

STAT7017

Big Data Statistics

This research-led course provides an introduction to recent developments in Random Matrix Theory and Online Learning that addresses the challenges and opportunities posed by the availability of large amounts of data. In the first instance, we will review some classic results from multivariate statistical theory, matrix analysis, and probability theory. Then we will present the salient statistical features of big data (e.g., heterogeneity, noise accumulation, spurious correlation, and incidental endogeneity) and show how this impacts on traditional statistical methods and theory. We follow with an introduction to modern Random Matrix theory and its application in statistics. Applications presented may include topics such as high-dimensional statistical inference, large covariance matrices, large-scale statistical learning through subsampling, sparsification of large matrices, principal component analysis, and dimension reduction. We conclude with an introduction to the theory of online learning (aka. sequential prediction) to handle the situation of streaming data.

Mode of Delivery	On campus
Prerequisites	You must have completed STAT6039 and STAT6038 and STAT7018.
Course Convener	Dr Dale Roberts
Phone	(02) 612 57336
Email	dale.roberts@anu.edu.au
Office location	CBE 26c Room 3.48
Office hours	10:00 – 11:00 Monday
Research Interests	Probability theory, stochastic processes, machine learning, and applications.
Student Administrator	Maria Lander Maria.lander@anu.edu.au

COURSE OVERVIEW

Learning Outcomes

Upon successful completion of the requirements of this course, students should have the knowledge and skills to:

- Demonstrate understanding of the theoretical and practical differences between the analysis of Big Data compared to the traditional small- or medium-scale data setting
- Demonstrate the ability to apply the techniques and theory covered to new problems in the area of big data statistics
- Explain in detail the phenomena of heterogeneity, noise accumulation, spurious correlation, and incidental endogeneity that occur with big data
- Explain in detail the required theory of random matrices and their application in the analysis of high-dimensional data sets and covariance matrices
- Understand and apply the mathematical techniques required to analyse random matrices
- Develop new codes to analyse large datasets in the statistical package R
- Derive and apply results from probability theory and linear algebra to obtain justification of statistical methodologies in the analysis of big data

Assessment Summary

Assessment Task	Value	Due Date
Homework 1	15%	Week 3
Homework 2	15%	Week 6
Homework 3	15%	Week 9
Final Project (FP)	55%	Week 12

Research-Led Teaching

This course is solely largely based on recent research papers and surveys on the topic of random matrices in statistics. The topic is rapidly advancing and recent results may be introduced into the course as they appear in the literature.

Feedback

Staff Feedback

Students will be given feedback in the form of written comments, verbal comments, and feedback to the whole class.

Student Feedback

ANU is committed to the demonstration of educational excellence and regularly seeks feedback from students. One of the key formal ways students have to provide feedback is through Student Experience of Learning Support (SELS) surveys. The feedback given in these surveys is anonymous and provides the Colleges, University Education Committee and Academic Board with opportunities to recognise excellent teaching, and opportunities for improvement.

For more information on student surveys at ANU and reports on the feedback provided on ANU courses, go to

<http://unistats.anu.edu.au/surveys/selt/students/> and
<http://unistats.anu.edu.au/surveys/selt/results/learning/>

Policies

ANU has educational policies, procedures and guidelines, which are designed to ensure that staff and students are aware of the University's academic standards, and implement them. You can find the University's education policies and an explanatory glossary at: <http://policies.anu.edu.au/>

Students are expected to have read the [Academic Misconduct Rules 2014](#) before the commencement of their course.

Other key policies include:

- Student Assessment (Coursework)
- Student Surveys and Evaluations

Required Resources

None

Examination material or equipment

A non-programmable calculator.

Recommended Resources

Research papers and lecture notes will be provided.

COURSE SCHEDULE

Week	Summary of Activities	Assessment
1	Introduction to the challenges of Big Data and overview of the course. Review of some prerequisite concepts.	
2	Further matrix analysis, eigenvalues and eigenvectors, the multivariate normal distribution.	
3	Fundamental tools for studying limiting spectral distributions, Marcenko-Pastur distributions, Fisher spectral distribution.	HW1
4	CLT for linear spectral statistics: Introduction and integration tools.	
5	Moments and statistics of the Marcenko-Pastur distribution	
6	CLT for linear spectral statistics: Sample covariance matrix, Bai and Silverstein's CLT, CLT for random Fisher matrices.	HW 2
7	Generalised variance in higher dimensions.	
8	Multiple correlation coefficient.	
9	Multivariate linear regression in the high-dimensional setting	HW3
10	PCA and high-dimensional spiked population models	
11	Applications: Optimisation of large financial portfolios	
12	Review and future directions	FP

ASSESSMENT REQUIREMENTS

The ANU is using Turnitin to enhance student citation and referencing techniques, and to assess assignment submissions as a component of the University's approach to managing Academic Integrity. For additional information regarding Turnitin please visit the [ANU Online](#) website.

Students may choose not to submit assessment items through Turnitin. In this instance you will be required to submit, alongside the assessment item itself, copies of all references included in the assessment item.

Assessment Tasks

Homework 1, 2, 3

Details of task: The homework tasks will be small take-home assessments that will typically involve one 'pen-and-paper' question and/or a 'computational' question. The question(s) will cover material that has been seen in previous lectures and are aimed at ensuring students are routinely studying the material. STAT7017 may include additional questions compared to co-taught course STAT3017.

Final Project (FP)

Details of task: The final project will be a mix of theoretical and computational tasks. The project will focus on either a recent research paper or a particular theme/application. STAT7017 may include additional questions compared to co-taught course STAT3017.

Extensions and penalties

Extensions and late submission of assessment pieces are covered by the Student Assessment (Coursework) Policy and Procedure.

No submission of assessment tasks without an extension after the due date will be permitted. If an assessment task is not submitted by the due date, a mark of 0 will be awarded.

Returning assessments

The homework assessments are to be submitted online (Wattle) in digital format. This can take the form of either: (1) a scanned hand-written document (2) properly typeset document using LaTeX or Rmarkdown (see RStudio).

Referencing requirements


Students must correctly reference material that they have used in their assessments.

Scaling

Your final mark for the course will be based on the **raw** marks allocated for each of your assessment items. However, your final mark may not be the same number as produced by that formula, as marks may be **scaled**. Any scaling applied will preserve the rank order of raw marks (i.e. if your raw mark exceeds that of another student, then your scaled mark will exceed the scaled mark of that student), and may be either up or down.

Tutorial and /or Seminar Registration

Enrolment in tutorials will be completed online using the CBE Electronic Teaching Assistant (ETA). To enrol, follow these instructions:

1. Go to <http://eta.fec.anu.edu.au>
2. You will see the Student Login page. To log into the system, enter your University ID (your student number) and password (your ISIS password) in the appropriate fields and hit the Login button.
3. Read any news items or announcements.
4. Select "Sign Up!" from the left-hand navigation bar.
5. Select your courses from the list. To select multiple courses, hold down the control key. On PCs, this is the Ctrl key; on Macs, it is the  key. Hold this key down while selecting courses with the mouse. Once courses are selected, hit the SUBMIT button.
6. A confirmation of class enrolments will be displayed. In addition, an email confirmation of class enrolments will be sent to your student account.
7. For security purposes, please ensure that you click the LOGOUT link on the confirmation page, or close the browser window when you have finished your selections.
8. If you experience any difficulties, please contact the School Office (see page 1 for contact details).
9. Students will have until 5pm February 25 to finalise their enrolment in tutorials. After this time, students will be unable to change their tutorial enrolment.

SUPPORT FOR STUDENTS

The University offers a number of support services for students. Information on these is available online from <http://students.anu.edu.au/studentlife/>

Privacy Notice

The ANU has made a number of third party, online, databases available for students to use. Use of each online database is conditional on student end users first agreeing to the database licensor's terms of service and/or privacy policy. Students should read these carefully.

In some cases student end users will be required to register an account with the database licensor and submit personal information, including their: first name; last name; ANU email address; and other information.

In cases where student end users are asked to submit 'content' to a database, such as an assignment or short answers, the database licensor may only use the student's 'content' in accordance with the terms of service – including any (copyright) licence the student grants to the database licensor.

Any personal information or content a student submits may be stored by the licensor, potentially offshore, and will be used to process the database service in accordance with the licensors terms of service and/or privacy policy.

If any student chooses not to agree to the database licensor's terms of service or privacy policy, the student will not be able to access and use the database. In these circumstances students should contact their lecturer to enquire about alternative arrangements that are available.