Betti numbers, models and experimental design

Bruyhooghe, David
London School of Economics, Statistics Department
Houghton Street
London WC2A 2AE,UK
E-mail: d.hawelleh@lse.ac.uk

Maruri-Aguilar, Hugo
Queen Mary, University of London, Mathematics Department
Mile End Road
London E1 4NS, UK
E-mail: H.Maruri-Aguillar@qmul.ac.uk

Sáenz de Cabezón, Eduardo
Department of Mathematics
26005-Logronó, Spain
E-mail: eduardo.saenz-de-cabezon@unirioja.es

Wynn, Henry
London School of Economics, Statistics Department
Houghton Street
London WC2A 2AE,UK
E-mail: h.wynn@lse.ac.uk

1 Introduction

This paper is a description of a research programme in main branch of what has become to be known "algebraic statistics". This is the branch which considers experimental designs and the models that can be fitted using a design, [6],[7]. The research programme is to look closely at the models from an algebraic point of view. At its simplest this is a theory of interactions. But not just the interactions in the model but also those excluded from the model. Some of these are interactions which, in a sense, had no hope of being estimated, because the sample size is two small. In classical factorial design one is familiar with looking at the full alias table and discussing which main effects are in the model and which not. The work can be seen as an extension of such discussion.

At the centre of the theory is the idea of a hierarchical model, one in which if a higher order interactions appears, then so do the lower order interactions: eg if $x_1x_2x_3$ appears then the constant terms (1), the main effects x_1, x_2, x_3 and two way interactions x_1x_2, x_1x_3, x_2x_3. Notice that we can code up this scheme as a triangle with x_1, x_2, x_3 as vertices, x_1x_2, x_1x_3, x_2x_3 as the edges and the face $x_1x_2x_3$. This generalises easily to the idea of a hierarchical model being coded as a simplicial complex in the "square free", ie multilinear case.

2 The algebraic method in experimental design

The algebraic method is to consider a design to be the zero set of a set of polynomials and regression models in d variables are considered to be members of the polynomial ring $\mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_d]$. For finite set of polynomials $f_1, \ldots, f_m$, the set $\{\sum_{i=1}^m f_i s_i : s_i \in \mathbb{R}[x]\} = \langle f_1, \ldots, f_m \rangle$ is an ideal. The ideal generated by a finite set of points representing the design, $\mathcal{D} \subset \mathbb{R}^d$ is

$I(\mathcal{D}) = \{f \in \mathbb{R}[x] : f(x) = 0, x \in \mathcal{D}\} \subset \mathbb{R}[x]$. For a set of polynomials $f_1, \ldots, f_m \in \mathbb{R}[x]$, the set r the subset of $\mathbb{R}[x]$ ideal $I = \langle f_1, \ldots, f_m \rangle$ A term order τ is a total ordering in monomials in $T^d = \{x^\alpha : \alpha \in \mathbb{Z}^d_{\geq 0}\}$, compatible with monomial simplification: $1 \prec x^\alpha, \alpha \neq 0$, $x^\alpha \prec x^\beta \Rightarrow x^{\alpha+\gamma} \prec x^{\beta+\gamma}$ for $x^\alpha, x^\beta, x^\gamma \in T^d$. Given a term order $\prec$, the leading term of the polynomial g, $\mathrm{LT}(g)$, is the highest term of g (under $\prec$) with non-zero coefficient. A Gröbner basis G_τ for $I(\mathcal{D})$ is a finite subset of $I(\mathcal{D})$ such that $\langle \mathrm{LT}(g) : g \in G_\tau \rangle = \langle \mathrm{LT}(f) : f \in I(\mathcal{D}) \rangle$. The leading term of g is $\mathrm{LT}(g)$. See the basic text [5].

The elements of $\mathbb{R}[\mathcal{D}]$ are in one to one correspondence with equivalence classes of polynomials modulo $I(\mathcal{D})$.As vector spaces and isomorphisms hold

$$\mathbb{R}[\mathcal{D}] \sim \mathbb{R}[x]/I(\mathcal{D}) \sim \mathbb{R}[x]/\langle LT(I(\mathcal{D}))\rangle \quad (1)$$

A basis (model) for $\mathbb{R}[x]/I(\mathcal{D})$ is given by those monomials that cannot be divided by any of $\mathrm{LT}(g)$ for $g \in G_\tau$ (Gröbner basis).

To summarize

1. For any $f \in \mathbb{R}[x]$, unique remainder r in division of f by $I(\mathcal{D})$

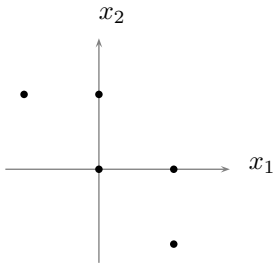
   $$f = \sum_{g \in G_\tau} gh + r \quad (2)$$

2. The remainder in (2) is known as the normal form of f (modulo $I(\mathcal{D})$), i.e. $\mathrm{NF}(f) = r$.

3. A model can be described through the (complementary) monomial ideal associated with it.

4. The computations depend on the term order selected τ. By varying τ over all term orders, the collection of models is called a design fan, see [1] and [5].

As an example take $D = \{(0,0), (1,0), (0,1), (-1,1), (1,-1)\}$, given below



If we take the term order with, term order $x_2 \prec x_1$, the G-Basis is $G_\prec = \{\underline{x_1^2} + 2x_1x_2 + x_2^2 - x_1 - x_2, \underline{x_2^3} - x_2, \underline{x_1x_2^2} - x_1x_2 - x_2^2 + x_2\}$ and the model basis is $\{1, x_1, x_2, x_1x_2, \underline{x_2^2}\}$. The underlined terms are the leading terms.

3 Hilbert Series

A degree-by-degree description of the monomials that generate $\mathbb{R}[x]/I$ is given by the Hilbert series of $\mathbb{R}[x]/I$:

$$HS(s) = \sum_{t=0}^{\infty} H(t)s^t,$$

where $H(t) = \dim \mathbb{R}[x]_t/I_t$ is the Hilbert function. When $I = I(\mathcal{D})$ then $HS(s)$ encodes degree-by-degree information of the model (L). The Hilbert Series can be written alternatively as

$$HS(s) = \sum_{\alpha \in L} s^{|\alpha|},$$

with $|\alpha|$ the sum of elements in α. The Hilbert function can be retrieved from the Hilbert Series

$$H(t) = \lim_{s\to 0} \frac{1}{t!}\frac{\partial^t HS(s)}{\partial s^t}$$

A finer description of the monomials generating $\mathbb{R}[x]/I$ is given by the multigraded Hilbert series. Let W be a matrix of d rows with integer entries (W need not be square, nor full rank). The multigraded Hilbert Series of the quotient ring $\mathbb{R}[x]/I$ is defined as

$$\text{HS}_W(x) = \sum_{\alpha\in L} x^{W^T\alpha}, \quad (3)$$

where L is the set of monomials which do not lie in $\langle LT(I)\rangle$. The standard Hilbert series is easily retrieved by using $W^T = (1,\ldots,1)$. Setting W as identity (d) then $\text{HS}_W(x) = \sum_{\alpha\in L} x^\alpha$, and the partition $\mathbb{Z}^d = L \cup (\mathbb{Z}^d \setminus L)$ is reflected in the following relation:

$$\begin{aligned} \frac{1}{\prod_{i=1}^d (1-x_i)} &= \textstyle\sum_{\alpha\in L} x^\alpha + \left(\frac{1}{\prod_{i=1}^d (1-x_i)} - \sum_{\alpha\in L} x^\alpha\right) \\ \\ HS(\mathbb{R}[x]/\langle 0\rangle) &= HS(\mathbb{R}[x]/I) + HS(I) \end{aligned}$$

As an example consider the following monomial ideal $I = \langle de, abe, ace, abcd\rangle$ The Hilbert Series of $\mathbb{R}[a,b,c,d,e]/I$ is

$$HS(s) = \frac{1+2s+2s^2}{(1-s)^3},$$

with Hilbert function $H(t) = \frac{5}{2}t^2 + \frac{3}{2}t + 1$ for $t \geq 0$, while the multigraded Hilbert Series of $\mathbb{R}[a,b,c,d,e]/I$ is

$$HS = \frac{1 - de - abe - ace + abce + abde + acde - abcd}{(1-a)(1-b)(1-c)(1-d)(1-e)}$$

4 The Stanley-Reisner ideal

A simplicial complex has a one to one relation with a special type of monomial ideal, called the Stanley-Reisner ideal. For a simplicial complex Δ, let I_Δ be the squarefree monomial ideal created by the non-faces of Δ: $I_\Delta = \langle x^a : a \notin \Delta\rangle$. The complexity of the model Δ can be studied by the Stanley-Reisner ring $R[x]/I_\Delta$. If Δ is a simplicial complex, then $\dim R[x]/I_\Delta = \dim \Delta + 1$.

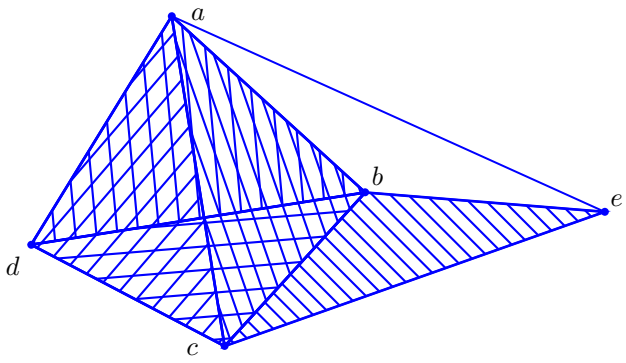
In the description of $R[x]/I_\Delta$, the Betti numbers play a central role. Graded Betti number is the minimal number of generators $e_{a,j}$ in degree $i+j$. The ideal of leading terms $\langle \text{LT}(I(\mathcal{D}))\rangle$ is related to the Stanley-Reisner ideal:

$$\langle \text{LT}(I(\mathcal{D}))\rangle = \bar{I}_\Delta,$$

where $\bar{I}_\Delta$ is the Artinian closure of I_Δ (see the example below for an explanation of this term).

As an example, consider the following simplicial complex model, below: $\Delta = \{1,\ a, b,\ c, d,\ e, ab, ac,\ ad, ae,\ bc, bd, be,\ cd, ce, abc,\ abd,\ adc,\ bcd, bce\}$ and the Stanley-Reisner ideal generated by it (see [8]):

$$I_\Delta = \langle de, abe, ace, abcd\rangle \subset \mathbb{R}[a,b,c,d,e]$$



We carry out the computation of the Betti numbers using the CoCoa package

```
Use T::=Q[a,b,c,d,e];
J:=Ideal(de,abe,ace,abcd);
Hilbert(T/J);
HilbertSeries(T/J);

H(t) = 5/2t^2 + 3/2t + 1   for t >= 0
-------------------------------
(1 + 2a + 2a^2) / (1-a)^3
-------------------------------

BettiDiagram(T/J);

        0    1    2    3
-------------------------
 0:     1    -    -    -
 1:     -    1    -    -
 2:     -    2    3    1
 3:     -    1    1    -
-------------------------
```

The Hilbert series, after simplification, is given by

$$HS = \frac{1 - de - abe - ace - abcd + abde + acde + abce}{(1-a)(1-b)(1-c)(1-d)(1-e)}$$

```
BettiDiagram(J);

        0    1    2
--------------------
 2:     1    -    -
 3:     2    3    1
```

```
 4:     1    1    -
--------------------
Tot:    4    4    1
---------------------
```

Alternatively we could use the Artinian closure of I_Δ which is given adjoining monomial power terms on the axes, giving

$$\bar{I}_\Delta = \langle de, abe, ace, abcd \rangle + \langle a^2, b^2, c^2, d^2, e^2 \rangle$$

```
K:=J+Ideal(a^2,b^2,c^2,d^2,e^2);
Hilbert(T/K);
HilbertSeries(T/K);
------------------------------
H(0) = 1
H(1) = 5
H(2) = 9
H(3) = 5
H(t) = 0   for t >= 4
------------------------------
(1 + 5a + 9a^2 + 5a^3)
------------------------------
```

In our working $K = J + Ideal(a^2, b^2, c^2, d^2, e^2)$.

```
K;
BettiDiagram(T/K);

Ideal(de, abe, ace, abcd, a^2, b^2, c^2, d^2, e^2)
------------------------------
        0    1    2    3    4
------------------------------
 2:     6    2    -    -    -
 3:     2   21   21    7    1
 4:     1    5   17   17    5
------------------------------
Tot:    9   28   38   24    6
------------------------------
```

Essentially this yields a degree-by-degree description of the model border, that is the "region" monomials on the frontier between terms in the model and terms not in the model.

5 Hierarchical models

We have seen above how to associate an important monomial ideals with models. In one case this is the leading term ideal associated with the G-basis of a design, and then in the special square-free we have the Stanley Reisner ideal, which leads to the leading term ideal with Artinian closure. The methodology works largely because the models are hierarchical: x^α in the model implies that x^β is in the model for any $\beta \leq \alpha$. In the square-free case with the model is described by a simplicial complex Δ.

The Betti numbers, or equivalently the multigraded Hilbert Series, provide a complete description of the model border terms. There is a theory of maximal Betti numbers and they are, roughly speaking, related to models of low degree. Some progress is made in [5] relating the theory to that on "linear aberration" studied in [1]. In the squarefree case (designs which are fractions of 2^d, whether regular or not), the important Hochster formula (see [8]) relates graded Betti numbers of the Stanley-Reisner ideal I_Δ with the reduced simplicial homology of the model itself S. What remains, whether for the leading terms ideal, I_Δ or Δ, is an interpretation of the Betti numbers, in terms of model complexity. Is there an interpretation for this "model homology" by analogy with the fast-growing area of persistent homology? Such an interpretation could make a valuable contribution to a general discussion of the interpretation of interactions.

Hierarchical models have a rich theory in graphical models and in the author's show how monomial ideals can be associated with very general statistical hierarchical models: [3]. As a simple example from binary log-linear multinomial models the model, with obvious notation

$$p(x_1, x_2, x_3) = C \exp(\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_1 x_3 + \theta_5 x_2 x_3)$$

has the simplicial complex Δ with maximal cliques $\{13, 23\}$ and $I_\Delta =< x_1 x_2 >$. This corresponds to the simple conditional independence model: $X_1 \coprod X_2 | X_3$. It is possible to map the structure of any hierarchical model defined by maximal cliques into monomial ideals. In the general case we would have for this example and a probability density $f(x_1, x_2, x_3)$:

$$f(x_1, x_2, x_3) = C \exp(h(x_1, x_3) + h(x_2, x_3)).$$

Although so straightforward it is possible to map such cases also into monomial ideals.

But the reverse direction shows promise too. That is, we can start with a class of monomial ideals studied in algebra and see what type of model it yields. An example is 2-linear ideals, in the square free case. In that case I_Δ is generated by simple interaction terms $x_i x_j$ not in the model, Δ and with an additional technical term "linear". It turns out that decomposable hierarchical models are 2-linear. Providing translation between the algebra and statistical models in this way is at the heart of the research programme, we have sketched.

REFERENCES (RÉFERENCES)

1. Berstein, Y., Maruri-Aguilar, H., Onn, S., Riccomagno, E., Wynn, H. (2008). Minimal average degree aberration and the state polytope for experimental design. *Ann. Inst. Stat. Math.***83**, 673-698
2. Cox, D. Little, J. and O'Shea, D (2077). *Ideals, varieties and algorithms.* Springer.
3. Bruynooghe, D and Wynn, H.P. (2011). Differential cumulants, hierarchical models and monomials ideals. arXiv:1102.2118v1
4. Maruri-Aguillar (2007). Ph.D. Thesis, University of Warwick. .
5. Maruri-Aguillar, Sáenz de Cabezön, E. and Wynn, H.P. (2011). The Betti numbers of polynomial hierarchicaal models for experimental design. Ann. Math. Art. Intel. (submitted)
6. Pistone, G and and Wynn, H.P. (1996). Generalised confounding with Gröbner bases. *Biometrika* **83**(3), 653-666.
7. Pistone, G Riccomagno, E and Wynn, H.P. (2000). *Algebraic Statistics.* CRC.
8. Miller, E., Sturmfels, B. (2005). *Combinatorial Commutative Algebra.* Springer.