Graphical Models for Complex Health Data (G8124 Fall 2021)

Due date: Tuesday December 21st, 8pm

FINAL PROJECT DESCRIPTION

This is an open-ended data analysis project, where you can gain some experience applying methods based on graphical models to real data. Some data sources are made available below. Students may elect to use their own data as long as it is publicly available or they have permission to use it (but should not expect any assistance from instructors or TAs on accessing, cleaning, or handling said data). You should define your own analysis objective: what do you want to learn/estimate from this data? What scientific or inferential question do you want to address? You can use whatever methods you like, as long as they are *related* to the graphical methods we discuss in the course. There is a bit of leeway here — you need not limit yourself to methods we've directly discussed (e.g., if you want to use a graphical structure learning algorithm we have not mentioned, or an estimation procedure that we have not explicitly gone over in class, a different Monte Carlo sampling method, etc.), but whatever methods you use should be clearly related to the course material. (*In particular: do not simply learn a neural net, or use K-means clustering, or some other standard machine learning (ML) method you've learned outside this course and call it a day! This would not be acceptable. You may, however, use other ML methods not discussed in class in conjunction with graphical methods.*) You may use whatever software is publicly available to do your analysis, or implement things yourself. You will write a short paper in the style of an ML conference paper. Remember to clearly introduce your problem, data, methods, approach, and results. In fact, a good way to organize your paper is into the following sections: Introduction, Data, Methods, Results, Discussion. (You do not need to discuss "related work" though you are welcome to use already published work as an inspiration, as long as you cite it. Pure re-implementations of already published work are to be avoided.) You should justify all your analysis choices — how you chose tuning parameters, why you chose certain parametric forms or model classes, etc. Make sure your work is reproducible, so someone using the same data could implement your method and achieve the same results.

Requirements:

1) Specify, either based on background knowledge, problem design, or learning from the data, at least one graphical model that will serve as the basis of your analysis. It can be any kind of graphical model that you think is appropriate. Then, perform some sort of "inference" task based on the model. Here I mean "inference" broadly construed: parameter estimation, hypothesis testing, prediction/ classification, causal effect estimation, sample generation, MAP inference, etc.
2) You should compare at least two approaches/methods/settings. That is, you should consider what someone else might do alternatively to your proposal, try it, and compare results. How you do this is up to you: the important thing is that you try more than one thing.
3) Write a paper using the formatting guidelines of the NeurIPS conference. Formatting guidelines and sample Latex document/style file can be found here: https://neurips.cc/Conferences/2021/PaperInformation/StyleFiles (Note: do not anonymize your submission!)
4) Minimum length: 4 pages. Maximum length: 8 pages, including all tables and figures (not including references). You may include additional supplementary material if you wish but the grading will be based entirely on the content of the main paper, and supplementary material will probably not be examined at all.
5) **A 500 word (approx) project proposal is due December 1st at 8pm** (on Courseworks). Describe which data, methods, and software do you intend to use, and the goal of your analysis. This does not need to contain all the details, but it should be clear that you have a well-formed idea and a rough plan for executing it. The final project should also be submitted via Courseworks. Late submissions will not be accepted.

Note: there is a list of GM-related software packages in R here: https://cran.r-project.org/web/views/gR.html This is of course not a complete list, new packages are added all the time and many are on Github or other sites. However, this list is a good place to start.

AVAILABLE DATA

1) **fMRI dataset**. Download here: https://www.dropbox.com/s/jh51ase6e5wpx7v/fmri%20data.zip?dl=0

This data set contains fMRI brain scan results for individuals with Autism Spectrum Disorder (ASD) as well as "neurotypical" controls. It was taken from the Autism Brain Data Exchange:

http://fcon_1000.projects.nitrc.org/indi/abide/abide_I.html

Specifically, we have the Carnegie Mellon University dataset, where the age range of subjects is between 19 and 40. The preprocessing pipeline used is NIAK with 160 Regions of Interest (ROIs):

http://preprocessed-connectomes-project.org/abide/Pipelines.html

On the above website you can see the preprocessing steps performed by NIAK (last column in the comparison tables). And in the last section of the website you can see the description of the parcellation to get the ROIs: "Dosenbach 160."

The folder contains two subfolders, one with data for 14 ASD individuals, and the other with 13 controls. The individuals all have the same number of ROIs/variables (columns), but potentially different number of samples (rows), because some samples are dropped by the data preprocessing to remove artifacts of head motion (akin to outliers), etc. The data was sampled every 2 seconds, and the size of the voxels is 3mm x 3mm x 3mm. There is also a file called "phenotypic_CMU.csv" which has some metadata on the individuals in the data set. *Note: due to a bug there is an extra ROI (column 161) in this data which you should just remove before analysis.*

2) **Genetics data.** The data used by Wang et al. (2016) in their "FastGGM" paper is available here: https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1425/ Another source of genetics data is here: https://jhubiostatistics.shinyapps.io/recount/ Note: Use the "TCGA" data and focus on a specific tissue (e.g., "lung").

The TCGA data is a relatively "clean" RNA-seq data which is easy to download (instructions on the above website). However, it is *not* very easy to understand if you don't already have some familiarity with data of this type. So, if you have no experience with such data, I would probably advise against using it.

3) **X-ray image data.** The ChestX-Ray8 from NIH contains >100K chest x-ray images of >30K unique patients, along with radiologist labels to indicate 14 common pathologies of the thorax. The data (both raw images and labels/other metadata) are available here: https://nihcc.app.box.com/v/ChestXray-NIHCC

4) **NHANES epidemiological data.** Data from the National Health and Nutrition Examination Survey (NHANES) is publicly available. Lots of info can be found here: https://www.cdc.gov/nchs/nhanes/index.htm. One relatively straightforward way to download the data directly in R is to use the package RNHANES which is described here: https://cran.r-project.org/web/packages/RNHANES/vignettes/introduction.html. There are a lot of resources/tutorials on this data floating around the web, for example: https://www.r-bloggers.com/2016/11/nhanes-made-simple-with-rnhanes/ and also tutorials on the main CDC website above.

GRADING RUBRIC

Meeting the basic requirements: 40 points

Clearly stated objectives: 5 points

Appropriateness of methods for the stated task(s): 10 points

Adequate description of the methods used (including all modeling choices, any tuning parameters, parametric forms, etc): 20 points

Informative presentation of the results: 10 points

Clear and understandable writing: 10 points

Creativity/novelty: 5 points

Total: 100 points