Syllabus :: Advanced Biological Statistics II: Bio 610, Winter 2018

Peter Ralph

9 January 2018

- **Class:** 10-12, Tuesday/Thursday.
- **Office Hours:** Wednesday 12-1 and Friday 11-12 in 282 Onyx
- **Instructor:** Peter Ralph (plr@uoregon.edu)
- **Office:** 6 Deady Hall and/or 282 Onyx (email to arrange)

This course is an introduction to advanced methods in modern statistics, with an emphasis on applied biological data analysis. After taking this course, you will be able to invent, implement, and apply probabilistic models that describe the signal and noise inherent in real datasets. We will cover modern statistical techniques important for dealing with large, noisy datasets of many variables, including regularization, dimensionality reduction, and deconvolution. Mostly, we will work in the Bayesian hierarchical modeling framework, using Stan as interfaced through R. We will use almost no mathematical analysis (equations) for understanding and implementing these methods, relying instead on simulation and the properties of the various kinds of randomness.

Textbook(s) and Software:

- The main textbook is Doing Bayesian Data Analysis : a tutorial with R, JAGS, and Stan, 2nd edition, by John Kruschke; PDF available at this link, at least on campus (see uo library entry).

- We will also refer to Experimental Design and Data Analysis for Biologists, Quinn and Keough, 2002.

- The Stan User's Guide will also be often useful.

Note: both JAGS and Stan do roughly the same job, and are very similar. Although Kruschke spends more time with JAGS (for historical reasons), we will only use Stan - so, you can skip the JAGS stuff, although you should make it up with the Stan manual.

- We will make constant use of R.

- You will also need Stan and it's R interface, rstan.

- To write reports, you need to be able to compile Rmarkdown documents, which uses knitr and pandoc (note that Rstudio includes it's own version of pandoc).

- In class, I will use Rstudio for demonstrations, although you don't have to use it.

To test your setup, you should be able to compile this Rmarkdown document into something that looks like this html file (but: the formatting might differ).

Assessment, and grading

The main course work will be a series of reports, some *short* and others *longer*. Short reports will be less than one page of text (more, with code), are mostly based on simulated data, and are designed to demonstrate you understand and can implement an important topic from that week. These will occur in weeks without a longer report due. The three longer reports will contain more narrative, describing and analyzing a real dataset that I will provide. All reports will be turned in as self-contained Rmarkdown documents, that I should be able to run myself.

I will assign grades to each of you, independently, based on how well your course work demonstrates completion of course goals.
I will not assign grades for attendance, but class participation and group work will be important: please come. The relative weights of each category will be:

- *short reports:* six, worth together 40%
- *longer reports:* three, worth 20% each

Preliminary drafts of longer reports will be handed in for feedback partway through.

Accessibility

I take seriously my responsibility to create inclusive learning environments. Please notify me if there are aspects of the instruction or design of this course that result in barriers to your participation! You are also encouraged to contact the Accessible Education Center in 164 Oregon Hall at 541-346-1155 or uoaec@uoregon.edu.

Course schedule

The schedule is available here.

.

Schedule :: Advanced Biological Statistics II: Bio 610, Winter 2018

Peter Ralph

9 January 2018

Course schedule

The tentative schedule (subject to adjustment, especially towards the end) is (**K** referes to Kruschke):

Week 1 (*1/9*) (slides) Recap of probability and likelihood; central limit theorem ($\sqrt{n}$); Bayes' rule. The beta-binomial distribution: putting a prior on the probability of success. **(K ch. 4, 5, 6)**

    Short report: (due 1/16) MLE for beta-binomial

Week 2 (*1/16*) Introduction to MCMC and Stan for sampling from posterior distributions, hierarchical models for binary responses, shrinkage. **(K ch. 7, 9 and Intro to Stan)**

Week 3 (*1/23*) Assessing power, model choice, and using simulation: looking more at shrinkage, posterior predictive sampling, Bayes factors. Logistic regression: robustly, including categorical factors. **(K ch 13 and 21, with a bit of chapters 10-12)**

Week 4 (*1/30*) Count data: using Poisson regression and hierarchical modeling to fit overdispersion. Model selection by crossvalidation. **(K ch 24)**

Week 5 (*2/6*) Continuous ("metric") data: groupwise means, univariate regression, robust regression by adjusting the noise distribution, multivariate regression and variable selection; the Horseshoe. **(K ch 16, 17, 18)**

Week 6 (*2/13*) Friends of ANOVA: sources of variance. An in-depth applied example, cumulative. **(K ch 19, 20)**

Week 7 (*2/20*) Dimension reduction: overview of the goal and taxonomy of various approaches (PCA, CCA, t-sne, etc). Deconvolution. Regularization as a penalty and as a (Bayesian) prior. **(Reference to Quinn & Keough)**

Week 8 *(2/27)* Clustering and categorization: softmax regression; applied example. **(K ch 22)**

Week 9 *(3/6)* Time series: modeling local dependency, smoothing. Conditional independence.

Week 10 *(3/13)* Spatial and network covariance: sharing power between related locations.