Syllabus, DSC 155 & Math 182, Winter 2022
Hidden Data in Random Matrices

Instructor: Ioana Dumitriu
E-mail: idumitriu@ucsd.edu

Course website: www.math.ucsd.edu/∼dumitriu/m182.html
All information in this syllabus can be found in more detail on the course website.

Lectures	MWF 11-11:50am;
Office Hours	M 3-5pm, Thu 6-7pm
TA:	Haixiao Wang (see course website).

About this class. DSC 155 & MATH 182 is a one quarter topics course on the theory of Principal Component Analysis (PCA), a tool from statistics and data analysis that is extremely wide-spread and effective for understanding large, high-dimensional data sets. PCA is a method of projecting a high-dimensional data set into a lower-dimensional affine subspace (of chosen dimension) that best fits the data (in least squares sense), or equivalently maximizes the preserved variance of the original data. It is a computationally efficient algorithm (utilizing effective numerical methods for the singular value decomposition of matrices), and so is often a first-stop for advanced data analytics of big data sets.

The goal of this course is to present and understand the PCA algorithm, and then analyze it to understand how (and when) it works. Time permitting, we will then use these ideas to apply to some current interesting problems in data science and computer science.

Announcements. All announcements will be made via Canvas. We are starting the quarter online, with a tentative date to return to in-person instruction during the third week. Should this change, we will adjust to the new instructions.

Piazza. One of our most useful tools for discussions is our Piazza website (visit the course website to find out how to join). We encourage you to follow and participate in discussions on Piazza; the TA and I will be monitoring and answering the questions raised there.

Lectures: We are starting the quarter online. The lectures will be given live, by Zoom (links provided on Canvas). The Zoom lectures will be recorded, and the video recordings will be uploaded to the Media Gallery. If and when we return to in-person instruction, I will lecture in CSB 001, and will podcast as well.

In addition to lectures, I will be posting my slides on Canvas–both in the "Before" and "After" format.

Homework: There will be a total of 4 homework assignments, with posting and due dates on the Calendar posted on the course website. Both homeworks and solutions will be posted on Canvas; solutions will appear in a special "Homework Solutions" directory in Canvas Files.

All homework will be due on the date indicated on Canvas, by **10:59pm**, in Gradescope. It is allowed (and even encouraged!) to discuss homework problems with your classmates and your

instructor and TA, but your final write up of your homework solutions must be your own work. If you collaborate on the homework, you must list the people you collaborated with on the write up.

Labs: The data science labs will be accessible through DataHub. There will be 5 of them, and the postings/due dates are available of the course website calendar. The turn-in components should be exported as pdf files and turned in through Gradescope; they are due at **10:59pm** on the dates indicated on the labs and the calendars. **Make sure to attend the first Lab/Discussion session on 01/06.**

Lab Project. You will choose a real-world high-dimensional data set, and implement the PCA algorithm to analyze it. You will use the tools explored in this class to give a careful analysis of how the PCA algorithm performed, what it discovered about the data, and what structural shortcomings were evidence in the analysis. Topics and data-sets to be approved by the instructor.

Take-Home Midterm Exam. There will be a single take-home midterm exam, available immediately after the lecture on Monday, February 7, due the following day before 10:59pm. You are free to use any paper / online resources you like during the exam, but **collaboration with other people is not allowed**. You will upload your midterm to Gradescope; all uploads must be done by 10:59m on February 8. You will sign an Academic Integrity Pledge (which will be provided, and must also be uploaded.) No exceptions.

Among the things you will agree to by signing the Academic Integrity Pledge will be the following: ***In case the instructor suspects academic misconduct in completing the exam, you will be invited to defend your solutions with the instructor and/or the TA. If you decline, or if you accept but fail to defend your solution(s) to the instructor's or the TA's satisfaction, the instructor will refer your case to the Academic Integrity Office for an investigation. If you accept and defend your solution(s) satisfactorily, the case will be closed and you will receive a grade for the exam.***

Final Exam. If in-person, the final exam will be held on March 14, from 11:30am-2:29pm, location TBA. If we are still remote by then, we will make other arrangements. The final will be open book.

Grading. Your total grade will be computed as follows.

- 20% Homework, 15% Labs, 20% Take-Home Midterm, 10% Final Project, 35% Final Exam
- 20% Homework, 15% Labs, 10% Take-Home Midterm, 20% Final Project, 35% Final Exam.

If you become sick and you cannot take the final exam, and had up to that point passing grades, you will get an Incomplete (see below).

Incomplete Grades. The only way to obtain an Incomplete is if a student had been doing satisfactory work up until the final exam, and then misses the final exam because of a good (preferrably documented) excuse. If the excuse is undocumented, it will be up to the instructor to decide whether to grant the Incomplete.