STAT 718: High-Dimensional Data

Spring 2021

Instructor:	Ray Bai	**Time:**	MWF 1:10 PM – 2:00 PM
Email:	RBAI@mailbox.sc.edu	**Place:**	TBA

Course Page:

https://blackboard.sc.edu/ (Check regularly for announcements and homework assignments)

Office Hours: By appointment. I am also very accessible by e-mail and will typically reply to e-mails within one business day of receiving them.

Course Description: Modern data mining problems often deal with large and complex data sets, otherwise known as "big data." Typically, these models contain a large number of parameters, the dimensionality of which may exceed the sample size. In this course, we will study these models, including high-dimensional regression and classification, grouped regression and nonparametric additive models, and graphical models. Time permitting, additional topics such as the matrix completion problem and sparse principal component analysis will also be discussed. The tentative schedule of topics is given on the last page of this syllabus.

We will examine key ideas from optimization theory (e.g. convexity and duality) and standard optimization algorithms, including gradient descent and majorization-minimization algorithms. Practical issues such as the bias-variance trade-off, evaluation of predictive accuracy, and the selection of tuning parameters are also discussed. Statistical computing and programming are key components of the course.

Prerequisites: STAT 512 and MATH 544 or equivalent. Students should also be comfortable with a programming language such as R, Python, MATLAB, or C++.

Learning Outcomes:

1. Demonstrate the ability to analyze high-dimensional data and compare the performance of different methods.
2. Understand the methodology and theory underlying high-dimensional data analysis.
3. Be able to implement algorithms for analyzing high-dimensional data.
4. Be able to communicate effectively through writing scientific reports and making/delivering presentations.

Course Materials: We will use typed handouts prepared by the instructor. Parts of these lecture notes are *not* complete and will be filled in during lecture. Thus, it is in your best interest to attend every lecture.

No textbook is required. However, the following books may be useful as supplementary references:

- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.
- Bühlman, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Computing: It is recommended that students use R for computing, although other programming languages like MATLAB or Python may also be used. However, please note that some homework problems may reference R packages, and if you want to use another language, you will need to find the analogous package in that other language on your own.

Attendance Policy: Attendance is expected but is not part of the grade.

Homework: There will be approximately five or six homework sets. Students will work in small groups, and each group will submit a single, typed report for each homework assignment. The homework will consist of both conceptual/theoretical exercises and questions that involve programming and data analysis. No late homework will be accepted.

All homework reports should be typed in LaTeX, including answers to math exercises and data analysis portions that may include figures, tables, etc. Code should be put in the appendix of the report. The first few weeks of the class, groups should scour the Internet for a homework template to use. It is expected that students will be able to learn Latex on their own by following existing templates and using the Internet.

Exams: There will be a take-home final. Students will work in small groups and each group will submit a single report for the final. The report should be typed and in the same format as the homework reports.

Final Project: Students will work in small groups to read one or two papers on a topic of their choosing, prepare a 15-20 minute presentation, and write a short report in the style of a journal article: abstract, introduction, method, simulations and data analysis, and a bibliography. The last week of the semester will be devoted to project presentations. Some potential examples of projects include:

- joint estimation of graphical models
- fused lasso for time series/temporal data
- multiclass sparse discriminant analysis
- regularized calibrated estimation in causal inference
- regularized Cox regression.

Students are encouraged to pursue projects that are relevant to their current research or their research interests. Projects must be approved in advance by the instructor, and no two groups may do the same topic for their project. If you have an idea of what you want to do for your project, please "claim" it early. Detailed instructions for the presentation and the report will be given at a later date.

Grading Policy: Homework (50%), take-home final (30%), group project and presentation (20%).

The tentative grading scale is as follows: 90-100 for an A, 85-89 for an A-, 80-84 for a B+, 70-79 for a B, 60-69 for a B-, 0-59 for a C.

Honor Code: See the Carolinian Creed in the *Carolina Community: Student Handbook and Policy Guide*. The *minimum* punishment for violations of the USC Honor Code is a grade of zero for the work in question. In accordance with university policy, there may be other punishments including an automatic F in the class and/or expulsion from the university.

Accommodation: If you need special accommodations for examinations or any other aspects of the course, please contact me before or during the first week of the semester. Note that reasonable accommodations are available for students with a documented disability. If you have a disability and may need accommodations to fully participate in this class, contact the Office of Student Disability Services by phone (803-777-6142), email sasds@mailbox.sc.edu, or stop by LeConte College Room 112A. All accommodations must be approved through the Office of Student Disability Services.

Tentative Schedule of Topics:

- **Weeks 1-2**: introduction, methods for assessing predictive accuracy, selection accuracy, uncertainty quantification, and tuning parameter selection
- **Weeks 3-4**: bias-variance trade-off, ordinary least squares (OLS), ridge regression, lasso regression, elastic net
- **Weeks 5-7**: overview of convex optimization, gradient descent, coordinate descent, majorization-minimization algorithms, stochastic approximation methods
- **Weeks 8-9**: generalized linear models and classification
- **Week 10**: grouped regression and sparse nonparametric additive models
- **Week 11**: Gaussian graphical models
- **Week 12**: matrix decompositions and completion
- **Week 13**: sparse multivariate methods, principal component analysis
- **Week 14**: group project presentations

We will definitely cover optimization, regression, and classification. If we end up spending a lot of extra time on these earlier topics, the topics tentatively planned for the last few weeks of the semester will be turned into potential group project presentation ideas.

If we need extra time for the group project presentations, we will use the university scheduled final exam date and time for this class to complete the remainder of the presentations, in lieu of a timed final. **All students are expected to attend their classmates' presentations**. The slides and reports will be shared in a central repository (e.g. on Dropbox or Google Drive) so that everyone can access them.