THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

PolyU 理大商學院 Business School
Innovation-driven Education and Scholarship

School of ACCOUNTING & FINANCE 會計及金融學院

Week 10: Principal Component Analysis and Factor Analysis

AF3214 Python Programming for Accounting and Finance

Vincent Y. Zhuang, Ph.D.
vincent.zhuang@polyu.edu.hk

School of Accounting and Finance
The Hong Kong Polytechnic University

R508, 8:30 am – 11:20 am, Wednesdays, Semester 2, AY 2024-25

Agenda

- Overview of principal component analysis (PCA) and factor analysis
- PCA methodology
- Component/factor retention
- Component/factor rotation (orthogonal vs. oblique)
- When to use PCA

What are PCA and Factor Analysis

- PCA and factor analysis are data reduction methods used to re-express multivariate data with fewer dimensions.
- Meaning that we have a lot of variables in the dataset and we wonder if all of them should be used in an analysis or some of them might be redundant and you can express all of those variables with fewer factors or components.
- 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 PCA 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 PCA does not make such an assumption.

The Methodology of PCA

- The goal of PCA is to find components $z = [z_1, z_2, \dots, z_p]$, which are a linear combination $u = [u_1, u_2, \dots, u_p]$ of the original variables $x = [x_1, x_2, \dots, x_p]$ that achieve maximum variance.

- u_1 is the coefficient (or weight) for the 1st principal component z_1. Each element in u_1 corresponds to the contribution of original variable x_1 to 1st principal component z_i.

- 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 z_1 and is also uncorrelated with the first component z_1 .

- PCA seeks to maximize the variance of the elements of $z = xu$, such that $u'u = 1$ (constraint, unit vector). So it is sensitive to scale differences in variables. It is best to standardize the data and work with correlations rather than covariance among the original variables.

The Methodology of PCA – Cont'd

- The solution (to find the principal components) 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.
- Suppose we have two original variable x_1 and x_2, data is like a scatter plot. Can we find a different dimension that would summarize this variation in the best possible way?
- 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 $D = diag(\lambda)$.

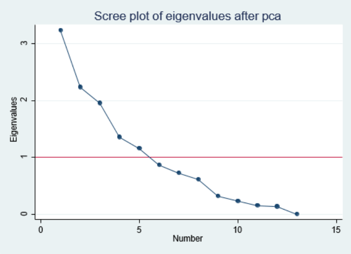
The Methodology of PCA – Cont'd

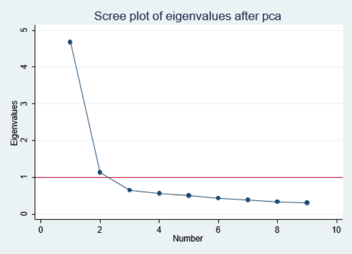
- The proportion of the variance in each original variable x_i accounted for by the first c factors is given by the sum of the squared factor loadings; that is: $\sum_{k=1}^{c} f_{ik}^2$. When c=p (all components are retained), $\sum_{k=1}^{c} f_{ik}^2 = 1$ (all variation in the data are explained).

- 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.

Component/Factor Retention

- Since PCA and factor analysis are data reduction methods, retain an appropriate number of factors based on the trade-off between simplicity (retaining as few factors as possible) and completeness (explaining most of the variation in the data) is important.
- The Kaiser's rule recommends retaining only factors with eigenvalues λ exceeding unity, meaning 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/elbow" in the plot with the remaining factors explaining considerably less variation.





Component/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. Highest possible correlations but fewest possible factors.
- 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–by focusing on the columns of the factor loading matrix.
- Quartimax rotation–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 PCA?

- PCA 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 values above 0.5 are considered satisfactory for a PCA.
- 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.

PCA Example

- Data: gross state product from Lattin, Carroll, and Green (2003)
- Data Structure: 50 observations (U.S. states) and 13 categories (agriculture, mining, trade, etc.) for the gross state product expressed as shares.

	State	Ag	Mining	Constr	Manuf	Manuf_nd	Transp	Comm	Energy	TradeW	TradeR	RE	Services	Govt
1	AL	2	1.5	4.2	10.5	11.8	2.9	2.9	3.6	6.3	9.9	12.8	16.1	15.5
2	AK	1.5	22.4	4.1	1.1	3.7	12.1	2	1.5	2.9	6.5	10.7	11.9	19.6
3	AZ	1.7	1.3	5.8	11.5	3	2.8	2.2	2.7	6.3	10.5	18.9	20.2	13
4	AR	5.1	1	4	12.8	11.8	4.4	2.4	4.2	6.1	10.2	11.4	14.8	11.8
5	CA	2.1	.6	3.3	9	5	2.6	2.5	1.8	6.8	8.9	22.7	23.1	11.5
6	CO	1.8	1.7	5.4	7.7	4.5	3.3	5.7	2.2	6.3	9.7	17	21.6	13.1
7	CT	.7	0	3.3	11	5.7	1.8	2.3	2.2	6.6	7.4	28.2	21.8	9
8	DE	1	0	3.4	4.5	16.6	1.6	1.3	2.4	4	6	35.4	14.3	9.4
9	FL	1.8	.2	4.7	4.6	3.5	3.1	3	2.8	7.3	11.2	21.8	23.4	12.4
10	GA	1.8	.4	3.9	7.4	10.7	4	4.5	2.7	8.8	8.9	16.4	18	12.5
11	HI	1.2	.1	4.8	.8	2.3	4.5	3.1	2.7	4	11.5	21.4	22.2	21.3
12	ID	6.3	.6	5.9	15	5.6	3.5	1.6	3.7	6.1	9.9	12.3	16.3	13.2
13	IL	1.4	.3	4.2	11.3	7.9	3.8	2.3	3.1	7.7	8.1	19.2	20.7	10
14	IN	1.8	.5	4.6	21.4	10.3	3.5	1.4	3.1	6	9.1	13.1	15.3	9.8
15	IA	7.6	.2	4.1	13.2	10.8	3.3	2	2.8	6.8	8.3	14.3	15.3	11.4
16	KS	4.4	1.4	4.2	10.4	7.9	3.9	3.6	3.4	7.8	9.6	12.7	16.7	14.1
17	KY	2.6	2.6	3.9	14.9	13.2	3.9	1.5	2.9	5.8	8.9	11.2	15	13.6
18	LA	1.2	14.8	4.2	3.7	15.3	3.3	1.9	3.6	5.3	7.8	12.1	15.7	10.9
19	ME	1.8	.1	4.5	7.9	10.6	2.3	2	3.1	6	11.1	18.5	18.7	13.5
20	MD	.9	.1	5	4.1	4.5	2.1	2.9	2.9	6.3	8.7	21.4	23.2	17.8

PCA Example

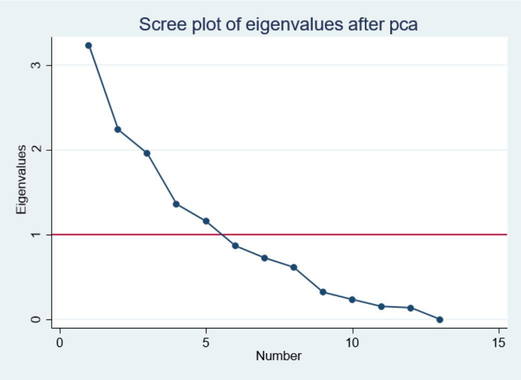
- Principal components, eigenvalues, proportion of variance explained…

Component	Eigenvalue	Difference between eigenvalues	Standard deviation	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).
- First 5 components have eigenvalues >1 and explain 77% of variation.
- The Kaiser's rule recommends retaining only factors with eigenvalues λ exceeding unity (e.g., 1).

PCA Example

- Let's plot all eigenvalues on a scree plot



- First 5 components have eigenvalues > 1 (the components explain at least 77% as much of the variation as the original variables).
- There is an "elbow" between components 3 & 5. We can either use 3 or 5 components for the rest of the analysis.

PCA Example

- 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.
- The sign of the loading simply indicates the direction of the correlation, not whether the variable is more or less important. Both large positive/negative loadings indicate variables that are important in defining that component.

PCA Example

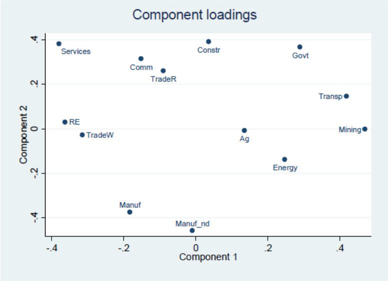
- Component rotations

	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 are "named" based on the highest loadings.
- Rotations on components to have the highest possible loadings on as few variables as possible to facilitate interpretations.

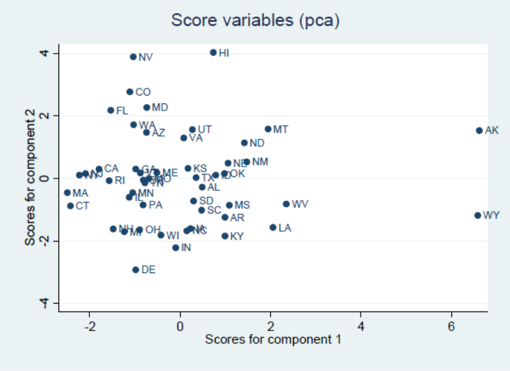
PCA Example

- Component loadings rotations – no rotation, varimax, and promax.



PCA Example

- Plot of principal component scores for first two components.



- 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.

PCA Example

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

State	pc1	pc2	pc3
AL	0.48	-0.28	0.91
AK	6.62	1.53	-2.70
AZ	-0.74	1.47	0.86
AR	0.99	-1.24	1.78
CA	-1.80	0.31	-1.06
CO	-1.11	2.77	-0.13

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

- 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 loadings)
- KMO measures show that values are less than 0.5, therefore overall the variables have little in common to warrant PCA.

The End