STAT–S426/626 Bayesian Theory & Data Analysis
Course Project

1 Introduction

The goal of the course project is for you to collaborate with another student to

- Apply Bayesian methods to a new dataset to answer a compelling scientific question in your research area; or
- Explore a new topic in Bayesian theory or data analysis.

2 Guideline

1. You are strongly encouraged to work in pairs on the project. If more than two persons want to work together or one student prefers to work alone, please notify me in advance. **There cannot be more than 3 people in each group.**

2. You are welcome to come up with your own idea, but you can also pick one from the list of ideas I provide. If more than one person/group wants to do the same idea from below, I'll talk to you to see which has a more suitable background, etc. I might also want to see whether there is a possibility that one idea could produce two projects, if two or more people/groups come up with somewhat different aspects of the idea to focus on.

3. Each person/group please meet with me no later than **Nov 3** to decide a project and complete a project proposal. In the proposal, please include:

   - Describe the existing model briefly, if it is not one that we have discussed in class. You may skip this step if the model is one that we have discussed in class.
   - State and describe the objective of your investigation. What is (are) the question(s) that you are trying to answer?
   - Sketch the mathematical or technical approach that you will use to address the stated objectives. Be specific about how you will modify the model (if at all), and what sorts of analyses you will conduct.
   - Sketch the type of results that you anticipate.

   The project proposal should be just one page and will count for 20% of your project grade.

4. Most projects should involve about the equivalent of two homeworks' worth of work. The report should contain:

   - A brief introduction.
   - Methods and results.
   - A discussion.

   You must properly cite all materials that you use for the project. You must also acknowledge anyone (apart from instructor) who helps you with the project. The final result should be no more than 10 pages including plots. The report is due on **Dec 8th** and counts for 80% of your project grade.

5. For the course project, it is enough to do a preliminary assessment – seeing whether the basic idea is promising. It's OK if it turns out to not be promising, as long as you did a good job demonstrating that.

3 Project Ideas

1. (UG) Probit regression. Read Hoff. pg 209-214 and implement the example of "Educational attainment". Study the codes from the author and add comments to every step. Try some different priors if possible.

2. Gaussian copula model. Read Hoff. pg 217-222 and implement the example of "Social mobility data".

3. BDA3: Noninformative prior distributions (page 61-65). Find noninformative prior distributions for some sampling models and compare with conjugate priors.

4. BDA3. Take one of the latest chapters in Gelman, Carlin, Stern Dunson, Vehtari and Rubin (2013) Bayesian Data Analysis, Chapman Hall, 3rd edition. We suggest one of section 19-22. Take one example in the section (preferably with data–all the datasets are on the web-site for the book). Explain clearly what the example is and detail the procedures used. Then do something: change the assumptions; vary the prior; where the authors say by the usual arguments put in the details; try to prove something; find something wrong with the analysis; or go get some more data... (You do not have to do all of this, of course: these are just suggestions.

5. MCMC for mixture models without component indicators. The MCMC methods I presented for mixtures include in the state indicators for each observation of which mixture component it comes from, which are updates during the MCMC run. This is what is commonly done, perhaps because it sometimes leads to nice simple Gibbs sampling algorithms.

   However, for finite mixtures, one can sum the probabilities of a data point coming from each component, and use the likelihood found this way to compute the posterior probability (up to an unknown factor) for a state consisting of only the mixing proportions and the parameters of each mixture. One could then use Metropolis, slice sampling, or other MCMC updates to sample for this state.

   The project would investigate whether this works better or worse than standard methods using component indicators. There are many possible variations, of course, so this isn't a straightforward assessment. One could also try to think of a way of handling infinite mixtures this way, or at least handling mixtures with a large number of components efficiently when many components are not actually used.

6. Behaviour of finite or infinite mixture models as the number of variables increases. What happens as the number of variables goes to infinity? Knowing this could be informative about performance on lots of bioinformatics problems with tens of thousands of genes.

   The answer will presumably depend on what the increasing number of variables are like. One could look at

   a) All variables are actually independent and have distributions in the family used for the mixtures components, so one mixture component is all that's needed.

   b) Some fixed number of variables have distributions that need more than one component to model, or have dependencies that need to be modeled using more than one component. An unlimited number of other variables are independent and have distributions that can be modeled by one component, as for (a).

c) All variables are dependent. There are many ways they could be dependent, of course...

You might investigate these questions entirely theoretically, or by doing numerical experiments, or both.

7. Applied Bayesian Analysis. The papers below present examples of Bayesian statistics in action. Choose one and study it carefully. Provide a clear summary of one of the models they consider; discuss the assumptions; how does the Bayesian approach translates in conclusions that might be different from the ones obtained within a classical framework? what do you think are the limitations of the analysis? Extend the work presented in one direction, either using new data, new model or priors, new computational methods, or studying some properties of the estimates that is not analyzed in the original paper.

   - James Albert and Siddhartha Chib (1993) Bayesian Analysis of Binary and Polychotomous Response Data, Journal of the American Statistical Association 88: 669679.
   - Gelman and King (1990) Estimating the Electoral Consequences of Legislative Redistrict- ing, JASA 85: 274282.
   - Efron and Thisted (1976) Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know? Biometrika 63:43547.

8. Disputed Authorship. For this project, you will recreate a classic Bayesian analysis of disputed authorship. Read the paper Inference in an Authorship Problem by Mosteller & Wallace (Journal of the American Statistical Association, 1963). Choose two contemporary authors who write articles/papers/blog posts for roughly similar audiences (e.g., Larry Summers and Paul Krugman). Their writing styles should not be so obviously different that distinguishing between them is trivial. Collect 50 works from each author and sample 5 from each to be disputed papers. Finally, apply the Bayesian methodology of Mosteller & Wallace to infer authorship of the disputed papers.

   Suggestions:

   - As in the original analysis, choose non-contextual words on the basis of intuition and/or features of the data.
   - The original paper was limited by the computational tools and resources of the time. For example, the authors choose to ignore uncertainty in the estimated rates M and H. In- stead, you might want to simulate from the posterior to evaluate the posterior probability of authorship for the disputed papers. You may also find that selecting a very small set of words, as the authors did, is not necessary.
   - Feel free to modify parts the original analysis, e.g., if you think a different prior distribution would be appropriate.
   - If you use R, the package tm might be useful for parsing and analyzing text files.

9. Take your own Bayesian peek at Feller vol. 1 Find a copy of the article A Bayesian peek at Feller volume 1, by Diaconis and Holmes. Take one of the examples and understand it in your own language: when the authors say standard computations show or it is easy to see... put in some detail. Then find a copy of W. Feller, An introduction to probability and its applications vol. 1 (3rd ed.). Take any other topic (eg. a homework problem or a section) and make a Bayesian version.

10. Big Bayesian Networks. Scaling algorithms for learning Bayesian networks
    Nice tutorial on David Heckermans home page http://www.research.microsoft.com/ heckerman/