Holes and aberration

Hugo Maruri-Aguilar and Henry Wynn
Queen Mary, University of London and London School of Economics
H.Maruri-Aguilar@qmul.ac.uk and H.Wynn@lse.ac.uk

Abstract

Algebraic techniques have been used to identify polynomial models in experimental design [6]. A feature of models identified with such techniques is that they are of low weighted degree. This lead to a generalisation of the idea of aberration, i.e. algebraic models are of low aberration. when designs are fractions of 2^k factorial designs, we can further describe models from an elementary topological viewpoint. That is, models are simplicial complexes and the quantity and quality of holes of such complexes is described by Betti numbers. In summary, our poster presents two forms of describing models: by degree and by connectivity. Some examples are included.

Design and design ideal

An experimental design $\mathcal{D}$ is a set of n points in d factors, $\mathcal{D} \subset \mathbb{R}^d$. For indeterminates $x_1, \ldots, x_d$ consider a subset of the polynomial ring $\mathbb{R}[x_1, \ldots, x_d]$ composed by all polynomials that vanish on the design points. This set is a polynomial ideal which is called the *design ideal* $I(\mathcal{D})$.

Identification of models with algebra

Two polynomials $f, g \in \mathbb{R}[x_1, \ldots, x_d]$ are congruent modulo $I(\mathcal{D})$ if $f - g \in I(\mathcal{D})$. The quotient ring $\mathbb{R}[x_1, \ldots, x_k]/I(\mathcal{D})$ is the set of equivalence classes for congruence modulo $I(\mathcal{D})$. When considered as real vector spaces, the following isomorphisms between quotient rings hold:

$$\mathbb{R}[x_1, \ldots, x_d]/I(\mathcal{D}) \cong \mathbb{R}[x_1, \ldots, x_d]/\langle \mathrm{LT}(I(\mathcal{D})) \rangle \cong \mathbb{R}[\mathcal{D}] \qquad (1)$$

Here $\langle \mathrm{LT}(I(\mathcal{D})) \rangle$ is the monomial ideal created by the leading terms of polynomials in $I(\mathcal{D})$, namely $\langle \mathrm{LT}(I(\mathcal{D})) \rangle = \langle \mathrm{LT}_{\prec}(f) : f \in I(\mathcal{D}) \rangle$; the symbol $\prec$ refers to a term ordering in T^d, the set of all monomials in $x_1, \ldots, x_d$; while $\mathbb{R}[\mathcal{D}]$ is the set of polynomial functions defined on $\mathcal{D}$, see [3].
Recall that the Gröbner basis for $I(\mathcal{D})$ is a subset $G \subset I(\mathcal{D})$ such that $\langle LT_{\prec}(g) : g \in G \rangle = \langle \mathrm{LT}(I(\mathcal{D})) \rangle$. For a fixed term ordering $\prec$, let G be a Gröbner basis for $I(\mathcal{D})$ and let $L_{\prec}(I(\mathcal{D}))$ be the set of all monomials that cannot be divided by the leading terms of the Gröbner basis G, that is

$$L_{\prec}(I(\mathcal{D})) := \{x^\alpha \in T^d : x^\alpha \text{ is not divisible by } \mathrm{LT}_{\prec}(g), g \in G\} \qquad (2)$$

The set $L(\mathcal{D}) = L_{\prec}(I(\mathcal{D}))$ is referred to as the model identified by the design. In statistical terms, $L(\mathcal{D})$ is a saturated, hierarchical model.
Example 1: For the design $\mathcal{D} = \{(0,0), (1,0), (0,1), (-1,1), (1,-1)\}$ and a monomial ordering $x_2 \prec x_1$, the set $G_{\prec} = \{x_1^2 + 2x_1x_2 + x_2^2 - x_1 - x_2, x_2^3 - x_2, x_1x_2^2 - x_1x_2 - x_2^2 + x_2\}$ is a Gröbner basis for $I(\mathcal{D})$ and the model identified is $L(\mathcal{D}) = \{1, x_1, x_2, x_1x_2, x_2^2\}$.

A column selection algorithm [1]

The identification of models with algebraic techniques is equivalent to column selection of the design model matrix using a term ordering for the candidate columns. By row elimination, the following algorithm retrieves a reduced Gröbner basis for $I(\mathcal{D})$. This algorithm is a variation of the FGLM algorithm for change of basis [4].

1. Compute the design-model matrix with design points and monomials from set $V_n^d := \{x \in \mathbb{Z}_{\geq 0}^d : \prod_{i=1}^d (x_i + 1) \leq n\}$. Columns are indexed by monomials from V_n^d.
2. Reorder the columns using a monomial ordering $\prec$.
3. Select the first n columns which form a linearly independent set

1	x_1	x_2	x_1^2	x_1x_2	x_2^2	x_1^3	$\cdots$
1	0	0	0	0	0	0	
1	1	0	1	0	0	1	
1	0	1	0	0	1	0	
1	1	−1	1	−1	1	1	
1	−1	1	1	−1	1	−1	

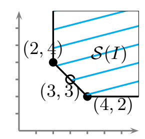
References

1. Babson, E., Onn, S., Thomas R. (2003). The Hilbert zonotope and a polynomial time algorithm for Universal Gröbner bases. *Adv. Appl. Math.* **30**(3), 529-544.
2. Bernstein, Y., Maruri-Aguilar, H., Onn, S., Riccomagno, E., Wynn, H. (2010). Minimal aberration and the state polytope for experimental designs. *AISM* (Accepted).
3. Cox, D., Little, J., O'Shea, D. (1996). *Ideals, varieties and algorithms.* Springer.
4. Faugère, J.C., Gianni, P., Lazard, D., Mora, T. (1993). Efficient computation of zero-dimensional Gröbner bases by change of ordering. *J. Symb. Comput.* **16**(4), 329-344.
5. Miller, E., Sturmfels, B. (2005). *Combinatorial Commutative Algebra.* Springer.
6. Pistone, G., Riccomagno, E., Wynn, H. P. (2001). *Algebraic Statistics*, Vol. 89 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, Boca Raton.
7. Plackett, R.L., Burman, J.P. (1946). The design of optimum multifactorial experiments. *Biometrika* 33(4), 305–325.

Design fan

The collection of models identified by $\mathcal{D}$ by changing all possible monomial orderings is called the algebraic fan on $\mathcal{D}$. The algebraic fan is a finite collection of models $A(\mathcal{D}) = \bigcup_{\prec} L_{\prec}(D)$, where the union is carried out over the infinite set of all term orderings $\prec$. In general this is a difficult computation for which there are some partial solutions. One alternative is to create a family of universal weighing ordering vectors using the *Hilbert zonotope*, see [1]. A different solution is software `Gfan` which computes $A(\mathcal{D})$ for a given design ideal. A third alternative, not fully explored, is a probabilistic search for models.

Models in $A(\mathcal{D})$ correspond to vertexes of the *state polytope* $\mathcal{S}(I)$, e.g. sum exponent vector in $L = \{1, x_1, x_2, x_1x_2, x_2^2\}$, $\bar{\alpha}_L = \sum_L \alpha = (2,4)$.
The state polytope is defined as $\mathcal{S}(I) = \mathrm{conv}(\bar{\alpha}_L : L \in A(\mathcal{D})) + \mathbb{R}_+^d$



Example 2: For $d = 3, \alpha = 2, n_0 = 1$, set $\mathcal{D}$ to be the central composite design (CCD) with points $\{(0,0,0), (\pm1, \pm1, \pm1), \pm e_i, i = 1,2,3\}$. There are three models identified by algebra: $1, x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3$ plus terms $x_i^3, x_i^4, x_i^2x_j, j \neq i, i = 1,2,3$.

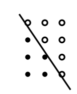
Aberration

Motivated by the term *generator aberration*, we would like to identify models with low degree. Linear aberration gives a weighted (average) degree:

$$A(w, L) = \frac{1}{n} \sum w_i \bar{\alpha}_{L_i}$$

with $w_i \geq 0$, $\sum w_i = 1$. Models in the algebraic fan minimize $A(w, L)$. This is because models in $A(\mathcal{D})$ lie in the boundary of $\mathcal{S}(I)$.

A special type of designs are *generic designs*, which minimize $A(w, L)$ over all weights w and over all sets of hierarchical monomials. Importantly, for generic designs, the models in $A(\mathcal{D})$ are *corner cut* models.



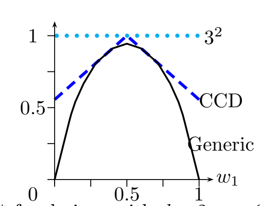
Corner cut
$\{1, x_1, x_2, x_1x_2\}$ is not *corner cut*

Bounds for minimal aberration [2]

For fixed w, minimal aberration is computed by minimization of aberration over all hierarchical models identifiable by the design: $A^* = \min_L A(w, L)$. Bounds were found for minimal aberration A^* over the family of generic designs:

$$A^+ - 1 \leq A^* \leq A^+ + 1$$

where $A^+ = k \cdot g(w)$.
Approximated minimal aberration: $\tilde{A} = db^{1/d} g(w) - a$ where a, b are fixed constants. We experimentally observe that $\tilde{A} \to A^*$



A^* for designs with $d = 2, n = 9$.

Fractions of factorial designs 2^d

The algebraic techniques exposed above have particular characteristics when the designs are fractions of factorial designs with two levels. The models are squarefree, i.e. (hierarchical) subsets of $\bigotimes_{i=1}^d \{1, x_i\}$. For this reason, models can be seen and studied as simplicial complexes. This is due to the squarefree nature of models and the hierarchical structure. We thus write Δ for the model (2) to emphasize its simplicial structure.
The ideal of leading terms $\langle \mathrm{LT}(I(\mathcal{D})) \rangle$ is related to a special ideal created by Δ:

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

here I_Δ is the squarefree monomial ideal created by the non-faces of Δ. This ideal is the well-known Stanley-Reisner ideal [5], and $\bar{I}$ is the artinian closure of I.
The complexity of the model Δ can be studied by the Stanley-Reisner ring $R[x_1, \ldots, x_d]/I_\Delta$. Note that this ring is not the same as the quotient ring of Equation (1), i.e. the artinian closure of I_Δ is required.

The role of Betti numbers[a]

Betti numbers stem from topology, and the Betti number of a space counts the maximum number of cuts that can be made without dividing the space into two pieces. Informally, the k-th Betti number refers to the number of unconnected k-dimensional surfaces. For instance, the first few Betti numbers have the following intuitive definitions: β_0 is the number of unconnected components, β_1 is the number of two dimensional circles "holes", and β_2 is the number of three dimensional "holes" or voids.
We aim to relate Betti numbers of the ideal of leading terms $\beta(\mathrm{LT}(I(\mathcal{D})))$ to models that minimize aberration. A starting point is that for a design fan, Betti numbers of the leading term ideal are maximal when using graded term ordering.
Currently Betti numbers provide a direct interpretation in terms of the complexity of the boundary between model Δ and its complement. Models are categorized by sharing the same Hilbert function.

[a] joint with E. Sáenz de Cabezón (La Rioja)

Plackett Burman (PB) designs [7]

Designs in this family consist of small fractions of 2^d with d factors and $n = d+1$ runs. The designs are constructed by circular shifts of a generator and are available for $d = 7, 11, 15, 19, 23, \ldots$ PB designs possess a complicated aliasing table, but they have an orthogonal design-model matrix for the linear model with all factors

$$E(y) = \theta_0 + \sum_{i=1}^d \theta_i x_i \qquad (3)$$

PB designs are a popular choice for screening in a first stage of experimentation. We study their algebraic fan and describe the structure of models with the aid of Betti numbers.
Example 3: PB8 Consider a Plackett-Burman design with eight runs, seven factors a, b, c, d, e, f, g and generator $+--+-++$. We refer to this design as PB8. Set a block term ordering $\prec_1$ for which monomials in c, d, e are smaller than monomials in a, b, f, g. The model identified by PB8 is

$$\Delta_{\prec_1} = \{1, e, d, de, c, ce, cd, a\},$$

that is a simplicial complex consisting of two components. The description of Δ (degree by degree) is found by computing the Hilbert series of the quotient ring $R[a, b, c, d, e, f, g]/\langle I(\mathcal{D}) \rangle$ which is $HS(s) = 1 + 4s + 3s^2$.
A further description of the model border shows that $\Delta_{\prec_1}$ contains one leading monomial of degree one (a) and three leading monomials of degree two (de, ce, cd). We now examine the monomial ideal generated by the leading terms of the design ideal $\langle I(\mathcal{D}) \rangle = \langle f, b, g, a^2, ac, ad, ae, c^2, e^2, d^2, cde \rangle \subset \mathbb{R}[a, b, c, d, e, f, g]$, which has three generators of degree one, seven of degree two and one of degree three. This precise description degree by degree is read from the Betti diagram of the ideal of leading terms, i.e. $\beta_{0,1} = 3, \beta_{0,2} = 7, \beta_{0,3} = 1$, and the border description of Δ is also read from the Betti diagram: $\beta_{6,8} = 1, \beta_{6,9} = 3$.

	0	1	2	3	4	5	6
1:	3	3	1	-	-	-	-
2:	7	30	52	47	24	7	1
3:	1	10	33	52	43	18	3
Tot:	11	43	86	99	67	25	4

Table 1: Betti diagram for PB8 and $\prec_1$.

In contrast, consider a term ordering $\prec_2$ which is `DegRevLex`. Now the Betti diagram contains much higher numbers $\beta_{0,2} = 28$ (number of leading terms of $\langle I(\mathcal{D}) \rangle$) and $\beta_{6,8} = 7$ (number of leading monomials of Δ). In other words, PB8 identifies a simplicial complex with seven disconected components $\Delta_{\prec_2} = \{1, a, b, c, d, e, f, g\}$ and a very complex ideal of leading terms. This is the model for Equation (3).
Now repeat the analysis with a `Lex` term ordering $\prec_3$. The Betti diagram now shows much smaller figures $\beta_{0,1} = 4, \beta_{0,2} = 3$ and $\beta_{6,10} = 1$. The model $\Delta_{\prec_3}$ is simpler, being one simplicial complex completely connected. Correspondingly, its border shares this simplicity.
The results for the 218 models in the algebraic fan of PB8 are summarized in Table 2, where representatives of six equivalence classes (up to permutation of factors) are shown. Note that models with the same Hilbert function do not neccesarily share the same Betti numbers. Take model $\Delta_{\prec_4}$ with numbers $\beta_{0,1} = 3, \beta_{0,2} = 7$ and $\beta_{6,9} = 3$, which are smaller than for the model $\Delta_{\prec_1}$ (Block).

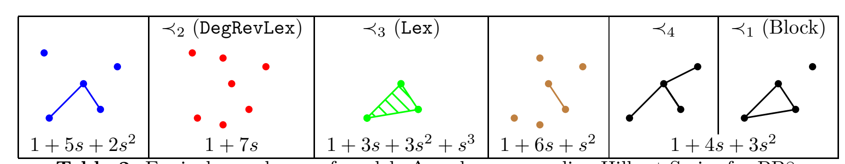


Table 2: Equivalence classes of models Δ and corresponding Hilbert Series for PB8.

Example 4: PB12 This design has 12 runs in 11 factors and generator $++-+++---+-$. The algebraic fan of PB12 is very complex, showing a rich variety of simplicial models. Despite its enormous size (around 3×10^5), models have been classified in nineteen classes (up to permutations of variables), which in turn share only ten distinct Hilbert Series. Among the different models, Betti numbers for the model with Hilbert Series $1 + 11s$ are the largest.