Session 9-10, Dimen ality Reduction and rived Attributes (Te cal Slides)

T. Evgeniou, A. Ovchinnikov, INSEA

What is Dimensionality Deduct Factor Analysis?

Derive new variables which are (linear) combir original ones and capture most of the informatior data.

Is often used as a first step in Data Anal

Can also be used to solve multicollinearity issues

Factor Analysis: Key ide

1. Transform the original selected variables into a factors

2. Understand the underlying structure of the da factors

3. Use the factors for subsequent analysis

Key Questions

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

2. How many factors should we use?

3. How good are the factors we found?

4. How interpretable and actionable are the facto

Dimensionality Reduction and
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

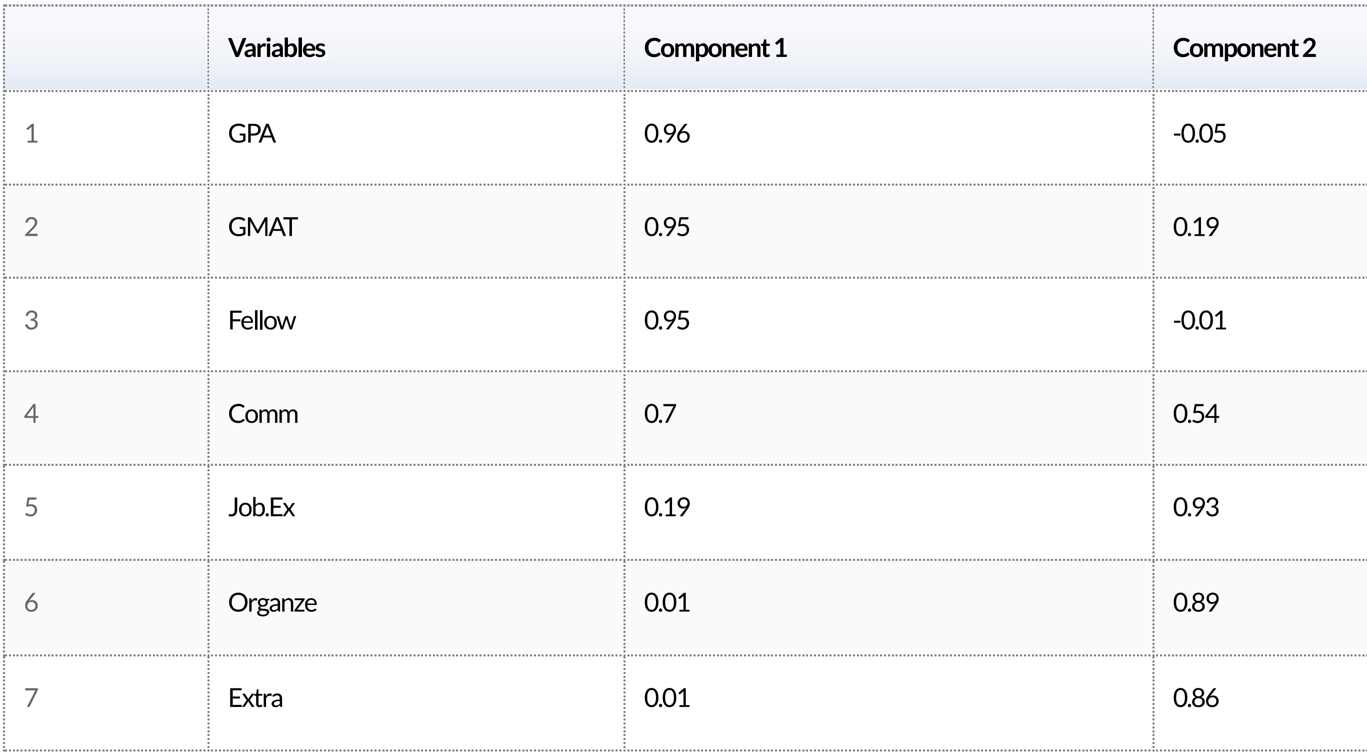
Applying Factor Analysis: Evalua Applications

Variables available:

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

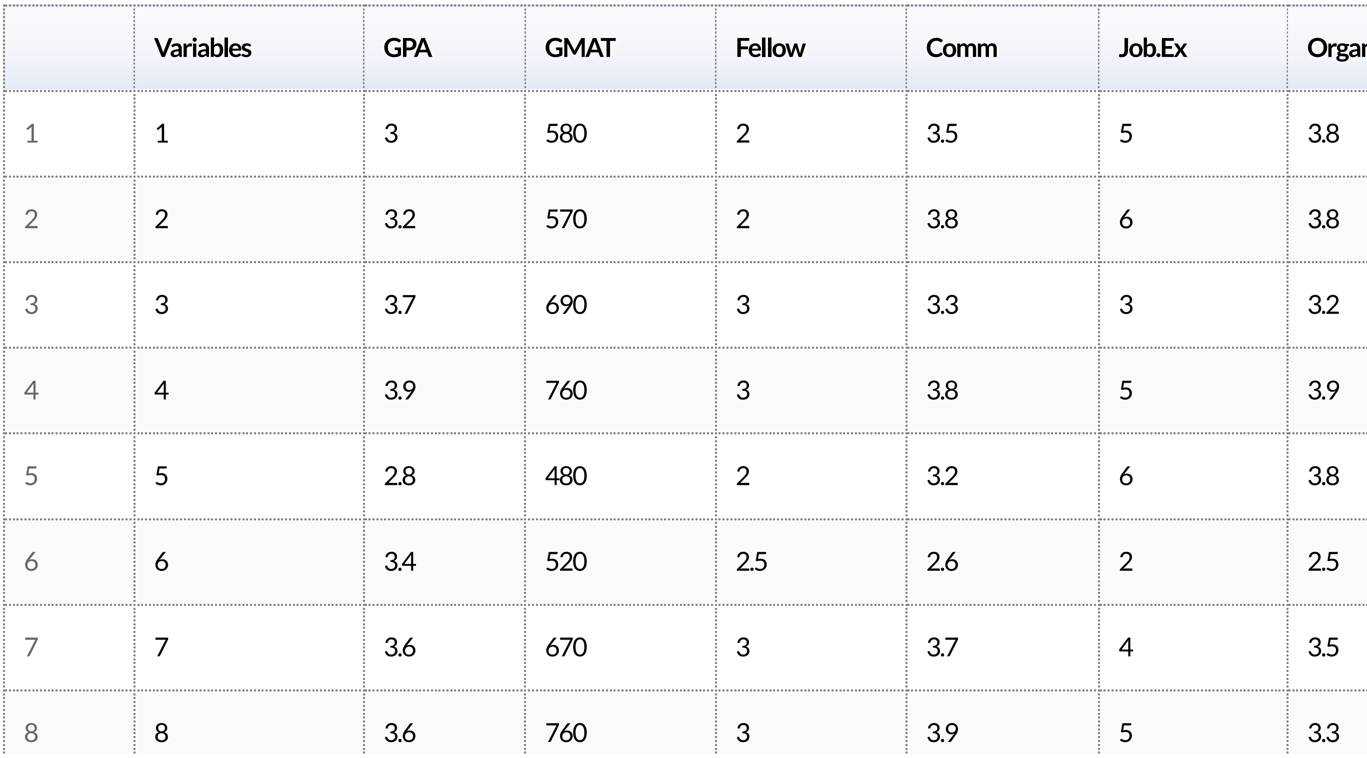
Which variables are correlated? What do thes capture?

Example Factors



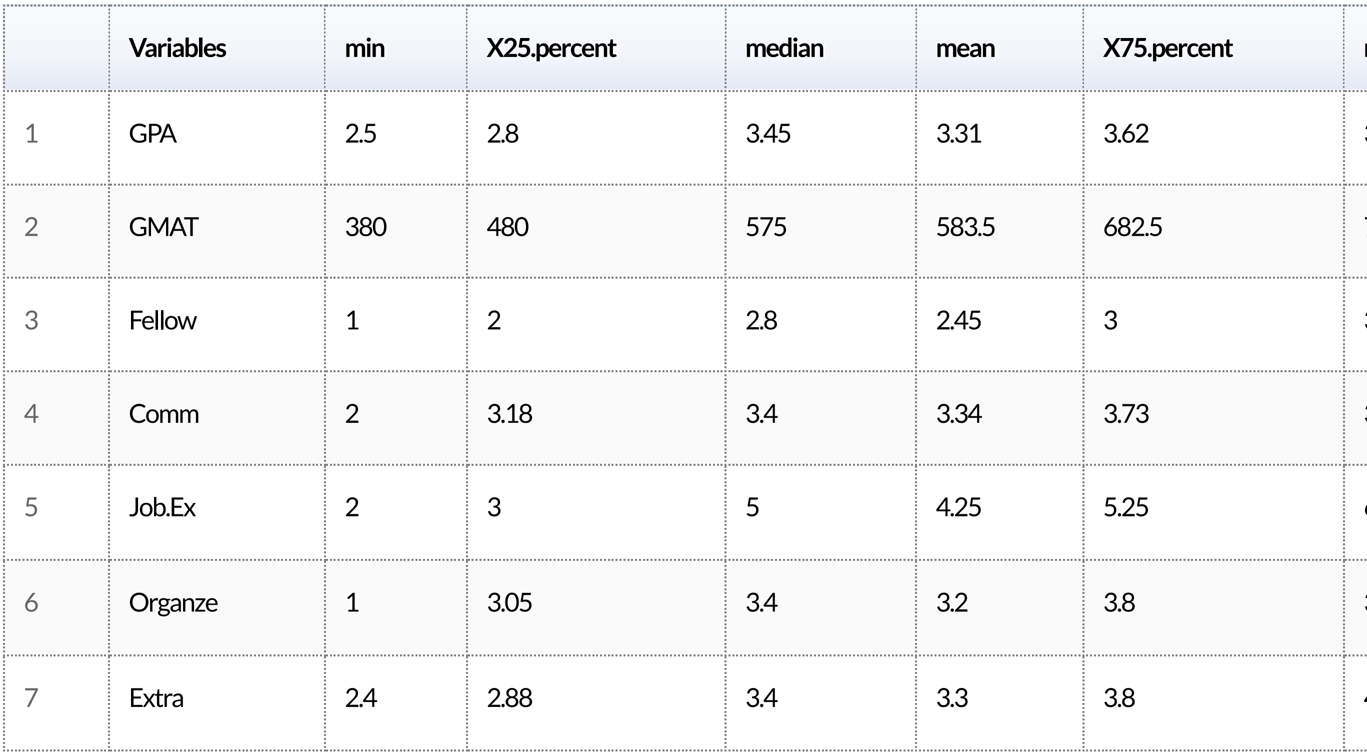
	Variables	Component 1	Component 2
1	GPA	0.96	-0.05
2	GMAT	0.95	0.19
3	Fellow	0.95	-0.01
4	Comm	0.7	0.54
5	Job.Ex	0.19	0.93
6	Organze	0.01	0.89
7	Extra	0.01	0.86

Step 1: Confirm data is me



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

Step 2: Scale the data

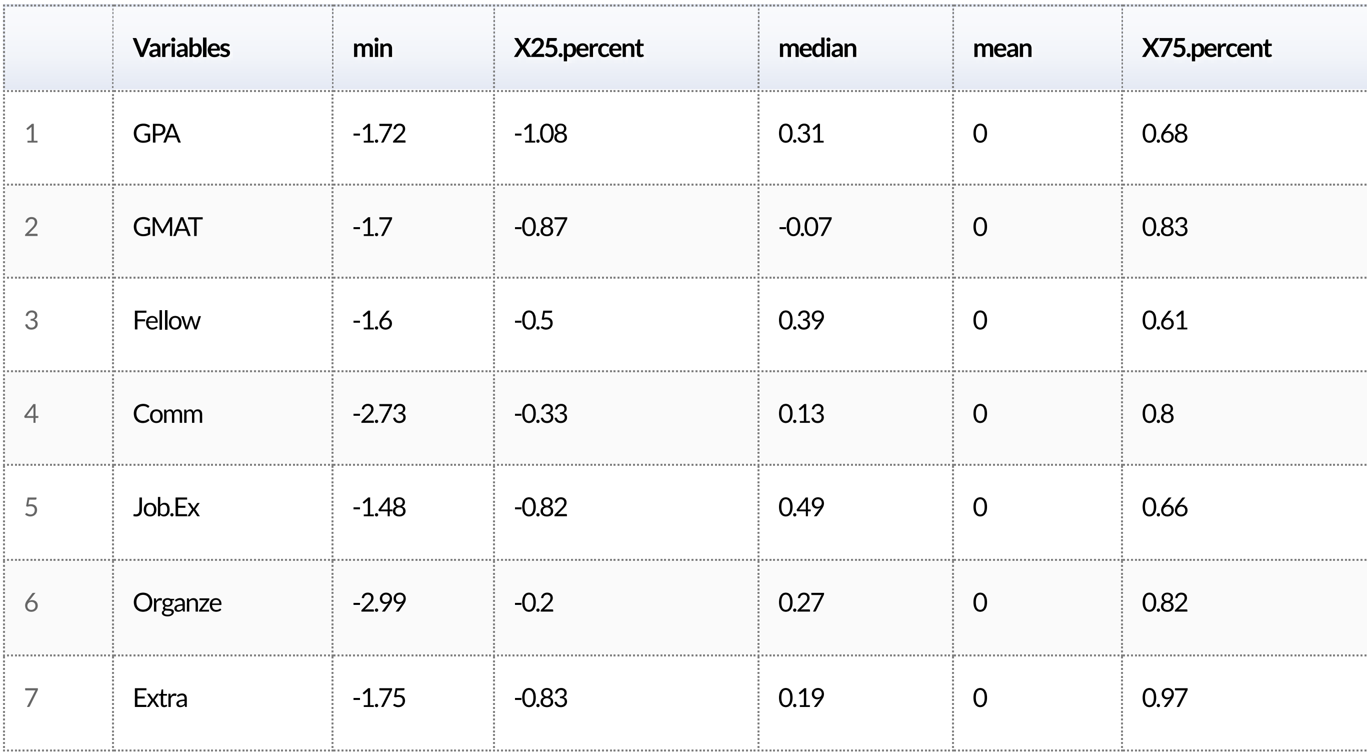


	Variables	min	X25.percent	median	mean	X75.percent
1	GPA	2.5	2.8	3.45	3.31	3.62
2	GMAT	380	480	575	583.5	682.5
3	Fellow	1	2	2.8	2.45	3
4	Comm	2	3.18	3.4	3.34	3.73
5	Job.Ex	2	3	5	4.25	5.25
6	Organze	1	3.05	3.4	3.2	3.8
7	Extra	2.4	2.88	3.4	3.3	3.8

Data Standardization: Exampl

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

Standardized Data: Summary S



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

Step 3: Check correlatio

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

Step 4. Choose number of fa

For the method considered here (Principal C
Analysis):

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

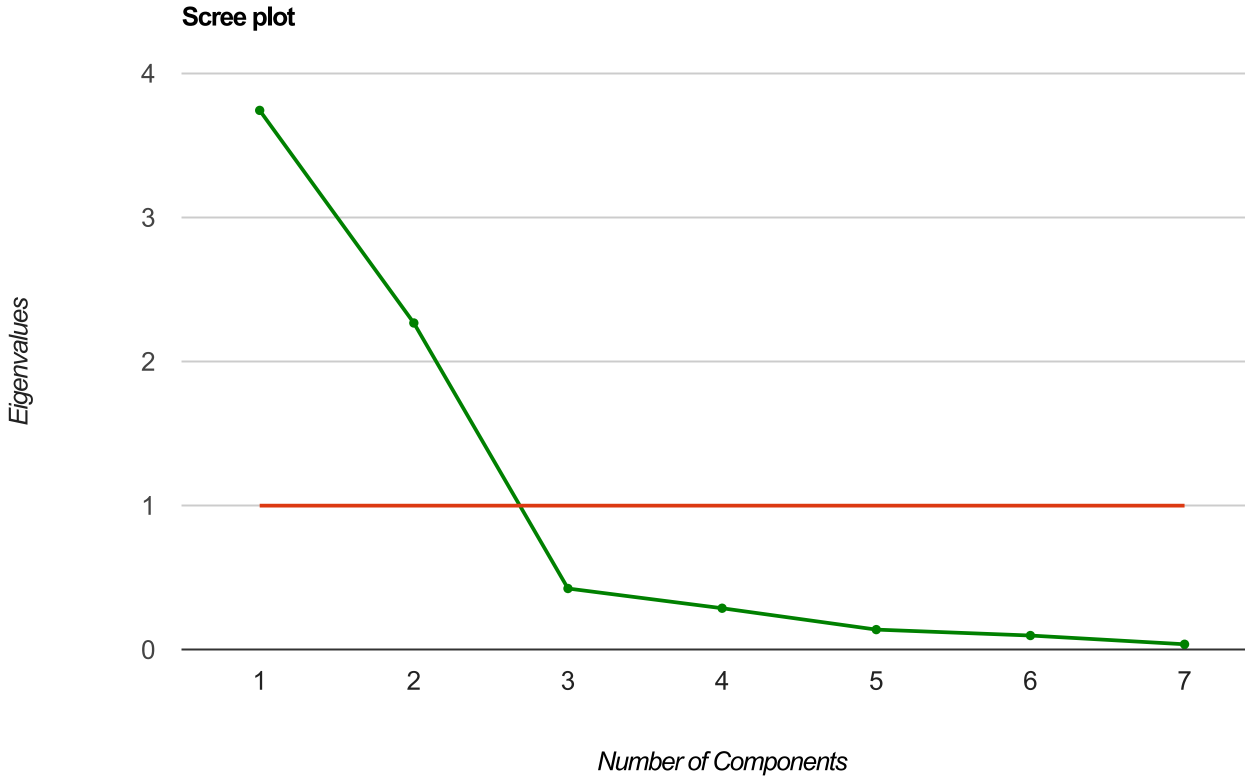
Eigenvalues and Variance Explained by F

- each factor will have an associated eigenvalue corresponds to the amount of variance explain factor
- with standardized variables each variable has a and the sum of all eigenvalues with n raw attrib
- we would like to capture as much of the total v possible, while keeping as few factors as possib

How Many Factors? Eigenvalues
ance Explained

Components	Eigenvalue	Percentage_of_explained_variance	Cumulative_percentage_of_expl
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



How many factors?

Three criteria to use:

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

Using the eigenvalue criterion we select 2

Step 5. Interpret the facto

Rotated Selected Factors using the varimax

	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 t
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

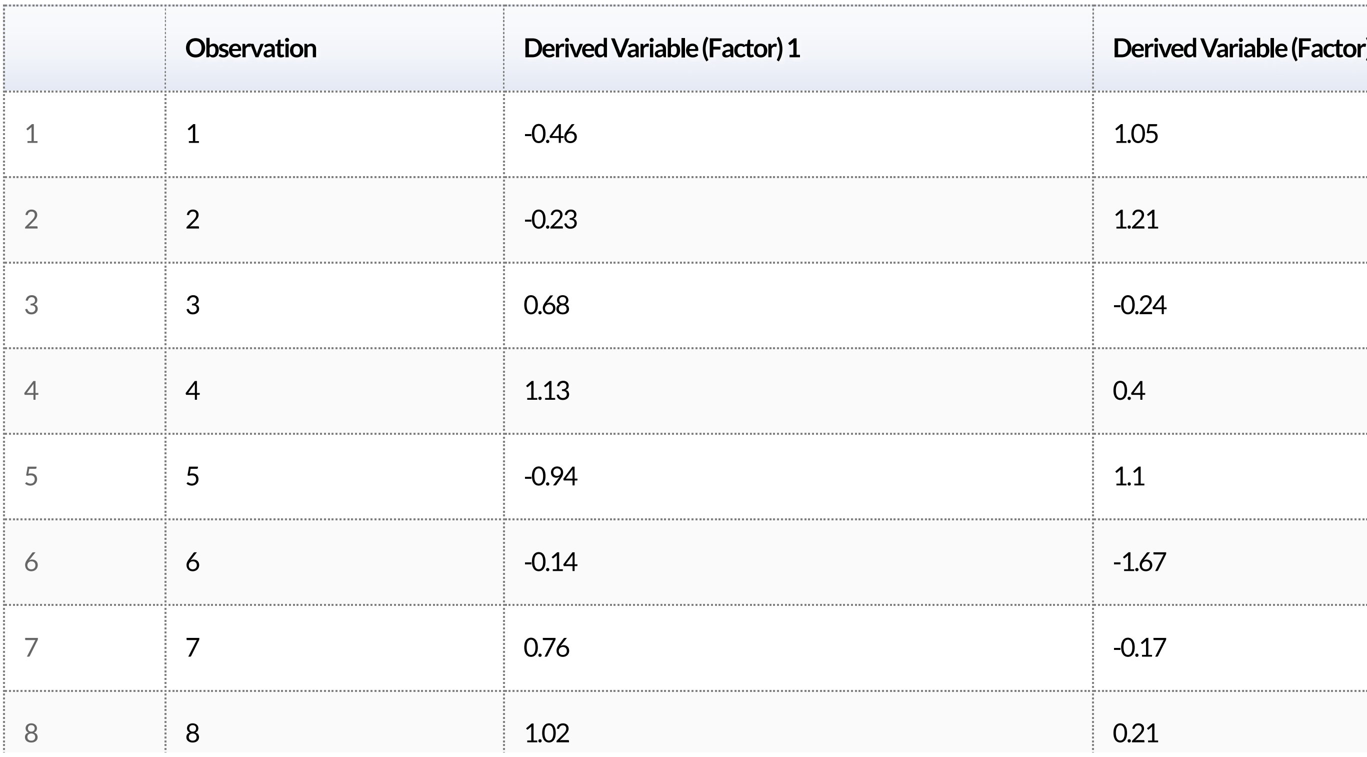
What Factor Loads ”Look Good
Technical Quality Criteri

1. For each factor (column) only a few loadings ar absolute value)

2. For each raw attribute (row) only a few loading absolute value)

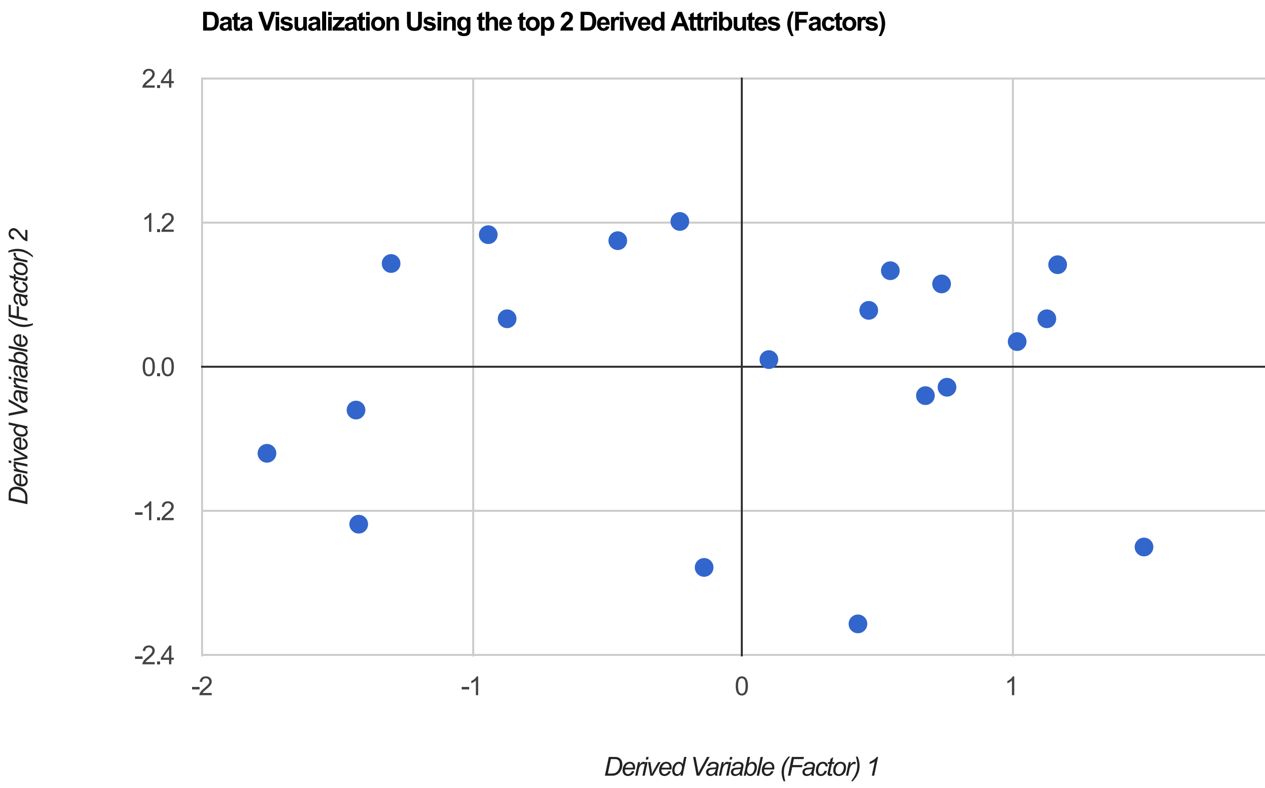
3. Any pair of factors (columns) should have diffe of loading

Step 6. Save factor scor



	Observation	Derived Variable (Factor) 1	Derived Variable (Factor
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: Percep



Factor Analysis: Some (Tech
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? Quan Qualitative criteria

2. How can we name and interpret the factors?

3. What are some issues to consider?