Introduction to Multivariate Analysis

Lecture 1

August 24, 2005

Multivariate Analysis

Today's Lecture

- Introductions.
- Syllabus and course overview.
- Chapter 1 (a brief review, really):
  - Data organization/notation.
  - Graphical techniques.
  - Distance measures.
- Introduction to SAS.

Who are you?

To help all of use get to know each other better, please tell us your:

- Name.
- Department and specialty.
- Where you are from originally.
- Where you did you undergraduate work.

Syllabus

Syllabus discussion...

Multivariate Statistics

A taxonomy of multivariate statistical analyses shows that most techniques fall into one of the following categories:

1. Data reduction or structural simplification.
2. Sorting and grouping.
3. Investigation of the dependence among variables.
4. Prediction.
5. Hypothesis construction and testing.

Data Organization

- As a precursor of things to come, here is a preview of the ways data are organized in this book/course.
- Multivariate data are a collection of observations (or measurements) of:
    - p variables $(k = 1, \ldots, p)$.
    - n "items" $(j = 1, \ldots, n)$.
        - "items" can also be though of as subjects/examinees/individuals or entities (when people are not under study) .
        - In some disciplines (such as educational measurement), "items" are considered the variables collected per individual.

Data Organization

- x_{jk} = measurement of the k^{th} variable on the j^{th} entity.

	Variable 1	Variable 2	$\dots$	Variable k	$\dots$	Variable p
Item 1:	x_{11}	x_{12}	$\dots$	x_{1k}	$\dots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\dots$	x_{2k}	$\dots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j:	x_{j1}	x_{j2}	$\dots$	x_{jk}	$\dots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n:	x_{n1}	x_{n2}	$\dots$	x_{nk}	$\dots$	x_{np}

Arrays

- To represent the entire collection of items and entities, a rectangular array can be constructed:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2k} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \dots & x_{jk} & \dots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} & \dots & x_{np} \end{bmatrix}$$

- In the next class, we will learn about how arrays like this have an algebra that makes life somewhat easier.

- All arrays will be symbolized by boldfaced font.

Array Example

- So, putting things all together, envision standing outside of the Kansas Union Bookstore, asking people for receipts.
- You are interested in looking at two variables:
  - Variable 1: the total amount of the purchase.
  - Variable 2: the number of books purchased.
- You find four people, and here is what you see observe (with notation:

$$x_{11} = 42 \quad x_{21} = 52 \quad x_{31} = 48 \quad x_{41} = 58$$

$$x_{12} = 4 \quad x_{22} = 5 \quad x_{32} = 4 \quad x_{42} = 3$$

Array Example (Continued)

- The data array would the look like:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

- Notice for any variable, x_{jk}:
  - The first subscript (j) represents the ROW location in the data array.
  - The second subscript (k) represents the COLUMN location in the data array.

Descriptive Statistics Review

- When we have a large amount of data, it is often hard to get a manageable description of the nature of the variables under study.
- For this reason (and as a way of introducing a review topics from previous courses), descriptive statistics are used.
- Such descriptive statistics include:
  - Means.
  - Variances.
  - Covariances.
  - Correlations.

Sample Mean

- For the k^{th} variable, the sample mean is:

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk}$$

- An array of the means for all p variables then looks like this (which we will come to know as the mean vector):

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \bar{x}_4 \end{bmatrix}$$

Sample Variance

- For the k^{th} variable, the sample variance is:

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2$$

- Note the “kk” subscript, this will be important because the equation that produces the variance for a single variable is a derivation of the equation of the covariance for a pair of variables.

- Also note the division by n. Reasons for this will become apparent in the near future.

- For a pair of variables, i and k, the sample covariance is:

$$s_{ik} = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

Sample Covariance Matrix

- Making an array of all sample covariances give us:

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Sample Correlation

- Sample covariances are dependent upon the scale of the variables under study.
- For this reason, the correlation is often used to describe the association between two variables.
- For a pair of variables, i and k, the sample correlation is found by dividing the sample covariance by the product of the standard deviation of the variables:

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}}$$

- The sample correlation:
  - Ranges from -1 to 1.
  - Measures linear association.
  - Is invariant under linear transformations of i and k.
  - Is a biased estimator.

Sample Correlation Matrix

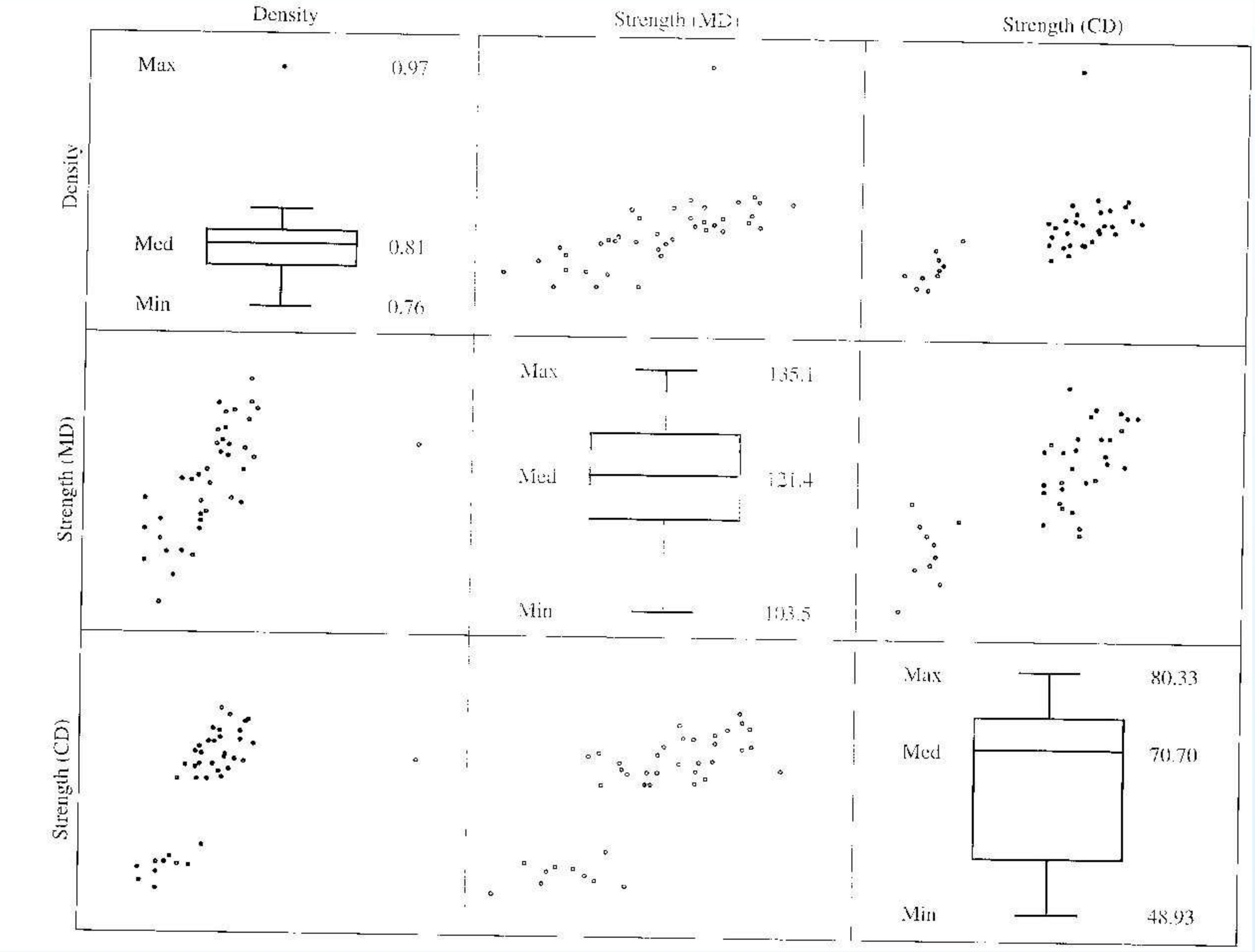
- Making an array of all sample covariances give us:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix}$$

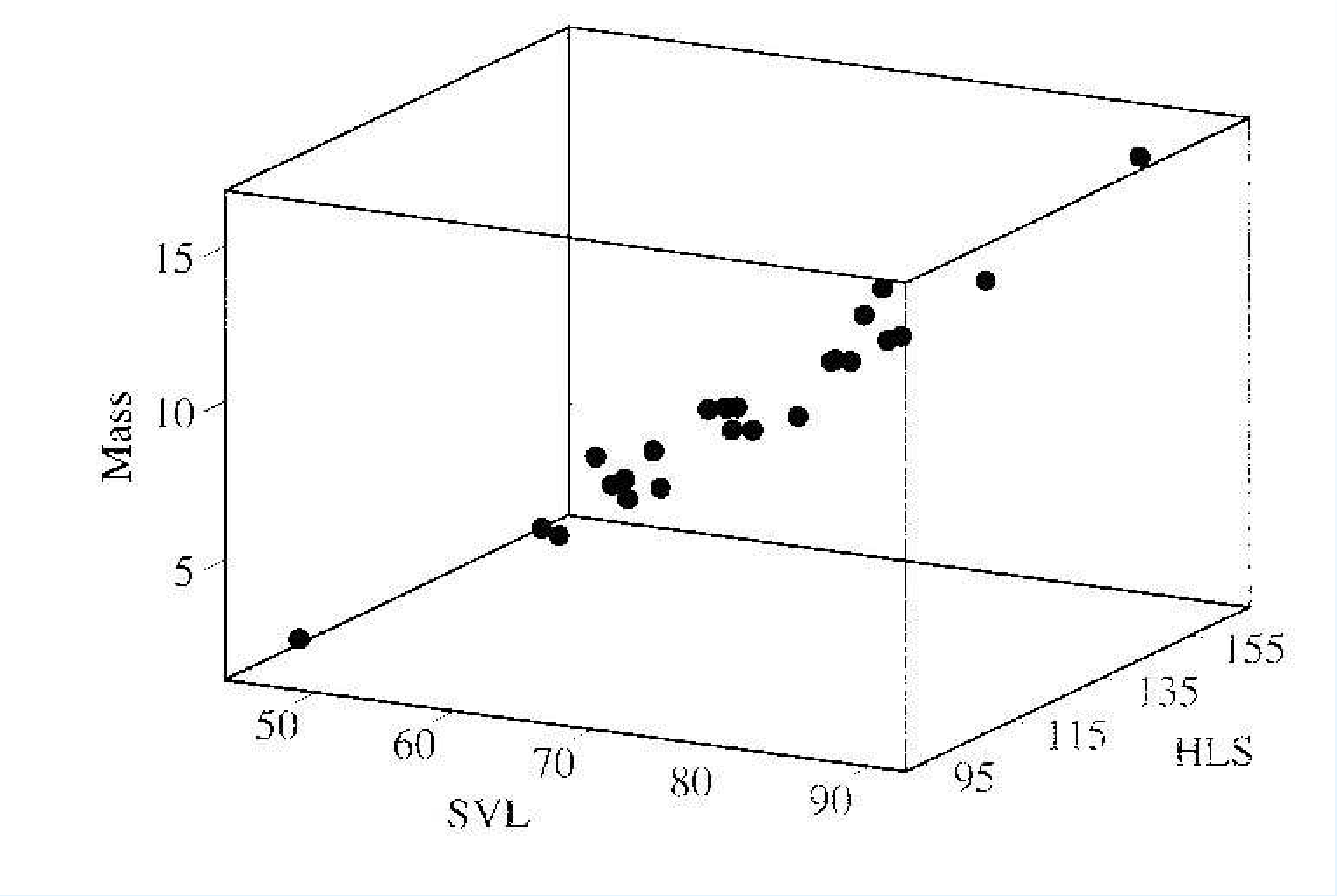
Graphical Techniques

- Displaying multivariate data can be difficult due to our natural limitations of 3-dimensions.
- Several simple ways of displaying data include:
  - Bivariate scatterplots.
  - Three-dimensional scatterplots.
- But you already know those, some plots that can be achieved by multivariate methods include:
  - "Stars."
  - Chernoff faces.

Bivariate Scatterplots



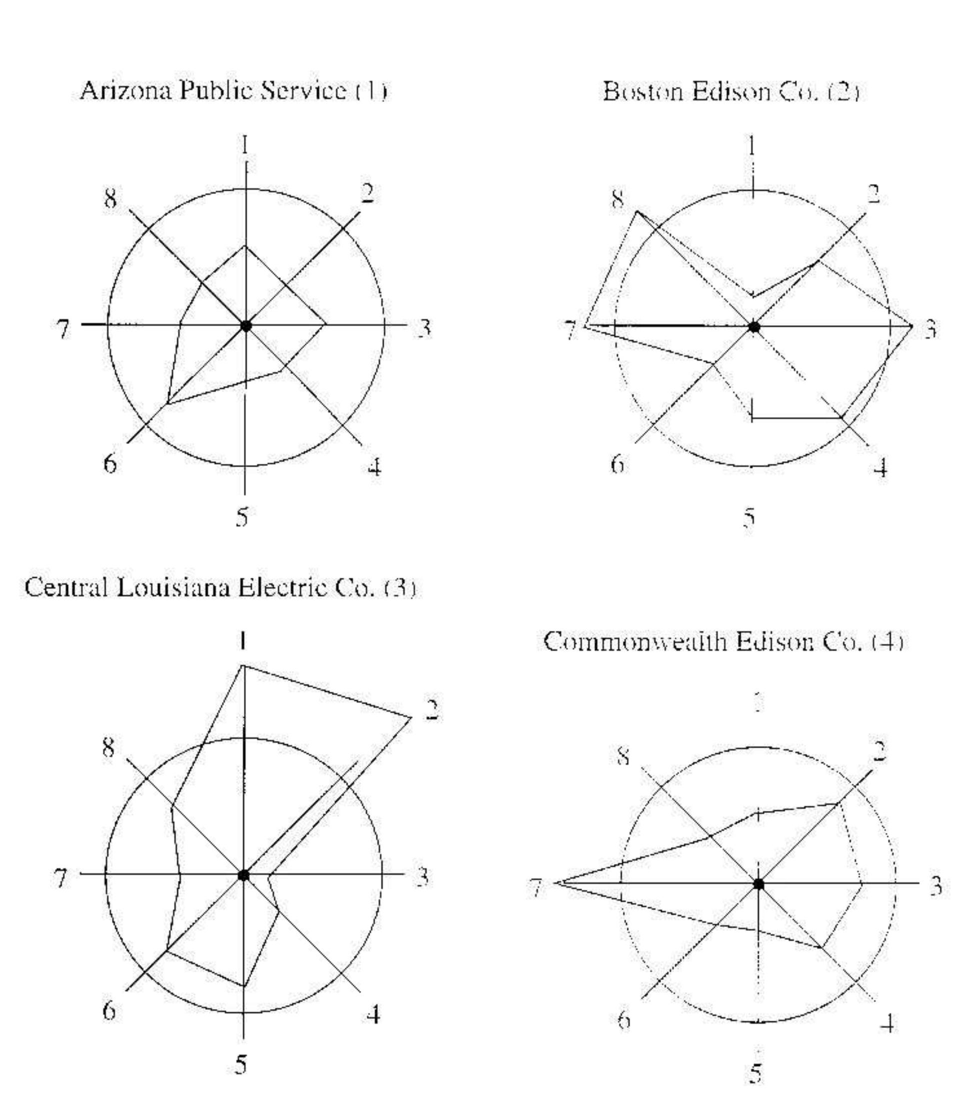
Trivariate Scatterplots



Graphical Techniques

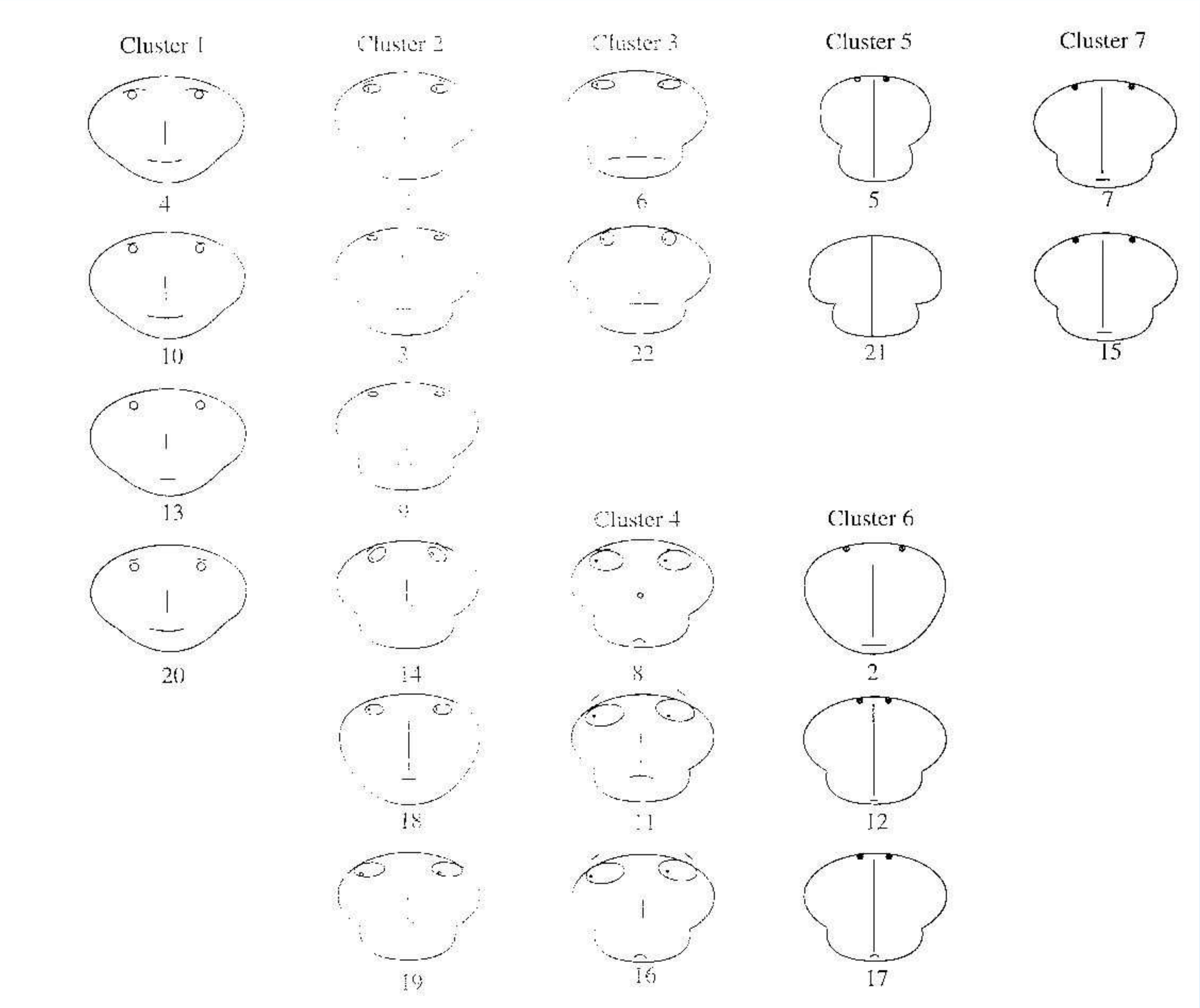
- But you already know those plots.
- Some plots that can be achieved by multivariate methods include:
  - “Stars.”
  - Chernoff faces.
  - Dendrograms.
  - Bivariate plots, but of the variable space.
  - Network graphs.

Stars



Arizona Public Service (1)
1
2
3
4
5
6
7
8
Boston Edison Co. (2)
1
2
3
4
5
6
7
8
Central Louisiana Electric Co. (3)
1
2
3
4
5
6
7
8
Commonwealth Edison Co. (4)
2
3
4
5
6
7
8

Chernoff Faces



Dendrograms

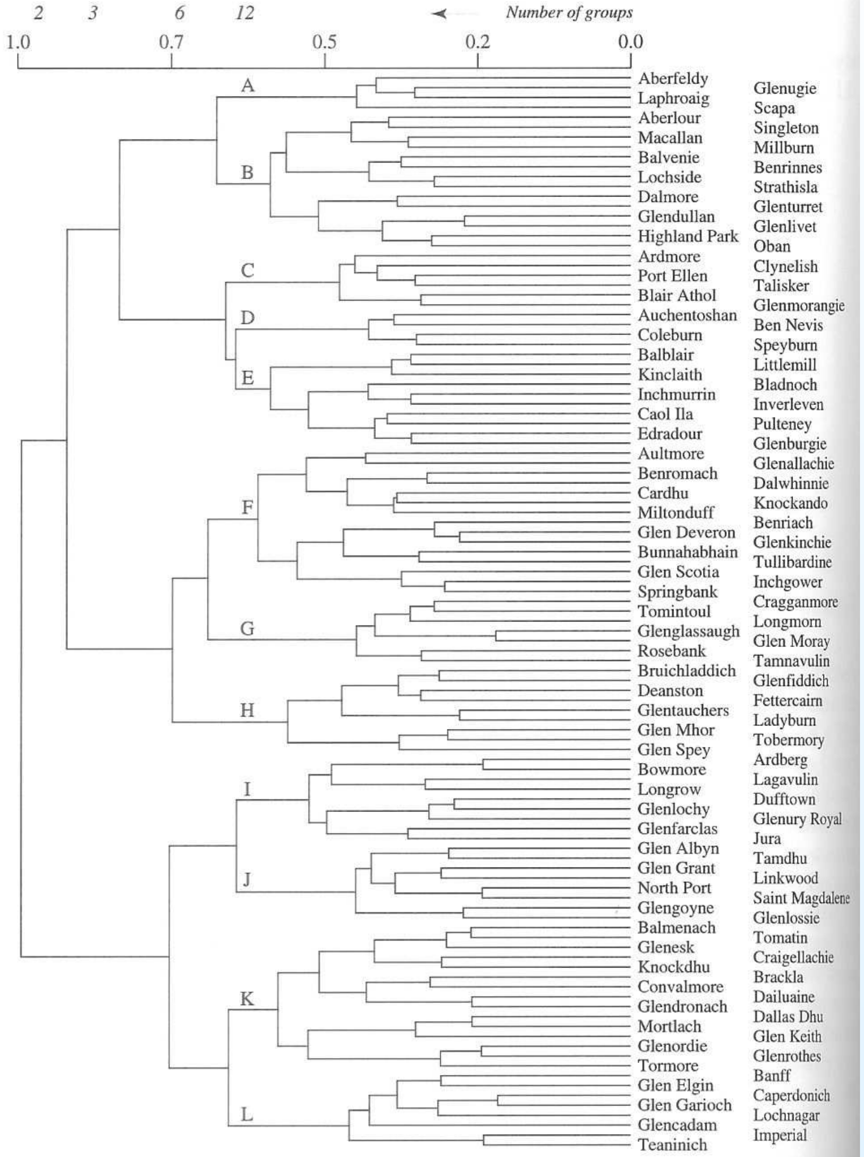


Figure 12.12 A dendrogram for similarities between 109 pure malt Scotch whiskies.

Variable Space Plots

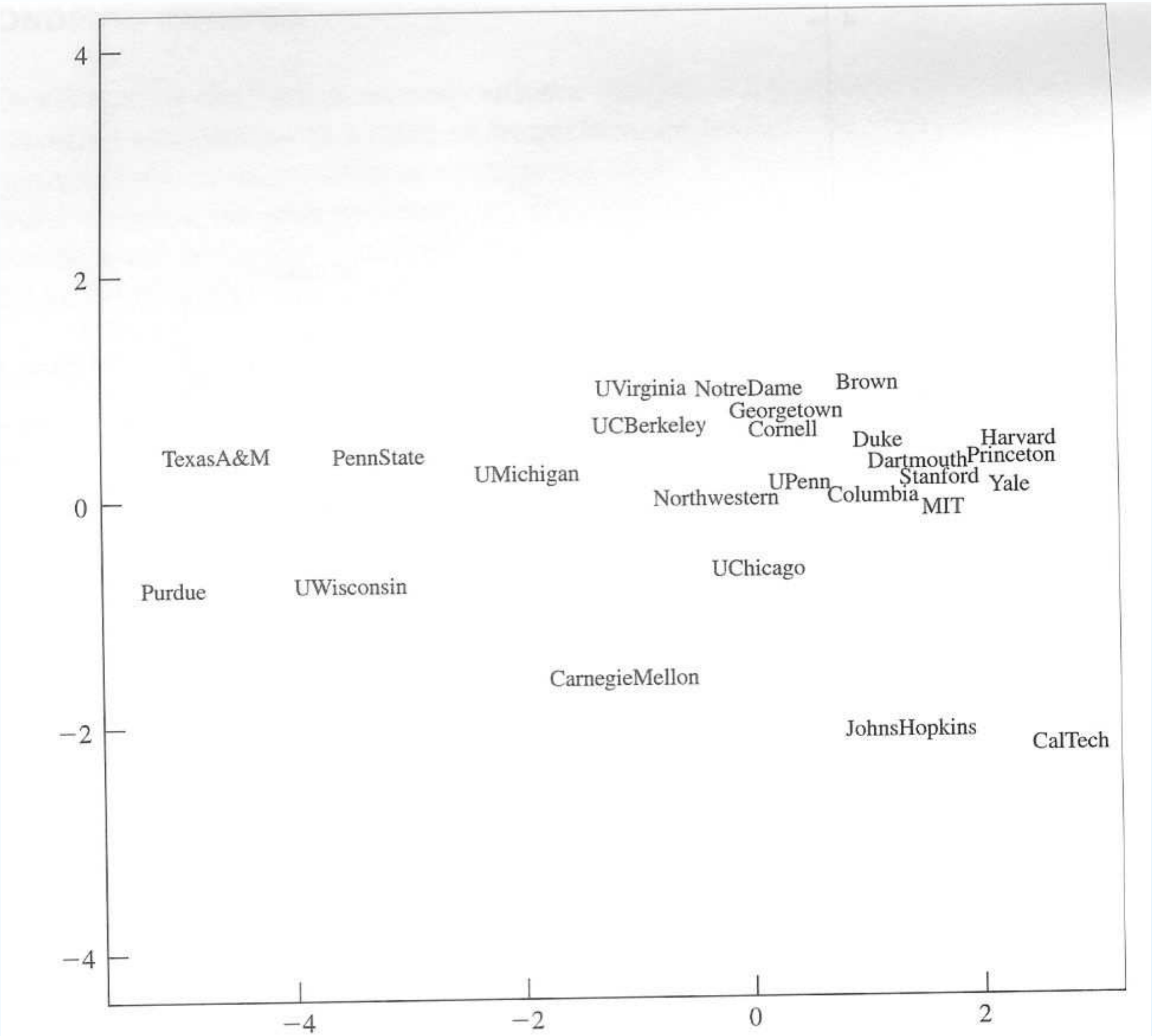
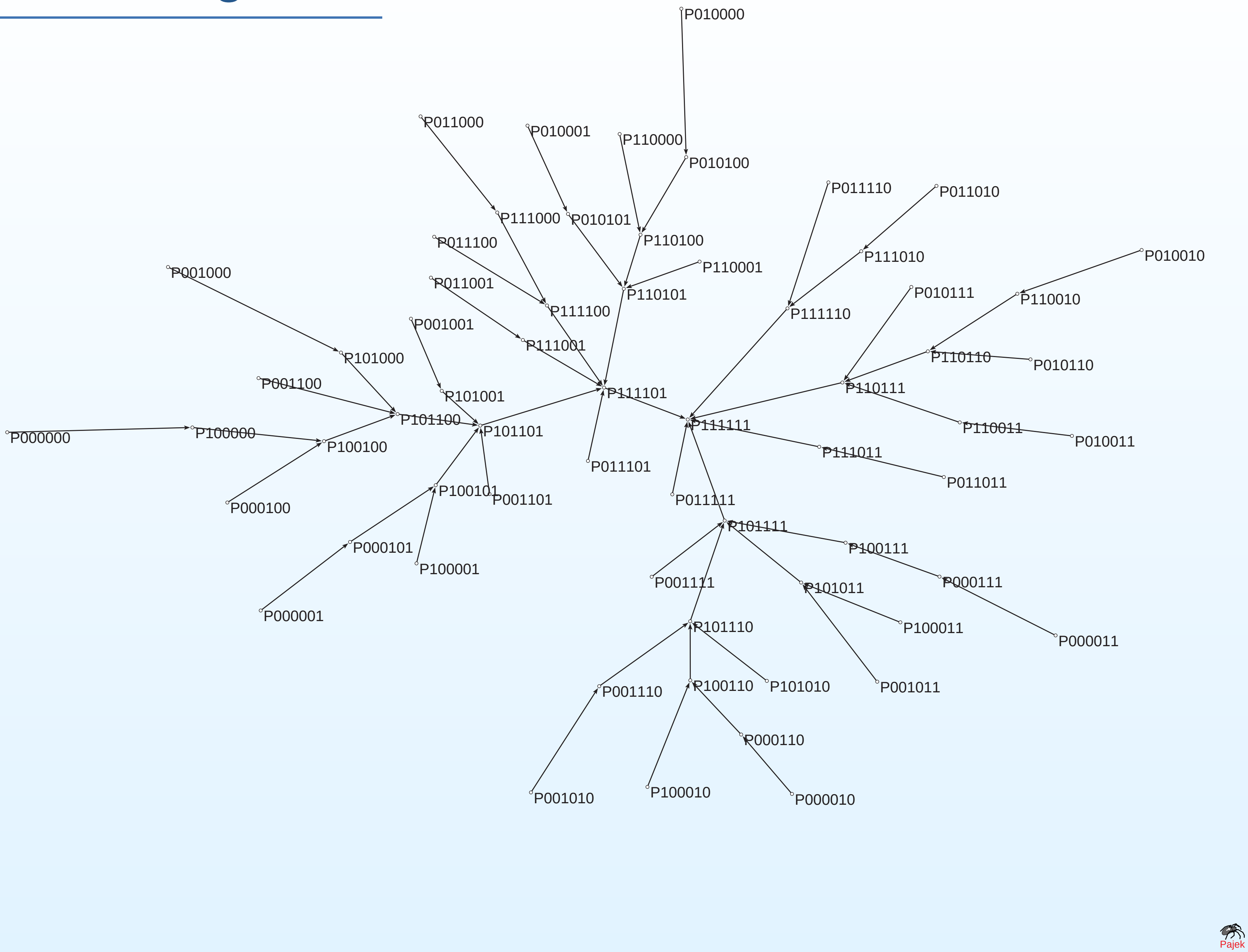


Figure 12.18 A two-dimensional representation of universities produced by metric multidimensional scaling.

Network Diagrams



Distance Measures

- A great number of multivariate techniques revolve around the computation of distances:
  - Distances between variables.
  - Distances between entities.
- The formula for the Euclidean distance formula between the coordinate pair $P = (x_1, x_2)$ and the origin $P = (0, 0)$:

$$d(O, P) = \sqrt{x_1^2 + x_2^2}$$

Distance Measures

- Elaborate discussions of distance measures will be found later in the class
- Just keep in mind that there are statistical analogs to distance measures, taking the variability of variables into account.
- Also be aware that there are literally an infinite number of distance measures!
- A distance measure must satisfy the following:
  - $d(P, Q) = d(Q, P)$
  - $d(P, Q) > 0$ if $P \neq Q$
  - $d(P, Q) = 0$ if $P = Q$
  - $d(P, Q) \leq d(P, R) + d(R, Q)$ (known as the triangle inequality)

Introduction to SAS

- SAS has a reputation for being...well...unliked by many in the social sciences.
- Why bother teaching it in this course?
  - New focus on SAS in our department.
  - Adding value to your degree (check out amstat.org or dice.com for details).
- For some good things about SAS, check out http://www.pbs.org/cringely/pulpit/pulpit20020411.html

Final Thought

- We just introduced what this course will be about.
- Things will become increasingly relevant as time progresses.

- Please be patient with SAS.
- We will now head down to the lab for a SAS introduction session.

Next Time

- Matrix algebra (Chapter 2, Supplement 2A)
- SAS *proc iml*