Course Policies and Syllabus

Updated January 13

Office e-mail Telephone Office Hours
Instructor: Pantelis Vlachos BH 232K vlachos@stat.cmu.edu x8-1883 TBA
TA: Alex Rojas BH229C arojas@stat.cmu.edu x8-7826 Fri. 2:00pm

Lectures:
MWF 11:30-12:20am — PH A18B

WWW: http://www.stat.cmu.edu/~vlachos/


Recommended Texts


Course Description
This course is divided in two parts. The first part attempts a fair coverage of Bayesian statistics, whereas the second part is meant as a fast survey of modern statistical techniques. What ties together the two parts is the emphasis on the computational and applied aspects of the various models and techniques we will study. If you are not a student in statistics and are interested in only one part of the course, please see me.

Bayesian Statistics. If you know Bayes’ theorem you know (almost) everything there is to know about Bayesian statistics. Why then a whole course in Bayesian methods? For a long time, Bayesian statistics was nothing more than an intellectual exercise, not good for data analysis and shunned by “real” statisticians. It wasn’t until the 1950s that the Bayesian approach started to gain some respect, thanks to Jeffreys, Savage, DeFinetti, and others. Even after it became more popular, the Bayesian approach was severely limited because of its computational infeasibility. Whereas frequentist estimation relies on maximization of the likelihood, Bayesian estimation depends on the posterior distribution, which can be derived only through integration, a more challenging operation. With the advent of numerical integration and Markov Chain Monte Carlo (MCMC) techniques, the posterior distribution can be explored via simulations and very complex problems can be tackled.
Therefore, this is an exciting time to study Bayesian statistics, both for the momentous advances and the research possibilities that are still open. The focus of this part of the course will be on parametric methods, and even though the Bayesian approach can be embedded in a decision-theoretic framework, the emphasis of the course is decidedly applied (but still making use of mathematical statistics). Topics will include:

- Philosophical considerations regarding the Bayesian and frequentist approaches
- Methods for deriving prior distributions, including noninformative priors
- Prior-posterior analysis, including Empirical Bayes methods, asymptotics, and
- Simulation methods, including Monte Carlo integration and MCMC
- Bayesian inference, including point estimation and credible intervals
- Diagnostics and model checking, including sensitivity analysis, Bayes factors and other model selection techniques
- Predictions

Computational Methods. Thanks to computational advances, statisticians have been able to venture outside the realm of linear models and other restrictive assumptions that were necessary 30 years ago. Possibilities include parametric and non-parametric models, simple, multiple and multivariate techniques, continuous (regression) and discrete (classification) problems, and more. However, it will be impossible to cover all such techniques in depth. Therefore, the goal for this part of the course is to get a flavor of the most common techniques and models, how, why, and when they work, so that you are conversant in the major directions of modern statistics. The emphasis is again applied and the focus will be computational data analysis (learning to implement procedures in the software and interpret related plots and results). These ideas will be grounded in the relevant theorems, but their proof is deferred to more advanced classes. I am also interested in tailoring the course to your interests, so let me know if there are any models/procedures you are committed to learn. Topics can include:

- The Bootstrap and other subsampling methods
- Modern regression techniques (e.g., splines and related models, wavelets, neural networks, kernel smoothers such as loess)
- Classification methods (e.g., logistic regression-type methods, discriminant analysis, trees, nearest neighbor methods)
- Multivariate methods for dimensionality reduction, such as principal components and factor analysis
- Others?
Objectives
At the highest level, the objectives for this course are:

1. Understand the research question(s)
2. Select and build appropriate models for data to answer the question(s)
3. Apply the relevant inferential techniques to extract evidence from the data
4. Write computer code to carry out required computations and simulations
5. Interpret the results
6. Evaluate the adequacy of the model
7. Communicate results in writing to stakeholders

These objectives in turn imply a non-superficial understanding of all the different techniques, how they work and their comparative advantages and disadvantages.

Prerequisites
I assume you are familiar with mathematical statistics at the level of 36-705 (Intermediate Statistics) and linear regression at the level of 36-707 (Linear Regression). This course also involves a fair amount of computing. While you are free to use any package or language you are comfortable with (Matlab, C, C++, BUGS or others), I will use S-plus in my handouts or in the code I will provide for you. For this reason, I will also assume familiarity with S-plus at the level of 36-711 (Statistical Computing).

Required Work.

Lectures. The nature of the course implies that lectures will often weave together multiple sources, therefore attendance is expected at all classes. You will be responsible for everything covered in class, so participation is in your best interest.

Homework. Five homework assignments will be handed out roughly every other Wednesday. Homework should be dropped off before lecture.

Projects. There will be two projects, one at the end of the Bayesian module and one at the end of the semester on computational methods. Statistics students can think of these projects as a preparation for the Data Analysis Exam in May. The first project will be structured in two parts – you will submit a draft first, then you will have a chance to revise the draft according to the feedback I will give you. The second project will only have a final version, for which you will be required to submit a technical report and an executive summary – very much like the data analysis exam. More details to follow in the project descriptions. The projects are due according to the following schedule:

March 19 Project 1 due
April 30 Project 2 due
Final Exam: There is no final exam.

Policy on Collaboration. Collaboration on homework among students is expected and even encouraged as an additional learning opportunity. However, the write-up of each problem has to be your own. My suggestion is to work on the assignment first and then ask around if you are stuck or to compare methods. Note: you are expected to work on your own for both projects. The purpose for this policy is two-fold: a) as a preparation for the data analysis exam; b) more importantly, the projects are very open-ended, so you can be very creative with the methods you will choose to use (which won’t necessarily be the same methods other students will use).

Cheating and Plagiarism: Cheating and/or plagiarism will not be tolerated. Please see the CMU Student Handbook, p. 7-8, for definitions of cheating and plagiarism, and the severe consequences of such behaviors or read the University Policy on Academic Integrity online at http://www.studentaffairs.cmu.edu/acad_integ/acad_index.htm

Writing: Statisticians never work alone: they always analyze data for other professionals, or are involved in article and grant writing (not to mention thesis writing!). Clear communication is crucial to interdisciplinarity and fund raising. For these reasons most homework assignments will include the writing of a report. You are expected to tie together all the evidence from your analysis in a coherent fashion, appropriate style and good grammar. Your grade will be affected by the clarity of your writing.

Final Grades. Final course grades will be based on the following breakdown:

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework Assignments</td>
<td>60%</td>
</tr>
<tr>
<td>Projects</td>
<td>40%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

Statistical Software

- I will assume everyone in the course has an account on the Statistics Department workstations. If you are not a member or guest of the department you should be able to get by with an Andrew account, but you will be somewhat on your own.

- BUGS (Bayesian inference Using Gibbs Sampling) is an S-Plus-like system for posterior simulation of hierarchical models. Information and documentation on BUGS, including download and install information is available at http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml.

- Most of the data sets used in the Carlin and Louis book are available online at http://www.biostat.umn.edu/~brad/data.html.

- All the handouts, assignments and solutions for this course will be stored on the class Web site, http://www.stat.cmu.edu/~vlachos under the section Downloads.