Keck School of Medicine of USC

PM 579: Statistical Analysis of High-Dimensional Data
Units: 4 units
Term–Day–Time: Summer 2021, Wed 1:00 PM-4:50 PM
Location: online

Course Coordinator(s):
Name: Kimberly Siegmund Contact: kims@usc.edu
Office: Office Hours: By Appointment

Course Instructor(s):

Teaching Assistant(s):

Course Description

The course is designed for M.S. and Ph.D. students in the biological or mathematical sciences, and is highly recommended for Biostatistics Ph.D. students in the Statistical Genetics track. The course aims to provide students with a broad overview of current statistical problems and approaches to high-dimensional data analysis. The content will cover methods for classification and class discovery, using data sets for gene expression and DNA methylation. This course will be taught with an emphasis on selecting the appropriate statistical method for data analysis and interpreting the results. We will learn and use the R statistical computing language and Bioconductor, open source software for the analysis of genomic data.

Learning Objectives

After completing this course, the student should be able to:

1. Create figures to visualize high-dimensional data sets
2. Apply statistical hypothesis testing in a high-dimensional data setting while controlling the error rate
3. Apply the proper statistical analysis method to high-dimensional data
4. Interpret the results from a statistical analysis of high-dimensional data
5. Present analysis results to readers who are not familiar with the statistical methods

Prerequisite(s): PM510 or equivalent

Co-Requisite(s):

Concurrent Enrollment:

Recommended Preparation: PM511a

Teaching & Assessment Methods

Teaching Methods

- Assigned reading/writing (texts)
- Assigned reading (journal or papers)
- Online activity
- Classroom lecture
- Group activity
- Student presentation
- Recorded lecture

Assessment Methods

- Quiz
- Oral presentation
- Team based learning
- Term paper
- Other

Course Notes

A letter grade will be given for this course. The instructor will use Blackboard and GitHub for file sharing.

Communication

Students are encouraged to contact the instructor by email and during office hours. The instructor will reply to emails within 48 hours, 72 hours over a weekend, and the work day following a holiday.

Technological Proficiency and Hardware/Software Required

Students will be asked to bring a laptop to each class. They will learn R programming and the use of R and Bioconductor packages freely available from the following websites: http://www.r-project.org/ http://www.bioconductor.org/ Computer code will be shared through GitHub, https://github.com/ksiegmund/PM579.

Required Materials

- The textbooks will be Bioconductor Case Studies, Springer Inc., editors Hahn F, Huber W, Gentleman R, Falcon S. 2008, and freely available from USC Library. (HHGF)

  Grolemund G, Wickham H. R for Data Science, 2016. available at: https://r4ds.had.co.nz/. (GW)

Optional Materials

- R scripts for textbook: Bioconductor Case Studies
  http://www.bioconductor.org/help/publications/books/bioconductor-case-studies/web-supplement/

Description and Assessment of Assignments

Weekly homework assignments will be given to provide experience in applications of standard methods to real data. During Summer offerings (12 week schedule), no more than 12 homeworks will be assigned. Students will give one oral presentation on a topic of their choice.

In lieu of a final exam, students will be asked to prepare a term paper, due on the last day of class. The term paper could be in the form of
(1) a grant application,
(2) a report on a statistical analysis of high-dimensional data, or
(3) a report on comparing a number of statistical techniques on a (high-dimensional) data set. The data may be your own, or some obtained from the public domain.

Grading Breakdown

Assignment	**% of Grade**
Participation	10
Homeworks	40
Oral Presentation	25
Written Project	25
Total	**100**

Grading Scale

Course final grades will be determined using the following scale.

A	90-100
A-	85-89
B+	82-84
B	79-81
B-	75-78
C+	71-74
C	67-70
C-	63-66
D+	59-62
D	57-58
F	56 and below

Course-specific Policies

Assignment Submission

Coding assignments will be submitted through GitHub Classroom and reports will be submitted through Blackboard.

Grading Timeline

Homework will be graded within 1 week of turning in.

Late work

Late assignments will be accepted, but no later than the last day of class. Late work will be penalized by 20% deduction in points for each week the assignment is late unless due to an emergency situation excused by the instructor. Email the instructor as soon as possible to discuss alternate arrangements due to an emergency.

Technology in the classroom

On ground students will bring a laptop to class each week for computer lab. Online students will be expected to attend live Zoom sessions through a computer/laptop allowing them to share their screen.

Academic integrity

A grade of zero will be applied to submitted work that does not comply with the USC standards of academic conduct. Such work may not be resubmitted for a new grade. Academic integrity is included at the end of the syllabus.

Attendance

Wednesday class discussions will be recorded for asynchronous viewing.

Classroom norms

Cameras are not required for the Zoom sessions. If you are able and willing to use video during the breakout sessions, you are encouraged to use it then.

I expect students to treat others with compassion and understanding, and to have a genuine interest in learning. Students are encouraged to offer tips and feedback to the instructor, as the class gets adapted to the online format.

Expectations on Student Engagement

Students are expected to actively participate in class discussions and work in small groups for in-class exercises scheduled during the synchronous Zoom sessions.

Course evaluation

Policy on Learning & Assessment Feedback (LAF)

Feedback on examinations will be provided using the following methods. Please indicate which method(s) you will use in the course.

- In-office review (with specific conditions to be defined for each assessment)

Course Schedule: A Weekly Breakdown

Date	Topic	Lecturer
Wed 05/19/21 01:00p - 04:50p	Introduction to molecular biology and high throughput technologies Learning Objectives: 1. Create a Reproducible Report 2. Manipulate high-dimensional data in R Reading: Class Material on Blackboard GW Welcome & Ch 1 Homework 1: (due 5/26) Watch Video: http://videolectures.net/cancerbioinformatics2010_baggerly_irrh/ turn in hw1 using GitHub Classroom	Kimberly Siegmund

Wed 05/26/21 01:00p - 04:50p	Data Visualization Learning Objectives: 1. Conduct at least 2 different dimension reduction techniques. 2. Create a figure to show the results from a dimension reduction method 3. Write the methods section for a journal article describing the data and your analysis Readings: Class handouts/videos provided on Blackboard Homework 2: Turn in using GitHub Classroom (due 6/1)	Kimberly Siegmund
Wed 06/02/21 01:00p - 04:50p	Unsupervised Learning Learning Objectives: 1. Compare and contrast different unsupervised learning methods 2. Apply at least two different unsupervised learning methods 3. List three decisions/actions of the method that influenced your analysis Reading: On Blackboard & HHGF Ch. 10 Homework 3: Turn in using GitHub Classroom (due 6/8)	Kimberly Siegmund
Wed 06/09/21 01:00p - 04:50p	Differential Expression Learning Objectives: 1. Explain why moderated t-tests are applied to gene expression data 2. Apply moderated t-tests to gene expression data Topics: Fold-change, volcano plots, moderated t tests, annotation, GSEA Reading: HHGF Ch. 3.4.1-3.4.4, 7.1-7.3, 7.5 Class handouts provided on Blackboard Apply methods using R Homework 4: Turn in using GitHub Classroom (due 6/15)	Kimberly Siegmund
Wed 06/16/21 01:00p - 04:50p	Multiple Testing I Learning Objective: 1. Interpret the family-wise error rate 2. Interpret the false-discovery rate Reading: Class Material on Blackboard References: Benjamini & Hochberg (1995); Storey & Tibshirani (2003)	Kimberly Siegmund
Wed 06/23/21 01:00p - 04:50p	Multiple Testing II Learning Objectives: 1. Interpret multiple comparison adjusted p-values, q-values and the posterior error probability	Kimberly Siegmund

	2. Describe a simple method to increase power for multiple hypothesis testing 3. List the statistical quantities required for power calculations Reading: Class Material on Blackboard References: Bourgon et al. (2010, PNAS), Jung (2005) Homework 5: Select Topic for Student Presentations (due: 6/30)	
Wed 06/30/21 01:00p - 04:50p	RNA Sequencing Learning Objectives: 1. Select the proper statistical model to differential gene expression (DGE) of RNA-seq data 2. Apply computational pipeline for DGE analysis Reading: Class notes & F1000Research 2016, 5:1408 Assignment: Execute code from F1000Research paper Homework 6: Select Topic for Final Project (due: 7/6)	Kimberly Siegmund
Wed 07/07/21 01:00p - 04:50p	Supervised Learning Learning Objectives: 1. Describe the bias-variance tradeoff for classification methods 2. Name two methods for evaluating a classification model Reading: HHGF Ch. 9 & Class Material on Blackboard Homework 6: Prepare in-class presentation (due 7/14)	Kimberly Siegmund
Wed 07/14/21 01:00p - 04:50p	Student Presentations Learning Objective: Present a topic on high-dimensional data to a general audience of scientists	
Wed 07/21/21 01:00p - 04:50p	Network Analysis Learning Objectives: 1. Describe hub genes, 2. Explain why hub genes are interesting biologically 3. Describe one approach to identify hub genes Reading: Class Material on Blackboard References: Zhang and Horvath (2005) https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/	Kimberly Siegmund
Wed 07/28/21 01:00p - 04:50p	Data Integration Learning Objective: 1. List three methods for integrating data from separate platforms Reading: Class Material on Blackboard References: Ritchie et al (2015) Nat Rev Genet; Witten et al. (2009) SAGMB; Shen et al. (2013) Ann. of Applied Stat	Kimberly Siegmund

Wed 08/04/21 01:00p - 04:50p	Data Integration with TCGA data Learning Objective: Conduct integrative analysis of TCGA colon cancer data In class activity	Kimberly Siegmund

Statement on Academic Conduct and Support Systems

Academic Conduct:
Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Part B, Section 11, "Behavior Violating University Standards" policy.usc.edu/scampus-part-b. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Support Systems:
Student Counseling Services (SCS) – (213) 740-7711 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention. engemannshc.usc.edu/counseling

National Suicide Prevention Lifeline – 1 (800) 273-8255
Provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week. www.suicidepreventionlifeline.org

Relationship and Sexual Violence Prevention Services (RSVP) – (213) 740-4900 – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender-based harm. engemannshc.usc.edu/rsvp

Sexual Assault Resource Center
For more information about how to get help or help a survivor, rights, reporting options, and additional resources, visit the website: sarc.usc.edu

Office of Equity and Diversity (OED)/Title IX Compliance – (213) 740-5086
Works with faculty, staff, visitors, applicants, and students around issues of protected class. equity.usc.edu

Bias Assessment Response and Support
Incidents of bias, hate crimes and microaggressions need to be reported allowing for appropriate investigation and response. studentaffairs.usc.edu/bias-assessment-response-support

The Office of Disability Services and Programs
Provides certification for students with disabilities and helps arrange relevant accommodations. dsp.usc.edu

Student Support and Advocacy – (213) 821-4710
Assists students and families in resolving complex issues adversely affecting their success as a student EX: personal, financial, and academic. studentaffairs.usc.edu/ssa

Diversity at USC
Information on events, programs and training, the Diversity Task Force (including representatives for each school), chronology, participation, and various resources for students. diversity.usc.edu

USC Emergency Information
Provides safety and other updates, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible. emergency.usc.edu

USC Department of Public Safety – UPC: (213) 740-4321 – HSC: (323) 442-1000 – 24-hour emergency or to report a crime.
Provides overall safety to USC community. dps.usc.edu