**Noname manuscript No.**
(will be inserted by the editor)

# A short history of algebraic statistics

**Eva Riccomagno**

Received: date / Accepted: date

**Abstract** In algebraic statistics, computational techniques from algebraic geometry become tools to address statistical problems. This, in turn, may prompt research in algebraic geometry. The basic ideas at the core of algebraic statistics will be presented. In particular we shall consider application to contingency tables and to design of experiments.

**Keywords** Algebraic Statistics · Design of experiments · Contingency tables

## 1 Introduction

Polynomials and ratios of polynomials appear in statistics and probability under various forms, in model representations as well as in inferential procedures. Algebraic geometry studies (ratios of) polynomials and the zero set of systems of polynomial equations. Recent developments in computational commutative algebra and their implementation made effective the application of algebraic geometry in statistics and probability, generating what is now called Algebraic Statistics.

Algebraic statistics uses techniques from (real, computational) algebraic geometry, commutative algebra and geometric combinatorics, to name a few, to gain insight into the structure and properties of statistical models and to advise in model analysis. In turn, these applications may prompt research in algebraic geometry. An example of this synergy is given by the special issues of two international journals dedicated to algebraic statistics, one is the Journal of Symbolic Computation [2] and the other one is Statistica Sinica [1].

Two papers set the foundations of algebraic statistics. Paper [24] circulated in manuscript form since 1993 and its abstract gives a clear indication of its aim. “We

E. Riccomagno
Department of Mathematics, Università di Genova
Via Dodecaneso 35
Tel.: +39-010-5646938
Fax: +39-010-564
E-mail: riccomagno@dima.unige.it

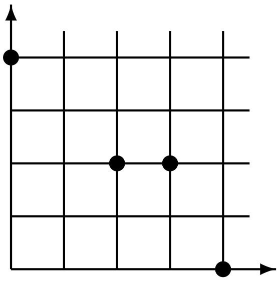

**Fig. 1** A four point design in two dimensions

construct Markov chain algorithms for sampling from discrete exponential families conditional on a sufficient statistic. Examples include contingency tables, logistic regression, and spectral analysis of permutation data. The algorithms involve computations in polynomial rings using Gröbner bases." The other paper [57] introduces the use of computational commutative algebra in design and analysis of experiments. Its abstract reads: "Many problems of confounding and identifiability for polynomial and multidimensional polynomial models can be solved using methods of algebraic geometry aided by the fact that modern computational algebra packages such as MAPLE can be used. [...]."

Another prime contribution of algebraic statistics is in statistical modelling. The basic idea of identifying some statistical models with algebraic varieties can be found in Chapter 6 of [55] and led first into the algebra of toric models [37] and is currently being developed in various directions.

Our paper gives a rather technical introduction to the very basics of algebraic statistics. Sections 2 presents the fundamental notions from algebraic geometry. It does so with reference to the application to design. The application to contingency tables is build on this in Section 3, hinting to a connection between these two main applications of algebraic statistics which is recently being developed. In Section 4 algebraic statistical models are discussed. Pointers to the literature are provided throughout and a reasoned list of references on some more advanced topics is collected in Section 5.

## 2 Designs and polynomials

Designs give settings for experiments. It is then of interest which models are identifiable/estimable from the outcome of the experiment. We consider polynomial models which give a (purely) mathematical description of all possible models for the experiment (this statement is properly qualified in Section 2.2). In common statistical practice, usually one uses some standard models and if needed modifies them. The first contribution of algebraic statistics is to provide a systematic approach to this issue, particularly useful for non-standard designs.

### 2.1 From designs to polynomial ideals

By a design here we mean a finite set of $n$ distinct points in $k$ dimensions on each of which a measurement is taken or the same number of measurements are taken at each point. In this case the outcomes at a location are averaged. The technology illustrated below becomes relevant for designs with no particular geometrical regularity and $n, k$ relatively large. But, here, consider $\mathcal{A} = \{(0,4),(2,2),(4,0),(3,2)\}$ in Figure 1 to illustrate the main points. The reader could try the computations below on the set of points $\{(-1,1),(0,0),(1,-1),(1,1)\}$.

First, observe that the points in $\mathcal{A}$ are solutions of the system of polynomial equations $g_1 = g_2 = g_3 = g_4 = 0$ with

$$\begin{cases} g_1 := x(x-2)(x-4)(x-3) \\ g_2 := (y-4)(y-2)y \\ g_3 := (y-2)(x+y-4) \\ g_4 := (x-3)(x+y-4) \end{cases} \tag{1}$$

and that this system has no other solution. Although the system $g_2 = g_3 = g_4 = 0$ has the same set of solutions, we keep $g_1$ to underline the fact that $g_3$ and $g_4$ "cut" a subset of points out of the grid generated by $g_1 = g_2 = 0$.

Next, consider a "polynomial" combination of the $g_i$, $i = 1, \ldots, 4$, namely $h(x,y) = \sum_{i=1}^4 f_i(x,y)g_i(x,y)$ with $f_i$, $i = 1, \ldots, 4$, a polynomial. Each point in $\mathcal{A}$ is a zero of $h$. Hence, $\mathcal{A}$ is a subset of the zero set of any polynomial in

$$\text{Ideal}(\mathcal{A}) = \left\{ \sum_{i=1}^{4} f_i(x,y)g_i(x,y) : f_i \text{ polynomial } i = 1, \ldots, 4 \right\}.$$

It can be shown that also the converse holds, i.e. if $\mathcal{A}$ is in the zero set of a polynomial, then that polynomial is in Ideal($\mathcal{A}$). In this sense Ideal($\mathcal{A}$) is the algebraic object corresponding to $\mathcal{A}$. It is called the design ideal of $\mathcal{A}$ and the polynomials $g_1, g_2, g_3, g_4$ are a set of generators of Ideal($\mathcal{A}$). Generator sets are not unique. The set Ideal($\mathcal{A}$) has the algebraic structure of an ideal; more specifically it is a polynomial ideal. A polynomial ideal is a subset of polynomials which is closed under summation and under product with any polynomial; for example, for $f, g \in$ Ideal($\mathcal{A}$) and for any polynomial $s$ then both $f + g$ and $sf$ belong to Ideal($\mathcal{A}$).

Before generalising the above to any design, observe that by considering only the support of a design measure, we implicitly assume that each point in the design is counted only once. The theory we are summarising can be generalised to designs with replicates, but we do not consider this to simplify the presentation.

In algebraic geometry the set of points in $k$ dimensions is called the affine $k$-dimensional space $\mathbb{A}^k$ over the ground field $\mathcal{K}$ (to which the point coordinates belong). A finite set of points in $\mathbb{A}^k$ is a zero-dimensional, reduced, algebraic set, where, roughly, "zero-dimensional" means "points", "reduced" stands for "distinct" and "algebraic set" means "zero set of a system of polynomial equations". A polynomial ideal is associated to an algebraic variety. As the points are distinct the ideal is radical. (For a technical definition of radical ideal and other algebraic notions see e.g. [22,44]).

Let $\mathcal{D} \subset \mathbb{A}^k$ be a finite set of $n$ distinct points. Assume that the point coordinates are in a field $\mathcal{K}$, for example the real numbers $\mathbb{R}$, rational numbers, or complex numbers $\mathbb{C}$. Let $\mathcal{K}[x_1, \ldots, x_k]$ be the set of all polynomials in the variables $x_1, \ldots, x_k$ with coefficients in $\mathcal{K}$. Then define

$$\text{Ideal}(\mathcal{D}) = \{f \in \mathcal{K}[x_1, \ldots, x_k] : f(d) = 0 \text{ for all } d \in \mathcal{D}\}.$$

Observe that Ideal($\mathcal{D}$) is a polynomial ideal. The important Hilbert basis theorem states that any polynomial ideal is finitely generated. This means that there exist

$g_1, \ldots, g_m \in \mathcal{K}[x_1, \ldots, x_k]$ such that $\mathcal{D}$ is the zero set of $g_1 = \ldots = g_m = 0$ and $\operatorname{Ideal}(\mathcal{D}) = \{\sum_{i=1}^m f_i g_i : f_i \in \mathcal{K}[x_1, \ldots, x_k]\}$.

Importantly, $\operatorname{Ideal}(\mathcal{D})$ is computable from the coordinates of the points in $\mathcal{D}$. In [57] it is shown how to perform this computation in Maple. The freely available computer algebra software CoCoA provides the function `IdealOfPoints` which receives in input the coordinates of the design points and returns the associated ideal [3]. In these systems of symbolic computations an ideal is represented by any of its generator sets, that is by a finite list of polynomials. The geometric structure of the design can provide insight on a generating set of the design ideal. For example, an obvious generator set of a full factorial grid in $k$-dimensions with levels $\{-1, 0, 1\}$ is a set of $k$ polynomials each one in a different variable, namely $\{t(t^2 - 1) : t = x_1, \ldots, x_k\}$. Actually, to perform most computations we do not need to know the point coordinates but a generating set is sufficient. This has potential for being exploited in the planning phase of an experiment, before taking the actual measurements, but a systematic treatment of this idea is not available yet in the literature.

## 2.2 The space of functions over a design

In Section 2.1 we identified a design with a polynomial ideal. Now, we consider the set of polynomial functions over $\mathcal{D}$ with values in $\mathcal{K}$, namely $\mathcal{L} = \{f : \mathcal{D} \longrightarrow \mathcal{K} : f \text{ function}\}$. As $\mathcal{D}$ is a finite set, $\mathcal{L}$ can be identified with the set of polynomial interpolating functions over $\mathcal{D}$. In algebraic geometry it is called the coordinate ring of $\mathcal{D}$ and is indicated with the symbol $\mathcal{K}[\mathcal{D}]$.

Importantly, $\mathcal{K}[\mathcal{D}]$ is isomorphic to a computable set of polynomials (e.g. [22, Ch.5§4]). This is the quotient ring modulo the design ideal, $\mathcal{K}[x_1, \ldots, x_k]/\operatorname{Ideal}(\mathcal{D})$, whose elements are the equivalence classes of polynomials defined by the equivalence relationship for which $f$ and $g \in \mathcal{K}[x_1, \ldots, x_k]$ are equivalent modulo $\operatorname{Ideal}(\mathcal{D})$ if and only if $f(d) = g(d)$ for all $d \in \mathcal{D}$. That is, two functions are in the same equivalence class, and hence are identified over $\mathcal{D}$, if they take the same values over the design points. Computations in the quotient ring can be performed using a Gröbner basis of $\operatorname{Ideal}(\mathcal{D})$.

Most computations in algebraic statistics rely upon Gröbner bases. Their definition seems to be purely algebraic and not to have a direct statistical counterpart. Four fundamental facts about Gröbner bases are: they are particular generating sets of $\operatorname{Ideal}(\mathcal{D})$; they are computable from every generating set of $\operatorname{Ideal}(\mathcal{D})$ (e.g. in Maple and CoCoA); their definition depends on the notion of monomial ordering; and, although there are infinitely many monomial orderings, for any polynomial ideal there are finitely many Gröbner bases [51]. Furthermore, if $G$ is a Gröbner basis there exists a finite subset of $G$ which is a Gröbner basis.

A monomial in $\mathcal{K}[x_1, \ldots, x_k]$ is a particular polynomial of the form $x^\alpha = x_1^{\alpha_1} \ldots x_k^{\alpha_k}$, with $\alpha_1, \ldots, \alpha_k$ non-negative integers. It is identified by the list of its exponents $\alpha = (\alpha_1, \ldots, \alpha_k)$, and hence, it can be visualised as a point in the $k$-dimensional grid of non-negative integers.

**Definition 1** A monomial ordering is a total order on this grid such that (a) $(0, \ldots, 0) \prec (\alpha_1, \ldots, \alpha_k)$ and (b) if $\alpha \prec \beta$ then $\alpha + \gamma \prec \beta + \gamma$ for all $\alpha, \beta, \gamma$ $k$-dimensional vectors with non-negative integer components.

Item (a) means $1 \prec x^\alpha$ for all $\alpha$ and Item (b) means that a monomial ordering is compatible with simplification of monomials, e.g. as $x_1^2x_2/x_1x_2 = x_1$ is a monomial, hence $x_1x_2 \prec x_1^2x_2$. Once a monomial ordering has been chosen, the leading term of a polynomial $f$ is the largest monomial in $f$, e.g. the leading term of $f = x_1^2x_2 + 3x_1x_2^2$ is $x_1^2x_2$ for any monomial ordering for which $x_2 \prec x_1$.

A polynomial ideal is called a monomial ideal if it admits a generating set formed by monomials. A version of the Hilbert basis theorem assures that this generating set is finite. We give the definition of a Gröbner basis for a general polynomial ideal $I \subseteq \mathcal{K}[x_1, \dots, x_k]$. A subset $G$ of $I$ is a Gröbner basis if two particular monomial ideals are equal, one constructed from $I$ and one from $G$.

**Definition 2** A Gröbner basis of $I$ is a generating set $G$ such that the ideal generated by the leading terms of the polynomials of $G$ is equal to the ideal generated by the leading terms of all polynomials in $I$.

So far the set of functions over $\mathcal{D}$ has been identified with the quotient space $\mathcal{K}[x_1, \dots, x_k]/\operatorname{Ideal}(\mathcal{D})$ and we indicated that Gröbner bases are a tool to perform computation over it. Observe that $\mathcal{L}$ is a vector space over $\mathcal{K}$ and that the indicator functions in $\mathcal{D}$ of each of its points form a vector space basis of $\mathcal{L}$. Theorem 1 describes how to compute vector space bases of $\mathcal{K}[x_1, \dots, x_k]/\operatorname{Ideal}(\mathcal{D})$, and hence of $\mathcal{L}$, formed by monomials. For the proof see e.g. [22, Ch.5§3Prop.1]. Clearly a subset of one such basis can be used as regression functions of an identifiable, linear, regression model.

**Theorem 1** *Let $G$ be a Gröbner basis of* $\operatorname{Ideal}(\mathcal{D})$ *with respect to a monomial ordering. The set of monomials which are not divisible by the leading terms of the elements of $G$ form a vector space basis of $\mathcal{L}$.*

There are three important facts about any basis obtained in Theorem 1, call it $E$. It is formed by monomials; it has as many elements as there are points in $\mathcal{D}$; and if it contains a monomial then it contains any monomial that divides it. We say that the model is hierarchical. In statistical practice often interaction terms are included in a regression models, only in the presence of linear terms.

*Example 1* It can be shown that $\{g_1, g_2, g_3, g_4\}$ in Equation (1) together with $g_5$ below form a Gröbner basis of Ideal($\mathcal{A}$) for any term ordering. The underlined terms below are the leading terms for any term ordering for which $x \prec y$ and $y \prec x^3$

$$\begin{cases} g_1 := \underline{x^4} - 9x^3 + 26x^2 - 24x \\ g_2 := \underline{y^3} - 6y^2 + 8y \\ g_3 := \underline{y^2} + yx - 6y - 2x + 8 \\ g_4 := \underline{yx} + x^2 - 3y - 7x + 12 \\ g_5 := \underline{x^3} - 6x^2 + 3y + 11x - 12 \end{cases}$$

For example, in $g_3$ we have $x \prec y$ by assumption, and we have $y \prec xy$ and $xy \prec y^2$ by Definition 1(b). Then according to Theorem 1, $E = \{1, x, x^2, y\}$ is a maximal set of linearly independent functions over $\mathcal{A}$. In Figure 2 the empty dots are the (exponents of) $E$ and the full dots are the leading terms of $g_3, g_4, g_5$; there is no need to draw the leading terms of $g_1, g_2$. Clearly we could take $\mathcal{K} = \mathbb{Q}$, the set of rational numbers. It can be noted that the polynomials $g_2, g_3, g_4$ form a Gröbner basis if $y \prec x$ and that the polynomial $g_1$ is worthless for Gröbner basis computation

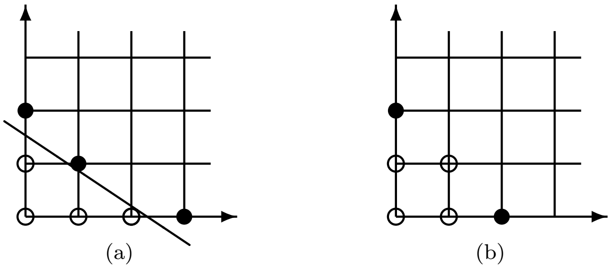

**Fig. 2** (a) A corner cut model and (b) Not a corner cut model

Display (2) gives a summary. In the left most column the design points are listed. In the next four columns the monomials in $E$ are evaluated at the design points giving a 4-by-4 invertible matrix. In the sixth column it is observed that the polynomial $x^3y - xy^3$ takes on $\mathcal{D}$ the same values as $30y + 30x - 120$ which is a linear combination of the elements of $E$ and hence belongs to $\mathbb{R}[x, y]/\operatorname{Ideal}(\mathcal{D})$.

|  |  |  |  |  | NormalForm$(x^3y - y^3x) =$ |  |
|---|---|---|---|---|---|---|
|  | 1 | $x$ | $x^2$ | $y$ | $30y + 30x - 120$ | $\frac{15}{8}2^{x+y} - 30$ |
| $(0, 4)$ | 1 | 0 | 0 | 4 | 0 | 0 |
| $(2, 2)$ | 1 | 2 | 4 | 2 | 0 | 0 |
| $(4, 0)$ | 1 | 4 | 16 | 0 | 0 | 0 |
| $(3, 2)$ | 1 | 3 | 9 | 2 | 30 | 30 |

(2)

The "normal form" operation, which again is implemented in standard software, allows the computation of the element of the quotient space equivalent to a given polynomial (see e.g. [55]). It generalizes the Euclidean division to multivariate polynomials and its usefulness is not to be underestimated, for example it is a fundamental tool in the change of bases in $\mathcal{L}$ from the indicator function basis to a monomial basis.

The last column suggests the obvious observation that the non-polynomial function $30 - \frac{15}{8}2^x2^y$ is identified with $x^3y - xy^3$ over $\mathcal{D}$. In general, the technology described here does not give us any information on the structure of the model outside the design points, although it can be adapted to some special types of regression models (e.g. see [17] for trigonometric regression models; see also [45] for wavelet models in signal processing).

## 2.3 Fans of a design

In Example 1 we chose a monomial ordering for which $x \prec y$ and $y \prec x^3$. The obtained model is a corner cut model; that is, a model where the leading terms of the Gröbner basis in Theorem 1 are separated by a hyperplane from the $E$-basis. See Figure 2(a).

If we choose a monomial ordering for which $y \prec x$, then by Theorem 1 we find $E = \{1, x, y, y^2\}$, again a corner cut model. For an ordering for which $x \prec y$ and $x^3 \prec y$ we would obtain $E = \{1, x, x^2, x^3\}$. These three cases cover all possible monomial orderings. Note however that $\mathcal{A} = \{(0, 4), (2, 2), (4, 0), (3, 2)\}$ identifies the (saturated) model $\{1, x, y, xy\}$ in Figure 2(b). That is, $\{1, x, y, xy\}$ is a monomial vector space basis of $\mathcal{L}$ but it is not corner cut. The set of all hierarchical models obtained by varying the monomial ordering is called the Gröbner fan of the design, whilst the set of saturated hierarchical models identified by a design is called the statistical fan of the design. The relationship between Gröbner and statistical fans is studied in [47].

A design in $\mathbb{R}^k$ with $n$ distinct points is called generic if its Gröbner fan is the set of all corner cuts formed by $n$ monomials in $\mathbb{R}[x_1, \ldots, x_k]$ [52]. As a rule of thumb, designs which present geometrical symmetries are not generic. For example, full factorial designs are not generic, and also their statistical and algebraic fans are equal.

## 2.4 Example: models for compositional data

A mixture design $\mathcal{D}$ in $\mathbb{R}^k$ is a finite set of points in $(k-1)$-simplex. That is, there is a functional constraints on the treatment combinations. Hence

$$\underline{x_1} + x_2 + \ldots + x_k - 1 \in \text{Ideal}(\mathcal{D}).$$

In particular with Theorem 1 we retrieve only slack models, where one factor is not present at all, equivalently it is totally confounded with the other factors.

The trick is to work within a projective framework as the focus is on the relative proportions of the components in the experimental settings. There are also mathematical reasons by which some theorems and ideas are better expressed and understood within a projective algebraic geometry set-up than an affine set-up, e.g. Bézout's theorem on the number of intersection points between two plane curves. The projective counterpart of the theory in Sections 2.1, 2.2 is fairly intuitive. For a full development, which address issues about experiments with mixtures raised in [65,66], see [48,49], while here we illustrate it with an example.

*Example 2* The mixture design $\mathcal{B} = \left\{(1,0,0), (0,1,0), (0,0,1), (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})\right\}$ is identified with its "cone", namely $\{(a,0,0), (0,b,0), (0,0,c), (d,d,d) : a,b,c,d$ real numbers$\}$. The role of Ideal$(\mathcal{B})$ is taken by the ideal generated by the homogeneous polynomials $g_1, g_2, g_3$ with $g_1 = \underline{x_1x_3} - x_2x_3, g_2 = \underline{x_1x_2} - x_2x_3, g_3 = \underline{x_2^2x_3} - x_2x_3^2$. This is the vanishing ideal of the cone. The projective analogue of Theorem 1 returns a set of linearly independent monomials over $\mathcal{B}$ all of the same degree. Underlined are the leading terms with respect to any monomial ordering for which $x_3 \prec x_2 \prec x_1$.

## 2.5 Indicator functions

Another polynomial representation of a design turned out to be useful. Take $\mathcal{K} = \mathbb{R}$. Here the design $\mathcal{F}$ of interest is seen as a subset of a larger design $\mathcal{D} \subset \mathbb{R}^k$and it is represented by the indicator function of $\mathcal{F}$ in $\mathcal{D}$, namely $F : \mathcal{D} \longrightarrow \{0,1\}$ defined as

$$F(d) = \begin{cases} 1 \text{ if } d \in \mathcal{F} \\ 0 \text{ if } d \in \mathcal{D} \setminus \mathcal{F}. \end{cases}$$

Let $E$ be a basis of $\mathbb{R}[x_1, \ldots, x_k]/\operatorname{Ideal}\mathcal{D}$ as in Theorem 1 and let $L$ be the set of exponent vectors of the elements of $E$, e.g. for $k = 2$ and $\{1, x_1, x_1x_2\} \subset E$ then $(0,0), (1,0), (1,1) \in L$. As $F$ is a function over $\mathcal{D}$, then there exists a polynomial representation of $F$ of the form $F = \sum_{\alpha \in L} b_\alpha x^\alpha \in \mathbb{R}[x_1, \ldots, x_k]/\operatorname{Ideal}(\mathcal{D})$.

Often $\mathcal{D}$ is a full factorial design and the levels of a factor with $k$ levels are coded using the $k$th (complex) roots of unity, solutions to $z^k = 1$. In this case, the indicator function is a real valued function with complex coefficients and $\mathcal{K} = \mathbb{C}$. The

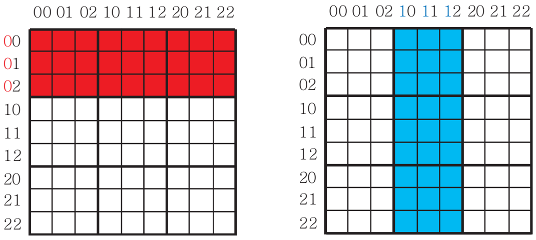

**Fig. 3** A band and a stack for $p = 3$.

complex coding is not unusual in design of experiments, e.g. [7], and the group structure implied by this coding can be exploited. From a polynomial algebra viewpoint it is advantageous as the set of complex numbers is algebraically closed; equivalently every univariate polynomial is the product of degree one polynomials. Theorems can be proved in algebraic geometry over algebraically closed fields which do not hold otherwise.

The coefficient of $x^\alpha$ can be computed from the values of the (response) function $x^\alpha$ at $\mathcal{F}$[1], namely

$$b_\alpha = \frac{1}{\#\mathcal{D}} \sum_{d \in \mathcal{F}} \overline{x^\alpha(d)}$$

where $\bar{c}$ is the conjugate of a complex number $c$ and $\#D$ is the sample size. More importantly, the coefficients of $F$ embed statistically relevant information on $\mathcal{F}$. For example, orthogonality among factors and interactions, projectivity, aberration and regularity correspond to the fact that suitable subsets of coefficients of $F$ are zero [34, 56, 70]. See also [35] for an overview.

### 2.6 Example: algebraic theory of sudoku

In a recent paper Bailey et al. [8] discuss the mathematics of sudoku puzzles and show that their solutions are examples of "gerechte design". These are particular Latin square designs. A square matrix of size $p^2 \times p^2$ is divided in $p^2$ regions, each containing $p^2$ cells; and $p^2$ symbols are to be allocated so that each symbol occurs once in each region. In the forthcoming paper [35], outlined below, Fontana and Rogantin study the indicator function for the solutions to sudoku puzzles.

A sudoku game is described by four factors with $p^2$ levels each: $R, C, B$ and $S$ for rows, columns, boxes and symbols, respectively. With a typical trick in design of experiment, a factor with $p^2$ levels is split into two pseudo-factors with $p$ levels each. Figure 3 shows that $R$ splits into $(R_1, R_2)$ where $R_1$ identifies the "band" and $R_2$ identifies a row within a "band". Analogously, $C = (C_1, C_2)$ where $C_1$ identifies the "stack" and $C_2$ the column within a "stack". Note that the three "position" factors are not linearly independent as the box factor $B$ is identified by $(R_1, C_1)$. The symbol $S$ could be coded resorting to pseudo factors $(S_1, S_2)$, but here this is not necessary.

*Example 3* For $p = 3$ the digit 5 in the top left corner of the grid in Figure 3 is encoded by $(r_1, r_2, c_1, c_2, s) = (0, 0, 0, 0, 5)$.

[1] The abuse of notation by which $x^\alpha$ is both a monomial in $\mathbb{R}[x_1, \ldots, x_k]$ and the corresponding monomial function in $\mathbb{R}[\mathcal{D}]$ is typical.

Solutions to sudoku puzzles are identified with particular fractions of a full factorial design $\mathcal{D}$ in 5 factors, where the first four factors have $p$ levels coded by the $p$th root of unity and the last factor has $p^2$ levels. Following Section 2.5 the set of exponents $L$ is $L = \{(\alpha_1, \dots, \alpha_5) : 0 \leq \alpha_i \leq 2 \text{ for } i = 1, \dots, 4, \text{ and } \alpha_5 = 0, \dots, 8\}$ and the indicator function of a solution to a sudoku puzzle is a linear combination of the type $F = \sum_{\alpha \in L} b_\alpha x^\alpha$.

The game rules translate into values of the coefficients of $F$, for example "each symbol appears exactly once in each row" if and only if $R_1 \times R_2 \times S$ is a full factorial design; that is, if and only if $b_{(r_1,r_2,0,0,s)} = 0$ for all $(r_1, r_2, 0, 0, s) \in L$. The other rules are similar and listed in [35]. Let $M$ be the collection of all these indices.

The solutions to the sudoku puzzle are all and only the solutions of the following system of polynomial equations in the coefficients on the indicator function $F$

$$\begin{cases} b_\alpha = \sum_{\beta \in L} b_\beta \, b_{[\alpha-\beta]} & \text{with } \alpha \in M \\ b_\alpha = 0 & \text{with } \alpha \in L \setminus M. \end{cases}$$

It is known that some operations, e.g. permutation of symbols, transforms a solution to a sudoku puzzle to another solution. In [35] it is shown that this is equivalent to keep fixed some margins of a contingency table which again represents a sudoku puzzle. This makes a natural link to the other paper at the foundations of algebraic statistics which is summarised in Section 3. On this link see also [5].

## 3 Markov Bases of Log-linear Models

Now we turn to the application of algebraic geometry techniques to exact conditional inference in contingency tables. Useful references are [24,26,58]. It provides an MCMC procedure to obtain a sample from a conditional distribution of a discrete exponential family given the sufficient statistics and can be used in a variety of applications, including hypothesis testing for log-linear models, hyper-geometric sampling, multinomial rule of succession with conditioning on incomplete information.

At the basis there is the so-called Diaconis-Sturmfels algorithm which provides a method for constructing a symmetric and irreducible Markov chain on an intersection of hyperplanes in $\mathbb{Z}^k_{\geq 0}$ and is used to handle practical problems in categorical tables. (Here $\mathbb{Z}_{\geq 0}$ is the set of non-negative integers and $\mathbb{Z}$ the set of integer numbers).

The finite set $\mathcal{D} \subset \mathbb{R}^k$ represents the cells of a contingency tables or any other finite set of interest. We consider contingency tables to be in a familiar framework, but $\mathcal{D}$ could be any finite set on which to define a statistical model. We assume the log-linear model

$$p(x; \theta) = Z(\theta) \exp\Big(\sum_{i=1}^{d} \theta_i T_i(x)\Big) \qquad \text{for } x \in \mathcal{D}$$

with parameter vector $\theta^t = [\theta_1, \dots, \theta_d] \in \mathbb{R}^d$ and with sufficient statistics the integer valued function $T : \mathcal{D} \longrightarrow \mathbb{Z}^d_{\geq 0} \setminus \{0\}$, where $A^t$ is the transposed matrix of $A$. In particular, for $N$ independent draws from $p(\cdot; \theta)$, the statistics $T = \sum_{i=1}^{N} T(x_i)$ is sufficient for $\theta$. At this stage of the research, it seems essential that the sufficient statistics take non-negative integer values. Before going into the formalism of [24], in Examples 4 and 5 we outline the very basics following Section 6 of [55].

*Example 4* For $\mathcal{D} = \{(0,0),(0,1),(1,0),(1,1)\} \in \mathbb{R}^2$, from Theorem 1 and for any monomial ordering we have $E = \{1, x, y, xy\}$. Then, the saturated exponential model on $\mathcal{D}$ is $\exp\{\theta_{00} + \theta_{10}x + \theta_{01}y + \theta_{11}xy\}$ where $\theta_{00}$ is the cumulant function depending on $\theta_{10}, \theta_{01}, \theta_{11}$. Assume the exponential submodel

$$p(x,y;\psi) = \exp\{\psi_{10}x + \psi_{01}y - K(\psi_{01},\psi_{10})\} \tag{3}$$

with sufficient statistics $T(x,y) = (x,y)$ and $K$ the normalising factor. Its intrinsically polynomial structure is evident in the parametrization $\zeta_{00} = \exp\{-K(\psi_{01},\psi_{10})\}$, $\zeta_{10} = \exp\{\psi_{10}\}$ and $\zeta_{01} = \exp\{\psi_{01}\}$ and it is

$$p(x,y) = \zeta_{00}\zeta_{10}^{x}\zeta_{01}^{y} \qquad \text{for}(x,y) \in \mathcal{D}.$$

*Example 5* Model (3) means that the integer valued vector $[\log p(x,y)]_{(x,y)\in\mathcal{D}}$ belongs to the span of the following (evaluation) matrix

|  | $1$ | $x$ | $y$ |
|---|---|---|---|
| $(0,0)$ | 1 | 0 | 0 |
| $(0,1)$ | 1 | 0 | 1 |
| $(1,0)$ | 1 | 1 | 0 |
| $(1,1)$ | 1 | 1 | 1 |

Call it $Z_1$ and consider its orthogonal, here $Z_2 = [1,-1,-1,1]^t$. Then, Model (3) is orthogonal to $Z_2$ and equivalently

$$Z_2^t[\log p(x,y)]_{(x,y)\in\mathcal{D}} = 0. \tag{4}$$

Now, consider the positive and negative parts of $Z_2$, so that $Z_2 = Z_2^+ - Z_2^-$; here $Z_2^+ = [1,0,0,1]^t$ and $Z_2^- = [0,1,1,0]^t$. Then, Equation (4) becomes

$$(Z_2^+)^t[\log p(x,y)]_{(x,y)\in\mathcal{D}} = (Z_2^-)^t[\log p(x,y)]_{(x,y)\in\mathcal{D}}$$

and taking logarithm it becomes $p(0,0)p(1,1) = p(1,0)p(0,1)$ which is a polynomial invariant of the considered model, under the assumption of strict positivity. This leads into the toric representation of exponential models which we shall consider again in Section 4.

### 3.1 Markov bases

For a value $t$ of $T$ consider $\mathcal{F}_t = \left\{f : \mathcal{D} \to \mathbb{Z}_{>0} : \sum_{x\in\mathcal{D}} f(x)T(x) = t\right\} \subset \mathbb{R}[\mathcal{D}]$ and $\mathcal{Y}_t = \left\{(x_1,\ldots,x_N) \in \mathcal{D}^N : T(x_1) + \ldots + T(x_N) = t\right\}$. The fiber $\mathcal{F}_t$ is the set of tables for which the sufficient statistics is equal to $t$ and $\mathcal{Y}_t$ is the set of samples with fixed values of the sufficient statistics.

One would like to enumerate $\mathcal{Y}_t$ or sample from the uniform distribution over $\mathcal{Y}_t$. This is difficult for reasonable size problems and instead one samples from the hyper-geometric distribution over $\mathcal{F}_t$. One could see that this is possible by considering the map which associates to a sample its contingency table, $\psi : \mathcal{Y}_t \longrightarrow \mathcal{F}_t$ defined by $\psi(x_1,\ldots,x_N) = \sum_{x\in\mathcal{D}} e_x \sum_{k=1}^{N} \mathbf{1}_x(x_i)$ where $(e_x)_{x\in\mathcal{D}}$ is the canonical basis in $\mathbb{R}^{\mathcal{D}}$ and $\mathbf{1}_x$ the indicator function of $x \in \mathcal{D}$ (see [13,14,58]). The Diaconis-Sturmfels algorithm works over $\mathcal{F}_t$ and hinges on Markov bases.

**Definition 3** A Markov basis is a set of functions $f_1, \dots, f_m : \mathcal{D} \longrightarrow \mathbb{Z}$ such that (a) $\sum_{x \in \mathcal{D}} f_i(x)T(x) = 0$ for $i = 1, \dots, m$ and (b) if $f, f' \in \mathcal{F}_t$ then there are $A \in \{1, \dots, m\}$ and $e_j = \pm 1$ $(j = 1, \dots, A)$ such that $f' = f + \sum_{i=1}^{A} e_j f_{i_j}$ and $f + \sum_{i=1}^{a} e_j f_{i_j} \geq 0$ for all $a \leq A$.

Item (b) means that there is a path from $f$ to $f'$ which preserves $\mathcal{F}_t$. From a Markov basis it is possible to construct a stationary Markov chain of $\mathcal{F}_t$ with transition matrix

$$\begin{aligned} \pi(f, f + f_i) &= 1/(2m) \quad \text{if } f + f_i \geq 0 \\ \pi(f, f - f_i) &= 1/(2m) \quad \text{if } f - f_i \geq 0. \end{aligned}$$

This is an irreducible, aperiodic, Markov chain with stationary distribution the hypergeometric distribution over $\mathcal{F}_t$ (see [24, Lemma 2.1]).

*Example 6* The margins of 2by2 tables are preserved by the basic move $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$, for example

$$\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

where $t = (6, 4, 5, 5)$, has the same row and column margins as $\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix}$. We could have chosen $\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$. Which other basic moves preserve the margin?

### 3.2 Algebraic formalism and toric ideals

To $x \in \mathcal{D}$ associate an indeterminate $p_x$, define $\mathbf{p} = (p_x)_{x \in \mathcal{D}}$ and consider the polynomial ring $\mathbb{R}[\mathbf{p}]$ which has as many variables as cells in the table. The following maps translate "statistical objects" into polynomials in $\mathbb{R}[\mathbf{p}]$.

1. A non-negative, integer valued function $f : \mathcal{D} \to \mathbb{N}$ is represented by the monomial $\prod_{x \in \mathcal{D}} p_x^{f(x)} = \mathbf{p}^{f(x)}$.

   *Example 7* For $\mathcal{D} = \{x_1, x_2, x_3, x_4\}$, the function $(x_1, x_2, x_3, x_4) \to (2, 4, 3, 1)$ goes into $p_1^2 p_2^4 p_3^3 p_4$. (We simplified the notation by setting $p_{x_i} = p_i$.) Here $\mathcal{D}$ could represent the cells of a 2by2 table.

2. An integer valued function $f : \mathcal{D} \to \mathbb{Z}$ is represented by the binomial $\mathbf{p}^{f^+(x)} - \mathbf{p}^{f^-(x)}$ where $f^+$ is the positive part of $f$ and $f^-$ the negative part.

   *Example 8* The function $(x_1, x_2, x_3, x_4) \to (1, -1, -1, 1)$ becomes $p_1 p_4 - p_2 p_3$.

3. We need to introduce other indeterminates $t_1, \dots, t_d$ and $\mathbf{t} = (t_1, \dots, t_d)$. The multi-valued function

   $$T : \mathcal{D} \longrightarrow \mathbb{N}^d \setminus \{0\}, \qquad x \longmapsto (T_1(x), \dots, T_d(x))$$

   is represented by the ring-homomorphism

   $$\phi_T : \mathbb{R}[\mathcal{D}] \longrightarrow \mathbb{R}[t_1, \dots, t_d], \qquad \mathbf{1}_x \longmapsto t_1^{T_1(x)} \dots t_d^{T_d(x)}$$

   where $x$ is a point in $\mathcal{D}$ and $\mathbf{1}_x$ the indicator function of $x$.[2] Write $\mathbf{t}^{T(x)}$ for the monomial $t_1^{T_1(x)} \dots t_d^{T_d(x)}$.

[2] We could have considered the monomial function $x \in \mathbb{R}[\mathcal{D}]$ instead of $\mathbf{1}_x$.

Let $I_T$ be the kernel of $\phi_T$, namely $I_T = \{f \in \mathbb{R}[\mathcal{D}] : \phi_T(f) = 0\}$. Three relevant facts about $I_T$ are: $\sum_x f(x)T(x) = 0$ if and only if $\mathbf{p}^{f^+(x)} - \mathbf{p}^{f^-(x)} \in I_T$; it is the set of polynomials in the $\mathbf{p}$ indeterminates that vanish on the monomials $\{\mathbf{t}^{T(x)} : x \in \mathcal{D}\}$; and it is a polynomial ideal, which is generated by homogeneous binomials. These are called toric ideals. Theorem 2 states that a set of generators of $I_T$ corresponds to a Markov basis. For the proof see [24, §3]. Many papers are devoted to the computations of Markov bases for specific applications [6,27].

**Theorem 2** *Let $\mathcal{D}$ be a finite set in $\mathbb{R}^k$. The set of functions $\{f_1, \ldots, f_m\} : \mathcal{D} \to \mathbb{Z}$ is a Markov basis if and only if the set of monomial differences $\mathbf{p}^{f_i^+(x)} - \mathbf{p}^{f_i^-(x)}$, $i = 1, \ldots, m$ generates the ideal $I_T$.*

### 3.3 Contingency tables: a summary

Section 3.2 is based on identifying an integer vector of length $k$ by a monomial difference in $k$ variables. Then, a contingency table with non-negative integer entries translates into (the exponents of) a monomial in as many variables as entries in the table. Also a normalised contingency table with $n = n_1 \times \ldots \times n_k$ cells can be translated into a symbolic, polynomial framework in various ways. This table can be viewed as

1. a $k$-dimensional array with entries in $[0, 1]$, e.g. $\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$,
2. a point in the $(n-1)$-simplex, e.g. $(p_{00}, p_{01}, p_{10}, p_{11})$,
3. a function defined on its cells into $[0, 1]$ satisfying the constraints, e.g. $p_{00} + p_{01} + p_{10} + p_{11} = 1$ and $p_{ij} \geq 0$.

Let $\mathcal{D}$ be the set of cells of the table. From Item 3. above a contingency table is a function $\mathcal{D} \longrightarrow [0, 1] \subset \mathbb{R}$. Hence there exists a unique polynomial in the quotient space identifying the table, call it $\sum_{\alpha \in L} \theta_\alpha x^\alpha \in \mathbb{R}[x_1, \ldots, x_k]/\operatorname{Ideal}(\mathcal{D})$. Note that Items 2. and 3., as well as Section 3, do not require a "cross-product" structure, e.g. cells with structural zeros can simply not be included in the representational framework. For pointers to applications of the Diaconis-Sturmfels algorithm in the presence of structural zeros see Section 5.

## 4 Algebraic statistical models

Densities, random variables, statistical models, probabilistic models, etc. are functions defined from a sample space $\mathcal{D}$ to a suitable space $\mathcal{K}$. We have seen in Section 2.1 that when $\mathcal{D}$ is a finite set, then for each of these statistical/probabilistic objects there is an object in the coordinate ring and a polynomial in $\mathcal{K}[x_1, \ldots, x_k]/\operatorname{Ideal}(\mathcal{D})$ representing it. This isomorphism actually holds also for other algebraic varieties, not necessarily a finite set of points [22, Ch.5§4]. Furthermore, different polynomial representations can be used for the same statistical object, e.g. contingency tables in Section 3.3.

Algebraic statistical models for discrete sample spaces, first introduced in [55, Ch.6], are defined as statistical models which are algebraic varieties with respect to some parametrization. It is required that there is a parametrization under which the model is the solution set of a system of polynomial equations. This ignores the condition of non-negativity required by some parametrizations and postpones the checking of

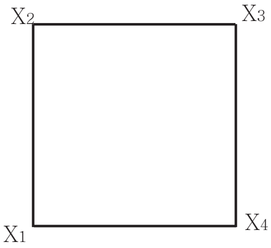

**Fig. 4** Four cycle model

non-negativity to a subsequent analysis phase. (Note that the sum-to-one condition of probability densities is algebraic, namely the polynomial equation $\sum_{x \in \mathcal{D}} p_x = 1$.) Example 10 considers a graphical model and indicates the role of the algebraic operation of elimination in change of parametrization.

The definition of an algebraic statistical model was refined in [30] and resorts to real algebraic geometry [15], which studies solutions over (the field of) real numbers of systems of polynomial equations and, in doing so, studies semi-algebraic sets. These are solutions to a finite number of polynomial equations and polynomial inequalities. A drawback of real algebraic geometry is that, at the moment, the computational advantages are not as developed and widely available as for computational commutative algebra.

**Definition 4** Consider $\Theta \subseteq \mathbb{R}^k$ with no-empty interior and $\{p(\cdot, \theta) : \theta \in \Theta\}$ a family of probability distributions on a sample space parametrised by $\theta$. Then, $\{p(\cdot, \theta) : \theta \in M\}$ is an algebraic statistical model if there exists a semi-algebraic set $A \subseteq \mathbb{R}^k$ such that $M = A \cap \Theta$.

*Example 9* Surfaces of independence are Segre varieties [68]. Bayesian networks, which can be visualised as directed acyclic graphs, whose nodes represent random variables and whose arcs describe a recursive factorisation for their joint distribution, are algebraic statistical models [10,36].

*Example 10* Consider $\mathcal{D} = \{0, 1\}^4$ and the random vector $\mathbf{X} = (X_1, X_2, X_3, X_4)$ on $\mathcal{D}$. Assume for the moment strict positivity of the joint distribution. Strict positivity is obtained algebraically via the operation of saturation [43, §3.5B].) The conditional independence model in Figure 4 states the two conditions: $X_2$ and $X_4$ independent given $X_1$ and $X_3$ and $X_1$ and $X_3$ independent given $X_2$ and $X_4$.

The quotient ring $\mathbb{R}[\mathcal{D}]$ is isomorphic to the polynomial space spanned by the monomials in $\big\{1, x_i, x_i x_j : i \neq j, x_i x_j x_k : i \neq k \neq j, x_1 x_2 x_3 x_4 : i, j, k = 1, \dots, 4\big\}$. Let $L$ be the corresponding exponent set. Consider $\mathcal{M} = \{1, x_1, x_2, x_3, x_4, x_1x_2, x_2x_3, x_3x_4, x_4x_1\}$, let $M$ be the corresponding exponent set and $M_0 = M \setminus \{(0, 0, 0, 0)\}$. It can be shown that the model in Figure 4 is an exponential model of the form

$$\begin{aligned} p(x; \psi) = {} & \exp(\psi_{0000} + \psi_{0100} x_2 + \psi_{0001} x_4) \exp(\psi_{1000} x_1 + \psi_{1100} x_1 x_2 + \psi_{1001} x_1 x_4) \\ & \exp(\psi_{0010} x_3 + \psi_{0110} x_2 x_3 + \psi_{0011} x_3 x_4) \end{aligned}$$

which involves only terms in $\mathcal{M}$.

We distinguish three parametrizations for the $\mathcal{M}$ model:

($p$) the raw probabilities $(p(d), d \in \mathcal{D}) \subset \Delta_{n-1}$;

($\theta$) the vector space parametrization

$$\begin{cases} p_\theta = \theta_{0000} + \sum_{\alpha \in M_0} \theta_\alpha x^\alpha \\ \theta_{0000} = 1 - \sum_{\alpha \in M_0} \theta_\alpha m_\alpha \end{cases}$$

where $m_\alpha = \mathrm{E}_0(X^\alpha)$ is the first moment of $X^\alpha$, $\alpha \in M_0$, with respect to the uniform distribution over $\mathcal{D}$; and

($\zeta$) the toric parametrization

$$\begin{cases} p(x;\psi) &= \exp\left(\sum_{\alpha \in M} \psi_\alpha x^\alpha\right) \\ &= \prod_{\alpha \in M} \exp\left(\psi_\alpha x^\alpha\right) \\ &= \zeta_0 \prod_{\alpha \in M_0} \zeta_\alpha^{x^\alpha} = p(x;\zeta) \end{cases}$$

where $\zeta_\alpha = \exp(\psi_\alpha)$. It is possible to switch parametrization using elimination theory (e.g. [22, Ch3]) which generalises Gaussian elimination for linear polynomials and can be handled with Gröbner basis algorithms (e.g. the macro Elim in CoCoA). For example, elimination of $\zeta$ from the $p$-$\zeta$ equations gets an implicit representation of the four cycle model, given by a set of binomials in the $p(x;\psi)$, $x \in \mathcal{D}$. This leads into the algebra of toric ideals, that is ideals which admit Gröbner bases formed by binomials with terms of the same degree. A full analysis of the four-cycle model with respect to the pairwise, local, global Markov properties for graphical models, can be found in [37]. Below we outline the main result of the related theory.

## 4.1 Toric algebra of graphical models

Let $\#\mathcal{D} = n$, $P = (p_1, \ldots, p_n) \in \Delta_{n-1}$ be a probability distribution on $\mathcal{D}$, $A = \left[a_{ij}\right]_{i=1,\ldots,d, j=1,\ldots,n}$ a $d \times n$-matrix with $a_{ij} \geq 0$ and $\sum_i a_{i1} = \ldots = \sum_i a_{in}$ and consider the map

$$\begin{array}{rccc} \Phi_A : & \mathbb{R}^d_{\geq 0} & \longrightarrow & \mathbb{R}^n_{\geq 0} \\ & (t_1, \ldots, t_d) & \longmapsto & \left(\prod_i t_i^{a_{i1}}, \ldots, \prod_i t_i^{a_{in}}\right). \end{array}$$

In Example 5 the $t_i$'s are the $\zeta$ parameters and $Z_1$ is the $A$ matrix, specifically the columns of $A$ coincide with the sufficient statistics. The change of notation is to follow the literature. A probability $P$ belongs to the model $A$ if and only if the following three equivalent conditions hold

1. $P \in \mathrm{Image}(\Phi_A)$,
2. $P$ factors according to model $A$,
3. $P$ is an exponential family

$$P_\theta(x) = Z(\theta) \exp(\sum_{i=1}^d \theta_i T_i(x))$$

where $Z(\theta)$ is a normalising constant and $T : \mathcal{D} \to \mathbb{Z}^d \setminus \{0\}$ is a sufficient statistics.

Undirected graphical models, log-linear models and other exponential models factor according to model $A$ for some matrix $A$. This is studied in [37] where it is also shown that for decomposable graphical models those conditions are equivalent to the Hammersley-Clifford theorem, while they are not for non-decomposable models. For non-decomposable models it is also shown that the maximum likelihood degree is a non rational number and it is given the following characterisation of distributions that factor according to $A$ in terms of toric varieties.

A subset $F$ of columns of $A$ is feasible if the support of any other column is not contained in the union of the supports of the columns in $F$; for $a_j$ the $j$-th column of $A$, and $F \subset \{1, \ldots, n\}$, $j \in \{1, \ldots, n\} \setminus F$ then $supp(a_j) \not\subset \cup_{l \in F} a_l$. Let $u - v \in \ker(A)$; i.e. $\sum_i u_i a_i = \sum_i v_i a_i$. The non-negative toric variety associated to $A$ is

$$X_A = \left\{(x_1, \ldots, x_n) \in \mathbb{R}^m_{\geq 0} : x_1^{u_1} \ldots x_n^{u_n} = x_1^{v_1} \ldots x_n^{v_n}\right\}.$$

**Theorem 3** *The probability $P$ factors according to the model $A$ if and only if $P \in X_A$ and the support if $P$ is $A$-feasible.*

The assumption of strict positivity in Example 10 can now be dropped.

## 4.2 Causal/intervention models

Discrete models for causal reasoning on Bayesian networks are algebraic statistical models. This observation allows to generalise ideas in [54, 67] to other graphical structures rather than Bayesian networks and to non graphical algebraic statistical models [40, 61–63]. We present the basic idea and refer for motivation and main results to the aforementioned papers.

Consider a single rooted probability tree $\mathcal{T}$ with transition probabilities $\pi$. These can be considered as labels on the directed edges of the tree which respect the laws of probabilities: $\pi = (\pi(w|v) \in [0,1] :$ for all $(v, w)$ directed edges in $\mathcal{T})$ and $\sum_w \pi(w|v) = 1$ where the sum runs over the children of $v$.

*Example 11* In the tree in Figure 5 there are two natural parametrizations: the transition probabilities $\pi_i$, $i = 1, \ldots, 22$ and the probabilities of the root-to-leaf paths: $p_i$, $i = 4, 6, 8, 10 - 12, 14 - 16, 18, 20 - 22$. Linear constraints are $1 = \pi_{20} + \pi_{21} + \pi_{22}$, $\sum_i p_i = 1$ and inequality constraints are clearly $\pi_i, p_i \in [0, 1]$. Polynomial equations map one parametrization into the other, e.g. $p_4 = \pi_1 \pi_4$, and the inverse mapping is defined by ratios of polynomials.

Algebraic statistical models, including some commonly used models, can be imposed by adjoining to the probability constraints some polynomial constraints, $q(\pi) = 0$ with $q$ polynomial, and some polynomial inequalities $r(\pi) > 0$, with $r$ polynomial. For example setting equal some transition probabilities, e.g. $\pi(v_4|v_1) = \pi(v_8|v_3)$ and $\pi_{10} = \pi_{14}$ in Figure 5 leads to a chain event graph [61–63].

A manipulation on the tree (or a submodel) is defined as a new family of probability densities $\hat{\pi} = (\pi(w|v))$ which is function of the transition probabilities $\hat{\pi} = f(\pi)$ where $f$ is a polynomial map. For example, in Figure 5 the equalities $\hat{\pi}(v_4|v_1) = \hat{\pi}(v_8|v_3) = 1$ and $\hat{\pi}(v_5|v_1) = \hat{\pi}(v_9|v_3) = 0$ give a typical manipulation considered in [54]. Compatibility conditions have to be respected between manipulation and submodel. Some (polynomial) functions of the $\hat{\pi}$, $m = m(\hat{\pi})$, are assumed observed or somehow known, and the interest is in identifying an effect of the manipulation. Namely in checking if a polynomial function $e = e(\pi)$ can be written as $e = e(m(\hat{\pi}))$. This type of problems can be handles with elimination theory, although often computations are forbidding.

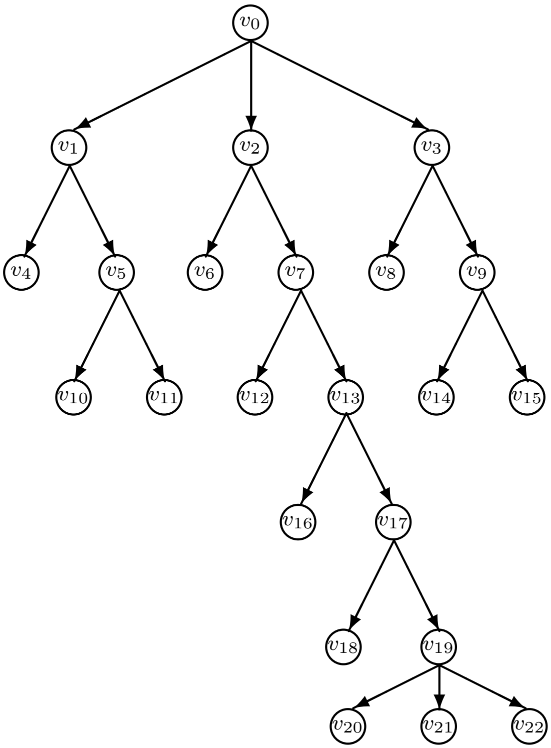

**Fig. 5** A tree with unusual stage set

## 5 Advanced topics

We conclude with hints to more advanced results and applications. The selected topics and the provided references are an incomplete list and follow the taste and interests of the author. Still we believe the provided pointers to the literature are useful indications of promising research areas which have not been fully developed yet.

There is much interest in the application of computational algebra to computational biology similar to the application in statistics in many ways. The volume [53] introduces to the applications of computational commutative algebra in computational biology and includes open problems, some of which have been solved in the meantime, while some remain unanswered. Some models in computational biology, for example those used in phylogenetic tree reconstruction, have a not dissimilar structure to causal models and the relevant statistical problems are often of the identifiability type outlined in Section 4.2. But the amount and type of available data are often very different and present different mathematical and inferential issues. Interesting work in computational biology has been developed by Allman and Rhodes [4] on evolutionary models and Casanellas and co-authors on phylogenetic invariants [19].

The connection between algebraic statistics and information geometry was, to our knowledge, first discussed by Professor G. Pistone at the 2nd International Symposium on Information Geometry and its Applications, Tokyo (2005). He noted that toric models are somehow located between the class of exponential models and the class of algebraic statistical models (see also [37]). This line of research, which considers simultaneously a model as a differential manifold and an algebraic variety, is currently being investigated and a contribution in that direction is the collected volume [38]; relevant is also the forthcoming thematic year at SAMSI 2008-09 “Algebraic Methods in Systems Biology and Statistics”.

In [12] the linear aberration for polynomial models is introduced. This is a new concept in experimental designs for quantitative factors which is an alternative to aberration and is strongly related to corner cut models and state polytopes. Algorithmic and computational aspects are related to non-linear matroid optimization [11]. Another novel idea in design of experiment is to interpret an algebraic variety, not necessarily a zero-dimensional one, as a (repository of potential) designs of experiments [9, 50]. An example is the sum-to-one condition for mixture designs. The challenge is to device methods to sample on the algebraic variety according to appropriate criteria. The baggage of statistical techniques to analyse designs whose point coordinates are “approximately” known, could be supplemented by transferring the algebraic notion of border bases [44] to statistics. The techniques in Section 2 are applied to reverse engineering in particular for the identification of biochemical networks [25].

The Diaconis-Sturmfels algorithm and Markov bases have been applied to address a number of questions related to categorical data, requiring Markov chain Monte Carlo techniques. These include the Bowker test for matched pair data [41], rater agreement test [59], calibration of Fourier analysis of ranking data [23], in model selection [42], sequential sampling in multi-way contingency tables with given constraints [21], for quasi-independence [39] and weaken independence [18], to mention a few. The computation of Markov bases in special cases has received much attention e.g. for decomposable graphical models [27] and in the presence of structural zeros [60]. Fast algorithms to compute Markov bases have been devised [46]. High degree binomials in the Markov bases correspond to long steps in the Markov chain which should be avoided in efficient algorithms. This technical difficulty should be overcome in order to make Markov basis technique effective in a wide variety of applications.

One of the first applications of algebraic statistics was to disclosure limitation problems developed by Professor S.E. Fienberg and co-authors [33]. The algebraic notion of maximum likelihood degree introduced in [20] is arousing some interest of algebraists [31]. Further topics within likelihood based inference have been studied in [29] and maximum likelihood estimator for latent class models in [32].

Although this paper has been mainly concerned with discrete sample space, algebraic techniques have been applied to Gaussian vectors. Polynomial conditions on the entries of the variance-covariance matrix generate a variety in the cone of semi-definitive positive matrices [28, 69].

**Acknowledgements** I am grateful to Professors R. Fontana and M.P. Rogantin for permission to use the material in Section 2.6; to two unknown referees for spotting far too many mistakes, misprints and misleading sentences in the first version of this paper; and to Professors R. Notari and G. Pistone for useful comments on and careful reading of the second version.

## References

1. Statistica Sinica: special issue on Algebraic Statistics and Computational Biology, 17:4 (2007).

2. Journal of Symbolic Computation: special issue on Computational Algebraic Statistics, 41:2 (2006).
3. J. Abbott and A. Bigatti and M. Kreutzer and L. Robbiano, Computing ideals of points, Journal of Symbolic Computation, 30:4, 341–356 (2000).
4. E.S. Allman and J.A. Rhodes, Identifying evolutionary trees and substitution parameters for the general Markov model with invariable sites, Mathematical Biosciences, 211:1, 18-33 (2008).
5. S. Aoki and A. Takemura, Markov bases for design of experiments with three-level factors, In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
6. S. Aoki, A. Takemura and R. Yoshida, Indispensable monomials of toric ideals and Markov bases, Journal of Symbolic Computation, 43, 490–507 (2008).
7. R.A. Bailey, The decomposition of treatment degrees of freedom in quantitative factorial experiments, J. R. Statist. Soc., B 44:1, 63–70 (1982).
8. R.A. Bailey, P.J. Cameron and R. Connelly, Sudoku, Gerechte Designs, Resolutions, Affine Space, Spreads, Reguli, and Hamming Codesread, Ameri- can Math. Monthly (May 2008).
9. T. Becker and V. Weispfenning, The Chinese remainder problem, multivariate interpolation, and Gröbner bases, in ISSAC 91: Proceedings of the 1991 international symposium on symbolic and algebraic computation, 64–69 (1991).
10. N. Beerenwinkel, N. Eriksson and B. Sturmfels, Conjunctive Bayesian networks, Bernoulli 13:4, 893-909 (2007).
11. Y. Berstein, J. Lee, H. Maruri-Aguilar, S. Onn, E. Riccomagno, R. Weismantel, H.P. Wynn, Nonlinear Matroid Optimization and Experimental Design, SIAM Journal on Discrete Mathematics 22:3, 901-919 (2008).
12. Y. Berstein, H. Maruri-Aguilar, S. Onn, E. Riccomagno, H.P. Wynn, Minimal average degree aberration and the state polytope for experimental design, MUCM report no. 07/07.
13. J. Besag and P. Clifford, Generalized Monte Carlo significance tests, Biometrika, 76, 633–642 (1989).
14. Y.M. Bishop, S.E. Fienberg and P.W. Holland, Discrete multivariate analysis: theory and practice, x+557. MIT Press, Cambridge, MA (1977).
15. J. Bochnak, M. Coste and M.F. Roy, Real algebraic geometry, x+430. Springer-Verlag, Berlin (1998).
16. CoCoATeam, CoCoA : a system for doing Computations in Commutative Algebra, Available at http://cocoa.dima.unige.it
17. M. Caboara and E. Riccomagno, An algebraic computational approach to the identifiability of Fourier models, Journal of Symbolic Computation, 26:2, 245–260 (1998).
18. E. Carlini and F. Rapallo, Algebraic modelling of category distinguishability, In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
19. M. Casanellas and J. Fernndez-Snchez, Performance of a new invariants method on homogeneous and non-homogeneous quartet trees, Molecular Biology and Evolution, 24:1, 288–293 (2007).
20. F. Catanese, S. Hoşten, A. Khetan and B. Sturmfels, The maximum likelihood degree, Amer. J. Math. 128:3, 671–697 (2006).
21. Y. Chen, I.H. Dinwoodie and S. Sullivant, Sequential importance sampling for multiway tables, The Annals of Statistics, 34:1, 523-545 (2006).
22. D. Cox, J. Little and D. O'Shea, Ideal, Varieties, and Algorithms, xvi+551. Springer-Verlag, New York (2008), Third Edition.
23. P. Diaconis and N. Eriksson, Markov bases for noncommutative Fourier analysis of ranked data, J. Symbolic Comput. 41:2, 182–195 (2006).
24. P. Diaconis and B. Sturmfels, Algebraic algorithms for sampling from conditional distributions, Annals of Statistics, 26:1, 363–397 (1998).
25. E. Dimitrova, A. Jarrah, R. Laubenbacher, and B. Stigler, A Gröbner fan method for biochemical network modeling, ISSAC Proceedings, 122-126 (2007).
26. I.H. Dinwoodie, The Diaconis-Sturmfels algorithm and rules of succession, Bernoulli 4:3, 401–410 (1998).
27. A. Dobra, Markov Bases for decomposable graphical models, Bernoulli, 9, 1093–1108 (2003).
28. M. Drton, Algebraic techniques for Gaussian models, In Prague Stochastics (M. Huskova, M. Janzura Eds.), 81–90 (2006).
29. M. Drton, Likelihood ratio tests and singularities, Annals of Statistics (to appear).

30. M. Drton and S. Sullivant, Algebraic statistical models, Statistica Sinica, 17:4, 1273–1297 (2007).
31. N. Eriksson, S.E. Fienberg, A. Rinaldo and S. Sullivant, Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. J. Symbolic Comput. 41:2, 222–233 (2006).
32. S.E. Fienberg, P. Hersh, A. Rinaldo and Y. Zhou, Maximum likelihood estimation in latent class models For contingency table data, In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
33. S.E. Fienberg and A. Slavkovic, Preserving the confidentiality of categorical statistical data bases when releasing information for association rules. Data Min. Knowl. Discov. 11:2, 155–180 (2005).
34. R. Fontana, G. Pistone and M.P. Rogantin, Classification of two-level factorial fractions, Journal of Statistical Planning and Inference, 87:1, 149–172 (2000).
35. R. Fontana and M.P. Rogantin, Indicator function and sudoku designs, In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
36. L.D. Garcia, M. Stillman and B. Sturmfels, Algebraic geometry of Bayesian networks, J. Symbolic Comput. 39:3-4, 331–355 (2005).
37. D. Geiger, C. Meek and B. Sturmfels, On the toric algebra of graphical models, The Annals of Statistics 34:3, 1463–1492 (2006).
38. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn (eds.), Geometric and algebraic methods in statistics. Cambridge University Press, Cambridge (forthcoming).
39. H. Hara, A. Takemura and R. Yoshida, A Markov basis for conditional test of common diagonal effect in quasi-independence model for two-way contingency tables, arXiv:0802.2603.
40. C. Kang and J. Tian, Polynomial constraints in causal Bayesian networks, In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), 2007.
41. A. Krampe and S. Kuhnt, Bowker's test for symmetry and modifications within the algebraic framework, Computational Statistics and Data Analysis, 51:9, 4124–4142 (2007).
42. A. Krampe and S. Kuhnt, Model selection for contingency tables with algebraic statistics, In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
43. M. Kreuzer and L. Robbiano, Computational commutative algebra, 1, x+321. Springer-Verlag, Berlin (2000).
44. M. Kreuzer and L. Robbiano, Computational commutative algebra, 2, x+586. Springer-Verlag, Berlin (2008).
45. Z. Lin, L. Xu and Q. Wu, Applications of Gröbner bases to signal and image processing: a survey, Linear Algebra and its Applications, 391, 169–202 (2004).
46. P. Malkin, Computing Markov bases, Gröbner bases, and extreme rays, x+223. Ph.D. thesis, Université Catholique de Louvain (2007).
47. H. Maruri-Aguilar, Algebraic statistics in experimental design, Department of Statistics, University of Warwick (2007).
48. H. Maruri-Aguilar, R. Notari and E. Riccomagno, On the description and identifiability analysis of mixture designs, Statistica Sinica 17:4, 1417–1440 (2007).
49. H. Maruri-Aguilar and E. Riccomagno, A model selection algorithm for mixture experiments including process variables, In Proceedings of Moda8 (J Lopez-Fidalgo, J Rodrguez-Daz, B Torsney eds.), 107-114 (2007).
50. H. Maruri-Aguilar and H.P. Wynn, Generalised design: interpolation and statistical modelling over varieties, In Geometric and Algebraic Methods in Statistics (eds. P. Gibilisco, E. Riccomagno, M.P. Rogantin and H.P. Wynn), Cambridge University Press, Cambridge (forthcoming).
51. T. Mora and L. Robbiano, The Gröbner fan of an ideal, Journal of Symbolic Computation, 6, 183–208 (1988).
52. S. Onn and B. Sturmfels, Cutting corners, Advances in Applied Mathematics, 23:1, 29–48 (1999).
53. L. Pachter and B. Sturmfels (eds.), Algebraic statistics for computational biology, 420. Cambridge University Press, Cambridge (2005).
54. J. Pearl, Causality. Models, reasoning, and inference, xvi+384. Cambridge University Press, Cambridge (2000).
55. G. Pistone, E. Riccomagno and H. P. Wynn, Algebraic Statistics, xvii+160. Chapman & Hall/CRC, Boca Raton (2001).

56. G. Pistone and M.P. Rogantin, Indicator function and complex coding for mixed fractional factorial designs, Journal of Statistical Planning and Inference, 138:3:1, 787–802 (2008).
57. G. Pistone and H.P. Wynn, Generalised confounding with Gröbner bases, Biometrika, 83:3, 653–666 (1996).
58. F. Rapallo, Algebraic Markov bases and MCMC for two-way contingency tables, Scandinavian Journal of Statistics, 30, 385–397 (2003).
59. F. Rapallo, Algebraic exact inference for rater agreement models, Stat. Methods Appl. 14:1, 45–66 (2005).
60. F. Rapallo, Markov bases and structural zeros, J. Symbolic Comput. 41:2, 164–172 (2006).
61. E. Riccomagno and J.Q. Smith, Identifying a cause in models which are not simple Bayesian networks, In Proc. IPMU, 1345–1322 (2004).
62. E. Riccomagno and J.Q. Smith, The geometry of causal probability trees that are algebraically constrained, In Search for Optimality in Design and Statistics: Algebraic and Dynamical System Methods (L Pronzato and A A Zigljavsky eds.), 95-129 (2008).
63. E. Riccomagno and J.Q. Smith, The causal manipulation of chain event graphs, http://arxiv.org/abs/0709.3380.
64. E. Riccomagno and J.Q. Smith, Algebraic causality: Bayes nets and beyond. CRiSM Paper No. 07-3 (2007).
65. H. Scheffé, Experiments with mixtures, J. Roy. Statist. Soc. Ser. B, 20, 344–360 (1958).
66. H. Scheffé, The simplex-centroid design for experiments with mixtures, J. Roy. Statist. Soc. Ser. B, 25, 235–263 (1963).
67. G. Shafer, The art of causal conjecture, 552. MIT Press, Cambridge (1996).
68. B. Sturmfels, Solving systems of polynomial equations, CBMS Reg. Conf. Ser. Math., Washington DC (2002).
69. S. Sullivant, Algebraic geometry of Gaussian Bayesian networks, Advances in Applied Mathematics (to appear).
70. K.Q. Ye, Indicator function and its application in two level factorial designs, The Annals of Statistics, 31:3, 984–994 (2003).