Toric Statistical Models:

Parametric and Binomial Representations

Fabio Rapallo

Department of Mathematics, University of Genova,

via Dodecaneso 35, 16146 Genova, Italy.

e-mail: rapallo@dima.unige.it

Abstract

Toric models have been recently introduced in the analysis of statistical models for categorical data. The main improvement with respect to classical log-linear models is shown to be a simple representation of structural zeros. In this paper we analyze the geometry of toric models, showing that a toric model is the disjoint union of a number of log-linear models. Moreover, we discuss the connections between the parametric and algebraic representations. The notion of Hilbert basis of a lattice is proved to allow a special representation among all possible parametrizations.

Keywords: Contingency tables, Hilbert basis, log-linear models, polynomial algebra, structural zeros, sufficient statistic, toric ideals.

1 Introduction

In the past few years, the application of new algebraic non-linear techniques to Probability and Statistics have been presented. Here we follow the polynomial representation of random variables on discrete sample spaces as introduced in Pistone et al. (2001*a*) and we study some algebraic and geometrical properties of a class of models introduced as toric models in Pistone et al. (2001*b*), showing the links between toric models and log-linear models. See also Diaconis & Sturmfels (1998), where algebraic techniques were used for the first time in the analysis of contingency tables.

The polynomial Algebra is used here to describe the geometrical structure of the statistical toric models on finite sample spaces. The first works in this direction, but limited to the analysis of graphical models, are Geiger et al. (2002) and Garcia et al. (2005).

In this paper we consider a general finite sample space and we use algebraic techniques in order to obtain a description of the notion of sufficiency. We show the connections between the parametric representation and the binomial representation of a toric model. In particular, we study the boundary of the toric model, and the problem of structural zeros. We present new results which lead to a parametrization with major properties, see Theorem 4. A number of classical examples of log-linear models (independence, quasi-independence, quasi-symmetry) are revisited in order to show the relevance of our analysis.

The term "toric" comes from Commutative Algebra, because of the algebraic

structure of the probabilities. In Commutative Algebra, toric ideals describe the algebraic relations among power products and in toric models the probabilities are expressed in terms of power products. See also Sturmfels (1996) and Bigatti & Robbiano (2001), where toric ideals are studied in details.

Working in the non-negative case, in Section 2 we recall some background material. In Section 3, we introduce the class of toric models, both parametric and binomial, and we present the first results on the relationships between the two representations. Moreover, we study the behavior of the sufficient statistic under the sampling, generalizing a result on the exponential models. In Section 4, we study the geometry of toric models and we show that toric models are not exponential models, but they are disjoint union of log-linear models, i.e., disjoint union of exponential models. In Section 5 we analyze the problem of structural zeros and its connection with the parametrization of the model. In Section 6, we show that the parametrization plays a fundamental role, and we define a special parametrization based on the notion of Hilbert basis of a lattice and we show its properties. Finally, in Section 7 we show a detailed example from the classical literature on log-linear models.

2 Notation and background material

Consider a statistical model on a finite sample space $\mathcal{X}$. Although in contingency tables the sample space is usually a Cartesian product (for example in the two-way case is a space of the form $\{(i,j) \mid i = 1,\dots I,\ j = 1,\dots,J\}$), we assume here,

without loss of generality, a generic list of points $\mathcal{X}$, with $\#\mathcal{X} = k$. The sample points are denoted by $x \in \mathcal{X}$.

A probability distribution on $\mathcal{X}$ is characterized by the value of the parameters $p_x = \mathbb{P}[x]$ for all $x \in \mathcal{X}$. In this paper we use the vector notation, that is p is the k-dimensional vector $(p_x, x \in \mathcal{X})$. The parameter space for the saturated model is given by the simplex $p_x \geq 0$, $x \in \mathcal{X}$ and $\sum_{x\in\mathcal{X}} p_x = 1$. A statistical model is a variety of that simplex. Here we assume that the model is described by a set of polynomial equations. This means that the model is defined through the conditions $f_1(p) = 0, \ldots, f_m(p) = 0$, where $f_1, \ldots, f_m$ are polynomials. The variety of the simplex is the subset of the simplex where all $f_1, \ldots, f_m$ vanish.

Let us consider a statistical model on the sample space $\mathcal{X}$ of the form

$$p_x = \mathbb{P}_\phi[x] = \phi(T(x)), \quad x \in \mathcal{X} \tag{1}$$

where $T : \mathcal{X} \longrightarrow \mathbb{N}^s$ is the vector of integer valued components of the sufficient statistic. Here $\mathbb{N}$ denotes the set of non-negative integer numbers. In general $\phi \in \Phi$, where Φ is a subset of functions from $\mathbb{N}^s$ to $\mathbb{R}$. In point of fact, the range of T is a finite subset $\mathcal{T} \subset \mathbb{N}^s$.

The set Φ defines a subset M of the space of the probabilities, and we refer to this subset as to a statistical model. In other words, M is a subset defined through

$$M = \{p \ : \ p = \phi T, \ \phi \in \Phi\}\,.$$

By the well known factorization theorem, T is a sufficient statistic of this model.

3 Parametric and binomial toric models

We specialize Equation (1), by assuming that there exists a parametrization with non-negative parameters $\zeta_1, \ldots, \zeta_s$ such that the probabilities assume the form

$$p_x = \frac{L(\zeta, x)}{\sum_{y \in \mathcal{X}} L(\zeta, y)}\,, \tag{2}$$

with $L(\zeta, x) = \zeta^{T(x)}$. Here $T(x) = (T_1(x), \ldots, T_s(x))$ and $\sum_{y \in \mathcal{X}} L(\zeta, y)$ is the normalizing constant. Consistently with the vector notation, $\zeta^{T(x)}$ denotes the monomial $\zeta_1^{T_1(x)} \cdots \zeta_s^{T_s(x)}$. Apart from the normalizing constant, the function $L(\zeta, x)$ is the likelihood of the statistical model.

It is known, see Pistone et al. (2001*a*) and Pistone et al. (2001*b*), that the probabilities expressed in the form (2) lead to a binomial representation of the statistical model, called toric model.

We recall briefly the construction of the relevant binomials. For the basic Commutative Algebra we refer to Kreuzer & Robbiano (2000).

Consider the polynomial ring $\mathbb{Q}[p, \zeta]$, i.e. the set of all polynomials in the indeterminates $p_1, \ldots, p_k, \zeta_1, \ldots, \zeta_s$. Define the binomials $p_x - L(\zeta, x)$, $x \in \mathcal{X}$ and take the ideal $\mathcal{I}$ generated by such binomials.

The relevant set of binomials $\mathcal{B}$ is obtained by elimination of the ζ indeterminates, see Pistone et al. (2001*a*). $\mathcal{B}$ is a set of generators of the ideal

$$\mathcal{I}_M = \mathrm{Elim}(\zeta, \mathcal{I}) \tag{3}$$

It is known that the set of generators is not unique. Among others, we consider here $\mathcal{B}$ as a Gröbner basis of the ideal $\mathcal{I}_M$, see Sturmfels (1996) for algebraic details.

The computation of the elimination ideal in Equation (3) and its Gröbner basis can be performed with symbolic software, such as CoCoA, see CoCoATeam (2004). In point of fact, there exist other methods for the computation of the relevant ideal. The elimination method cited here is the simplest, while other methods based on the algebraic technique of saturation are faster and computationally feasible. For a review on such methods, see for example Rapallo (2003).

Definition 1. If a statistical model M consists of all probability functions of the form in Equation (2), then we say that the model is a parametric toric model.

Example 1. Following Pistone et al. (2001*a*), we show the computation of the binomial for the classical 4-cycle, the conditional independence model for 4 binary random variables X_1, X_2, X_3, X_4, see Lauritzen (1996). Here $\mathcal{X} = \{1, 2\}^4$. The model is defined through the conditional independence statements $X_1 \perp X_3 | \{X_2, X_4\}$ and $X_2 \perp X_4 | \{X_1, X_3\}$. The parametric toric model is expressed for example by the set of equations below.

$$
\begin{array}{ll}
p_{1111} = \zeta_0 & p_{1221} = \zeta_0\zeta_2\zeta_3\zeta_7 \\
p_{2111} = \zeta_0\zeta_1 & p_{1212} = \zeta_0\zeta_2\zeta_4 \\
p_{1211} = \zeta_0\zeta_2 & p_{1122} = \zeta_0\zeta_3\zeta_4\zeta_8 \\
p_{1121} = \zeta_0\zeta_3 & p_{2221} = \zeta_0\zeta_1\zeta_2\zeta_3\zeta_5\zeta_7 \\
p_{1112} = \zeta_0\zeta_4 & p_{2212} = \zeta_0\zeta_1\zeta_2\zeta_4\zeta_5\zeta_6 \\
p_{2211} = \zeta_0\zeta_1\zeta_2\zeta_5 & p_{2122} = \zeta_0\zeta_1\zeta_3\zeta_4\zeta_6\zeta_8 \\
p_{2121} = \zeta_0\zeta_1\zeta_3 & p_{1222} = \zeta_0\zeta_2\zeta_3\zeta_4\zeta_7\zeta_8 \\
p_{2112} = \zeta_0\zeta_1\zeta_4\zeta_6 & p_{2222} = \zeta_0\zeta_1\zeta_2\zeta_3\zeta_4\zeta_5\zeta_6\zeta_7\zeta_8
\end{array}
$$

In the statistical model, the coding $\{1, 2\}$ is merely a notational fact. Moreover, the conditional independence statements are invariant under the shift of indices. Thus, all equations have to be invariant under the action of the group $(\mathcal{S}_2)^4 \times \mathcal{C}_4$, where $\mathcal{S}_2$ is the group of permutations over the set of codings $\{1, 2\}$ and $\mathcal{C}_4$ is the cyclic sub-group of $\mathcal{S}_4$ generated by the permutation $(2, 3, 4, 1)$. $(\mathcal{S}_2)^4$ acts componentwise on the indices, while $\mathcal{C}_4$ naturally acts as permutation over four elements. The use of symmetries and group actions in this context is carefully discussed in Aoki & Takemura (2005). We exploit this fact in order to give a synthetic description of the ideal. In particular, the following binomials and their orbits give a system of generators of the elimination ideal in Equation (3):

- $p_{1111}p_{1212} - p_{1211}p_{1112}$ (orbit of cardinality 8);
- $p_{1112}p_{1221}p_{2122} - p_{1121}p_{2112}p_{1222}$ (orbit of cardinality 16);
- $p_{1111}p_{1221}p_{1212}p_{2122} - p_{1211}p_{1121}p_{2112}p_{1222}$ (orbit of cardinality 32);
- $p_{2111}p_{1221}p_{1212}p_{2122} - p_{1211}p_{2121}p_{2112}p_{1222}$ (orbit of cardinality 8).

For a discussion on the meaning of such binomials, see Pistone & Wynn (2003) and Sturmfels (2002). In this paper we investigate deeply the relationships between the parametric representations and the binomial representation.

In Definition 1, the ζ parameters are unrestricted, except the non-negativity constraint. Note that in general a toric model is bigger than the exponential model, as we do not assume positive probabilities, and with the constraint $p_x > 0$ for all $x \in \mathcal{X}$ it is a log-linear model. Let $M_{>0}$ be the subset of the toric model M with

the restriction $p_x > 0$ for all $x \in \mathcal{X}$. Then

$$\log p_x = \sum_j (\log \zeta_j) T_j(x) + \log \zeta_0 \,,$$

where $\zeta_0 = (\sum_x L(\zeta, x))^{-1}$ is the normalizing constant. $M_{>0}$ is a log-linear, and thus an exponential model with sufficient statistic T and canonical parameters $\log \zeta_j$, $j = 1, \ldots, s$.

Note that the representation of the toric model in Equation (2) points to the notion of multiplicative form of a log-linear model, see for example Goodman (1979).

Example 2. Denote by $\mathbb{I}_A$ the indicator function of the set A (i.e. $\mathbb{I}_A(x) = 1$ if $x \in A$ and $\mathbb{I}_A = 0$ otherwise). A first example of toric model is a model with sufficient statistic consisting of the counts over the sets $A_1, \ldots, A_s \subseteq \mathcal{X}$, possibly overlapping:

$$T : x \longmapsto (\mathbb{I}_{A_1}(x), \ldots, \mathbb{I}_{A_s}(x))$$

In most examples in the literature $\mathcal{X} \subseteq \{1, \ldots, I_1\} \times \cdots \times \{1, \ldots, I_d\}$ is a d-way array, possibly incomplete.

Example 3. A second example is a log-linear model of the type

$$\log \mathbb{P}_\psi(x) = \sum_{i=1}^{s} \psi_i T_i(x) - k(\psi) \tag{4}$$

with integer valued T_i's and parameters $\psi = (\psi_1, \ldots, \psi_s)$, see Fienberg (1980). Note that in Equation (1) the strict positivity implied by the log-linear model in Equation (4) is not assumed.

For simplicity we state the following definition for the binomial representation of a toric model.

Definition 2. Given a parametric toric model, the corresponding binomial toric model is the zero set of the polynomial ideal defined in Equation (3).

The connections between the two representations of toric models will be analyzed in the next sections. The following Lemma shows a first relationship among parametric toric models and binomial toric models.

Lemma 1. *Given a parametric toric model, the corresponding binomial toric model contains the parametric model.*

Proof. Let $\gamma : \mathbb{R}^s_{\geq 0} \longrightarrow \mathbb{R}^n$ be the function such that $\gamma(\zeta) = (p_x(\zeta), x \in \mathcal{X})$ and let $V(\mathcal{I}_M)$ be the zero set of the ideal $\mathcal{I}_M$. It is enough to prove that $\gamma(\mathbb{R}^s_{\geq 0}) \subseteq V(\mathcal{I}_M)$. By definition $\mathcal{I}_M = \text{Elim}(\zeta, \mathcal{I})$, where $\mathcal{I} = \text{Ideal}(p_x - p_x(\zeta), x \in \mathcal{X})$. If $p = p(\gamma) \in \gamma(\mathbb{R}^s_{\geq 0})$, for all polynomial $g \in \mathcal{I}_M \subset \mathcal{I}$ there exist polynomials $q_x(p, \zeta)$, $x \in \mathcal{X}$ such that

$$g = \sum_{x \in \mathcal{X}} q_x(p, \zeta)(p_x - p_x(\zeta))$$

and thus $g = 0$, as all terms $(p_x - p_x(\zeta))$ vanish. □

It is known that different parametrizations can lead to the same set of binomials, i.e. to the same binomial toric model. Thus, we state the following definition.

Definition 3. Two parametric toric models are said to be b-equivalent if they have the same binomial representation.

A number of parametrizations exist for the same binomial toric model. However, the different parametrizations allow differences only on the boundary of the model.

Lemma 2. *Two b-equivalent parametric toric models restricted to $M_{>0}$ define the same variety.*

Proof. In the strictly positive case the parametric toric model assumes the form $\log p = T \log \zeta$, where T is the matrix of the exponents of the monomials. Let

$$\log p = T \log \zeta \quad \text{and} \quad \log p = \tilde{T} \log \tilde{\zeta} \tag{5}$$

be two b-equivalent parametric toric models. From Theorem 2.10 in Bigatti & Robbiano (2001) follows that $\text{Image}(T) = \text{Image}(\tilde{T})$ and then there exists parameters ζ and $\tilde{\zeta}$ such that the relationships in Equation (5) are both verified. □

Example 4. Consider the simple case of the independence model for 2×2 tables. The binomial toric model is given by the binomial

$$p_{11}p_{22} - p_{12}p_{21}$$

and the following two parametric toric models are *b*-equivalent:

$$(p_{11}, p_{12}, p_{21}, p_{22}) = (\zeta_0, \zeta_0\zeta_1, \zeta_0\zeta_2, \zeta_0\zeta_1\zeta_2) \quad \text{or} \quad (\zeta_0\zeta_2, \zeta_0\zeta_3, \zeta_1\zeta_2, \zeta_1\zeta_3)\,. \tag{6}$$

The second parametrization is used in many books, see for example Agresti (2002).

In view of Lemmas 1 and 2, it is clear why we use the analysis of parametric toric models in order to study the structural zeros.

Consider the problem of sampling. It is easy to generalize a result well-known in the case of exponential models, that is when we suppose $p_x > 0$ for all $x \in \mathcal{X}$. We denote by $(x_1, \ldots, x_N)$ a sample of size N drawn from a vector $(X_1, \ldots, X_N)$ of

independent and identically $\mathbb{P}$-distributed random variables with values in $\mathcal{X}$. As the probabilities can be written in the form

$$p_x = \zeta_0 \zeta_1^{T_1(x)} \cdots \zeta_s^{T_s(x)}$$

the probability of a sample $(x_1, \ldots, x_N)$ is

$$p_{x_1} \cdots p_{x_N} = \zeta_0^N \zeta_1^{\sum_i T_1(x_i)} \cdots \zeta_s^{\sum_i T_s(x_i)}$$

i.e., the sufficient statistic for the sample of size N is the sum of the sufficient statistics of the N components of the sample. Note that this result is formally the same as in the positive case where the theory of exponential models applies. The proof can be carried out by straightforward computation. In fact, the j-th component of the sufficient statistic for the sample of size N can be written as

$$\sum_i T_j(x_i) = \sum_{a \in \mathcal{X}} T_j(a) F_a(x_1, \ldots, x_N)$$

where

$$F_a(x_1, \ldots, x_N) = \sum_{i=1}^{N} \mathbb{I}_a(x_i)$$

is the count of the cell a.

4 Geometry of toric models

Define the matrix A_T in the following way. A_T is a matrix with k rows and s columns and its generic element $A_T(i, j)$ is $T_j(x_i)$ for all $i = 1, \ldots k$ and $j = 1, \ldots, s$.

If $\eta_j = E(T_j)$ are the expectation parameters and p is the row vector of the probabilities, then

$$\eta = pA_T \tag{7}$$

The matrix A_T can also be used in order to describe the geometric structure of the statistical model. First, we can state the following result.

Proposition 1. *Choose a parameter ζ_j and take the set $\mathcal{X}'$ of points $x \in \mathcal{X}$ such that $T_j(x) > 0$. Suppose that $\mathcal{X}' \neq \mathcal{X}$. If we set $\zeta_j = 0$, we obtain a model on the remaining sample points and this model is again a toric model.*

Proof. Without loss of generality, suppose $j = 1$ and $T_1(x) = 0$ for $i = 1, \ldots, k'$ and $T_1(x) > 0$ for $i = k' + 1, \ldots, k$. The matrix A_T can be partitioned as

$$A_T = \left(\begin{array}{c|c} 0 & \\ \vdots & A'_T \\ 0 & \\ \hline * & \\ \vdots & \\ * & \end{array}\right)$$

where $*$ denotes non-zero entries. Now, the matrix A'_T is non-zero and it is the representation of a toric model on the first k' sample points. □

Thus, the geometric structure of the toric model is an exponential model and at most s toric models with $(s-1)$ parameters on appropriate subsets of the sample space $\mathcal{X}$. Moreover, by applying recursively the above theorem we obtain the following result.

Theorem 1. *Let $\zeta_{j_1}, \ldots, \zeta_{j_r}$ be a set of parameters and take the set $\mathcal{X}'$ of points $x \in \mathcal{X}$ such that $T_{i_b}(x) > 0$ for at least one $b \in \{1, \ldots, r\}$ and $T_{i_b}(x) = 0$ for all*

$b = 1, \ldots, r$ and $x \in \mathcal{X} - \mathcal{X}'$. Suppose that $\mathcal{X}' \neq \mathcal{X}$. If we set $\zeta_{j_b} = 0$ for all $b = 1, \ldots, r$, we obtain a toric model on the remaining sample points.

Proof. Apply b times Proposition 1. □

These results lead to a geometrical characterization of a toric model.

Theorem 2. *A toric model is the disjoint union of exponential models.*

Proof. The result follows from the application of Theorem 1 for all possible sets of parameters $\zeta_{j_1}, \ldots, \zeta_{j_b}$ for which $\mathcal{X}' \neq \mathcal{X}$. □

Example 5. Consider the independence model for 3×2 tables. The matrix representation of the sufficient statistic is

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

and the toric model is, apart from the normalizing constant,

$$(p_{11}, p_{12}, p_{21}, p_{22}, p_{31}, p_{32}) = (\zeta_1\zeta_4, \zeta_1\zeta_5, \zeta_2\zeta_4, \zeta_2\zeta_5, \zeta_3\zeta_4, \zeta_3\zeta_5)\,.$$

If $\zeta > 0$ we have the log-linear model. Moreover, we can choose the following sets $\mathcal{X}'$ in Theorem 1:

- $\mathcal{X}' = \{(1,1), (1,2)\}$ corresponding to $\zeta_1 = 0$. In this cases we obtain the independence model for the 2×2 table $\mathcal{X} - \mathcal{X}'$. Similarly for $\zeta_2 = 0$ and $\zeta_3 = 0$;

- $\mathcal{X}' = \{(1,1),(2,1),(3,1)\}$ corresponding to $\zeta_4 = 0$. In this case we obtain the multinomial model for the second column. Similarly for $\zeta_5 = 0$.

- $\mathcal{X}' = \{(1,1),(1,2),(2,1),(2,2))\}$ corresponding to $\zeta_1 = \zeta_2 = 0$. In this case we obtain the Bernoulli model for the third row. Similarly for $\zeta_1 = \zeta_3 = 0$ and $\zeta_2 = \zeta_3 = 0$.

- $\mathcal{X}' = \{(1,1),(1,2),(2,1),(3,1)\}$ corresponding to $\zeta_1 = \zeta_4 = 0$ and similarly for $\zeta_1 = \zeta_5 = 0$, $\zeta_2 = \zeta_4 = 0$, $\zeta_2 = \zeta_5 = 0$, $\zeta_3 = \zeta_4 = 0$ and $\zeta_3 = \zeta_5 = 0$. In this case we obtain 6 Bernoulli models on the columns of the independence models found above.

- corresponding to the six conditions $\zeta_1 = \zeta_2 = \zeta_4 = 0$, $\zeta_1 = \zeta_2 = \zeta_5 = 0$, $\zeta_1 = \zeta_3 = \zeta_4 = 0$, $\zeta_1 = \zeta_3 = \zeta_5 = 0$, $\zeta_2 = \zeta_3 = \zeta_4 = 0$ and $\zeta_2 = \zeta_3 = \zeta_5 = 0$ we obtain 6 trivial distributions on one sample point.

The toric model is then formed by 21 models: the log-linear model on 6 points, 3 models on 4 points, 2 models on 3 points, 9 models on 2 points and 6 trivial models on 1 point.

5 Structural zeros

The procedure in Theorem 1, applied as in Example 5 can also be used to define the admissible sets of structural zeros.

Definition 4. A subset $\mathcal{X}' \subset \mathcal{X}$ is an admissible set of structural zeros for a parametric toric model with parameters $\zeta_1, \ldots, \zeta_s$ if there exist parameters $\zeta_{i_1}, \ldots, \zeta_{i_r}$

such that the condition of Theorem 1 holds.

Now, we consider the behavior of the η-parametrization in the different exponential models. Starting from the representation of the η parameters as function of the ζ parameters in Equation (7), we can easily prove that the η parametrization is coherent, as stated in the following proposition.

Proposition 2. *The η parameters for the reduced models are the same as in the exponential case, provided that a component is fixed to zero.*

Proof. For the proof, it is enough to combine the linear relation between η and ζ given in Equation (7) and Theorem 1. □

Related work about the relationships among the parametrizations and the exponential family are presented in Geiger et al. (2001). Moreover, in the strictly positive case, the geometry of independence and conditional independence models is analyzed in the context of graphical models in Geiger et al. (2002).

6 Analysis of parametric toric models

The geometric representation of the structural zeros as presented in the previous definition needs some further discussions. Let us consider a simple example.

Example 6. Consider the independence model for 2×2 contingency tables and its two parametrizations in Equation (6). The parametric toric models are different. In fact, the second parametrization contains for example the Bernoulli model on the two points $(1, 1)$ and $(1, 2)$, while the first parametrization does not.

Now, consider the binomial equations obtained from a toric model by elimination of the ζ indeterminates as described in Section 3. In this way we obtain a set of binomials which defines a statistical model, but in general the binomial toric model differs from the parametric toric model. The binomial model is independent on the parametrizations and it can allow some boundaries excluded from the parametric model, as in the previous example.

Moreover, we can also restate the definition of set of structural zeros for a binomial toric model.

Definition 5. Let $\mathcal{B}$ be the set of binomials defined by elimination as described in Section 3. A subset $\mathcal{X}' \subset \mathcal{X}$ is an admissible set of structural zeros independent from the parametrization if the polynomial system $\mathcal{B} = 0$ together with $p_x = 0$ for all $x \in \mathcal{X}'$ has a non-negative normalized solution.

In order to show the difference between Definitions 4 and 5, let us discuss the following example.

Example 7. Consider the parametric toric model of independence for 2×2 contingency tables with the parametrization

$$(p_{11}, p_{12}, p_{21}, p_{22}) = (\zeta_0, \zeta_0\zeta_1, \zeta_0\zeta_2, \zeta_0\zeta_1\zeta_2)$$

and binomial representation

$$\mathcal{B} = \{p_{11}p_{22} - p_{12}p_{21}\}.$$

The set $\mathcal{X}' = \{(1,1),(1,2)\}$ is an admissible set of structural zeros independent from the parametrization as $(p_{11}, p_{12}, p_{21}, p_{22}) = (0, 0, 1/2, 1/2)$ is a non-negative

normalized solution of $\mathcal{B} = 0$, but it is not an admissible set of structural zeros for this parametrization as $p_{11} = 0$ implies $\zeta_0 = 0$ which in turn implies $(p_{11}, p_{12}, p_{21}, p_{22}) = (0, 0, 0, 0)$.

In general, it is difficult to find a parametrization such that the parametric form of the toric model contains all the exponential sub-models.

Definition 6. If a parametrization defines a parametric toric model which contains all the exponential sub-models, we say that the parametrization is a full parametrization.

In view of Example 6 and Definition 6, it follows that not all matrix representations of the sufficient statistic are equivalent. In general there is an infinite number of non-negative integer valued bases of the sub-space spanned by T, but not all of these contains all the exponential sub-models. This happens because the columns of the matrix A_T are defined in the vector space framework, but in the power product representation we need non-negative exponents and then we need linear combinations with non-negative integer coefficients. In general, it is difficult to find a full parametrization, but it is easy to characterize all the parametrizations which lead to a given binomial toric model.

Theorem 3. *Let A_T be the matrix representation of the sufficient statistic of a toric model. If $v_1, \ldots, v_s$ is a non-negative integer system of generators of the image of A_T, then*

$$p_x = \zeta_1^{v_1(x)} \cdots \zeta_s^{v_s(x)}$$

for $x \in \mathcal{X}$ is a parametrization and all parametrizations of this kind lead to the same binomial toric model.

Proof. Let A_v be the matrix formed by the column vectors $v_1, \ldots, v_s$, i.e., $A_v = (v_1, \ldots, v_s)$. By the definition of the v's it follows that $\text{Image}(A_T) = \text{Image}(A_s)$ and thus the kernels of the systems $A_v = 0$ and $A_T = 0$ are the same. As a consequence of the construction of the toric ideals presented in Section 3, the toric models defined through A_T and A_v are represented by the same binomials. □

Among others, one can consider the image of A_T as a lattice and then consider a Hilbert basis of such lattice. We look at A_T as an operator from $\mathbb{N}^k$ to $\mathbb{N}^s$. The corresponding linear system has natural coefficients and we are interested in its solution with natural components.

Definition 7. Let $S \subseteq \mathbb{N}^k$ be the set of integer solutions of a diophantine system $pA_T = 0$. A set of integer vectors $\mathcal{H} = \{v_1, \ldots, v_h\}$ is a Hilbert basis of S if for all $\beta \in S$

$$\beta = \sum_{v \in \mathcal{H}} c_v v$$

where $c_v \in \mathbb{N}$.

The notion of Hilbert basis has major applications both in combinatorics and integer programming. It is known that such a set $\mathcal{H}$ exists and is unique. The number of elements in $\mathcal{H}$ in general differs from the dimension of the image of $\mathcal{A}_T$ as vector sub-space. For details about the properties of the Hilbert basis and the algorithms for its computation, see Sturmfels (1993) and Kreuzer & Robbiano (2005).

Theorem 4. *Let $v_1, \ldots, v_s$ be the columns of the matrix A_T and suppose that $\{v_1, \ldots, v_s\}$ is the Hilbert basis of the image of A_T. Then the parametrization*

$$p_x = \zeta_1^{v_1(x)} \cdots \zeta_s^{v_s(x)}$$

for $x \in \mathcal{X}$ is the bigger parametrization of the toric model.

Proof. Consider another parametrization

$$p_x = \theta_1^{u_1(x)} \cdots \theta_t^{u_t(x)}$$

As $\{v_1, \ldots, v_s\}$ is a Hilbert basis, any $u_i(x)$ can be written in the form

$$u_i(x) = \sum_{j=1}^{s} c_{i,j} v_j(x)$$

with non-negative integer coefficients $c_{i,j}$. Thus,

$$p_x = \theta_1^{u_1(x)} \cdots \theta_t^{u_t(x)} = \theta_1^{\sum_{j=1}^{s} c_{1,j} v_j(x)} \cdots \theta_t^{\sum_{j=1}^{s} c_{t,j} v_j(x)} =$$

rearranging the exponents

$$= \left(\prod_{i=1}^{t} \theta_i^{c_{i,1}} \right)^{v_1(x)} \cdots \left(\prod_{i=1}^{t} \theta_i^{c_{i,s}} \right)^{v_s(x)}$$

and the result is proved. □

The Hilbert basis can be computed using symbolic software, for example the free software 4ti2, see Hemmecke et al. (2005).

Example 8. Consider again the simple independence model for 2 tables of Examples 4 and 6. The image of A_T is generated by the three vectors $u_1 = (1,1,1,1)^t$, $u_2 = (1,0,1,0)^t$ and $u_3 = (0,0,1,1)^t$. The vectors u_1, u_2, u_3 generate the first

parametrization in Equation (6). The Hilbert basis of the image of A_T consists of the four vectors $v_1 = (1,1,0,0)^t$, $v_2 = (0,0,1,1)^t$, $v_3 = (1,0,1,0)^t$ and $v_4 = (0,1,0,1)^t$. These vectors lead to the second parametrization in Equation (6).

7 A final example

Classical books on log-linear models state that the quasi-independence model is equivalent to the quasi-symmetry model in the case of 3×3 contingency tables, see for example Agresti (2002), page 427. We shortly recall two such models.

In the usual notation for log-linear models, see for example Bishop et al. (1975), the quasi-independence model has the form

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \delta_i \mathbb{I}_{\{i=j\}} \tag{8}$$

where the μ_{ij}'s are the expected frequencies, the λ_i^X's are the effects of the first variable X with values $1, \ldots, 3$, the λ_j^Y's are the effects of the second variable Y with values $1, \ldots, 3$ and the δ_i's are the effects of the diagonal cells (here the indicator $\mathbb{I}_{\{i=j\}}$ equals 1 when $i = j$ and 0 otherwise).

The quasi-symmetry model has the form

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij} \tag{9}$$

with $\lambda_{ij} = \lambda_{ji}$ for all $i, j = 1, \ldots, 3$.

Using the technique described in Example 3, one can found the parametric toric

models. For the quasi-independence model the following parametrization comes out:

$$(10)\quad (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33}) = \\ = (\zeta_1\zeta_4\zeta_7, \zeta_1\zeta_5, \zeta_1\zeta_6, \zeta_2\zeta_4, \zeta_2\zeta_5\zeta_8, \zeta_2\zeta_6, \zeta_3\zeta_4, \zeta_3\zeta_5, \zeta_3\zeta_6\zeta_9)\,.$$

The binomial toric model is represented by one binomial. In fact, applying the elimination algorithm, we find:

$$(11)\qquad \mathcal{B}_{qi} = \{p_{12}p_{23}p_{31} - p_{13}p_{21}p_{32}\}$$

For the quasi-symmetry, the parametric representation is

$$(12)\quad (p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33}) = \\ = (\zeta_1\zeta_4\zeta_7, \zeta_1\zeta_5\zeta_{10}, \zeta_1\zeta_6\zeta_{11}, \zeta_2\zeta_4\zeta_{10}, \zeta_2\zeta_5\zeta_8, \zeta_2\zeta_6\zeta_{12}, \zeta_3\zeta_4\zeta_{11}, \zeta_3\zeta_5\zeta_{12}, \zeta_3\zeta_6\zeta_9)$$

and the binomial toric model is again:

$$(13)\qquad \mathcal{B}_{qs} = \mathcal{B}_{qi} = \{p_{12}p_{23}p_{31} - p_{13}p_{21}p_{32}\}$$

The two parametric toric models are b-equivalent, but the boundaries differ. For instance, the set $\mathcal{X}' = \{(1,3),(3,1)\}$ is a set of structural zeros for the quasi-symmetry model, but it does not for the quasi-independence model.

Acknowledgments

We want to acknowledge Professor Giovanni Pistone (Politecnico of Turin, Italy) for his helpful suggestions. This work is partially supported by MIUR grant COFIN03.

References

Agresti, A. (2002), *Categorical Data Analysis*, 2 edn, Wiley, New York.

Aoki, S. & Takemura, A. (2005), The largest group of invariance for Markov bases and toric ideals, Technical Report METR 2005-14, Department of Mathematical Informatics, The University of Tokio, Tokio.

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

Bishop, Y. M., Fienberg, S. & Holland, P. W. (1975), *Discrete multivariate analysis: Theory and practice*, MIT Press, Cambridge.

CoCoATeam (2004), *CoCoA, a system for doing Computations in Commutative Algebra*, 4.0 edn, Available at `http://cocoa.dima.unige.it`.

Diaconis, P. & Sturmfels, B. (1998), 'Algebraic algorithms for sampling from conditional distributions', *Annals of Statistics* **26**(1), 363–397.

Fienberg, S. (1980), *The Analysis of Cross-Classified Categorical Data*, MIT Press, Cambridge.

Garcia, L. D., Stillman, M. & Sturmfels, B. (2005), 'Algebraic geometry of Bayesyan networks', *Journal of Symbolic Computation* **39**, 331–355.

Geiger, D., Heckerman, D., King, H. & Meek, C. (2001), 'Stratified exponential families: Graphical models and model selection', *Annals of Statistics* **29**(3), 505–529.

Geiger, D., Meek, C. & Sturmfels, B. (2002), On the toric algebra of graphical models. Microsoft Research Report MSR-TR-2002-47.

Goodman, L. A. (1979), 'Multiplicative models for square contingency tables with ordered categories', *Biometrika* **66**(3), 413–418.

Hemmecke, R., Hemmecke, R. & Malkin, P. (2005), '4ti2 version 1.2—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more', Available at www.4ti2.de.

Kreuzer, M. & Robbiano, L. (2000), *Computational Commutative Algebra 1*, Springer, Berlin.

Kreuzer, M. & Robbiano, L. (2005), *Computational Commutative Algebra 2*, Springer, Berlin.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford University Press, New York.

Pistone, G., Riccomagno, E. & Wynn, H. P. (2001*a*), *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman&Hall/CRC, Boca Raton.

Pistone, G., Riccomagno, E. & Wynn, H. P. (2001*b*), Computational commutative algebra in discrete statistics, *in* M. A. G. Viana & D. S. P. Richards, eds, 'Algebraic Methods in Statistics and Probability', Vol. 287 of *Contemporary Mathematics*, American Mathematical Society, pp. 267–282.

Pistone, G. & Wynn, H. P. (2003), 'Statistical toric models', Lecture Notes, Grostat VI, Menton, France.

Rapallo, F. (2003), 'Algebraic Markov bases and MCMC for two-way contingency tables', *Scandinavian Journal of Statistics* **30**(2), 385–397.

Sturmfels, B. (1993), *Algorithms in Invariant Theory*, Texts and Monographs in Symbolic Computation, Springer, New York.

Sturmfels, B. (1996), *Gröbner bases and convex polytopes*, Vol. 8 of *University lecture series (Providence, R.I.)*, American Mathematical Society.

Sturmfels, B. (2002), *Solving Systems of Polynomial Equations*, Vol. 97 of *CBMS Regional Conference Series in Mathematics*, American Mathematical Society, Providence, RI.