STAT 730 Chapter 1 Background

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 730: Multivariate Analysis

Logistics

- Course notes hopefully posted evening before lecture, based on Mardia, Kent, and Bibby.
- We will cover Chapters 1–14, skipping 7, in varying amounts of detail, spending about one week on each chapter.
- If time permits we will discuss additional topics, e.g. models specific to spatiotemporal data; decison trees; ensemble methods such as boosting, bagging, & random forests; Bayesian approaches.
- Your course grade will be computed based on graded homework problems.

Multivariate data

Multivariate data abounds. Often we have many variables recorded on each experimental subject.

- (f_i, q_i): beer drinking frequency and drinking quantity on 4-point ordinal scale for New Mexican DUI offenders along with abuse history, age, and gender $\mathbf{x}_i$.
- Temperature T_{ijkt} (^{o}C) at different locations and depths of pig heads in MRI coil over time. Essentially functional data.
- MRI $(T_{1\rho,i}, T_{2\rho,i})$ measurements on Parkinson's and non-Parkinson's patients, along with race, gender, and age $\mathbf{x}_i$.
- Survival times T_{ij} from diagnosis of prostrate cancer patients by county for South Carolinian men, along with marriage status, stage, grade, race, and age $\mathbf{x}_{ij}$.
- Fisher's iris data.
- Dental data from STAT 771.

Fisher's iris data

Fisher (1936) introduced these data which consist of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four measurements taken on each flower: the length and the width of the sepals and petals (cm). R code to obtain scatterplot matrix:

```
data(iris)
iris # examine data
plot(iris[,1:4],pch=c(rep(1,50),rep(2,50),rep(3,50)))
```

Dental data

Pothoff and Roy (1964) present data on $n = 27$ children, 16 boys and 11 girls. On each child, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was made at ages 8, 10, 12, and 14 years of age. R code to obtain spaghetti plot:

```
library(heavy) # data are in this package
data(dental) # attach data
dental # look at data
library(ggplot2) # used to make profile plot
ggplot(data=dental,aes(x=age,y=distance,group=Subject,color=Sex))+geom_line()
```

Each connected curve represents one child. Note boys typically larger than girls at each time point, but not perfect. How to model these data? How to test whether boys and girls are different?

Related topics/courses

We will study the theory of classical multivariate analysis in STAT 730, most of it based on a multivariate normal assumption. Classical techniques widely used. Courses related to STAT 730:

- STAT 517 & STAT 704, flexible regression models.
- STAT 771 Longitudinal Data Analysis (multivariate regression, MANOVA, non-normal extenstions)
- STAT 718 Statistical Learning (shrinkage priors in regression, dimension reduction, classification, smoothing, model selection and averaging, boosting, SVM & discrimination, clustering, random forests)
- CSCE 822 Data Mining and Warehousing (classification, clustering, dimension reduction, feature selection)
- CSCE 883 Machine Learning (dimension reduction, clustering, decision trees, discrimination, SVM, HMM)
- CSCE 768 Pattern Recognition (discrimination, neural networks, genetic algorithms, clustering, dimension reduction, HMM)
- PSYC 821 Theory of Psychological Measurement (scaling and factor theory, item analysis).
- PSYC 823 Multivariate Analysis of Behavioral Data (multiple linear regression, canonical correlation, discriminant functions, multidimensional scaling)
- PSYC 824 Special Topics in Quantitative Psychology (longitudinal data analysis, multilevel modeling, structural equation modeling).

Also closely related are the fields of Functional Data Analysis, where the data are curves or surfaces, and Shape Analysis.

Matrix review

Knowledge of elemetary linear algebra is assumed. You may want to review.

- A.2 Matrix operations
- A.3 Types of matrices
- A.4 Vector spaces, rank, linear equations
- A.5 Linear transformations
- A.6 Eigenvalues and eigenvectors, spectral decomposition
- A.7 Quadratic forms
- A.8–A.10 Generalized inverse, matrix differentiation, maximization, geometric ideas

Especially important: spectral decomposition of symmetric $\boldsymbol{\Sigma} > 0$ (p. 469). Also related: SVD.

Data matrix

To start, we have n objects (if human, subjects), each having p variables measured on each. Let x_{ij} be the jth variable measured on the ith object. The $n \times p$ data matrix is

$$\mathbf{X} = \left[\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{array} \right].$$

Each row has all p variables on an object and each column has all measurements on one variable across all objects.

Data vector

The *i*th object has the data vector

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}.$$

Then we can write $\mathbf{X}$ as

$$\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_n]' = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}.$$

Measurement vector

All n measurements of the the jth variable is

$$\mathbf{x}_{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}.$$

Then we can write $\mathbf{X}$ as

$$\mathbf{X} = [\mathbf{x}_{(1)}\mathbf{x}_{(2)} \cdots \mathbf{x}_{(p)}].$$

Univariate summary statistics

Let $\mathbf{1}_a = (1, 1, \ldots, 1)'$ be an $a \times 1$ vector of all 1's. The sample mean of the jth measurement is

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} = \frac{1}{n} \mathbf{1}_n' \mathbf{x}_{(j)}.$$

The p sample means are collected into the sample mean vector

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \frac{1}{n} (\mathbf{1}_n' \mathbf{X})' = \frac{1}{n} \mathbf{X}' \mathbf{1}_n.$$

The sample variance of the jth measurement is

$$s_{jj} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 = \frac{1}{n} (\mathbf{x}_{(j)} - \mathbf{1}_n \bar{x}_j)' (\mathbf{x}_{(j)} - \mathbf{1}_n \bar{x}_j) = \frac{1}{n} ||\mathbf{x}_{(j)} - \mathbf{1}_n \bar{x}_j||^2.$$

Covariance & correlation

The sample covariance of the ith and jth measurements is

$$s_{ij} = \frac{1}{n}\sum_{r=1}^{n}(x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j) = \frac{1}{n}(\mathbf{x}_{(i)} - \mathbf{1}_n\bar{x}_i)'(\mathbf{x}_{(j)} - \mathbf{1}_n\bar{x}_j).$$

The sample correlation between measurements i and j is

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

The sample covariance matrix is

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

Clever ways of writing $\mathbf{S}$

Note that the ijth element of $n\mathbf{S}$ is $(\mathbf{x}_{(i)} - \mathbf{1}_n\bar{x}_i)'(\mathbf{x}_{(j)} - \mathbf{1}_n\bar{x}_j)$. Thus

$$n\mathbf{S} = \begin{bmatrix} (\mathbf{x}_{(1)} - \mathbf{1}_n\bar{x}_1)' \\ \vdots \\ (\mathbf{x}_{(p)} - \mathbf{1}_n\bar{x}_p)' \end{bmatrix}_{p\times n} \left[\mathbf{x}_{(1)} - \mathbf{1}_n\bar{x}_1 \cdots \mathbf{x}_{(p)} - \mathbf{1}_n\bar{x}_p\right]_{n\times p}.$$

One matrix is the transpose of the other. Verify that the matrix on the right is $(\mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}$ where $\mathcal{I}_n$ is the $n \times n$ identity matrix. Therefore

$$n\mathbf{S} = [(\mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}]'[(\mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}] = \mathbf{X}'(\mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X} \overset{def}{=} \mathbf{X}'\mathbf{H}\mathbf{X},$$

as $(\mathcal{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')$ is *idempotent* (verify this). So

$$\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{H}\mathbf{X}.$$

Clever ways of writing $\mathbf{S}$

Note that the ijth element of the $p \times p$ matrix $(\mathbf{x}_r - \bar{\mathbf{x}})(\mathbf{x}_r - \bar{\mathbf{x}})'$ is $(x_{ri} - \bar{x}_i)(x_{rj} - \bar{x}_j)$.

Verify that

$$\begin{aligned} n\mathbf{S} &= \sum_{r=1}^{n}(\mathbf{x}_r - \bar{\mathbf{x}})(\mathbf{x}_r - \bar{\mathbf{x}})' \\ &= \sum_{r=1}^{n}\mathbf{x}_r\mathbf{x}_r' - n\bar{\mathbf{x}}\bar{\mathbf{x}}'. \end{aligned}$$

Correlation and covariance matrices

Let $\mathbf{R} = [r_{ij}]$ by the $p \times p$ correlation matrix. Let $\mathbf{D} = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}})$. Then

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1} \text{ and } \mathbf{S} = \mathbf{D}\mathbf{R}\mathbf{D}.$$

Recall that the inverse of a diagonal matrix simply takes the reciprocal of each entry.

Two measures of multivariate scatter

S measures scatter about the mean, as well as pairwise association among variables about the mean.

Two univariate measures of scatter derived from **S** are

- The generalized variance $|\mathbf{S}| = |\mathbf{R}| \prod_{j=1}^{p} s_j^2$.
- The total variation $\text{tr}(\mathbf{S}) = \sum_{j=1}^{p} s_j^2$.

The larger either is, the more the values of $\mathbf{x}_i$ are scattered about $\bar{\mathbf{x}}$.

Cork data

Rao (1948) gives the weight of cork deposits in centigrams for $n = 28$ trees on the north, south, east, and west sides of each tree.

The following R code reads the data and produces the sample mean vector, covariance matrix, and correlation matrix.

```
d=read.table("http://www.stat.sc.edu/~hansont/stat730/cork.txt",header=T)
d  # show the data matrix
mean(d)  # sample mean
cov(d)  # a bit off from what's in your book
S=cov(d)*27/28  # book uses n instead on (n-1) in denominator
cor(d)   # correlation matrix
det(S)  # generalized variance
sum(diag(S))  # total variation
```

Linear combinations

Let $\mathbf{a} \in \mathbb{R}^p$. A linear combination of the $\mathbf{x}_i$ is given by

$$y_i = \mathbf{a}'\mathbf{x}_i = a_1 x_{i1} + \cdots a_p x_{ip}.$$

Then

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}'\mathbf{x}_i = \mathbf{a}'\bar{\mathbf{x}}.$$

Note that $(y_i - \bar{y})^2 = \underbrace{\mathbf{a}'(\mathbf{x}_i - \bar{\mathbf{x}})}_{\in\mathbb{R}}\underbrace{(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{a}}_{\in\mathbb{R}} = \mathbf{a}'[(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})']\mathbf{a}$.

Then

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{a}'(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{a} = \mathbf{a}'\mathbf{S}\mathbf{a}.$$

A more general transformation

Now consider $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{q\times p}$ and $\mathbf{b} \in \mathbb{R}^p$. Then

$$\bar{\mathbf{y}} = \mathbf{A}\bar{\mathbf{x}} + \mathbf{b}$$

and

$$\mathbf{S}_{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})' = \mathbf{A}\mathbf{S}\mathbf{A}'.$$

You should show this. Finally, if $\mathbf{A} \in \mathbb{R}^{p\times p}$ and $|\mathbf{A}| \neq 0$ then

$$\mathbf{S} = \mathbf{A}^{-1}\mathbf{S}_{\mathbf{y}}(\mathbf{A}')^{-1}.$$

Standardizing and Mahalanobis transformations

Each measurement can be standardized via

$$\mathbf{y}_i = \mathbf{D}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}), \quad \mathbf{D} = \text{diag}(s_1, \dots, s_p),$$

where $s_j = \sqrt{s_{jj}}$ is the std. deviation of $\{x_{1j}, \dots, x_{nj}\}$. Each standardized measurement in $\{y_{1j}, \dots, y_{nj}\}$ has mean zero and variance one. Note that $\mathbf{S}_\mathbf{y} = \mathbf{R}$.

If $\mathbf{S} > 0$ (i.e. $|\mathbf{S}| \neq 0$) then $\mathbf{S}^{-1}$ has a unique pos. def. symmetric square root $\mathbf{S}^{-1/2}$. What is it? Hint: use the spectral decomposition. The Mahalanobis transformation is

$$\mathbf{y}_i = \mathbf{S}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}).$$

Note that $\mathbf{S}_\mathbf{y} = \mathcal{I}$; the transformed measurements are uncorrelated.

Principle components transformation

The spectral decomposition of $\mathbf{S}$ is

$$\mathbf{S} = \mathbf{GLG}',$$

where $\mathbf{G}$ has (orthonormal) eigenvectors of $\mathbf{S}$ as columns and the diagonal matrix $\mathbf{L} = \mathrm{diag}(l_1, \dots, l_p)$ has the corresponding eigenvalues of $\mathbf{S}$. The principle components rotation is

$$\mathbf{y}_i = \mathbf{G}'(\mathbf{x}_i - \bar{\mathbf{x}}).$$

Again, the transformed measurements are uncorrelated as $\mathbf{S}_\mathbf{y} = \mathbf{G}'\mathbf{SG} = \mathbf{L}$.

Note that the two measures of multivariate scatter are then $|\mathbf{S}| = \prod_{i=1}^{p} l_i$ and $\mathrm{tr}(\mathbf{S}) = \sum_{i=1}^{p} l_i$.

Geometry

- Comparing columns of $\mathbf{X}$, i.e. variables, use R-techniques: principle components analysis, factor analysis, canonical correlation analysis.
- Comparing rows of $\mathbf{X}$, i.e. objects, use Q-techniques: discriminant analysis, clustering, multidimensional scaling.

Example from R-space

Consider

$$\mathbf{Y} = \mathbf{HX},$$

the centered data matrix, i.e. each variable is shifted to be mean-zero. Then the cosine of the angle θ_{ij} between $\mathbf{y}_{(i)}$ and $\mathbf{y}_{(j)}$ is

$$\cos\theta_{ij} = \frac{\mathbf{y}'_{(i)}\mathbf{y}_{(j)}}{||\mathbf{y}_{(i)}||||\mathbf{y}_{(j)}||} = \frac{s_{ij}}{s_i s_j} = r_{ij}.$$

Note than when $\cos\theta_{ij} = 0$ the vectors are orthogonal, when $\cos\theta_{ij} = 1$ they coincide.

Example from Q-space

The n rows of $\mathbf{X}$ are points in $\mathbb{R}^p$. The Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is

$$||\mathbf{x}_i - \mathbf{x}_j||^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j).$$

A distance that takes correlation into account is the Mahalanobis distance

$$D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_j).$$

Mahalanobis distance underlies the Hotelling T^2 test and discriminant analysis; it is also used for comparing populations with different means but similar covariance.

Visualizing multivariate data

- Univariate displays of each variable: histograms, dotplots, boxplots, etc. Do not show association.
- Scatterplot matrix. Shows pairwise association.
- Parallel coordinate plots, or profile plots, or spaghetti plots.
- Harmonic curves.
- Chernoff faces, star plots, others. See examples on course webpage.

Profile plots & harmonic curves

Profile plot: Each measurement variable is a "point" on the x-axis. A connected line represents one row of **X**. Can color-code the lines to show different groups.

Harmonic curve: each measurement is a coefficient in a harmonic expansion

$$f_{\mathbf{x}_i}(t) = \frac{x_{i1}}{\sqrt{2}} + x_{i2} \sin t + x_{i3} \cos t + x_{i4} \sin 2t + x_{i5} \cos 2t + \cdots$$

In either case, similar curves implies similar sets of measurements.

Visualize Iris data

Scatterplot matrices:

```
data(iris)
plot(iris[,1:4],pch=c(rep(1,50),rep(2,50),rep(3,50)))

library(lattice)
splom(~iris[1:4],groups=Species,data=iris)
```

Harmonic curves:

```
library(andrews)
data(iris)
andrews(iris,type=1,clr=5,ymax=3)
```

Profile plots:

```
library(reshape) # to turn data matrix into list
library(ggplot2) # easier to use than lattice directly
iris$id=1:150 # add id variables for each iris
iris2=melt(iris,id.var=5:6,measure.vars=1:4) # one measurement per row
ggplot(data=iris2,aes(x=variable,y=value,group=id,color=Species))+geom_line()
```