1

Algebraic and geometric methods in statistics

Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin, Henry P. Wynn

1.1 Introduction

It might seem natural that where a statistical model can be defined in algebraic terms it would be useful to use the full power of modern algebra to help with the description of the model and the associated statistical analysis. Until the mid 1990s this had been carried out, but only in some specialised areas. Examples are the use of group theory in experimental design and group invariant testing, and the use of vector space theory and the algebra of quadratic forms in fixed and random effect linear models. The newer area which has been given the name 'algebraic statistics' is concerned with statistical models that can be described, in some way, via polynomials. Of course, polynomials were there from the beginning of the field of statistics in polynomial regression models and in multiplicative models derived from independence models for contingency tables, or to use a more modern terminology, models for categorical data. Indeed these two examples form the bedrock of the new field. (Diaconis and Sturmfels 1998) and (Pistone and Wynn 1996) are basic references.

Innovations have entered from the use of the apparatus of polynomial rings: algebraic varieties, ideals, elimination, quotient operations and so on. See Appendix 1.7 of this chapter for useful definitions. The growth of algebraic statistics has coincided with the rapid developments of fast symbolic algebra packages such as CoCoA, SINGULAR, 4ti2 and Macaulay 2.

If the first theme of this volume, algebraic statistics, relies upon computational commutative algebra, the other one is pinned upon differential geometry. In the 1940s Rao and Jeffreys observed that Fisher information can be seen as a Riemannian metric on a statistical model. In the 1970s Čencov, Csiszár and Efron published papers that established deep results on the involved geometry. Čencov proved that Fisher information is the only distance on the simplex that contracts in the presence of noise (Čencov 1982).

The fundamental result by Čencov and Csiszár shows that with respect to the scalar product induced by Fisher information the relative entropy satisfies a Pythagorean equality (Csiszár 1975). This result was motivated by the need to minimise

Algebraic and Geometric Methods in Statistics, ed. Paolo Gibilisco, Eva Riccomagno, Maria Piera Rogantin and Henry P. Wynn. Published by Cambridge University Press.

relative entropy in fields such as large deviations. The differential geometric counterparts are the notions of divergence and dual connections and these can be used to give a differential geometric interpretation to Csiszár's results.

Differential geometry enters in statistical modelling theory also via the idea of exponential curvature of statistical models due to (Efron 1975). In this 'exponential' geometry, one-dimensional exponential models are straight lines, namely geodesics. Sub-models with good properties for estimation, testing and inference, are characterised by small exponential curvature.

The difficult task the editors have set themselves is to bring together the two strands of algebraic and differential geometry methods into a single volume. At the core of this connection will be the exponential family. We will see that polynomial algebra enters in a natural way in log-linear models for categorical data but also in setting up generalised versions of the exponential family in information geometry. Algebraic statistics and information geometry are likely to meet in the study of invariants of statistical models. For example, on one side polynomial invariants of statistical models for contingency tables have long been known (Fienberg 1980) and in phylogenetic algebraic invariants were used from the very beginning in the Hardy–Weinberg computations (Evans and Speed 1993, for example) and are becoming more and more relevant (Casanellas and Fernández-Sánchez 2007). While on the other side we recall with Shun-Ichi Amari[1] that 'Information geometry emerged from studies on invariant properties of a manifold of probability distributions'. The editors have asked the dedicatee, Giovanni Pistone, to reinforce the connection in a final chapter. The rest of this introduction is devoted to an elementary overview of the two areas, avoiding too much technicality.

1.2 Explicit versus implicit algebraic models

Let us see with simple examples how polynomial algebra may come into statistical models. We will try to take a transparent notation. The technical, short review of algebraic statistics in (Riccomagno 2009) can complement our presentation.

Consider quadratic regression in one variable:

$$Y(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \epsilon(x). \tag{1.1}$$

If we observe (without replication) at four distinct *design points*, $\{x_1, x_2, x_3, x_4\}$ we have the usual matrix form of the regression

$$\eta = \mathrm{E}[Y] = X\theta, \tag{1.2}$$

where the X-matrix takes the form:

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{pmatrix},$$

and Y, θ are the observation, parameter vectors, respectively, and the errors have

[1] Cited from the abstract of the presentation by Prof Amari at the LIX Colloquium 2008, Emerging Trends in Visual Computing, 18th-20th November 2008, Ecole Polytechnique.

zero mean. We can give algebra a large role by saying that the design points are the solution of $g(x) = 0$, where

$$g(x) = (x - x_1)(x - x_2)(x - x_3)(x - x_4). \tag{1.3}$$

In algebraic terms the design is a zero-dimensional variety. We shall return to this representation later.

Now, by eliminating the parameters θ_i from the equations for the mean response: $\{\eta_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2,\ i = 1, \dots 4\}$ we obtain an equation just involving the η_i and the x_i:

$$\begin{aligned} &-(x_2 - x_3)(x_2 - x_4)(x_3 - x_4)\eta_1 + (x_1 - x_3)(x_1 - x_4)(x_3 - x_4)\eta_2 \\ &-(x_1 - x_2)(x_1 - x_4)(x_2 - x_4)\eta_3 + (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)\eta_4 = 0, \end{aligned} \tag{1.4}$$

with the conditions that none of the x_i are equal. We can either use formal algebraic elimination (Cox *et al.* 2008, Chapter 3) to obtain this or simply note that the linear model (1.2) states that the vector η belongs to the column space of X, equivalently it is orthogonal to the orthogonal (kernel, residual) space. In statistical jargon we might say, in this case, that the quadratic model is equivalent to setting the orthogonal cubic contrast equal to zero. We call model (1.2) an *explicit* (statistical) algebraic model and (1.4) an *implicit* (statistical) algebraic model.

Suppose that instead of a linear regression model we have a Generalized Linear Model (GLM) in which the Y_i are assumed to be independent Poisson random variables with means $\{\mu_i\}$, with log link

$$\log \mu_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2, \quad i = 1, \dots, 4.$$

Then, we have

$$\begin{aligned} &-(x_2 - x_3)(x_2 - x_4)(x_3 - x_4)\log \mu_1 + (x_1 - x_3)(x_1 - x_4)(x_3 - x_4)\log \mu_2 \\ &-(x_1 - x_2)(x_1 - x_4)(x_2 - x_4)\log \mu_3 + (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)\log \mu_4 = 0. \end{aligned} \tag{1.5}$$

Example 1.1 Assume that the x_i are integer. In fact, for simplicity let us take our design to be $\{0, 1, 2, 3\}$. Substituting these values in the Poisson case (1.5) and exponentiating we have

$$\mu_1 \mu_3^3 - \mu_2^3 \mu_4 = 0.$$

This is a special variety for the μ_i, a toric variety which defines an implicit model. If we condition on the sum of the 'counts': that is $n = \sum_i Y_i$, then the counts become multinomially distributed with probabilities $p_i = \mu_i / n$ which satisfy $p_1 p_3^3 - p_2^3 p_4 = 0$.

The general form of the Poisson log-linear model is $\eta_i = \log \mu_i = X_i^\top \theta$, where $^\top$ stands for transpose and $X_i^\top$ is the i-th row of the X-matrix. It is an exponential family model with likelihood:

$$\begin{aligned} L(\theta) = \prod_i p(y_i, \mu_i) &= \prod_i \exp(y_i \log \mu_i - \mu_i - \log y_i!) \\ &= \exp\left(\sum_i y_i \sum_j X_{ij} \theta_j - \sum_i \mu_i - \sum_i \log y_i! \right), \end{aligned}$$

where y_i is a realization of Y_i. The sufficient statistics can be read off in the usual way as the coefficients of the parameters θ_j:

$$T_j = \sum_i X_{ij} y_i = X_j^\top Y,$$

and they remain sufficient in the multinomial formulation. The log-likelihood is

$$\sum_j T_j \theta_j - \sum_{i=1}^n \mu_i - \sum_{i=1}^n \log y_i!$$

The interplay between the implicit and explicit model forms of algebraic statistical models has been the subject of considerable development; a seemingly innocuous explicit model may have a complicated implicit form. To some extent this development is easier in the so-called *power product*, or *toric* representation. This is, in fact, very familiar in statistics. The Binomial(n, p) mass distribution function is

$$\binom{n}{y} p^y (1-p)^{n-y}, \ y = 0, \ldots, n.$$

Considered as a function of p this is about the simplest example of a power product representation.

Example 1.2 (Example 1.1 cont.) For our regression in multinomial form the power product model is

$$p_i = \xi_0 \xi_1^{x_i} \xi_2^{x_i^2}, \quad i = 1, \ldots, 4,$$

where $\xi_j = e^{\theta_j}$, $j = 0, \ldots, 2$. This is algebraic if the design points $\{x_i\}$ *are integer*. In general, we can write the power product model in the compact form $p = \xi^X$. Elimination of the p_i, then gives the implicit version of the toric variety.

1.2.1 Design

Let us return to the expression for the design in (1.2). We use a quotient operation to show that the cubic model is naturally associated to the design $\{x_i : i = 1, \ldots, 4\}$. We assume that there is no error so that we have exact interpolation with a cubic model. The quadratic model we chose is also a natural model, being a sub-model of the saturated cubic model. Taking any polynomial interpolator $\tilde{y}(x)$ for data $\{(x_i, y_i),\ i = 1, \ldots, 4\}$, with distinct x_i, we can quotient out with the polynomial

$$g(x) = (x - x_1)(x - x_2)(x - x_3)(x - x_4)$$

and write

$$\tilde{y}(x) = s(x)g(x) + r(x),$$

where the remainder, $r(x)$, is a univariate, at most cubic, polynomial. Since $g(x_i) = 0$, $i = 1, \ldots, 4$, on the design $r(x)$ is also an interpolator, and is the unique cubic interpolator for the data. A major part of algebraic geometry, exploited in

algebraic statistics, extends this quotient operation to higher dimensions. The design $\{x_1, \ldots, x_n\}$ is now multidimensional with each $x_i \in \mathbb{R}^k$, and is expressed as the unique solution of a set of polynomial equations, say

$$g_1(x) = \ldots = g_m(x) = 0 \tag{1.6}$$

and the quotient operation gives

$$\tilde{y}(x) = \sum_{i=1}^{m} s_i(x) g_i(x) + r(x). \tag{1.7}$$

The first term on the right-hand side of (1.7) is a member of the *design ideal*. This is defined as the set of all polynomials which are zero on the design and is indicated as $\langle g_1(x), \ldots, g_m(x) \rangle$. The remainder $r(x)$, which is called the *normal form* of $\tilde{y}(x)$, is unique if the $\{g_j(x)\}$ form a Gröbner basis which, in turn, depends on a given *monomial ordering* (see Section 1.7). The polynomial $r(x)$ is a representative of a class of the quotient ring modulo the design ideal and a basis, as a vector space, of the quotient ring is a set of monomials $\{x^\alpha,\ \alpha \in L\}$ of small degree with respect to the chosen term-ordering as specified in Section 1.7. This basis provides the terms of e.g. regression models. It has the *order ideal* property, familiar from statistics, e.g. the hierarchical property of a linear regression model, that $\alpha \in L$ implies $\beta \in L$ for any $\beta \leq \alpha$ (component-wise). The set of such bases as we vary over all term-orderings is sometimes called the *algebraic fan* of the design. In general it does not give the set of all models which can be fitted to the data, even if we restrict to models which satisfy the order ideal property. However, it is, in a way that can be well defined, the set of models of minimal average degree. See (Pistone and Wynn 1996) for the introduction of Gröbner bases into design, (Pistone *et al.* 2001) for a summary of early work and (Berstein *et al.* 2007) for the work on average degree.

Putting all the elements together we have half a dozen classes of algebraic statistical models which form the basis for the field: (i) linear and log-linear explicit algebraic models, including power product models (ii) implicit algebraic models derived from linear, log-linear or power product models (iii) linear and log-linear models and power product models suggested by special experimental designs.

An explicit algebraic model such as (1.1) can be written down, before one considers the experimental design. Indeed in areas such as the optimal design of experiments one may choose the experimental design using some optimality criterion. But the implicit models described above are design dependent as we see from Equation (1.4). A question arises then: is there a generic way of describing an implicit model which is not design dependent? The answer is to define a polynomial of total degree p as an analytic function all of whose derivatives of higher order than p vanish. But this is an infinite number of conditions.

We shall see that the explicit–implicit duality is also a feature of the information geometry in the sense that one can consider a statistical manifold as an implicit object or defined by some parametric path or surface.

1.3 The uses of algebra

So far we have only shown the presence of algebraic structures in statistical models. We must try to answer briefly the question: what real use is the algebra? We can divide the answer into three parts: (i) to better understand the structure of well-known models, (ii) to help with, or innovate in, statistical methodology and inference and (iii) to define new model classes exploiting particular algebraic structures.

1.3.1 Model structure

Some of the most successful contributions of the algebra are due to the introduction of ideas which the statistical community has avoided or not had the knowledge to pursue. This is especially true for toric models for categorical data. It is important to distinguish two cases. First, for probability models all the representations: log-linear, toric, power product are essentially equivalent in the case that all probabilities are restricted to be *positive*. This condition can be built into the toric analysis via the so-called *saturation*. Consider our running Example 1.2. If ξ is a dummy variable then the condition $p_1p_2p_3p_4v + 1 = 0$ is violated if any of the p_j is zero. Adding this condition to the conditions obtained via the kernel method and eliminating v turns out to be equivalent to directly eliminating the ξ in the power product (toric) representation.

A considerable contribution of the algebraic methods is to handle boundary cases where probabilities are allowed to be zero. Zero counts are very common in sparse tables of data, such as when in a sample survey respondents are asked a large number of questions, but this is not the same as zero probabilities. But we may in fact have special models with zero probabilities in some cells. We may call these models *boundary models* and a contribution of the algebra is to analyse their complex structure. This naturally involves considerable use of algebraic ideas such as *irreducibility, primary decompositions, Krull dimension* and *Hilbert dimension*.

Second, another problem which has bedevilled statistical modelling is that of identifiability. We can take this to mean that different parameter values lead to different distributions. Or we can have a data-driven version: for a given data set (the one we have) the likelihood is locally invertible. The algebra is a real help in understanding and resolving such problems. In the theory of experimental design we can guarantee that the remainder (quotient) models (or sub-models of remainder models), $r(x)$, are identifiable given the design from which they were derived. The algebra also helps to explain the concept of *aliasing*: two polynomial models $p(x)$ and $q(x)$ are aliased over a design $\mathcal{D}$ if $p(x) = q(x)$ for all x in $\mathcal{D}$. This is equivalent to saying that $p(x) - q(x)$ lies in the design ideal.

There is a generic way to study identifiability, that is via elimination. Suppose that $h(\theta)$, for some parameter $\theta \in \mathbb{R}^u$ and $u \in \mathbb{Z}_{>0}$, is some quantity of interest such as a likelihood, distribution function, or some function of those quantities. Suppose also that we are concerned that $h(\theta)$ is over-parametrised in that there is a function of θ, say $\phi(\theta) \in \mathbb{R}^v$ with dimension $v < u$, with which we can parametrise the model

but which has a smaller dimension than θ. If all the functions are polynomial we can write down (in possibly vector form): $r - h(\theta) = 0,\ s - \phi(\theta) = 0$, and try to eliminate θ algebraically to obtain the (smallest) variety on which (r, s) lies. If we are lucky this will give r explicitly in terms as function of s, which is then the required reparametrisation.

As a simple example think of a 2×2 table as giving probabilities p_{ij} for a bivariate binary random vector (X_1, X_2). Consider an over-parametrised power product model for independence with

$$p_{00} = \xi_1\xi_3,\ p_{10} = \xi_2\xi_3,\ p_{01} = \xi_1\xi_4,\ p_{11} = \xi_2\xi_4.$$

We know that independence gives zero covariance so let us seek a parametrisation in terms of the non-central moments $m_{10} = p_{10}+p_{11},\ m_{01} = p_{01}+p_{11}$. Eliminating the ξ_i (after adding $\sum_{ij} p_{ij} - 1 = 0$), we obtain the parametrisation: $p_{00} = (1-m_{10})(1-m_{01}),\ p_{10} = m_{10}(1-m_{01}),\ p_{01} = (1-m_{10})m_{01},\ p_{11} = m_{10}m_{01}$. Alternatively, if we include $m_{11} = p_{11}$, the unrestricted probability model in terms of the moments is given by $p_{00} = 1 - m_{10} - m_{01} + m_{11},\ p_{10} = m_{10} - m_{11},\ p_{01} = m_{01} - m_{11}$, and $p_{11} = m_{11}$, but then we need to impose the extra *implicit* condition for zero covariance: $m_{11} - m_{10}m_{01} = 0$. This is another example of implicit–explicit duality.

Here is a Gaussian example. Let $\delta = (\delta_1, \delta_2, \delta_3)^\top$ be independent Gaussian unit variance input random variables. Define the output Gaussian random variables as

$$\begin{aligned} Y_1 &= \theta_1\delta_1 \\ Y_2 &= \theta_2\delta_1 + \theta_3\delta_2 \\ Y_3 &= \theta_4\delta_1 + \theta_5\delta_3, \end{aligned} \tag{1.8}$$

It is easy to see that this implies the conditional independence of Y_2 and Y_3 given Y_1. The covariance matrix of the $\{Y_i\}$ is

$$C = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix} = \begin{pmatrix} \theta_1^2 & \theta_1\theta_2 & \theta_1\theta_4 \\ \theta_1\theta_2 & \theta_2^2 + \theta_3^2 & \theta_2\theta_4 \\ \theta_1\theta_4 & \theta_2\theta_4 & \theta_4^2 + \theta_5^2 \end{pmatrix}.$$

This is invertible (and positive definite) if and only if $\theta_1\theta_3\theta_5 \neq 0$. If we adjoin the saturation condition $\theta_1\theta_3\theta_5 v - 1 = 0$ and eliminate the θ_j and we obtain the symmetry conditions $c_{12} = c_{21}$ etc. plus the single equation $c_{11}c_{23} - c_{12}c_{13} = 0$. This is equivalent to the (2,3) entry of C^{-1} being zero. The linear representation (1.8) can be derived from a graphical simple model: $2 - 1 - 3$, and points to a strong relationship between graphical models and conditions on covariance structures. The representation is also familiar in time series as the moving average representation. See (Drton *et al.* 2007) for some of the first work on the algebraic method for Gaussian models.

In practical statistics one does not rest with a single model, at least not until after a considerable effort on diagnostics, testing and so on. It is better to think in terms of hierarchies of models. At the bottom of the hierarchy may be simple models. In regression or log-linear models these may typically be additive models. More complex models may involve interactions, which for log-linear models may be representations of conditional independence. One can think of models of higher

polynomial degree in the algebraic sense. The advent of very large data sets has stimulated work on model choice criteria and methods. The statistical kit-bag includes AIC, BIC, CART, BART, Lasso and many other methods. There are also close links to methods in data-mining and machine learning. The hope is that the algebra and algebraic and differential geometry will point to natural model structures be they rings, complexes, lattices, graphs, networks, trees and so on and also to suitable algorithms for climbing around such structures using model choice criteria.

In latent, or hidden, variable methods we extended the model top 'layer' with another layer which endows parameters from the first layer with distributions, that is to say *mixing*. This is also, of course, a main feature of Bayesian models and classical random effect models. Another generic term is hierarchical models, especially when we have many layers. This brings us naturally to *secant varieties* and we can push our climbing analogy one step further. A secant variety is a bridge which walks us from one first-level parameter value to another, that is it provides a support for the mixing. In its simplest form secant variety takes the form

$$\{r : r = (1-\lambda)p + \lambda q,\ 0 \le \lambda \le 1\}$$

where p and q lie in varieties P and G respectively (which may be the same). See (Sturmfels and Sullivant 2006) for a useful study.

In probability models distinction should be made between a zero in a cell in data table, a zero *count*, and a structural zero in the sense that the model assigns zero probability to the cell. This distinction becomes a little cloudy when it is a cell which has a count but which, for whatever reason, could not be observed. One could refer to the latter as censoring which, historically, is when an observation is not observed because it has not happened yet, like the time of death or failure. In some fields it is referred to as having *partial information*.

As an example consider the toric idea for a simple balanced incomplete block design (BIBD). There are two factors, 'blocks' and 'treatments', and the arrangement of treatment in blocks is given by the scheme

$$\binom{1}{2}\binom{1}{3}\binom{1}{4}\binom{2}{3}\binom{2}{4}\binom{3}{4}$$

e.g. $\binom{1}{2}$ is the event that treatment 1 and 2 are in the first block. This corresponds to the following two-factor table where we have inserted the probabilities for observed cells, e.g. p_{11} and p_{21} are the probabilities that treatments one and two are in the first block,

p_{11}	p_{12}	p_{13}			
p_{21}			p_{24}	p_{25}	
	p_{32}		p_{34}		p_{36}
		p_{43}		p_{45}	p_{46}

The additive model $\log p_{ij} = \mu_0 + \alpha_i + \beta_j$ (ignoring the $\sum p_{ij} = 1$ constraint) has nine degrees of freedom (the rank of the X-matrix) and the kernel has rank 3 and one solution yields the terms:

$$\begin{aligned} p_{12}p_{21}p_{34} - p_{11}p_{24}p_{32} &= 0 \\ p_{24}p_{36}p_{45} - p_{25}p_{34}p_{46} &= 0 \\ p_{11}p_{25}p_{43} - p_{13}p_{21}p_{45} &= 0. \end{aligned}$$

A Gröbner basis and a Markov basis can also be found. For work on Markov bases for incomplete tables see (Aoki and Takemura 2008) and (Consonni and Pistone 2007).

1.3.2 Inference

If we condition on the sufficient statistics in a log-linear model for contingency tables, or its power-product form, the conditional distribution of the table does not depend on the parameters. If we take a classical test statistic for independence such as a χ^2 or likelihood ratio (deviance) statistics, then its conditional distribution, given the sufficient statistics T, will also not depend on the parameters, being a function of T. If we are able to find the conditional distribution and perform a conditional test, e.g. for independence, then (Type I) error rates will be the same as for the unconditional test. This follows simply by taking expectations. This technique is called an *exact conditional test*. For (very) small samples we can find the exact conditional distribution using combinatorial methods.

However, for tables which are small but too large for the combinatorics and not large enough for asymptotic methods to be accurate, algebraic Markov chain methods were introduced by (Diaconis and Sturmfels 1998). In the tradition of Markov Chain Monte Carlo (MCMC) methods we can simulate from the true conditional distribution of the tables by running a Markov chain whose steps preserve the appropriate margins. The collection of steps forms a *Markov basis* for the table. For example for a complete $I \times J$ table, under independence, the row and column sums (margins) are sufficient. A table is now a state of the Markov chain and a typical move is represented by a table with all zeros except values 1 at entry (i, i') and (j, j') and entry -1 at entries (j, i') and (i, j'). Adding this to or subtracting this from a current table (state) keeps the margins fixed, although one has to add the condition of non-negativity of the tables and adopt appropriate transition probabilities. In fact, as in MCMC practice, derived chains such as in the Metropolis–Hastings algorithm are used in the simulation.

It is not difficult to see that if we set up the X-matrix for the problem then a move corresponds to a column orthogonal to all the columns of X i.e. the kernel space. If we restrict to all probabilities being positive then the toric variety, the variety arising from a kernel basis and the Markov basis are all the same. In general the kernel basis is smaller than the Markov basis which is smaller than the associated Gröbner basis. In the terminology of ideals:

$$I_K \subset I_M \subset I_G,$$

with reverse inclusion for the varieties, where the sub-indices $_{K,\,M,\,G}$ stands for Kernel, Markov and Gröbner, respectively.

Given that one can carry out a single test, it should be possible to do multiple testing, close in spirit to the model-order choice problem mentioned above. There are several outstanding problems such as (i) finding the Markov basis for large problems and incomplete designs, (ii) decreasing the cost of simulation itself for example by repeat use of simulation, and (iii) alternatives to, or hybrids, simulation, using linear, integer programming, integer lattice theory (see e.g. Chapter 4).

The algebra can give insight into the solutions of the Maximum Likelihood Equations. In the Poisson/multinomial GLM case and when $p(\theta)$ is the vector of probabilities, the likelihood equations are

$$\frac{1}{n}X^\top Y = \frac{1}{n}T = X^\top p(\theta),$$

where $n = \sum_{x_i} Y(x_i)$ and T is the vector of sufficient statistics or generalised margins. We have emphasised the non-linear nature of these equations by showing that p depends on θ. Since $m = X^\top p$ are the moments with respect to the columns of X and $\frac{1}{n}X^\top Y$ are their sample counterpart, the equations simply equate the sample non-central moments to the population non-central moments. For the example in (1.1) the population non-central moments are $m_0 = 1,\ m_1 = \sum_i p_i x_i,\ m_2 = \sum_i p_i x_i^2$. Two types of result have been studied using algebra: (i) conditions for when the solution have closed form, meaning a rational form in the data Y and (ii) methods for counting the number of solutions. It is important to note that unrestricted solutions, $\hat{\theta}$, to these equations are not guaranteed to place the probabilities $p(\hat{\theta})$ in the region $\sum_i p_i = 1,\ p_i > 0, i = 1, \ldots, n$. Neither need they be real. Considerable progress has been made such as showing that decomposable graphical models have a simple form for the toric ideals and closed form of the maximum likelihood estimators: see (Geiger *et al.* 2006). But many problems remain such as in the study of non-decomposable models, models defined via various kinds of marginal independence and marginal conditional independence, and distinguishing real from complex solutions of the maximum likelihood equations.

As is well known, an advantage of the GLM formulation is that quantities which are useful in the asymptotics can be readily obtained, once the maximum likelihood estimators have been obtained. Two key quantities are the score statistic and the Fisher information for the parameters. The score (vector) is

$$U = \frac{\partial l}{\partial \theta} = X^\top Y - X^\top \mu,$$

where $j = (1, \ldots, n)^\top$ and we recall $\mu = \mathrm{E}[Y]$. The (Fisher) information is

$$\mathcal{I} = -\,\mathrm{E}\left[\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right] = X^\top \mathrm{diag}(\mu) X,$$

which does not depend on the data.

As a simple exercise let us take the 2×2 contingency table, with the additive Poisson log-linear model (independence in the multinomial case representation) so that, after reparametrising to $\log \mu_{00} = \theta_0$, $\log \mu_{10} = \theta_0 + \theta_1$, $\log \mu_{01} = \theta_0 + \theta_2$ and