STEVENS INSTITUTE of TECHNOLOGY
School of Business

Amir H Gandomi; PhD
Assistant Professor
Stevens Institute of Technology
a.h.gandomi@stevens.edu

Multivariate Data Analysis – BIA 652

Class 10 – Factor Analysis

Overview of Class 11

- Factor Analysis – Chapter 15
- A Case Study
- Discussing about the Exam and sample questions.
- Important Dates:
  - Due date to submit Poster draft: **April 10, 2019**
  - Late term exam (in class for BIA-652-A): **April 15, 2019**
  - Submitting the final version of your poster: **April 20, 2019**
  - In class and oral presentation: **April 22 and 29, 2019**
  - Poster Event (Corporate Networking Event): **April 30, 2019**
  - Deadline to submit your complete project report: **May 6, 2019**

Where we are:

- If there is an outcome variable:
  - Perform a regression or discriminant analysis or logistical regression.
- To group observations:
  - Perform Cluster Analysis
- To restructure a group of variables:
  - Perform PCA or SVD
  - (Now) Factor Analysis

Factor Analysis

(A special case of Structural Equation Modeling)

Goals

- Generalization of Principal Components Analysis (similar areas of application)
- Explain interrelationships among a set of variables
- Explore the underlying structure of a set of variables.
- Select a small number of factors to convey essential information
- Perform additional analysis to improve interpretation

Exploratory and Confirmatory Factor Analysis

- **EFA vs CFA:**
  - EFA: Explore for a new model
  - CFA: Confirm a tested model
- More information:
  http://www2.sas.com/proceedings/sugi31/200-31.pdf

Factor Model

- **Start with *P* standardized variables (In the remainder we assume each x_i is a standardized variable)**
- **Express each variable as a linear combination of *m* common factors plus a unique factor**
- ***m* << *P*. Ideally *m* is known in advance**

Examples

- **Fifty test scores:**
  - Each is a function of $m = 3$ factors
  - Verbal, quantitative, analytical skills

- **Center for Epidemiologic Studies Depression (CESD) items:**
  - Each response is a function of some factors of depression

Model Equations

$$X_1 = l_{11} F_1 + l_{12} F_2 + \ldots + l_{1m} F_m + e_1$$

$$X_2 = l_{21} F_1 + l_{22} F_2 + \ldots + l_{2m} F_m + e_2$$

$$\ldots$$

$$X_p = l_{p1} F_1 + l_{p2} F_2 + \ldots + l_{pm} F_m + e_p$$

Factor model is the mirror image of the principal components model!

Terms

$$X_i = \sum l_{ij} F_j + e_i$$

F_j = Common or latent factors

e_i = Unique factors

l_{ij} = Coefficients of common factors
= Factor Loadings

Implications

- Variance of any original (X) variable is composed of:
  - Communality: part due to common factors, and
  - Specificity: part due to unique factor

- Variance X_i = V (X_i) = communality + specificity
  - V (X_i) = h_i^2 + u_i^2
  - V (X_i) = 1 when X's are standardized

Assumptions

- Each V (F_j) = 1
- F_j 's are uncorrelated
- F_j 's and e_i 's are uncorrelated

Steps on Factor Analysis

- **Initial factor extraction:**
  - **Estimate the loadings and communalities**
- **Factor "rotations" to improve interpretation**

Example data model (known)

100 data points generated from five variables with multivariate normal distribution.

$$X_1 = 1 * F_1 + 0 * F_2 + e_1$$
$$X_2 = 1 * F_1 + 0 * F_2 + e_2$$
$$X_3 = 0 * F_1 + 0.5 * F_2 + e_3$$
$$X_4 = 0 * F_1 + 1.5 * F_2 + e_4$$
$$X_5 = 0 * F_1 + 2 * F_2 + e_5$$

Example - Implications

- F_1 , F_2 and all e_i 's are independent, normal variables
- Therefore: the first 2 X's are inter-correlated, and the last 3 X's are inter- correlated
- And: The first 2 X's are not correlated with the last 3 X's

Means and Correlations

Means: 0.163, 0.142, 0.098, -0.039, -0.013

Correlation Matrix

1.0				
0.757	1.0			
0.047	0.054	1.0		
0.115	0.176	0.531	1.0	
0.279	0.322	0.521	0.942	1.0

Steps on Factor Analysis

- **Initial factor extraction:**
  - **Estimate the loadings and communalities**
- **Factor "rotations" to improve interpretation**

Initial Factor Extraction

Initial Factor Extraction

- **Principal Component Factor Model**
- **Iterated Principal Factor Model**
- **Maximum Likelihood Model**

Principal Component Factor Model

- **Recall – Principal Component Model:**
  - **C = AX, PC's are Functions of X's**

    Basic Idea:

    X_1 , X_2 – correlated

    Transform them into:

    C_1 , C_2 – uncorrelated

- **We want: X's = Functions of F's**

Inverse Model

- **If C = 5X, then X = (1/5)C, or $X = 5^{-1} C$**

- **$C = AX \rightarrow X = A^{-1} C$**
  - The Inverse of Principal Components Model is $X = A^{-1} C$
  - In this case, A is an orthogonal matrix, Therefore: $A^{-1} = A^T$

and

$$x_1 = a_{11} C_1 + a_{21} C_2 + \ldots + a_{p1} C_p$$

. . .

$$x_m = a_{m1} C_1 + a_{m2} C_2 + \ldots + a_{mp} C_p$$

PC Factor Model Derivation

$$x_i = \sum_{j=1}^{p} a_{ji} C_j$$

$$x_i = \sum_{j=1}^{m} a_{ji} C_j + \sum_{j=m+1}^{p} a_{ji} C_j$$

$$x_i = \sum_{j=1}^{m} l_{ji} F_j + e_i$$

F_j = Common or latent factors

e_i = Unique factors

l_{ij} = Coefficients of common factors
= Factor Loadings

Interpretation

- Var (C_j) = λ_j NOT 1
- Transform: $F_j = C_j / \lambda_j^{1/2}$
- Therefore: Var (F_j) = 1
- And loadings are: $l_{ij} = (\lambda_j^{1/2})(a_{ji})$
  - l_{ij} is the correlation coefficient between variable i and factor j
- $x_i = \sum_{j=1}^{m} (\lambda_j^{1/2})(a_{ji}) \; C_j / \lambda_j^{1/2} + e_i$
- $x_i = \sum_{j=1}^{m} l_{ji} \; F_j + e_i$

Selecting a number of factors

- **Common Approach**
  - The number of factors is the number of eigenvalues greater than one
- **In Previous Example:**
  - Variances of the principal components are:

    2.578, 1.567, 0.571, 0.241, 0.043
  - Select m = 2 factors

Previous Example PC Method (p 386)

Initial factor analysis summary for hypothetical data set from principal components extraction method

Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2	Specificity u_i^2
x_1	0.511	0.782	0.873	0.127
x_2	0.553	0.754	0.875	0.125
x_3	0.631	−0.433	0.586	0.414
x_4	0.866	−0.386	0.898	0.102
x_5	0.929	−0.225	0.913	0.087
Variance explained	2.578	1.567	$\sum h_i^2 = 4.145$	$\sum u_i^2 = 0.855$
Percentage	51.6	31.3	82.9	17.1

Reading the Output (p 385)

- **The factor model:**

$$X_1 = 0.511F_1 + 0.782F_2 + e_i$$
$$X_2 = 0.553F_1 + 0.754F_2 + e_2$$
...

- **Communality:**

For X_1 : $h_1^2 = 0.873$
For X_2 : $h_2^2 = 0.875$

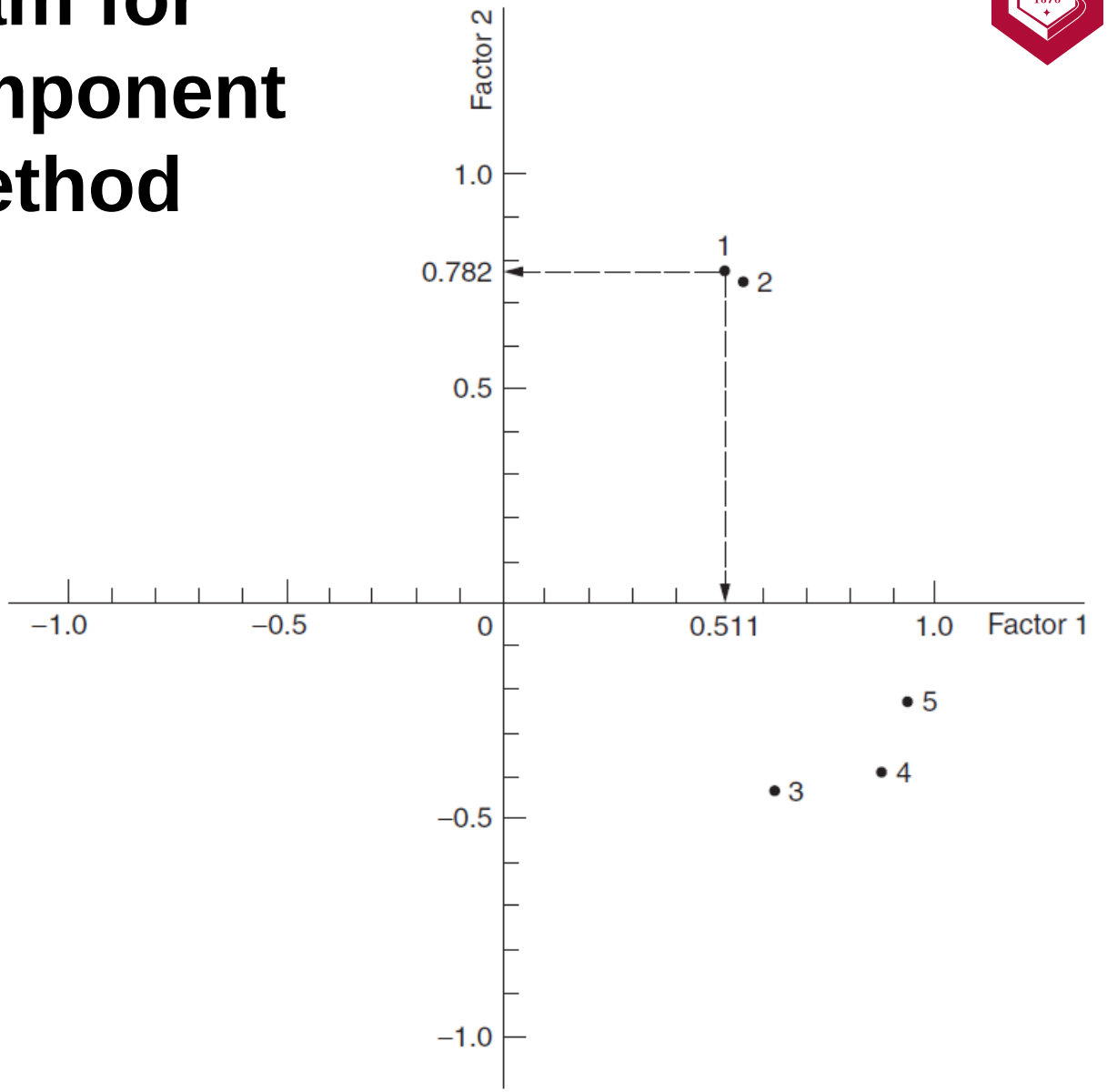
- **Specificity = 1 - Communality**

Implications

- **1st row: $h_1^2 = 0.87 = .51^2 + .78^2$, etc.**
- **1st column: Variance Explained**

  $= 2.58 = .51^2 + .55^2 + \ldots$, etc.
- **Variance Explained = eigenvalue**
- **$\sum$ Variance Explained = $\sum h_i^2$ = total variance explained by common factors = 4.145 = 83% of total variance.**

Factor Diagram for Principal Component Extraction Method (p 387)



Initial Factor Extraction Method 2 Iterated Principal Factors (IPF)

- **Select common factors to maximize the total communality**

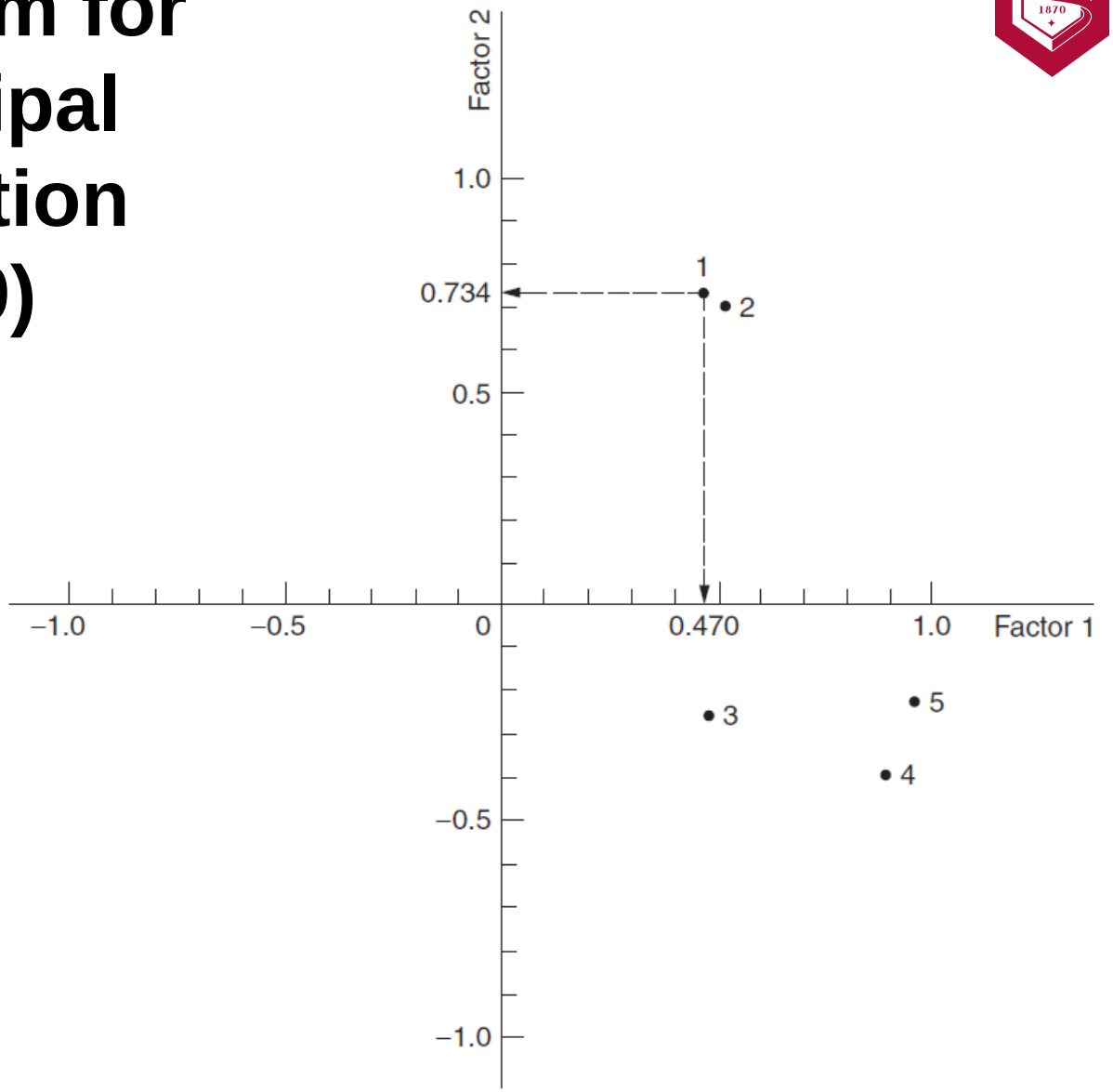
- Use iterative procedure:
  1. Get initial communality estimates
  2. Use these (instead of original variances) to get Principal Components (from the modified matrix) and factor loadings (From multiplying the principal components coefficients by the standard deviation of the respective principal components)
  3. Get new communality estimates
  4. If appreciable change, go to step 2,
  5. Else, stop.

Example: IPF Method (p 388)

Initial factor analysis summary for hypothetical data set from iterated principal factor extraction method

Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2	Specificity u_i^2
x_1	0.470	0.734	0.759	0.241
x_2	0.510	0.704	0.756	0.244
x_3	0.481	−0.258	0.598	0.702
x_4	0.888	−0.402	0.949	0.051
x_5	0.956	−0.233	0.968	0.032
Variance explained	2.413	1.317	$\sum h_i^2 = 3.730$	$\sum u_i^2 = 1.270$
Percentage	48.3	26.3	74.6	25.4

Factor Diagram for Iterated Principal Factor Extraction Method (p 389)



Comparison: PCF vs IPF

PCF:

Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2	Specificity u_i^2
x_1	0.511	0.782	0.873	0.127
x_2	0.553	0.754	0.875	0.125
x_3	0.631	−0.433	0.586	0.414
x_4	0.866	−0.386	0.898	0.102
x_5	0.929	−0.225	0.913	0.087
Variance explained	2.578	1.567	$\sum h_i^2 = 4.145$	$\sum u_i^2 = 0.855$
Percentage	51.6	31.3	82.9	17.1

IPF:

Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2	Specificity u_i^2
x_1	0.470	0.734	0.759	0.241
x_2	0.510	0.704	0.756	0.244
x_3	0.481	−0.258	0.598	0.702
x_4	0.888	−0.402	0.949	0.051
x_5	0.956	−0.233	0.968	0.032
Variance explained	2.413	1.317	$\sum h_i^2 = 3.730$	$\sum u_i^2 = 1.270$
Percentage	48.3	26.3	74.6	25.4

Steps on Factor Analysis

- **Initial factor extraction:**
  - **Estimate the loadings and communalities**
- **Factor "rotations" to improve interpretation**

Factor Rotations

Find new factors that are easier to interpret

Orthogonal Rotations

- **Find new uncorrelated (orthogonal) factors that are easier to interpret:**

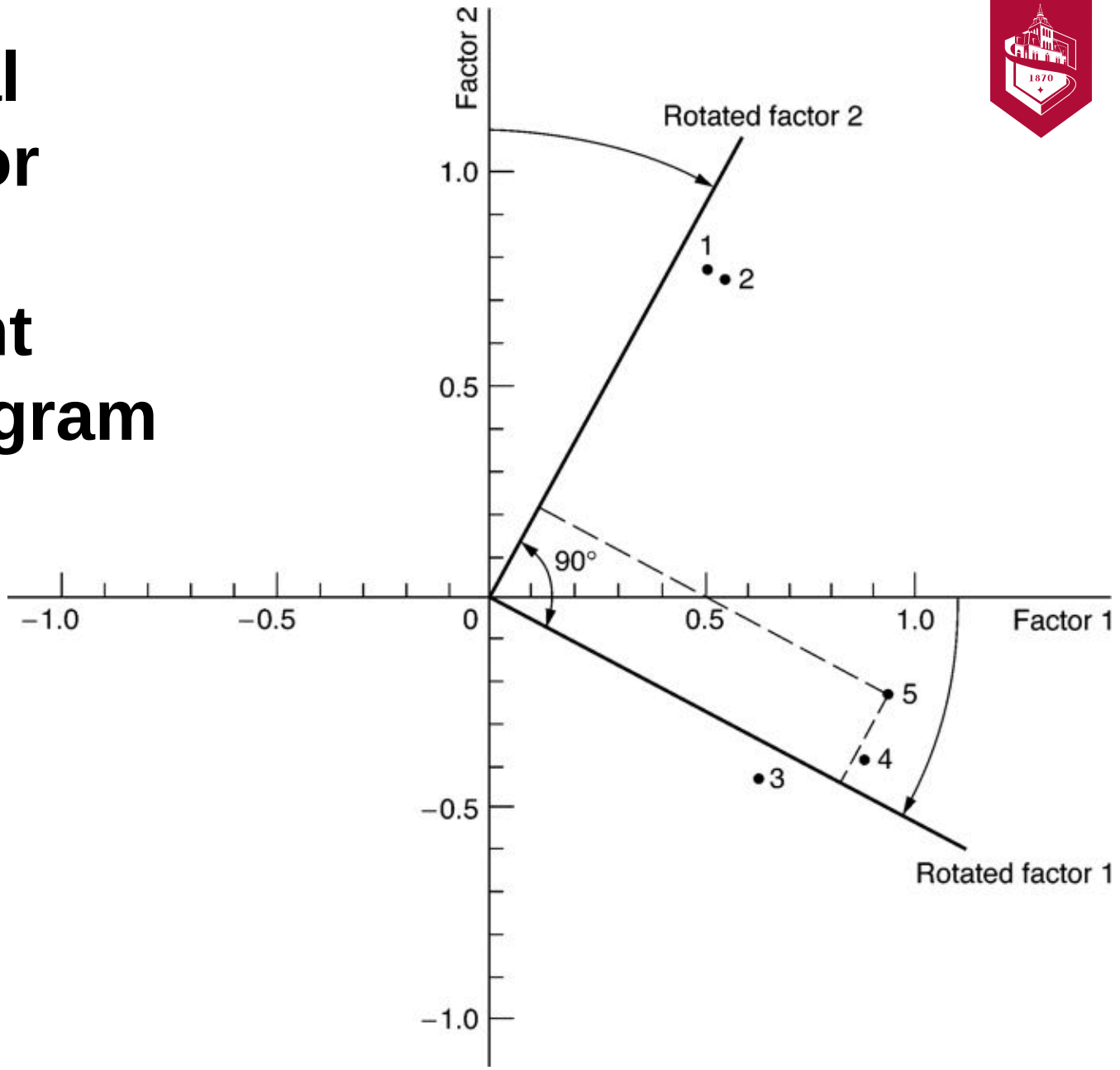
- **Varimax orthogonal rotation:**

  Maximize $\sum$ Var ($l_{ij}^{\,2} \mid F_j$)
  Therefore vary l_{ij} within each factor

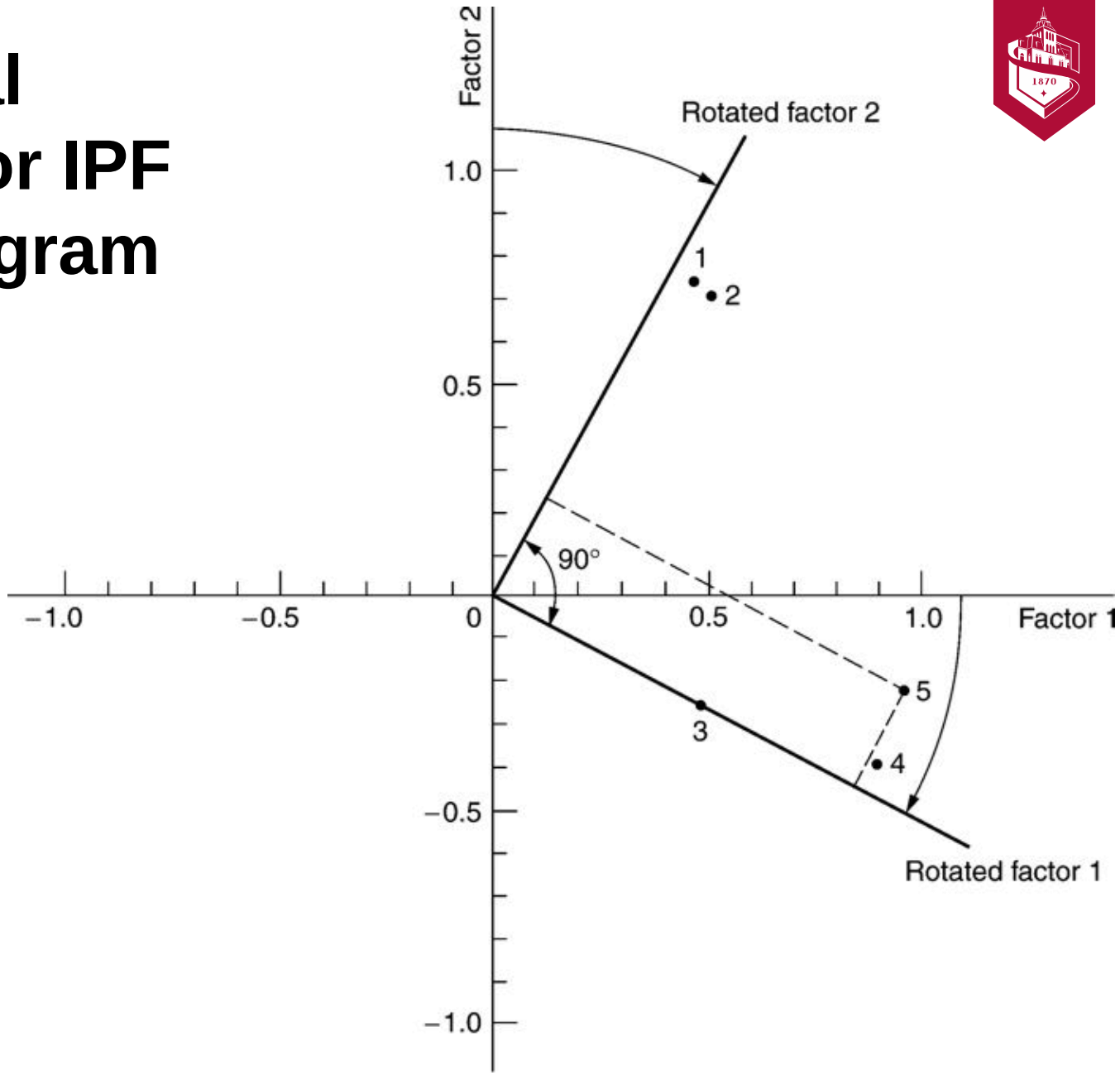
- **Quartimax orthogonal rotation:**

  Maximize Var (all $l_{ij}^{\,2}$)
  Therefore vary all l_{ij}

Orthogonal Rotation for Principal Component Factor Diagram (p 391)



Orthogonal Rotation for IPF Factor Diagram (p 392)



Varimax for PCF (p 393)

Varimax rotated factors: Principal components extraction

Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2
x_1	0.055	**0.933**	0.873
x_2	0.105	**0.929**	0.875
x_3	**0.763**	−0.062	0.586
x_4	**0.943**	0.095	0.898
x_5	**0.918**	0.266	0.913
Variance explained	2.328	1.817	4.145
Percentage	46.6	36.3	82.9

The **communalities are unchanged** after the varimax rotation

Varimax for IPF (p 393)

Varimax rotated factors: Iterated principal factors extraction

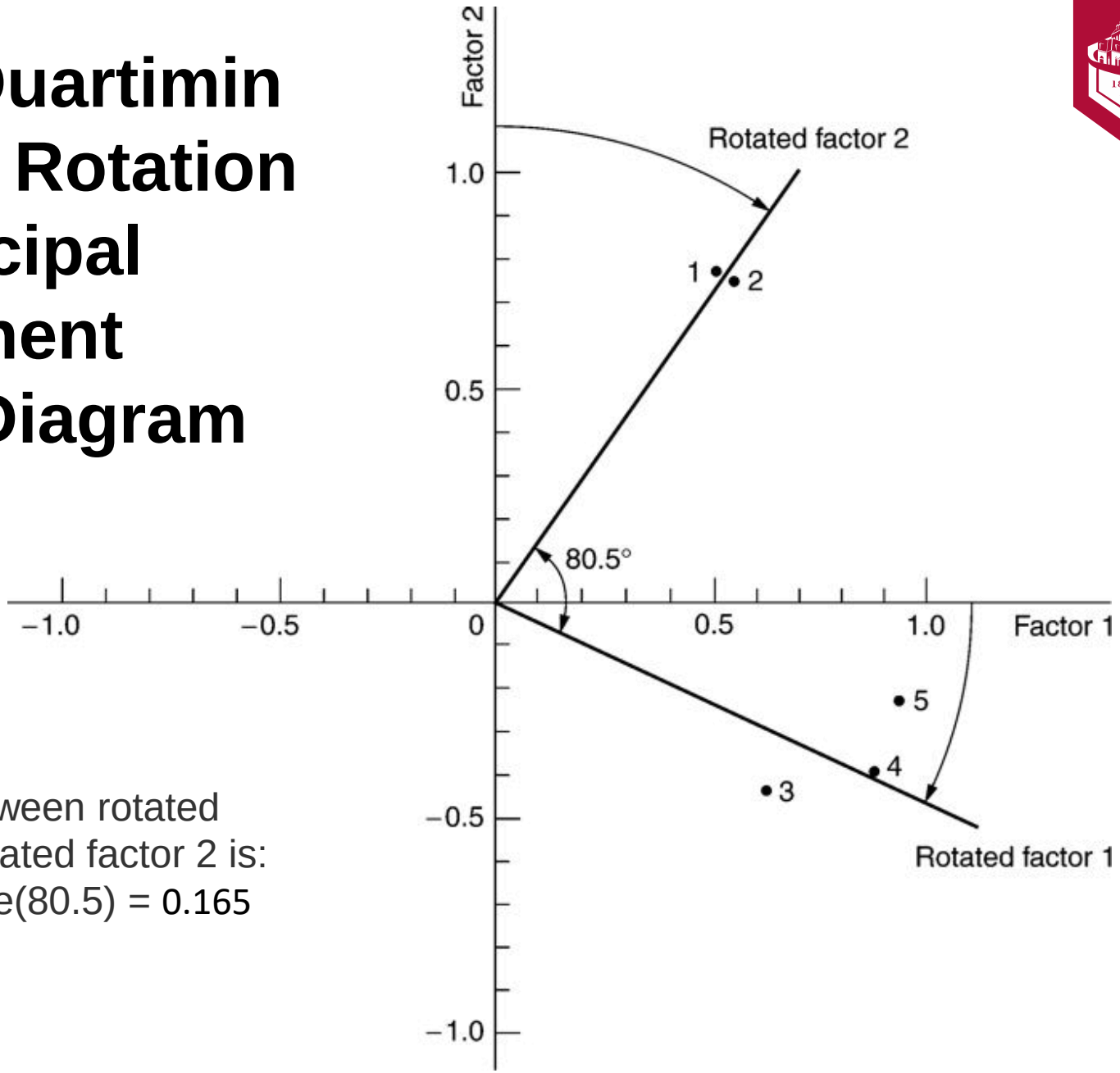
Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2
x_1	0.063	**0.869**	0.759
x_2	0.112	**0.862**	0.756
x_3	**0.546**	0.003	0.298
x_4	**0.972**	0.070	0.949
x_5	**0.951**	0.251	0.968
Variance explained	2.164	1.566	3.730
Percentage	43.3	31.3	74.6

The **communalities are unchanged** after any orthogonal rotation

Oblique Rotations

- Nonorthogonal rotations are called oblique rotations
- Oblique rotated factors are correlated
- The cosine of the angle between the two factor axes is the sample correlation between them
- Permitting a further degree of flexibility.
- The origin of the term oblique lies in geometry
- Most common method: direct quartimin procedure

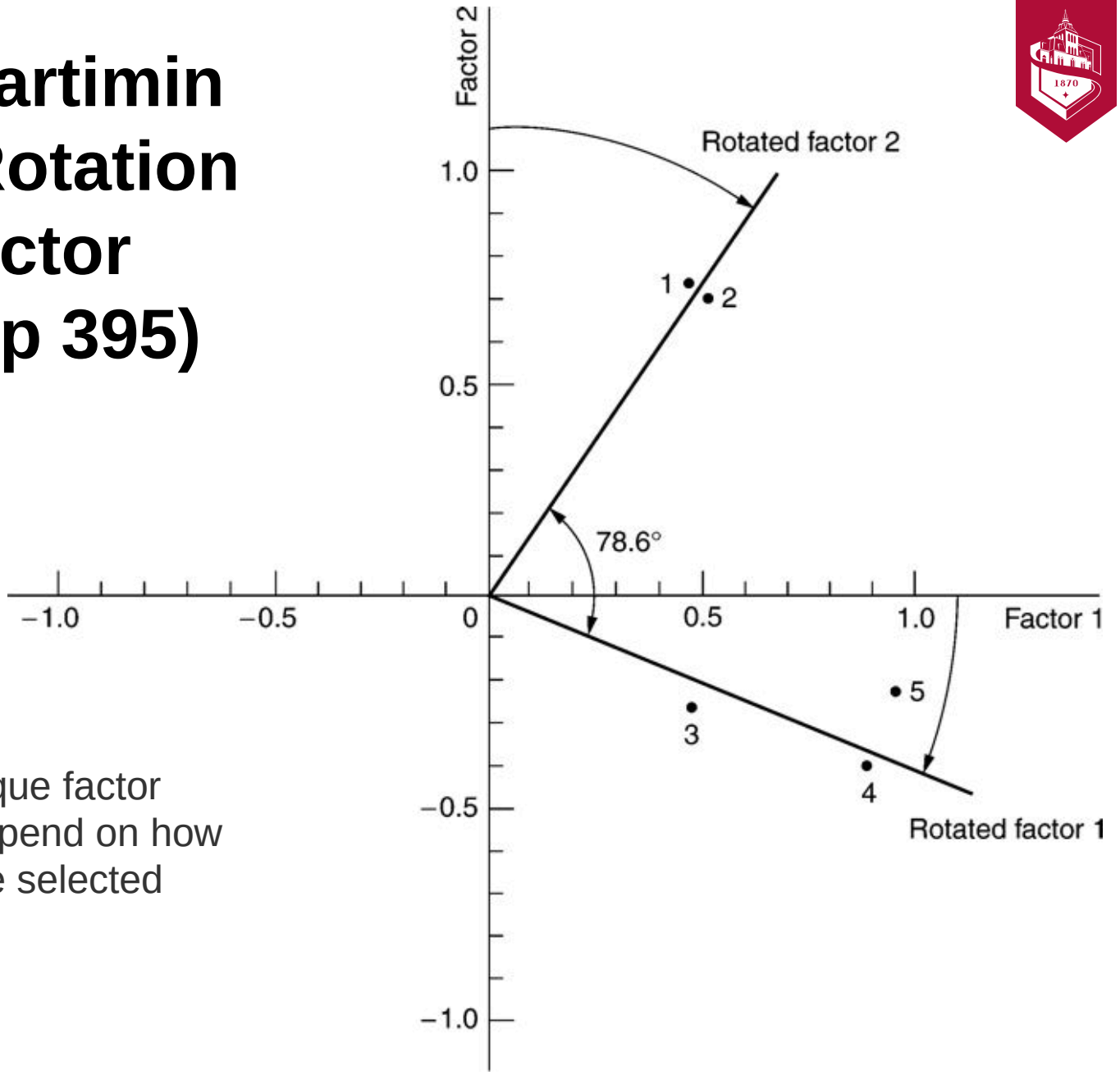
Direct Quartimin Oblique Rotation for Principal Component Factor Diagram (p 394)



Correlation between rotated factor 1 and rotated factor 2 is:

Cosine(80.5) = 0.165

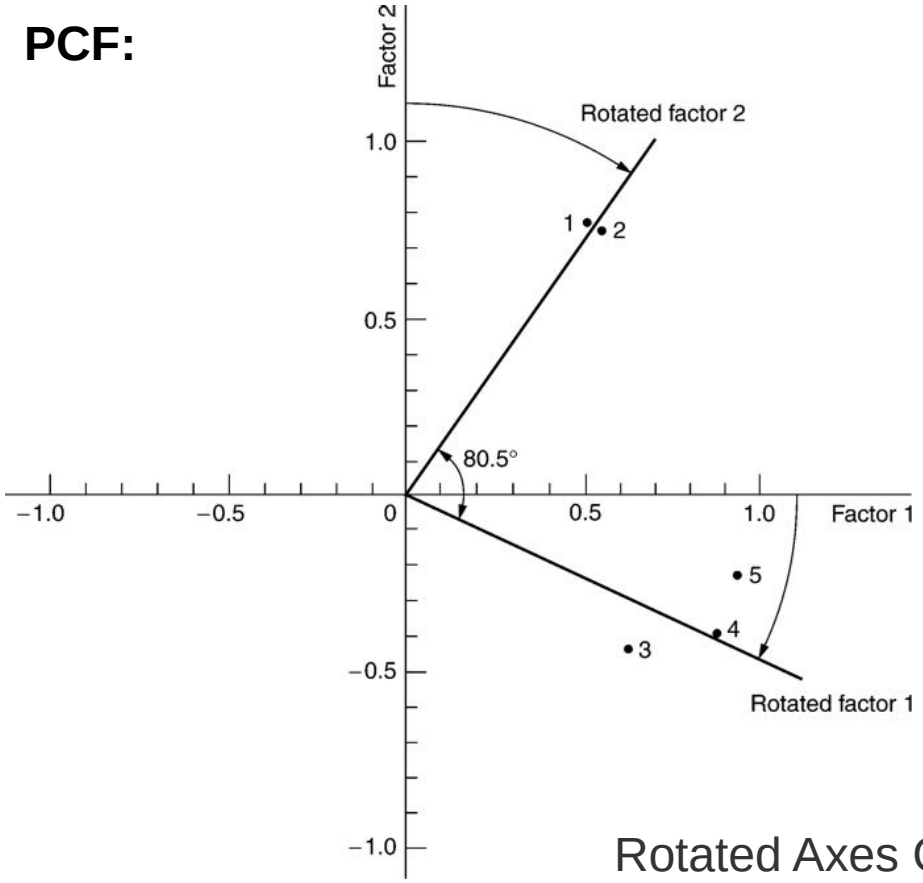
Direct Quartimin Oblique Rotation for IPF Factor Diagram (p 395)



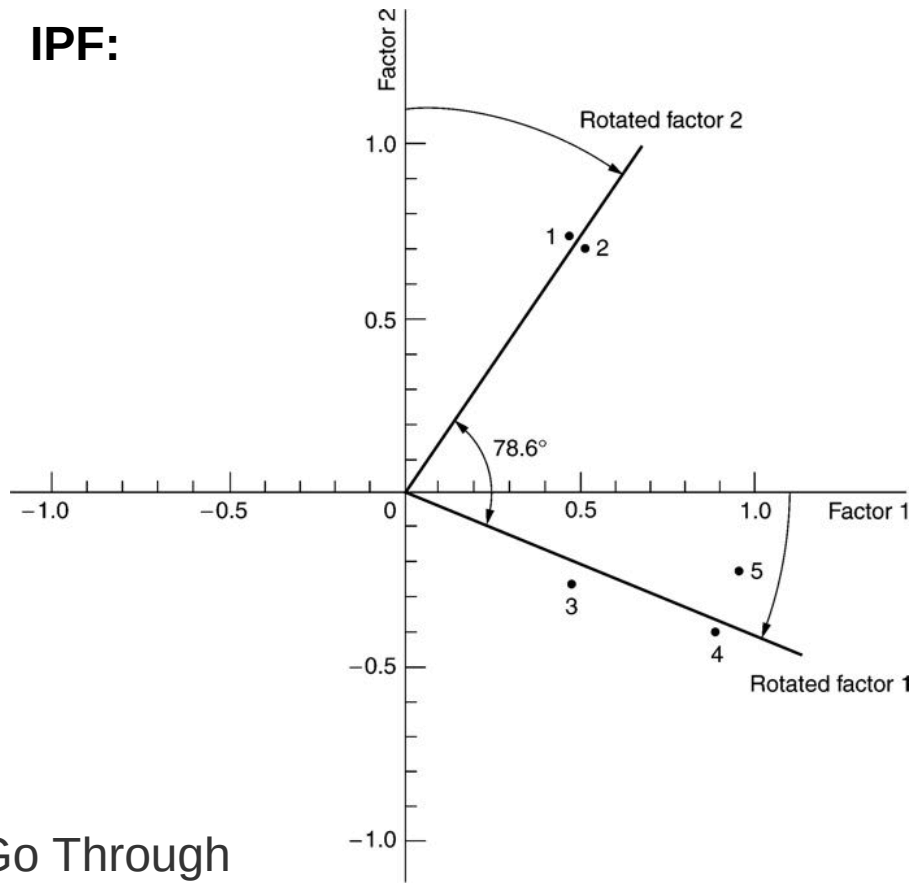
The rotated oblique factor loadings also depend on how many factors are selected

Comparison

PCF:



IPF:



Rotated Axes Go Through Clusters of Variables But Are Not Required to Be Perpendicular to Each Other

Comparison
Orthogonal vs Oblique Rotations

	Advantages	Disadvantages
Orthogonal	• Factors Independent • Communalities Preserved	• Interpretation slightly less clear
Oblique	• Better Interpretation • More flexible	• Factors are Correlated • Communalities Change

Factor Scores

Factor Scores

- FA: each X = function of F's
- Express each F = function of X's.

Computing Factor Scores

- **Recall Multiple Linear Regression:**

$$Y = A + B_1 X_1 + B_2 X_2 + \ldots.$$

- $B = S_{xx}^{-1} S_{yx}$
- **In Factor Analysis, target is:**

$$F = A + B_1 X_1 + B_2 X_2 + \ldots.$$

- $B = S_{xx}^{-1} S_{Fx}$
- **Then S_{xx} and S_{Fx} is the column of Factor Loadings**

Uses of Principal Component and Factor Scores

- 1st Principal Component Score can summarize several variables.
  - Can be used as dependent or independent variable in other analyses
- Factor scores can be used as dependent or independent variables in other analyses

Example: Varimax rotated factors of Principal components extraction

Variable	Factor loadings F_1	Factor loadings F_2	Communality h_i^2
x_1	0.055	**0.933**	0.873
x_2	0.105	**0.929**	0.875
x_3	**0.763**	−0.062	0.586
x_4	**0.943**	0.095	0.898
x_5	**0.918**	0.266	0.913
Variance explained	2.328	1.817	4.145
Percentage	46.6	36.3	82.9

$$\text{factor score } 1 = -0.076x_1 - 0.053x_2 + 0.350x_3 + 0.414x_4 + 0.384x_5$$

large factor score coefficients for x_3, x_4, and x_5 correspond to the large factor loadings

Example
CESD
See Page 398

Varimax rotation, PC factors for standardized CESD (p. 397)

Item		F_1	F_2	F_3	F_4	h_i^2
	Negative affect					
1.	I felt that I could not shake off the blues even with the help of my family or friends.	0.638	0.146	0.268	0.280	0.5784
2.	I felt depressed.	0.773	0.296	0.272	−0.003	0.7598
3.	I felt lonely.	0.726	0.054	0.275	0.052	0.6082
4.	I had crying spells.	0.630	−0.061	0.168	0.430	0.6141
5.	I felt sad.	0.797	0.172	0.160	0.016	0.6907
6.	I felt fearful.	0.624	0.234	−0.018	0.031	0.4448
7.	I thought my life had been a failure.	0.592	0.157	0.359	0.337	0.6173
	Positive affect					
8.	I felt that I was as good as other people.	0.093	−0.051	0.109	0.737	0.5655
9.	I felt hopeful about the future.	0.238	0.033	0.621	0.105	0.4540
10.	I was happy.	0.557	0.253	0.378	0.184	0.5516
11.	I enjoyed life.	0.498	0.147	.407	0.146	0.4569
	Somatic and retarded activity					
12.	I was bothered by things that usually don't bother me.	0.449	0.389	−0.049	−0.065	0.3600
13.	I did not feel like eating; my appetite was poor.	0.070	0.504	−0.173	0.535	0.5760
14.	I felt that everything was an effort.	0.117	0.695	0.180	0.127	0.5459
15.	My sleep was restless.	0.491	0.419	−0.123	−0.089	0.4396
16.	I could not "get going."	0.196	0.672	0.263	−0.070	0.5646
17.	I had trouble keeping my mind on what I was doing.	0.270	0.664	0.192	0.000	0.5508
18.	I talked less than usual.	0.409	0.212	−0.026	0.223	0.2628
	Interpersonal					
19.	People were unfriendly.	−0.015	0.237	0.746	−0.088	0.6202
20.	I felt that people disliked me.	0.358	0.091	0.506	0.429	0.5770
	Variance explained	4.795	2.381	2.111	1.551	10.838
	Percentage	24.0	11.9	10.6	7.8	54.2

Interpretation

Principal Component Analysis identified 5 PCs, but previous literature used 4 factors, not 5:

- **Factor 1: loads heavily on items 1 – 7**
- **Factor 2: items 12 – 18 (somatic and retarded activity)**
- **Factor 3: items 19 – 20 (interpersonal), plus items 9 and 11 (positive-affect)**
- **Factor 4: no clear pattern: item 8 (positive affect) has highest loading**

Caveats

- Number of Factors should be chosen with care – check default options
- There should be at least two variables with non-zero weights per factor
- If Factors are correlated, try Oblique Factor Analysis
- Results usually evaluated by “reasonableness to investigator” as opposed to formal tests
- Motivate theory, not replace it.

A Case study for Multivariate Data Analysis

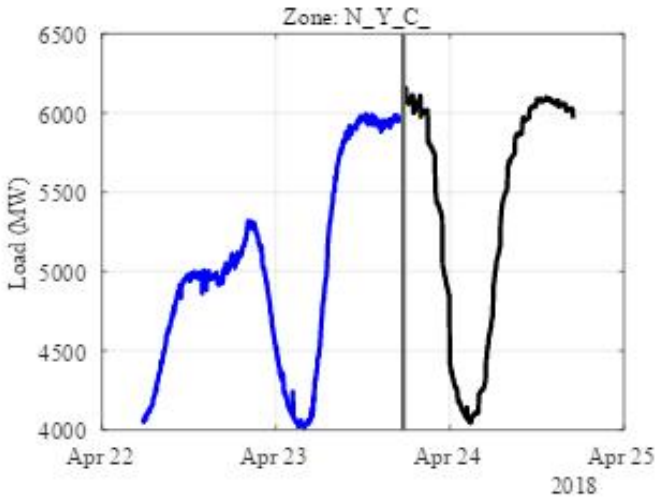
Load Forecasting Case Study

Ref.: https://www.mathworks.com/

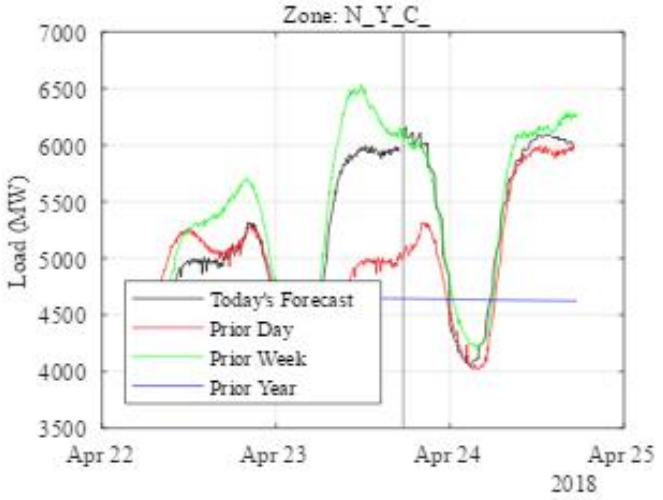
Zone NYISO|J-N.Y.C

Generate Forecast

Model Diagnostics Report



Comparison



Load Data Preparation - 1

- Download historical load data (New York ISO) in a range that we would like data for

```
Downloading 8 of 12. 20050901pal_csv.zip
Downloading 7 of 12. 20050801pal_csv.zip
Downloading 6 of 12. 20050701pal_csv.zip
Downloading 5 of 12. 20050601pal_csv.zip
Downloading 2 of 12. 20050301pal_csv.zip
Downloading 1 of 12. 20050201pal_csv.zip
Downloading 4 of 12. 20050501pal_csv.zip
Downloading 3 of 12. 20050401pal_csv.zip
Downloading 10 of 12. 20051101pal_csv.zip
Downloading 9 of 12. 20051001pal_csv.zip
Downloading 11 of 12. 20051201pal_csv.zip
Downloading 12 of 12. 20060101pal_csv.zip
```

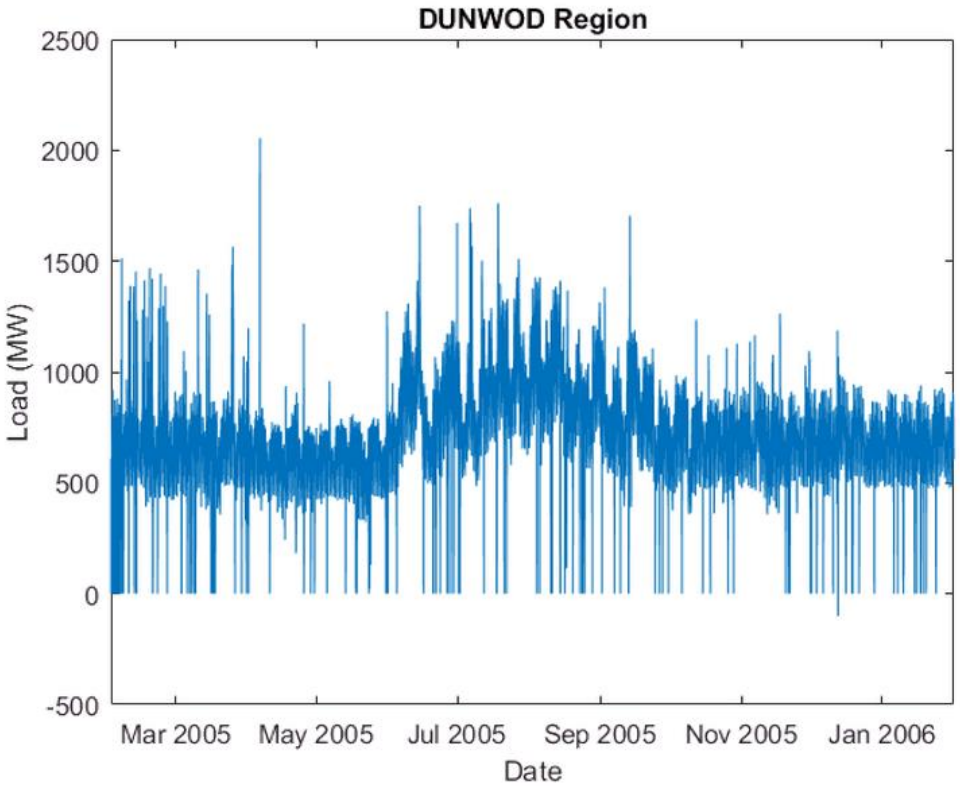
- Aggregate files from different stations!

TimeStamp	Name	Load
02/01/2005 00:00:00	CAPITL	1250.4
02/01/2005 00:00:00	CENTRL	1942.1
02/01/2005 00:00:00	DUNWOD	612.2
02/01/2005 00:00:00	GENESE	1131.6
02/01/2005 00:00:00	HUD VL	1154.9
02/01/2005 00:00:00	LONGIL	2270.5
02/01/2005 00:00:00	MHK VL	901.3
02/01/2005 00:00:00	MILLWD	272.3

Load Data Preparation - 2

- Unstack the data so that each location is a column in the table.

Date	CAPITL	CENTRL	DUNWOD
02/01/2005 00:00:00	1250.4	1942.1	612.2
02/01/2005 00:05:00	0	0	0
02/01/2005 00:10:00	1232.5	1978.1	581.1
02/01/2005 00:15:00	1236.4	1971.6	569.5
02/01/2005 00:20:00	1213.5	1955.6	599.3
02/01/2005 00:25:00	1202.1	1972.8	581.1
02/01/2005 00:30:00	1219.4	1976.9	586
02/01/2005 00:35:00	1227.8	1964.4	584.8

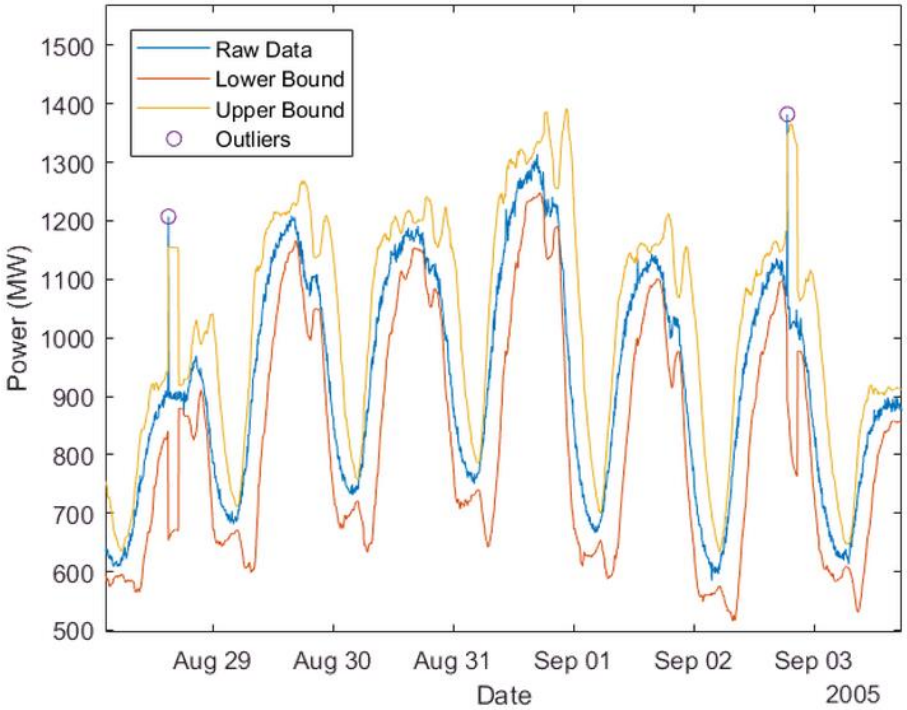


- Problems:
  - Zero values
  - Large spikes

Load Data Preparation - 3

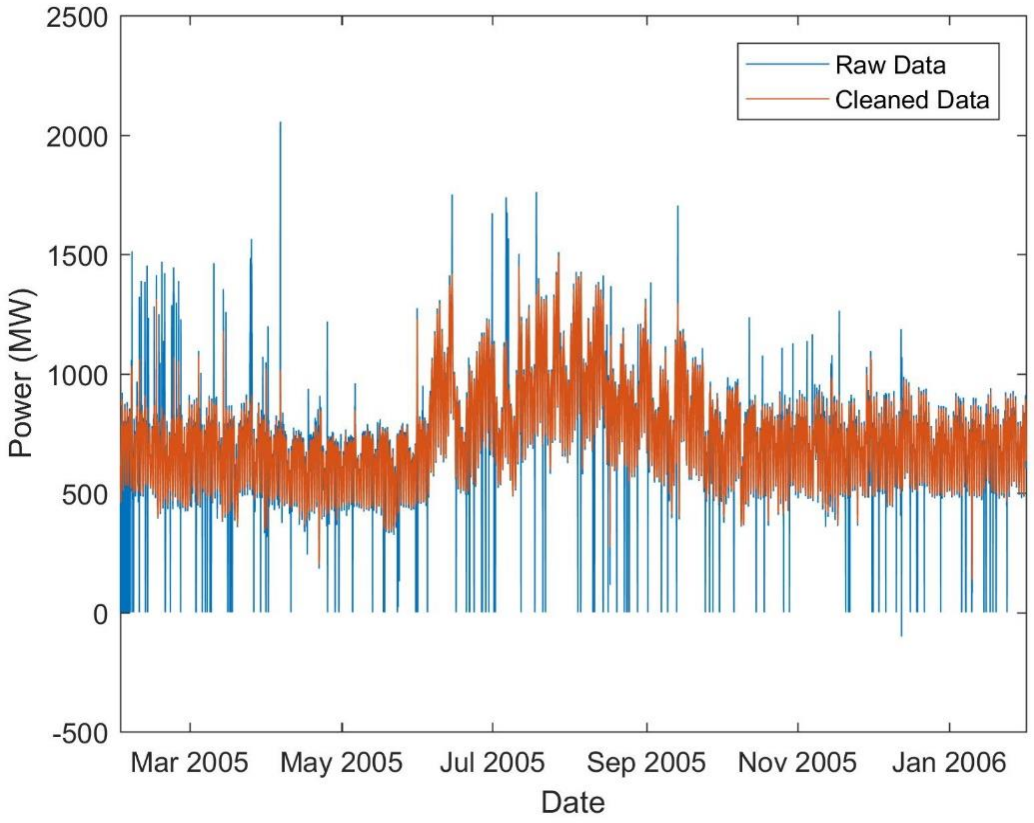
- Imputing the zero values (missing values): Here linear interpolation of neighboring is used

- Identify outliers



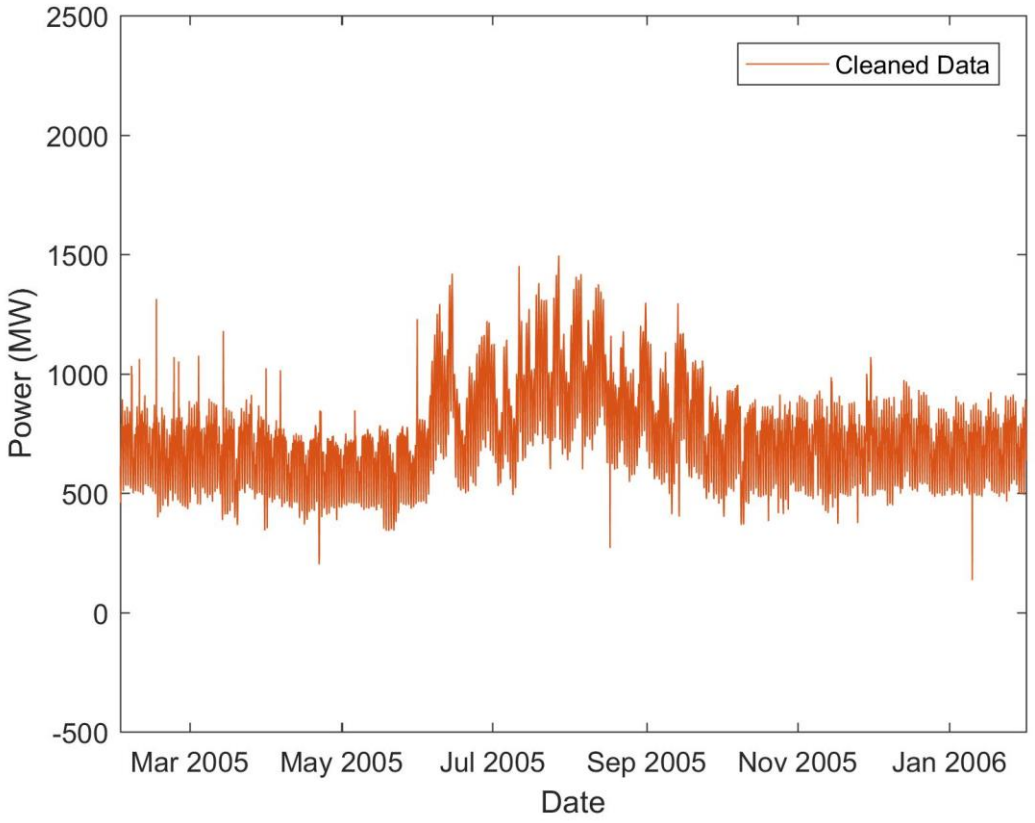
Load Data Preparation - 4

- Replace outliers and smooth noise



Load Data Preparation - 4

- Replace outliers and smooth noise



Weather Data Preparation - 1

- Weather data is publicly available at: National Climatic Data Center

WBAN	Date	Time	DryBulbFarenheit	DewPointFarenhei
3011	"20070501"	"0050"	45	30
3011	"20070501"	"0150"	45	30
3011	"20070501"	"0250"	45	30
3011	"20070501"	"0350"	43	30
3011	"20070501"	"0450"	43	30
3011	"20070501"	"0550"	43	30
3011	"20070501"	"0650"	46	30
3011	"20070501"	"0750"	54	28

- Big Data!
- Extract the WBAN stations in the area.

Weather Data Preparation - 2

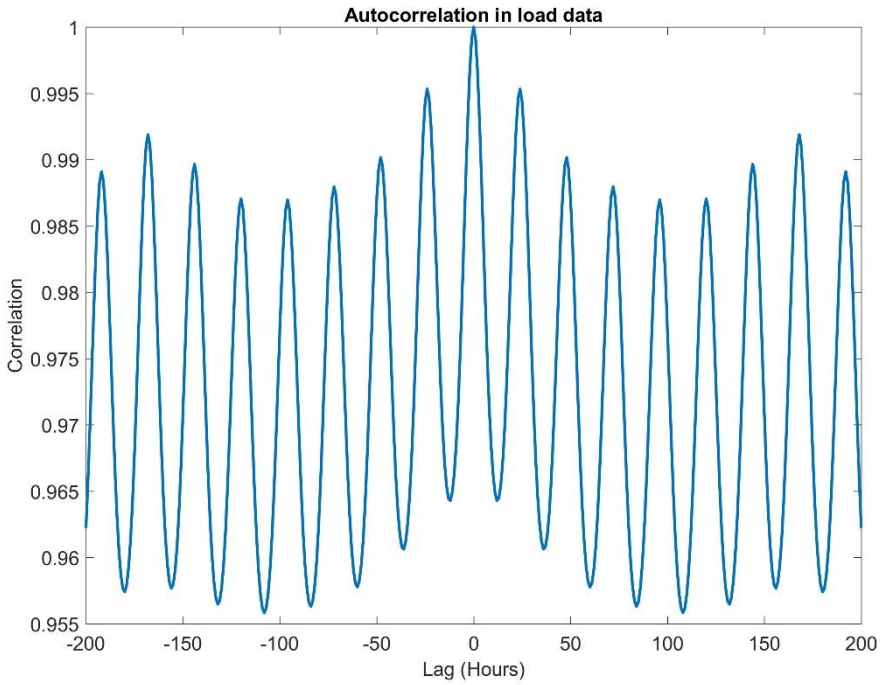
- Different weather stations report at different times:

Date	x4725		x4781		x14714		x14732	
01-May-2007 00:45:00	NaN	NaN	NaN	NaN	50	34	NaN	NaN
01-May-2007 00:51:00	NaN	NaN	NaN	NaN	NaN	NaN	58	30
01-May-2007 00:53:00	46	26	NaN	NaN	NaN	NaN	NaN	NaN
01-May-2007 00:54:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
01-May-2007 00:55:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
01-May-2007 00:56:00	NaN	NaN	51	32	NaN	NaN	NaN	NaN
01-May-2007 01:45:00	NaN	NaN	NaN	NaN	46	34	NaN	NaN
01-May-2007 01:51:00	NaN	NaN	NaN	NaN	NaN	NaN	59	28
01-May-2007 01:53:00	44	26	NaN	NaN	NaN	NaN	NaN	NaN
01-May-2007 01:54:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
01-May-2007 01:55:00	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
01-May-2007 01:56:00	NaN	NaN	50	33	NaN	NaN	NaN	NaN

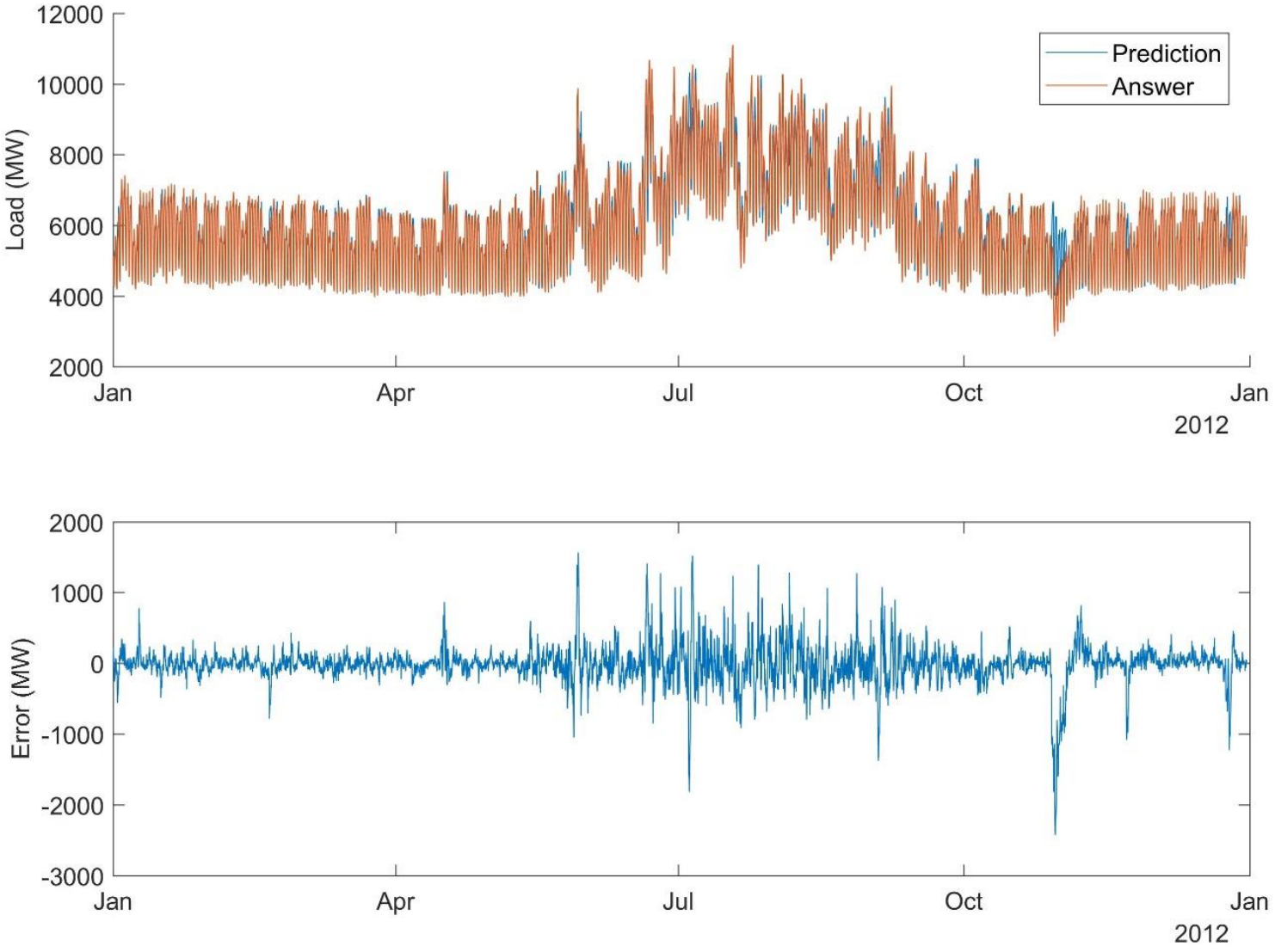
- Retime and clean up this dataset!

Data set for analysis

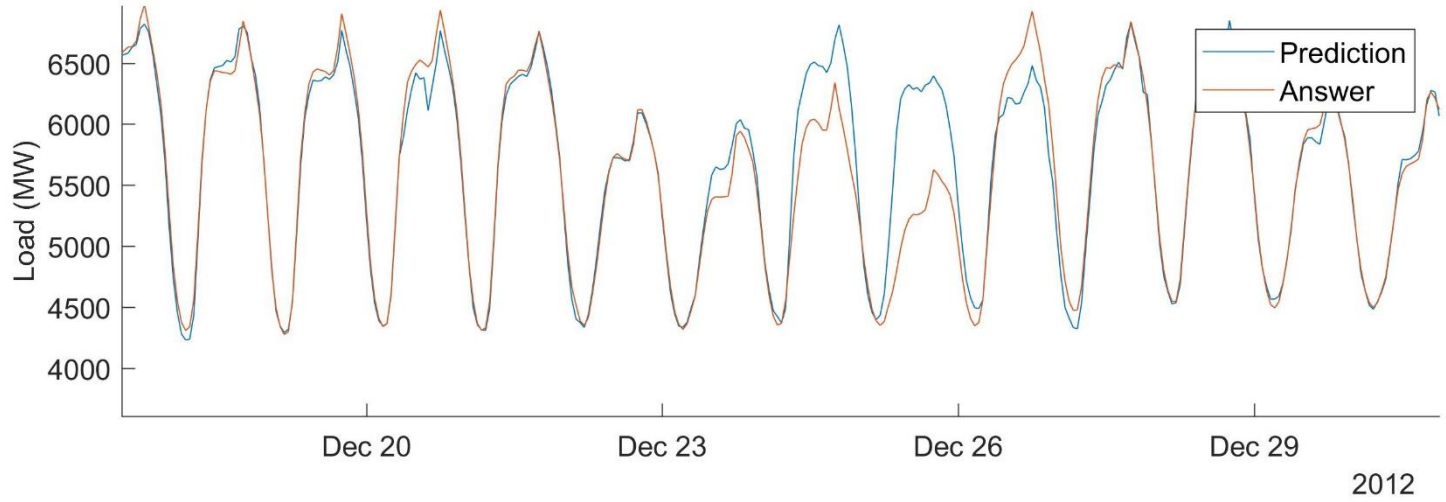
- Merge load and weather data sets
- Create predictors
  - Weather variables (Temperature, etc.)
  - Load History

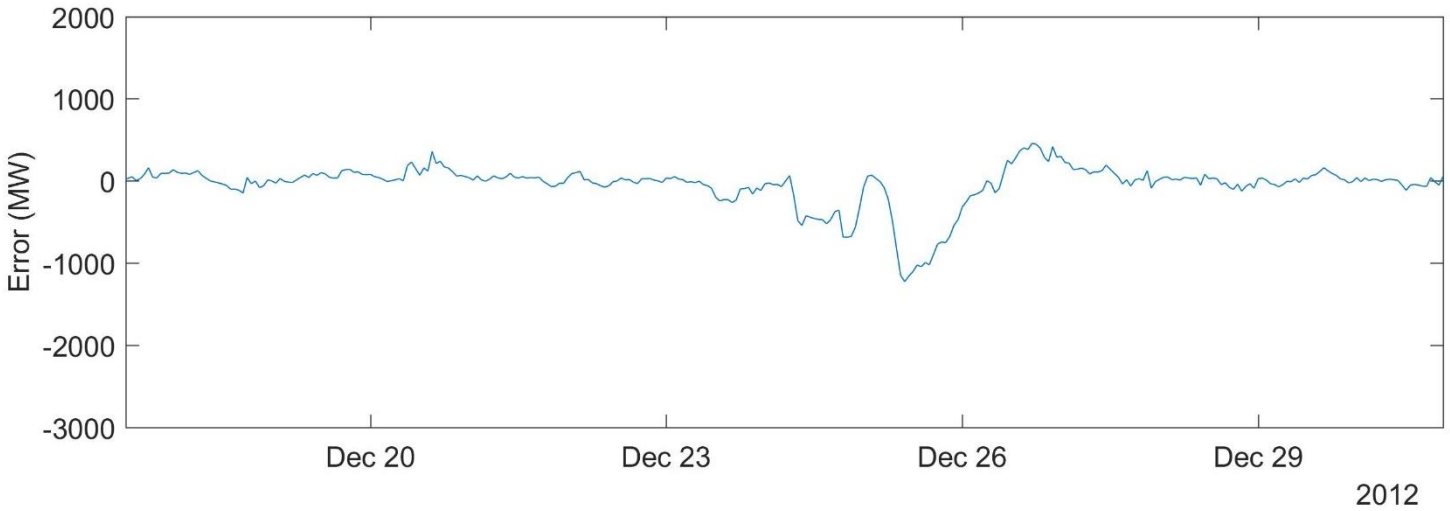


Load forecast on a zone

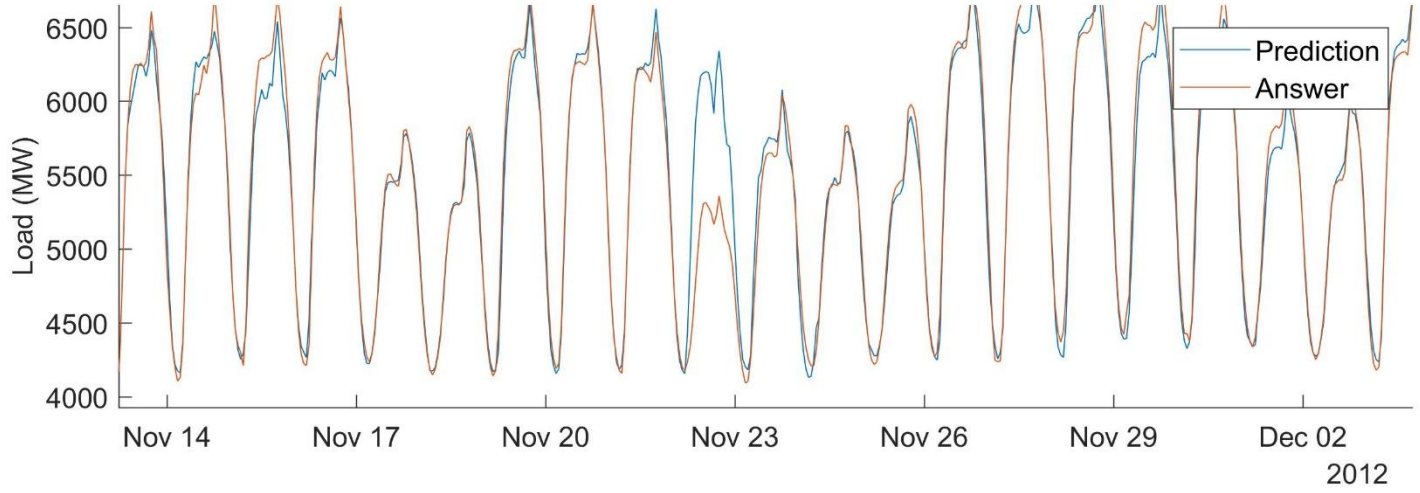


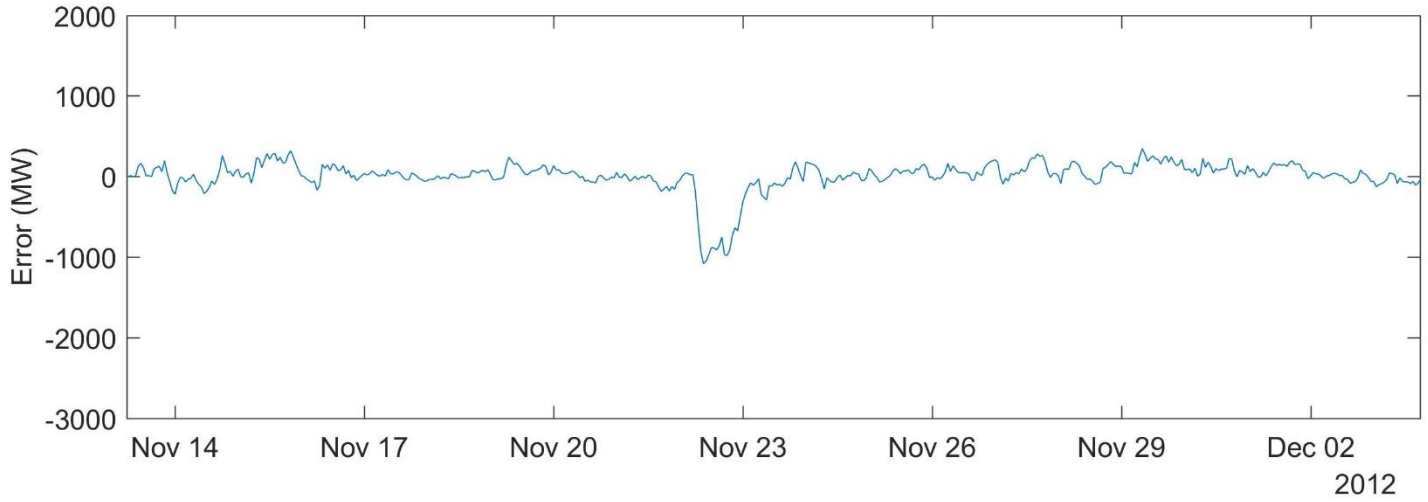
Interpretations & Discussions - 1



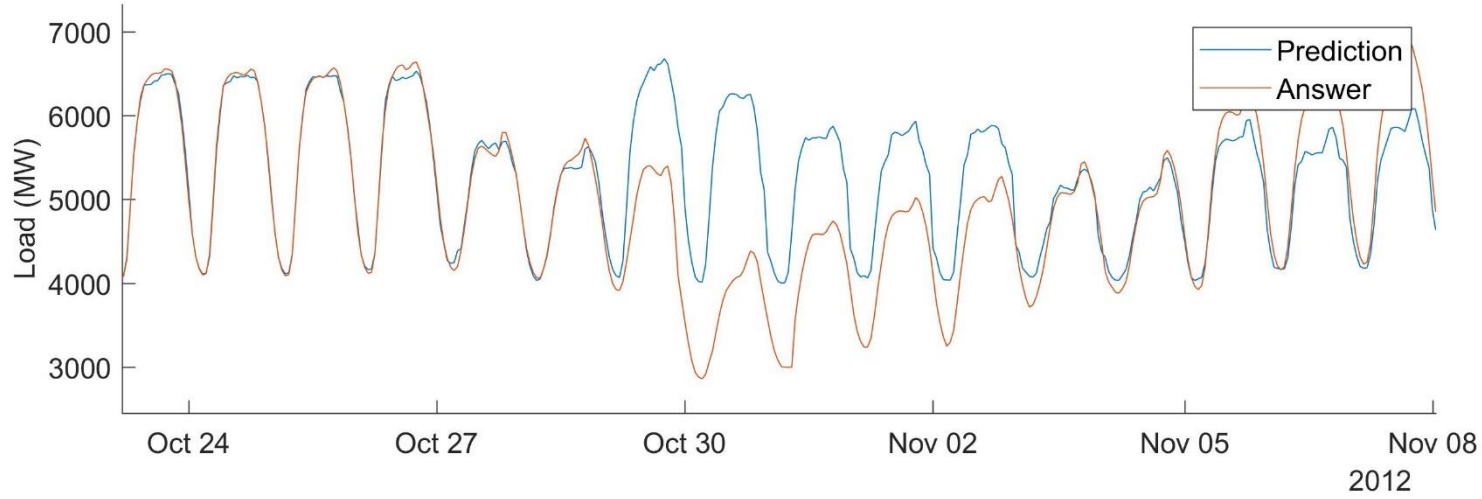


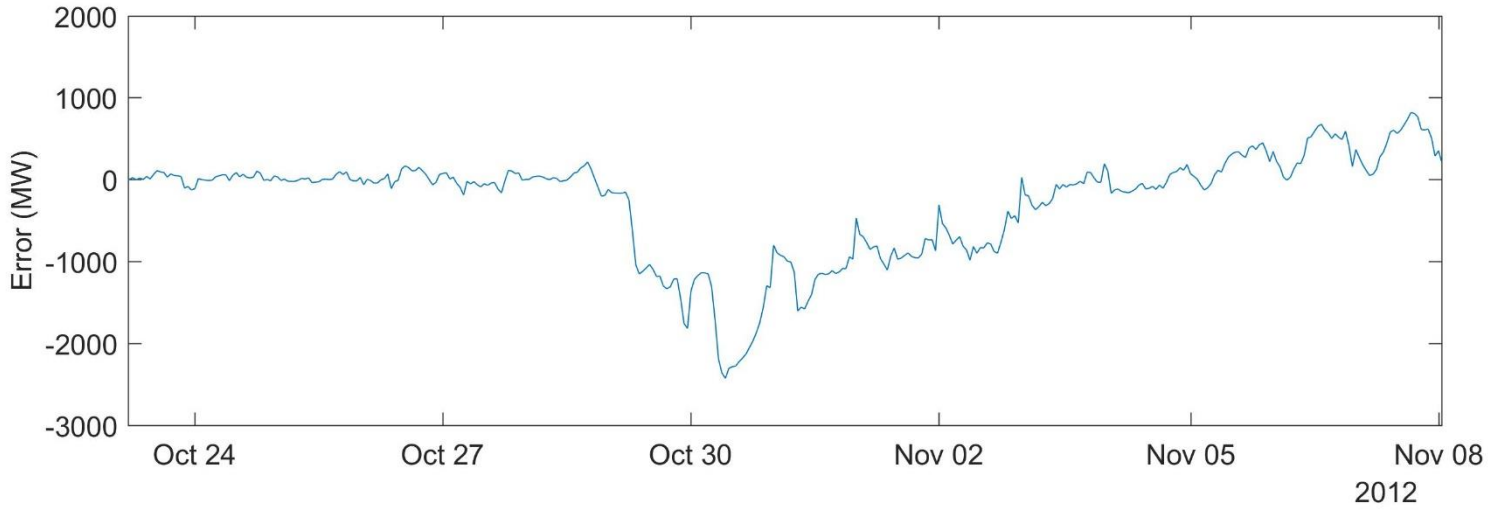
Interpretations & Discussions - 2





Interpretations & Discussions - 3





Late term exam (April 15, 2019)

- In your class time (6:15 – 8:45 pm)
- It will be in class for BIA-652-A
- Do not forget to have your laptop ready
- You cannot use internet during the exam
- No cell-phone!
- Open book (it is not a formula memorization activity)
- Also the required data will be provided as appendix
- Coding won't be required

A Bit of Review of the exam topics!

- Preparation of Data: Outliers, Missing (Null) Values, etc.
- Regression – Simple and Multiple
- Classification
  - Discriminant Analysis
  - Logistic Regression
  - Naïve Bayes
  - KNN
  - Ensembles
- Cluster Analysis – Hierarchical, K-Means, Density
- Dimension Reduction
  - Principal Component Analysis
  - Singular Value Decomposition
  - Factor Analysis

* Reds won't be part of the exam

Sample Problems for the late term exam: Data Preparation

- Sample:
- Complete an ANOVA TABLE
- Given MANOVA and ask for the F statistic

Source of variation	Sums of squares	df	Mean square	*F*
Regression	21.0570	2	10.5285	36.81
Residual	42.0413	147	0.2860	
Total	63.0983			

- Explaining the significance of the ANOVA, R^2, and adj R^2
- Finding the candidate outliers and explain why

Sample Problems for the late term exam: Regression and Classification

- Sample:
- Predict the output of a multi-variate regression model for a new observation
- LDA/LR model is given and then it will ask you to classify a new observation along with explanation
- Discussing about the Confusion Matrix (FN,FP,TN,TP)
- Discussing about the ROC
- Find a Cut probability for having high TP and low FN
- Compare two algorithms results

Sample Problems for the late term exam: Clustering

- Sample:
- Give a distance for different samples and ask to find each observation cluster
- For K-mean, k will be given
- How you can interpret the results
- Drawing the dendrogram (tree graph) for give data
- Interpret a Profile Plot

Sample Problems for the late term exam: Dimension Reduction

- Sample:
- PCA results will be given
- Ask you to choose number of PC and also PCs those retain at least % of the variance
- Calculate a PC value for a give observation

- Factor Analysis results will be given
- Calculate loading using Eigen values and vectors
- Describe a factor in terms of loading
- Find highly loaded variable

Project Presentation

- The project will be presented in the order scheduled.
- Please find the template for your presentation here
- You may use your laptop to make the project presentation.
- The project presentations will last 15 minutes and then there will be 5 minutes for questions.
- Presentations will be strictly timed, just as if being done at a professional conference.
- You are advised to practice your presentation before giving it, with the overhead materials you will actually use, to make sure it does not take too long. Don't try to memorize the presentation -- use the overheads as clues about what you should be talking about, but also (very important) don't just READ the overheads.
- All group members should participate in the presentation. For example, if you are in a group of three, each of you should present about five minutes.

Term Project

- Report (both PDF and Word files)
- Presentation (PPT file)
- Code

You need to submit them up to May 6, 2019

The presentation and code only need to be submitted! and your project credit comes from how you present the project and how is the final written report. Report needs to be complete and you should not mention “it is mentioned/presented in the code or presentation”. All points should be explained and discussed in the report and do not leave any figure or table without explanations or discussion. Also, you can expand/modify the report and address the points that you missed in your presentation.

Term Paper

The report needs detail the following:

- Title
- Names
- Abstract
- Introduction
- Problem description
- Evaluation of database (no. variables, instances, etc.)
- Data processing and preparation
- Methods used (regression, classification, and clustering) and why, perhaps you tried many which should also be reported
- Dimension reduction method(s)
- Results (plots and tables), Comparisons, and Discussions
- Conclusion
- Future research
- References

Write this as if you were trying to publish results in a professional conference/journal. Your evaluation will be based on this viewpoint!

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

stevens.edu

Amir H Gandomi; PhD
Assistant Professor of Analytics & Information Systems
a.h.gandomi@stevens.edu