# TORIC IDEALS IN ALGEBRAIC STATISTICS

ALESSIO D'ALÌ

## 1. What is a... *toric ideal*?

For a very concrete introduction to commutative algebra and algebraic geometry, the reader is invited to consult [2].

Let us start by an example. Assume we are given the map

$$\begin{array}{ccc} \mathbb{K}^2 & \to & \mathbb{K}^3 \\ (x,y) & \mapsto & (x^2, xy, y^2). \end{array}$$

(Here $\mathbb{K}$ is a field. In what comes next, we will generally use $\mathbb{R}$.)

**Question.** *How can we decide if a point* $(a,b,c) \in \mathbb{K}^3$ *lies in the image of this map or not?*

The map itself gives us a parametric description of the image, but we would like to know the implicit equations.

**Answer.** *The implicit equations are given by the* toric ideal *associated with the monomials* $x^2$, $xy$, *and* $y^2$.

Let us give a precise definition.

Fix a polynomial ring $\mathbb{K}[x_1, \dots, x_n]$ and take a list of $k$ monomials

$$\{x_1^{a_{11}} x_2^{a_{21}} \dots x_n^{a_{n1}}, \; \dots, \; x_1^{a_{1k}} x_2^{a_{2k}} \dots x_n^{a_{nk}}\},$$

where the exponents $a_{ij}$ lie in $\mathbb{N}$. This nonnegativity assumption is just for simplicity's sake: one can also (with some tricks from the computational point of view) deal with exponents in $\mathbb{Z}$.

- The toric ideal associated with our list of monomials is the kernel of the map of $\mathbb{K}$-algebras

  $$\mathbb{K}[z_1, \dots, z_k] \to \mathbb{K}[x_1, \dots, x_n]$$

  that sends each polynomial variable $z_i$ to the $i$-th monomial in the list. In other words, we are interested in all the polynomials in $z_1, \dots, z_n$ that become zero after the indicated substitutions.
- Alternatively, take the $n$-by-$k$ matrix $A$ filled with the exponents $a_{ij}$ appearing in the given list of monomials. The toric ideal associated with $A$ is the ideal generated by

  $$\{\mathbf{z}^{\mathbf{u}} - \mathbf{z}^{\mathbf{v}} \mid \mathbf{u}, \mathbf{v} \in \mathbb{N}^n, \; A\mathbf{u} = A\mathbf{v}\}.$$

One checks that the two definitions agree, see for instance [1].

*Date*: August 11, 2017.

*Example* 1. The toric ideal associated with $\{x^2, xy, y^2\}$, or equivalently with the matrix $\begin{pmatrix} 2 & 1 & 0 \\ 0 & 1 & 2 \end{pmatrix}$, is

$$\langle z_1 z_3 - z_2^2 \rangle.$$

This equation arises because $(x^2)(y^2) = (xy)^2$ or, in additive notation, because $\begin{pmatrix} 2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 2 \end{pmatrix} = 2 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

Toric ideals have many desirable properties:

- they are prime (and hence the corresponding varieties are irreducible: these are the so-called *affine toric varieties*);
- they are generated by binomials, more specifically differences of monomials;
- from a computational point of view, they behave better (i.e. are less costly) than general kernels of $\mathbb{K}$-algebra maps.

## 2. Toric ideals and statistical models

In this talk we are interested in analyzing some *discrete* random variable that takes a finite number $r$ of states. This amounts to identifying a single point in the *probability simplex*

$$\Delta_{r-1} := \{(p_1, \dots, p_r) \in \mathbb{R}^r \mid p_i \geqslant 0 \ \forall i \in \{1, \dots, n\}, \sum_{i=1}^{r} p_i = 1\}.$$

If we consider the full simplex, we can always identify the observed data with a single point therein. From the point of view of statistics, this is not very interesting, since our data are not exact to start with! We want to take into account some noise and recognize our data as the (perturbed) manifestation of some easier distribution.

A *statistical model* is a (nice) subset of a probability simplex, often given in parametrized form. For our purposes, such a model will be semialgebraic, i.e. cut by polynomial equations and inequalities. Here we will try to get some information on the equations by resorting to algebraic geometry and commutative algebra.

Many interesting models in statistics arise from monomial parametrizations: thus, their vanishing ideals are toric!

*Example* 2.

$$(\theta, \psi) \mapsto (\theta^2, 2\theta\psi, \psi^2)$$

is the monomial parametrization associated with two trials of a binomial random variable: basically, we are tossing twice a biased coin that gives us heads with probability $\theta$ and tails with probability $\psi$ (if we really want to interpret these as probabilities, we should also impose that $\theta$ and $\psi$ are nonnegative and sum to 1). Then $\theta^2$ is the probability of getting two heads, $\psi^2$ of getting two tails and $2\theta\psi$ of having one heads and one tails overall (notice the "2").

*Remark* 3. Attaching coefficients to the monomials does not play a fundamental role when trying to determine the equations of the associated parametrization: the effect on the equations is kept under control simply by rescaling suitably the indeterminates. For this reason we can drop these coefficients from our considerations, hence coming back to the toric ideal setting.

If we are dealing with $d$ rolls of an $n$-sided biased die, the associated monomial parametrization is going to be associated with the Veronese variety of degree $d$ in $n$ variables.

**Question.** *What about Segre?*

## 3. Investigating independence in contingency tables

Imagine we are sampling from a certain population and taking note of hair colour and eye colour, considering then the occurrences of all possible combinations. What we are doing is investigating a *joint* (discrete) random variable, i.e. a discrete random variable that registers the simultaneous states of several "smaller" discrete random variables. When it comes to organize and visualize these data, the standard object to deal with is the so-called *contingency table*. When exactly two variables are in play, such a table is a matrix, see Table 1 below.

Table 1. A two-way contingency table from [3] (original data in [4]).

| ↓ Eye colour \ Hair colour → | Black | Brunette | Red | Blonde | Total |
|---|---|---|---|---|---|
| Brown | 68 | 119 | 26 | 7 | 220 |
| Blue | 20 | 84 | 17 | 94 | 215 |
| Hazel | 15 | 54 | 14 | 10 | 93 |
| Green | 5 | 29 | 14 | 16 | 64 |
| Total | 108 | 286 | 71 | 127 | 592 |

The $(i, j)$-th entry of a contingency matrix counts the number of occurrences of the $i$-th state of the variable $X$ and, at the same time, of the $j$-th state of the variable $Y$. Normalizing, i.e. rescaling suitably, we can interpret this entry as the probability of the joint event. A very natural question then arises:

**Question.** *Are* $X$ *and* $Y$ independent *of each other?*

Two variables are independent if and only if the probability of any joint event is the product of the probabilities of the two single events that compose it: in formulas,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

for all possible choices of $x$ and $y$.

In our context, the variables $X$ and $Y$ are independent precisely when the normalized contingency matrix has rank one. In this case the matrix is the product of the column vector $[\mathbb{P}(X = 1), \dots, \mathbb{P}(X = m)]^T$ by the row vector $[\mathbb{P}(Y = 1), \dots, \mathbb{P}(Y = n)]$.

As a consequence, our *independence model* is parametrically given by all the rank-one matrices inside the probability simplex $\Delta_{mn-1}$. Hence, the equations for the model are given by the vanishing of all 2-by-2 subdeterminants of a generic $m$-by-$n$ matrix of indeterminates. In other words, we are dealing with a Segre variety!

Note that our model is parametrized by monomials and hence the vanishing ideal is again toric. The associated matrix is an $(m+n)$-by-$mn$ matrix whose columns are indexed by the $mn$ states of the joint variable and whose rows are indexed by the $m$ states of the variable $X$ and then by the $n$ states of the variable $Y$. The $(i,j)$-th entry of this matrix is 1 if the row label is *compatible* with the column label and 0 otherwise. See Table 2 for an example when $X$ has two possible states $\{x_1, x_2\}$ and $Y$ has three possible states $\{y_1, y_2, y_3\}$.

Table 2. The matrix defining the independence model for two variables $X$ and $Y$ with respectively two and three states.

| | $\{x_1,y_1\}$ | $\{x_1,y_2\}$ | $\{x_1,y_3\}$ | $\{x_2,y_1\}$ | $\{x_2,y_2\}$ | $\{x_2,y_3\}$ |
|---|---|---|---|---|---|---|
| $\{x_1,*\}$ | 1 | 1 | 1 | 0 | 0 | 0 |
| $\{x_2,*\}$ | 0 | 0 | 0 | 1 | 1 | 1 |
| $\{*,y_1\}$ | 1 | 0 | 0 | 1 | 0 | 0 |
| $\{*,y_2\}$ | 0 | 1 | 0 | 0 | 1 | 0 |
| $\{*,y_3\}$ | 0 | 0 | 1 | 0 | 0 | 1 |

*Remark* 4. This construction can be generalized in several directions: one can deal with several variables at a time and ask for more complicated patterns of (conditional) independence. This gives rise to *hierarchical log-linear models* (see for instance [5, Chapter 9]), where the desired patterns of (in)dependence are encoded into combinatorial objects called simplicial complexes. For the great majority of such models, the problem of listing explicitly the generators of the associated toric ideal is wide open!

**Problem.** *Develop new tools and techniques to understand toric ideals associated with hierarchical log-linear models.*

*Remark* 5. As a final remark, let us note that a deeper understanding of minimal generating sets of such toric ideals would help us performing tests whose aim is to find out whether our model fits well the data or not (via sampling, since in general the environment is too big to run an exact test). This was first noticed by Diaconis and Sturmfels in 1998, see [3].

## References

[1] Bigatti, A., & Robbiano, L. (2001). *Toric ideals*. Matemática Contemporânea, **21**, 1–25.

[2] Cox, D., Little, J., & O'Shea, D. (1992). *Ideals, varieties, and algorithms*. New York: Springer.

[3] Diaconis, P., & Sturmfels, B. (1998). *Algebraic algorithms for sampling from conditional distributions*. The Annals of Statistics, **26**, 363–397.

[4] Snee, R. D. (1974). *Graphical display of two-way contingency tables. The American Statistician*, **28**, 9–12.

[5] Sullivant, S. (2017+). *Algebraic Statistics*. In preparation.