STAT 666:
MULTIVARIATE STATISTICAL METHODS

Dr. William F. Christensen
2197 WVB
Brigham Young University
801-422-7057
william@stat.byu.edu

Webpage (Syllabus):
`Learning Suite`

I. INTRODUCTION & OVERVIEW

I.A. Overview of Multivariate Concepts and Techniques

Concepts at the Heart of Multivariate Analysis

- Dimension reduction/simplification
- Latent structure
- Classification & Clustering
- Multivariate distance

Tools for Multivariate Analysis

- Linear combinations of variables
- Multivariate inference (to control experimentwise error rate)
- Eigenstructure of a matrix

Multivariate Analysis Techniques

- MANOVA & Hotelling's T^2
- Multivariate Regression
- Canonical Correlation
- Principal Components
- Factor Analysis
- Structural Equation Modeling
- Discriminant (& Classification) Analysis
- Cluster Analysis

MANOVA & Hotelling's T^2

- Many physiological and learning comprehension tests are given to students after being assigned to various eye exercise and chiropractic manipulation groups.
- Is there a difference in overall physiology or learning ability between groups?

Multivariate Regression

- At each of n business organizations, several business environment variables were measured along with several variables related to innovation.
- Can business environment variables predict differences in the multi-dimensional notion of innovation?

Canonical Correlation

- Various measures of risk-taking propensity and socioeconomic success characteristics for top-level business executives were collected.
- What combination of risk-taking behaviors is most closely related to another combination of success characteristics?

Principal Components

- For the "artificial nose" created at Tufts University with the Dept. of Defense, a large number of fiber optic cables are used to "sniff" the chemical structure of different odors.
- How many "essential dimensions" are there to smell for the artificial nose?

Factor Analysis

- Multiple 57-variate pollution observations are gathered over the course of two months.
- Despite the presence of a variety of point source and non-point source polluters, is there evidence that the 57-dimensional pollution process is being driven by only 3 pollution processes (industrial emissions, auto exhaust, and liquid gasoline evaporation)?

Structural Equation Modeling

- Several surrogate measures of family stability are obtained along with several measures of at-risk behavior in youth.
- Is there evidence to support a sociological theory that there is a significant negative relationship between an unobserved two-dimensional delinquency factor and an unobserved one-dimensional stability factor?

Discriminant & Classification Analysis

- After conducting an experiment to assess the effect of chiropractic manipulation (C.M.) on learning comprehension, it is determined that some subjects respond favorably, while others do not (i.e., we observe "responders" and "non-responders").
- Using the pre-treatment physiological measures of the subjects...
  - ...can we describe the nature of the separation of responders and non-responders in terms of physiology? [discriminant analysis]
  - ...can we construct a classification rule for predicting which subjects would have increased short-term learning comprehension do to the C.M.? [classification analysis]

Cluster Analysis

- For each of 81,000 3mm $\times$ 3mm $\times$ 5mm measured locations (voxels) in a functional magnetic resonance imaging (fMRI) experiment, hemodynamic (blood flow) response is measured while several cognitive tasks are carried out by the patient.
- Using only these multivariate responses, can we cluster voxels according to brain function?

I.B. Visualizing Multivariate Concepts

- Dimension reduction & latent structure
- Clustering
- Multivariate distance (unusual observations)

Dimension reduction & latent structure

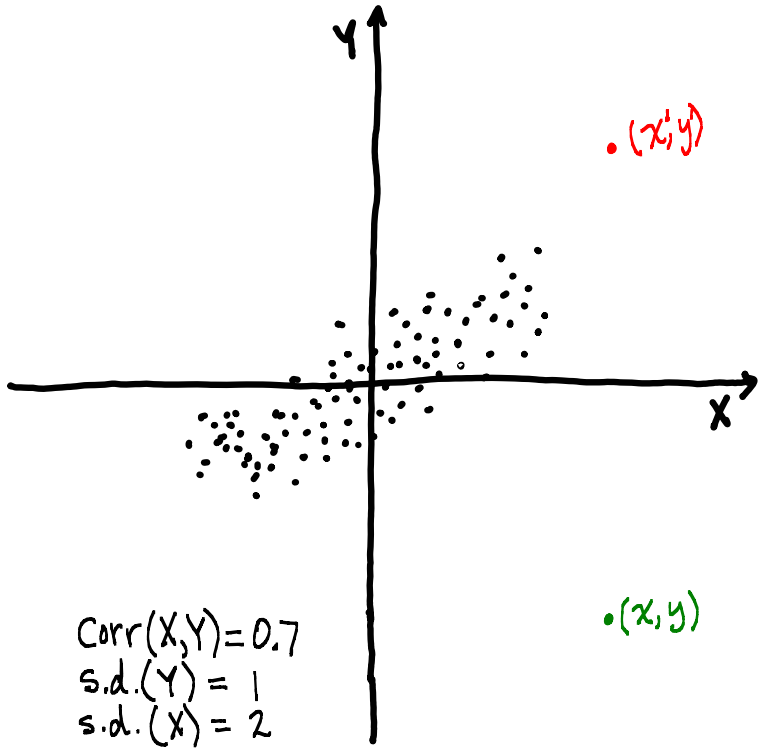
- Examples: SAT scores, IQ, dimensions of personality
- Visualizing 2-D, 3-D, 10-D structure (can we do that???)
- Essential dimensionality
- Data examples: Pollution data, Precision Ag. data

Clustering

- "Data Mining" issue – What kinds of customers are similar? How many species are there?
- Data example: Flea beetles data

Multivariate distance (for describing similarity among observations and for identifying unusual observations)

- "Data Mining" issue – Which transaction record is abnormal or suspicious?
- Distance



– Euclidean distance: defines distance from (x, y) to origin as

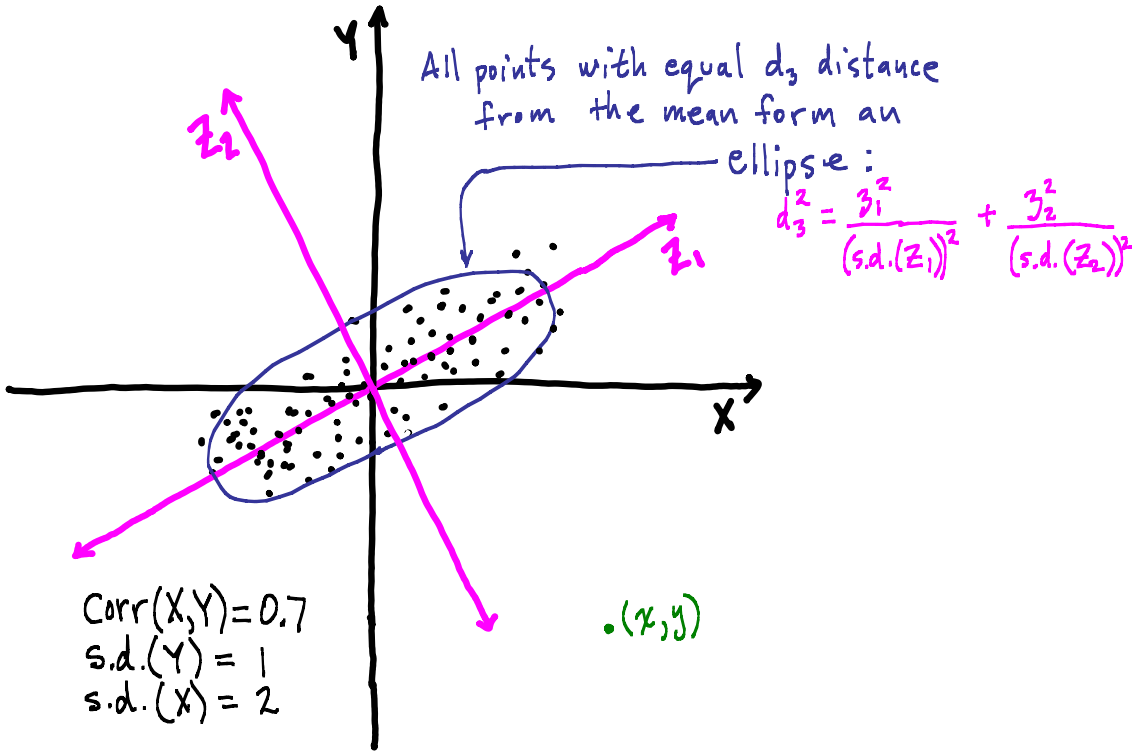
$$d_1(x, y) = \sqrt{(x-0)^2 + (y-0)^2}$$

  * Not a good choice....gives equal weight to x-coordinate and y-coordinate. Prefer to have more weight on the variable with lower s.d. $\Longrightarrow$ divide by s.d. before squaring to obtain...

– Standardized distance: defines distance from (x, y) to origin as

$$d_2(x, y) = \sqrt{\left(\frac{x-0}{2}\right)^2 + \left(\frac{y-0}{1}\right)^2}$$

  * Not a good choice....acts as if x and y are independent. Note that $d_2(x, y) = d_2(x, y)$ even though (x, y) is clearly more "weird" than (x, y)!



– Multivariate (Mahalanobis') distance

  * considers rotating axes until variables are uncorrelated and then redefining all points on this new (Z_1, Z_2) scale

$*$ defines distance from (x, y) (a.k.a. (z_1, z_2)) to origin as

$$
\begin{aligned}
d_3(x, y) &= \sqrt{\text{Mahalanobis' distance}} \\
&= \sqrt{\left(\frac{z_1 - 0}{\text{s.d.}(Z_1)}\right)^2 + \left(\frac{z_2 - 0}{\text{s.d.}(Z_2)}\right)^2} \\
&= \sqrt{\left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right]' \begin{bmatrix} 2^2 & 1.4 \\ 1.4 & 1^2 \end{bmatrix}^{-1} \left[\begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix}\right]}
\end{aligned}
$$

- What makes an observation an outlier?
- "Seeing" an outlier in multivariate data
- Data examples: Cities data, Pollution data

I.C. Data Displays

- Univariate analyses (histograms, boxplots, etc.)
- Star plots
- Chernoff faces
- Scatterplot Matrix
- Linked scatterplots & brushing
- "Grand Tour" (animated series of 2-D projections)

I.D. Multivariate Notation and Descriptive Statistics

Multivariate Notation

$$\begin{aligned} n &= \text{number of observations} \\ p\,(>1) &= \text{dimensionality} \\ &\quad (\text{sometimes referenced by } d) \end{aligned}$$

Vector notation for ith observation:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix} \leftarrow \text{a single vector of observations}$$

Data matrix:

$$\mathbf{X} = (x_{ij}) = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \underbrace{\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}}_{\text{columns are variables}} \leftarrow \begin{array}{l} \text{rows} \\ \text{are} \\ \text{indi-} \\ \text{viduals} \end{array}$$

x_{ij} = the jth variable of the ith item.

Descriptive Statistics

Example: Cities data

Mean vector:

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \frac{1}{n}\sum_{i=1}^{n} x_{i2} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \frac{1}{n}\mathbf{X}'\mathbf{1}$$

Covariance matrix (unbiased estimator):

$$\mathbf{S} = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ s_{21} & \cdots & s_{2p} \\ \vdots & & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} = \tfrac{1}{n-1}\mathbf{X}'(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X},$$

where

$$s_{ii} = s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2$$

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Covariance matrix (m.l. estimator):

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ s_{21} & \cdots & s_{2p} \\ \vdots & & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix} = \tfrac{1}{n}\mathbf{X}'(\mathbf{I} - \tfrac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X},$$

where

$$s_{ii} = s_i^2 = \frac{1}{n}\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2$$

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

Correlation matrix:

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

where $r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$

$$\mathbf{D} = \text{diag}(s_{11}, s_{22}, \ldots, s_{pp})$$

$$\mathbf{D}^{-1/2} = \begin{bmatrix} \left(\frac{1}{s_{11}}\right)^{1/2} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \left(\frac{1}{s_{pp}}\right)^{1/2} \end{bmatrix}.$$

$$\mathbf{R} = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

Properties:

1. $-1 \leq r \leq 1$
2. $r = 0 \Rightarrow$ No linear relationship
   $r > 0 \Rightarrow$ Positive linear relationship
   $r < 0 \Rightarrow$ Negative linear relationship
3. $\text{corr}\{x_1, x_2\} =$
   $\text{cov}\{(x_1 - \bar{x}_1)/\sqrt{s_{11}}, (x_2 - \bar{x}_2)/\sqrt{s_{22}}\}$
4. $\text{corr}\{x_1, x_2\} = \text{corr}\{ax_1 + b, cx_2 + d\}$, if a and c have same sign.