ALGEBRAIC STATISTICS

SABRINA MI

Abstract. In this article, we review the basics of Algebraic Geometry and toric varieties, and their relation to certain kinds of statistical models. We explain how to use these algebraic tools to describe a random walk on a set of tables sharing the same statistical results. Lastly, we discuss how the same ideals that define toric varieties also show that every statistical model has a basis that connects observed data to related data.

Contents

1. Introduction

The Algebraic Geometry dictionary provides a translation between ideals and varieties. Geometric information about a variety can provide algebraic meaning through its corresponding ideal, and an ideal can provide geometric information about its variety. This will be the topic of Section 2.

In Section 3, we will introduce cones and their associated toric varieties through several intermediate structures, including dual cones, monoids, and coordinate rings. The generators of the monoid S_σ, associated to a rational cone σ, also provide the binomial relations of an ideal in the polynomial ring over $\mathbb{C}$. Then, an affine toric variety associated to a cone is defined by these binomials.

In Section 4 we will introduce Markov bases. They are an important link between Commutative Algebra and Statistics. A Markov basis is a set of vectors that allows us to generate synthetic tables based off an observed table. The resulting sample of tables gives evidence for or against a proposed statistical hypothesis. The Metropolis-Hastings Algorithm describes a random walk to test a sequence of potential tables connected to the observed table through moves in the Markov basis. In addition, the moves in a Markov basis can be decomposed to define an ideal generated by binomials. The toric ideal of a matrix associated with a log-linear model is also generated by binomials. The Fundamental Theorem of Markov Bases, which states that these ideals are identical, ensures that there exists a Markov basis for every statistical model.

2. Generalities on Coordinate Rings

The collection of all polynomials over a field k forms a ring, a set with addition and multiplication.

Definition 2.1. A *polynomial* f in n variables $x_1, \ldots, x_n$ with coefficients in k is the sum of a finite number of monomials of the form $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$:

$$f = \sum_\alpha \lambda_\alpha x^\alpha, \lambda_\alpha \in k.$$

The set of all polynomials in $x_1, \ldots, x_n$ with coefficients in k is the polynomial ring $k[x] = k[x_1, \ldots, x_n]$

Definition 2.2. The n-dimensional *affine space* over k is the set

$$k^n = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \in k\}$$

So each polynomial $f \in k[x]$ gives a function $f : k^n \to k$. Take polynomials $f_1, \ldots, f_t$ in $k[x]$. This set defines an affine variety in k^n.

Definition 2.3. The *affine variety* defined by $f_1, \ldots, f_t$ is

$$V(f_1, \ldots, f_t) = \{(a_1, \ldots, a_n) \in k^n \mid f_i(a_1, \ldots, a_n) = 0,\ i = 1, \ldots, t\}.$$

Next, we introduce ideals, a special subset of a ring.

Definition 2.4. A subset $I \subset k[x]$ is an *ideal* if it satisfies

(1) $0 \in I$
(2) If $f, g \in I$, then $f + g \in I$
(3) If $f \in I$ and $h \in k[x]$, then $hf \in I$

One example is the ideal generated by a finite number of polynomials:

Definition 2.5. Let $f_1, \ldots, f_s \in k[x]$. Then we define the ideal generated by $f_1, \ldots, f_s$

$$\langle f_1, \ldots, f_s \rangle = \{\textstyle\sum_{i=1}^s h_i f_i \mid h_1, \ldots, h_s \in k[x]\}$$

$\langle f_1, \ldots, f_s \rangle$ is an ideal because if $f = \sum_{i=1}^s p_i f_i$, $g = \sum_{i=1}^s q_i f_i$, and $h \in k[x]$, it satisfies:

(1) $0 = \sum_{i=1}^s 0 \cdot f_i \in \langle f_1, \ldots, f_s \rangle$
(2) $f + g = \sum_{i=1}^s (p_i + q_i) f_i \in \langle f_1, \ldots, f_s \rangle$
(3) $hf = \sum_{i=1}^s (hp_i) f_i \in \langle f_1, \ldots, f_s \rangle$

Now, we can consider the set of all polynomials that vanish on a given variety.

Definition 2.6. Let $V \subset k^n$ be an affine variety. We define the ideal of V

$$I(V) = \{f \in k[x] \mid f(a_1, \ldots, a_n) = 0,\ (a_1, \ldots, a_n) \in V\}$$

We can confirm $I(V)$ is an ideal. Let $f, g \in I(V)$ and $h \in k[x]$. By definition, $f(a_1, \ldots, a_n) = g(a_1, \ldots, a_n) = 0$ for all $(a_1, \ldots, a_n)$ in V.

(1) $0 \in I(V)$
(2) $f + g \in I(V)$, since $f(a_1, \ldots, a_n) + g(a_1, \ldots, a_n) = 0$ for $(a_1, \ldots, a_n) \in V$
(3) $hf \in I(V)$, since $h(a_1, \ldots, a_n) f(a_1, \ldots, a_n) = 0$ for $(a_1, \ldots, a_n) \in V$

Now, we have a relationship between polynomials, varieties, and ideals: $f_1, \ldots, f_s \to V(f_1, \ldots, f_s) \to I(V(f_1, \ldots, f_s))$. The map I also has the property:

Proposition 2.7. *Let V and W be affine varieties in k^n. Then*

(1) $V \subset W$ *if and only if* $I(V) \supset I(W)$
(2) $V = W$ *if and only if* $I(V) = I(W)$

We also have an assignment V, from ideals to varieties, as a consequence of the following:

Theorem 2.8. (Hilbert Basis Theorem) *Every ideal* $I \subset k[x]$ *has a finite generating set, so that* $I = \langle g_1, \ldots, g_t \rangle$ *for some* $g_1, \ldots, g_t \in I$.

In geometric terms, this theorem says that any algebraic set is the set of common roots of finitely many polynomial equations.

Corollary 2.9. *Let* $I \subset k[x]$ *be a ideal. Then it is generated by some set* $f_1, \ldots, f_s \subset I$, *so that* $V(I) = V(f_1, \ldots, f_s)$ *is an affine variety.*

The variety defined by a set depends only on ideals. The Algebraic Geometry dictionary connects ideals to varieties, so that an algebraic analysis of an ideal can be translated geometrically in the variety. The relationship between polynomials, varieties, and ideals is clarified in the Nullstellensatz.

Theorem 2.10. (The Weak Nullstellensatz) *Let* k *be an algebraically closed field, and* $I \subset k[x]$ *be an ideal such that* $V(I) = \emptyset$. *Then* $I = k[x]$.

The Weak Nullstellensatz is consistent with the inclusion-reversing relation between ideals and varieties (**Proposition 2.7**), so that the smallest possible variety is associated with the largest possible ideal. The Nullstellensatz illustrates that every maximal ideal corresponds to a point in affine space.

Definition 2.11. An ideal I in a ring R is a *maximal ideal* if for every ideal J such that $I \subseteq J$, $J = I$ or $J = R$.

Let E be a set of polynomial equations in $k[x]$ such that the set of common roots is a single point, $a = (a_1, \ldots, a_n) \in k^n$, $V(E) = a$. Then its ideal is $I(a) = \langle x_1 - a_1, \ldots, x_n - a_n \rangle$. $I(a)$ is a maximal ideal, which we denote M_a.

Theorem 2.12. (Hilbert's Nullstellensatz)*Let* k *be an algebraically closed field. Every maximal ideal in* $k[x]$ *can be expressed* $M_a = \langle x_1 - a_1, \ldots, x_n - a_n \rangle$ *for some point* $a \in k^n$.

Now we have a relationship between points and maximal ideals through the function $M(a) = M_a$.

Corollary 2.13. *The correspondence* M *is a bijection between points in* k^n *and maximal ideals* M *of* $k[x]$.

$$M : k^n \to \{M \subset k[x] \mid M \text{ maximal ideal}\}$$
$$M(a) = M_a,\ a \in k^n$$

We can also learn about the geometric properties of a variety V by restricting the polynomial functions. The set of all polynomial functions $f : V \to k$ is denoted $k[V]$. It can be constructed as a quotient ring. To start, we define an equivalence relation, $\sim$, to describe if two polynomial functions are the same function on $V = V(f_1, \ldots, f_s)$.

Definition 2.14. Let $V \subset k^n$ be an affine variety, $f, g \in k[x]$. $f \sim g$ if $f - g \in I(V)$

If $f \sim g$, for all $a \in V$, $f(a) - g(a) = 0$. f and g represent the same polynomial function on V.

Definition 2.15. The *equivalence class* of the element f in $k[x]$ under this equivalence relation is denoted $[f] = \{f + h \mid h \in I(V)\}$. The set of all equivalence classes forms the quotient ring $k[x]/I(V)$.

In other words, $k[x_1, \ldots, x_n]/I(V)$ describes the set of distinct polynomials on V. So the coordinate ring $k[V] \cong k[x_1, \ldots, x_n]/I(V)$.

For example, let $f(x, y) = x^3 - y^2$, with $V = V(f) = \{(x, y) \in k^2 \mid x^3 - y^2 = 0\}$. Then $k[x, y]/\langle x^3 - y^2 \rangle$ is the ring of functions on V.

Definition 2.16. We call $k[V] = k[x_1, \ldots, x_n]/I(V)$ the *affine coordinate ring* of V. It is a k-algebra generated by the equivalence classes $[x_1], \ldots, [x_n]$.

Definition 2.17. A *finitely generated k-algebra* R is a commutative structure with addition, multiplication, and scalar multiplication such that there exists a finite set $x_1, \ldots, x_n \in R$ such that every element $f \in R$ can be expressed as a polynomial $f = \sum_\alpha \lambda_\alpha x^\alpha,\ \lambda_\alpha \in k$.

The correspondence between finitely generated algebras and affine varieties is strengthened in the Algebraic Geometry dictionary, which offers an alternate definition:

Theorem 2.18. *R is a finitely generated k-algebra if and only if it is isomorphic to a quotient ring of the of the form $k[x_1, \ldots, x_n]/I$, where I is an ideal in $k[x_1, \ldots, x_n]$*

The affine coordinate ring $k[V] = k[x_1, \ldots, x_n]/I(V)$ allows us to generalize **Corollary 2.13**:

Corollary 2.19. *There is a one-to-one correspondence between points in a variety V and maximal ideals of the coordinate ring $k[V]$.*

3. Affine Toric Varieties

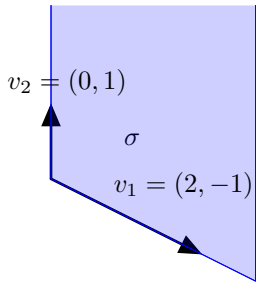
The intersection of commutative algebra and statistics can be studied with affine toric varieties. From a cone σ, we can associate the dual cone $\check{\sigma}$, then a monoid S_σ, coordinate ring R_σ, and lastly, the toric variety X_σ. The ideals that represent R_σ will be important later, when discussing Markov bases.

Definition 3.1. Let $A = \{v_1, \ldots, v_r\}$ be a finite set of vectors in $\mathbb{R}^n$. Then the set

$$\sigma = \{x \in \mathbb{R}^n \mid x = \lambda_1 v_1 + \cdots + \lambda_r v_r, \lambda_i \in \mathbb{R},\ \lambda_i \geq 0\}$$

is called a *polyhedral cone*.

$A = \emptyset$ generates the zero cone, $\sigma = \{0\}$. For another example, consider the cone generated by $v_1 = 2e_1 - e_2$ and $v_2 = e_2$.



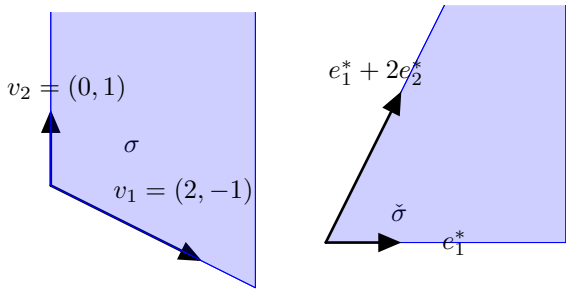
Each cone σ can be associated to a dual cone in $(\mathbb{R}^n)^*$, the set of all linear maps $f : \mathbb{R}^n \to \mathbb{R}$. $(\mathbb{R}^n)^*$ is the dual space of $\mathbb{R}^n$ with $\langle u, v\rangle$ as the dual pairing between $v \in \mathbb{R}^n$ and $u \in (\mathbb{R})^*$. We can see that $\mathbb{R}^n$ and $(\mathbb{R}^n)^*$ are isomorphic through the standard basis $\{e_1, \ldots, e_n\}$ and standard dual basis $\{e_1^*, \ldots, e_n^*\}$, where

$$\langle e_i^*, e_j\rangle = \begin{cases} 1 & i = j \\ 0 & otherwise \end{cases}$$

Definition 3.2. The *dual cone* $\check{\sigma}$ associated to the cone σ is defined by

$$\check{\sigma} := \{u \in (\mathbb{R}^n)^* \mid \langle u, v\rangle \geq 0 \text{ for all } v \in \sigma\}$$

So for the cone σ with generators $v_1 = 2e_1 - e_2$ and $v_2 = e_2$, its dual cone $\check{\sigma}$ is generated by e_1^* and $e_1^* + 2e_2^*$:



Now, we introduce lattices and dual lattices. A lattice is a group isomorphic to $\mathbb{Z}^n$. Let N be a lattice, $N \cong \mathbb{Z}^n$. Then set its dual lattice $M = Hom_{\mathbb{Z}}(N; \mathbb{Z}) \cong \mathbb{Z}^n$ in $(\mathbb{R}^n)^*$. Toric varieties can be constructed when the generators of a cone are in a fixed lattice.

Definition 3.3. A cone σ is a *rational* cone if all the generators v_i are in N.

If we set $N = \mathbb{Z}^2$, our cone σ is rational since its generators $v_1 = 2e_1 - e_2$ and $v_2 = e_2$ are in N.

Definition 3.4. A cone σ is *strongly convex* if there are no lines that can be drawn through the origin that are contained in σ, or $\sigma \cap (-\sigma) = \{0\}$.

From the figure, we can see that σ is a strongly convex and rational cone. Also, $\check{\sigma}$ is a rational cone with respect to $(\mathbb{Z}^n)^*$.

If a cone is rational, then its dual cone is rational as well. However, the dual cone of a strongly convex cone is not necessarily strongly convex. For example, the dual of the zero cone in $\mathbb{R}^2$, $\sigma = \{0\}$, is $\check{\sigma} = (\mathbb{R}^2)^*$, which is not strongly convex.

Next, we can construct monoids from cones. The generators of these monoids will later be relevant when generating the ideal to represent the cones as coordinate rings.

Definition 3.5. A set S with a binary operation $+ : S \times S \to S$ is a *monoid* if it satisfies

(1) Associativity: $(a + b) + c = a + (b + c)\ \forall\ a, b, c \in S$
(2) Identity element: $0 + a = a$

$\mathbb{N}$ forms a monoid under addition, with 0 as the identity element. A monoid can also be formed from a cone and a lattice.

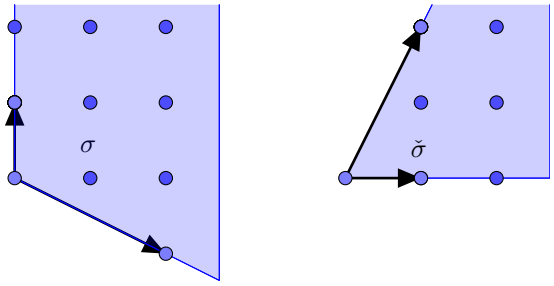
Lemma 3.6. *If σ is a cone, then $\sigma \cap N$ is a a submonoid of N.*

Note that if $x, y \in \sigma \cap N$, then $x + y \in \sigma \cap N$.

Definition 3.7. A monoid S is *finitely generated* if there are elements $a_1, \ldots, a_k \in S$ such that

$$\forall\ s \in S,\ s = \lambda_1 a_1 + \cdots + \lambda_k a_k \text{ with } \lambda_i \in \mathbb{N} \cup \{0\}$$

Elements $a_1, \ldots, a_k$ are called generators of the monoid.



The dots drawn in cones σ and $\check{\sigma}$ represent the monoids $\sigma \cap N$ and $\check{\sigma} \cap M$. $\sigma \cap N$ is generated by $\{e_1,\ e_2,\ 2e_1 - e_2\}$ and $\check{\sigma} \cap M$ is generated by $\{e_1^*,\ e_1^* + e_2^*,\ e_1^* + 2e_2^*\}$. When studying toric varieties, it is necessary to restrict to rational cones because of the following lemma:

Lemma 3.8. (Gordon's Lemma) *If σ is a polyhedral lattice cone, then $\sigma \cap N$ is a finitely generated monoid.*

To construct toric varieties from cones, we apply Gordon's lemma to the dual cone $\check{\sigma}$, focusing on the monoid $S_\sigma := \check{\sigma} \cap M$. The generators of S_σ correspond to Laurent monomials, which is important when associating a coordinate ring R_σ to the cone σ.

The variables in a Laurent monomial can have positive or negative powers. A Laurent monomial in $\mathbb{C}$ is written as $\lambda x^\alpha = \lambda x_1^{\alpha_1} \cdots x_n^{\alpha_n}$, with $\lambda \in \mathbb{C}^* = \mathbb{C} \setminus \{0\}$ and $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n$. The set of all Laurent polynomials forms a ring, which is denoted $\mathbb{C}[x, x^{-1}] = \mathbb{C}[x_1, \ldots, x_n, x_1^{-1}, \ldots, x_n^{-1}]$.

The additive group $\mathbb{Z}^n$ is isomorphic to the multiplicative group of Laurent monomials with coefficient 1 through the map:

$$\theta : \mathbb{Z}^n \to \mathbb{C}[x, x^{-1}]$$
$$\theta(\alpha) = x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n},\ \alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n$$

θ is an isomorphism because it is bijective and satisfies $\theta(a+b) = z^{a+b} = z^a \cdot z^b = \theta(a) \cdot \theta(b)$ for all $a, b \in \mathbb{Z}^n$.

Definition 3.9. Let f be a Laurent polynomial with finite terms. The *support* of a Laurent polynomial $f = \sum_{finite} \lambda_\alpha z^\alpha$ is
$$\text{supp}(f) = \{\alpha \in \mathbb{Z}^n \mid \lambda_\alpha \neq 0\}$$

Proposition 3.10. *Let σ be a rational cone.*
$$R_\sigma := \left\{ f \in \mathbb{C}[x, x^{-1}] \mid \text{supp}(f) \subset \check{\sigma} \cap M \right\}$$
is a $\mathbb{C}$-algebra finitely generated by Laurent monomials.

This follows directly from Gordon's lemma, since $\check{\sigma} \cap M$ is a finitely generated monoid. So the generators of $\check{\sigma} \cap M$ give the monomials that generate R_σ through the mapping θ. The S_σ in our example is generated by $\{e_1^*,\ e_1^* + e_2^*,\ e_1^* + 2e_2^*\}$, so the $\mathbb{C}$-algebra is generated by $\{x_1,\ x_1x_2,\ x_1x_2^2\}$.

Recall that a finitely generated $\mathbb{C}$-algebra can be written as a coordinate ring $\mathbb{C}[x_1, \ldots, x_n]/I$ for some ideal I. Depending on which set of generators of S_σ is chosen, R_σ can be represented as a coordinate ring $\mathbb{C}[x_1, \ldots, x_n]/I_\sigma$, where $I_\sigma = \langle r_1, \ldots, r_k \rangle$ for relations r_i. The relations between generators of S_σ are written $\sum_j u_j a_j = \sum_j v_j a_j$. The relations $r_1, \ldots, r_k$ allow us to assign a variety, denoted by $V(I_\sigma)$, to the cone σ.

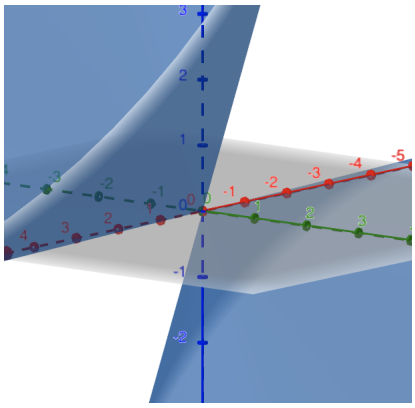
Definition 3.11. The *affine toric variety* corresponding to a rational polyhedral, strongly convex cone σ is $X_\sigma := \{M \subset R_\sigma \mid M \text{ maximal ideal}\}$.

The choice of generators of S_σ determines the representation of R_σ as a finitely generated algebra $R_\sigma = \mathbb{C}[x_1, \ldots, x_n]/I_\sigma$, and I_σ then defines the toric variety X_σ.

Theorem 3.12. *Let σ be a rational cone in $\mathbb{R}^n$ and $A = (a_1, \ldots, a_n)$ be a system of generators of S_σ. Then the corresponding toric variety X_σ is represented by the affine toric variety $V(I_\sigma) \subset \mathbb{C}^n$, where $I_\sigma \subset \mathbb{C}[x_1, \ldots, x_n]$ is generated by finitely many binomials of the form*
$$x_1^{u_1} \cdots x_n^{u_n} = x_1^{v_1} \cdots x_n^{v_n}$$
Through θ, these binomials correspond to relations between the generators of S_σ:
$$\sum_j u_j a_j = \sum_j v_j a_j$$

One consequence of the theorem is that a point $x = (x_1, \ldots, x_n) \in \mathbb{C}^n$ is a point in the variety X_σ if and only it satisfies all binomial relations
$$x_1^{u_1} \cdots x_n^{u_n} = x_1^{v_1} \cdots x_n^{v_n}.$$

Now, we will construct X_σ from our example cone generated by $v_1 = 2e_1 - e_2$ and $v_2 = e_2$, with $a_1 = e_1^*$, $a_2 = e_1^* + e_2^*$, and $a_3 = e_1^* + 2e_2^*$ set as the system of generators for S_σ. The isomorphism $\theta : \mathbb{Z}^2 \to \mathbb{C}[x_1, x_2, x_1^{-1}, x_2^{-1}]$ gives the Laurent monomials $x_1 = z_1$, $x_2 = z_1z_2$, and $x_3 = z_1z_2^2$ that generate R_σ. We have the relation $a_1 + a_3 = 2a_2$, which provides the binomial relation $x_1x_3 = x_2^2$. Then we represent $R_\sigma = \mathbb{C}[x_1, x_2, x_3]/I_\sigma$, where the ideal I_σ is generated by the equation $x_1x_3 = x_2^2$, namely $I_\sigma = \langle x_1x_3 - x_2^2 \rangle$. Then the corresponding toric variety is given by $X_\sigma = V(I_\sigma) = \{x = (x_1, x_2, x_3) \in \mathbb{C}^3 \mid x_1x_3 = x_2^2\}$. X_σ is a quadratic cone in $\mathbb{C}^3$. The real part of it in $\mathbb{R}^3$ is shown below:



Consider the cone $\sigma = \{0\}$, with dual cone $\check{\sigma} = (\mathbb{R}^n)^*$. Both systems generate S_σ:

$$A_1 = (e_1^*, \ldots, e_n^*, -e_1^*, \ldots, -e_n^*)$$
$$A_2 = (e_1^*, \ldots, e_n^*, -(e_1^* + \cdots + e_n^*))$$

If we denote the first system of generators, $A_1 = (a_1, \ldots, a_{2n})$, we have the following relation between elements of A_1:

$$a_1 + a_{n+1} = 0$$
$$\vdots$$
$$a_n + a_{2n} = 0$$

So R_σ can be written $\mathbb{C}[x_1, \ldots, x_n, x_{n+1}, \ldots, x_{2n}]/I_\sigma$, where the corresponding ideal is $I_\sigma = \langle x_1x_{n+1} - 1, x_2x_{n+2} - 1, \ldots, x_nx_{2n} - 1\rangle$. Then the toric variety is $X_\sigma = V(x_1x_{n+1} - 1, x_2x_{n+2} - 1, \ldots, x_nx_{2n} - 1)$

In addition, since R_σ is also generated by $z_1, \ldots, z_n, z_1^{-1}, \ldots, z_n^{-1}$, the ring of Laurent polynomials over $\mathbb{C}$ can also be represented as

$$\mathbb{C}[z_1, \ldots, z_n, z_1^{-1}, \ldots, z_n^{-1}] = \frac{\mathbb{C}[x_1, \ldots, x_{2n}]}{\langle x_1x_{n+1} - 1, x_2x_{n+2} - 1, \ldots, x_nx_{2n} - 1\rangle}$$

A_2 gives the relation:

$$a_1 + \cdots + a_n + a_{n+1} = 0$$

so that $R_\sigma = \mathbb{C}[x_1, \ldots, x_n, x_{n+1}]/\langle x_1 \cdots x_nx_{n+1} - 1\rangle$ and the toric variety is $X_\sigma = V(x_1 \cdots x_nx_{n+1} - 1)$.

4. Statistical Models

A contingency table shows the cross-section of observed cases between two or more discrete variables:

$$U = \begin{pmatrix} u_{11} & \cdots & u_{1c} \\ \vdots & & \vdots \\ u_{r1} & \cdots & u_{rc} \end{pmatrix}$$

where u_{ij} is the number of observations where $X_1 = i$, $X_2 = j$. For example:

Gender	Handedness		Total by Gender
	Right-handed	Left-handed	
Male	3	1	4
Female	5	1	6
Total by Handedness	8	2	10

One question about this table that could be studied is whether gender is independent of handedness. In hypothesis testing in statistics, we ask how likely it is to get the data we observed if our hypothesis that gender is independent of handedness is true.

Definition 4.1. Random variables X_1 and X_2 have *joint probabilities*

$$p_{ij} = P(X_1 = i, X_2 = j),$$

and *marginal probabilities*

$$p_{i+} := \sum_{j=1}^{c} p_{ij}, \quad p_{+j} := \sum_{i=1}^{r} p_{ij},$$

for $i \in [r], j \in [c]$

Definition 4.2. X_1 and X_2 are *independent* if $p_{ij} = p_{i+}p_{+j}$ for all $i \in [r]$, $j \in [c]$

In our example, we assumed gender and handedness were independent. This defines a statistical model. The joint probability matrix $p = (p_{ij})$ is unknown, so a statistical model $\mathcal{M}$ represents the set of all possible p. It is a subset of

$$\left\{ q \in \mathbb{R}^{r \times c} \mid q_{ij} \geq 0 \text{ and } \sum_{i=1}^{r} \sum_{j=1}^{c} q_{ij} = 1 \right\}.$$

An algebraic statistical model can be represented by the image of a polynomial map

$$f : \mathbb{R}^d \to \mathbb{R}^m$$
$$f(\theta) = (f_1(\theta), \ldots, f_m(\theta), \theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d.$$

The variables $\theta_1, \ldots, \theta_d$ represent the model parameters and $f_1(\theta), \ldots, f_m(\theta)$ make up the the probability distribution determined by the model. We require $f_1(\theta) + \cdots + f_m(\theta) = 1$ for all θ in the parameter space, and each $f_i(\theta)$ is a polynomial with finitely many terms:

$$f_i(\theta) = \textstyle\sum_{\alpha \in \mathbb{N}^d} \lambda_{i,\alpha} \theta_1^{\alpha_1} \cdots \theta_d^{\alpha_d}$$

Two important classes of models are linear models and log linear models.

Definition 4.3. An algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$ is a *linear model* if each polynomial $f_i(\theta)$ is a real linear function of the paramenters $\theta_1, \ldots, \theta_d$,

$$f_i(\theta) = \textstyle\sum_{j=1}^{d} a_{ij}\theta_j + b_i$$

One example of a linear model might involve conditional probability. Suppose there are two events A and B. Let $P(A) = \theta$ and $P(A^C) = 1 - \theta$ be unknown. If

$P(B|A) = .25$, $P(B^C|A) = .75$, $P(B|A^C) = .6$, $P(B^C|A^C) = .4$, then the statistical model can be written as a linear function of the parameter θ:

$$\begin{aligned} f_1(\theta) &:= P(B) = P(B|A)P(A) + P(B|A^C)P(A^C) \\ &= .25\theta + .6(1-\theta) \\ &= -.35\theta + .6 \\ f_2(\theta) &:= P(B^C) = P(B^C|A)P(A) + P(B^C|A^C)P(A^C) \\ &= .75\theta + .4(1-\theta) \\ &= .35\theta + .4 \end{aligned}$$

On the other hand, the coordinate functions f_i in a log linear model are monomials, so that $\log(f_i(\theta))$ is a linear function of $\log(\theta_1), \ldots, \log(\theta_d)$. The independence model is a type of log linear model.

Let $A = (a_{ij})$ be a non-negative integer $d \times m$ matrix with all columns summing to the same value

$$\sum_{i=1}^{d} a_{i1} = \cdots = \sum_{i=1}^{d} a_{im}.$$

Each column a_j of A determines a monomial

$$\theta^{a_j} = \prod_{i=1}^{d} \theta_i^{a_{ij}}$$

Definition 4.4. An algebraic statistical model $f : \mathbb{R}^d \to \mathbb{R}^m$ is a *log linear model* if each $f_j(\theta)$ has the form:

$$f_j(\theta) = \frac{1}{\sum_{i=1}^{m} \theta^{a_i}} \cdot \theta^{a_j}$$

Scaling by $\sum_{i=1}^{m} \theta^{a_i}$ ensures that $f_1(\theta) + \cdots + f_m(\theta) = 1$. Taking the logarithm of both sides, we have the linear relationship:

$$\log(f_j(\theta)) = \textstyle\sum_{i=1}^{d} a_{ij} \log(\theta_i)$$

The independence model is a log linear model because each joint probability $p_{ij} = p_{i+}p_{+j}$, so that taking the logarithm gives the linear function

$$\log(p_{ij}) = \log(p_{i+}) + \log(p_{+j}).$$

Statistical hypotheses about contingency tables can be tested by performing random walks on a constrained set of tables with non-negative integer entries. Markov bases are useful because they ensure that the random walk connects every pair of tables in the considered set.

The following matrix A represents the model of independence.

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

It satisfies the identity

$$Au = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_{11} \\ u_{12} \\ u_{21} \\ u_{22} \end{pmatrix} = \begin{pmatrix} u_{1+} \\ u_{2+} \\ u_{+1} \\ u_{+2} \end{pmatrix}$$

Observe that the product Au has the marginal counts. We say the marginal counts make up the sufficient statistics for the independence model because knowing the probability of observing a contingency table u means knowing the marginal probabilities p_{i+} and p_{+j}. Under the assumptions in the independence model, the marginal counts u_{i+} and u_{+j} provide sufficient information.

We want to generate a set of tables v that has the same sufficient statistics, or that

$$Au = Av = \begin{pmatrix} 4 \\ 6 \\ 8 \\ 2 \end{pmatrix}$$

Definition 4.5. The set of tables

$$\mathcal{F}(u) = \{v \in \mathbb{N} \,|\, Av = Au\}$$

is called the *fiber* of a contingency table u.

Markov bases help generate a sequence of tables v, such that $Au = Av$, and allow us to create a Markov chain to sample from the fiber. Moves in the Markov basis connect the fiber. The Metropolis-Hastings Algorithm draws on running a random walk from u as a starting point.

Definition 4.6. Let $\mathcal{M}_A$ be the log-linear model associated with matrix A. A finite subset $\mathcal{B} \subset \ker_{\mathbb{Z}}(A)$ is a *Markov basis* for $\mathcal{M}$ if for all u and all pairs v, $v' \in \mathcal{F}(u)$ there exists a sequence $u_1, \ldots, u_L \in \mathcal{B}$ such that

$$v' = v + \textstyle\sum_{k=1}^{L} u_k \text{ and } v + \sum_{k=1}^{l} u_k \geq 0 \text{ for all } l = 1, \ldots, L$$

The Markov basis represents the set of possible moves from the current contingency table to the next so that they produce the same results under the model. We know a Markov basis $\mathcal{B}$ is in $\ker_{\mathbb{Z}}(A)$, because if we consider a single step $v = v' + u$, $u \in \mathcal{B}$, it follows from $Av = Av'$ that $u = v' - v \in \ker_{\mathbb{Z}}(A)$.

For the example with gender and handedness, where we assumed independence, we have a useful proposition to find its Markov basis.

Proposition 4.7. *The unique minimal Markov basis $\mathcal{B}$ for an independence model has $2 \cdot \binom{r}{2}\binom{c}{2}$ moves, and*

$$\mathcal{B} = \{\pm(e_{ij} + e_{kl} - e_{il} - e_{kj}) \mid 1 \leq i < k \leq r,\ 1 \leq j < l \leq c\}$$

So we have these two moves for our model:

$$\mathcal{B} = \left\{ \pm \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right\}$$

We can confirm that $\begin{pmatrix} 2 & 2 \\ 6 & 0 \end{pmatrix}$ and $\begin{pmatrix} 4 & 0 \\ 4 & 2 \end{pmatrix}$ produce the same statistics as the observed table $\begin{pmatrix} 3 & 1 \\ 5 & 1 \end{pmatrix}$.

For tables with larger fibers where sampling is difficult, we can use the Metropolis-Hastings Algorithm. Given a contingency table u and a Markov basis $\mathcal{B}$, it generates a sequence of tables $\{v_t\}_{t=1}^{\infty}$ in $\mathcal{F}(u)$ and their chi-square statistic values.

Start with $v_1 = u$, $t = 1$. Select at random $u_t \in \mathcal{B}$. If any entry in $v_t + u_t$ is negative, reject the table, keeping $v_{t+1} = v_t$. Otherwise, move so that $v_{t+1} = v_t + u_t$

with probability q or keep $v_{t+1} = v_t$ with probability $q - 1$, where

$$q = \min\left\{1, \frac{P(U = v_t + u_t \mid AU = Au)}{P(U = v_t \mid AU = Au)}\right\}.$$

We repeat for $t = 2, \ldots$ and compare $\chi^2(u)$ to every $\chi^2(v_t)$. So if we move to a candidate table that is more probable than the current table, it is set as the next table in the sequence, which will converge. The hypothesis is rejected if the observed $\chi^2(u)$ is very unlikely.

We can also take the moves in a Markov basis and construct an ideal to help determine if two tables u, v in the fiber are connected. Let $\mathcal{M}$ be any finite set in $\ker_{\mathbb{Z}}(A)$. Each $m \in \mathcal{M}$ can be decomposed into positive and negative components, $m = m^+ - m^-$, where each index $m_i^+ = \max\{m_i, 0\}$ and $m_i^- = \max\{-m_i, 0\}$. Then we can define a binomial $x^{m^+} - x^{m^-}$ in polynomial ring $k[x]$. So for the set $\mathcal{M}$ we define the ideal

$$I_{\mathcal{M}} := \langle x^{m^+} - x^{m^-} \mid m \in \mathcal{M} \rangle$$

Both $I_{\mathcal{M}}$ and the ideal I_σ that determines the toric variety X_σ are generated by binomials.

Proposition 4.8. *There exists a nonnegative walk $v_1, \ldots, v_t$ between v and v' if and only if $x^u - x^v \in I_{\mathcal{M}}$.*

If we take a sequence $v_1, \ldots, v_t$ that is nonnegative and connects u and v, we can confirm the proposition by rewriting $x^u - x^v$:

$$x^u - x^v = x^u - x^{v_1} + x^{v_1} + \ldots - x^{u_t} + x^{u_t} - x^v.$$

Since each step $u_{i-1} - u_i$ is a move in $\mathcal{M}$, $x^{u_{i-1}} - x^{u_i}$ is in $I_{\mathcal{M}}$ and $x^u - x^v \in I_{\mathcal{M}}$.

The matrix A associated to a model also determines an ideal:

Definition 4.9. Let $A \in \mathbb{Z}^{d \times n}$. The *toric ideal* for A is

$$I_A = \langle x^u - x^v \mid u, v \in \ker_{\mathbb{Z}}(A) \rangle$$

The Fundamental Theorem ties I_A to $I_{\mathcal{M}}$ to show that all models have a Markov Basis.

Theorem 4.10. (The Fundamental Theorem of Markov Bases) *$\mathcal{M} \subset \ker_{\mathbb{Z}}(A)$ is a Markov basis if and only if the corresponding set of binomials*

$$\{x^{m^+} - x^{m^-} \mid m \in \mathcal{M}\}$$

generates the toric ideal I_A, or $I_{\mathcal{M}} = I_A$.

Proof. ($\Rightarrow$): Let $\mathcal{M} \subset \mathbb{Z}^n$ be a Markov basis. If $x^u - x^v \in I_A$, then $u, v \in \ker_{\mathbb{Z}}(A)$. It follows that $Au = Av$, so there is a walk connecting u and v. By **Proposition 4.8**, $x^u - x^v \in I_{\mathcal{M}}$, and $I_A \subset I_{\mathcal{M}}$
($\Leftarrow$): Since $\mathcal{M} \subset \ker_{\mathbb{Z}}(A)$, then

$$I_{\mathcal{M}} = \langle x^{m^+} - x^{m^-} \mid m \in \mathcal{M} \rangle \subset \langle x^{m^+} - x^{m^-} \mid m \in \ker_{\mathbb{Z}}(A) \rangle \subset I_A.$$

□

Recall the **Hilbert Basis Theorem**, which states that every ideal in a polynomial ring is finitely generated. Then the toric ideal I_A is finitely generated. We can take the generators of I_A to find the Markov basis. Most importantly, this means that there exists a Markov basis for any log-linear statistical model!

5. Acknowledgments

I am so appreciative of my mentor, Ignacio Darago for helping me pick this topic, explaining the tougher concepts, and guiding me throughout the project. This project was an engaging introduction to Algebraic Geometry and helped branch my interest in the computational tools that stem from algebra. I also want to thank Peter May for organizing the REU.

References

[1] Jean-Paul Brasselet. Introduction to Toric Varieties. https://www2.math.ethz.ch/education/bachelor/lectures/fs2015/math/alg_geom/brasselet

[2] Mathias Drton, Bernd Sturmfels, Seth Sullivant. Lectures on Algebraic Statistics. https://math.berkeley.edu/~bernd/owl.pdf

[3] Thomas Kahle. What is a Markov basis and what is it good for?. http://gta.math.unibuc.ro/~dumi/sna2016/kahle1bis.pdf

[4] Lior Pachter and Bernd Sturmfels. Algebraic Statistics for Computational Biology. http://yaroslavvb.com/papers/pachter-algebraic.pdf

[5] David Cox, John Little, Donal O'Shea. Ideals, Varieties, and Algorithms. http://people.dm.unipi.it/caboara/Misc/Cox,%20Little,%20O%27Shea%20-%20Ideals,%20varieties%20and%20algorithms.pdf