Session 9-10, Dimensionality Reduction and Derived Attributes (Technical Slides)

T. Evgeniou, A. Ovchinnikov, INSEAD

What is Dimensionality Deduction and Factor Analysis?

Derive new variables which are (linear) combinations of the original ones and capture most of the information in the original data.

Is often used as a first step in Data Analytics

Can also be used to solve multicollinearity issues in regression

Factor Analysis: Key idea

1. Transform the original selected variables into a smaller set of factors

2. Understand the underlying structure of the data and the new factors

3. Use the factors for subsequent analysis

Key Questions

1. Can we really simplify the data by grouping the raw attributes?

2. How many factors should we use?

3. How good are the factors we found?

4. How interpretable and actionable are the factors we found?

Dimensionality Reduction and Factor Analysis: 6 (Easy) Steps

1. Confirm data is metric

2. Scale the data

3. Check correlations

4. Choose number of factors

5. Interpret the factors

6. Save factor scores

Applying Factor Analysis: Evaluating MBA Applications

Variables available:

- GPA
- GMAT score
- Scholarships, fellowships won
- Evidence of Communications skills
- Prior Job Experience
- Organizational Experience
- Other extra curricular achievements

Which variables are correlated? What do these variables capture?

Example Factors

	Variables	Component 1	Component 2
1	GPA	0.96	-0.05
2	GMAT	0.95	0.19
3	Fellow	0.95	-0.01
4	Comm	0.7	0.54
5	Job.Ex	0.19	0.93
6	Organze	0.01	0.89
7	Extra	0.01	0.86

Step 1: Confirm data is metric

	Variables	GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
1	1	3	580	2	3.5	5	3.8	4
2	2	3.2	570	2	3.8	6	3.8	3.8
3	3	3.7	690	3	3.3	3	3.2	3.6
4	4	3.9	760	3	3.8	5	3.9	3.2
5	5	2.8	480	2	3.2	6	3.8	3.8
6	6	3.4	520	2.5	2.6	2	2.5	2.4
7	7	3.6	670	3	3.7	4	3.5	2.9
8	8	3.6	760	3	3.9	5	3.3	3.2

Step 2: Scale the data

	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	GPA	2.5	2.8	3.45	3.31	3.62	3.9	0.47
2	GMAT	380	480	575	583.5	682.5	760	119.44
3	Fellow	1	2	2.8	2.45	3	3.8	0.91
4	Comm	2	3.18	3.4	3.34	3.73	3.9	0.49
5	Job.Ex	2	3	5	4.25	5.25	6	1.52
6	Organze	1	3.05	3.4	3.2	3.8	3.9	0.73
7	Extra	2.4	2.88	3.4	3.3	3.8	4	0.52

Data Standardization: Example Code

```
ProjectDatafactor_scaled=apply(ProjectDataFactor,2,
 function(r) {
  if (sd(r)!=0) {
    res=(r-mean(r))/sd(r)
    } else {
      res=0*r; res
      }
  })
```

Standardized Data: Summary Statistics

	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	GPA	-1.72	-1.08	0.31	0	0.68	1.27	1
2	GMAT	-1.7	-0.87	-0.07	0	0.83	1.48	1
3	Fellow	-1.6	-0.5	0.39	0	0.61	1.49	1
4	Comm	-2.73	-0.33	0.13	0	0.8	1.16	1
5	Job.Ex	-1.48	-0.82	0.49	0	0.66	1.15	1
6	Organze	-2.99	-0.2	0.27	0	0.82	0.95	1
7	Extra	-1.75	-0.83	0.19	0	0.97	1.36	1

Step 3: Check correlations

GPA	GMAT	Fellow	Comm	Job.Ex	Organze	Extra
1	0.9	0.92	0.56	0.15	-0.03	0.01
0.9	1	0.86	0.78	0.33	0.19	0.16
0.92	0.86	1	0.59	0.18	0.01	0.02
0.56	0.78	0.59	1	0.6	0.47	0.39

Step 4. Choose number of factors

For the method considered here (Principal Component Analysis):

- If there are n variables we will have n factors in total
- First factor will explain most of the variance, second next and so on.

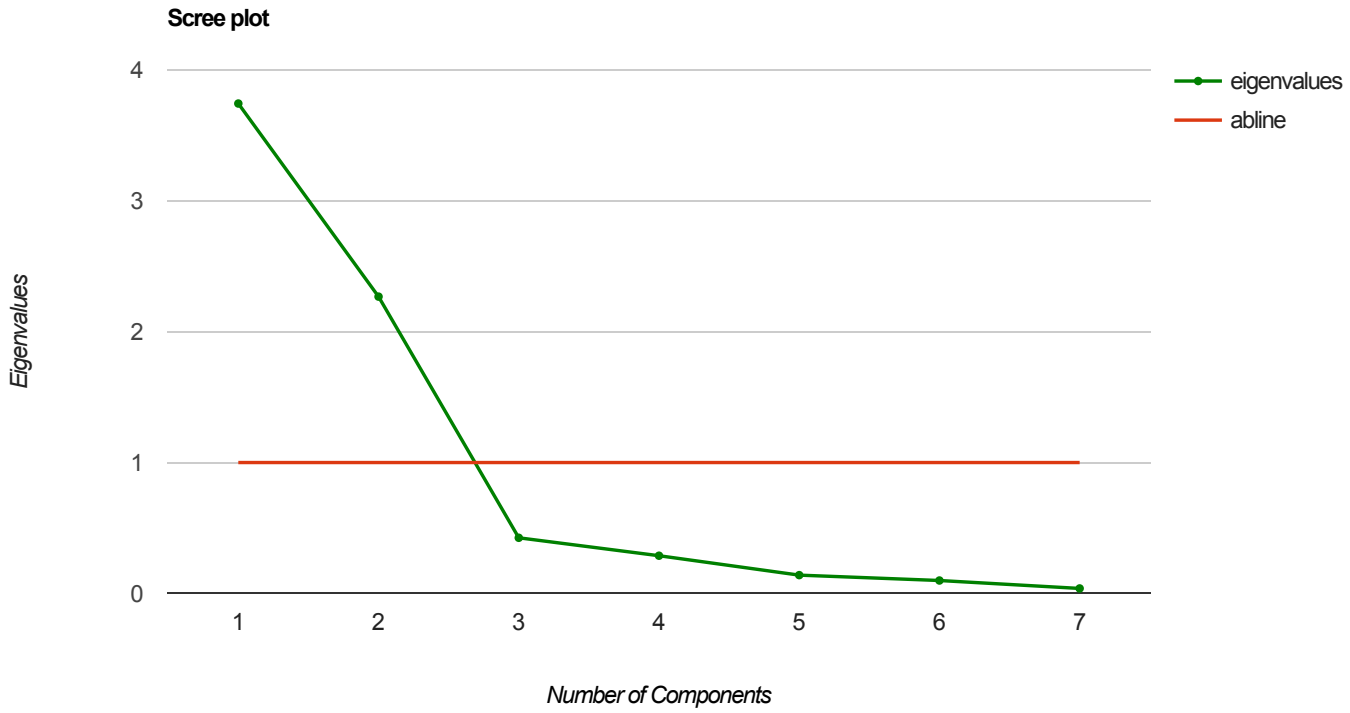
Eigenvalues and Variance Explained by Factors

- each factor will have an associated eigenvalue - which corresponds to the amount of variance explained by that factor
- with standardized variables each variable has a variance of 1, and the sum of all eigenvalues with n raw attributes is n
- we would like to capture as much of the total variance as possible, while keeping as few factors as possible

How Many Factors? Eigenvalues and Variance Explained

Components	Eigenvalue	Percentage_of_explained_variance	Cumulative_percentage_of_explained_variance
Component No:1	3.74	53.48	53.48
Component No:2	2.27	32.4	85.88
Component No:3	0.42	6.07	91.95
Component No:4	0.29	4.11	96.06
Component No:5	0.14	1.99	98.05
Component No:6	0.1	1.41	99.46
Component No:7	0.04	0.54	100

How Many Factors? Scree Plot



How many factors?

Three criteria to use:

- Eigenvalue > 1
- Cumulative variance explained
- “Elbow” in the Scree plot

Using the eigenvalue criterion we select 2 factors.

Step 5. Interpret the factors

Rotated Selected Factors using the varimax rotation.

	Variables	Component 1	Component 2
1	GPA	0.96	-0.05
2	GMAT	0.95	0.19
3	Fellow	0.95	-0.01
4	Comm	0.7	0.54
5	Job.Ex	0.19	0.93
6	Organze	0.01	0.89
7	Extra	0.01	0.86

For visualization, let's supress the small numbers...

	Variables	Component 1	Component 2
1	GPA	0.96	
2	GMAT	0.95	
3	Fellow	0.95	
4	Comm	0.7	0.54
5	Job.Ex		0.93
6	Organze		0.89
7	Extra		0.86

What Factor Loads ”Look Good”? Three Technical Quality Criteria

1. For each factor (column) only a few loadings are large (in absolute value)

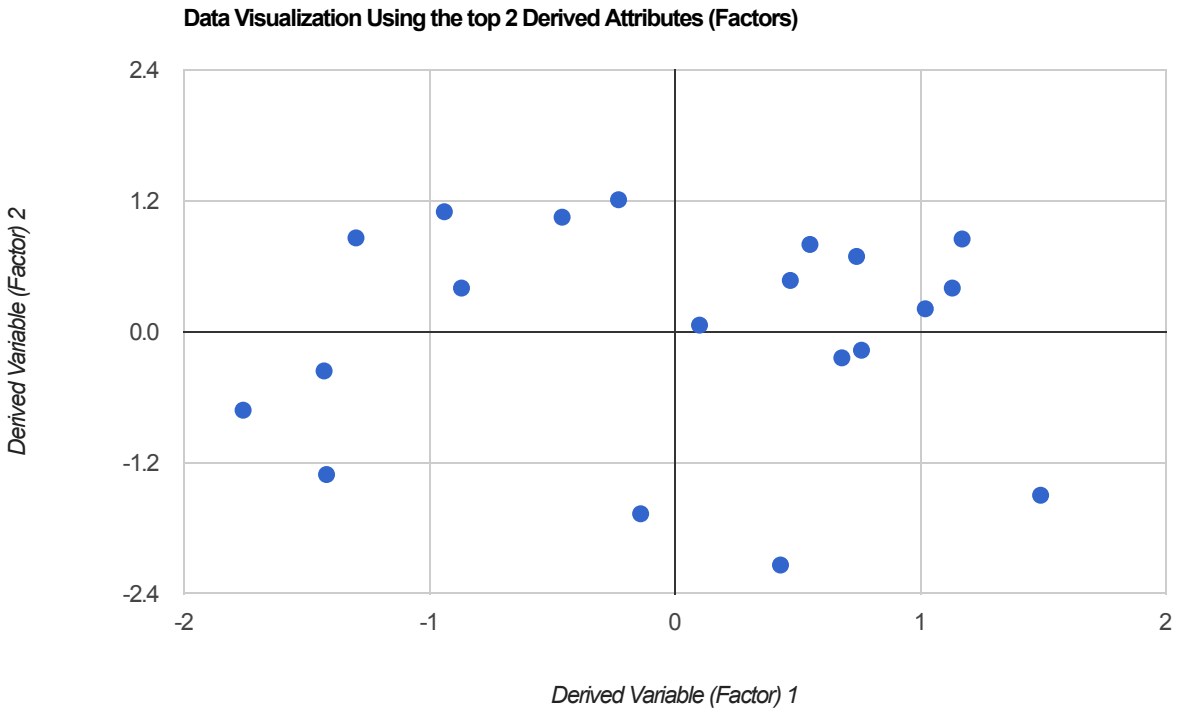
2. For each raw attribute (row) only a few loadings are large (in absolute value)

3. Any pair of factors (columns) should have different “patterns” of loading

Step 6. Save factor scores

	Observation	Derived Variable (Factor) 1	Derived Variable (Factor) 2
1	1	-0.46	1.05
2	2	-0.23	1.21
3	3	0.68	-0.24
4	4	1.13	0.4
5	5	-0.94	1.1
6	6	-0.14	-1.67
7	7	0.76	-0.17
8	8	1.02	0.21

Using the Factor Scores: Perceptual Maps



Factor Analysis: Some (Technical) Concepts

1. Correlation
2. Variance explained (eigenvalues)
3. Scree plot
4. varimax rotation
5. Factor Loadings (“components”)
6. Factor scores

Key Questions

1. How many factors should we use? Why? Quantitative and Qualitative criteria

2. How can we name and interpret the factors?

3. What are some issues to consider?