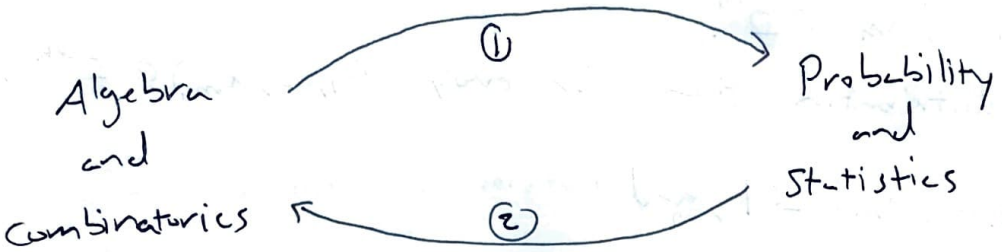
Combinatorial and Algebraic Statistics VT21

Lesson 1: What is Algebraic Statistics?

- "Algebraic Statistics" primarily describes the field of study that emerged in the 1990's centered on the connection between algebraic geometry and probability and statistics. Over the decades the "algebraic geometry" bit has become broadly interpreted to include commutative algebra and combinatorics.



① How can algebra and combinatorics be useful in probability and statistics?

② How can ideas from probability and statistics drive new research in algebra and combinatorics?

- Today, we will go through a first example, analyzing how algebra and statistics arise when thinking about statistical problems.
- We will try and note some of the "Big Ideas" that make up the popular research initiatives in the field. A key goal of this course is to give you a solid intro to these popular topics in algebraic statistics.

A First Example

- Let $X_1, \dots, X_n$ be random variables on the (finite) state space Ω.

Ex. Let $\Omega = \{0,1\}$. X_i is a function $X_i : \Omega \to \mathbb{R}$ such that there is a function $P : 2^{\Omega} \to [0,1]$ such that for every measurable subset $B \subseteq \mathbb{R}$

$P(X_i^{-1}(B))$ = the probability of the outcomes in $X_i^{-1}(B)$.

If $X_i = id$ then B contains either

- only 0 $\longrightarrow P(X_i^{-1}(B)) = P(0)$
- only 1 $\longrightarrow = P(1)$
- both 1 and 0 $\longrightarrow = P(1 \text{ or } 0) = P(0) + P(1) = 1$
- neither $\longrightarrow = P(\text{neither}) = 0.$

— P is the probability distribution associated to X_i.

- The distributions of each X_i naturally induce a distribution on all of them considered together.
- The sequence $(X_1, \dots, X_m)$ has outcomes $\Omega^m = \{(1,0,1,1\dots,1,0), \dots\}$ and and associated probability distribution P such that

$$P(X_1 = x_1, \dots, X_m = x_m) = \text{probability of } X_1 = x_1, \dots, \text{ and } X_m = x_m$$

for every $(x_1, \dots, x_m) \in \Omega^m$.

- Since P a distribution then for every $(x_1, \dots, x_m) \in \Omega^m$:

(1) $0 \leq P(x_1, \dots, x_m) \leq 1$, and

(2) $\sum_{(x_1,\dots,x_m) \in \Omega^m} P(x_1, \dots, x_m) = 1.$

Definition ① $(X_1, \dots, X_m)$ is called a <u>Markov chain</u> if for all $i = 3, \dots, m$ and all $(x_1, \dots, x_i) \in \Omega^i$

$$P(X_i = x_i \mid X_1 = x_1, \dots, X_{i-1} = x_{i-1}) = P(X_i = x_i \mid X_{i-1} = x_{i-1}).$$

That is, $X_i \perp\!\!\!\perp X_{\{1,\dots,i-2\}} \mid X_{i-1}$.

- In words: the probability of the outcome $X_i = x_i$ given every preceding outcome depends only on the immediately preceding outcome. In stats, we say, "X_i is (conditionally) independent of $X_1, \dots, X_{i-2}$ given X_{i-1}."
- We represent a Markov chain with a directed graph

$$I_m = (1) \longrightarrow (2) \longrightarrow (3) \longrightarrow \cdots \longrightarrow (m-1) \longrightarrow (m)$$

- info coming from $(1), (2), (3), \dots, (i-2)$ must pass through $(i-1)$ to get to (i)...

Big Idea (Graphical Models) The combinatorial structure of graphs can be used to capture important information (e.g. conditional independence relations) that hold among random variables.

Definition ② A statistical model is any subset of distributions that could be associated to a set of random variables.

- Ex. The model $M(I_m) := \{$all distributions satisfying Def ① for $X_1, \dots, X_m\}$.

- Like many popular statistical models, the underlying description of $M(I_m)$ is inherently algebraic.

- Fix $m=3$ and $\Omega = \{0,1\}$.

- Any distribution P for X_1, X_2, X_3 is a point in $\mathbb{R}^8$:

$$P_{011} = P(X_1=0, X_2=1, X_3=1) \in \mathbb{R}.$$

$P = (p_{000}, p_{100}, p_{010}, p_{001}, p_{110}, p_{101}, p_{011}) \in \mathbb{R}^8$.

- By (1) and (2)

$$P \in \Delta_{|\Omega^m|-1} = \{p \in \mathbb{R}^8 \mid 0 \le p_i \le 1, \ \textstyle\sum_{i \in \Omega^m} p_i = 1\}$$

(The $(|\Omega^m|-1)$-dim'l probability simplex)

$\Rightarrow M(I_3) \subset \Delta_7 \subset \mathbb{R}^8$.

- Let $\sum_{k \in \Omega = \{0,1\}} p_{ijk} =: p_{ij+}$. By def of conditional probability:

$$P(X_3 = k \mid X_1 = i, X_2 = j) = \frac{p_{ijk}}{p_{ij+}}.$$

- Def ① $\Rightarrow \forall i,j,k \in \{0,1\}$,

$$\frac{p_{ijk}}{p_{ij+}} = \frac{p_{+jk}}{p_{+j+}}, \text{ or equivalently for all } i,i',j,k \in \{0,1\}, \ \frac{p_{ijk}}{p_{ij+}} = \frac{p_{i'jk}}{p_{i'j+}}.$$

- Clearly denominators yields a purely algebraic description of $M(I_3)$:

$$P \in M(I_3) \iff \text{(1) } p_{ijk} \ge 0 \ \forall ijk \in \{0,1\}$$
$$\text{(2) } \textstyle\sum_{i,j,k \in \{0,1\}} p_{ijk} = 1$$
$$\text{(3) } p_{000}p_{101} - p_{001}p_{100} = 0$$
$$\text{(4) } p_{010}p_{111} - p_{011}p_{110} = 0 \quad \text{(★)}$$

- $M(I_3)$ is a semialgebraic set; i.e. a collection of points satisfying a system of polynomial equations and inequalities.

Big Idea: Which popular statistical models are representable as "nice" semialgebraic sets? What does this "nice" structure capture about our model?

- $\mathcal{M}(\mathcal{I}_m)$ is an example of a conditional independence model as it is defined by constraints of the form $X_A \perp\!\!\!\perp X_B \mid X_C$.

$$\text{(★)} \Rightarrow \mathcal{M}(\mathcal{I}_3) = \Delta_7 \cap \underbrace{V(p_{000}p_{101} - p_{001}p_{100},\ p_{010}p_{111} - p_{011}p_{110})}_{\text{a conditional independence variety.}}$$

"vanishing set of..."

Big Idea What is the algebraic structure of a given conditional independence (ideal or) variety?

- While many useful statistical models turn out to be conditional independence models, it is most common in statistics to define a model parametrically. The standard parametrization of $\mathcal{M}(\mathcal{I}_m)$ follows from the chain rule in probability.

The Chain Rule:

$$P(X_1 = x_1, \dots, X_m = x_m) = \prod_{i \in [m]} P(X_i = x_i \mid X_1 = x_1, \dots, X_{i-1} = x_{i-1})$$

$$= \prod_{i \in [m]} P(X_i = x_i \mid X_{i-1} = x_{i-1}) \quad \text{(By Def ①)}$$

- For $\mathcal{M}(\mathcal{I}_3)$: $P(X_1 = i, X_2 = j, X_3 = k) = \alpha_i \beta_{ji} \gamma_{kj}$ where

$\alpha_i = P(X_1 = i)$, $\beta_{ji} = P(X_2 = j \mid X_1 = i)$ and $\gamma_{kj} = P(X_3 = k \mid X_2 = j)$.

Big Idea Many statistical models have parametric representations in terms of polynomial functions of the parameters. What are their algebraic properties?

- The algebraic structure of a statistical model translates into algebraic answers to statistical questions when we start to incorporate data.

Statistical Inference Questions:

① How can we fit a statistical model to given data?

② Does the model fit the given data well?

Starting (frequentist) Assumption S'pose there is some "true" unknown distribution $\mathbb{P}$ from which all our data are independently drawn samples.

- Example: For $\mathcal{M}(I_3)$ we could have data

$\mathbb{D} = \{000, 010, 110, 000, 101, 110, 100, 010, 110, 111, 000, 000, 010\}$

13 independent and identically distributed samples (iid).

- iid assumption $\Rightarrow$ probability of observing $\mathbb{D}$ really only depends on the vector of counts: $\underline{u} = (u_{ijk} \mid ijj,k \in \{0,1\}) \in \mathbb{R}^8$, where

$u_{ijk} :=$ # of ijk's in $\mathbb{D}$.

- e.g. $\underline{u} = (\underset{000}{4}, \underset{001}{0}, \underset{010}{3}, \underset{100}{1}, \underset{011}{0}, \underset{101}{1}, \underset{110}{3}, \underset{111}{1})$.

- The likelihood function records the probability of observing the data given the model (parameters):

$$L(p \mid \underline{u}) := P(\mathbb{D}) = \underbrace{\frac{n!}{\prod_{ijk} u_{ijk}!}}_{\binom{n}{\underline{u}}} \prod_{ijk \in \{0,1\}^3} p_{ijk}^{u_{ijk}},$$

$\binom{n}{\underline{u}}$ ← summing over all possible orderings of $\mathbb{D}$.

where $n = |\mathbb{D}|$. Here $\underline{u}$ is given and p is unknown...

- To find $\hat{p}$ that maximizes the likelihood of seeing $\mathbb{D}$ (i.e. the maximum likelihood estimate (MLE)) we solve

$$\arg\max L(p \mid \underline{u}) \quad \text{subject to} \quad p \in \mathcal{M}(I_3), \text{ or}$$

$$\arg\max \prod_{(ijj,k) \in \{0,1\}^3} (\alpha_i \beta_{ji} \gamma_{kj})^{u_{ijk}} \quad \text{subject to} \quad \alpha_i, \beta_{ji}, \gamma_{kj} \geq 0,$$
$$\sum_i \alpha_i = 1, \ \sum_j \beta_{ji} = 1, \text{ and } \sum_k \gamma_{kj} = 1.$$

- Equivalently, we can solve the same problem for $\log L(p \mid \underline{u})$.

- To do so, we take the partial derivatives of and set equal to 0.

$\hat{p}$ is the solution to a system of polynomial equations.

(the score (or critical) equations)

- For $\mathcal{M}(I_3)$ this solution is unique!

$$\hat{p} = \left(\hat{p}_{ijk} = \frac{u_{ij+}}{u_{+++}} \cdot \frac{u_{+jk}}{u_{+j+}} \;\middle|\; ijk \in \{0,1\}^3 \right).$$

Big Idea For which statistical models do the score equations admit a unique solution (ML-degree 1)? When does this solution admit a closed-form, rational expression as above?

Testing Goodness-of-fit

- Our second main statistical inference question is whether we can reject the hypothesis:

H_0: the data-generating distribution is in $\mathcal{M}(I_3)$.

Question If we assume the data-generating distribution $P \in \mathcal{M}(I_3)$, among all data sets $\mathbb{D}'$ that could have been generated, what proportion of such $\mathbb{D}'$ are more likely to be observed than $\mathbb{D}$?

→ If proportion large, reject H_0.

↳ depends on unknown distribution P ⇒ can't compute! ☹

Observation The value of $L(P|\underline{u})$ depends only on some lower-order functions of $\mathbb{D}$ (sufficient statistics):

$$L(P|\underline{u}) = P(\underline{u} \mid \alpha_i, \beta_{ji}, \gamma_{kj}) = \binom{n}{\underline{u}} \prod_{i,j,k \in \{0,1\}^3} (\alpha_i \beta_{ji} \gamma_{kj})^{u_{ijk}}$$

$$= \binom{n}{\underline{u}} \prod_{i,j \in \{0,1\}^2} (\alpha_i \beta_{ji})^{u_{ij+}} \prod_{k,j \in \{0,1\}^2} (\gamma_{kj})^{u_{+jk}}.$$

sufficient statistics $= \{u_{ij+}, u_{+jk}\}$.

Fisher's Question Among all $\mathbb{D}'$ with the same sufficient statistics as $\mathbb{D}$, what proportion of such $\mathbb{D}'$ are more likely to occur than $\mathbb{D}$?

$$P(\underline{v} \mid \alpha_i, \beta_{ji}, \gamma_{kj}, u_{ij+} = v_{ij+}, u_{+jk} = v_{+jk}) = \frac{\binom{n}{\underline{v}} \prod_{ij} (\alpha_i \beta_{ji})^{u_{ij+}} \prod_{jk} \gamma_{kj}^{u_{+jk}}}{\sum_{\underline{v} :} \binom{n}{\underline{v}} \prod_{ij} (\alpha_i \beta_{ji})^{u_{ij+}} \prod_{jk} \gamma_{kj}^{u_{+jk}}}$$

$$= \frac{\binom{n}{\underline{v}}}{\sum_{\underline{v}} \binom{n}{\underline{v}}}.$$

- Define the convex polytope

$$\mathcal{P}(\underline{u}) := \left\{ \underline{v} \in \mathbb{R}^{\mathcal{I}} \;\middle|\; v_{ij+} = u_{ij+},\ v_{+jk} = u_{+jk},\ v_{ijk} \geq 0 \right\}.$$

- The proportion of interest (Fisher's p-value) is:

$$p = \frac{\sum_{\underline{v} \in \mathcal{P}(\underline{u}) \cap \mathbb{Z}^{\mathcal{I}}} \binom{n}{\underline{v}} \mathbb{1}\left(\binom{n}{\underline{v}} > \binom{n}{\underline{u}}\right)}{\sum_{\underline{v} \in \mathcal{P}(\underline{u}) \cap \mathbb{Z}^{\mathcal{I}}} \binom{n}{\underline{v}}}.$$

Big Idea Testing goodness-of-fit (via Fisher's exact test) amounts to enumerating lattice points in the convex polytope $\mathcal{P}(\underline{u})$.

Next time... Kathlén will give us a review of some of the key algebraic concepts needed in this course. In the third lecture, I will do the same for probability and statistics. After that, the lectures will be focused on exploring the "big ideas" from today.

These lectures you will give: each person will be assigned a lecture and given associated reading. They will give their lecture to one of us the monday before their friday lecture.