Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Syllabus and Lecture 1 for Applied Multivariate Analysis

Outline

1. Course Description

2. Applied Multivariate Analysis-Outline

What's this course about

Applied Multivariate Analysis

Multivariate data through experimentation and observation occur quite often in engineering, business, social sciences, as well as biological and physical sciences. This is a course in applied multivariate data analysis. It will cover descriptive and graphical methods for continuous multivariate data, the multivariate normal, multivariate tests of means, covariances and equality of distributions, univariate and multivariate regression and their comparisons, multivariate analysis of variance, covariance structure models, and discrimination and classification. Furthermore it should be emphasized that this course and hence the chosen text, is designed around the application of multivariate techniques to continuous data, time allowing we will endeavor to discuss methods of discrete multivariate analysis from prepared class notes.

Software

Students will learn how to use statistical software to facilitate the understanding of the foundations of multivariate analysis. Statistical packages will include R, MatLab, SAS.

Prerequisites:

Students taking MthSc 807 are expected to have had a course in regression or intermediate data analysis (perhaps 805) and exposure to some statistical software although it is not required. The equivalence of MthSc 403/603 (Intro to Statistical Theory) can be considered preparatory for those students interested in multivariate data analysis. Prerequisites are a working knowledge of general linear models, statistical inference concerning these types of models, and hypothesis testing, and elementary matrix operations.

Attendance Policy:

All classes should be attended, but, if you are ill stay at home. I will accept e-mail or phone messages to that effect. Note that this does not exempt you from turning in homework/projects on time nor taking quizzes at their proposed times. Legitimate excuses must be offered with respect to the day(s) missed. Attendance will be monitored. It is to the instructors discretion whether an excuse is legitimate or not. Accordingly, the university's policy on religious holidays will be acknowledged and honored.

Tardy Professor Policy:

If the instructor is more than 15 minutes late for any class you may leave.

Examination Policy:

There will be **two** 50 minutes in class **closed book** examinations and a **closed book final** examination consisting of short answer questions requiring no more than a calculator. **No makeup examinations will be given**. Any student who misses an examination without a **legitimate excuse**,*ie, a documented medical excuse*, will receive a score of **zero** for that exam. A student with a **legitimate excuse**, will receive a final score based on all other class work. More than one missed exam with require withdrawal from the course and/or the receipt of a failing final grade.

Homework and/or Take Home Projects:

There will also be several homework sets and/or take home projects assigned from the text as well as from material covered during class. It is imperative that each student be completely comfortable with these assigned problems and projects. Homework is a critical part of the course. It will include both theoretical and applied work. Late homework will not be accepted (although you may turn in assignments early).

Homework Constraints

Homework should be detailed enough to adequately demonstrate your solution. You may discuss the homework problems with other students, however, the final work you turn in must be your own. **Don't copy other students' solutions.** Questions regarding them can be asked during office hours and via email. **All homework assignments must be sent electronically. This requires that solutions to homework sets and take home projects be typed. No hand written assignments will be accepted. All data sets will be available electronically to facilitate this being done. Homework sets will be due two class periods after they have been assigned.**

Grading Policy:

The two regular exams (20% each)(midterms if you wish) and a final exam (10%) which will count as 50% of the final grade, homework sets 10%, several multivariate exploratory data cases 20%, and a final multivariate data analysis project 20%. The final exam will cover the more important topics covered during the semester and will consist of several short answer questions requiring no more than a calculator.

Grading Scale:

A 100 - 90

B 89 - 80

C 79 - 70

D 69 - 60

F 59 -

Final Multivariate Data Analysis project

The purpose of this project is to have the student conduct a complete and thorough analysis of a multivariate data set. Data sets can be selected from journal articles or other referenced sources. The analysis should be completely novel from the approach taken in the source. The final project should include:

1. A one page description of the data set, which should contain at least 50 observations and 5 variables, including the source.
2. A one page proposed analysis, methods, and rational for the methods used.
3. The results of the analysis, including a printouts and graphics. Graphics should be incorporated into results. Source code should be included in the appendices along with a reference page sourcing the problem and the data.
4. Data from the text book will not be allowed, although simulation of data similar to those in the text will be allowed.

Final Multivariate Data Analysis project

Reports should be neatly typed, well-organized and attractive. Graphical displays (either computer-generated or hand-drawn) are encouraged. Generally, graphs are more effective if they are incorporated into the text, rather than hidden at the end of the report. You may also use a computer package to aid in the data analysis. If you do so, the results should be discussed in the text of your report, and the computer output itself may be included in an appendix.

A typed rough draft of the final report will be due approximately 2 weeks before the final report is due (last day of class).

The project is worth 100 points. Grades will be based on:

- Appropriate and correct procedures 50 pts;
- Well-written and attractive presentation 20 pts;
- Grammar, spelling and punctuation 20 pts;
- Complexity 10 pts.

Cell phones

Cell phones should be unseen and unheard during class.

You **are not** the exception to this rule. Students who fail to comply will be asked to leave the classroom. Use of them during class is prohibited. Please turn your phone off and store it away when you enter the classroom. It should not be left on your desk during class.

Academic Integrity:

"As members of the Clemson University community, we have inherited Thomas Green Clemson's vision of this institution as a "high seminary of learning." Fundamental to this vision is a mutual commitment to truthfulness, honor, and responsibility, without which we cannot earn the trust and respect of others. Furthermore, we recognize that academic dishonesty detracts from the value of a Clemson degree. Therefore, we shall not tolerate lying, cheating, or stealing in any form." "When in the opinion of a faculty member, there is evidence that a student has committed an act of academic dishonesty, the faculty member shall make a formal written charge of academic dishonesty including a description of the misconduct, to the Dean of the Graduate School. At the same time, the faculty member may, but is not required to, inform each involved student privately of the nature of the alleged charge."

Disability Access Statement

Students with disabilities who need accommodations should make an appointment with me to discuss specific needs within the first two weeks of classes. Students should present a Faculty Accommodation Letter(FAL) from Student Disabilities Services when we meet. Student Disability Services is located in G-20 Redfern (www.clemson.edu/sds/). Please be aware that accommodations are not retroactive and new FAL must be presented each semester.

Lecture Topics

Aug 24-31	3.1-3.12	Characterization of Multivariate Data
Sep 2-7	4	Multivariate Normal Distribution
Sep 9-16	5.1-5.3	Test on Mean Vectors
Sep 16	5.4-5.9	Comparing Two mean vectors
Sep 19-21	6.1-6.2	Multivariate Analysis of Variance
Sep 23-26	6.5-6.7	Two-Way Classification
Sep 28-30	6.8-6.11	Profile Analysis & Growth Curves
Oct 3-5	7	Tests on Covariance Matrices
Oct 7-14	8	Discriminant Analysis
Oct 17	Fall Break	
Oct 19-24	9	Classification Analysis
Oct 26-Nov 4	10	Multivariate Regression
Nov 7-14	12	Principal Component Analysis
Nov 16-21	13	Factor Analysis
Nov 23-25	Thanksgiving Break	
Nov 28-Dec 9	14	Cluster Analysis
Dec 9	14	(Last Day of Classes)

Aims of the next couple of meetings

- To explain what we mean by multivariate methods
- To provide an overview of the material we will cover
- To introduce the notation, to start thinking about linear combinations
- To provide some ideas on how to conduct a multivariate exploratory data analysis (eda)

Univariate Measures of Center and Variability

I. Mean and Variances of Univariate Data-Common measures of center and variability

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Generally, $\overline{y}$ will never be equal to μ; by this we mean that the probability is zero that a sample will ever arise in which $\overline{y}$ is exactly equal to μ. However, $\overline{y}$ is considered a good estimator for μ because $E(\overline{y}) = \mu$ and $\text{var}(\overline{y}) = \frac{\sigma^2}{n}$, where σ^2 is the variance of y. In other words, $\overline{y}$ is an unbiased estimator of μ and has a smaller variance than a single observation y. The variance σ^2 is defined shortly.

The notation $E(\overline{y})$ indicates the mean of all possible values of $\overline{y}$; that is, conceptually, every possible sample is obtained from the population, the mean of each is found, and the average of all these sample means is calculated. If every y in the population is multiplied by a constant a, the expected value is also multiplied by a: $E(ay) = aE(y) = a\mu$.

The variance of the population is defined as $\mathrm{var}(y) = \sigma^2 = E\left(y-\mu\right)^2$. This is the average squared deviation from the mean and is thus an indication of the extent to which the values of y are spread or scattered. It can be shown that $\sigma^2 = E\left(y^2\right) - \mu^2$. The sample variance is defined as

$$s^2 = \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}{n-1}$$

The sample variance s^2 is generally never equal to the population variance σ^2 (the probability of such an occurrence is zero), but it is an unbiased estimator for σ^2; that is, $E\left(s^2\right) = \sigma^2$. Again the notation $E\left(s^2\right)$ indicates the mean of all possible sample variances. The square root of either the population variance or sample variance is called the standard deviation. If each y is multiplied by a constant a, the population variance is multiplied by a^2, that is, $\text{var}(ay) = a^2\sigma^2$.

Diagnostic I

I. If $Z \sim N(0,1)$, what is the distribution of Z^2?

(a) F
(b) t
(c) χ^2_1
(d) χ^2_p

Diagnostic I

I. If $Z \sim N(0,1)$, what is the distribution of Z^2?

(a) F
(b) t
(c) χ^2_1
(d) χ^2_p

Diagnostic I

I. If $Z \sim N(0, 1)$, what is the distribution of Z^2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic I

I. If $Z \sim N(0,1)$, what is the distribution of Z^2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic I

I. If $Z \sim N(0, 1)$, what is the distribution of Z^2?

(a) F
(b) t
(c) χ^2_1
(d) χ^2_p

Diagnostic I

I. If $Z \sim N(0,1)$, what is the distribution of Z^2?

(c) χ_1^2

Diagnostic I

I. If $Z \sim N(0,1)$, what is the distribution of Z^2?

(c) χ_1^2

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0,1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(a) $N(0, 1^p)$

(b) t

(c) χ_1^2

(d) χ_p^2

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0,1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(a) $N(0, 1^p)$

(b) t

(c) χ_1^2

(d) χ_p^2

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0, 1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(a) $N(0, 1^p)$

(b) t

(c) χ_1^2

(d) χ_p^2

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0,1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(a) $N(0, 1^p)$

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0,1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(a) $N(0, 1^p)$
(b) t
(c) χ^2_1
(d) χ^2_p

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0,1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(d) χ_p^2

Diagnostic II

II. If $Z_1, Z_2, \ldots, Z_p \sim_{iid} N(0,1)$, what is the distribution of $Z_1^2 + Z_2^2 + \ldots + Z_p^2$?

(d) χ_p^2

Diagnostic III

III. What is:

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})?$$

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic III

III. What is:

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})?$$

(a) Variance

(b) Covariance

(c) Correlation

(d) Standard deviation

Diagnostic III

III. What is:

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})?$$

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic III

III. What is:

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})?$$

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic III

III. What is:

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})?$$

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic III

III. What is:

$$r = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})?$$

(b) Covariance

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(a) Variance

(b) Covariance

(c) Correlation

(d) Standard deviation

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(a) Variance

(b) Covariance

(c) Correlation

(d) Standard deviation

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(a) Variance
(b) Covariance
(c) Correlation
(d) Standard deviation

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(c) Correlation

Diagnostic IV

IV. What is the sample estimate of $\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$ also known as?

(c) Correlation

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(a) *Z*-test

(b) *t*-test

(c) *F*-test

(d) ANOVA

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(a) Z-test

(b) t-test

(c) F-test

(d) ANOVA

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(a) Z-test

(b) t-test

(c) F-test

(d) ANOVA

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(a) Z-test
(b) t-test
(c) F-test
(d) ANOVA

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(a) Z-test

(b) t-test

(c) F-test

(d) ANOVA

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(b) *t*-test

Diagnostic V

V. If $\bar{x}_1$ is the mean of group 1, and $\bar{x}_2$ is the mean of group 2, what test do we use to find whether $\mu_1 = \mu_2$?

(b) *t*-test

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

(a) *Z*-test

(b) *t*-test

(c) *F*-test

(d) *ANOVA*

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

(a) *Z*-test

(b) *t*-test

(c) *F*-test

(d) *ANOVA*

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

(a) Z-test

(b) t-test

(c) F-test

(d) *ANOVA*

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

(a) Z-test

(b) t-test

(c) F-test

(d) *ANOVA*

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

(a) Z-test

(b) t-test

(c) F-test

(d) $ANOVA$

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

(d) *ANOVA*

Diagnostic VI

VI. What test do you use if you have more than 2 groups?

  (d) *ANOVA*

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

(b) t

(c) χ^2_1

(d) χ^2_p

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

Diagnostic VII

last but not least

VII. If $X_1 \sim \chi^2_{\nu_1}$ and $X_2 \sim \chi^2_{\nu_2}$, what is the distribution of X_1/X_2?

(a) F

Summarising the data

Let's start with a gentle introduction to some matrix terminology.

- The values observed on the i^{th} individual can be written as the vector $\mathbf{y}_i^T = (y_{i1}, y_{i2}, \ldots, y_{ip})$
- If all the individual vectors are written as the rows of a table, the result is called the data matrix and is usually denoted by $\mathbf{Y}$. The element y_{ij} of this matrix contains the value of variable y_j observed on individual i.

And now, lets fix our ideas by thinking about some real data.

- To describe a multivariate data set it would help to be able to look at it, i.e. to have a pictorial representation.
- First, let us assume that all the variables are purely numerical.
- When there are just two such variables, a scatter diagram is a familiar idea for representing the data.
- The n individuals are represented by n points on a graph; the coordinates of point i are given by (y_{i1}, y_{i2}).

Life Lines

Points and Dimensions

A natural extension to a two-dimensional scatter diagram (scatterplot) where there are two variables in which a relationship is sought is the extension to p-dimensions. Let $\mathbf{y}$ represent a random vector of p variables measured on a sampling unit (subject or object). If there are n individuals in the sample, the n denoted by $\mathbf{y}_1$,$\mathbf{y}_2, \ldots$,$\mathbf{y}_n$, where each of the $\mathbf{y}_i$'s is a vector of length p and $\mathbf{y}_{i1}$,$\mathbf{y}_{i2}, \ldots$,$\mathbf{y}_{ip}$ coordinate axes are taken to correspond to the variables so that the i^{th} point is y_{i1} units along the first axis, y_{i2} units along the second,... and y_{ip} units along the p^{th} axis. Generally, this resulting "scatterplot" will result in a distinct pattern of variability, but should also reflect any similarities and well as dissimilarities among the n observations. Clustering may also be observed.

What is Applied Multivariate Analysis?

Multivariate random variables

Uppercase boldface letters are used for matrices of random variables or constants, lowercase boldface letters represent vectors of random variables or constants, and univariate random variables or constants are usually represented by lowercase nonbolded letters.

Representing Data

The Data Matrix

$$\mathbf{Y} \quad = \quad \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix}$$

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{ip} \end{pmatrix}$$

Mean Vectors

$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \\ \vdots \\ \bar{y}_p \end{pmatrix}$$

Let’s look at some data!