Stat 27850/30850: final project

For your final project in this course, you can work in a group of size 2–4 (larger groups should develop a more substantial project). Project timeline:

- Week 6: group meets with instructor to discuss ideas for the final project (sign up for a time slot by email).
- Due Thursday week 7 = May 10: 1–2 page project proposal outlining your topic, and goals and plans for the project
- Due Thursday week 9 = May 24: 2–4 page progress report describing what you've found so far (problems or questions addressed, early empirical results, etc)
- During class on week 10 = May 29&31: final project presentations for all groups.
- Due Thursday of week 10 = May 31: final project report due for all groups with any members graduating this quarter.
- Due Thursday of finals week = June 7: final project report due for all other groups.

Guidelines for the presentation:

- The presentation should be 5–10 minutes, with slides, where you introduce the problem, present your approach, and show empirical or theoretical results.
- All team members should play a role in the presentation. Please be sure you can all attend to present on either Tuesday or Thursday of Week 10.
- It's fine if some parts of the project are not complete since your report is not yet due, but you should be able to present an interesting exploration of the problem and some preliminary results.

Guidelines for the final report:

- There is no page length requirement. Around 10–15 pages is typical, including the writing plus plots/tables/diagrams to display your results.
- Your res Please also hand in your code, clearly organized and labeled so that we can see which parts of the code generates which plots/tables/etc in the paper. Alternately, your report can be written in R markdown so that the code and the report is handed in together.
- Your final report should describe the problems and questions you posed, the details of any methods you implemented / models fitted / hypotheses tested, and should discuss some interesting issues relating to inference (for example, multiple testing / appropriately controlling for confounding factors / reducing a high dimensional model to a manageable size / etc)

Each project should be developed around one of the following data sets. We will provide scripts to download and clean the data so that it's ready for R and organized in a simple format.

1. fMRI data, from `http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/`. A human subject's brain activity is measured with a fMRI (functional MRI) scanner, during a task that involves looking at pictures of scenes and sentences describing those scenes, and determining whether the picture and sentence match. The picture is presented separately from the sentence, in either order. Each recorded image is tagged with a time stamp indicating whether the picture, the sentence, or neither is currently being displayed, along with the activity level measured at each of $>$4000 voxels (3D locations) in the brain.
2. Gene expression data, from `https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq#`. Gene expression data on 20531 genes is gathered from 801 patients, who each have one of five tumor types.

3. Bikeshare data, from `https://www.capitalbikeshare.com/system-data`. This data comes from Washington D.C.'s bikeshare program, which records every individual ride taken. Each ride is tagged with its start and end time, start and end station (there are hundreds of locations where bikes can be rented across the city), as well as whether the rider has bought a one-time rental or is a member of the bikeshare program.

4. If you would like to use a different data set from the three suggested here, please bring details to your Week 6 project planning meeting.

Your assignment for the project is completely flexible—find some interesting questions to analyze, which connect in some way to any of the issues we've discussed in the course, or to any other topics or challenges in modern inference. Here are a few examples:

1. For the fMRI data set:
   - Which voxels show significantly different activity level depending on the stimulus (picture/sentence/neither)?
   - Can we identify voxels that react to the stimulus initially, versus voxels whose activity increases later on and perhaps is caused by the early-responder voxels instead of directly by the stimulus itself?
   - Are there differences between the brain activity patterns when the picture is presented without having seen a sentence earlier, versus when a picture is presented after having seen the sentence already? (I.e. the difference between simply viewing an image, versus viewing it to determine some particular piece of information.)
2. For the gene expression data set:
   - In our class examples using gene expression data, we tested questions like, "Is gene j's expression level different for tumor type A vs type B?", but in this data set, the tumor types are very different (e.g. breast vs prostate cancer) and so the differences are very clear. A more challenging version of this question would be to ask, for instance, whether the association between genes j and k is different for the different tumor types (e.g. perhaps for one tumor type, the two genes are noticeably correlated, but not for the other tumor types).
   - Another question might be, which genes show different expression levels for the different tumor types, even after controlling for the effects of other genes? That is, maybe gene j has higher expression level for tumor type A vs type B, but after you regress out the effects of some other gene k, this difference vanishes—that might indicate that gene k is the main one that interacts with the tumor type.
   - You can also ignore tumor type, and just take the portion of the data from a single tumor type, to ask questions about gene-gene interactions. Which pairs of genes appear to be significantly correlated even after controlling for the potential effects of all other genes?
3. For the bikeshare data set:
   - One interesting direction would be to identify changes in user behavior over time. While some changes are obvious (e.g. a higher number of users), it might be interesting to explore changing patterns in routes traveled (e.g. which station to users travel to from Station A; whether typical ride time or distance has changed over time; etc).
   - A primary practical goal when collecting this type of data set is in prediction rather than inference—can you predict how many bikes will be needed at a particular location over the course of the day/week? Since this is a high dimensional data set, selecting relevant features to help this prediction problem, and predicting reliably without overfitting, is challenging.
   - You can gather additional data to inspire more questions—for example, daily weather in D.C.; calendar of events in D.C.; etc.