Algebraic & Geometric Methods in Statistics

Outline and some illustrative examples in nonlinear statistics

Sonja Petrović
Created for Math/Stat 561

Jan 9, 2023.

Goals

After the course, you can:

- **list topics** in algebraic statistics
- **recognize problems** in statistics that are answerable by algebraic methods
- **assess which algebraic methods** are suitable for solving a problem
- **apply basic algebraic tools** to solve a problem

Tentative course outline:

0. **What is** algebraic statistics? An invitation / introduction / overview
1. **Exponential families** 1.1. Statistical foundations 1.2. Underlying algebra
2. **Conditional independence and graphical models** 2.1. Statistical foundations 2.2. Underlying algebra
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
4. **Parameter identifiability** 4.1. Overview 4.2. Graphical models 4.3. Phylogenetics and evolutionary biology 4.4. Model selection: learning a causal graph
5. **Maximum likelihood estimation** 5.1. Introduction 5.2. Deciding existence of ML estimators 5.3. Algorithms for MLE: convex and non-convex optimization

Materials

Books and resources

Main textbook: Seth Sullivant *"Algebraic Statistics"*. It is avaiable in the bookstore. (I will check with the library for an e-copy.)
General course syllabus is here

Homework and grade

Approximately 6-7 assignments, expect a usual weekly workload.

Project

Reading a paper, working on a small research project, or applying algebraic methods on a data set, and writing a report on it. Timeline will be determined soon; project will take place during second half of semester. Groups up to 2 students.
Alternative: Participate (team) in the Eric and Wendy Schmidt Center's cancer immunotherapy data science challenge:
https://go.topcoder.com/schmidtcentercancerchallenge/

Communication

We will *not* use Blackboard. Email is not efficient. So.. ??

- Campuswire
- GitHub
- Your input please as I decide. Decision will be made THIS WEEK.

Saving this information:

Course homepage will be created here: sonjapetrovicstats.com/teaching, and the syllabus will be posted there.

Student input time!

AhaSlides.com/NLASTAT1

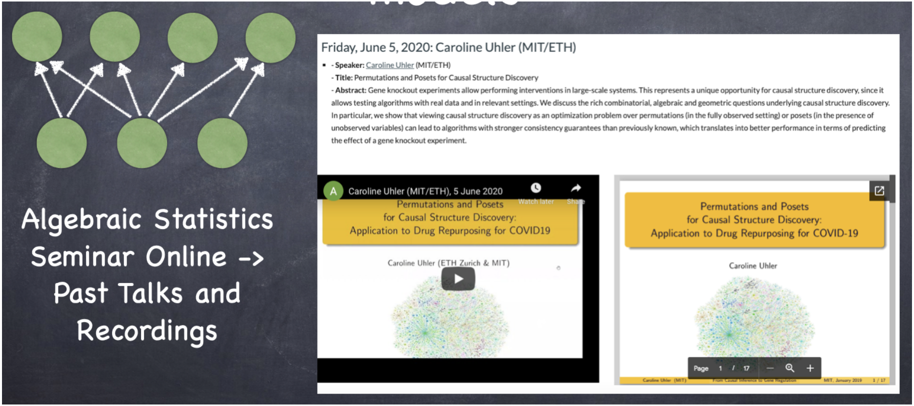
Motivating example 1: Discrete Markov chain

Section 1.1. of the textbook.

Lecture on board.

Action item: derive the two polynomial equations for the Markov chain model.

Motivating example 2: Graphical models



Algebraic Statistics Seminar Online -> Past Talks and Recordings

Friday, June 5, 2020: Caroline Uhler (MIT/ETH)

- - **Speaker:** Caroline Uhler (MIT/ETH)

  - **Title:** Permutations and Posets for Causal Structure Discovery

  - **Abstract:** Gene knockout experiments allow performing interventions in large-scale systems. This represents a unique opportunity for causal structure discovery, since it allows testing algorithms with real data and in relevant settings. We discuss the rich combinatorial, algebraic and geometric questions underlying causal structure discovery. In particular, we show that viewing causal structure discovery as an optimization problem over permutations (in the fully observed setting) or posets (in the presence of unobserved variables) can lead to algorithms with stronger consistency guarantees than previously known, which translates into better performance in terms of predicting the effect of a gene knockout experiment.

Caroline Uhler (MIT/ETH), 5 June 2020

Permutations and Posets
for Causal Structure Discovery:
Application to Drug Repurposing for COVID19

Caroline Uhler (ETH Zurich & MIT)

Permutations and Posets
for Causal Structure Discovery:
Application to Drug Repurposing for COVID-19

Caroline Uhler

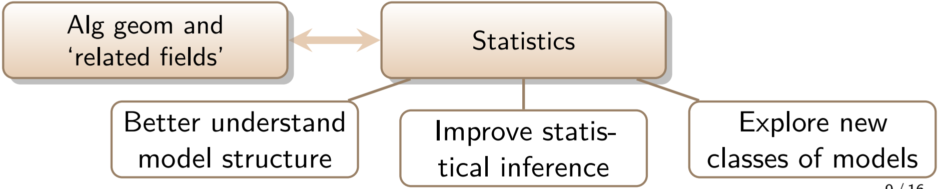
Caroline Uhler (MIT) From Causal Inference to Gene Regulation MIT, January 2019 1 / 17

What *is* algebraic statistics?

Fact [Guiding principle]

Many statistical models are defined by (semi-)algebraic sets of parameters.

- The geometry of these parameter spaces determines the behavior of widely used statistical inference procedures.
- "Shape" of a statistical model: intuitive notion of fundamental importance to statistical inference.
  - E.g: is the likelihood function multimodal?
  - Does the model have singularities (is non-regular)?



Alg geom and ‘related fields’

Statistics

Better understand model structure

Improve statistical inference

Explore new classes of models

Probability / statistics

- Probability distribution
- Statistical model
- (Discrete) exponential family
- Conditional inference
- Maximum likelihood estimation
- model selection
- Multivariate Gaussian model
- Phylogenetic model
- MAP estimates

Algebra/geometry

- Point
- (Semi)algebraic set
- Toric variety / ideal
- Lattice points in polytopes
- Polynomial optimization
- Geometry of singularities
- Spectrahedral geometry
- Tensor networks
- Tropical geometry

Lecture plan

We will continue now with the following topics:

- Probability Primer (Chapter 2) and
- Conditional Independence (Chapter 4)

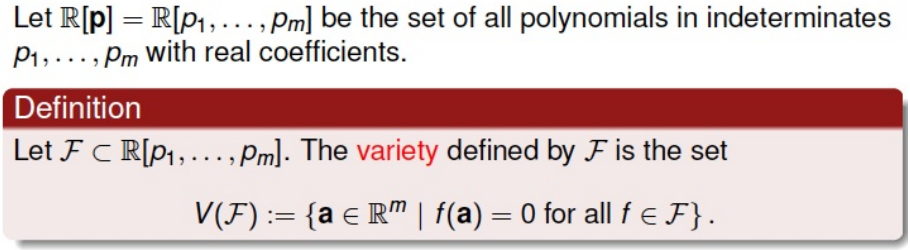
License

This document is created for Math/Stat 561, Spring 2023, at Illinois Tech.

Appendix

Following is a 3-slide "intro" to algebraic geometry; these were slides by S. Sullivant given at a colloquium a long long time ago. They are meant to just give you a glimpse into the vocabulary... not to digest this immediately.

Introduction to algebraic geometry

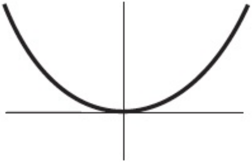


Let $\mathbb{R}[\mathbf{p}] = \mathbb{R}[p_1, \ldots, p_m]$ be the set of all polynomials in indeterminates $p_1, \ldots, p_m$ with real coefficients.

Definition

Let $\mathcal{F} \subset \mathbb{R}[p_1, \ldots, p_m]$. The variety defined by $\mathcal{F}$ is the set

$$V(\mathcal{F}) := \{\mathbf{a} \in \mathbb{R}^m \mid f(\mathbf{a}) = 0 \text{ for all } f \in \mathcal{F}\}.$$

$V(\{p_2 - p_1^2\}) =$ 

Definition

Given $S \subset \mathbb{R}^m$

$$\mathcal{I}(S) := \{f \in \mathbb{R}[\mathbf{p}] \mid f(\mathbf{a}) = 0 \text{ for all } \mathbf{a} \in S\}$$

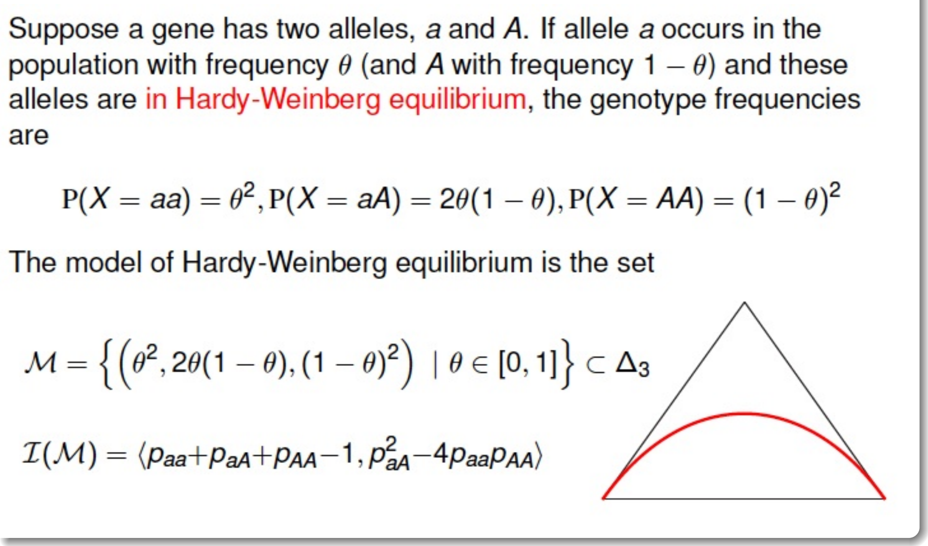
is the vanishing ideal of S.

Theorem (Hilbert's Basis Theorem)

Every ideal $I \subset \mathbb{R}[\mathbf{p}]$ has a finite generating set. That is, for each ideal I, there exists a finite set $\mathcal{F} \subset I$ such that $\langle \mathcal{F} \rangle = I$, where

$$\langle \mathcal{F} \rangle := \left\{ \sum h_i f_i \mid h_i \in \mathbb{R}[\mathbf{p}], f_i \in \mathcal{F} \right\}.$$

Example: Hardy-Weinberg Equilibrium



Suppose a gene has two alleles, *a* and *A*. If allele *a* occurs in the population with frequency θ (and *A* with frequency $1-\theta$) and these alleles are in Hardy-Weinberg equilibrium, the genotype frequencies are

$$\mathrm{P}(X = aa) = \theta^2, \mathrm{P}(X = aA) = 2\theta(1-\theta), \mathrm{P}(X = AA) = (1-\theta)^2$$

The model of Hardy-Weinberg equilibrium is the set

$$\mathcal{M} = \left\{ \left(\theta^2, 2\theta(1-\theta), (1-\theta)^2 \right) \mid \theta \in [0,1] \right\} \subset \Delta_3$$

$$\mathcal{I}(\mathcal{M}) = \langle p_{aa} + p_{aA} + p_{AA} - 1, p_{aA}^2 - 4 p_{aa} p_{AA} \rangle$$