MAS465 Multivariate Data Analysis
MAS6011(1) Dependent Data

0. Introductory slides

School of Mathematics and Statistics

Dr Frazer Jarvis (J12)
a.f.jarvis@sheffield.ac.uk

Plan

- General information about the module
- Introduction to Multivariate Data Analysis

General information about the module

- This is MAS465 Multivariate Data Analysis for undergraduate students
- It is the first half of MAS6011 Dependent Data (with Time Series in Semester 2) for MSc students.

20 lectures

- Monday 3pm in Hicks Lecture Theatre 3
- Wednesday 12 noon in Hicks Lecture Theatre 6
- Week 7 (November 7–11) is a *Reading Week*, when there will be no lectures.

General information about the module

MOLE

- You will all be registered on the MAS6011 MOLE page.
  - Teaching materials will be available there at the appropriate point
  - There is also a discussion board
- There's no MAS465 MOLE page.

General information about the module

Lecture Notes

- Full lecture notes will be provided on MOLE in advance of the lectures.
  - This is the fourth time I have given the course. I will put complete notes from last year onto MOLE at the start of the module, but don’t be surprised if I edit material as I go along, possibly changing the order of sections etc.

General information about the module

There will be weekly *Task Sheets*, intended to be simple and quick short exercises to reinforce the lecture material.

- Not for handing in, although you should feel free to ask me – or better, to use the MOLE discussion board – if you have problems

There are more substantial *Exercise Sheets*, issued in Weeks 3, 6 and 9.

- These should be submitted in Weeks 5, 8 and 11.
- To be returned in Weeks 6, 10 and 11 (UG) or at the start of February (MSc)
- Not for formal assessment.

General information about the module

Assessment – MAS465

- 75% of the credit for the module comes from the exam, taken in January/February by the undergraduates
- The remaining 25% of the credit will be awarded for a project
  - The UG project will be set in Week 6 for submission by Friday 16 December, and feedback will be available from the start of the exam period

General information about the module

Assessment – MAS6011

- 70% of the credit for the module comes from the exam, taken in May/June for the MSc students. The exam covers material from both semesters.
- 15% of the credit will be awarded for a project in Multivariate Data Analysis, and 15% of the credit will be awarded for a project in Time Series (semester 2)
  - The MSc project will be set in Week 11 for submission at the end of January, and feedback will be given in mid-February (you should have had more precise dates already)

General information about the module

Other resources

- *Podcasts*
- *Books*

General information about the module

R

Modern-day statistics would be just about impossible without a computer package, and there are many to choose from. However, R has emerged as just about the universal choice amongst professional statisticians, and we will make no mention of any alternative.

Be aware, however, that professional data analysts might use other packages.

Make sure that you have access to a computer with R installed: the projects in the module will use data sets in R, and you will be expected to be able to perform your analysis in R.

Introduction to Multivariate Data Analysis

When you go to the doctor for a check-up, generally many things will be tested: height, weight, blood pressure, blood/urine samples etc.

This is very typical of making observations; it is rare that only one variable is tested and recorded. The corresponding data is therefore multivariate.

Multivariate data sets are far more common than univariate sets.

The aim of multivariate data analysis is to analyse multivariate data sets.

Introduction to Multivariate Data Analysis

Multivariate sets arise everywhere: archaeology, biology, economics, environment, health/medical, nature, social science, sports, weather etc.

Techniques of multivariate data analysis are used in all sorts of places, e.g., pattern recognition, machine learning, facial identification, bioinformatics etc.

There are many online sources of data sets.

Introduction to Multivariate Data Analysis

Given a data set to analyse, there are two phases.

1. There is an exploratory phase, where the analyst tries to spot patterns and other interesting features in the data;
2. If interesting features appear, and *if* there is reason to suppose the existence of some underlying statistical model for the data, then the analyst can form hypotheses, and carry out hypothesis tests by evaluating test statistics. This is the statistical phase.

Introduction to Multivariate Data Analysis

In the univariate case, the exploratory phase is generally rather uninteresting, and one moves quickly into the statistical phase.

In the multivariate case, however, the exploratory phase is much more interesting.

Although we will spend most of the second half of the course on the statistical phase, the exploratory phase will take up most of the early part of the course.

- Hardly any statistics
- Almost pure mathematical in nature

Introduction to Multivariate Data Analysis

Indeed, because there is often no obvious underlying statistical model, the exploratory phase (also known as data mining) is often much more useful than the statistical phase – there may be no statistical test available if the data is distributed in an unknown manner.

The exception comes with hypotheses about the mean; there is a multivariate version of the Central Limit Theorem, saying that the sample mean of a number of i.i.d. observations tends towards a normal distribution as the number of observations increases – this means that we can develop formal statistical tests for the mean, whatever the distribution of the data.

Introduction to Multivariate Data Analysis

A quotation:

- “Much classical and formal theoretical work in Multivariate Analysis rests on assumptions of underlying *multivariate normality* – resulting in techniques of very limited value”

(from R.Gnanadesikan, *Methods for Statistical Data Analysis of Multivariate Observations*)

Introduction to Multivariate Data Analysis

Problems in exploratory analysis

- **Data visualisation**
  - How do we visualise data in *p* dimensions if *p* is large?
    - very hard in lots of dimensions
    - even in 3 dimensions, looking at all 2-dimensional views misses information
  - With high dimensional data, can we find good ways to visualise it in fewer dimensions?
    - e.g., if two variables are strongly correlated, we could miss one of them out
  - What about different sorts of data, e.g., contingency tables, similarity matrices, etc.?

Introduction to Multivariate Data Analysis

Problems in exploratory analysis

- **Classification problems**
  - Given a data set, can we guess whether the data arose from a homogeneous source, or from a variety of different sources?
    For example, does a medical condition have subvariants?
    - Can we *cluster* the data into sets in a natural way?
    - There are many techniques for forming these groups; this is **cluster analysis**

Introduction to Multivariate Data Analysis

Problems in exploratory analysis

- **Discrimination problems**
  - Given a data set which we know come from different groups, and get a new sample, can we place the new sample in one of the groups, based on the observation?
    For example, if you know that a medical condition has subvariants, and a new patient is diagnosed with the condition, can you predict from the observation which subvariant the patient has?
    - This called **discriminant analysis**

Introduction to Multivariate Data Analysis

Problems in statistical analysis

- **Multidimensional statistics**
  - We need to understand multidimensional analogues of e.g., the binomial or normal distributions.
    - Can we recognise when data is/should be derived from a probability distribution of a particular type?
    - If so, can we develop multidimensional analogues of univariate statistics? (e.g., hypothesis testing, ANOVA, regression, linear models etc.)
    - Are there interesting new tests that we can develop which only work in a multivariate setting?

Introduction to Multivariate Data Analysis

Just like univariate data, multivariate data can be discrete, continuous, categorical, binary,. . . .

For example, responses to a survey might be age, height, marital status, sex,. . . .

We usually write n for the number of observations, each recording the values of p variables – for multivariate statistics, $p \geq 2$. p is the dimension of the data.

But we might easily have p being very large! A sequence of images might have Red-Green-Blue colour data recorded for each pixel – so each 5MP image might have 15000000 associated data values.

Generally, statistical methods work best when $n > p$, but the most useful exploratory methods are those which work more generally.