Graphical Models for Complex Health Data
P8124

COURSE SCHEDULE

Tues & Thurs 4:00-5:20pm Hammer LL 207

INSTRUCTOR

Prof. Daniel Malinsky, PhD
dsm2128@cumc.columbia.edu
Allen Rosenfeld Building, R649; Office Hours Thursday: 2-3pm

TA: Angel Garcia de la Garza
ag3914@cumc.columbia.edu
Allen Rosenfeld Building, 6th floor conf room; Office Hours Tuesday: 1-2pm

COURSE DESCRIPTION

This is a course at the intersection of statistics and machine learning, focusing on graphical models. In complex systems with many (perhaps hundreds or thousands) of variables, the formalism of graphical models can make representation more compact, inference more tractable, and intelligent data-driven decision-making more feasible. We will focus on representational schemes based on directed and undirected graphical models and discuss statistical inference, prediction, and structure learning. We will emphasize applications of graph-based methods in areas relevant to health: genetics, neuroscience, epidemiology, image analysis, clinical support systems, and more. We will draw connections in lecture between theory and these application areas. The final project will be entirely "hands on," where students will apply techniques discussed in class to real data and write up the results.

PREREQUISITES

P8105: Data Science I and P8109: Statistical Inference

FOUNDATIONAL PUBLIC HEALTH KNOWLEDGE

This course addresses concepts and topics essential to public health, including the following:

- Explain the role of quantitative and qualitative methods and sciences in describing and assessing a population's health
- Explain the critical importance of evidence in advancing public health knowledge

DEGREE COMPETENCIES

This course is designed to help students attain mastery of the following degree competencies. Student achievement of these competencies will be measured through performance on the corresponding assessments.

Competency	Primary Assessment(s)	Secondary Assessments
Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software, as appropriate	Homework Assignments 1-4 and the Final Project	
Interpret results of data analysis for public health research, policy or practice	Readings and discussion from "Application Focus" sessions	Class Participation

COURSE LEARNING OBJECTIVES

By the time you complete this course, you should be able to

- Select the model class / representation appropriate to a given data problem
- Explain the semantics of different graphical model classes, e.g., directed and undirected graphs
- Perform inference and learning/estimation tasks for multiple model classes and data types
- Analyze real data using graphical methods

COURSE REQUIREMENTS

Required Course Materials

Students are not required to purchase any texts for this course. All readings, including book chapters and journal articles, will be made available for download from Courseworks. Selections from two textbooks will be used:

- Steffen L. Lauritzen (1996), *Graphical Models*, Clarendon Press.
- Kevin P. Murphy (2012), *Machine Learning: A Probabilistic Perspective*, MIT Press.

Some additional texts may serve as useful references:

- Daphne Koller & Nir Friedman (2009), *Probabilistic Graphical Models*, MIT Press.
- David Edwards (2000), *Introduction to Graphical Modelling* 2nd Ed., Springer.
- Martin J. Wainwright & Michael I. Jordan (2008), Graphical Models, Exponential Families, and Variational Inference, in *Foundations and Trends in Machine Learning* 1(1-2): 1-305.

COURSE STRUCTURE

The class will meet twice a week for lecture. In class meetings we will discuss both aspects of theory and examine some papers which apply techniques based on graphical models to applied scientific problems. Students will complete homework assignments individually, and eventually begin work on their final data analysis projects. Each student will meet with the instructor at least once toward the beginning of their data analysis project to discuss their proposed plan. The instructor and teaching assistants will be available for office hours.

ASSESSMENT AND GRADING POLICY

Student grades will be based on:

Homework Assignment 1 (problem set)	15%
Homework Assignment 2 (problem set)	15%
Homework Assignment 3 (problem set)	15%
Homework Assignment 4 (problem set)	15%
Class Participation	5%
Final Project Proposal	5%
Final Project (data analysis & report)	30%

Homework assignments will include theoretical problems, data analysis, and programming in R. **(Proficiency in R, at least at the level taught in P8105, is required for this course.)** The assignments must be typeset, preferably in LaTeX. They will be submitted online via Courseworks. The final project will be a research project that requires students to apply methods learned in the class to real data. Several public (and relatively "clean") data sets will be made available, spanning multiple scientific areas: computational biology, neuroscience, social science, etc. Students will write a report in the style of a short research paper, about 4-7 pages, applying graphical methods to this data. The report will be graded according to a rubric, which will be provided to the students. The mandatory project proposal, worth 5% of the grade, is a brief write-up which will be due several weeks before the final deadline, describing the data set, the methods, and software involved. The point of the proposal is to incentivize planning, and to identify and potential problems or pitfalls ahead of time. Additional details regarding the expectations for this project will be made available in class. Class Participation grades will be evaluated based on 3 factors: class attendance, participation in class-time discussions, and participation in online discussions via the discussion board (in Courseworks).

All assignments in this course are individual, not group, assignments. You may freely discuss homework assignments with your fellow classmates. The final solutions, however, must be written entirely on your own. This includes programming: you must implement any programming task on your own. Copying someone else's code (and then subsequently making minor changes) constitutes plagiarism. So, if you need to discuss programming assignments, you may discuss general strategy but should write the code by yourself.

Late assignments will be penalized by one letter grade for every day they are late. If an assignment is submitted more than two days late it will be given a zero. Late final projects will not be accepted except for a condition or circumstance documented by the Office of Student Affairs.

Assignment	Description	Due date
Homework Assignment #1	Theoretical analysis of graphical model properties, programming in R	TBA
Homework Assignment #2	Theoretical problems covering parameter learning and backdoor adjustment, programming in R	TBA
Homework Assignment #3	Theoretical problems and programming in R related to structure learning	TBA
Homework Assignment #4	Theoretical problems and programming in R related to mixture models and approximate inference	TBA
Final Project	Analysis of real data with software of student's choosing, written research report	TBA (finals period)

Grading

A+	Reserved for highly exceptional achievement.
A	Excellent. Outstanding achievement.
A-	Excellent work, close to outstanding.
B+	Very good. Solid achievement expected of most graduate students.
B	Good. Acceptable achievement.
B-	Acceptable achievement, but below what is generally expected of graduate students.
C+	Fair achievement, above minimally acceptable level.
C	Fair achievement, but only minimally acceptable.

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath.

Academic Integrity

Students are required to adhere to the Mailman School Community Standards and Conduct handbook, which includes the Code of Academic Integrity.

Disability Access

In order to receive disability-related academic accommodations, students must first be registered with the Office of Disability Services (ODS). Students who have or think they may have a disability are invited to contact ODS for a confidential discussion at 212.854.2388 (V) 212.854.2378 (TTY), or by email at disability@columbia.edu. If you have already registered with ODS, please speak to your instructor to ensure that they have been notified of your recommended accommodations by Meredith Ryer (mr4075@cumc.columbia.edu), Assistant Director of Student Support and Mailman's liaison to the Office of Disability Services.

Bias Response and Support System

Our community at Columbia University's Mailman School of Public Health is committed to creating an inclusive working, learning, and living environment where all are respected. The

occurrence of bias related incidents, involving conduct, speech, or expressions reflecting prejudice are an opportunity for learning and growing as a community. At the request of faculty, students, and staff, and in partnership with Student Conduct and Community Standards, the School has developed a Bias Response & Support System (BRSS) that will aid us in:

- Addressing bias-related concerns
- Ensuring that community members receive necessary support
- Identifying patterns of bias and best practices in promoting a bias-free environment that will be used to inform future programming and further our growth as an inclusive community

BRSS will ensure that we hold each other and ourselves accountable to the School's commitment to diversity, equity, and inclusion by acknowledging and addressing bias-related concerns.

You can access the BRSS here.

COURSE SCHEDULE

Please see the modules and files sections of Courseworks to download the readings and lecture slides.

Session 1 – Introduction to graphical models and conditional independence
[Sep.9] Learning Objectives: You will be able to 1. Describe key applications of graphical models in several scientific domains 2. Explain and distinguish properties of correlation, association, and conditional independence Assignment: None

Session 2 – Directed Acyclic Graphs (DAGs) / Bayesian Networks
[Sep.14-16] Learning Objectives: You will be able to 1. Define and explain the properties of DAG models: factorization, local Markov, d-separation 2. Identify and analyze key concepts: colliders, Markov equivalence, faithfulness, context-specific independence 3. Recognize and explain some special case DAG models (HMMs, DBNs, etc.) Assignment: 1. (Read) Lauritzen, chapter 3 (and section 2.1.1 for relevant notation)

Session 3 – Undirected Graphs (UGs) / Markov Random Fields

[Sep.21-23] Learning Objectives: You will be able to

1. Define and explain the properties of UG models: factorization, local Markov, pairwise Markov, undirected separation
2. Identify and explain some special case UG models (pairwise MRF, log-linear MRF) and applications of these
3. Distinguish between "discriminative" and "generative" modeling for prediction

Assignment:

1. (Finishing reading or re-read) Lauritzen, chapter 3
2. Homework Assignment #1 (due before Session 5)

Session 4 – Parameter learning: maximum likelihood and Bayesian estimation

[Sep.28-30] Learning Objectives: You will be able to

1. Analyze maximum likelihood estimation (MLE) of discrete and continuous (Gaussian) DAG models
2. Analyze the MLE for Gaussian MRFs and explain the contrast with a DAG parameterization
3. Contrast the frequentist properties of the MLE with Bayesian estimates of graph parameters

Assignment:

1. (Read) Hastie et al. (2009), *Elements of Statistical Learning*, chapter 17, Sections 17.3 –17.4.1
2. (Optional additional reading on Courseworks*)*

Session 5 – Application focus: epidemiology

[Oct.5-7] Learning Objectives: You will be able to

1. Explain how DAGs are used to assess possible sources of bias, design issues, and modeling choices in epidemiological studies
2. Apply the backdoor criterion for sufficient confounding adjustment with DAGs

Assignment:

1. (Read) Greenland et al. (1999), Causal diagrams for epidemiological research, *Epidemiology* 10(1): 37-48.
2. (Read) Johnson et al. (2019), Structure and control of healthy worker effects in studies of pregnancy outcomes, *American Journal of Epidemiology* 188(3): 562-569.

3. (Optional additional readings on courseworks)
4. Homework Assignment #2 (due before Session 7)

Session 6 – Structure Learning (part 1): constraint-based and score-based learning
[Oct.12-14] Learning Objectives: You will be able to 1. Explain and apply graphical lasso and neighborhood selection algorithms for learning UGs 2. Describe constraint-based and score-based learning for DAGs (PC algorithm and GES) Assignment: 1. (Read) Drton and Maathuis (2017), Structure learning in graphical modeling, *Annual Review of Statistics and Its Application* 4: 365-393. [you may skim]

Session 7 – Structure Learning (part 2): advanced learning methods
[Oct.19-21] Learning Objectives: You will be able to 1. Describe and apply methods for learning with hidden variables (FCI algorithm + others) 2. Explain learning approaches based on semiparametric structural assumptions (LiNGAM + others) Assignment: 1. Homework Assignment #3 (due before Session 10)

Session 8 – Application focus: genetics
[Oct.26-28] Learning Objectives: You will be able to 1. Explain how several recent papers that apply graphical methods to the study of genetic regulatory networks handle domain-specific challenges: high-dimensions, non-Gaussian distributions, background structural knowledge, tuning parameter selection Assignment: 1. (Read) Stekhoven et al. (2012), Causal stability ranking, *Bioinformatics* 28(21): 2819-2823. 2. (Read) Wang et al. (2016), FastGGM: An efficient algorithm for the inference of Gaussian graphical model in biological networks, *PLOS Computational Biology* 12(2): e1004755. 3. (Read) Ma et al. (2018), Constructing tissue-specific transcriptional regulatory networks via a Markov random field, *BMC Genomics* 19(10): 65-77.

Session 9 – Application focus: neuroscience
[Nov.4-9] Learning Objectives: You will be able to 1. Explain how several recent papers which apply graphical methods to the study of brain data handle domain-specific challenges: heterogeneity, time series, feature selection/definition, tuning parameter selection Assignment: 1. (Read) Dajani et al. (2019), Parsing heterogeneity in autism spectrum disorder and attention-deficit/hyperactivity disorder with individual connectome mapping, *Brain Connectivity*, 9(9): 673-691. 2. (Read) Dubois et al. (2020), Causal mapping of emotion networks in the human brain: Framework and initial findings, *Neuropsychologia* 145: 106571.

Session 10 – Mixture Models, EM Algorithm, and Missing Data
[Nov.11-16] Learning Objectives: You will be able to 1. Analyze mixture models, their estimation and identifiability 2. Explain the expectation-maximization (EM) algorithm and its relevance to mixture models 3. Explain different assumptions (MAR, MNAR) relevant to analyzing missing data 4. Explain how to use the EM algorithm and multiple imputation to handle MAR data Assignment: 1. (Read) Murphy, chapter 11.

Session 11 – Approximate Inference (part 1): variational methods
[Nov.18-23] Learning Objectives: You will be able to 1. Define approximate inference and formulate inference as an optimization problem 2. Derive ELBO and mean field update equations 3. Describe and explain example applications in image analysis (e.g., image denoising) Assignment: 1. (Read) Murphy, chapter 21. 2. Homework Assignment #4 (due before Session 13)

Session 12 – Approximate Inference (part 2): Markov Chain Monte Carlo, Gibbs sampling
[Nov.30-Dec.2] Learning Objectives: You will be able to 1. Describe and distinguish sampling methods, e.g., importance sampling, Gibbs sampling 2. Analyze the stationary distribution of a Markov Chain 3. Apply MCMC diagnostics Assignment: 1. (Read) Murphy, chapters 23 (up to 23.5) and 24 (up to 24.5)

Session 13 – Graphs + Deep Learning: Restricted Boltzmann Machines, autoencoders
[Dec.7-9] Learning Objectives: You will be able to 1. Compare/contrast several neural network architectures from a graphical perspective 2. Analyze and explain variational autoencoders (VAEs) and Restricted Bolzmann Machines (RBMs) Assignment: 1. (Read) Murphy, chapter 28

Session 14 – Application focus: image analysis
[Dec.14-16] Learning Objectives: You will be able to 1. Explain the role(s) of deep learning methods in medical image applications Assignment: 1. (Read) Yu et al. (2020), An auto-encoder strategy for adaptive image segmentation, *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, PMLR 121: 881-891. 2. (Read) Shen et al. (2017), Deep learning in medical image analysis, *Annual Reviews of Biomedical Engineering* 19: 221-248.

BRIEF FINAL PROJECT DESCRIPTION (additional details to be provided later)

This is an open-ended data analysis project, where you will gain some experience applying methods based on graphical models to real data. Several data sources will be made available for use; students may use alternative data sources with permission from the instructor. You should define your own analysis objective(s): what do you want to learn/estimate from this data? What scientific or inferential question do you want to address? You may 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 different Monte Carlo sampling method, a graphical structure learning algorithm we have not mentioned, or an estimation procedure that we have not explicitly gone over in class, etc.), but whatever methods you use should be clearly related to the course material. 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 a machine learning conference paper or a short journal article. Remember to clearly explain your problem, data, methods, approach, and results. (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. Some specific additional project parameters/constraints, along with data sources, will be described in a document that will be posted to Courseworks later in the semester.

FINAL PROJECT 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