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 which 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 which 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 which 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, e.g.) 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 2008) 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 form of algebraic statistical models have 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

$$\left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}, \; y = 0, \ldots, n.$$

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 $\xi_j = e^{\theta_j}$, $i = 0, \ldots, 3,$. 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 because of 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 $v < u$, with which we can parametrise the

models 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_1 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 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 terms 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

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

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

$$\begin{pmatrix} 1 \\ 2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} \begin{pmatrix} 1 \\ 4 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

e.g. $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$ 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, 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 to, 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 graphic 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

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

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

$\log \mu_{11} = \theta_0 + \theta_1 + \theta_2$, we have the rank 3 X-matrix:

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

In the power product formulation it becomes $\mu_{00} = \xi_0$, $\mu_{10} = \xi_0\xi_1$, $\mu_{01} = \xi_0\xi_2$, and $\mu_{11} = \xi_0\xi_1\xi_2$, and if we algebraically eliminate the ξ_i we obtain the following variety for the entries of $\mathcal{I} = \{\mathcal{I}_{ij}\}$, the information matrix for the θ_j

$$\mathcal{I}_{13} - \mathcal{I}_{33} = 0,\ \mathcal{I}_{12} - \mathcal{I}_{22} = 0,\ \mathcal{I}_{11}\mathcal{I}_{23} - \mathcal{I}_{22}\mathcal{I}_{33} = 0.$$

This implies that the $(2,3)$ entry in $\mathcal{I}^{-1}$, the asymptotic covariance of the maximum likelihood estimation of the parameters, is zero, as expected from the orthogonality of the problem.

1.3.3 Cumulants and moments

A key quantity in the development of the exponential model and associated asymptotics is the cumulant generating function. This is embedded in the Poisson/multi/-nomial development as is perhaps most easily seen by writing the multinomial version in terms of repeated sampling from a given discrete distribution whose support is what we have been calling the 'design'. Let us return to Example 1.1 one more time. We can think of this as arising from a distribution with support $\{0,1,2,3\}$ and probability mass function:

$$p(x;\theta_1,\theta_2) = \exp(\theta_1 x + \theta_2 x^2 - K(\theta_1,\theta_2)),$$

where we have suppressed θ_0 and incorporated it into $K(\theta_1,\theta_2)$. We clearly have

$$K(\theta_1,\theta_2) = \log(1 + e^{\theta_1+\theta_2} + e^{2\theta_1+4\theta_2} + e^{3\theta_1+9\theta_2}).$$

The moment generating function is

$$M_X(s) = \mathrm{E}_X[e^{sX}] = e^{K(\theta_1+s,\theta_2)}e^{-K(\theta_1,\theta_2)},$$

and the cumulant generating function is

$$K_X(s) = \log M_X(s) = K(\theta_1+s,\theta_2) - K(\theta_1,\theta_2).$$

The expression for $K''(s)$ in terms of $K'(s)$ is sometime called the *variance function* in GLM theory and we note that $\mu = K'(0)$ and $\sigma^2 = K''(0)$ give the first two cumulants, which are respectively the mean and variance. If we make the power parametrisation $\xi_1 = e^{\theta_1}$, $\xi_2 = e^{\theta_2}, t = e^s$ and eliminate t from the expressions for K' and K'' (suppressing s), which are now rational, we obtain, after some algebra, the implicit representation

$$\begin{aligned} -8K'^2 + 24K' + (-12 - 12K' + 4K'^2 - 12K'\xi_2^2 + 36K'\xi_2^2)H \\ +(8 - 24\xi_2^2)H^2 + (-9\xi_2^6 - 3\xi_2^4 + 5\xi_2^2 - 1)H^3 \end{aligned}$$

where $H = 3K' - K'^2 - K''$. Only at the value $\xi_2 = 1/\sqrt{3}$ the last term is zero and there is then an explicit quadratic variance function:

$$K'' = \frac{1}{3}K'(3 - K').$$

All discrete models of the log-linear type with integer support/design have an implicit polynomial relationship between K' and K'' where, in the multivariate case these are respectively a $(p-1)$-vector and a $(p-1) \times (p-1)$ matrix, and as in this example, we may obtain a polynomial variance function for special parameter values. Another interesting fact is that because of the finiteness of the support higher order moments can be expressed in terms of lower order moments. For our example we write the design variety $x(x-1)(x-2)(x-3) = 0$ as

$$x^4 = 6x^3 - 11x^2 + 6x$$

multiplying by x^r and taking expectation we have for the moments $m_r = \mathrm{E}[X^r]$ the recurrence relationship

$$m_{4+r} = 6m_{3+r} - 11m_{2+r} + 6m_{r+1}.$$

See (Pistone and Wynn 2006) and (Pistone and Wynn 1999) for work on cumulants.

This analysis generalises to the multivariate case and we have intricate relations between the defining Gröbner basis for the design, recurrence relationships and generating functions for the moments and cumulants, the implicit relationship between K and K' and implicit relation for raw probabilities and moments, arising from the kernel/toric representations. There is much work to be done to unravel all these relationships.

1.4 Information geometry on the simplex

In information geometry a statistical model is a family of probability densities (on the same sample space) and is viewed as a differential manifold. In the last twenty years there has been a development of information geometry in the non-parametric (infinite-dimensional) case and non-commutative (quantum) case. Here we consider the finite-dimensional case of a probability vector $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$. Thus we may take the sample space to be $\Omega = \{1, ..., n\}$ and the manifold to be the interior of the standard simplex:

$$\mathcal{P}_n^1 = \{p : p_i > 0, \ \sum p_i = 1\}$$

(other authors use the notation $\mathcal{M}_>$). Each probability vector $p \in \mathcal{P}_n^1$ is a function from Ω to $\mathbb{R}^n$ and $f(p)$ is well defined for any reasonable real function f, e.g. any bounded function.

The tangent space of the simplex can be represented as

$$T_p(\mathcal{P}_n^1) = \{u \in \mathbb{R}^n : \sum_i u_i = 0\} \tag{1.9}$$

because the simplex is embedded naturally in $\mathbb{R}^n$. The tangent space at a given p can be also identified with the p-centered random variables, namely random variables with zero mean with respect to the density p

$$T_p(\mathcal{P}_n^1) = \{u \in \mathbb{R}^n : \mathrm{E}_p[u] = \sum_i u_i p_i = 0\}. \tag{1.10}$$

With a little abuse of language we use the same symbol for the two different representations (both will be useful in the sequel).

1.4.1 Maximum entropy and minimum relative entropy

Let p and q be elements of the simplex. Entropy and relative (Kullback–Leibler) entropy are defined by the following formulas

$$\mathrm{S}(p) = -\sum_i p_i \log p_i, \tag{1.11}$$

$$\mathrm{K}(p, q) = \sum_i p_i (\log p_i - \log q_i), \tag{1.12}$$

which for $q_0 = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ simplifies to $\mathrm{K}(p, q_0) = \sum_i p_i \log p_i - \sum_i p_i \log \frac{1}{n} = -\mathrm{S}(p) + \log n$.

In many applications, e.g. large deviations and maximum likelihood estimation, it is required to minimise the relative entropy, namely to determine a probability p on a manifold M that minimises $\mathrm{K}(p, q_0)$, equivalently that maximises the entropy $\mathrm{S}(p)$. Here Pythagorean-like theorems can be very useful. But the relative entropy is not the square of a distance between densities. For example, it is asymmetric and the triangle inequality does not hold. In Section 1.4.2 we illustrate some geometries on the simplex to bypass these difficulties.

In (Dukkipati 2008) the constrained maximum entropy and minimum relative entropy optimisation problems are translated in terms of toric ideals, following an idea introduced in (Hoşten *et al.* 2005) for maximum likelihood estimation. The key point is that the solution is an exponential model, hence a toric model, under the assumption of positive integer valued sufficient statistics. This assumption is embedded in the constraints of the optimisation, see e.g. (Cover and Thomas 2006). Ad hoc algorithms are to be developed to make this approach effective.

1.4.2 Paths on the simplex

To understand a geometry on a manifold we need to describe its geodesics in an appropriate context. The following are examples of curves that join the probability vectors p and q in $\mathcal{P}_n^1$:

$$(1 - \lambda)p + \lambda q, \tag{1.13}$$

$$\frac{p^{1-\lambda}\, q^{\lambda}}{C}, \tag{1.14}$$

$$\frac{((1 - \lambda)\sqrt{p} + \lambda\sqrt{q})^2}{B}, \tag{1.15}$$

where $C = \sum_i p_i^{1-\lambda}\, q_i^{\lambda}$ and $B = 2\sum_i [(1-\lambda)\sqrt{p_i} + \lambda\sqrt{q_i}]^2$ are suitable normalisation constants. We may ask which is the most 'natural' curve joining p and q. In the case (1.15) the answer is that the curve is a geodesic with respect to the metric defined by the Fisher information. Indeed, all the three curves above play important roles in the geometric approach to statistics.

1.5 Exponential–mixture duality

We consider the simplex and the localised representation of the tangent space. Define a parallel transport as

$$U^m_{pq}(u) = \frac{p}{q}u$$

for $u \in T_p(\mathcal{P}^1_n)$. This shorthand notation must be taken to mean $\left(\frac{p_1}{q_1}u_1, \ldots, \frac{p_n}{q_n}u_n\right)$. Then $\frac{p}{q}u$ is q-centred and composing the transports $U^m_{pq}U^m_{qr}$ gives $U^m_{p,r}$. The geodesics associated to this parallel transport are the mixture curves in (1.13).

The parallel transport defined as

$$U^e_{pq}(u) = u - \mathrm{E}_q[u]$$

leads to a geometry whose geodesics are the exponential models as in (1.14). In the parametric case this can be considered arising from local representation of the models via their differentiated log-density or *score*.

There is an important and general duality between the mixture and exponential forms. Assume that v is p-centred and define

$$\langle u, v\rangle_p = \mathrm{E}_p[uv] = \mathrm{Cov}_p(u, v).$$

Then we have

$$\langle U^e_{pq}(u), U^m_{pq}(v)\rangle_q = \mathrm{E}_q\left[(u - \mathrm{E}_q[u[]\frac{p}{q}v\right] = \\ \mathrm{E}_p[uv] - \mathrm{E}_q[u]\,\mathrm{E}_p[v] = \mathrm{E}_p[uv] = \langle u, v\rangle_p. \quad (1.16)$$

1.6 Fisher information

Let us develop the exponential model in more detail. The exponential model is given in the general case by

$$p_\theta = \exp(u_\theta - K(u_\theta))p$$

where we have set $p = p_0$ and u_θ is a parametrised class of functions. In the simplex case we can write the one-parameter exponential model as

$$p_{\lambda,i} = \exp(\lambda(\log q_i - \log p_i) - \log(C))p_i.$$

Thus with θ replaced by λ, the ith component of u_θ by $\lambda(\log q_i - \log p_i)$ and $K = \log C$, we have the familiar exponential model. After an elementary calculation the

Fisher information at p in terms of the centred variable $\bar{u} = u - \mathrm{E}_p[u]$ is

$$\mathcal{I}_p = \sum_{i=1}^{n} \bar{u}_i^2 p_i$$

where $\bar{u} \in T_p(\mathcal{P}_n^1)$ as in Equation (1.10). Analogously, the Fisher metric is $\langle u, v \rangle_p = \sum_{i=1}^n \bar{u}_i \bar{v}_i p_i$. In the representation (1.9) of the tangent space the Fisher matrix is

$$\langle \bar{\bar{u}}, \bar{\bar{v}} \rangle_{p,FR} = \sum_i \frac{\bar{\bar{u}}_i \bar{\bar{v}}_i}{p_i}$$

with $\bar{\bar{u}}_i = u_i - \sum_i u_i/n$ where n is the total sample size.

The duality in (1.16) applies to the simplex case and exhibits a relationship endowed with the Fisher information. Let $u = \log \frac{q}{p}$ so that for the exponential model

$$\dot{p}_\lambda = \frac{\partial p_\lambda}{\partial \lambda} = u - \mathrm{E}_\lambda[u].$$

Now the mixture representative of the models is $\frac{p_\lambda}{p} - 1$, whose differential (in the tangent space) is $\frac{u}{p_\lambda} = \frac{p}{q} v$, say. Then putting $\lambda = 1$ the duality in (1.16) becomes

$$\langle \bar{u}, \bar{v} \rangle_p = \langle \bar{\bar{u}}, \bar{\bar{v}} \rangle_{p,FR} = \mathrm{Cov}_p(u, v).$$

Note that the manifold $\mathcal{P}_n^1$ with the Fisher metric is isometric with an open subset of the sphere of radius 2 in $\mathbb{R}^n$. Indeed, if we consider the map $\varphi : \mathcal{P}_n^1 \to S_2^{n-1}$ defined by

$$\varphi(p) = 2(\sqrt{p_1}, ..., \sqrt{p_n})$$

then the differential on the tangent space is given by

$$D_p\varphi(u) = \left(\frac{u_1}{\sqrt{p_1}}, ..., \frac{u_n}{\sqrt{p_n}} \right).$$

(Gibilisco and Isola 2001) shows that the Fisher information metric is the pull-back of the natural metric on the sphere.

This identification allows us to describe geometric objects of the Riemannian manifold, namely $(\mathcal{P}_n^1, \langle \cdot, \cdot \rangle_{p,FR})$, using properties of the sphere S_2^{n-1}. For example, as in (1.15), we obtain that the geodesics for the Fisher metric on the simplex are

$$\frac{\left(\lambda\sqrt{p} + (1-\lambda)\sqrt{q}\right)^2}{B}.$$

As shown above, the geometric approach to Fisher information demonstrates in which sense mixture and exponential models are dual of each other. This can be considered as a fundamental paradigm of information geometry and from this an abstract theory of statistical manifolds has been developed which generalises Riemannian geometry, see (Amari and Nagaoka 2000).

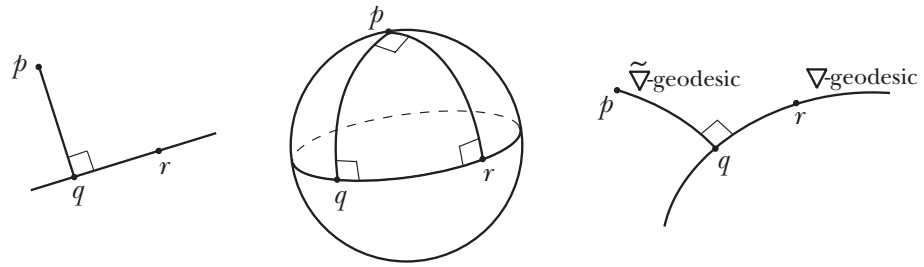

Fig. 1.1 Pythagora theorem: standard (left), geodesic triangle on the sphere (centre) and generalised (right).

1.6.1 The generalised Pythagorean theorem

We formulate the Pythagorean theorem in a form suitable to be generalised to a Riemannian manifold. Let p, q, r be points of the real plane and let $D(p|q)$ be the square of the distance between p and q. If γ is a geodesic connecting p and q, and δ is a geodesic connecting q with r, and furthermore if γ and δ intersect at q orthogonally, then $D(p|q)+D(q|r) = D(p|r)$, see Figure 1.1 (left). Figure 1.1 (centre) shows that on a general Riemannian manifold, like the sphere, $D(p|q) + D(q|r) \neq D(p|r)$, usually. This is due to the curvature of the manifold and a flatness assumption is required. The flatness assumption allows the formulation of the Pythagorean theorem in a context broader than the Riemannian one.

A *divergence* on a differential manifold M is a non-negative smooth function $D(\cdot|\cdot)$: $M \times M \to \mathbb{R}$ such that $D(p|q) = 0$ if, and only if, $p = q$ (note that here D stands for divergence and not derivative). A typical example is the Kullback-Leibler divergence, which we already observed is not symmetric hence it is not a distance.

It is a fundamental result of Information Geometry, see (Eguchi 1983, Eguchi 1992, Amari and Nagaoka 2000), that to any divergence D one may associate three geometries, namely a triple $\left(\langle\cdot,\cdot\rangle^D, \nabla^D, \tilde{\nabla}^D\right)$ where $\langle\cdot,\cdot\rangle^D$ is a Riemannian metric while $\nabla^D, \tilde{\nabla}^D$ are two linear connections in duality with respect to the Riemannian metric.

A statistical structure $\left(\langle\cdot,\cdot\rangle^D, \nabla^D, \tilde{\nabla}^D\right)$ is *dually flat* if both ∇ and $\tilde{\nabla}$ are flat. This means that curvature and torsion are (locally) zero for both connections. This is equivalent to the existence of an affine coordinate system. The triple given by the Fisher information metric, the mixture–exponential connection pair, whose geodesics are given in Equations (1.13) and (1.14), is an example of a dually flat statistical structure. The generalised Pythagorean theorem can be stated as follows.

Let $D(\cdot|\cdot)$ be a divergence on M such that the induced statistical structure is dually flat. Let p, q, $r \in M$, let γ be a ∇^D-geodesic connecting p and q, let δ be a $\tilde{\nabla}^D$-geodesic connecting q with r, and suppose that γ and δ intersect at q orthogonally with respect to the Riemannian metric $\langle\cdot,\cdot\rangle^D$. Then, as shown in Figure 1.1 (right),

$$D(p|q) + D(q|r) = D(p|r).$$

Summarising, if the divergence is the squared Euclidean distance, this is the usual Pythagorean theorem and if the divergence is the Kullback–Leibler relative entropy, this is the differential geometric version of the result proved in (Csiszár 1975), see also (Grünwald and Dawid 2004). In a quantum setting, (Petz 1998) proved a Pythagorean-like theorem with the Umegaki relative entropy instead of Kullback–Leibler relative entropy. Here as well the flatness assumption is essential.

1.6.2 General finite-dimensional models

In the above we really only considered the one-parameter exponential model, even in the finite-dimensional case. But as it is clear from the early part of this introduction more complex exponential models of the form

$$p_\theta = \exp\left(\sum \theta_i u_i - K(\theta)\right) p$$

are studied. Here the u_i are columns of the X-matrix, and we can easily compute the cumulant generating functions, as explained for the running example. More such examples are given in Chapter 21. A log-linear model becomes a flat manifold in the information geometry terminology. There remain problems, even in this case, for example when we wish to compute quantities of interest such as $K(\theta)$ at a maximum likelihood estimator and this does not have a closed form, there will be no closed form for K either.

More serious is when we depart from the log-linear formulation. To repeat: this is when u_θ is not linear. We may use the term *curved* exponential model (Efron 1975). As we have seen, the dual (kernel) space to the model is computable in the linear case and, with the help of algebra, we can obtain implicit representation of the model. But in the non-linear finite-dimensional case there will be often severe computational problems. Understanding the curvature and construction of geodesics may help both with the statistical analysis and also the computation e.g. those relying on gradients. The infinite-dimensional case requires special care as some obvious properties of submanifolds and, hence, tangent spaces could be missing. Concrete and useful examples of infinite-dimensional models do exists e.g. in the framework of Wiener spaces, see Chapter 21.

One way to think of a finite-dimensional mixture model is that it provides a special curved, but still finite-dimensional, exponential family, but with some attractive duality properties. As mentioned, mixture models are the basis of latent variable models (Pachter and Sturmfels 2005) and is to be hoped that the methods of secant varieties will be useful. See Chapter 2 and the on-line Chapter 22 by Yi Zhou. See also Chapter 4 in (Drton *et al.* 2009) for an algebraic exposition on the role of secant varieties for hidden variable models.

1.7 Appendix: a summary of commutative algebra (with Roberto Notari)

We briefly recall the basic results from commutative algebra we need to develop the subject. Without any further reference, we mention that the sources for the material in the present section are (Atiyah and Macdonald 1969) and (Eisenbud 2004).

Let $\mathcal{K}$ be a ground field, and let $R = \mathcal{K}[x_1, \ldots, x_k]$ be the polynomial ring over $\mathcal{K}$ in the indeterminates (or variables) $x_1, \ldots, x_k$. The ring operations in R are the usual sum and product of polynomials.

Definition 1.1 A subset $I \subset R$ is an *ideal* if $f + g \in I$ for all $f, g \in I$ and $fg \in I$ for all $f \in I$ and all $g \in R$.

Polynomial ideals

Proposition 1.1 *Let $f_1, \ldots, f_r \in R$. The set $\langle f_1, \ldots, f_r \rangle = \{f_1 g_1 + \cdots + f_r g_r : g_1, \ldots, g_r \in R\}$ is the smallest ideal in R with respect to the inclusion that contains $f_1, \ldots, f_r$.*

The ideal $\langle f_1, \ldots, f_r \rangle$ is called the *ideal generated by* $f_1, \ldots, f_r$. A central result in the theory of ideals in polynomial ring is the following Hilbert's basis theorem.

Theorem 1.1 *Given an ideal $I \subset R$, there exist $f_1, \ldots, f_r \in I$ such that $I = \langle f_1, \ldots, f_r \rangle$.*

The Hilbert's basis theorem states that R is a Noetherian ring, where a ring is Noetherian if every ideal is finitely generated.

As in the theory of $\mathcal{K}$-vector spaces, the intersection of ideals is an ideal, while the union is not an ideal, in general. However, the following proposition holds.

Proposition 1.2 *Let $I, J \subset R$ be ideals. Then,*

$$I + J = \{f + g : f \in I, g \in J\}$$

is the smallest ideal in R with respect to inclusion that contains both I and J, and it is called the sum of I and J.

Quotient rings

Definition 1.2 Let $I \subset R$ be an ideal. We write $f \sim_I g$ if $f - g \in I$ for $f, g \in R$.

Proposition 1.3 *The relation $\sim_I$ is an equivalence relation in R. Moreover, if $f_1 \sim_I f_2, g_1 \sim_I g_2$ then $f_1 + g_1 \sim_I f_2 + g_2$ and $f_1 g_1 \sim_I f_2 g_2$.*

Definition 1.3 The set of equivalence classes, the cosets, of elements of R with respect to $\sim_I$ is denoted as R/I and called the *quotient space* (*modulo* I).

Proposition 1.3 shows that R/I is a ring with respect to the sum and product it inherits from R. Explicitly, if $[f], [g] \in R/I$ then $[f] + [g] = [f + g]$ and $[f][g] = [fg]$. Moreover, the ideals of R/I are in one-to-one correspondence with the ideals of R containing I.

Definition 1.4 If J is ideal in R, then I/J is the ideal of R/J given by $I \supseteq J$ where I is ideal in R.

Ring morphisms

Definition 1.5 Let R, S be two commutative rings with identity. A map $\varphi : R \to S$ is a *morphism of rings* if (i) $\varphi(f + g) = \varphi(f) + \varphi(g)$ for every $f, g \in R$; (ii) $\varphi(fg) = \varphi(f)\varphi(g)$ for every $f, g \in R$; (iii) $\varphi(1_R) = 1_S$ where $1_R, 1_S$ are the identities of R and S, respectively.

Theorem 1.2 *Let $I \subset R$ be an ideal. Then, the map $\varphi : R \to R/I$ defined as $\varphi(f) = [f]$ is a surjective (or onto) morphism of commutative rings with identity.*

An isomorphism of rings is a morphism that is both injective and surjective.

Theorem 1.3 *Let I, J be ideals in R. Then, $(I + J)/I$ is isomorphic to $J/(I \cap J)$.*

Direct sum of rings

Definition 1.6 Let R, S be commutative rings with identity. Then the set

$$R \oplus S = \{(r, s) : r \in R, s \in S\}$$

with component-wise sum and product is a commutative ring with $(1_R, 1_S)$ as identity.

Theorem 1.4 *Let I, J be ideals in R such that $I + J = R$. Let*

$$\phi : R \to R/I \oplus R/J$$

be defined as $\phi(f) = ([f]_I, [f]_J)$. It is an onto morphism, whose kernel is $I \cap J$. Hence, $R/(I \cap J)$ is isomorphic to $R/I \oplus R/J$.

Localisation of a ring

Let $f \in R, f \neq 0$, and let $S = \{f^n : n \in \mathbb{N}\}$. In $R \times S$ consider the equivalence relation $(g, f^m) \sim (h, f^n)$ if $gf^n = hf^m$. Denote with $\frac{g}{f^n}$ the cosets of $R \times S$, and R_f the quotient set.

Definition 1.7 The set R_f is called the localisation of R with respect to f.

With the usual sum and product of ratios, R_f is a commutative ring with identity.

Proposition 1.4 *The map $\varphi : R \to R_f$ defined as $\varphi(g) = \frac{g}{1}$ is an injective morphism of commutative rings with identity.*

Maximal ideals and prime ideals

Definition 1.8 An ideal $I \subset R$, $I \neq R$, is a *maximal ideal* if I is not properly included in any ideal J with $J \neq R$.

Of course, if $a_1, \ldots, a_k \in \mathcal{K}$ then the ideal $I = \langle x_1 - a_1, \ldots, x_k - a_k \rangle$ is a maximal ideal. The converse of this remark is called Weak Hilbert's Nullstellensatz, and it needs a non-trivial hypothesis.

Theorem 1.5 *Let $\mathcal{K}$ be an algebraically closed field. Then, I is a maximal ideal if, and only if, there exist $a_1, \ldots, a_k \in \mathcal{K}$ such that $I = \langle x_1 - a_1, \ldots, x_k - a_k \rangle$.*

Definition 1.9 An ideal $I \subset R, I \neq R$, is a prime ideal if $xy \in I, x \notin I$ implies that $y \in I$, where $x, y \in \{x_1, \ldots, x_k\}$.

Proposition 1.5 *Every maximal ideal is a prime ideal.*

Radical ideals and primary ideals

Definition 1.10 Let $I \subset R$ be an ideal. Then,

$$\sqrt{I} = \{f \in R : f^n \in I, \text{ for some } n \in \mathbb{N}\}$$

is the *radical* ideal in I.

Of course, I is a radical ideal if $\sqrt{I} = I$.

Definition 1.11 Let $I \subset R, I \neq R$, be an ideal. Then I is a *primary ideal* if $xy \in I, x \notin I$ implies that $y^n \in I$ for some integer n, with $x, y \in \{x_1, \ldots, x_k\}$.

Proposition 1.6 *Let I be a primary ideal. Then, $\sqrt{I}$ is a prime ideal.*

Often, the primary ideal I is called $\sqrt{I}$-primary.

Primary decomposition of an ideal

Theorem 1.6 *Let $I \subset R, I \neq R$, be an ideal. Then, there exist $I_1, \ldots, I_t$ primary ideals with different radical ideals such that $I = I_1 \cap \cdots \cap I_t$.*

Theorem 1.6 provides the so-called primary decomposition of I.

Corollary 1.1 *If I is a radical ideal, then it is the intersection of prime ideals.*

Proposition 1.7 links morphisms and primary decomposition, in a special case that is of interest in algebraic statistics.

Proposition 1.7 *Let $I = I_1 \cap \cdots \cap I_t$ be a primary decomposition of I, and assume that $I_i + I_j = R$ for every $i \neq j$. Then the natural morphism*

$$\varphi : R/I \to R/I_1 \oplus \cdots \oplus R/I_t$$

is an isomorphism.

Hilbert function and Hilbert polynomial

The Hilbert function is a numerical function that 'gives a size' to the quotient ring R/I.

Definition 1.12 Let $I \subset R$ be an ideal. The *Hilbert function* of R/I is the function

$$h_{R/I} : \mathbb{Z} \to \mathbb{Z}$$

defined as $h_{R/I}(j) = \dim_{\mathcal{K}}(R/I)_{\leq j}$, where $(R/I)_{\leq j}$ is the subset of cosets that contain a polynomial of degree less than or equal to j, and $\dim_{\mathcal{K}}$ is the dimension as $\mathcal{K}$-vector space.

The following (in)equalities follow directly from Definition 1.12.

Proposition 1.8 *For every ideal $I \subset R, I \neq R$, it holds: (i) $h_{R/I}(j) = 0$ for every $j < 0$; (ii) $h_{R/I}(0) = 1$; (iii) $h_{R/I}(j) \leq h_{R/I}(j+1)$.*

Theorem 1.7 *There exists a polynomial $p_{R/I}(t) \in \mathbb{Q}[t]$ such that $p_{R/I}(j) = h_{R/I}(j)$ for j much larger than zero, $j \in \mathbb{Z}$.*

Definition 1.13 (i) The polynomial $p_{R/I}$ is called the *Hilbert polynomial* of R/I. (ii) Let $I \subset R$ be an ideal. The *dimension of* R/I is the degree of the Hilbert polynomial $p_{R/I}$ of R/I.

If the ring R/I has dimension 0 then the Hilbert polynomial of R/I is a non-negative constant called the degree of the ring R/I and indicated as $\deg(R/I)$. The meaning of the degree is that $\deg(R/I) = \dim_{\mathcal{K}}(R/I)_{\leq j}$ for j large enough. Moreover, the following proposition holds

Proposition 1.9 *Let $I \subset R$ be an ideal. The following are equivalent: (i) R/I is $0-$dimensional; (ii) $\dim_{\mathcal{K}}(R/I)$ is finite. Moreover, in this case, $\deg(R/I) = \dim_{\mathcal{K}}(R/I)$.*

Term-orderings and Gröbner bases

Next, we describe some tools that make effective computations with ideals in polynomial rings.

Definition 1.14 A *term* in R is $x^a = x_1^{a_1} \dots x_k^{a_k}$ for $a = (a_1, \dots, a_k) \in (\mathbb{Z}_{\geq 0})^k$. The set of terms is indicated as $\mathbb{T}^k$.

The operation in $\mathbb{T}^k$, of interest, is the product of terms.

Definition 1.15 A term-ordering is a *well ordering* $\preccurlyeq$ on $\mathbb{T}^k$ such that $1 \preccurlyeq x^a$ for every $x^a \in \mathbb{T}^k$ and $x^a \preccurlyeq x^b$ implies $x^a x^c \preccurlyeq x^b x^c$ for every $x^c \in \mathbb{T}^k$.

A polynomial in R is a linear combination of a finite set of terms in $\mathbb{T}^k$: $f = \sum_{a \in A} c_a x^a$ where A is a finite subset of $\mathbb{Z}^k_{\geq 0}$.

Definition 1.16 Let $f \in R$ be a polynomial, A the finite set formed by the terms in f and $x^b = \max\{x^a : a \in A\}$. Let $I \subset R$ be an ideal.

(i) The term $\mathrm{LT}(f) = c_b x^b$ is called the *leading term* of f.
(ii) The ideal generated by $\mathrm{LT}(f)$ for every $f \in I$ is called the *order ideal* of I and is indicated as LT(I).

Definition 1.17 Let $I \subset R$ be an ideal and let $f_1, \ldots, f_t \in I$. The set $\{f_1, \ldots, f_t\}$ is a *Gröbner basis* of I with respect to $\preccurlyeq$ if $\mathrm{LT}(I) = \langle \mathrm{LT}(f_1), \ldots, \mathrm{LT}(f_t) \rangle$.

Gröbner bases are special sets of generators for ideals in R. Among the many results concerning Gröbner bases, we list a few, to stress their role in the theory of ideals in polynomial rings.

Proposition 1.10 *Let $I \subseteq R$ be an ideal. Then, $I = R$ if, and only if, $1 \in \mathcal{F}$, where $\mathcal{F}$ is a Gröbner basis of I, with respect to any term-ordering $\preccurlyeq$.*

Proposition 1.11 *Let $I \subset R$ be an ideal. The ring R/I is 0–dimensional if, and only if, $x_i^{a_i} \in \mathrm{LT}(I)$ for every $i = 1, \ldots, k$.*

Proposition 1.11, known as Buchberger's criterion for 0–dimensionality of quotient rings, states that for every $i = 1, \ldots k$, there exists $f_{j(i)} \in \mathcal{F}$, Gröbner basis of I, such that $\mathrm{LT}(f_{j(i)}) = x_i^{a_i}$..

Definition 1.18 Let $I \subset R$ be an ideal. A polynomial $f = \sum_{a \in A} c_a x^a$ is in *normal form* with respect to $\preccurlyeq$ and I if $x^a \notin \mathrm{LT}(I)$ for each $a \in A$.

Proposition 1.12 *Let $I \subset R$ be an ideal. For every $f \in R$ there exists a unique polynomial, indicated as $\mathrm{NF}(f) \in R$, in normal form with respect to $\preccurlyeq$ and I such that $f - \mathrm{NF}(f) \in I$. Moreover, $\mathrm{NF}(f)$ can be computed from f and a Gröbner basis of I with respect to $\preccurlyeq$.*

Gröbner bases allow us to compute in the quotient ring R/I, with respect to a term-ordering, because they provide canonical forms for the cosets. This computation is implemented in many software for symbolic computation.

As last result, we recall that Gröbner bases simplify the computation of Hilbert functions.

Proposition 1.13 *Let $I \subset R$ be an ideal. Then R/I and $R/\mathrm{LT}(I)$ have the same Hilbert function. Furthermore, a basis of the $\mathcal{K}$–vector space $(R/\mathrm{LT}(I))_{\leq j}$ is given by the cosets of the terms of degree $\leq j$ not in $\mathrm{LT}(I)$.*

References

4ti2 Team (2006). *4ti2 – A software package for algebraic, geometric and combinatorial problems on linear spaces* (available at `www.4ti2.de`).

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*, (American Mathematical Society/Oxford University Press).

Aoki, S. and Takemura, A. (2008). The largest group of invariance for Markov bases and toric ideals, *Journal of Symbolic Computing* **43**(5), 342–58.

Atiyah, M. F. and Macdonald, I. G. (1969). *Introduction to Commutative Algebra*, (Addison-Wesley Publishing Company).

Berstein, Y., Maruri-Aguilar, H., Onn, S., Riccomagno, E. and Wynn, H. P. (2007). Minimal average degree aberration and the state polytope for experimental design (available at arXiv:stat.me/0808.3055).

Casanellas, M. and Fernández-Sánchez, J. (2007). Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees, *Molecular Biology and Evolution* **24**(1), 288–93.

Čencov, N. N. (1982). *Statistical decision rules and optimal inference* (Providence, RI, American Mathematical Society). Translation from the Russian edited by Lev J. Leifman.

Consonni, G. and Pistone, G. (2007). Algebraic Bayesian analysis of contingency tables with possibly zero-probability cells, *Statistica Sinica* **17**(4), 1355–70.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory* 2nd edn (Hoboken, NJ, John Wiley & Sons).

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems, *Annals of Probability* **3**, 146–58.

Cox, D., Little, J. and O'Shea, D. (2008). *Ideals, Varieties, and Algorithms* 3rd edn (New York, Springer-Verlag).

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics* **26**(1), 363–97.

Drton, M., Sturmfels, B. and Sullivant, S. (2007). Algebraic factor analysis: tetrads pentads and beyond. *Probability Theory and Related Fields* **138**, 463–93.

Drton, M., Sturmfels, B. and Sullivant, S. (2009). *Lectures on Algebraic Statistics* (Vol. 40, Oberwolfach Seminars, Basel, Birkhäuser).

Dukkipati, A. (2008). Towards algebraic methods for maximum entropy estimation (available at arXiv:0804.1083v1).

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second–order efficiency) (with discussion), *Annals of Statistics* **3**, 1189–242.

Eisenbud, D. (2004). *Commutative Algebra*, GTM 150, (New York, Springer-Verlag).

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family, *Annals of Statistics* **11**, 793–803.

Eguchi, S. (1992). Geometry of minimum contrast, *Hiroshima Mathematical Journal* **22**(3), 631–47.

Evans, S. N. and Speed, T. P. (1993). Invariants of some probability models used in phylogenetic inference, *Annals of Statistics* **21**(1), 355–77.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data* 2nd edn (Cambridge, MA, MIT Press).

Grayson, D. and Stillman, M. (2006). *Macaulay 2, a software system for research in algebraic geometry* (available at `www.math.uiuc.edu/Macaulay2/`).

Geiger, D., Meek, C. and Sturmfels, B. (2006). On the toric algebra of graphical models, *Annals of Statistics* **34**, 1463–92.

Gibilisco, P. and Isola, T. (2001). A characterisation of Wigner-Yanase skew information among statistically monotone metrics, *Infinite Dimensional Analysis Quantum Probability and Related Topics* **4**(4), 553–7.

Greuel, G.-M., Pfister, G. and Schönemann, H. (2005). SINGULAR *3.0. A Computer Algebra System for Polynomial Computations. Centre for Computer Algebra* (available at `www.singular.uni-kl.de`).

Grünwald, P. D. and Dawid, P. (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory, *Annals of Statistics* **32**(4), 1367–433.

Hoşten, S., Khetan, A. and Sturmfels, B. (2005). Solving the likelihood equations, *Foundations of Computational Mathematics* **5**(4), 389–407.

Pachter, L. and Sturmfels, B. eds. (2005). *Algebraic Statistics for Computational Biology* (New York, Cambridge University Press).

Petz, D. (1998). Information geometry of quantum states. In *Quantum Probability Communications*, vol. X, Hudson, R. L. and Lindsay, J. M. eds. (Singapore, World Scientific) 135–58.

Pistone, G., Riccomagno, E. and Wynn, H. P. (2001). *Algebraic Statistics* (Boca Raton, Chapman & Hall/CRC).

Pistone, G. and Wynn, H. P. (1996). Generalised confounding with Gröbner bases, *Biometrika* **83**(3), 653–66.

Pistone, G., and Wynn, H. P. (1999). Finitely generated cumulants, *Statistica Sinica* **9**(4), 1029–52.

Pistone, G., and Wynn, H. P. (2006). Cumulant varieties, *Journal of Symbolic Computing* **41**, 210–21.

Riccomagno, E. (2008). A short history of Algebraic Statisitcs, *Metrika* (in press).

Sturmfels, B. and Sullivant, S. (2006). Combinatorial secant varieties, *Pure and Appl Mathematics Quarterly* **3**, 867–91.