Monte Carlo Methods for Optimization
STATS 202C
Professor: Mark S. Handcock

At Home Final Project

Overview

The project can take one of two forms. You should choose one only.

The **first** is to review a topic on the use of Monte Carlo techniques on a statistical problem, such as distributional estimation or optimization. This topic should extend or expand on a topic covered in the class, but not be just that topic. The report should be stand-alone, starting from the level of the lecture notes and expanding out to review the existing literature on that topic. It need not analyze real data, but may use real or simulated data to illustrate the main points. If you choose this form you must pass the topic by me first by sending me a description in an email, chatting with me after class or coming to my office hours.

The **second** is to undertake an analysis of a complex data set for which Monte Carlo methods are useful for that analysis. The analysis should be using one or more of the methods covered in class. The idea is that he statistical model should be appropriate for the data and the Monte Carlo methods be an appropriate choice for implementing the model-based inference.

For graduate students, it is natural to choose data related to your graduate work or thesis area. If could be data you are currently working on for your thesis, or data in areas you might go with your thesis work. If this is not yet specified, you can look in the literature for studies that address questions in you area and see if you can get access to the data via groups like:

A great places to browse for data is the Dataverse at Harvard `https://thedata.harvard.edu/dvn/`. There are an amazing and fast accumulating resources for data out there. Please spend some time looking through them.

I do not want a quick and routine analysis; a good project will show understanding of the problem and possible solutions and techniques to consider. The technical conclusions should be stated clearly.

Your solution to the problem should be submitted in the form of a report (either data analysis or literature review). The report should be typed and at most 25 pages of 12 pt or greater type, including any appendices, tables, and figures. Please include margins of at least one inch on all sides of the paper. It should be submitted on the class website under the "Homework" section.

The report should be self-contained and suitable for a non-statistician with a M.S. level of knowledge of statistics.

Advanced modeling methods should be defined briefly in the text to the extent necessary for understanding of the results, along with a reference. By advanced I mean methods that were not prerequisite for the course.

If you choose the data analysis form, it is recommended that the following outline be followed in preparing the report:

i. *Abstract.* This should consist of a brief statement of the results of your analysis. This should be like that of a research paper analyzing the data.

ii. *Introduction.* Here include a clear statement of the scientific questions addressed by the your analysis of the data. The goal of the statistical analysis and the scientific context of the problem should be clear to all who read the introduction.

iii. *Analysis and Results.* Describe your analysis and its results clearly and concisely. If necessary, use graphical displays and tables to convey the results. Describe methods used, approaches taken to examine the underlying assumptions and so on. Explain why the methodology is appropriate. Avoid using highly technical language.

iv. *Discussion.* This section should describe the scientific and statistical issues raised by the results described in the previous section. Limitations of the study and of the analysis should be discussed here. There may be scientific issues that you would have liked to discuss with the investigator if this had been a real collaboration with a non-statistical scientist. If so, describe these issues and why they would be relevant to the analysis and/or interpretation of results. If appropriate, provide suggestions for further analysis or collection of additional data. Summarize your conclusions about the issues of scientific concern.

v. *References.* List the key books and articles you consulted that are reflected in your report, as well as references that might be useful to the reader if they want to know more.

vi. *Tables and Figures.* These should be included in the text, but can also be included in a separate section at the end. All should have clear titles/captions, and figures should have explanatory legends.

vii. *Appendices.* There can be one or more appendices. You may include more technical discussion of your methodology or any theory developed to implement your models. Appendices are not required (in fact they are discouraged), and should only be included only if you feel they add something important to your report.

You should spend the majority of your time thinking about the scientific and statistical issues and writing the report rather than spending all your time carrying out the statistical analysis. You may do a wonderful job of analysis, but it is of no use unless you can communicate the results to your audience.

Evaluation

Criteria used to judge performance will include the following three factors, each given equal weight:

i. *Statistical Appropriateness.* Appropriateness of the analyses and models for the data and questions. Technical execution of the analysis.

ii. *Scientific appropriateness.* Thoughtfulness and simplicity of your analysis. Does your analysis really answer the scientific questions of interest?

iii. *Quality of the written report.* The report will be judged based upon its organization, clarity, and accuracy. Simple, concise sentences are preferred to sentences which are convoluted or otherwise confusing.

Choice of your data set

The data set should be sufficient rich in information to not just be an illustration. This will depend on the application. It will likely be multivariate. Do not merely use data from a textbook – the world is an interesting place! All data sources must be cited, and described.

If you find interesting data but are unsure if it fits the class requirements, see me after class or in my office hours to discuss it.

Your statistical task is to model the structure in the data and describe it. To do this you should consider the various forms of social structure considered in the course.

Here are some suggestions:

- Start by presenting visual summaries of the data, followed by numerical summary measures.

- Describe the impact of the covariates on the outcome variable(s) via simple models first.

- For any model be sure to consider diagnostics of the appropriateness of the model.

- Construct summaries of the overall quality of the fit, and a comparison between alternatives to the model finally used.

This is an open ended question, so feel free to experiment.