Statistics and Computing (2001) **11,** 37–46

Gröbner bases and factorisation in discrete probability and Bayes

G. PISTONE*, E. RICCOMAGNO† and HENRY P. WYNN†

**Department of Mathematics, Polytechnic in Turin, Corso Duce Degli Abruzzi, 10129 Turin, Italy*
pistone@calvino.polito.it
†*Department of Statistics, University of Warwick, CV4 7AL Coventry, United Kingdom*
Riccomagno@eurandm.tue.nl, H.P.Wynn@warwick.ac.uk

Received August 1998 and accepted August 1999

Gröbner bases, elimination theory and factorization may be used to perform calculations in elementary discrete probability and more complex areas such as Bayesian networks (influence diagrams). The paper covers the application of computational algebraic geometry to probability theory. The application to the Boolean algebra of events is straightforward (and essentially known). The extension into the probability superstructure is via the polynomial interpolation of densities and log densities and this is used naturally in the Bayesian application.

Keywords: Gröbner bases, probability distributions, Bayes

1. Introduction

There have been two recent demonstrations of the usefulness of computational algebraic geometry in statistics and probability, the work of Diaconis and Sturmfels on contingency tables (see Diaconis and Sturmfels 1998) and the introduction of Gröbner basis techniques into the design of experiments, by Pistone and Wynn (1996). In this paper we introduce the application to elementary probability and statistics.

The development of probability has two basic components, the algebra of events and the superstructure of probability and random variables. The algebra of events is, of course, the Boolean algebra of $\cap$ and $\cup$, or $\wedge$ and $\vee$ in logic, translated into the algebra of indicator functions of events. The situation of probability is harder if we want to incorporate it *into the same algebraic environment* as the algebra of events. Following a literature review, it surprised the authors that George Boole himself was concerned with the use of elementary algebra in probability, trying to automatise calculations in probability in the same way as he successfully had for logic (see Hailperin 1976). Of course in areas such as quantum probability, algebraic structures are used via projections and, for example, the idempotency of projections $P^2 = P$ uses the analogue of the idempotency of a simple indicator function $X^2 = X$ in elementary probability. For an application of Gröbner bases to logic see Chazarain *et al.* (1991).

We study here simple discrete probability, random variables and applications to elementary discrete statistics. We first present two basic approaches which are essentially duals of each other. The most obvious method is to give every elementary event its own indicator function so that probabilities can be manipulated in an unambiguous way because of the disjointness of elementary events. The alternative is to *interpolate* discrete probability functions or their logarithms with polynomial interpolators. The interpolation method is powerful in its ability to cover the territory of exponential families and Bayesian networks (influence diagrams) and the conditional probability structures.

The algebraic calculations in this paper have been carried out on Maple.

2. Boolean statements

Consider a set Ω and a ring of subsets (events) $\{A_i\}_{i=1}^n$ consisting of all A_i, all complements $A_i^c = \Omega \backslash A_i$ and all unions and intersections. This gives 2^n elementary events of the form

$$E_J = \bigcap_{i \in J} A_i \bigcap \bigcap_{i \in J^c} A_i^c$$

where $J \subseteq \{1, \ldots, n\}$ is a subset of indices and $J^c = \{1, \ldots, n\} \backslash J$ (J can be the empty set of indices).

0960-3174

For an outcome $\omega \in \Omega$ we define the indicator function

$$I_{A_i} = \begin{cases} 1 & \text{if } \omega \in A_i \\ 0 & \text{otherwise} \end{cases}$$

Note immediately that the I_{A_i}'s form a Boolean algebra that is in one-to-one correspondence with the quotient of the ring $\mathbb{R}[x_1, \ldots, x_n]$ with respect to the relationships $x_i(x_i - 1) = 0, i = 1, \ldots, n$ where x_i represents I_{A_i}.

We should note that other representations have been used in the literature. Thus for example

$$I_{A_i} = \begin{cases} 1 & \text{if } \omega \in A_i \\ -1 & \text{otherwise} \end{cases}$$

and the alternative

$$I_{A_i} = \begin{cases} -1 & \text{if } \omega \in A_i \\ 0 & \text{otherwise} \end{cases}$$

each of which has particular algebraic advantages. For example the first turns the system into an Abelian group with all non-unit elements of order 2: $x_i^2 = 1$. The second (sometimes called the Stone representation) has the advantage of removing minus signs in many computations for example $A \cap B$ translates to $I_A + I_B + I_A I_B$. However, we shall stick to the indicator representation more familiar in probability and statistics.

Note that with our representation

$$\begin{aligned} I_{A\cap B} &= I_A I_B \\ I_{A\cup B} &= I_A + I_B - I_{A\cap B} \\ I_{A^c} &= 1 - I_A \end{aligned}$$

and

$$A \subseteq B \to I_A(I_B - 1) = 0$$

These representations and their generalisations allow all statements about events to be converted to Boolean algebra.

3. Gröbner bases

There are various ways to interpolate polynomials on a set of points in an n dimensional space. For example the three points $(1, 1, 0), (0, 0, 0), (1, 0, 1)$ in $\mathbb{R}^3$ are interpolated by either of the following sets of curves and surfaces

$$\begin{cases} x - y - z = 0 \\ y^2 - y = 0 \\ z^2 - z = 0 \\ yz = 0 \end{cases} \quad \text{or} \quad \begin{cases} z + y - x = 0 \\ y^2 - y = 0 \\ yx - y = 0 \\ x^2 - x = 0. \end{cases} \tag{1}$$

Gröbner basis theory deals with this issue. A basic reference is Cox, Little and O'Shea (1996). The above points to the fact that there are various equivalent ways to represent a given set of polynomials according to different orderings on the monomials. Let $P = \{p_1, \ldots, p_s\}$ be a set of polynomials in the indeterminates $x = (x_1, \ldots, x_n)$. Let $k[x]$ be the set of all polynomials in the indeterminates x with coefficients in the field k. A monomial in $k[x]$ is a polynomial of the form $x_1^{\alpha_1}, \ldots, x_n^{\alpha_n}$ with $\alpha = (\alpha_1, \ldots, \alpha_n)$, a vector with non negative integer components. A term ordering, τ, is a total ordering relation among monomials such that $1 < x^\alpha$ for all $\alpha \in \mathbb{Z}_+^n$ and for all α, β, γ such that $x^\alpha < x^\beta$ we have $x^\alpha x^\gamma < x^\beta x^\gamma$, that is compatibility with the division of monomials. A monomial ordering allows us to define the leading (monomial) term for any polynomial. A polynomial ideal I is a subset of $k[x]$ such that for all $f, g \in I$ and for all $m, n \in k[x]$ also $mf + ng \in k[x]$. A monomial ideal is a polynomial ideal that admits a basis of monomials. The set of polynomials $G = \{g_1, \ldots, g_k\}$ is a Gröbner basis for a polynomial ideal I with respect to τ if the monomial ideal generated by the leading terms of G, with respect to τ, is equal to the monomial ideal formed by all the leading terms of the elements of the ideal generated by I.

The most interesting property of Gröbner bases, for us, is that given a term ordering and a corresponding G-basis $\{g_1, \ldots, g_k\}$, for all polynomials $f \in k[x]$ there exists a unique polynomial r whose leading term is not divisible by the leading terms of the g_i's such that

$$f(x) = \sum_{i=1}^{k} s_i(x) g_i(x) + r(x)$$

Let us consider a full factorial design in n factors and with levels $0, 1, \ldots, N_i - 1$ in the ith-factor, namely the product set

$$D = \prod_{i=1}^{n} \{x_{i,0}, \ldots, x_{i,N_i-1}\}$$

There is a unique interpolator for any set of observations attached (as "y-values") to the design D. In fact the G-basis for the design D above with respect to any term-ordering is given by

$$\left\{ \prod_{j=0}^{N_i-1} (x_i - x_{ij}), \quad i = 1, \ldots, n \right\}$$

and the interpolator is always of the form

$$r(x) = \sum_{0 \le \alpha \le \bar{\alpha}} \theta_\alpha x^\alpha$$

and $0 \le \alpha \le \bar{\alpha}$, where $\bar{\alpha} = (N_1 - 1, \ldots, N_n - 1)$, means $0 \le \alpha_i \le N_i - 1$ for $i = 1, \ldots, n$. The polynomial r is simply the full factorial model over a full factorial design, in statistical terminology.

For a general design, it is not true in general that the interpolator is unique. This is discussed in detail in Caboara *et al.* (1997). In general the monomial terms in r are $\{x^\alpha, \alpha \in L\}$ consisting of all monomials not divisible by the leading terms of the G-basis elements. Moreover the number of such terms is precisely the same as the number of design points. Note that L is an order ideal: $\alpha \in L \Rightarrow \beta \in L$ for all $0 \le \beta \le \alpha$.

The Gröbner basis method can be used to construct an interpolator for values $y_j = y(x^{(j)})$ for each $x^{(j)} = (x_{j,0}, \ldots, x_{j,N_j-1})$ in a general design, D. This is done by constructing a Gröbner basis for the combined ideal given by

$$\begin{cases} x - x^{(j)} \\ y - y_j \end{cases}$$

for the points $(x^{(j)}, y_j)$, $x^{(j)} \in D$, in the extended space of variables $(x_1, \dots, x_n, y)$. If this is done for example with respect to the term ordering `plex` with y highest in the initial ordering then the Gröbner basis will exhibit the interpolator as one of the G-basis elements

$$g = y - f(x)$$

for which $y_j - f(x^{(j)}) = 0$ for $x^{(j)} \in D$. In this case we shall write a general interpolator in the form

$$f(x) = \sum_{\alpha \in L} \phi_\alpha x^\alpha$$

Gröbner bases can be used to carry out Boolean operations. We give only a brief description here relevant to the rest of our discussion. Note first that in the binary case the statements $x_i(x_i - 1) = 0$ $(i = 1, \dots, n)$ constitute an algebraic variety with corresponding ideal

$$I_0 = \langle x_i(x_i - 1) : i = 1, \dots, n \rangle$$

In fact I_0 is already a total G-basis, that is a Gröbner basis whose leading terms are the same with respect to any ordering. Any complex Boolean statement can be reduced by taking the quotient with respect to this ideal. The remainder is the reduction we seek. Since the leading terms of I_0 are all of the form x_i^2 the remainder will be multi-linear in the x_i by the nature of the remainder r described above. Of course the same result will be achieved by multiplying out the Boolean expression substituting $x_i^2 = x_i$ wherever it occurs and collecting terms in the usual way. This is the standard Boolean reduction of logic.

The quotient operation can be used effectively when the statement is intersected with (conditioned on) several others. Thus let $F(x)$ be a Boolean function and let

$$\{G_j(x) = 0 : j = 1, \dots, m\}$$

be several simultaneous statements. Let

$$\langle g_j(x) : j = 1, \dots, k \rangle$$

be the G-basis for the combined ideal

$$I = \langle G_j(x) : j = 1, \dots, m \rangle \cap I_0$$

with respect to a suitable monomial ordering. Expanding

$$F(x) = \sum_{j=1}^{k} s_j(x) g_j(x) + r(x)$$

we can claim that on the variety $\{G_j(x) = 0 : j = 1, \dots, m\}$ it holds that

$$F(x) \equiv r(x)$$

A somewhat more general version is to insert the *value* of the $G_j(x)$ by writing the variety as $G_j(x) = v_j$ or $G_j(x) - v_j = 0$ $(j = 1, \dots, m)$. Then, recompute the G-basis considering the v_j as unknown coefficients and the remainder r will be a function also of these values

$$r(x, v_1, \dots, v_m) \tag{2}$$

As an example consider

$$F(x) = (1 - x_1 - x_4 + x_1 x_4)(x_1 + x_2 - x_1 x_2) x_3$$
$$G_1(x) = x_1(1 - x_4)$$
$$G_2(x) = x_1 x_3$$

where F corresponds to the Boolean statement $(A_1 \cup A_2) \cap A_3 \setminus (A_1 \cup A_4)$. Then $\{G_1(x), G_2(x), x_i(x_i - 1), i = 1, \dots, 4\}$ is already a total G-basis. Using the `normalf` procedure in Maple, which computes the remainder, we obtain

$$F(x) \equiv r(x) \equiv x_2 x_3 (1 - x_4)$$

on $G_1 = G_2 = 0$.

4. Attaching the probability: Binary random variables

As mentioned, any elementary event takes the form

$$E_J = \bigcap_{i \in J} A_i \bigcap \bigcap_{j \in J^c} A_j^c$$

for an index set J. It is enough to assign to each such event a probability

$$P(E_J) \geq 0 \ \text{(for all } J)$$

such that

$$\sum_J P(E_J) = 1$$

Additivity is then used to extend this to all Boolean expressions. Thus, suppose we have an expression F written in canonical form

$$F(x) = \psi_0 + \sum_{i=1}^{n} \psi_i x_i + \sum_{i<j} \psi_{ij} x_i x_j + \dots + \psi_{12,\dots,n} x_1, \dots, x_n = \sum_\alpha \psi_\alpha x^\alpha$$

where $\alpha = (\alpha_1, \dots, \alpha_n)$ is the binary string indexing the elementary monomial term and ψ_α is the corresponding parameter. Under $P(\cdot)$, each indicator function of the sets A_i $(i = 1, \dots, n)$, that is each x_i, can be considered as a random variable and we write X_i $(i = 1, \dots, n)$. Then the corresponding random variable is

$$F(X) = \psi_0 + \sum_{i=1}^{n} \psi_i X_i + \sum_{i<j} \psi_{ij} X_i X_j + \dots + \psi_{12,\dots,n} X_1, \dots, X_n = \sum_\alpha \psi_\alpha X^\alpha$$

Also since $F(x) = 0$ or 1, because of the indicator property

$$\text{Prob}(F(X) = 1) = E(F(X)) = \sum_{\alpha} \psi_\alpha E(X^\alpha)$$

Thus if $\alpha = (\alpha_1, \ldots, \alpha_n)$ and we define J_α so that $\alpha_i = 1$ for $i \in J_\alpha$ and $\alpha_i = 0$ otherwise then

$$X^\alpha = \prod_{i \in J_\alpha} X_i$$

and

$$m_\alpha = E(X^\alpha) = \text{Prob}\left(\omega \in \bigcap_{i \in J_\alpha} A_i\right)$$

is the moment corresponding to the index set α. However, this development does not have the complete algebraic flavour we require. The probability seems very much an addition, not particularly to be manipulated automatically.

We introduce two other methods of representing the probability. The first method concentrates on elementary events. Thus define an indicator function on the power set of indices $\{J\}$

$$Y_J = \begin{cases} 1 & \text{if } \omega \in E_J \\ 0 & \text{otherwise} \end{cases}$$

where E_J is an elementary event. Note that for all J

$$\begin{cases} Y_J(Y_J - 1) = 0 \\ \Sigma_J Y_J = 1 \end{cases}$$

The second expression forces the events E_J to be disjoint. Now attach the probabilities by defining

$$l(\omega) = \sum_J P(E_J) Y_J$$

This implies that when $\omega \in E_J$ we capture $l(\omega) = P(E_J)$.

The second method carries out essentially the same interpolation but using the indicator variables $\{x_i\}$ for the random variables X_i. This is equivalent to fitting the (complete) multi-linear function

$$p(x) = \sum_\alpha \psi_\alpha x^\alpha \tag{3}$$

to the values $P(E_J)$. Note that $p(x)$ can be decomposed uniquely into the weighted sum of special interpolators, one over each J. Thus

$$p(x) = \sum_J P(E_J) p_J(x)$$

where

$$p_J(x) = \prod_{i \in J} x_i \prod_{i \in J^c} (1 - x_i)$$

Considering the lattice of indices α under the usual partial ordering of set inclusion ($J_\alpha \subset J_\beta \Rightarrow \alpha \leq \beta$) the relationship between the $P(E_J)$ and the ψ_α is that of Möbius inversion (see for example Constantine 1987, Chapter 9). Thus we have

$$P(E_{J_\beta}) = \sum_{0 \leq \alpha \leq \beta} \psi_\alpha \prod_{i \in J_\alpha} x_i = \sum_{0 \leq \alpha \leq \beta} \psi_\alpha$$

and

$$\psi_\beta = \sum_{0 \leq \alpha \leq \beta} (-1)^{|\beta| - |\alpha|} P(E_{J_\alpha})$$

where $|\alpha|$ is the number of elements in J_α. We have a similar expression for moments in terms of $P(E_J)$

$$m_\beta = E(X^\beta) = \sum_{0 \leq \alpha \leq \beta} P(E_{J_\alpha})$$

$$P(E_{J_\beta}) = E(p_{J_\beta}(X)) = \sum_{0 \leq \alpha \leq \beta} (-1)^{|\beta| - |\alpha|} m_\alpha$$

One of the benefits of the polynomial representation is that the structure of the joint distribution is intimately connected to the structure of the interpolator (3). For example if the random variables X_i (and the events A_i) are independent then

$$\text{Prob}\left(\prod_{i=1}^n X_i = 1\right) = \prod_{i=1}^n \text{Prob}\{X_i = 1\}$$

$$= \prod_{i=1}^n p_i, \text{ say}$$

Then by the uniqueness of the interpolator, which follows by the G-basis theory, we have

$$p(x) = \prod_{i=1}^n (1 - p_i + (2p_i - 1)x_i)$$

We shall return to the notation of factorisation in more detail in Section 6.

We can summarise the joint distribution of X_i by the variety

$$\begin{cases} x_i(x_i - 1) = 0 \;\; (i = 1, \ldots, n) \\ t - p(x) = 0 \end{cases} \tag{4}$$

where t is a new indeterminate. This will be the standard form of expression for distributions.

5. Discrete random variables

5.1. *Univariate case*

The above dual approach carries over to general discrete random variables with finite support. We first study the case of a single univariate variable Z taking values $z_0, \ldots, z_{N-1}$ with probabilities $p_0, \ldots, p_{N-1}$ respectively ($p_i > 0, i = 0, \ldots, N-1$, $\sum_{i=0}^{N-1} p_i = 1$).

First take indicators Z_r for the values z_r

$$\begin{cases} Z_r(Z_r - 1) = 0 \;\; (r = 0, \ldots, N-1) \\ \sum_{r=0}^{N-1} Z_r = 1 \end{cases}$$

Then we can express Z as

$$Z = \sum_{r=0}^{N-1} z_r Z_r$$

and interpolate the probabilities by

$$l = \sum_{r=0}^{N-1} p_r Z_r$$

The alternative is to interpolate directly the probabilities p_r. We can use the G-basis method of constructing the interpolator described in Section 3 and the result will simply be the usual Lagrange interpolator

$$p(z) = \sum_{i=0}^{N-1} p_i \frac{\prod_{j\neq i}(z - z_j)}{\prod_{j\neq i}(z_i - z_j)}$$

Notice that the individual terms $p_j(z) = \prod_{j=i}(z - z_j)/\prod_{j\neq i}(z_i - z_j)$ are simply the polynomial interpolators of the individual indicator functions of the points $z_0, \ldots, z_{N-1}$ that is Z_r. Thus in analogy to equation (4) the variety

$$\begin{cases} \prod_{j=0}^{N-1}(z - z_j) = 0 \\ t - p(z) = 0 \end{cases} \tag{5}$$

holds all the information about the distribution.

The discrete finite nature of the support implies many relationships between moments and similar quantities. The following simple example illustrates this. Let the support for a random variable Z be $S = \{0, 1, 2\}$ which is held by the solution to

$$z(z-1)(z-2) = 0$$

or equivalently

$$z^3 = 3z^2 - 2z$$

Multiplying by z^r ($r > 0$, integer) we have

$$z^{3+r} = 3z^{2+r} - 2z^{1+r}$$

Now this same relationship holds for the random variable Z itself, so replacing z by Z and taking expectation we have

$$m_{3+r} = 3m_{2+r} - 2m_{r+1} \quad (r = 1, 2, \ldots)$$

where $m_r = E(Z^r)$ is the non-central moment of Z. We refer to this property as *moment aliasing*.

Now consider the general case in (5) with support $S = \{z_0, \ldots, z_{N-1}\}$. We can capture the moment aliasing by interpolating e^{sz}

$$e^{sz} = \sum_{j=0}^{N-1} b_j(s) z^j$$

Again taking Z random with support S we have the moment generating function

$$M_Z(s) = E(e^{sZ}) = \sum_{r=0}^{\infty} m_r \frac{s^r}{r!} = \sum_{j=0}^{N-1} b_j(s) m_r$$

This expresses all higher order moments in terms of $m_0 = 1, m_2, \ldots, m_{N-1}$.

We may also consider a rational version

$$\frac{1}{1 - sz} = \sum_{j=0}^{N-1} c_j(s) z^j$$

giving

$$H_Z(s) = E\left(\frac{1}{1 - sz}\right) = \sum_{r=0}^{\infty} m_j s^r = \sum_{j=0}^{N-1} c_j(s) m_j$$

Any function $g(\cdot)$ taking values on the support $S = \{z_0, \ldots, z_{N-1}\}$ can be interpolated on the support so that

$$g(z) = \sum_{\alpha\in S} \psi_\alpha z^\alpha \quad z \in S$$

Now again consider Z as random and we have a representation for any random variable which is function of Z as

$$G(Z) = \sum_{\alpha\in S} \psi_\alpha Z^\alpha$$

Taking expectations we have

$$E(G(Z)) = \sum_{\alpha\in S} \psi_\alpha m_\alpha$$

an expression for expectations in terms of the moments of Z which generalises moment aliasing.

5.2. *Multivariate case*

The extension to several random variables continues the ideas of the last two sections. We assume first that $Z = (Z_1, \ldots, Z_n)$ have a joint distribution on the product support

$$S = \prod_{i=1}^{n} \{z_{i,0}, \ldots, z_{i,N_i-1}\} \tag{6}$$

which can be written in ideal form

$$\left\{ \prod_{j=0}^{N_i-1} (z_i - z_{ij}) = 0 \quad (i = 1, \ldots, n) \right\}$$

We have the two methods of interpolating the distribution

$$p_{j_1 j_2 \ldots j_n} = \mathrm{Prob}(Z_i = z_{ij_i}, i = 1, \ldots, n)$$

First define for each Z_i

$$Z_i = \sum_{j=0}^{N_i-1} z_{ij} Z_{ij}$$

where

$$\begin{cases} Z_{ij}(1 - Z_{ij}) = 0 \\ \sum_{j=0}^{N_i-1} Z_{ij} = 1 \end{cases}$$

($i = 1, \ldots, n$, $j = 0, \ldots, N_i - 1$). Then the multivariate random variable Z can be written as

$$Z = \sum_{j_1,\ldots,j_n} (z_{j_i}, \ldots, z_{j_n}) \prod_{i=1}^{n} Z_{ij_i}$$

The first interpolation of the distribution takes the form

$$l = \sum_{j_1,\dots,j_n} p_{j_i,\dots,j_n} \prod_{i=1}^{n} Z_{ij_i}$$

The second interpolation uses as a basis all monomials of the form

$$z^\alpha = z_1^{\alpha_1}, \dots, z_n^{\alpha_n}$$

with $0 \le \alpha_i \le N_i - 1$. This can be written in the compact form

$$p(z) = \sum_{0 \le \alpha \le \bar{\alpha}} \phi_\alpha z^\alpha$$

(see Section 4 Equation (3)).

Let $\beta = (j_1, \dots, j_n)$, then the quantity $\prod_{i=1}^{n} Z_{ij_i}$ appearing in the first interpolator is simply the indicator function for the point $(z_{j_1}, \dots, z_{j_n})$. Each such indicator function can itself be interpolated

$$p_\beta(z) = \sum_\alpha \theta_{(\alpha,\beta)} z^\alpha$$

and

$$p(z) = \sum_\beta p_\beta p_\beta(z)$$

Conversely each z^α has an expansion in terms of the $p_\beta(z)$

$$z^\alpha = \sum_\beta \psi_{(\alpha,\beta)} p_\beta(z)$$

Taking expectations we have a relationship between moments and the probabilities $p_\beta = p_{j1j2\dots j_n}$

$$m_\alpha = E(Z^\alpha) = \sum_\beta \psi_{(\alpha,\beta)} E(p_\beta(Z))$$

$$= \sum_\beta \psi_{(\alpha,\beta)} p_\beta$$

and

$$p_\beta = \sum_\alpha \theta_{(\alpha,\beta)} m_\alpha$$

These transformations can be expressed in terms of the special matrix

$$X = [x^\alpha]_{x \in S, 0 \le \alpha \le \bar{\alpha}}$$

Thus, if $[m]$, $[\phi]$ and $[p]$ are the vectors corresponding to m_α, θ_α and p_β respectively we see easily that

$$[m] = X^t[p] = X^t X[\phi]$$

Thus, $[\psi_{(\alpha,\beta)}] = X^t$ and $[\theta_{(\alpha,\beta)}] = (X^t)^{-1}$. The matrix X is familiar as the X-matrix for the product support design and full factorial model. It is instructive to recapture the binary case of Section 4 using this notation. It also points to simplifications in this case using the indicator function with values $\{-1, 1\}$ rather than {0,1}. In that case $X^t X = 2nI$ and $[m] = 2n[\phi]$.

One of the benefits of the G-basis approach is that all this generalises to the case of arbitrary support, S, using the general version of the interpolation and the monomial terms for a particular monomial ordering $\tau\{x^\alpha : \alpha \in L\}$ with corresponding X-matrix, $X = [x^\alpha]_{x \in S, \alpha \in L}$.

The moment aliasing of the last section also generalises to completely arbitrary supports again using the generalised interpolation based on τ and L.

Thus interpolate $e^{\sum s_i z_i}$

$$e^{\sum s_i z_i} = \sum_{\alpha \in L} b_\alpha(s) z^\alpha$$

where $\alpha = (\alpha_1, \dots, \alpha_d)$ and $z^\alpha = z_1^{\alpha_1}, \dots, z_d^{\alpha_d}$. Then the full moment generating function is

$$M_Z(s) = E\left(e^{\sum s_i z_i}\right) = \sum_{\beta \ge 0} m_\beta \frac{s^\beta}{\beta!} = \sum_{\alpha \in L} b_\alpha(s) m_\alpha$$

where $\beta = (\beta_1, \dots, \beta_d)$, $m_\beta = E(Z_1^{\beta_1}, \dots, Z_d^{\beta_d})$, $s = (s_1, \dots, s_d)$, $s^\beta = s_1^{\beta_1}, \dots, s_d^{\beta_d}$ and $\beta! = \beta_1! \dots \beta_d!$.

The rational version is given first from the interpolation

$$\frac{1}{\prod_{i=1}^{d}(1 - s_i z_i)} = \sum_{\alpha \in L} c_\alpha(s) z^\alpha,$$

giving

$$H_Z(s) = E\left(\frac{1}{\prod_{i=1}^{d}(1 - s_i z_i)}\right) = \sum_{\beta \ge 0} m_\beta s^\beta = \sum_{\alpha \in L} c_\alpha(s) m_\alpha$$

In both cases we can express any higher order moment in terms of lower order moments in a unique way, given s and the chosen monomial ordering τ. The parameters $b_\alpha(s)$ and $c_\alpha(s)$ depend only on the support not otherwise on the distribution. These then are general forms of moment aliasing.

5.3. *Conditioning*

In Section 4 we saw how in the Boolean-binary case G-basis methods can be used to express conditioning statements. This extends to general supports.

Let J and K be two disjoint index sets with $J \cup K = 1, \dots, d$. Let $Z^{(1)}$ represent the random variable z_j for a j in J and $Z^{(2)}$ represents z_k, $k \in K$. The conditional distribution of $Z^{(1)}$ given $Z^{(2)} = z^{(2)}$ is

$$f\left(z^{(1)} \mid z^{(2)}\right) = \frac{f\left(z^{(1)}, z^{(2)}\right)}{m_2\left(z^{(2)}\right)}$$

where $f(\cdot, \cdot)$ is the joint distribution on support S and $m_2(z^{(2)})$ is the marginal distribution. For fixed $z^{(2)} = c$ the pair $(z^{(1)}, z^{(2)})$ lies on a sub-support given by restricting the support to the *section*, $z^{(2)} = c$. Suppose now that we have an interpolator

$$p(z) = \sum_{\alpha \in L} \phi_\alpha z^\alpha$$

of the joint distribution. Then we may simply set $z^{(2)}$ equal to a constant. Alternatively we may adjoin $z^{(2)} = c$ to the representation of the support and interpolate

$$p\big(z^{(1)}, c\big) = \sum_{\alpha \in L(c)} \phi_\alpha(c) z^\alpha$$

It is possible to show that $L(c) \subseteq L$ so that the restricted form uses a subset of the original monomial terms (provided the same monomial ordering is used).

To represent $m_2(z^{(2)})$ we take the projection of the original support into the $z^{(2)}$ hyper-plane. This can be carried out using elimination `(plex)`. In this case it is necessary to capture the marginal probabilities at the values of $z^{(2)}$.

Let $I_c(z^{(1)}, z^{(2)})$ be the indicator polynomial function which takes the value unity whenever $z^{(2)} = c$ and is zero at other values. Then

$$m_2(c) = E\big(I_c\big(z^{(1)}, z^{(2)}\big)\big)$$

where expectation is with respect to the full joint distribution.

6. Factorisation and exponential families

In several places in the above analysis we have seen that for the uniqueness of the interpolation of the joint distribution a product support is critical. The description of a support that is not of product form depends on G-basis theory: with respect to any term ordering the representation of the distribution in the quotient ring with respect to the support ideal is unique. It is this uniqueness which allows us to switch backwards and forwards between discrete distributions and their interpolators.

The uniqueness is particularly valuable for expressing independence: factorisation of the probability carries over perfectly to factorisation of the interpolating polynomials. Thus in two dimensions if we express the joint distribution as

$$\begin{cases} \prod_{i=1}^{N_1}(z_1 - z_{1i}) = 0 \\ \prod_{i=1}^{N_2}(z_2 - z_{2i}) = 0 \\ t - p(z_1, z_2) = 0 \end{cases}$$

and if Z_1 and Z_2 are independent random variables

$$p(z_1, z_2) = m_1(z_1) m_2(z_2)$$

then the marginal distributions have the representations

$$\begin{cases} \prod_{i=1}^{N_1}(z_1 - z_{1i}) = 0 \\ s - m_1(z_1) = 0 \end{cases}$$

$$\begin{cases} \prod_{i=1}^{N_2}(z_2 - z_{2i}) = 0 \\ t - m_2(z_2) = 0 \end{cases}$$

both of which are themselves unique representations.

Rewrite the product support as

$$S = S_1 \times \cdots \times S_n$$

where the variable z_j takes values in S_j $(j = 1, \ldots, n)$ and for simplicity $S_i = \{0, 1, 2, \ldots, N_i - 1\}$ so that $|S_i| = N_i$ and $|S| = \prod_{i=1}^{N} N_i$. Let $\{f(z) > 0, z \in S\}$ be the discrete probability function of the random variable Z, with support S. Then, in fact, there are two interpolators of interest in statistics: the interpolator of $f(z)$, discussed in the previous section, and of $\log f(z)$. Thus let

$$p(z) = \sum_{\alpha \in S} \phi_\alpha z^\alpha$$

with $p(z) = f(z)$ $(z \in S)$ and

$$\tilde{p}(z) = \sum_{\alpha \in S} \theta_\alpha z^\alpha$$

with $\tilde{p}(z) = \log f(z)$, $(z \in S)$. Here $\log f(z)$ is well defined because $f(z) > 0$ for all $z \in S$. In this section we concentrate on $\tilde{p}(z)$ and its relationship to the exponential family.

If we exponentiate $\tilde{p}(z)$ we obtain

$$\begin{aligned} f(z) &= \exp(\tilde{p}(z)) \\ &= \exp\left(\sum_{\substack{\alpha \in S \\ \alpha \neq (0,\ldots,0)}} \theta_\alpha z^\alpha - K(\theta)\right) \quad (z \in S) \end{aligned} \tag{7}$$

Here the constant term $\exp(\theta_{(0,\ldots,0)})$ has been replaced by the exponential term $\exp(-K(\theta))$. We call such a model a saturated exponential model. As we range over all possible $f(z)$ we obtain a full exponential family with the z^α as sufficient statistics. The term $K(\theta)$ is the cumulant generating function of the random variables

$$Z^\alpha = Z_1^{\alpha_1}, \ldots, Z_n^{\alpha_n} \quad (\alpha \in S \backslash (0, \ldots, 0))$$

with respect to the base distribution, in this case the uniform. Model (7) is the largest exponential model which can be built on the support S. It has $N - 1$ parameters as one would expect because of the restriction $\sum_{z \in S} p(z) = 1$.

From exponential family theory (or by direct evaluation) allowing θ_α to vary we have that

$$\begin{aligned} E_\theta(Z^\alpha) &= \frac{\partial K(\theta)}{\partial \theta_\alpha} \\ Cov_\theta(Z^\alpha, Z^\beta) &= \frac{\partial^2 K(\theta)}{\partial \theta_\alpha \partial \theta_\beta} \end{aligned} \tag{8}$$

for $\alpha, \beta \in \tilde{S} = S \backslash (0, \ldots, 0)$, where these moments are with respect to the distribution (7).

The variance function of an exponential family relates the variance-covariance terms above directly to the expectation terms. The first contribution of the interpolation method is a short method for expressing this.

Let

$$\{g_1(z), \ldots, g_k(z)\}$$

be a G-basis for the support S under a monomial ordering τ and write

$$Cov_\theta(Z^\alpha, Z^\beta) = E_\theta(Z^{\alpha+\beta}) - E_\theta(Z^\alpha) E_\theta(Z^\beta)$$

With S of product form two situations arise

$$\alpha + \beta \in S \quad \text{or} \quad \alpha + \beta \notin S$$

When $\alpha + \beta \in S$ we can write

$$Cov_\theta(Z^\alpha, Z^\beta) = E_\theta(Z^{\alpha+\beta}) - E_\theta(Z^\alpha)E_\theta(Z^\beta)$$

When $\alpha + \beta \notin S$ we instead interpolate $Z^{\alpha+\beta}$. Thus suppose

$$z^{\alpha+\beta} = \sum_{i=1}^{k} s_i(z)g_i(z) + r(z)$$

where $r(z) = \sum_{\delta \in S} \psi_\delta z^\delta$. That the remainder uses monomials from the set used to interpolate, is critical and follows from the algebraic theory. But

$$z^{\alpha+\beta} = r(z) \text{ for } z \in S$$

since by construction $g_i(z) = 0$ for $z \in S$ $(i = 1, \ldots, t)$. Thus, applying the interpolator directly to the random variables

$$\begin{aligned} Cov_\theta(Z^\alpha, Z^\beta) &= E_\theta(Z^{\alpha+\beta}) - E_\theta(Z^\alpha)E_\theta(Z^\beta) \\ &= E_\theta\left(\sum_{\delta \in S} \psi_\delta Z^\delta\right) - E_\theta(Z^\alpha)E_\theta(Z^\beta) \\ &= \sum_{\delta \in S} \psi_\delta E_\theta(Z^\delta) - E_\theta(Z^\alpha)E_\theta(Z^\beta) \end{aligned}$$

we have thus obtained the variance function, namely an expression of the covariance in terms of the means of Z^α's with $\alpha \in S$. Note also that the relationship is multi-linear. We collect this result in the following theorem.

Theorem 1. *For a discrete distribution expressed as a saturated exponential model with respect to a uniform distribution on a product set $S = S_1 \times \cdots \times S_n$, the variance function for monomials Z^α, $\alpha \in \tilde{S} = S \backslash (0, \ldots, 0)$ is multi-linear in $E(Z^\alpha)$.*

Now consider the more difficult case of an arbitrary discrete support S. We may still use an interpolator but in general the remainder $r(Z)$ is not unique and hence the construction of the interpolator is also not unique. We can, however, define the exponential model, that is the interpolator $\tilde{p}(z)$, in a consistent way. Thus, following the design theory, select a monomial ordering τ and construct the G-basis $G = \{g_i : i = 1, \ldots, k\}$ for the zero-dimensional ideal corresponding to S, which is now general. The quotient ring is generated (spanned) by a set of monomials $\{z^\alpha : z \in L\}$ where we recall that L is an order ideal, that is $z^\alpha \in L \Rightarrow z^\beta \in L$ $(\beta \in L)$ whenever z^β divides z^α. All such ideals satisfy the divisibility condition. Note in particular that 1 is included since $1 = z_1^0, \ldots, z_n^0$. We can then define both types of interpolator which use terms $z^\alpha (\alpha \in L)$

$$p(z) = \sum_{\alpha \in L} \phi_\alpha z^\alpha$$

$$\tilde{p}(z) = \sum_{\alpha \in L} \theta_\alpha z^\alpha$$

and by construction

$$f(z) = p(z), \quad \log f(z) = \tilde{p}(z) \ (z \in S)$$

Now we are in a position to generalise the exponential family model to

$$f(z) = \exp\left(\sum_{\alpha \in \tilde{L}} \theta_\alpha z^\alpha - K(\theta)\right)$$

where $\tilde{L} = L \backslash (0, \ldots, 0)$, absorbing the constant term into $K(\theta)$.

Now (8) holds with $\tilde{S}$ replaced by $\tilde{L}$. Again we can interpolate $Z^{\alpha+\beta}$ over S

$$Cov_\theta(Z^\alpha, Z^\beta) = \sum_{\delta \in L} \psi_\delta E_\theta(Z^\delta) - E_\theta(Z^\alpha)E_\theta(Z^\beta) \tag{9}$$

where $z^{\alpha+\beta} = \sum_{\delta \in L} \psi_\delta z^\delta (z \in S)$. Theorem (1) thus extends to general support.

Theorem 2. *A discrete distribution expressed as a saturated exponential model on an arbitrary support with respect to the uniform distribution in $\mathbb{R}^n$ has a (possibly non–unique) multilinear variance function.*

We should re-emphasize the important point that the form (9) depends on the choice of monomial ordering. Only for certain supports will the interpolator $r(z)$ be unique, that is L is unique.

Consider as an example the distribution on the five points $(\pm 1, \pm 1)$ and (2, 2) with saturated exponential model. Under the `tdeg`$(z_1 > z_2)$ ordering the G-basis for this support is

$$z_1^2 - z_2^2, z_2^3 - 2z_2^2 - z_2 + 2, z_1 z_2^2 - 2z_2^2 - z_1 + 2$$

It follows that to interpolate we use the terms

$$\{1, z_1, z_2, z_1 z_2, z_2^2\}$$

so that

$$L = \{(0, 0), (1, 0), (0, 1), (1, 1), (0, 2)\}$$

Thus the saturated exponential model is

$$\exp\big(\theta_1 z_1 + \theta_2 z_2 + \theta_3 z_1 z_2 + \theta_4 z_2^2 - K(\theta)\big)$$

First we compute the quadratic variance function. For this we need to compute the interpolators of the higher order monomials, $E(Z^{\alpha+\beta})$ which will appear in the covariance terms. We do this by computing the remainder using the `normalf` command in Maple

$$\begin{aligned} z_1^2 &= z_2^2 \\ z_1^2 z_2 &= 2z_2^2 + z_2 - 2 \\ z_1 z_2^2 &= 2z_2^2 + z_1 - 2 \\ z_2^3 &= 2z_2^2 + z_2 - 2 \\ z_1^2 z_2^2 &= 5z_2^2 - 4 \\ z_1 z_2^3 &= 4z_2^2 + z_1 z_2 - 4 \\ z_2^4 &= 5z_2^2 - 4 \end{aligned}$$

from which the variance function can easily be computed as described.

The reduction of $z^{\alpha+\beta}$ used in the above analysis is in effect the same as that used in the construction of the moment generating function of the last section. A version more tailored to the exponential family development is useful. Thus, the moment generating function (mgf) of the quantities Z^α, $\alpha \in \tilde{L}$ under the full exponential model $f(z) = \exp(\sum_{\alpha\in\tilde{L}} \theta_\alpha z^\alpha - K(\theta))$ is

$$\begin{aligned} M_\theta(s) &= \sum_{z\in S} \exp\left(\sum_{\alpha\in\tilde{L}} s_\alpha z^\alpha + \sum_{\alpha\in\tilde{L}} \theta_\alpha z^\alpha - K(\theta)\right) \\ &= M_0(\theta)^{-1} M_0(s+\theta) \end{aligned}$$

where $M_0(\cdot)$ is the mgf at $\theta = 0$.

Again interpolating

$$\exp\left(\sum_{\alpha\in\tilde{L}} \theta_\alpha z^\alpha\right) = \sum_{\beta\in L} e_\beta(\theta) z^\beta \quad (x \in S)$$

and taking expectation (at $\theta = 0$) we have

$$M_0(\theta) = \sum_{\beta\in L} e_\beta(\theta) m_\beta^0$$

where m_β^0 are the moments at $\theta = 0$. We thus have a complete development of all moments in terms of quantities e_β, m_β which depend only on the support, S, and the monomial ordering τ.

We can make the above analysis a little more general by changing the base distribution to a more general distribution than uniform

$$\begin{aligned} f(z) &= \exp(\tilde{p}(z)) \\ &= \exp\left(\sum_{\alpha\in\tilde{L}} \theta_\alpha z^\alpha - K(\theta)\right) f_0(z) \end{aligned}$$

where $f_0(z)$ is the distribution at $\theta = 0$ and $K(\theta)$ is its cumulant generating function. Taking logs we see that this is equivalent to interpolating

$$\log f(z) - \log f_0(z)$$

We may also interpolate $\log f_0(z)$ separately so that $\tilde{p}_0(z) = \log f_0(z)$ $(z \in S)$ and $f_0(z) = \exp(\tilde{p}_0(z))$ $(z \in S)$. Theorems 1 and 2 are thus extended to this case, replacing the uniform distribution by the general $f_0(z)$.

7. Bayes and submodels

In Bayesian theory, and particularly Bayes networks, conditional independence plays an important role. Thus

$$Z_1 \perp Z_2 \mid Z_3$$

gives the factorisation of densities

$$f(z_1, z_2, z_3) = \frac{f(z_1, z_3) f(z_2, z_3)}{m(z_3)}$$

where $m(z_3)$ is the marginal distribution of Z_3. We have seen above that pure independence carries over nicely to factorisation of densities. The situation with conditional independence is more complex, but considerable progress can be made by considering the exponential family representation of the last section. The conditional independence structure is held by the non-zero θ_α in the structure of the interpolator $\tilde{p}(z) = \exp(\sum_{\alpha\in\tilde{L}} \theta_\alpha z^\alpha - K(\theta))$.

We say that z_i and z_j $(i \neq j)$ *appear together in an interaction term* in $\tilde{p}(z)$ if there is an α with $\theta_\alpha \neq 0$ such that $\alpha_i, \alpha_j > 0$. This is by direct analogy with polynomial regression.

Now construct an undirected graph G in which a node j is associated with each random variable. Draw an edge $i \leftrightarrow j$ if z_i and z_j appear together in at least one interaction term. This graph is then precisely the usual undirected graph developed in Bayes networks and all conditional independence structures can be read off it. Thus

$$Z_1 \perp Z_2 \mid Z_3$$

is equivalent to node 3 (dis)connecting node 1 and node 2 in the sense that if node 3 is removed (together with all its edges) nodes 1 and 2 are disconnected.

This corresponds to conditional additivity in the exponential term

$$\sum_{\alpha\in\tilde{L}} \theta_\alpha z^\alpha = \sum_{\alpha\in\tilde{L}_1} \theta_\alpha z^\alpha + \sum_{\alpha\in\tilde{L}_2} \theta_\alpha z^\alpha$$

where the first term on the right hand side has all appropriate monomials in z_1 and z_3 but not z_2 and the second term in z_2 and z_3 but not z_1. The computations under conditioning can be carried out by adapting the arguments in Section 5.3 to the log-interpolation case.

All the usual properties hold provided $f(z) > 0$ for all z in the product support. For example G has the global Markov property with respect to $Z_1, \ldots, Z_n$ (defined in the previous sections) and the Hammersley-Clifford theorem relating the pairwise Markov property applies (see Lauritzen 1996, Chapter 3).

A more complex question is whether a pure polynomial representation, $p(z)$, of the conditional independence is possible generalising the easier factorisation for simple independence mentioned in the last section. One way of proceeding is to interpolate each term in the exponential representation with a polynomial

$$p_\alpha(z) = \exp(\theta_\alpha z^\alpha) \quad (z \in S)$$

(for $\theta_\alpha \neq 0$). Then we can construct an interpolator $p^\star(z)$ of $f(z)$

$$f(z) = p^\star(z) = c \prod_{\{\alpha:\theta_\alpha\neq 0\}} p_\alpha(z) \quad (z \in S, \quad c > 0)$$

Note that since $p(z)$ of the last section also interpolates $f(z)$ we have

$$p(z) = p^\star(z) \quad (z \in S)$$

It is important to note that this is a factorisation of $p(z)$ on S rather than on R^n (where the factorization would be unique) and typically $p^\star(z)$ has higher degree than $p(z)$. To summarize, it is possible to construct polynomial factorisations of $p(z)$ which hold on S, and reflect the conditional independence structure, which may not hold on $\mathbf{R}^n$.

A more complex theory arises when we have both $\theta_\alpha = 0$ for some α and also $f(z) = 0$ for some $z \in S$. This happens when we consider special submodels or for example when we have structural zeros in contingency tables. This impacts both on the usual maximum likelihood theory which may be developed from the exponential family representation in the last section and the existence of, and manipulations with, the conditional independence structure. The challenge is to extend both theories using the general support and the interpolation theory based on the order ideal of monomials, L, of the last section. Exponential submodels are exhibited as submodels $L' \subset L$ in which certain $\theta_\alpha = 0$ for α in L. The benefit of the algebraic theory is that once a term-ordering is specified the basic order ideal L is unique, so that the starting point for such a theory is well-defined. This submodel theory will appear in a forthcoming monograph by the authors.

References

Caboara M., Pistone G., Riccomagno E., and Wynn H.P. 1997. The fan of an experimental design. SCU Report 33, Department of Statistics, University of Warwick.

Capani A., Niesi G., and Robbiano L. 1995. CoCoA, a system for doing computations in commutative algebra. Available via anonymous ftp from lancelot.dima.unige.it.

Char B., Geddes K., Gonnet G., Leong B., Monogan M., and Watt S. 1991. MAPLE V Library Reference Manual. Springer-Verlag, New York.

Chazarain J., Riscos A., Alonso J.A., and Briales E. 1991. Multi-valued logic and Gröbner bases with applications to modal logic. J. Symbolic Computation 11: 181–194.

Constantine G.M. 1987. Combinatorial Theory and Statistical Design. John Wiley and Sons, New York.

Cox D., Little J., and O'Shea D. 1996. Ideal, Varieties, and Algorithms, 2nd ed. Springer-Verlag, New York.

Diaconis P. and Sturmfels B. 1998. Algebraic algorithms for sampling from conditional distributions. The Annals of Statistics 26(1): 363–397.

Hailperin T. 1976. Boole's logic and probability. Studies in Logic and Foundations of Mathematics. Springer, Amsterdam.

Lauritzen S.L. 1996. Graphical Models. Clarendon Press, Oxford.

Pistone G., Riccomagno E., and Wynn H.P. 2000. Algebraic Statistics, Chapman and Hall/CRC, London.

Pistone G. and Wynn H.P. 1996. Generalised confounding with Gröbner bases. Biometrika 83(3): 653–666.