

Gröbner basis methods for structuring and analysing complex industrial experiments

GIOVANNI PISTONE

*Dipartimento di Matematica, Politecnico di Torino
Corso Duca degli Abruzzi 24, 10129 Torino (Italy)
pistone@calvino.polito.it*

EVA RICCOMAGNO

*EURANDOM
Den Dolech 2, 5612 AZ, Eindhoven (NL)
Riccomagno@eurandom.nl*

and

HENRY P. WYNN

*Department of Statistics, University of Warwick
Coventry, CV4 7AL (UK)
hpw@stats.warwick.ac.uk*

This work extends the research programme of the authors into the design and analysis of complex experiments.^{10,11,12} It is seen how the special algebraic structures studied in the polynomial ring algebra and Gröbner basis environment can be exploited for situations in which there is blocking, nesting, crossing and so on, or where groups of factors are “favoured” over others. The connection is made between the Gröbner basis methods and the more classical symbolic formalism associated for example with Generalised Linear Model packages, such as Splus and Glim.¹⁶ Examples are given from physical experiments in engine mapping.

1. Gröbner basis method

The algebra leading to the theory described in this paper assumes that the observations and the design points are without error, that is we study the identifiability of the GLM component of a given model. In other words Gröbner basis methods treat the identifiability problem for the deterministic part of linear regression models given a design. We recall very briefly some basic notions. A design is here a finite set of distinct points in \mathbf{R}^m where m is the number of independent factors in

the experiment. We consider models of the type

$$\eta(x) = \sum \theta_\alpha x^\alpha + \varepsilon(x) \quad (1)$$

where the sum is finite, x runs over \mathbf{R}^m and x^α 's are monomials or terms: $x^\alpha = x_1^{\alpha_1} \dots x_m^{\alpha_m}$, where the α_i 's for $i = 1, \dots, m$ are non-negative integers. In particular 1 represents the constant component of GLM models where $\alpha_i = 0$, $i = 1, \dots, m$. The θ 's are parameters, that is unknown constants. Finally $\varepsilon(x)$ is some error, for example $\varepsilon(x)$ is normally distributed around zero with variance σ^2 . We concentrate on the non-stochastic part of the model, namely $\sum \theta_\alpha x^\alpha$.

Let the design points and the model terms be ordered as $D = [d_1, \dots, d_N]$ and $[x^\alpha]_{\alpha \in L}$ respectively. Here L is a list of m -dimensional, non-negative integer vectors giving the exponents α of the model terms. We shall see in the next section that the methods require us to order the model terms. The matrix form of the deterministic part of Model (1) can be written as

$$Y = X\Theta$$

where Y is the observation vector, $\Theta = [\theta_\alpha]_{\alpha \in L}$ is the vector of the parameters, X is the design matrix with as many rows as design points and as many columns as terms in the model and (i, j) -entry of X is the j -th term of the model evaluated at the i -th design point.

Given a design D the aim is to determine a set of terms $\{x^\alpha\}_{\alpha \in L}$ such that the matrix X is full rank.

The starting point of the algebraic method is to express the set of design points, $D = \{d_1, \dots, d_N\} \subset \mathbf{R}^m$, as the solution to a set of polynomial equations in as many variables as statistical factors. This set of equations forms the basis of a zero-dimensional polynomial ideal, $\text{Ideal}(D) \subset \mathbf{R}[x_1, \dots, x_m]$, called a design ideal, and thus the theory of polynomial ideals can be applied.³ The Gröbner basis methods lead to a very natural class of identifiable models.¹¹

2. Term-orderings

One of the main objects in this paper is a term-ordering. A term-ordering τ is a total ordering on the set of all monomials, x^α or GLM components of any possible polynomial model in a given number of factors. Moreover such a total ordering has to be compatible with the simplification of monomials, that is $x^\alpha \succ_\tau x^\beta$ if x^β divides x^α , equivalently if there exists x^γ , $\gamma \neq (0, \dots, 0)$ such that $x^\alpha = x^\gamma x^\beta$. In particular $x^\alpha \succ_\tau 1$ for all $\alpha \neq (0, \dots, 0)$.

Given a term-ordering, τ , a finite subset G of a polynomial ideal I is a Gröbner basis of I with respect to τ if

$$(i) \ G \text{ generates } I, \text{ that is for all } f \in I, f = \sum_{g \in G} s_g g \text{ for certain polynomials } s_g$$

$$(ii) \ \langle \text{LT}_\tau(f) : f \in I \rangle = \langle \text{LT}_\tau(g) : g \in G \rangle$$

where for a set of polynomials A , $\langle A \rangle$ is the ideal generated by A according to (i) and $\text{LT}_\tau(f)$ is the largest term in f with respect to τ .

Given a term-ordering, τ , a Gröbner basis of the design ideal, $G = \{g_1, \dots, g_t\}$ is computed. A \mathbf{R} -vector space basis of the quotient space $\mathbf{R}[x_1, \dots, x_m]/\langle G \rangle = \mathbf{R}[x_1, \dots, x_m]/\text{Ideal}(D)$ is given by the monomials not dividing any of the leading terms of the Gröbner basis. We call this set of monomials Est_τ where the subscript τ stresses the dependency on the term-ordering. Sometimes we write $\text{Est}_\tau(D)$ where D is the design. The set Est_τ is a saturated set, that is it has as many elements as design points, and can be used as support for an identifiable, saturated linear model as in (1)

$$\ell(\eta(x)) = \sum_{x^\alpha \in M \subset \text{Est}_\tau} \theta_\alpha x^\alpha$$

where θ_α 's are identifiable parameters, M holds the model terms and $\ell(\eta(x))$ is the deterministic component of the model $\eta(x)$. The set Est_τ has the structure of an order ideal, equivalently it gives a hierarchical model structure. McCullagh and Nelder speak of functional marginality.⁹ That is if a higher order term is included in Est_τ then its factors are also included. In particular the intercept is always identifiable. Notice that the matrix X for the design D and the model M is full rank and for $M = \text{Est}$ the matrix X is invertible.

The above technology makes use of computer algebra packages such as Maple and CoCoA.^{2,4}

As an example consider the two-dimension five-point design

$$D = \{(0, 0), (1, 0), (2, 0), (1, 1), (-2, 2)\}$$

We chose the term-ordering τ given by the matrix A (see Section 3)

$$A = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$$

In particular $x_2 \succ_\tau x_1$. The Gröbner basis with respect to τ is given by the following three polynomials

$$\begin{aligned} x_1 x_2 + 3x_2^2 - 4x_2 \\ x_2^3 - 3x_2^2 + 2x_2 \\ x_1^3 - 3x_1^2 + 12x_2^2 + 2x_1 - 12x_2 \end{aligned}$$

and the leading terms are $x_1 x_2, x_2^3, x_1^3$. The above equations can be easily checked in Maple or CoCoA. The set Est is given by the monomials not divisible by those leading terms. Thus the set Est is given by the following five terms

$$1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_2^2$$

For the full saturated model the matrix X below is invertible

$$X = \begin{array}{c|ccccc} & 1 & x_1 & x_2 & x_1^2 & x_2^2 \\ \hline (0,0) & 1 & 0 & 0 & 0 & 0 \\ (1,0) & 1 & 1 & 0 & 1 & 0 \\ (2,0) & 1 & 2 & 0 & 4 & 0 \\ (1,1) & 1 & 1 & 1 & 1 & 1 \\ (-2,2) & 1 & -2 & 2 & 4 & 4 \end{array}$$

More examples are given below.

3. Special monomial orderings: blocking

In the practical applications to date use has been made of the basic monomial orderings like `tdeg` and `plex`, which are available in standard packages. The basic definitions are in Robbiano (1985).¹⁴ However it is also possible to use a very general definition of a monomial ordering based on a special matrix construction. Associated to a (total rational) monomial ordering τ there is a matrix A such that A is square, with integer entries, full rank and the first non-zero entry in each column is positive. Then the monomial $x^\alpha = x_1^{\alpha_1} \dots x_m^{\alpha_m}$ is larger than x^β with respect to the term-ordering τ associated with A if and only if

$$A(\alpha_1 \dots \alpha_m)^t \succ_{lex, \mathbf{Z}^m} A(\beta_1 \dots \beta_m)^t \quad (2)$$

where $\succ_{lex, \mathbf{Z}^m}$ is the lexicographic ordering over \mathbf{Z}^m . Recall that for two integer vectors (a_1, \dots, a_m) and (b_1, \dots, b_m) we have $(a_1, \dots, a_m) \succ_{lex, \mathbf{Z}^m} (b_1, \dots, b_m)$ if and only if $a_1 > b_1$ or $a_i = b_i$ for all $i = 1, \dots, p$ and $a_{p+1} > b_{p+1}$ for some $p < m$.

The key to the present work is to translate the kind of structural requirements of the design (blocking, favourite factors, etc.) into the structure of the matrix A .

Note first that a term-ordering has associated an initial order on the factors: x_j precedes x_i in the initial order if $x_i \succ x_j$ in the term-ordering. Next, under a lexicographic ordering such as `plex` all possible monomial terms in x_1 would appear in Est_τ before those involving x_2 etc. Under a “graded” ordering such as `tdeg` all possible terms of total degree r would appear before those of degree s , $r < s$. Recall that the total degree of x^α is $\sum_{i=1}^m \alpha_i$. We can combine these two ideas together by blocking the ordering matrix

$$A = \begin{pmatrix} A_1 & 0 & \dots & \\ 0 & A_2 & 0 & \\ 0 & 0 & \ddots & 0 \\ \dots & \dots & 0 & A_k \end{pmatrix}$$

and blocking the variables $[x^{(1)} : \dots : x^{(k)}]$. It is important to note that the term “blocking” here is different from the usual meaning in experimental design. Between blocks a `plex` type ordering operates but within a block, $x^{(t)}$ the monomial ordering would be as given by A_t . For example the “within-block” ordering could be `tdeg`.

Thus factors in block A_k will appear in Est before factors in other blocks. A simple but important example is where one factor has a complex output relationship and the other factors are simpler, something which occurs in repeated measurements.⁷ Then, factor one would be in a block on its own and the remaining factors treated with, say, **tdeg**.

One could build up total term-orderings in m factors in more general ways which include the above blocking method. It is clear that a term-ordering on m factors induces an ordering on each subset of factors by projection. The blocking method is one way of building orderings from the bottom up. Referring to the integer representation of term-ordering defined in Equation (2), separated orderings on \mathbf{Z}^p and \mathbf{Z}^q are combined to define an ordering on \mathbf{Z}^{p+q} .

4. Blocking and classical structures

Simple crossed experiments correspond to the so-called minimal fan designs and have a unique Est, whatever the monomial ordering. In this case Est consists of crossed models and we write, for example, $A \times B$. Thus for a full factorial design with l_i levels in the i th dimension ($i = 1, \dots, m$) the corresponding Est is the “full factorial” model with terms $x_1^{k_1} \dots x_m^{k_m}$ with $k_i = 0, \dots, l_i - 1$. A generalisation is given by the so-called echelon designs. A design D is an echelon design if for any design point (d_1, \dots, d_m) all points of the form (y_1, \dots, y_m) with $0 \leq y_i \leq d_i$ ($i = 1, \dots, m$) belong to D . For an echelon design the identifiable set is given by a similar structure to that of the design

$$x_1^{d_1} \dots x_m^{d_m} \quad \text{for all} \quad (d_1, \dots, d_m) \in D$$

This result can be extended to generalised echelon designs, that is designs such that if a point (d_1, \dots, d_m) is in the design then also all the points of the form (y_1, \dots, y_m) with $0 \leq |y_i| \leq |d_i|$ ($i = 1, \dots, m$) belong to the design, where $|a| = a$ if $a \geq 0$ and $|a| = -a$ if $a < 0$.¹

Let now fix a term-ordering τ . For a pair of designs such that $D_1 \subseteq D_2$ then $\text{Est}(D_1) \subseteq \text{Est}(D_2)$. Moreover the terms added to $\text{Est}(D_1)$ to obtain $\text{Est}(D_2)$ are the smallest terms with respect to τ for which the design matrix is invertible.¹³

In particular it follows that $\text{Est}(D_1) \cup \text{Est}(D_2) \subseteq \text{Est}(D_1 \cup D_2)$. Indeed $D_2 \subseteq D_1 \cup D_2$ and $D_1 \subseteq D_1 \cup D_2$ imply $\text{Est}(D_2) \subseteq \text{Est}(D_1 \cup D_2)$ and $\text{Est}(D_1) \subseteq \text{Est}(D_1 \cup D_2)$ respectively and thus $\text{Est}(D_1) \cup \text{Est}(D_2) \subseteq \text{Est}(D_1 \cup D_2)$. The equality in the last inclusion is in general false, for an example consider the union design to be $\{(0, 0), (1, 1)\}$ and any term-ordering.

Similarly we have $\text{Est}(D_1 \cap D_2) \subseteq \text{Est}(D_1) \cap \text{Est}(D_2)$. This inclusion may be strict as the following trivial example shows. Let the designs be $D_1 = \{(0, 0)\}$ and $D_2 = \{(1, 1)\}$. Now $\text{Est}(D_1 \cap D_2)$ is the empty set and $\text{Est}(D_1) = \text{Est}(D_2) = \{1\}$ the constant term.

Before proceeding we should introduce the notion of a compatible term-ordering. The term-ordering τ in $m + p$ dimensions is compatible with the term-ordering τ_1 in m dimension when $x^\alpha \succ_{\tau_1} x^\beta$ if and only if $x^\alpha \succ_\tau x^\beta$ for all monomials x^α, x^β

in the smaller space of m factors. For example consider the term-orderings τ_1 in m factor with corresponding matrix A_1 and τ_2 in other p factors with matrix A_2 . Then the block term-ordering described on Section 3 on $m + p$ factors operating on the first factors according to τ_1 and on the second factors according to τ_2 and corresponding to the matrix

$$\begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

is compatible with both τ_1 and τ_2 .

Consider now designs on different sample spaces. Let D_1 be a design in \mathbf{R}^m and D_2 a design in \mathbf{R}^p and let the cross-product design be defined as

$$D_1 \times D_2 = \{(d_1, d_2) : d_1 \in D_1, d_2 \in D_2\}$$

Let G_1 be a Gröbner basis for D_1 with respect to the term-ordering τ_1 and G_2 a Gröbner basis for D_2 with respect to τ_2 . The set of polynomials $G = \{g, h : g \in G_1, h \in G_2\}$ is an interpolating set for $D_1 \times D_2$. It is also a Gröbner basis with respect to any term-ordering on the monomials in $m + p$ factors compatible with τ_1 and τ_2 . This can be proved with a very useful tool in Gröbner basis theory called the S-polynomial test. We refer to Cox, Little and O'Shea (1997)³ and only mention the main definition and theorem as S-polynomials are the main tool to prove that a polynomial set is a Gröbner basis. Given a term-ordering, τ the S-polynomial associated with two polynomials f and g is

$$\frac{\text{LCM}(\text{LT}_\tau(f), \text{LT}_\tau(g))}{\text{LT}_\tau(f)} f - \frac{\text{LCM}(\text{LT}_\tau(f), \text{LT}_\tau(g))}{\text{LT}_\tau(g)} g$$

where LCM stands for least common multiple. One can prove that a set G is a Gröbner basis for an ideal I with respect to τ if $G \subset I$ and the remainder of the division of the S-polynomial for each pairs (f, g) with $f, g \in G$ is zero. More importantly the structure of the Gröbner basis G implies that the crossed model $\text{Est}(D_1) \times \text{Est}(D_2)$ is identifiable by $D_1 \times D_2$ because $\text{Est}(D_1) \times \text{Est}(D_2)$ is the set of all monomials in $m + p$ variables not divisible by the leading terms of G_1 and G_2 , that is the leading terms of G .

There is a link between cross-product designs and echelon designs. If D_1 and D_2 are echelon then $D_1 \times D_2$ is echelon. This follows directly from the definition of echelon design. Moreover $\text{Est}(D_1 \times D_2) = \text{Est}(D_1) \times \text{Est}(D_2)$.

Similarly, purely additive models would be written $A + B$. For complex experiments it is very difficult to prescribe the allowable interactions, additive terms, etc. However the Gröbner basis methods can reveal the relationship between the allowable model structures and the designs (as varieties) more transparently via the ‘‘blocking’’ of the monomial ordering.

For two designs D_1 in \mathbf{R}^m and D_2 in \mathbf{R}^p define the design $D_1 \cup D_2$ as

$$D_1 \cup D_2 = \{(d_1, a), (b, d_2) : d_1 \in D_1, d_2 \in D_2\}$$

where a and b are some fixed points in \mathbf{R}^p and \mathbf{R}^m respectively. The model given by the union of $\text{Est}(D_1)$ and $\text{Est}(D_2)$ (the intercept is counted only once) is identifiable by $D_1 \cup D_2$ because for $a = 0, b = 0$ a set of interpolating polynomials, or more precisely a Gröbner basis with respect to a compatible ordering, is given by

$$\{g, h - f, x_i y_j : g \in G_1, h \in G_2, i = 1, \dots, m \text{ and } j = 1, \dots, p\}$$

where the x_i 's run over \mathbf{R}^m and the y_j 's over \mathbf{R}^p and f are some suitable polynomials of the x 's. For generic a and b invariance of shifted designs applies. Usually $\text{Est}(D_1) \cup \text{Est}(D_2) \subset \text{Est}(D_1 \cup D_2)$. For example for $D_1 = \{1, 2, 3\}$, $D_2 = \{1, 2\}$ and $a = b = 0$ we have $\text{Est}(D_1) = \{1, x_1, x_1^2\}$ and $\text{Est}(D_2) = \{1, y_1\}$ and $\text{Est}(D_1 \cup D_2) = \{1, x_1, x_1^2, y_1, y_1^2\}$. The above statements can also be understood by checking the appropriate X -matrices.

The relationship between term-ordering blocking, described above, and the classical structures makes itself felt in the analysis of variance, via special orthogonal polynomials associated with the term-ordering. We simply construct orthogonal polynomials in the usual way considering terms in the ordering determined by the term-ordering.⁶ Finally we refer to Fontana, Pistone and Rogantin (1999) for the analysis of two-level factorial fraction designs.⁵

5. Applications

5.1. Favoured factors

We present now an example where certain factors are favoured. The design is a subset of a much larger design for an engine emission experiment. In the considered design there are eight factors and 105 distinct data points. Two factors are speed and load and the other six factors are engine design factors. For twelve combination of speed/load there are 26 combinations of the design factors. In general in engine mapping, speed and load can be favoured over design factors. This is achieved by choosing a term-ordering where engine speed and load are in a smaller block and design factors in a higher block. Within a block a `tddeg` term-ordering applies. This choice of term-ordering will make higher order terms in speed-load favoured over those for the design factors. The design was given as a matrix whose first column corresponds to speed, the second to load and so on. The matrix for the term-ordering is then

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Notice that the “ordering equations” of (2) are

$$\begin{aligned}
 & \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \alpha_7 + \alpha_8 \\
 & -\alpha_8 \\
 & -\alpha_7 \\
 & -\alpha_6 \\
 & -\alpha_5 \\
 & -\alpha_4 \\
 & \alpha_1 + \alpha_2 \\
 & -\alpha_2
 \end{aligned}$$

for a generic monomial x^α , $\alpha = (\alpha_1, \dots, \alpha_8)$. At first glance this term-ordering does not look like the block term-ordering introduced in Section 3. But clearly it is possible to rearrange the factors and obtain a block diagonal matrix of the type in Section 3 and still corresponding to the above set of ordering equations. Specifically this is achieved by moving to the end the first two columns of A and by considering the rearranged vectors $(\alpha_3, \dots, \alpha_8, \alpha_1, \alpha_2)$ for the exponents of monomials.

The Gröbner basis computation was performed using the computer algebra package CoCoA² and it took less than 35.81s of cpu time on a SiliconGraphics machine (see <http://euridice.tue.nl>). Tables 1 and 2 give the set of leading terms of the Gröbner basis and the set of identifiable terms respectively. Notice that the linear and the constant terms are all identifiable. Also all the second order interactions involving x_1 and x_2 are identifiable, except x_1^2 .

Higher powers of x_2 are favoured over powers of x_1 . This follows from the fact that amongst the leading terms there is a lower power of x_1 , namely x_1^2 and a higher power of x_2 . The initial ordering corresponding to the term-ordering given by A above is $x_3 \succ \dots \succ x_8 \succ x_1 \succ x_2$. By swapping the first and the second column in the ordering matrix A for the ordering above the initial ordering on the factors changes to $x_3 \succ \dots \succ x_8 \succ x_2 \succ x_1$. In this case the factor x_1 would have been favoured over x_2 and higher powers of x_1 would appear in Est instead of powers of x_2 .

The set Est can be used as a starting set of terms on which to operate a backward stepwise regression or similar methods in order to fit a model to the data set. Above we have suggested that by changing the ordering different Est sets can be obtained. In Section 5.2 this idea is developed leading to the notion of fan of a design. Sometimes it is useful to orthogonalise the obtained Est set of identifiable terms with respect to the term-ordering used to compute Est itself.⁶ In our example the terms involving x_1 and x_2 only would be orthogonalised first as they are smaller in the chosen ordering. For completeness Table 3 shows the set Est ordered with respect to the ordering given by A , smaller terms come last.

5.2. Maximal fan designs

The Gröbner basis method provides a background theory for constructing designs which maximise the number of different identifiable models, a type of robustness.

Table 1. Set of leading terms for the engine emission experiment

x_1^2	$x_1x_2^6$	x_2^8	$x_2^4x_8$	$x_2^4x_7$	$x_2^4x_6$	$x_1x_2^5x_5$	$x_2^4x_5$	$x_1x_2^4x_4$
$x_2^6x_4$	$x_2^6x_3$	$x_1x_2^5x_3$	$x_1x_2^4x_8^2$	$x_2^6x_8^2$	$x_2^5x_7x_8$	$x_1x_2^4x_7x_8$	x_6x_8	x_5x_8
x_4x_8	x_3x_8	x_7^2	x_6x_7	x_5x_7	x_4x_7	x_3x_7	x_6^2	x_5x_6
x_4x_6	x_3x_6	x_5^2	x_4x_5	x_3x_5	x_4^2	x_3x_4	x_3^2	x_8^3
								$x_7x_8^2$

Table 2. Set of identifiable terms for the engine emission experiments

1								
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	
x_1x_2	x_1x_3	x_1x_4	x_1x_5	x_1x_6	x_1x_7	x_1x_8	x_2^2	
x_2x_3	x_2x_4	x_2x_5	x_2x_6	x_2x_7	x_2x_8	x_7x_8	x_8^2	
$x_1x_2^2$	$x_1x_2x_3$	$x_1x_2x_4$	$x_1x_2x_5$	$x_1x_2x_6$	$x_1x_2x_7$	$x_1x_2x_8$	$x_1x_7x_8$	$x_1x_8^2$
x_2^3	$x_2^2x_3$	$x_2^2x_4$	$x_2^2x_5$	$x_2^2x_6$	$x_2^2x_7$	$x_2^2x_8$	$x_2x_7x_8$	$x_2x_8^2$
$x_1x_2^3$	$x_1x_2^2x_3$	$x_1x_2^2x_4$	$x_1x_2^2x_5$	$x_1x_2^2x_6$	$x_1x_2^2x_7$	$x_1x_2^2x_8$	$x_1x_2x_7x_8$	$x_1x_2x_8^2$
x_2^4	$x_2^3x_3$	$x_2^3x_4$	$x_2^3x_5$	$x_2^3x_6$	$x_2^3x_7$	$x_2^3x_8$	$x_2^2x_7x_8$	$x_2^2x_8^2$
$x_1x_2^4$	$x_1x_2^3x_3$	$x_1x_2^3x_4$	$x_1x_2^3x_5$	$x_1x_2^3x_6$	$x_1x_2^3x_7$	$x_1x_2^3x_8$	$x_1x_2^2x_7x_8$	$x_1x_2^2x_8^2$
x_2^5	$x_2^4x_3$	$x_2^4x_4$	$x_2^4x_5$	$x_2^4x_6$	$x_2^4x_7$	$x_2^4x_8$	$x_2^3x_7x_8$	$x_2^3x_8^2$
$x_1x_2^5$	$x_1x_2^4x_3$	$x_1x_2^4x_4$	$x_1x_2^4x_5$	$x_1x_2^4x_6$	$x_1x_2^4x_7$	$x_1x_2^4x_8$	$x_1x_2^3x_7x_8$	$x_1x_2^3x_8^2$
$x_2^5x_3$	$x_2^5x_4$	$x_2^5x_5$	$x_2^5x_6$	$x_2^5x_7$	$x_2^5x_8$	$x_2^4x_7x_8$	$x_2^4x_8^2$	x_2^4
$x_1x_2^5x_6$	$x_1x_2^5x_7$	$x_1x_2^5x_8$	x_2^6	$x_2^5x_5$	$x_2^6x_6$	$x_2^5x_7$	$x_2^6x_8$	$x_2^5x_8^2$

Table 3. Set of identifiable terms ordered according to A

$x_1x_2^3x_7x_8$	$x_2^4x_7x_8$	$x_1x_2^2x_7x_8$	$x_2^3x_7x_8$	$x_1x_2x_7x_8$	$x_2^2x_7x_8$	$x_1x_7x_8$	$x_2x_7x_8$	x_7x_8
$x_2^5x_8^2$	$x_1x_2^3x_8^2$	$x_2^4x_8^2$	$x_1x_2^2x_8^2$	$x_2^3x_8^2$	$x_1x_2x_8^2$	$x_2^2x_8^2$	$x_1x_8^2$	$x_2x_8^2$
x_8^2	$x_1x_2^4x_3$	$x_2^5x_3$	$x_1x_2^3x_3$	$x_2^4x_3$	$x_1x_2^2x_3$	$x_2^3x_3$	$x_1x_2x_3$	$x_2^2x_3$
x_1x_3	x_2x_3	x_3	$x_2^2x_4$	$x_1x_2^3x_4$	$x_2^4x_4$	$x_1x_2^2x_4$	$x_2^3x_4$	$x_1x_2x_4$
$x_2^2x_4$	x_1x_4	x_2x_4	x_4	$x_2^2x_5$	x_1x_5	x_2x_5	x_5	$x_1x_2^3x_5$
$x_1x_2^2x_5$	$x_2^3x_5$	$x_1x_2x_5$	$x_2^2x_5$	x_1x_5	x_2x_5	x_5	$x_1x_2^2x_6$	$x_2^2x_6$
$x_1x_2^4x_6$	$x_2^5x_6$	$x_1x_2^3x_6$	$x_2^4x_6$	$x_1x_2^2x_6$	$x_2^3x_6$	$x_1x_2x_6$	$x_2^2x_6$	x_1x_6
x_2x_6	x_6	$x_1x_2^2x_7$	$x_2^3x_7$	$x_1x_2^2x_7$	$x_2^2x_7$	$x_1x_2x_7$	$x_2^2x_7$	$x_1x_2^2x_7$
$x_2^3x_7$	$x_1x_2x_7$	$x_2^2x_7$	x_1x_7	x_2x_7	x_7	$x_1x_2^2x_8$	$x_2^2x_8$	$x_1x_2^2x_8$
$x_2^5x_8$	$x_1x_2^3x_8$	$x_2^4x_8$	$x_1x_2^2x_8$	$x_2^3x_8$	$x_1x_2x_8$	$x_2^2x_8$	x_1x_8	x_2x_8
x_8	x_7	$x_1x_2^5$	x_2^6	$x_1x_2^4$	x_2^5	$x_1x_2^3$	x_2^4	$x_1x_2^2$
x_2^3	x_1x_2	x_2^2	x_1	x_2	1			

Such designs spread points in a space-filling manner and are suitable for computer and other highly complex experiments, where little is known about the model in advance.

Given a design D on \mathbf{R}^m one could vary all the possible term-orderings and determine all the Est-type models identifiable by D . The set so obtained is called the algebraic fan and is a subset of all the models identifiable by D , with an order ideal structure and with exactly as many terms as there are design points. This last set is called the statistical fan. A classical result in algebraic geometry tells us that the algebraic fan is finite, despite the fact that there are infinite term-orderings.¹⁵ Moreover the fan defines an equivalence relation over the set of term-orderings, that is two term-orderings τ_1 and τ_2 belong to the same class of equivalence with respect to a design D if they give rise to the same identifiable set. The identifiable sets are called leaves of the fan.

Designs may be classified according to how many leaves are in their fans. Thus we talk of minimal fan designs for designs whose statistical fan has only one leaf. Examples are full factorial designs and echelon designs. The term maximal fan designs refers to those designs whose fan (both algebraic and statistical) includes all the models with an order ideal structure and with exactly as many terms as there are design points. There always exists a design whose statistical fan is maximal.¹ The algorithms to compute directly the algebraic fan of a design, and known to the authors, are computationally too intensive to actually be useful in large complex designs. Nevertheless it has been noted that a random design is maximal fan with probability one.¹ Work on maximal fan designs is in progress.

5.3. *Mixture experiments*

Understanding confounding and orthogonality in mixture experiments remains a problematic area in the design of experiments. Mixture experiments present a particularly interesting class of problems for application of the Gröbner basis method. This is because the simplex condition $\sum_{i=1}^m x_i = 1$ (sum of proportions is one) imposes an additional algebraic condition and the polynomial $\sum_{i=1}^m x_i - 1$ belongs to the design ideal. By applying directly the Gröbner basis method at least one of the factors will never be in the identifiable set because of the simplex condition. The polynomial models which arise as models under this new condition are not fully determined in that the condition allows transformation between different representations.

As an example let us consider the following three factor 6-point design

$$(1, 0, 0) \quad (0, 1, 0) \quad (0, 0, 1) \quad \left(\frac{1}{2}, \frac{1}{2}, 0\right) \quad \left(\frac{1}{2}, 0, \frac{1}{2}\right) \quad \left(0, \frac{1}{2}, \frac{1}{2}\right)$$

With respect to `tdeg` and the initial ordering $x_1 \succ x_2 \succ x_3$ a Gröbner basis is

$$\begin{aligned} & x_1 + x_2 + x_3 - 1 \\ & x_3^3 - 3/2x_3^2 + 1/2x_3 \\ & x_2x_3^2 - 1/2x_2x_3 \\ & x_2^2x_3 - 1/2x_2x_3 \\ & x_2^3 - 3/2x_2^2 + 1/2x_2 \end{aligned}$$

Notice that the mixture condition $x_1 + x_2 + x_3 - 1$ comes directly as a G-basis element, and shows that this holds on the design points. For some design/term-ordering pair, namely when more than one linear term does not appear in Est, the simplex condition is not an element of the (reduced) Gröbner basis but still it belongs to the design ideal.

The leading terms for the previous example are

$$x_1 \quad x_3^3 \quad x_2x_3^2 \quad x_2^2x_3 \quad x_2^3$$

giving the model

$$1 \quad x_3 \quad x_2 \quad x_3^2 \quad x_2x_3 \quad x_2^2$$

This is equivalent to the full 6-parameter homogeneous quadratic model. By the symmetry of the design we deduce that the models

$$1 \quad x_3 \quad x_1 \quad x_3^2 \quad x_1x_3 \quad x_1^2$$

and

$$1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_2x_1 \quad x_2^2$$

are also identifiable. Actually, they come from other initial orderings. This can be formalised by a homogenisation process. Here homogenisation means that an extra indeterminate is added and, elevated to a suitable power, is multiplied to each term in Est so that the total degree (sum of exponents) of each term is the same. For example the homogenisation of the first model with respect to x_1 gives

$$x_1^2 \quad x_1x_3 \quad x_1x_2 \quad x_3^2 \quad x_2x_3 \quad x_2^2$$

and dehomogenising with respect to x_2 we obtain the second model. Notice that also the Scheffé polynomial model

$$x_1 \quad x_2 \quad x_3 \quad x_1x_2 \quad x_1x_3 \quad x_2x_3$$

is identifiable as the design matrix is invertible. But the Scheffé model will never be recovered directly as an Est set because it has not an order ideal structure. For example it does not include the constant term 1. Nevertheless it can be retrieved by exploiting the aliasing conditions defined by the Gröbner basis equations.¹² Note that the four models above, including the Scheffé model, are vector space bases of the quotient space $\mathbf{R}[x_1, x_2, x_3]/\text{Ideal}(D)$.

the first column of the matrix B is multiplied by x_1 , the second column by x_2 and so on according to (2).

It is advisable to use statistical techniques to help in the choice of the initial term-ordering and the term-ordering in general. For example, a prior screening on the initial factors can suggest an initial ordering. In the example above such an analysis suggested that x_1 and x_4 were the less influential factors. This, coupled with some knowledge of the specific case study, makes Gröbner bases a powerful tool in the analysis of non standard experiments.

Table 4. A 9-factor 26-point mixture design

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
0.2106	0.0000	0.1098	0.4119	0.0183	0.0366	0.0915	0.0366	0.0848
0.2014	0.0000	0.1050	0.3940	0.0175	0.0350	0.0875	0.0350	0.1247
0.2565	0.1062	0.1062	0.1681	0.0177	0.1062	0.0885	0.0354	0.1152
0.2553	0.1057	0.1057	0.1673	0.0176	0.1057	0.0881	0.0352	0.1194
0.1953	0.1020	0.1020	0.2124	0.0170	0.1020	0.0170	0.1020	0.1502
0.1950	0.1017	0.1017	0.2119	0.0170	0.1017	0.0170	0.1017	0.1523
0.2663	0.1103	0.0368	0.2480	0.0184	0.0368	0.0919	0.1103	0.0811
0.2635	0.1090	0.0363	0.2453	0.0182	0.0363	0.0909	0.1090	0.0915
0.2501	0.0000	0.1072	0.1430	0.0894	0.1073	0.0894	0.1073	0.1062
0.2710	0.0000	0.1161	0.1549	0.0968	0.1161	0.0968	0.1161	0.0322
0.2647	0.0000	0.0365	0.3559	0.0182	0.1095	0.0182	0.1095	0.0876
0.2586	0.0000	0.1069	0.2763	0.0891	0.0356	0.0891	0.0356	0.1089
0.1856	0.0967	0.0322	0.1371	0.0806	0.0967	0.0806	0.0967	0.1939
0.2239	0.0000	0.1075	0.2598	0.0179	0.1075	0.0896	0.0896	0.1042
0.2184	0.0000	0.1049	0.3232	0.0874	0.0350	0.0175	0.0874	0.1261
0.2791	0.0000	0.0385	0.4522	0.0192	0.1154	0.0192	0.0385	0.0380
0.2779	0.0000	0.0383	0.4504	0.0192	0.1150	0.0192	0.0383	0.0417
0.2297	0.0951	0.0951	0.1506	0.0793	0.0951	0.0159	0.0317	0.2075
0.2092	0.0000	0.0364	0.4092	0.0909	0.1091	0.0182	0.0364	0.0906
0.2616	0.1082	0.0361	0.2435	0.0902	0.0361	0.0180	0.1082	0.0981
0.2615	0.1081	0.0360	0.2434	0.0901	0.0360	0.0180	0.1081	0.0989
0.1862	0.0972	0.0324	0.3887	0.0162	0.0405	0.0162	0.0324	0.1901
0.1968	0.1028	0.0343	0.2140	0.0856	0.1028	0.0856	0.0343	0.1437
0.2583	0.1070	0.1070	0.2406	0.0891	0.0357	0.0178	0.0357	0.1087
0.2262	0.0522	0.0696	0.2784	0.0522	0.0696	0.0522	0.0696	0.1300
0.2362	0.0545	0.0726	0.2906	0.0545	0.0726	0.0545	0.0726	0.092

References

1. M. Caboara, G. Pistone, E. Riccomagno and H. P. Wynn, “The fan of an experimental design”, *Eurandom Preprint n. 99-038* (1999).
2. A. Capani, G. Niesi and L. Robbiano, CoCoA, *a system for doing Computation in Commutative Algebra*. Available via anonymous ftp from `cocoa.dima.unige.it` (1998).
3. D. Cox, J. Little and D. O’Shea, *Ideal, Varieties and Algorithms* (Springer-Verlag, New York 1997, second edition).
4. B. Char, K. Geddes, G. Gonnet, B. Leong and M. Monogan, *MAPLE V Library Reference Manual* (Springer-Verlag, New York, 1991).
5. R. Fontana, G. Pistone, and M.-P. Rogantin, “Classification of two-level factorial fractions” *JSPI 1999* (to appear).
6. B. Giglio, E. Riccomagno and H. P. Wynn, “Gröbner basis strategies in regression” *Applied Statistics, 2000* (to appear).
7. T. Holliday, A. J. Lawrance and T. P. Davis, *Technometrics* **40**,2:120-126 (1998).
8. T. Holliday, G. Pistone, E. Riccomagno and H. P. Wynn, *Computational Statistics* **14**-2:213-231 (1999).
9. P. McCullagh and J. A. Nelder, *Generalised linear models* (Chapman and Hall, London, 1989, second edition).
10. G. Pistone, E. Riccomagno and H. P. Wynn, *Algebraic Statistics* (in progress).
11. G. Pistone and H. P. Wynn, *Biometrika* **83**,3:653–666 (1996).
12. E. Riccomagno, Ph.D. Thesis, Department Of Statistics, University of Warwick, 1997.
13. E. Riccomagno and H. P. Wynn, “G-Basis, order ideals and a generalised divided difference formula” (Submitted, 1999).
14. L. Robbiano, *EUROCAL’85, LNCSeries* **204**:513-517 (1985).
15. B. Sturmfels, *Gröbner Bases and Convex Polytopes* (AMS, University Lecture Series, Providence R I, 1995).
16. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-PLUS* (Springer, New York, 1994 second edition).