COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

Graphical Models for Complex Health Data (P8124)

COURSE SCHEDULE
Tues & Thurs 4:00-5:20pm in Hammer LL204

INSTRUCTOR
Prof. Daniel Malinsky, PhD
dsm2128@cumc.columbia.edu
Office Hours Tuesdays at 2-3pm; Allan Rosenfield Building, R649

TAs: Safiya Sirota and Ting-Hsuan Chang
sns2819@cumc.columbia.edu and tc3255@cumc.columbia.edu
Office Hours Wednesdays at 10-11am & 2-3pm; room TBA

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" sections	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 Assignments	60% (15% each x 4)
Short Quiz (In-Class)	10%
Class Participation/Online Discussions	5%
Final Project Proposal	5%
Final Project (data analysis & report)	20%

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, epidemiology, 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 student's final grade*, is a brief write-up which will be due several weeks before the final deadline, describing the student's research plan. The point of the proposal is to incentivize advance 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. (*Note: if the class size grows, the final project may pivot to a group project instead of an individual one.*)

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.

AI Policy: Permitted for HW Assignments with Attribution; not Permitted for Final Projects

In this course, students are permitted to use Generative AI Tools such as ChatGPT for help with homework assignments, as designated by the instructor. To maintain academic integrity, students must disclose any use of AI-generated material. As always, students must properly use attributions, including in-text citations, quotations, and references. **Generative AI Tools are not permitted for completing the final project.**

A student should include the following statement in assignments to indicate use of a Generative AI Tool: "The author(s) would like to acknowledge the use of [Generative AI Tool Name], a language model developed by [Generative AI Tool Provider], in the preparation of this assignment. The [Generative AI Tool Name] was used in the following way(s) in this assignment [e.g., brainstorming, grammatical correction, citation, which portion of the assignment]."

Assignment	Description	Due date
Homework Assignment #1	Theoretical analysis of graphical model properties, programming in R	Sep.20
Homework Assignment #2	Theoretical analysis of graphical model properties, programming in R	Oct.4
Homework Assignment #3	Theoretical problems covering parameter learning and backdoor adjustment, programming in R	Oct.18
Homework Assignment #4	Theoretical problems and programming in R related to structure learning	Nov.8
Short Quiz	Conceptual Questions	TBA
Final Project	Analysis of real data with software of student's choosing & written research report	Dec.19 (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.
C-	Very low performance.
F	Failure. Course usually may not be repeated unless it is a required course.

Please refer to the School Handbook for further details on grading and good academic standing.

Courseworks

Individual- and course-level activity data are collected and maintained in CourseWorks, Panopto and other educational technology tools, and may be analyzed or monitored by the course faculty, teaching team, and/or the Office of Education to improve course experience and student support. Details about the information collected can be found here.

MAILMAN SCHOOL POLICIES AND EXPECTATIONS

Students and faculty have a shared commitment to the School's mission, values and oath. mailman.columbia.edu/about/mission-history

Academic Integrity

Students are required to adhere to the Mailman School Conduct and Community Standards, which includes the Code of Academic Integrity. Columbia Mailman and Columbia University take academic integrity very seriously. This instructor and course are no different. Should any student be suspected of an academic integrity violation, there will be a report submitted to the Center for Student Success & intervention/Student Conduct. After these offices conduct their process, if a student is found responsible for violating an academic integrity policy (see Standards & Discipline/Academic Violations and Student Honor Code & Professional Guidelines), they will be assigned a grade penalty, with a possible outcome being a 0% on the assignment. Please review the university, school, and course policies, as you are responsible for behaving according to the outlined expectations.

Personal Support

Students sometimes experience life challenges that require additional support and connection to resources. If you are experiencing difficult circumstances, please reach out for help and support. Student Support Services in the Office of Student Affairs is poised to connect with students, provide resource referrals, and provide ongoing, non-clinical support. They are a good place to start if you do not know where to turn.

If you would like to make a referral to student support, whether for yourself or someone else, the best way to do so is to fill out the form linked below, and student support will reach out directly. You may also email Meurcie Zignoli at mz3047@cumc.columbia.edu to connect to Student Support Services. Please note that this form is separate from other reporting structures in place for code of conduct violations or Title VI/ Title IX concerns. This form is also not an emergency response mechanism. In case of an emergency, please contact Public Safety at 212-305-7972: Mailman Student Support Team Referral Form (maxient.com)

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 Incidents

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.

As part of our efforts to create as inclusive a community as possible, when bias incidents occur at Columbia, we provide an opportunity for those involved to engage in education, advocacy and conversation. In this way, we work to address the incident and minimize the potential for future occurrences. Our community's tools to address bias include a reporting process and the Bias Incident Resource Team, plus resources within schools and various offices. You can access information about the Bias Reporting Process and FAQs here.

Why Reporting Matters and How to File a Report
Our priority is ensuring that Columbia University is a safe community and workplace where we can learn, live, work and express ourselves. As members of the community, we have a shared responsibility to uphold these standards and report behavior that violates these standards. The reporting options provide Columbia University community members an opportunity to share important information directly with appropriate offices. If you or a member of the community needs support please take the time to complete a report so we may provide support, care, and accountability: https://universitylife.columbia.edu/report

COURSE SCHEDULE

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

Section 1 – Introduction to graphical models and conditional independence	
[Sep.3]	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 Reading: None

Section 2 – Directed Acyclic Graphs (DAGs) / Bayesian Networks	
[Sep.5,10,12]	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.) and applications of these Reading: 1. (Read) Lauritzen, chapter 3 (and section 2.1.1 for relevant notation)

Section 3 – Undirected Graphs (UGs) / Markov Random Fields	
[Sep.17,19]	Learning Objectives: You will be able to 1. Define and explain the properties of UG models: factorization, local Markov, pairwise Markov, undirected separation 2. Understand the differences between directed and undirected models 3. Recognize and explain some special case UG models (pairwise MRF, log-linear MRF) and applications of these Reading: 1. (Finishing reading or re-read) Lauritzen, chapter 3

Section 4 – Graph parameter learning: maximum likelihood and Bayesian estimation
[Sep.24, 26] 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 Reading: 1. (Read) Hastie et al. (2009), *Elements of Statistical Learning*, chapter 17, Sections 17.3 –17.4.1

Section 5 – Application focus: epidemiology
[Oct.1,3] 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 Reading: 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. Required: post 2 discussion points/questions on courseworks.** 4. (Optional additional readings on courseworks)

Section 6 – Structure Learning (part 1): constraint-based and score-based learning
[Oct.8,10] 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) Reading: 1. (Read) Drton and Maathuis (2017), Structure learning in graphical modeling, *Annual Review of Statistics and Its Application* 4: 365-393. [you may skim]

Section 7 – Structure Learning (part 2): advanced learning methods	
[Oct.15,17]	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) Reading: None

Section 8 – Application focus: genetics	
[Oct.22,24]	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 Reading: 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. **Required: post 2 discussion points/questions on courseworks.** 4. (Optional) Ma et al. (2018), Constructing tissue-specific transcriptional regulatory networks via a Markov random field, *BMC Genomics* 19(10): 65-77.

Section 9 – Application focus: neuroscience	
[Oct.29,31]	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 Reading: 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. 3. **Required: post 2 discussion points/questions on courseworks.**

Section 10 – Latent Variable Models	
[Nov.7,12]	Learning Objectives: You will be able to 1. Analyze mixture models, their estimation and identifiability 2. Analyze other latent variable models: factor analysis and latent causal models ("representation learning") Reading: 1. (Read) Murphy, chapter 11.

Section 12 – Approximate Inference (part 1): Markov Chain Monte Carlo, Gibbs sampling	
[Nov.14,19]	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 Reading: 1. (Read) Murphy, chapters 23 (up to 23.5) and 24 (up to 24.5)

Section 11 – Approximate Inference (part 2): variational methods	
[Nov.21,26]	Learning Objectives: You will be able to 1. 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) Reading: 1. (Read) Murphy, chapter 21.

Section 13 – Graphs + Deep Learning: Restricted Boltzmann Machines, autoencoders	
[Dec.3,5]	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 Boltzmann Machines (RBMs) Reading: 1. (Read) Murphy, chapter 28

Section 14 – Application focus: image analysis
[Dec.10,12] <u>Learning Objectives</u>: You will be able to 1. Explain the role(s) of deep learning methods in medical image applications <u>Reading</u>: 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.