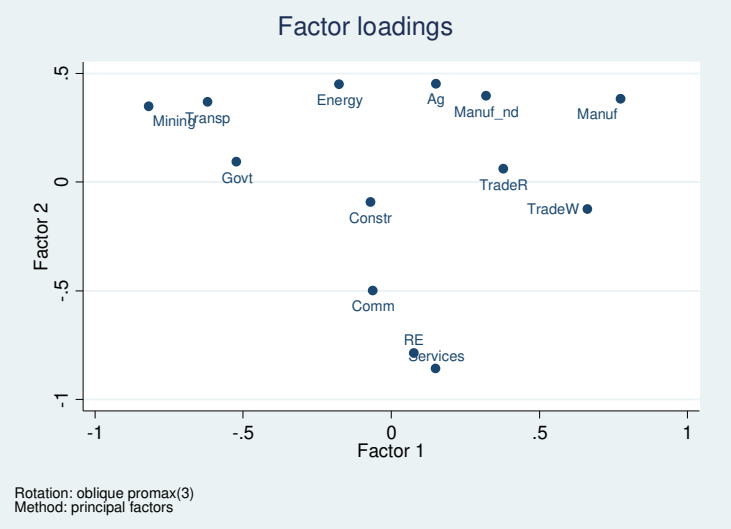
	No rotations (Stata)			Varimax (Stata)			Promax (Stata)		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Ag			0.54		0.42	0.40		0.45	0.44
Mining	0.85		-0.36	-0.81	0.39		-0.82	0.35	
Constr		0.58	0.36			0.67			0.66
Manuf	-0.33	-0.56	0.52	0.76	0.36		0.77	0.38	
Manuf_nd		-0.69		0.33	0.41	-0.45	0.32	0.40	-0.40
Transp	0.75			-0.65	0.38		-0.62	0.37	
Comm		0.47			-0.50			-0.50	
Energy	0.44				0.46			0.45	
TradeW	-0.57		0.41	0.65			0.66		
TradeR		0.39	0.71	0.32		0.76	0.38		0.79
RE	-0.65		-0.63		-0.76	-0.47		-0.79	-0.52
Services	-0.68	0.58			-0.87			-0.86	
Govt	0.52	0.55		-0.56		0.52	-0.52		0.49

- SAS produces different factor loadings after rotation.
- You can name factors: factor 1 is manufacturing and trade and facto 3 is ag and mining? Not clear groupings in this example.

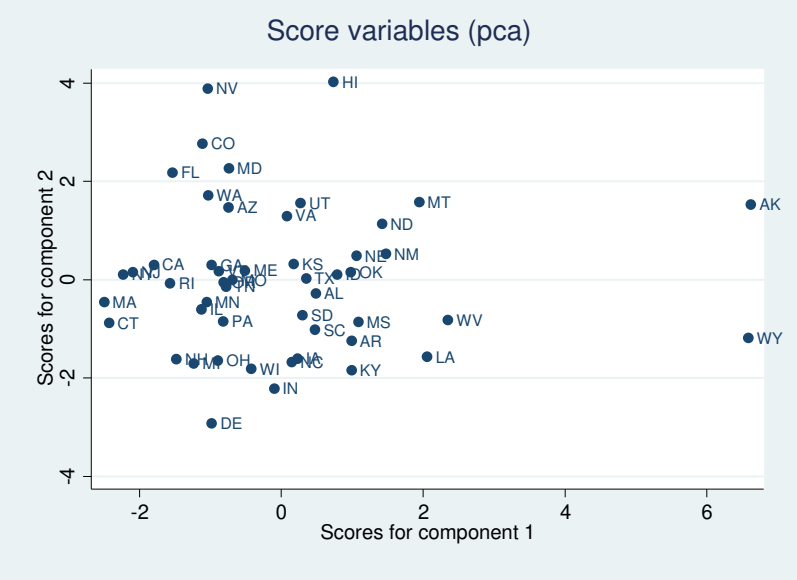
Factor loading rotations – no rotation, varimax, and promax

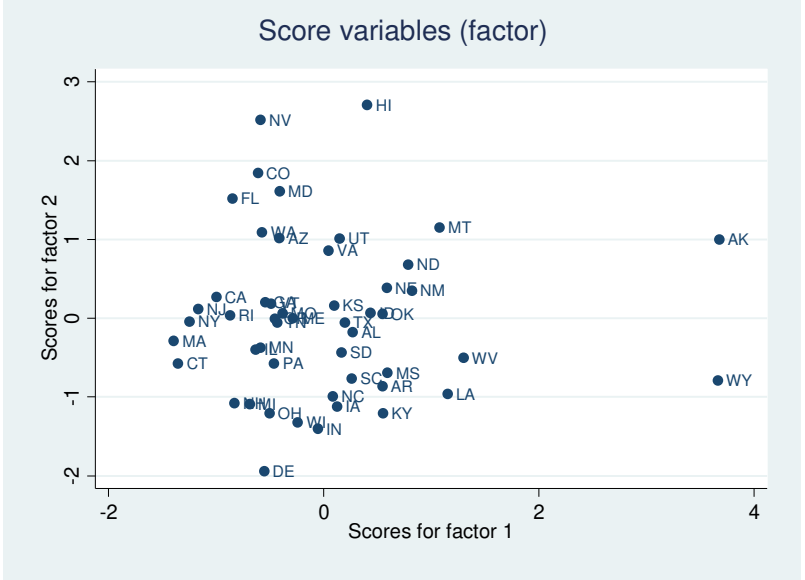






Plot of principal component scores for first two components (Stata)





- This gives an idea about the location of observations in the principal component space.
- Note that AK and WY are two outliers – high values on mining and transportation.

Scoring coefficients for the components and factors

	Comp 1	Comp 2	Comp 3	Factor 1	Factor 2	Factor 3
Ag	0.13	-0.01	0.39	0.02	-1.57	-1.42
Mining	0.47	0.00	-0.26	0.13	-3.81	-4.32
Constr	0.04	0.39	0.26	0.00	-0.33	-0.46
Manuf	-0.18	-0.38	0.38	-0.21	-3.35	-3.09
Manuf_nd	-0.01	-0.46	0.04	-0.10	-2.88	-2.77
Transp	0.42	0.15	0.01	0.19	-1.00	-1.19
Comm	-0.15	0.32	-0.08	-0.10	-0.35	-0.66
Energy	0.25	-0.14	0.07	0.11	-0.85	-0.77
TradeW	-0.32	-0.03	0.29	-0.20	-0.87	-0.72
TradeR	-0.09	0.26	0.51	-0.08	-0.64	-0.52
RE	-0.36	0.03	-0.45	-0.32	-3.41	-4.05
Services	-0.38	0.38	-0.13	-0.30	-2.18	-2.73
Govt	0.29	0.37	0.09	0.10	-1.56	-1.90

- The scoring coefficients are used to predict the values of the 3 principal components (or factors) for every observation in the data.

Predicting principal components and factors in the data (first few lines of data set)

State	pc1	pc2	pc3	f1	f2	f3
AL	0.48	-0.28	0.91	0.27	-0.18	0.67
AK	6.62	1.53	-2.70	3.67	1.00	-1.95
AZ	-0.74	1.47	0.86	-0.41	1.02	0.65
AR	0.99	-1.24	1.78	0.55	-0.86	1.25
CA	-1.80	0.31	-1.06	-0.99	0.27	-0.69
CO	-1.11	2.77	-0.13	-0.61	1.85	-0.05

- Instead of using the 13 variables, now 3 components or factors can be used to summarize the data.
- Note that the predicted values are for each observation (not the 13 categories like the component/factor loadings)

Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy

Ag	0.03
Mining	0.12
Constr	0.04
Manuf	0.06
Manuf_nd	0.04
Transp	0.10
Comm	0.04
Energy	0.04
TradeW	0.07
TradeR	0.05
RE	0.10
Services	0.11
Govt	0.07
Overall	0.07

- The values are less than 0.5 so overall the variables have little in common to warrant PCA.