MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Design and Analysis of Fractional Factorial Experiments from the Viewpoint of Computational Algebraic Statistics

Satoshi AOKI and Akimichi TAKEMURA

METR 2010–08 April 2010

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t-u-tokyo.ac.jp/research/techrep/index.html

Design and analysis of fractional factorial experiments from the viewpoint of computational algebraic statistics

Satoshi Aoki*†and Akimichi Takemura‡†

April, 2010

Abstract

We give an expository review of applications of computational algebraic statistics to design and analysis of fractional factorial experiments based on our recent works. For the purpose of design, the techniques of Gröbner bases and indicator functions allow us to treat fractional factorial designs without distinction between regular designs and non-regular designs. For the purpose of analysis of data from fractional factorial designs, the techniques of Markov bases allow us to handle discrete observations. Thus the approach of computational algebraic statistics greatly enlarges the scope of fractional factorial designs.

1 Introduction

Application of Gröbner bases theory to designed experiments is an attractive topic in a relatively new field in statistics, called *computational algebraic statistics*. After the first work by Pistone and Wynn ([24]) this topic is vigorously studied both by algebraists and statisticians. These developments are stimulated by advancements in algebraic algorithms. By recent algorithms some practical computations are becoming feasible for statistical applications. For these backgrounds see [5].

In this paper we revisit some fundamental results in this field, mainly in the treatments of two-level fractional factorial designs. The most important part of the long history of studies of two-level fractional factorial designs is the theory of regular designs. As explained in many classical works (e.g. [8], [9]), properly chosen regular fractional factorial designs have many desirable properties, mainly due to orthogonality and balancedness of designs. In addition, an elegant theory based on the linear algebra over $GF(2)$ is well established for regular two-level fractional factorial designs (e.g. [20]).

*Graduate School of Science and Engineering (Science Course), Kagoshima University.

†JST, CREST.

‡Department of Mathematical Informatics, Graduate School of Information Science and Technology, University of Tokyo.

On the other hand, there are still many open problems on general structures of non-regular fractional factorial designs, except for some specific designs such as Plackett-Burman designs. In computational algebraic statistics, designs are simply characterized as the solutions of a set of polynomial equations. Therefore we need not distinguish between regular designs and non-regular designs. In fact, this algebraic treatment yields many new results. For example, concepts defined for regular designs, such as resolution and aberration, can be naturally generalized to non-regular designs in the framework of computational algebraic statistics.

The construction of this paper is as follows. In Section 2 we review some algebraic definitions for handling fractional factorial designs. In Section 3 we present a method of representing the confounding relations between factor effects algebraically, which is one of the results in [24]. In Section 4 an indicator function defined in [15] is given. The indicator function is a valuable tool to characterize non-regular designs. In Section 5 we show a simple result related to the indicator function when new factors are added or when interaction effects are formally considered as factors. In Section 6 we consider the problem of selecting optimal designs. In Section 7 we discuss Markov chain Monte Carlo approach for testing factor effects, when observations are discrete random variables. We end the paper with some discussions in Section 8.

Throughout this paper, we use terminology of Gröbner bases theory without definition. In particular, we omit explanations for term orders and division algorithms. For these notions, see [11] or [1], for example.

2 Design ideals

Consider fractional factorial designs of m controllable factors. We assume that the levels of each factor are coded as elements of a field K, which is a finite extension of the field $\mathbb{Q}$ of rational numbers. For the case of factors with two levels we can take $K = \mathbb{Q}$. However for the case of factors with more than two levels, it is advantageous to code the levels by complex numbers ([22], [23]). In this case K contains some complex roots of unity.

A fractional factorial design (without replication) is identified with a finite subset of K^m. In computational algebraic statistics, this set is considered as the set of solutions of polynomial equations, called an *algebraic variety*, and the set of polynomials vanishing on all the solutions is called an *ideal*. For the rest of this section we only consider the case of two-level factors. For more general case see [21].

The full factorial design of m factors with two levels is expressed as

$$\mathcal{D} = \{(x_1, \ldots, x_m) \mid x_1^2 = \cdots = x_m^2 = 1\} = \{-1, +1\}^m,$$

where we write -1 and 1 as the two levels. We call a subset $\mathcal{F} \subset \mathcal{D}$ a fractional factorial design. Let $K[x_1, \ldots, x_m]$ be the polynomial ring of indeterminates $x_1, \ldots, x_m$ with the coefficients in K. Then the set of polynomials vanishing on the points of $\mathcal{F}$

$$I(\mathcal{F}) = \{f \in K[x_1, \ldots, x_m] \mid f(x_1, \ldots, x_m) = 0 \text{ for all } (x_1, \ldots, x_m) \in \mathcal{F}\}$$

is the design ideal of $\mathcal{F}$.

An ideal $I \subset K[x_1, \dots, x_m]$ is generated by a (finite) basis $\{g_1, \dots, g_k\} \subset I$ if for any $f \in I$ there exist polynomials $s_1, \dots, s_k \in K[x_1, \dots, x_m]$ such that

$$f(x_1, \dots, x_m) = \sum_{i=1}^{k} s_i(x_1, \dots, x_m) g_i(x_1, \dots, x_m).$$

The above $s_1, \dots, s_k$ are not unique in general. We write $I = \langle g_1, \dots, g_k \rangle$ if I is generated by a basis $\{g_1, \dots, g_k\}$. For example, for the full factorial design of two factors with two levels (2^2-design), the design ideal of $\mathcal{D} = \{-1, +1\}^2$ is written as

$$I(\mathcal{D}) = \langle x_1^2 - 1, x_2^2 - 1 \rangle.$$

Every ideal has a finite basis by the Hilbert basis theorem. In addition, if $\{g_1, \dots, g_k\}$ is a basis of $I(\mathcal{F})$, then $\mathcal{F}$ coincides with the solutions of the polynomial equations $g_1 = 0, \dots, g_k = 0$.

Suppose there are n runs (i.e. points) in a fractional factorial design $\mathcal{F} \subset \mathcal{D}$. A general method to derive a basis of $I(\mathcal{F})$ is to make use of the algorithm for calculating the intersection of the ideals. By definition, the design ideal of the design consisting of a single point, $(a_1, \dots, a_m) \in \{-1, +1\}^m$, is written as

$$\langle x_1 - a_1, \dots, x_m - a_m \rangle \subset K[x_1, \dots, x_m].$$

Therefore the design ideal of the n-runs design, $\mathcal{F} = \{(a_{i1}, \dots, a_{im}),\ i = 1, \dots, n\}$, is given as

$$I(\mathcal{F}) = \bigcap_{i=1}^{n} \langle x_1 - a_{i1}, \dots, x_m - a_{im} \rangle. \tag{1}$$

To calculate the intersection of ideals, we can use the theory of Gröbner bases. In fact, by introducing the indeterminates $t_1, \dots, t_n$ and the polynomial ring $K[x_1, \dots, x_m, t_1, \dots, t_n]$, equation (1) is written as

$$I(\mathcal{F}) = I^* \cap K[x_1, \dots, x_m],$$

where

$$I^* = \langle t_i(x_1 - a_{i1}), \dots, t_i(x_m - a_{im}),\ i = 1, \dots, n,\ t_1 + \cdots + t_n - 1 \rangle \tag{2}$$

is an ideal of $K[x_1, \dots, x_m, t_1, \dots, t_n]$. Therefore we can obtain a basis of $I(\mathcal{F})$ as the reduced Gröbner basis of I^* with respect to a term order satisfying $\{t_1, \dots, t_n\} \succ \{x_1, \dots, x_m\}$. This argument is known as the elimination theory, one of the important applications of Gröbner bases ([11]).

Example 2.1 (A 2^{7-4}_{III} design)**.** *Consider the design known as the orthogonal array $L_8(2^7)$ of resolution III with the defining relations*

$$x_3 = -x_1x_2,\ x_5 = -x_1x_4,\ x_6 = -x_2x_4,\ x_7 = x_1x_2x_4 \tag{3}$$

given as follows.

run\factor	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	-1	-1	-1	-1	-1	-1	-1
2	-1	-1	-1	1	1	1	1
3	-1	1	1	-1	-1	1	1
4	-1	1	1	1	1	-1	-1
5	1	-1	1	-1	1	-1	1
6	1	-1	1	1	-1	1	-1
7	1	1	-1	-1	1	1	-1
8	1	1	-1	1	-1	-1	1

For this design a basis of $I(\mathcal{F})$ is obtained by omitting elements containing the indeterminates $t_1, \ldots, t_8$ from the reduced Gröbner basis of (2). For the lexicographic term order with $x_1 \succ \cdots \succ x_7$, the reduced Gröbner basis is given as

$$\{x_7^2 - 1, x_6^2 - 1, x_5^2 - 1, x_3 + x_5x_6, x_2 + x_5x_7, x_1 + x_6x_7, x_4 - x_5x_6x_7\}, \qquad (4)$$

while for the graded reverse lexicographic term order, the reduced Gröbner basis is given as

$$\begin{aligned}&\{x_7^2 - 1, x_6^2 - 1, x_5^2 - 1, x_4^2 - 1, x_3^2 - 1, x_2^2 - 1, x_1^2 - 1,\\ &x_2x_3 + x_1, x_4x_5 + x_1, x_6x_7 + x_1, x_1x_3 + x_2, x_4x_6 + x_2, x_5x_7 + x_2,\\ &x_1x_2 + x_3, x_4x_7 + x_3, x_5x_6 + x_3, x_1x_5 + x_4, x_2x_6 + x_4, x_3x_7 + x_4,\\ &x_1x_4 + x_5, x_2x_7 + x_5, x_3x_6 + x_5, x_1x_7 + x_6, x_2x_4 + x_6, x_3x_5 + x_6,\\ &x_1x_6 + x_7, x_2x_5 + x_7, x_3x_4 + x_7\}.\end{aligned} \qquad (5)$$

Hereafter, we write a monomial of the indeterminates $x_1, \ldots, x_m$ as $\mathbf{x}^{\mathbf{a}} = x_1^{a_1} \cdots x_m^{a_m}$. It is sufficient to consider $\mathbf{a} = (a_1, \ldots, a_m) \in \{0, 1\}^m$ since the two levels are coded as $\{-1, +1\}$. The results in Example 2.1 indicate the relation between the defining relation and the design ideal for regular fractional factorial designs, i.e., designs obtained from the defining relations such as (3). In fact, for the design $\mathcal{F}$ obtained by the defining relation

$$\mathbf{x}^{\mathbf{a}_\ell} = c_\ell,\ c_\ell \in \{-1, 1\},\ \ell = 1, \ldots, s,$$

the design ideal is written as

$$I(\mathcal{F}) = \langle x_1^2 - 1, \ldots, x_m^2 - 1, \mathbf{x}^{\mathbf{a}_1} - c_1, \ldots, \mathbf{x}^{\mathbf{a}_s} - c_s \rangle.$$

For example, the design ideal in Example 2.1 is also written as

$$I(\mathcal{F}) = \langle x_1^2 - 1, \ldots, x_7^2 - 1, x_1x_2x_3 + 1, x_1x_4x_5 + 1, x_2x_4x_6 + 1, x_1x_2x_4x_7 - 1 \rangle. \qquad (6)$$

As we see here, an obvious basis of the design ideal of a regular fractional factorial design $\mathcal{F} \subset \mathcal{D}$ consists of defining relations in addition to $x_1^2 - 1, \ldots, x_m^2 - 1$. Also for non-regular designs we can consider the set of polynomials (in addition to $x_1^2 - 1, \ldots, x_m^2 - 1$) which forms a basis of $I(\mathcal{F})$. This set of the polynomials is called a set of defining equations of $\mathcal{F}$ in [15]. This is a generalized concept of defining relations from regular to non-regular designs. Note that the above *obvious* basis of a regular design is not a Gröbner basis in general. In fact, the right hand side of (6) is not a Gröbner basis for any term order. In the arguments above we used the elimination theory as a general method to obtain a basis of the design ideal and obtained a reduced Gröbner basis as a result. However, it is important *in itself* to obtain a Gröbner basis, which we see in the next section.

3 The confounding relation and the ideal membership problem

In this section we see from the Gröbner bases theory that the confounding relation can be generalized from regular to non-regular designs and expressed concisely. This is one of the merits to consider the design ideal $I(\mathcal{F})$. In fact, the problem of judging whether two factor effects or interaction effects are confounded or not is equivalent to the ideal membership problem, which is solved by calculation of a Gröbner basis of the design ideal. We give an overview of this fact. For details see [24] or [16].

First we give some necessary notation and definitions. $\mathrm{LT}_\tau(f)$ denotes the leading term of the polynomial $f \in K[x_1, \ldots, x_m]$ with respect to the term order τ. For an ideal $I \subset K[x_1, \ldots, x_m]$, we write the set of the leading terms of the elements in I as $\mathrm{LT}_\tau(I) = \{\mathrm{LT}_\tau(f) \mid f \in I\}$. A monomial is called a standard monomial if it does not belong to $\mathrm{LT}_\tau(I)$. From the definition of Gröbner basis, the set of standard monomials is also characterized as the set of monomials which is not divisible by any leading term of the element of the Gröbner basis with respect to the term order τ. We write the set of standard monomials $\{\mathbf{x}^{\mathbf{a}} \mid \mathbf{x}^{\mathbf{a}} \notin \mathrm{LT}_\tau(I(\mathcal{F}))\}$ as $\mathrm{Est}_\tau(\mathcal{F})$. The following is a basic theorem (Proposition 1.1 of [25]) in the theory of Gröbner bases.

Theorem 3.1. *$K[x_1, \ldots, x_m]/I(\mathcal{F})$ is isomorphic as a K-vector space to $\mathrm{Span}(\mathrm{Est}_\tau(\mathcal{F}))$. $\mathrm{Est}_\tau(\mathcal{F})$ is a basis of this vector space.*

$\mathrm{Est}_\tau(\mathcal{F})$ represents one of the identifiable sets of main and interaction effects under the design $\mathcal{F}$ and the number of the monomials in $\mathrm{Est}_\tau(\mathcal{F})$ is always the same as the run size n for any τ.

Example 3.1. *For the two reduced Gröbner bases in Example 2.1, $\mathrm{Est}_\tau(\mathcal{F})$ is written as follows.*

- *lexicographic:* $\{1, x_5, x_6, x_7, x_5x_6, x_5x_7, x_6x_7, x_5x_6x_7\}$
- *graded reverse lexicographic:* $\{1, x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$

Now we consider the relation between the confounding relation and the design ideal. We identify a monomial $\mathbf{x}^{\mathbf{a}}$ with a *main effect* if $\sum_{i=1}^m a_i = 1$, and a *two-factor interaction effect* if $\sum_{i=1}^m a_i = 2$ and so on. Then two main or interaction effects are confounded in the design $\mathcal{F}$ if $\mathbf{x}^{\mathbf{a}_1}\mathbf{x}^{\mathbf{a}_2}$ is identically equal to $+1$ (or -1) for all the points in $\mathbf{x} \in \mathcal{F}$. The confounded effects cannot be estimated simultaneously. Therefore a design has to be chosen such that main effects are confounded with higher order interaction effects under the hierarchical assumption. This is the concept of resolution. The confounding relation is expressed in terms of the design ideal as follows.

Proposition 3.1. *Let $c \in \{-1, +1\}$. Then the following two conditions are equivalent.*

$$\text{(i)}\ \ \mathbf{x}^{\mathbf{a}_1}\mathbf{x}^{\mathbf{a}_2} = c \ \text{ for all } \ \mathbf{x} \in \mathcal{F} \qquad \text{(ii)}\ \ \mathbf{x}^{\mathbf{a}_1} - c\mathbf{x}^{\mathbf{a}_2} \in I(\mathcal{F})$$

In general, we have to calculate a Gröbner basis to judge whether a given polynomial belongs to a given ideal or not, i.e., to solve the ideal membership problem.

Example 3.2. *Consider the design in Example 2.1 again. Since the defining relation $x_3 = -x_1x_2$ exists, the main effect of x_3 and two-factor interaction effect of x_1 and x_2 are confounded. For the reduced Gröbner basis (5), $x_1x_2 + x_3 \in I(\mathcal{F})$ is obvious since the basis includes $x_1x_2 + x_3$. On the other hand for the reduced Gröbner basis (4), it is shown as follows*

$$x_1x_2 + x_3 = (x_1 + x_6x_7)x_2 - (x_2 + x_5x_7)x_6x_7 + (x_7^2 - 1)x_5x_6 + (x_3 + x_5x_6).$$

For the case that the factor has s ($s > 2$) levels, similar relation as in Proposition 3.1 holds if we code the levels as the sth root of unity. For example of three-level factors, the levels are coded as $\{1, \omega, \omega^2\}$, $\omega = \exp(2\pi i/3)$. See [23] and [4] for details.

4 Indicator functions

In this section, we introduce an *indicator function*, which is defined by [15]. The indicator function of a design $\mathcal{F} \subset \mathcal{D}$ is a polynomial $f \in K[x_1, \ldots, x_m]$ satisfying

$$f(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \mathcal{F} \\ 0, & \text{if } \mathbf{x} \in \mathcal{D} \setminus \mathcal{F}. \end{cases}$$

The indicator function has a unique square-free representation under the constraint $x_i^2 = 1, i = 1, \ldots, m$, and in a one-to-one correspondence to the design $\mathcal{F}$. Many important results in the field of computational algebraic statistics are related to the indicator function. For example, some classes of fractional factorial designs can be classified by the coefficients of their indicator functions. It is also shown that some concepts of designed experiments such as confounding, resolution, orthogonality and estimability, are related to the structure of the indicator function of a design. Since the indicator function is defined for any design, some classical notions for regular designs, such as confounding and resolution, can be generalized to non-regular designs naturally by the notion of the indicator function. See [15] or [27] for details.

In addition, since the indicator function is a polynomial, it can be incorporated into the theory of computational algebraic statistics naturally. For example, the design ideal $I(\mathcal{F})$ is simply written as

$$I(\mathcal{F}) = \langle x_1^2 - 1, \ldots, x_m^2 - 1, f(\mathbf{x}) - 1 \rangle,$$

where $f(\mathbf{x})$ is the indicator function of $\mathcal{F}$. In other words, the indicator function forms a set of defining equations by itself.

We list some characteristics of the indicator function. The indicator function of the full factorial design $\mathcal{D}$ is $f(\mathbf{x}) \equiv 1$. The constant term of the indicator function of a fractional factorial designs is equal to the fraction $n/2^m$. The indicator function of a regular fractional factorial designs is simply written as a product of its defining relations (see [15] or [27]). For example, the indicator function of the 2^{7-4} design in Example 2.1 is written as

$$f(\mathbf{x}) = \frac{1}{16}(1 - x_1x_2x_3)(1 - x_1x_4x_5)(1 - x_2x_4x_6)(1 + x_1x_2x_4x_7). \quad (7)$$

The absolute values of the coefficients in the indicator function do not exceed the constant term. In particular, the absolute values of all the coefficients in the indicator function of regular designs coincide with the constant term.

Concerning non-regular designs, one of the results on the coefficients of the indicator function of a non-regular design is related to the existence of a regular design including the non-regular design.

Example 4.1 (The indicator function of the non-regular fractional factorial designs). *Consider the following three fractional factorial designs.*

$\mathcal{F}_1$

x_1	x_2	x_3
1	1	1
1	−1	−1
−1	1	−1
−1	−1	1

$\mathcal{F}_2$

x_1	x_2	x_3
1	1	1
1	−1	−1
−1	1	−1

$\mathcal{F}_3$

x_1	x_2	x_3
1	1	1
1	1	−1
1	−1	1
−1	1	1

$\mathcal{F}_1$ is a 2^{3-1} regular design defined by $x_1x_2x_3 = 1$, and $\mathcal{F}_2$ is a non-regular design which is a proper subset of $\mathcal{F}_1$. $\mathcal{F}_3$ is also a non-regular design, but there does not exist a regular design which includes $\mathcal{F}_3$ as a proper subset. The indicator functions of these three designs are given as follows.

$$\begin{aligned}
\mathcal{F}_1 &: f(\mathbf{x}) = \frac{1}{2} + \frac{1}{2}x_1x_2x_3 \\
\mathcal{F}_2 &: f(\mathbf{x}) = \frac{3}{8} + \frac{1}{8}(x_1 + x_2 + x_3 - x_1x_2 + x_1x_3 + x_2x_3) + \frac{3}{8}x_1x_2x_3 \\
\mathcal{F}_3 &: f(\mathbf{x}) = \frac{1}{2} + \frac{1}{4}(x_1 + x_2 + x_3 - x_1x_2x_3)
\end{aligned}$$

An important observation is that the terms whose coefficients are equal to to the constant term in the indicator function of $\mathcal{F}_2$ (i.e., $x_1x_2x_3$) coincide with the terms of the indicator function of $\mathcal{F}_1$, and there are no such terms in the indicator function of $\mathcal{F}_3$. These characteristics of indicator functions hold in general.

The absolute value of a coefficient in the indicator function represents a complete confounding relation if it is equal to the constant term, and a partial confounding relation if it is smaller than the constant term. Considering this point, we gave in [3] a new class of two-level non-regular fractional factorial designs, called an *affinely full-dimensional factorial design*, which has a desirable property for the identifiability of parameters. We present it briefly in Section 6.

5 Indicator function for adding factors

In this section, as a simple application of the indicator function, we consider the design ideal for adding factors. The additional factors may be real controllable factors, whose levels are determined by some defining relations. For the purpose of Markov bases in

Section 7 the additional factors are formal and correspond to interaction effects included in a null hypothesis.

Let $\mathcal{F}_1$ be a fractional factorial design of the factors $x_1, \ldots, x_m$. Consider adding factors $y_1, \ldots, y_k$ to $\mathcal{F}_1$. We suppose the levels of the additional factors are determined by the defining relations among $x_1, \ldots, x_m$ as

$$y_1 = e_1 \mathbf{x}^{\mathbf{b}_1}, \ldots, y_k = e_k \mathbf{x}^{\mathbf{b}_k},$$

where $e_1, \ldots, e_k \in \{-1, 1\}$. Write this new design of $x_1, \ldots, x_m, y_1, \ldots, y_k$ as $\mathcal{F}_2$. The run sizes of $\mathcal{F}_1$ and $\mathcal{F}_2$ are the same.

Let f_1 and f_2 be the indicator functions of $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively. Then we have

$$f_2(x_1, \ldots, x_m, y_1, \ldots, y_k) = \frac{1}{2^k}(1 + e_1 y_1 \mathbf{x}^{\mathbf{b}_1}) \cdots (1 + e_k y_k \mathbf{x}^{\mathbf{b}_k}) f_1(x_1, \ldots, x_m). \quad (8)$$

In fact, for $(x_1, \ldots, x_m, y_1, \ldots, y_k) \in \mathcal{F}_2$, $(x_1, \ldots, x_m) \in \mathcal{F}_1$ and

$$e_1 y_1 \mathbf{x}^{\mathbf{b}_1} = \cdots = e_k y_k \mathbf{x}^{\mathbf{b}_k} = 1$$

hold, which yields $f_2 = 1$. Conversely, if $(x_1, \ldots, x_m, y_1, \ldots, y_k) \notin \mathcal{F}_2$, then $(x_1, \ldots, x_m) \notin \mathcal{F}_1$ or some of $e_1 y_1 \mathbf{x}^{\mathbf{b}_1}, \ldots, e_k y_k \mathbf{x}^{\mathbf{b}_k}$ has to be -1, which yields $f_2 = 0$. Note that (8) generalizes the indicator function of regular fractional factorial designs (7), by taking $f_1 \equiv 1$, i.e. by assuming the full factorial design for $x_1, \ldots, x_m$.

From the above result, we have an expression of $I(\mathcal{F}_2)$

$$I(\mathcal{F}_2) = \langle x_1^2 - 1, \ldots, x_m^2 - 1, y_1^2 - 1, \ldots, y_k^2 - 1, f_1 - 1, f_2 - 1 \rangle.$$

If we fix the term order τ on $x_1, \ldots, x_m$ and σ on $x_1, \ldots, x_m, y_1, \ldots, y_k$, $\mathrm{Est}_\tau(\mathcal{F}_1)$ and $\mathrm{Est}_\sigma(\mathcal{F}_2)$ are defined. $\mathrm{Est}_\tau(\mathcal{F}_1)$ and $\mathrm{Est}_\sigma(\mathcal{F}_2)$ contain the same number of monomials since the run sizes of $\mathcal{F}_1$ and $\mathcal{F}_2$ are the same. In particular, if we use a term order σ such that $\{y_1, \ldots, y_k\} \succ_\sigma \{x_1, \ldots, x_m\}$, then $\mathrm{Est}_\tau(\mathcal{F}_1) = \mathrm{Est}_\sigma(\mathcal{F}_2)$ holds.

6 Consideration of selection of optimal designs

This section is based on two works by the authors related to optimal selection of non-regular designs.

In [3] we defined a new class of two-level non-regular fractional factorial designs as follows.

Definition 6.1 (Definition 2.1 of [3]). *A non-regular fractional factorial design $\mathcal{F}$ is called an affinely full-dimensional factorial design if there is no regular fractional factorial design $\mathcal{F}'$ satisfying $\mathcal{F} \subsetneq \mathcal{F}'$. Conversely, a non-regular fractional factorial design $\mathcal{F}$ is called a subset fractional factorial design if there is some regular fractional factorial design $\mathcal{F}'$ satisfying $\mathcal{F} \subsetneq \mathcal{F}'$.*

At a glance, the merit of this definition might not be clear. One of the properties of the affinely full-dimensional factorial design is the simultaneous identifiability of the parameters. In fact, if $\mathcal{F}$ is an affinely full-dimensional factorial design, then the parameters of the main effect model are simultaneously identifiable and vice versa. See Lemma 2.1 and Lemma 2.2 of [3]. As another interesting property of the affinely full-dimensional factorial design, we have the following conjecture (Conjecture 3.1 of [3]) on the D-optimality of affinely full-dimensional designs.

Conjecture 6.1. *Consider the main effect model for the observations obtained in a fractional factorial design of m factors. Then D-optimal design is affinely full-dimensional factorial if and only if $m = 5, 6, 7$ (mod 8).*

Though this conjecture is not proved in general, it is shown to be true when some known bounds for the maximal determinant problem ([7], [14] and [26]) are attained. See [3] for details.

In [2], we considered more realistic situation that the model is unknown. In this case, we cannot rely on model-based criterion such as D-optimality and have to evaluate the model-robustness. In [2], we considered the situation where (i) all the main effects are of primary interest and their estimates are required, (ii) an experimenter assumes that there are certain number of active two-factor interaction effects and certain number of active three-factor interaction effects, but it is unknown which of two- and three-factor interactions are active, and (iii) all the four-factor and higher-order interactions are negligible. This is a natural extension of the setting considered in [10]. In [3], we presented some optimality criteria to evaluate model-robustness of non-regular two-level fractional factorial designs. Our approach was based on minimizing the sum of squares of all the off-diagonal elements in the information matrix, and considering expectation under appropriate distribution functions for unknown contamination of the interaction effects. We also compared our criterion to a generalized minimum aberration criterion by [12] and affinely full dimensionality.

7 Markov bases and conditional tests by Markov chain Monte Carlo method

In this section we introduce another topic of application of Gröbner basis theory to the designed experiments, called *Markov bases for designed experiments*. The notion of Markov bases was introduced in [13]. They established a procedure for sampling from discrete conditional distributions by constructing a connected Markov chain on a given sample space. Since then many works have been published on the topic of Markov bases by both algebraists and statisticians. This constitutes another main branch of the field of computational algebraic statistics. See [5] for the history of this topic. It is of interest to investigate statistical problems which are related to both designed experiments and Markov bases. In [6] and [4] we considered applying Markov bases for discrete observations from designed experiments. In this section we review the results of these works.

Suppose we have nonnegative integer observations for each run of a regular fractional design. For simplicity, we also suppose that the observations are counts of some events and only one observation is obtained for each run. In this case it is natural to consider the Poisson model, in the framework of generalized linear models ([19]). Write the observations as $\mathbf{y} = (y_1, \dots, y_n)'$, where n is the run size and $'$ denotes the transpose. The observations are realizations from n Poisson random variables $Y_1, \ldots, Y_n$, which are mutually independently distributed with the mean parameter $\mu_i = E(Y_i), i = 1, \ldots, n$. We express the mean parameter μ_i as

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{\nu-1} x_{i\nu-1}, \tag{9}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_{\nu-1})'$ is the ν-dimensional parameter and $x_{i1}, \ldots, x_{i\nu-1}$ are the $\nu - 1$ covariates. We write the covariate matrix A as

$$A = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1\nu-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n\nu-1} \end{pmatrix}. \tag{10}$$

Note that the expression (9) can be treated as the *null model* H_0. Since the saturated model has n-dimensional parameter, various goodness-of-fit tests with the saturated model as the alternative H_1 can be written as

$$\begin{array}{ll} \mathrm{H}_0: & (\beta_\nu, \ldots, \beta_n) = (0, \ldots, 0) \\ \mathrm{H}_1: & (\beta_\nu, \ldots, \beta_n) \neq (0, \ldots, 0) \end{array}$$

by introducing additional parameter $(\beta_\nu, \ldots, \beta_n)$. Various other hypotheses can be written in a similar way. Under the null model (9) the sufficient statistic for the parameter β is given by $A'\mathbf{y} = (\sum_{i=1}^n y_i, \sum_{i=1}^n x_{i1} y_i, \ldots, \sum_{i=1}^n x_{i\nu-1} y_i)'$.

Once we specify the null model and a test statistic, our purpose is to calculate the p value. In this stage, Markov chain Monte Carlo procedure is a valuable tool, especially when the fitting of the traditional large-sample approximation is poor and the exact calculation of the p value is infeasible. To perform the Markov chain Monte Carlo procedure, the key notion is a Markov basis over the conditional sample space given the values of the sufficient statistic

$$\{\mathbf{y} \mid A'\mathbf{y} = A'\mathbf{y}^o,\ y_i \text{ is a non-negative integer},\ i = 1, \ldots, n\}, \tag{11}$$

where $\mathbf{y}^o$ is the observed vector. Once a Markov basis is calculated, we can construct a connected, aperiodic and reversible Markov chain over the space in (11), which can be modified so that the stationary distribution is the conditional distribution under the null model by the Metropolis-Hastings procedure. See [13] and [17] for details.

In the arguments above, an important step is to construct a covariate matrix A in (10). In this step, we have to construct A so that all the parameters in (9) are simultaneously estimable. This problem corresponds to the ideal membership problem in Section 3.

We illustrate the above setup with two examples.

We first consider a 2^{7-3} fractional factorial design. Suppose we have observations $\mathbf{y} = (y_1, \ldots, y_{16})'$ for each run of the fractional factorial design with the defining relation

$$x_1 x_2 x_4 x_5 = x_1 x_3 x_4 x_6 = x_2 x_3 x_4 x_7 = 1. \tag{12}$$

The design and the observation is written as follows.

	Factor							$\mathbf{y}$
Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
1	1	1	1	1	1	1	1	y_1
2	1	1	1	−1	−1	−1	−1	y_2
3	1	1	−1	1	1	−1	−1	y_3
4	1	1	−1	−1	−1	1	1	y_4
5	1	−1	1	1	−1	1	−1	y_5
6	1	−1	1	−1	1	−1	1	y_6
7	1	−1	−1	1	−1	−1	1	y_7
8	1	−1	−1	−1	1	1	−1	y_8
9	−1	1	1	1	−1	−1	1	y_9
10	−1	1	1	−1	1	1	−1	y_{10}
11	−1	1	−1	1	−1	1	−1	y_{11}
12	−1	1	−1	−1	1	−1	1	y_{12}
13	−1	−1	1	1	1	−1	−1	y_{13}
14	−1	−1	1	−1	−1	1	1	y_{14}
15	−1	−1	−1	1	1	1	1	y_{15}
16	−1	−1	−1	−1	−1	−1	−1	y_{16}

In this case, there are several models to be considered. If our interest is only on the main effects for seven factors, we may define

$$\beta = (\beta_0, \beta_1, \dots, \beta_7)'$$

and

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix}' \tag{13}$$

i.e., the covariate matrix A is constructed as the design matrix and the column vector $(1, \dots, 1)'$. In this case, the parameter β_j is interpreted as the parameter contrast for the main effect of the factor x_j for $j = 1, \dots, 7$.

We can also consider models containing interaction effects. For example, if we want to estimate the two-factor interaction effect among x_1 and x_2 along with all the main effects, we may add the column

$$(1, 1, 1, 1, -1, -1, -1, -1, -1, -1, -1, -1, 1, 1, 1, 1)' \tag{14}$$

to the above A. Note that this corresponds to adding a factor in Section 5. In this case, β is 9-dimension and the element corresponding to the additional row (14) is interpreted as the parameter contrast for the two factor interaction effect among x_1 and x_2.

If we want to estimate another interaction, we have to consider the confounding relations. For example, two two-factor interaction effects among $x_1 \times x_2$ and $x_4 \times x_5$ cannot be estimated simultaneously since they are confounded. This confounding relation is shown in (12). In the terms of algebra, this confounding relation is shown as the ideal membership,

$$x_1x_2 - x_4x_5 \in \langle x_1^2 - 1, \ldots, x_7^2 - 1, x_1x_2x_4x_5 - 1, x_1x_3x_4x_6 - 1, x_2x_3x_4x_7 - 1\rangle,$$

as discussed in Section 3.

As our second example, we consider the case of three-level factors. We also indicate how complex coding simplifies the specification of the conditional sample space (11). Consider the following 3^{3-1} fractional factorial design, where the levels are coded as $\{0, 1, 2\}$.

	Factor			$\mathbf{y}$
Run	x_1	x_2	x_3	
1	0	0	0	y_1
2	0	1	2	y_2
3	0	2	1	y_3
4	1	0	2	y_4
5	1	1	1	y_5
6	1	2	0	y_6
7	2	0	1	y_7
8	2	1	0	y_8
9	2	2	2	y_9

In this case, each main and interaction factor has more than one degree of freedom and has to be parameterized by more than one parameters. For example, the main effect of x_1 can be expressed $(\alpha_1, \alpha_2, \alpha_3)$ with one constraint. One of the simplest constraints is $\alpha_3 = 0$, i.e., to treat x_3 as the baseline, which is expressed as the two columns

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}'$$

in the covariate matrix A. We can also consider a symmetric constraint $\alpha_1 + \alpha_2 + \alpha_3 = 0$. In this case, we may include the two columns

$$\begin{pmatrix} 2 & 2 & 2 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 & 2 & -1 & -1 & -1 \end{pmatrix}'$$

to A. Note that the conditional sample space (11) is invariant for the constraints since the total $\sum y_i$ is fixed. As for the interaction effects, see [4].

Another equivalent expression is given directly from the complex coding

	Factor			$\mathbf{y}$
Run	x_1	x_2	x_3	
1	1	1	1	y_1
2	1	ω	ω^2	y_2
3	1	ω^2	ω	y_3
4	ω	1	ω^2	y_4
5	ω	ω	ω	y_5
6	ω	ω^2	1	y_6
7	ω^2	1	ω	y_7
8	ω^2	ω	1	y_8
9	ω^2	ω^2	ω^2	y_9

where $\omega = \exp(2\pi i/3)$. If we allow A to be a complex matrix and consider the real and the complex parts, the conditional sample space (11) defined from

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & \omega & \omega^2 \\ 1 & 1 & \omega^2 & \omega \\ 1 & \omega & 1 & \omega^2 \\ 1 & \omega & \omega & \omega \\ 1 & \omega & \omega^2 & 1 \\ 1 & \omega^2 & 1 & \omega \\ 1 & \omega^2 & \omega & 1 \\ 1 & \omega^2 & \omega^2 & \omega^2 \end{pmatrix}$$

is also the same. This follows from the following basic fact of the theory of cyclotomic polynomials (e.g. Section 6.3 of [18]): for $q_1, q_2, q_3 \in \mathbb{Q}$

$$q_1 + q_2\omega + q_3\omega^2 = 0 \;\Leftrightarrow\; q_1 = q_2 = q_3.$$

The same result holds for s-level factors, where s is a prime number.

8 Some discussions

In this paper, we review several topics related to designed experiments and computational algebraic statistics. As we have stated, there are two works as the beginning of computational algebraic statistics, i.e., the work by Pistone and Wynn ([24]) and by Diaconis and Sturmfels ([13]). It is important to study whether a closer connection can be established between these two branches, i.e., the design ideal and the Markov basis (toric ideal). Our works [4] and [6] are motivated by this goal, although we do not yet have a result of some general nature.

The following argument suggests that there should be some general results relating these two branches. Recall the covariate matrix A in (13). For the main effect model, except for the first element 1, each row of A is just a point in the design and therefore A

itself can be considered as a design. In the theory of toric ideals, the set of rows of A is often called a configuration defining the toric ideal. Therefore for the main effect model, the design simultaneously defines the design ideal and the toric ideal. Note that this relation also holds for three-level factors (or more generally for prime number of levels), if the levels are coded by complex numbers as indicated at the end of Section 7.

If we include some interaction effects in the null model for two-level case, this corresponds to adding factors as in Section 5. Therefore again there is a very simple relation between the design (without added factors) and the configuration for the toric ideal (with added factors). This argument suggests that some algebraic properties of the design ideal should be reflected in algebraic properties of the toric ideal. This is an important topic for further research.

References

[1] W. W. Adams and P. Loustaunau (1994). *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI.

[2] S. Aoki (2010). Some optimal criteria of model-robustness for two-level non-regular fractional factorial designs. *Annals of the Institute of Statistical Mathematics*, to appear.

[3] S. Aoki and A. Takemura (2009a). Some characterizations of affinely full-dimensional factorial designs. *Journal of Statistical Planning and Inference*, **139**, 3525–3532.

[4] S. Aoki and A. Takemura (2009b). Markov basis for design of experiments with three-level factors. in *Algebraic and Geometric Methods in Statistics* (dedicated to Professor Giovanni Pistone on the occasion of his sixty-fifth birthday), edited by P. Gibilisco, E. Riccomagno, M. P. Rogantin and H. P. Wynn, Cambridge University Press, 225–238.

[5] S. Aoki and A. Takemura (2009c). Statistics and Gröbner bases — The origin and development of computational algebraic statistics, in *Selected Papers on Probability and Statistics*. American Mathematical Society Translations. Series 2, Volume 227, 125–145.

[6] S. Aoki and A. Takemura (2010). Markov chain Monte Carlo tests for designed experiments. *Journal of Statistical Planning and Inference*, **140**, 817–830.

[7] G. Barba (1933). Intorno al teorema di Hadamard sui determinanti a valore massimo, *Giorn. Mat. Battaglini*, **71**, 70–86.

[8] G. E. P. Box and J. S. Hunter (1961). The 2^{k-p} fractional factorial design. *Technometrics*, **3**, 311–351, 449–458.

[9] G. E. P. Box, W. C. Hunter and J. S. Hunter (1978). *Statistics for Experiments*, Wiley, New York.

[10] Cheng, C. S., Deng, L. Y. and Tang, B. (2002). Generalized minimum aberration and design efficiency for nonregular fractional factorial designs. *Statistica Sinica*, **12**, 991–1000.

[11] D. Cox, J. Little and D. O'Shea (2007). *Ideal, Varieties, and Algorithms*, 3rd edition. Springer, New York.

[12] L. Y. Deng and B. Tang (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs. *Statistica Sinica*, **9**, 1071–1082.

[13] P. Diaconis and B. Sturmfels (1998). Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, **26**, 363–397.

[14] H. Ehlich (1964). Determinantenabschätzungen für binäre Matrizen, *Math. Z.*, **83**, 123–132.

[15] R. Fontana, G. Pistone and M. P. Rogantin (2000). Classification of two-level factorial fractions. *Journal of Statistical Planning and Inference*, **87**, 149–172.

[16] F. Galetto, G. Pistone and M. P. Rogantin (2003). Confounding revisited with commutative computational algebra. *Journal of Statistical Planning and Inference*, **117**, 345–363.

[17] W. K. Hastings (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

[18] S. Lang (2002). *Algebra.* 3rd ed. Graduate Texts in Mathematics, 211. Springer, New York.

[19] P. McCullagh and J. A. Nelder (1989). *Generalized linear models.* 2nd ed. Chapman & Hall, London.

[20] R. Mukerjee and C. F. J. Wu (2006). *A Modern Theory of Factorial Designs.* Springer Series in Statistics.

[21] G. Pistone, E. Riccomagno and H. P. Wynn (2001). *Algebraic Statistics: Computational Commutative Algebra in Statistics.* Chapman & Hall, London.

[22] G. Pistone and M. P. Rogantin (2008a). Algebraic statistics of level codings for fractional factorial designs. *Journal of Statistical Planning and Inference*, **138**, 234–244.

[23] G. Pistone and M. P. Rogantin (2008b). Indicator function and complex coding for mixed fractional factorial designs. *Journal of Statistical Planning and Inference*, **138**, 787–802.

[24] G. Pistone and H. P. Wynn (1996). Generalised confounding with Gröbner bases. *Biometrika*, **83**, 653–666.

[25] B. Sturmfels (1996). *Gröbner Bases and Convex Polytopes.* American Mathematical Society, Providence, RI.

[26] W. Wojtas (1964). On Hadamard's inequality for the determinants of order non-divisible by 4, *Colloq. Math.*, **12**, 73–83.

[27] K. Q. Ye (2003). Indicator function and its application in two-level factorial designs. *Annals of Statistics*, **31**, 984–994.