M392C/CSE392: Geometric Methods in Data Science

The University of Texas at Austin, Fall 2021

Times: TuTh 2-3:30PM CT
Location: Face-to-face in PMA 11.176
Meetings: 28 classes, Th Aug 26 – Th Dec 2, with Th Nov 25 off
Instructor: Joe Kileel, jkileel@math.utexas.edu
Office hours: Wed 1-3PM CT, simultaneously in POB 3.434 and on *Zoom*
Supporting media: *Canvas* with lectures (hopefully) streamed and recorded

Description: This is a graduate topics course on the mathematics of data science. We will mostly follow the book-in-progress called "Mathematics of Data Science" by Bandeira, Singer and Strohmer. I will provide the class with a private up-to-date version of the book project, kindly provided by the authors. Please do not circulate this draft. Time permitting, I plan to supplement the text with survey lectures on tensor decomposition, group actions in data science, and Riemannian optimization.

Where possible, we plan to emphasize geometric aspects in data science. The motivation for doing this is that in many applications the data itself carries an intrinsic geometry (for example, the data consists of 3D volumes or 2D images), while in other applications, the totality of the data set is well-modeled by a low-dimensional space. We will see that it is useful to take this information into account.

Prerequisites: The course's main prerequisites are linear algebra, basic probability, and mathematical maturity. Basic programming familiarity (or a willingness to learn) will help with some of the homework and some of the projects. A few elements of differential and algebraic geometry will be sketched along the way.

Grading: Since this is a graduate topics course, we would prefer to focus on how this course can help you to achieve your research goals. That said, grading will be based 50% on homework and 50% on a final project, with a generous curve for letters.

Homework: In the first half of the course, there will be four assignments. They will be released on the Tuesdays Aug 31, Sept 14, Sept 28, Oct 12, and then due on the Thursdays Sept 9, Sept 23, Oct 7, Oct 21, respectively. Use of computers may help with a few of the problems. Since I do not have a grader for this course, we will try to use a double-blind peer review system where each of your homework assignments is graded by at least two of your classmates, in-between the Thursday the assignment was due and the next Tuesday. In this setup, I will serve as the editor/meta-reviewer. Each assignment will count for 10% and each peer review cycle for 2.5%.

Final project: In the second half of the course, students will focus on a final term project. This is to be on a relevant topic (related to but not necessarily contained in the course). I must approve the topic you choose. You will produce a written report (rough guideline: 10 pages) and give a short presentation to the rest of the class. Projects may be computational and primarily applying methods to real data sets. Or they may be more theoretical, summarizing or comparing theoretical papers. You may collaborate with one other student in the course, or work individually. Where possible, originality is encouraged. The best outcome (although not required) is that the class project sparks a research project which you are motivated to work on later. A proposal (1-2 pages) for the final project is due on Tuesday, Oct 26. A first draft (ungraded, but I will provided comments/suggestions) is due on Tuesday, Nov 16. Presentations with constructive Q&A's will occur in class during Thursday, Nov 18 – Thursday, Dec 2. The final writeup is due Sunday, Dec 12 at 5PM CT. The proposal is worth 5%, the presentation 10% and the final writeup 35%. I will grade the projects.

Lectures: Lectures will be face-to-face as long as the university policy permits this. I will set up streaming and recording of the lectures for students who prefer that medium. Here is a tentative schedule of topics for the lectures, subject to change:

Lecture #	Date	Topics	Refs
1	Th, Aug 26	High dimensions	BSS Ch 2
2	Tu, Aug 31	High dimensions	BSS Ch 2
3	Th, Sept 2	SVD & PCA	BSS Ch 3
4	Tu, Sept 7	SVD & PCA	BSS Ch 3
5	Th, Sept 9	Graphs & clustering	BSS Ch 4
6	Tu, Sept 14	Graphs & clustering	BSS Ch 4
7	Th, Sept 16	Diffusion maps	BSS Ch 5
8	Tu, Sept 21	Diffusion maps	BSS Ch 5, Belkin-Niyogi
9	Th, Sept 23	Concentration	BSS Ch 6
10	Tu, Sept 28	Concentration	BSS Ch 6
11	Th, Sept 30	Max cut	BSS Ch 7
12	Tu, Oct 5	Community detection & SDPs	BSS Ch 8
13	Th, Oct 7	Community detection & SDPs	BSS Ch 8
14	Tu, Oct 12	Random projections	BSS Ch 9
15	Th, Oct 14	Random projections	BSS Ch 9
16 (guest lecturer: J. Pereira)	Tu, Oct 19	Statistics	BSS Ch 13
17 (guest lecturer: J. Pereira)	Th, Oct 21	Statistics	BSS Ch 13
18	Tu, Oct 26	Statistics	BSS Ch 13, other
19	Th, Oct 28	Estimation under group actions	various papers
20	Tu, Nov 2	Nonlinear algebra	Michalek-Sturmfels
21	Th, Nov 4	Tensor decomposition	Landsberg
22	Tu, Nov 9	Tensor decomposition	Kolda-Bader
23	Th, Nov 11	Riemannian optimization	Boumal
24	Tu, Nov 16	Riemannian optimization	Boumal
25	Th, Nov 18	Project presentations	–
26	Tu, Nov 23	Project presentations	–
–	Th, Nov 25	Thanksgiving holiday	–
27	Tu, Nov 30	Project presentations	–
28	Th, Dec 2	Project presentations	–
–	Sun, Dec 12	Project writeup due	–

(BSS = Bandeira, Singer, Strohmer; the other references will be explained in *Canvas*)

Covid: You must follow all university and governmental rules. Please also exercise common sense. Legally, I am only allowed to encourage vaccine and mask use.