Mel and Enid Zuckerman College of Public Health  
University of Arizona

SYLLABUS  
BIOS 648 ANALYSIS OF HIGH DIMENSIONAL DATA  
Spring Semester 2020

Time: Tuesday and Thursday 11:00am – 12:15pm

Location: Drachman Hall Room A112

Instructor: Chengcheng Hu, Ph.D. 
Office: Drachman A228 
Phone: 520-626-9308 
E-mail: hucc@email.arizona.edu

Instructor Availability: After class or by appointment

Teaching Assistant: TBD

TA Office Hours: TBD

Catalog Description: This course deals with the analysis of high dimensional data. It will cover multiple comparison, clustering and classification of high dimensional data, and regression methods involving high dimensional variables. Students will also learn the corresponding computer software.

Course Prerequisites: One year of college calculus, a course in matrix algebra, BIOS 576A and BIOS 576B, or permission of instructor. Please check with me if you have doubts about how well your background prepares you for this course.

Course Learning Outcomes: At the end of the course, you should be able to:
1. Identify studies and data sets to which high dimensional methods should be applied.
2. Identify appropriate methods to analyze high dimensional data.
3. Use the statistical software package R to implement high dimensional methods.
4. Interpret and critique medical and scientific journal articles which involve high dimensional data.
Course Objectives and Expected Learning Outcomes:

- **Course Objectives:** During this course students will learn modern methods for multiple comparisons, unsupervised clustering methods for high dimensional data, and classification and prediction models involving high dimensional information. Students will also learn the corresponding computer software packages.

- **Learning Outcomes:** Upon completion of this course students will be able to:
  1. Identify and apply appropriate statistical tools for high dimensional data, including inferential and predictive methodologies, for answering a particular research question.
  2. Recognize strengths and weaknesses of proposed statistical approaches.
  3. Communicate understanding of the assumptions necessary for a given statistical procedure as well as the ability to determine if the assumptions are met for a given study design or data set.
  4. Suggest preferred methodological alternatives to commonly used statistical methods when assumptions are not met.
  5. Manage data to handle a variety of practical problems in data format and structure.
  6. Apply appropriate statistical software in data analysis involving high dimensional outcomes.
  7. Demonstrate advanced competencies in areas of professional expertise and scholarship enabling advancement to further postgraduate study in statistics or biostatistics.
  8. Demonstrate the ability to identify, articulate and implement sound methodological and computational strategies for addressing scientific questions.
  9. Demonstrate the ability to communicate effectively in writing reports.
  10. Interpret and critique medical and scientific journal articles which involve high dimensional data.

MEZCOPH Program Competencies Being Addressed:

- Biostatistics MPH competencies:
  1. Ability to identify appropriate statistical tools to address specific scientific questions.
  2. Demonstrate skills in data management to handle a variety of practical problems in data format and structure.
  3. Demonstrate advanced working skills in application of computer systems and appropriate statistical software.

- Biostatistics MS competencies:
  1. Recognize strengths and weaknesses of proposed statistical approaches, including alternative designs, data sources, and analytical methods.
  2. Suggest preferred methodological alternatives to commonly used statistical methods when assumptions are not met.
3. Demonstrate advanced competencies in areas of professional expertise and scholarship enabling advancement to further postgraduate study in statistics or biostatistics.

- **Biostatistics PhD competencies:**
  1. Apply descriptive and inferential methodologies according to the type of study design for answering a particular research question.
  2. Communicate understanding of the assumptions necessary for a given statistical procedure as well as the ability to determine if the assumptions are met for a given study design or data set.
  3. Demonstrate the ability to identify, articulate and implement sound study design, methodological and computational strategies for addressing scientific questions.
  4. Demonstrate the ability to communicate effectively in writing reports, giving oral presentations, and teaching basic statistical material in a formal classroom or seminar setting.

**Course Notes:** Notes will be posted at the D2L website (see below) before lecture.

**Course Website:** A webpage is created for this class using the Desire 2 Learn (D2L) interface. This course website will contain the syllabus, class notes, sample programs and outputs, and datasets (used in lectures and homework assignments). Class announcements will also be posted on this site, so it is a good idea to check the site regularly to stay current.

To access the course website, login at: [http://d2l.arizona.edu/](http://d2l.arizona.edu/)
- Click ‘UA NetID Login’;
- Enter your NetID and password, and complete the NetID+ requirement, if needed, as you would to access your UA email account;
- Under ‘My Courses’, click on the link to BIOS 647.

For further information on how to use the D2L interface, go to: [http://help.d2l.arizona.edu](http://help.d2l.arizona.edu)

**Required Texts or Readings:** Lecture slides and notes

**Recommended Texts/Readings:** In addition to lecture notes and selected papers posted on the course website, the following textbooks are recommended:


- **D** Multiple Testing Procedures with Applications to Genomics, Sandrine Dudoit and Mark J. van der Laan, Springer 2008.
Course Requirements: In addition to reading the texts and attending lectures, the primary course requirements consist of five homework assignments and one final report. Homework must be turned in by 5pm on the due date, which is indicated on each assignment. Electronic submission to the d2l website is required (click the Assignments tab right below the course banner at the top of the page then select the appropriate folder). For the final report, you will be asked to analyze a dataset and write a report in five and half weeks.

Grading: Homework and the final report contribute to your final grade as follows:

- Homework: 60% (each assignment is weighted the same, even if length differs)
- Final Report: 40%

Late submissions will receive no or partial credit. Final grades are based on the following point system:

- A = 90-100%
- B = 80-89%
- C = 70-79%
- D = 60-69%
- E = 59% or less

Requests for incomplete (I) and withdrawal (W) must be made in accordance with University policies. University policy regarding grades and grading systems is available at: http://catalog.arizona.edu/policy-type/grade-policies

Required examination: The final report is due by 5pm on Tuesday 5/12. Below is the link to the final exam regulations:
http://www.registrar.arizona.edu/staff/courses/final-exams?audience=staff&cat1=10

R and Computer Lab: R is a free statistical package available at http://www.r-project.org, and can be accessed at the Drachman Hall Computing Lab (A319) on weekdays from 8-5. A few user-contributed packages for R are also needed. An introduction to R will be given in Lecture 2 and 3.
Description of each Assessment and Competencies Covered by the Assessment

HW #1: an exercise with R (covering Biostatistics MPH competency 3);

HW #2, #3, #4, #5: statistical analysis to answer scientific questions (covering Biostatistics MPH competencies 1-3, MS competencies 1-3, and PhD competencies 1-3).

Final project: you will be asked to use one or more methods covered in this course to analyze a dataset with high-dimensional information. Two options will be provided by the instructor, or you can work on your own data upon instructor’s approval (covering Biostatistics MPH competencies 1-3, MS competencies 1-3, and PhD competencies 1-4).

Communications: You are responsible for reading emails sent to your UA account from your instructor and the announcements that are placed on the course web site. Information about readings, news events, your grades, assignments and other course related topics will be communicated to you with these electronic methods. The official policy can be found at: [https://www.registrar.arizona.edu/personal-information/official-student-email-policy-use-email-official-correspondence-students](https://www.registrar.arizona.edu/personal-information/official-student-email-policy-use-email-official-correspondence-students)

UA Smoking and Tobacco Policy:
The purpose of this Policy is to establish the University of Arizona’s (University) commitment to protect the health of University faculty, staff, students, and visitors on campuses and in its vehicles. The official policy can be found at: [http://policy.arizona.edu/ethics-and-conduct/smoking-and-tobacco-policy](http://policy.arizona.edu/ethics-and-conduct/smoking-and-tobacco-policy)

University Course Policies: (please see the following URL):
[https://academicaffairs.arizona.edu/syllabus-policies](https://academicaffairs.arizona.edu/syllabus-policies)

Plagiarism: What counts as plagiarism?
- Copying and pasting information from a web site or another source, and then revising it so that it sounds like your original idea.
- Doing an assignment/essay/take home test with a friend and then handing in separate assignments that contain the same ideas, language, phrases, etc.
- Quoting a passage without quotation marks or citations, so that it looks like your own.
- Paraphrasing a passage without citing it, so that it looks like your own.
- Hiring another person to do your work for you, or purchasing a paper through any of the on- or off-line sources.

Course Schedule: Any changes to the following schedule will be announced in lecture. You are responsible for obtaining information on any changes, even if you miss class.

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Reading</th>
<th>Homework Assignment</th>
<th>Homework Due</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thu, 1/16</td>
<td>Logistics and introduction</td>
<td>H Ch 1, B 1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 1/21</td>
<td>Introduction to R</td>
<td>V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 1/23</td>
<td></td>
<td></td>
<td>#1</td>
<td>1/30</td>
</tr>
<tr>
<td>Date</td>
<td>Topic</td>
<td>Reading</td>
<td>Homework Assignment</td>
<td>Homework Due</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------------------------------------</td>
<td>--------------------------------</td>
<td>---------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Tue, 1/28</td>
<td>Multiple comparison methods</td>
<td>D 3.2, 3.4, selected papers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 1/30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 2/4</td>
<td>K-means clustering</td>
<td>H 13.2.1, 14.3.6-8 B 9.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 2/6</td>
<td>Hierarchical clustering</td>
<td>H 14.3.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 2/13</td>
<td>Linear discriminant analysis</td>
<td>H 4.3, B 4.1</td>
<td>#3</td>
<td>2/20</td>
</tr>
<tr>
<td>Thu, 2/18</td>
<td>Classification trees</td>
<td>Z Ch 1, 2, 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 2/20</td>
<td></td>
<td>H 9.2.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 2/25</td>
<td>Variable selection review: forward, backward,</td>
<td>H 2.9, 3.3, 7.1-7.7 B 1.4, 4.4.1</td>
<td></td>
<td>#4</td>
</tr>
<tr>
<td>Thu, 2/27</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 3/3</td>
<td>Final project (abstract due 3/26; final report</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 3/10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 3/12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 3/17</td>
<td>Regression trees</td>
<td>Z Ch 9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 3/19</td>
<td></td>
<td>H 9.2.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 3/24</td>
<td>Random forest</td>
<td>H Ch 15</td>
<td>#5</td>
<td>3/31</td>
</tr>
<tr>
<td>Thu, 3/26</td>
<td>Kernel smoothing</td>
<td>H 6.1, 6.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 3/31</td>
<td>Spline smoother and additive models</td>
<td>H 9.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 4/2</td>
<td>Projection pursuit regression</td>
<td>H 11.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 4/7</td>
<td>Ridge regression</td>
<td>H 3.4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 4/9</td>
<td>Lasso and elastic net</td>
<td>H 3.4.2, 3.4.3, 3.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 4/14</td>
<td>Principal component regression</td>
<td>H 3.5.1, 14.5.1 B 12.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 4/16</td>
<td>Partial least squares regression</td>
<td>H 3.5.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 4/21</td>
<td>Support vector machines</td>
<td>H 12.1-12.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 4/23</td>
<td></td>
<td>B 7.1.3, 7.1.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 4/28</td>
<td>Neural networks</td>
<td>H 11.1, 11.3-11.8 B 5.1-5.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 4/30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 5/5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, 5/7</td>
<td>Reading Day, No Class</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, 5/12</td>
<td>Final Report Due at 5pm</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>