STA 705 Project

General Instructions

Each project requires you to understand and implement at least one advanced statistical computing algorithm in order to fit a model to data. Three things will be required from each project:

Written report

This is to be a short report with

- no more than 8 pages double spaced with 12 point font and one inch margins excluding references and figures/tables
- no more than 4 combined figures and tables

that describes (i) the problem, (ii) the computational approach to solve the problem, and (iii) the results and conclusions/discussion. You should write the report as if it were a short article for *Journal of the American Statistical Association* applications and case studies section.[1] Your report should follow all of the protocols that a paper actually submitted to the journal should, e.g. you need to cite and attribute previous work (and *not* plaigarize). The report should effectively be a 'mini' scientific paper. You will be graded on the overall quality of this report. No matter how impressive your research is, you will not be able to get it published if you are unable to effectively communicate this to others.

Presentation

During the last week of class, each group will present a 20 minute talk to the class explaining the problem as you would at a conference such as the Joint Statistical Meetings (JSM). Your talk should be written in LaTeX using the `beamer` package. After your talk there will be 5 minutes for questions.[2] You should ensure the timing of your talk as I will be strict in keeping to time (as we will have two talks in each class).

R package

Your written report should be supplemented with a working R package that others can use to repeat your results and look at other data/models (within reason). The R code in this package must be your

[1] If you are unfamiliar with the journal and/or this section you need to read some of the recent articles published in this section. You can also read the information for authors on the JASA website to help understand the role of this section.

[2] It is expected that the students who are not in the group presenting will ask questions about the work.

own work. You may use other functions written by others – but these must not be counted among your most important functions.

Each group should meet with me approximately one month before the final week of classes to ensure that your work is progressing in the right direction.

Description of projects

There are four possible projects that can be undertaken. As I mentioned in class, your performance in a quiz about New Zealand will assign the choosing order for these projects.

Project 1: Variational Bayes[3]

Reference Ormerod and Wand (2010)

Data and model

Data from Givens and Hoeting (2012) on the reported number of risky sexual encounters of gay men[4]. Suppose $N = 1500$ gay men were surveyed and each was asked how many risky sexual encounters he had in the previous 30 days. Let n_i denote the number of respondents reporting i encounters, for $i = 1, \ldots, 16$. Table 1 summarizes the responses. These data are fitted poorly by a Poisson

Encounters	Frequency
0	379
1	299
2	222
3	145
4	109
5	95
6	73
7	59
8	45
9	30
10	24
11	12
12	4
13	2
14	0
15	1
16	1

Table 1: Frequency of risky sexual encounters reported by 1500 gay men.

[3]This project cannot be undertaken by the groups containing Hongyuan or Woody

[4]I am pretty sure this is not real data. For the purposes of the report you should behave as if it is.

model. To improve our estimation we envisage M groups of respondents. The first group are men who, for whatever reason, report zero risky encounters even if this is not true. A respondent has a probability of π_1 of belonging to this group. The remaining $M-1$ groups are assumed to follow a Poisson distribution. That is, group j ($j > 1$) follows a Poisson(λ_j) distribution. A respondent has probability of π_j of belonging to this group (with the constraint $\sum_{h=1}^{M} \pi_h = 1$). You should assume a Gamma(α_j, β_j) prior for λ_j, a Dirichlet($\alpha, \alpha, \ldots, \alpha$) prior for $\pi_1, \ldots, \pi_M$ and a prior mass function $\propto \frac{1}{M}$ for $2 \leq M \leq 10$. You may extend (or change) the model if desired. You are also encouraged (but not required) to compare your results to those using MCMC algorithms we will discuss in class.

Project 2: Hamiltonian Monte Carlo

Group Yifan, Wei, Zhiheng

Reference Neal (2011)

Data and model

This project does not contain data, but instead focuses on sampling from a density with high correlation. That is, your goal will be to use Hamiltonian Monte Carlo to sample from a 50-variate normal distribution, with mean $\mathbf{0}$, correlation of 0.999 between all variates and standard deviations of $1:50$ (i.e. the first variate has an sd of 1, the second has an sd of 2, ..., the 50th has an sd of 50). You should compare the Hamiltonian Monte Carlo to (i) the multivariate Metropolis algorithm, and (ii) the Gibbs sampler. You should also explore how the performance of the algorithm changes with the various "working parameters" (such as (a) the number of leapfrog steps, (b) the step size, and (c) the specification of the mass matrix). One of the reasons this algorithm is not commonly used is the difficulty in efficiently specifying these quantities. Recent work dealing with this includes Hoffman and Gelman (2013) and Girolami and Calderhead (2011).

Project 3: Slice sampling

Group Frank, Hongyuan, Meng

Reference Neal (2003)

Data and model

Implement slice sampling to sample from the posterior distribution for a model based on the example

in homework 3, question 4 (the titanic data). You should fit the same model you did for that example. Compare (i) univariate slice sampling[5], (ii) rejection sampling within a Gibbs algorithm, (iii) univariate Metropolis within a Gibbs algorithm and (iv) multivariate slice sampling. You are free to choose any priors you like[6]. You are also encouraged (but not required) to explore recent approaches such as Murray et al. (2010)[7].

Project 4: Evolutionary MC

Group Grady, Edward, Woody

Reference Liang et al. (2011), chapter 5

Data and model

This project does not contain data, but instead focuses on sampling from a density with multiple modes (this is a good test for how the algorithm would perform with a multi-modal posterior distribution). That is, your goal will be to use Evolutionary Monte Carlo to sample from a 20 part trivariate normal distribution mixtures (similar to the example mentioned in the book). Assume that each of the parts are equally likely (i.e. the mixing proportions are all 0.05), and the means for the 20 parts should be randomly drawn on $[-10, 10] \times [-10, 10] \times [-10, 10]$ with each part having a common variance of $0.1I_3$. . You should compare the Evolutionary Monte Carlo approach to (i) parallel tempering, and (ii) a multivariate metropolis updater. You should also explore how the performance of the algorithm changes with the various "working parameters" (such as (a) the temperatures, (b) the mutation rate, (c) the population size, and (d) the mutation scale parameter).

References

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Givens, G. H. and Hoeting, J. A. (2012). *Computational Statistics, 2nd ed.* John Wiley & Sons.

Hoffman, M. D. and Gelman, A. (2013). The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, In press.

[5] You should implement univariate slice sampling for ALL parameters.

[6] The choice of priors may help or hurt your ability to implement some of the algorithms described.

[7] I am not very familiar with this paper, which means that I am not 100% that this approach will work for this model – but I think it should provided the priors are appropriately specified.

Liang, F., Liu, C., and Carroll, R. (2011). *Advanced Markov chain Monte Carlo methods: learning from past samples*, volume 714. Wiley. com.

Murray, I., Adams, R. P., and Mackay, D. (2010). Elliptical slice sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 541–548.

Neal, R. (2003). Slice sampling (with discussion). *Annals of Statistics*, 31:705–767.

Neal, R. M. (2011). *Handbook of Markov Chain Monte Carlo*, chapter 5 (MCMC using Hamiltonian dynamics), pages 113 – 162. Chapman & Hall/CRC.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.