Graphical Models for Complex Health Data (P8124 Fall 2025)

Due date: Friday December 18th, 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. **Two data sources are made available below.** You may elect to use your own data **only if you obtain permission** from Professor Malinsky. (You must also have permission to use the data; publicly available data is ok only if the data is of high quality.) 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 train 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 report following the template attached. **You must justify all your analysis choices — how you chose tuning parameters, why you chose certain parametric forms or model classes, etc.** Your work must be reproducible, so someone using the same data could implement your method and achieve the same results. You will be required to share your code on a private GitHub repository. TAs/instructor should be able to reproduce your findings. In some cases, we will run your code to make sure it works and does what is claimed.

Requirements:

1) You must use some statistical methods based on graphical models, closely related to the content of the course. Whatever graph(s) you specify or learn, you must also do some statistical test or parameter estimation based on the graph(s) to answer a substantive scientific question. (That is, I want you to *do something* with graphs, not just learn some graphs and call it a day.)
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 report in LaTeX following the provided template. You must use full sentences and paragraphs, not bullet points. The writing should be clear and grammatical.
4) Minimum length: 3 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) Below, we provide two data sources and a list of possible scientific questions that you may endeavor to answer with the data. If you choose to either (1) use data different from the sources provided below or (2) answer a scientific question / pursue a scientific goal different from one of the provided questions, you must submit a one-page project proposal describing your proposed analysis to Prof. Malinsky by **December 2nd at 8pm**. 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.
6) 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/GraphicalModels.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://rutgers.box.com/s/imgbdaqhzlkbunf52ia8xfmiu97b05xh

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."

There are two diagnostic categories of patients, 14 ASD individuals and 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 a file called "phenotypic_CMU.csv" which has some metadata on the individuals in the data set, including which diagnostic category each individual belongs to (1=ASD, 2=control). *Note: due to a bug there is an extra ROI (column 161) in this data which you should just remove before analysis. Also, the files are really .csv files but for some reason the file extension was dropped. You might need to manually rename each data file as "CMU_a_005[….]1D.csv" so that your computer has an easier time opening and reading the files.*

Alternative: the fMRI data from the Dajani et al. (2017) paper is available here: https://github.com/cheninstitutecaltech/Caltech_DATASAI_Neuroscience_23/blob/main/07_20_23_day9_causal_modeling/code/solutions/exercise3.ipynb (instructions described in a Python notebook)

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").

*Note: 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.*

SCIENTIFIC QUESTIONS

(i) Compare ASD vs NT (neurotypical) control subjects on the basis of their estimated connectivity networks. Is the difference between a random ASD individual and a random NT individual larger/smaller, on average, than than the difference between individuals within same group category?

(ii) Is there something reliably different in the connectivity structures when comparing ASD vs NT subjects? Is there reason to believe this has something to do with ASD vs NT status or could it be explained another way?

(iii) Can you somehow reliably "predict" ASD status on the basis of learned connectivity networks? Can you use learned networks to classify individuals into ASD vs NT categories? What features, if any, of learned connectivity help predict ASD status?

(iv) Are genetic regulatory networks really different in different disease categories? What is different about them?

(v) Can differences in genetic regulatory networks be used to classify or cluster individuals into categories that reflect disease status?

(vi) Is it possible to identify genes that have a strong causal effect on some phenotype of interest? Is it possible to distinguish between genes that have a causal effect vs gene expressions that are associated with phenotype due to confounding?

(Note: there are many ways to interpret these questions and specific hypotheses that may be generated under these headings. You have some flexibility here. If you want to do something different than one of these questions, you must submit a proposal by December 2nd — see requirements above.)

GRADING RUBRIC

Meeting the basic requirements: 40 points

Clearly stated objectives: 10 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): 25 points

Informative presentation of the results: 10 points

Clear and understandable writing: 5 points

(Creativity/novelty: up to 5 points extra)

Total: 100 points