Math 561

Math 5xy - Algebraic and geometric methods in statistics

Course Description (Bulletin): Algebraic structures are present in a broad variety of statistical contexts, involving both parametric and non-parametric statistical models for continuous and discrete random variables. A broad range of algebraic tools is used to better understand model structure, improve statistical inference, and explore new classes of models. The course offers an overview of fundamental theoretical constructions relevant to some of the more popular recent applications in the field: exact conditional test for discrete data, likelihood geometry, parameter identifiability and model selection, network models with applications to social sciences and neuroscience, and phylogenetics and tree-based evolutionary models in biology.

Enrollment: Graduate elective.

Textbook(s): *Algebraic statistics,* (2018+), Seth Sullivant, American Mathematical Society, Graduate Studies in Mathematics [in press].

Other required material: Several research papers in various mathematics and statistics journals published in the last 20 years, depending on topic emphasis. Use of a statistical software, such as R, and related algebraic computation packages. (All software will be free/open source.)

Other references: *Lectures in algebraic statistics*, (2008) Mathias Drton, Bernd Sturmfels, and Seth Sullivant. Oberwolfach lecture series, Birkauser, ISBN 978-3-7643-8905-5 (available online).
Markov bases in algebraic statistics, (2012) Akimichi Takemura, Hisayuki Hara, and Satoshi Aoki, Springer series in statistics, ISBN 978-1-4614-3719-2.

Prerequisites: Undergraduate course in mathematical statistics (such as MATH 476 or beyond), basic proficiency in either computational algebra, graduate-level linear algebra, discrete optimization, or combinatorics. Or by instructor approval.

Objectives:

1. Students will develop the ability to recognize geometric structure in statistical models, realize its impact on statistical inference procedures in practice, and identify open challenges.
2. Students will become familiar with the major tools used in the field of algebraic statistics, including some basics of computational algebra, optimization, graph theory, and matroid theory.
3. Students will understand the basic notions of model geometry and geometry of parameter spaces and how the statistical problems can be translated to abstract mathematical problems which can often be solved directly with tools from other fields of mathematics.

4. Students will practice their knowledge of these techniques through the required use of a statistical computing software.
5. Students will further explore the field, and develop their research and communication skills through a course project and presentation. The project will concern a topic approved by the instructor. Topics can include (computational) applications of the course material to student's own research area, and expository talks (with proofs) on material not covered in class (such as from a research paper).

Lecture schedule: 2 75-minute lectures per week

Course Outline:

Topic	*Hours*
0. What is algebraic statistics? An invitation / introduction / overview of the field.	2
1. Exponential families 1.1. Statistical foundations 1.2. Underlying algebra	6
2. Conditional independence and graphical models 2.1. Statistical foundations 2.2. Underlying algebra	6
3. Goodness-of-fit testing of models for discrete data 3.1. Overview 3.2. Chromosome clusters in cancer cells 3.3. Network data 3.4. Challenges of large, sparse data sets	9
4. Parameter identifiability 4.1. Overview 4.2. Graphical models 4.3. Phylogenetics and evolutionary biology 4.4. Model selection: learning a causal graph	6
5. Maximum likelihood estimation 5.1. Introduction 5.2. Deciding existence of ML estimators 5.3. Algorithms for MLE: convex and non-convex optimization	6
Exams and overflow	3

Assessment:
Homework 25-50%
Quizzes and the mid-term exam 25-40%

Final Exam 25-40%
Project 10-20%

Syllabus prepared by: Sonja Petrović and Despina Stasi, 8/7/2018