The Application of Computational Algebraic Geometry to the Analysis of Designed Experiments: A Case Study

Tim Holliday[1], Giovanni Pistone[2], Eva Riccomagno[1] and Henry P. Wynn[1]

[1] Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[2] Dipartimento di Matematica, Politecnico di Torino, Turin, 10129, Italy

Summary

Computational algebraic geometry can be used to solve estimability/identifiability problems in the design of experiments. The key is to replace the design as a set of points by the polynomials whose solutions are the design points. The theory and application of Gröbner bases allows one to find a unique saturated model for each so-called monomial ordering of the independent factors. A case study in engine mapping is fully worked out and employs a simple plotting method for modelling.

Keywords: Computer algebra; Design and analysis of experiments; Identifiability

1 Introduction

The ideas of confounding and aliasing in experimental design can be said to be well understood in the context of the construction of symmetric and asymmetric factorial design. In these methods a number of factors $a, b, c, \ldots$ have a fixed number of levels $m_1, m_2, m_3, \ldots$ and Abelian group theory is used to generate orthogonal fractions giving a particular alias structure. For special situations such as screening experiments additional series of orthogonal fractions such as Plackett-Burman designs are available and more complex experimentation may employ response surface designs or optimum designs. Much of this is available via computer generation although careful 'by hand' construction is often employed to good effect.

None of these methods addresses the problem of identifiability for general polynomial models such as the following. Consider a two-factor design with independent variables (x_1, x_2) consisting of 5 points in arbitrary position in $\mathbb{R}^2$. What polynomial models are estimable? The general problem is to establish for a particular design the set of all estimable polynomial models. We shall return to this question in Section 4 in the light of the case study.

There are many situations where the actual experiment performed does not belong to a class for which the alias structure can easily be computed. For example observations may have been lost from a well structured experiment, the physical situations or geometry may prevent observation at certain points, sensors may be located in physically convenient positions, and so on.

In Pistone and Wynn (1996) the authors introduce ideas from algebraic geometry to tackle such situations in addition to recapturing the structure of classical designs. The paper uses the technology of Gröbner bases which is accessible via computational algebra packages such as Maple (Char, Geddes, Gonnet, Leong, Monogan, & Watt, 1991) and the more specialised package CoCoA (Capani, Niesi and Robbiano, 1995).

The key is to replace the natural notion of a design as a set of points in a design space (treatment combinations in classical terms) by the *polynomials whose solutions are the design points.* In algebraic geometry terms the points are a variety and the set of *all* polynomials passing through the points is an ideal. The methodology then consists, roughly, of factoring out the collection of all polynomial models by the experimental design ideal leaving a remainder which holds information about the estimable, or identifiable, models.

The purpose of this paper is to demonstrate the power of the method on a real case study in which the experimental design used is non-standard. We address the question: if we have such a design what models may we fit? The algebraic geometry method goes most of the way to answer this question.

2 Estimable models as remainders

Some motivation is provided by the one-dimensional case (see Pistone and Wynn, 1996). Let $y = p(x)$ be a polynomial interpolating the points (x_1, y_1), ..., (x_N, y_N) where the x_i's are all different. Let $d(x) = (x - x_1) \cdots (x - x_N)$ and write

$$p(x) = s(x)d(x) + r(x)$$

where $\deg(r(x)) < N$. Then $r(x)$ is the unique remainder on dividing $p(x)$ by $d(x)$ and is also the unique polynomial of degree $< N$ interpolating the points $y_i = p(x_i) = r(x_i)$ for $i = 1, \ldots, N$. The need for the computational algebraic geometry is because in more than one dimension the forms of $d(x)$, $r(x)$ and $s(x)$ are not in general unique. The idea of aliasing can be interpreted in the following way: two polynomial models are aliased when they have the same remainder.

Let $\mathbb{Q}[x_1, \ldots, x_m]$ be the set of all polynomials in the indeterminates $x_1, \ldots, x_m$ with rational coefficients. We may think of this as the class of all polynomial models in m factors linear on rational coefficients. We note at this point that the computational algebra packages work over the rational numbers because of the finiteness of computer models. In practice real numbers are satisfactorily approximated by rationals.

Polynomials of the form $x_1^{\alpha_1} \cdots x_m^{\alpha_m}$, with $\alpha_i \in \mathbb{Z}^+$, are called monomials or terms and play an important role in the theory. The polynomial ideal generated by $G = \{g_1, \ldots, g_k\}$ is the set of all polynomial combinations of the g_i's, that is

$$I = I(G) = < g_1, \ldots, g_k > = \left\{ \sum_{i=1}^{k} q_i g_i : \quad q_i \in \mathbb{Q}[x_1, \ldots, x_m] \right\}. \qquad (1)$$

The set G is called a basis of I. Notice that by equating to zero the g_i's a system of polynomial equations is obtained.

Let $D = \{(a_{i_1}, \ldots, a_{i_m}) : i = 1, \ldots, N\}$ be a design with N points with rational coordinates in m factors. Let us consider a system of equations whose solutions are exactly the design points. There are different ways to determine this set, see Pistone and Wynn (1996) and Caboara and Riccomagno (1997). We proceed as follows. The design ideal for the one point design variety of coordinates $(a_1, \ldots, a_m)$ is

$$< (x_1 - a_1), \cdots, (x_m - a_m) > .$$

The variety for a design with more points is the union of the single point varieties and the ideal is the intersection of the corresponding one-point ideals. All this says is that the set of polynomials through all the design points is the intersection of the larger sets through each individual point. We call $I(D)$ the design ideal generated by D. The procedure for intersecting ideals is implemented efficiently in CoCoA, a computer algebra package for commutative algebra. A less computationally effective but nonetheless useful implementation is possible in Maple (see Pistone and Wynn, 1996). In order to extend the division idea above to the case of m indeterminates we need to impose an order on the monomial terms.

Definition 1 *A monomial ordering is a total order, $\succ$, on the set of all monomials satisfying the following condition*

$$x_1^{\alpha_1}\dots x_m^{\alpha_m} \succ x_1^{\beta_1}\dots x_m^{\beta_m} \text{ implies } x_1^{\alpha_1+\gamma_1}\dots x_m^{\alpha_m+\gamma_m} \succ x_1^{\beta_1+\gamma_1}\dots x_m^{\beta_m+\gamma_m}$$

for all $(\alpha_1,\dots,\alpha_m)$, $(\beta_1,\dots,\beta_m)$ and $(\gamma_1,\dots,\gamma_m) \in \mathbb{Z}_+^m$.

In the case study we use reverse total degree ordering (`DegRevLex`). Given an initial order on the indeterminates $x_1 \succ \dots \succ x_m$ then in `DegRevLex` $x_1^{\alpha_1}\dots x_m^{\alpha_m} \succ x_1^{\beta_1}\dots x_m^{\beta_m}$ if its total degree (sum of the exponents) is greater or its total degree is equal and it precedes $x_1^{\beta_1}\dots x_m^{\beta_m}$ in the inverse lexicographic order. Thus when `DegRevLex` finds the same total degree it looks at the smallest variable in the initial ordering and favors the smaller power. Not only is `DegRevLex` the most efficient ordering for many computations but also it is a sensible ordering for response surface studies as it favors lower total degree terms over higher total degree terms.

The choice of initial ordering should probably be made on the basis of scientific knowledge or prior information. In this paper, however, we are able to consider all initial orderings systematically because of the low dimension of the factor space.

In the case study below we shall have $m = 4$ and name the factors e, l, r, a. With the initial ordering $e \succ l \succ r \succ a$ we have, for example, $a^5 r \succ e^3 l$, $e^4 a^3 \succ e^3 a^4$ and $e^2 lra \succ a^2 elr$ with respect to the `DegRevLex` ordering.

As mentioned an ordering allows us to select the leading term, $LT(f)$ of the polynomial f. This leading term is simply the largest monomial in f with respect to the ordering.

The main tool in this theory is the so-called Gröbner basis. An ideal I, which will here typically arise from the experimental design, may be generated by various different bases. A set of generators $G = \{g_1,\dots,g_k\}$ is a Gröbner basis for I if and only if it meets the following condition

$$< LT(f) : f \in I >=< LT(g_i) : \ g_i \in G >$$

namely the ideal generated by the leading monomials of G is the ideal generated by the leading monomials of I. See Section 3.1 for examples of Gröbner bases. The available packages for Gröbner basis computations (like CoCoA and Maple) usually return reduced Gröbner bases. A Gröbner basis is called reduced when any polynomial in the basis and any monomial in such a polynomial is necessary in that it cannot be reduced without destroying the ideal:

Definition 2 *A reduced Gröbner basis is a Gröbner basis such that*
(i) the coefficient of $LT(g)$ is 1 for all $g \in G$,
(ii) for all $g \in G$, no monomial of g lies in $< LT(G - \{g\}) >$.

Note that Gröbner bases with respect to different monomial orderings may differ. The following theorems are proved in, for example, Cox, Little, O'Shea (1996) Chapter 2.

Theorem 1 *Given a monomial ordering, any proper ideal of* $\mathbb{Q}[x_1, \ldots, x_m]$ *has a unique reduced Gröbner basis and there is an algorithm, the Buchberger's algorithm to calculate it.*

The main result we need from the Gröbner basis theory is that the remainder is unique under division by a Gröbner basis.

Theorem 2 *Let* $I \subset \mathbb{Q}[x_1, \ldots, x_m]$ *be an ideal,* $G = \{g_1, \ldots, g_k\}$ *a Gröbner basis for* I *and* $f \in \mathbb{Q}[x_1, \ldots, x_m]$. *Then there exist unique* $r, q_i \in \mathbb{Q}[x_1, \ldots, x_m]$ *(*$i = 1, \ldots, k$*) such that*
(i) $f = \sum_{i=1}^{k} q_i g_i + r$,
(ii) no term of r *is divisible by any of* $LT(g_1), \ldots, LT(g_k)$.

The quotient field $\mathbb{Q}[x_1, \ldots, x_m]/I$ is isomorphic as a $\mathbb{Q}$-vector space to the set of remainders r, the isomorphism being given by the division algorithm. We are interested in the set $r(x)$ when the ideal is that arising from a particular experimental design: $\mathbb{Q}[x_1, \ldots, x_m]/I(D)$. It represents the set of estimable models given by the design together with the particular term-ordering.

Theorem 3 *If* N *is the number of design points in a design* D *then the dimension of* $\mathbb{Q}[x_1, \ldots, x_m]/I(D)$ *as a* $\mathbb{Q}$*-vector space is* N *and it has a vector-space basis spanned by a set of* N *monomials which depends uniquely on* D *and the monomial ordering.*

The algorithm we use in the case study returns a monomial basis of $\mathbb{Q}[x_1, \ldots, x_m]/I(D)$. It is given by the set of all monomials not divisible by any leading term of the reduced Gröbner basis of $I(D)$ with respect to the chosen ordering. The basis is exactly the set of monomials identifiable, that is estimable, using the design.

Furthermore let f be a polynomial model and $G = \{g_1, \ldots, g_k\}$ be the reduced Gröbner basis for $I(D)$ and let r be the remainder of f with respect to G, then the confounding structure of f under D is read in the monomials of r in that $f = r$ at the design points and any polynomial p such that $p = f$ at the design points has r as remainder with respect to G.

3 Case Study

This case study is based on an existing experiment. The purpose of the paper is to exhibit the Gröbner basis method rather than give a critique of the experimental design itself.

Engine mapping is the term used in the automotive industry for the process of developing models to describe and predict engine behaviour. In its simplest form it consists of collecting data and fitting models, and is largely empirical. One application for such models is the calibration of memory based electronic engine control systems. When developing the models for such an application a typical engine mapping experiment will consist of varying five factors: engine load, engine speed, spark advance, air-fuel ratio and exhaust gas recycling while observing four responses: the power output of the engine and the emissions of the three exhaust products covered by legislation. The experiment usually takes the form of a repeated measures experiment; four of the factors are held constant while spark advance is varied, the data collected are known as a spark sweep. Many spark sweeps are collected at different levels of the other four factors. Holliday et al. (1997) describe the advantages of using a two-stage regression approach to this type of engine mapping data.

In this paper we consider only one response: carbon monoxide emissions. This response has the advantage that spark advance only has a very minor effect on it and therefore little is lost by treating the mean level of carbon monoxide emissions for each spark sweep as the response. Carbon monoxide emissions are affected by engine load, engine speed and the air-fuel ratio but exhaust gas recycling, which is used to control the emission of nitrous oxides, has no significant effect on carbon monoxide emissions (Heywood, 1988, Chapter 15). Draper et al (1994) use carbon monoxide data from an engine mapping experiment where the data were not collected as spark sweeps as a case study. They found a reasonable model for carbon monoxide for their data set to contain the linear effects of engine load, engine speed and air-fuel ratio.

The experiment described was an initial attempt at using designed experiment based on spark sweeps for engine mapping. When developing engine mapping models for electronic engine control any interactions between air-fuel ratio and exhaust gas recycling are unimportant because of the constraints on the strategy employed by the engine controller. Therefore one is only interested in the effect of exhaust gas recycling when the air-fuel ratio is stoichiometric ($A = 1$ in our experiment) and only in the air-fuel ratio effect when exhaust gas is not being recycled ($E = -1$ in our experiment). For simplicity the design used was the union of two 3^3 designs, which allows estimation of the full quadratic model excluding those terms featuring air-fuel ratio and exhaust gas recycling.

In the event the experiment was stopped because of a combination of a lack of resources and a difficulty in obtaining some of the experimental data. As this was the first attempt at using designed experiments for spark sweep based engine mapping data there was no precedent to follow. Normally when engine mapping designed experiments are not used, the analysis takes the form of a linear regression where the model is fixed a priori.

To summarise the intended experiment was a 45 run experiment in the four factors r, l, e, a in which only 23 runs were completed.

3.1 Identifiability Procedure

The ideal corresponding to the *intended* 45 run experiment is the ideal corresponding to a standard 3^4 experiment

$$\begin{array}{r} a(a^2-1)=0 \\ e(e^2-1)=0 \\ l(l^2-1)=0 \\ r(r^2-1)=0 \end{array} \tag{2}$$

with the additional equation

$$(a-1)(e+1)=0 \tag{3}$$

to give points at level $a=1$ *or* at level $e=-1$.

Note that we have used lower case letters throughout. The design points used in the intended experiment were precisely the solutions to Equations (2) and (3). It may be helpful to try to translate these back to traditional group theoretic methods. The set of Equations (2) would become

$$A^3 \equiv I, E^3 \equiv I, L^3 \equiv I, R^3 \equiv I$$

to represent the full 3^4 factorial. The last equation, (3), however is harder. It says take one block from the experiment formed by confounding $A \equiv I$ and union this with the block from $E \equiv -I$. Trying to write $A \equiv I$ and $E \not\equiv -I$ then produces a confusion between the group theory and its numerical realisation. For more complex experiments it becomes infeasible to retain the group structure and the traditional aliasing method breaks down. We will return to this point at the end of this section.

The identifiability procedure shows that this experiment allows the estimation of all the terms that can be estimated from a 3^4 experiment except those involving ae interactions. First we input the design ideal

$$ea-e+a-1, a^3-a, e^3-e, l^3-l, r^3-r.$$

Then we considered its leading terms

$$ea, a^3, e^3, l^3, r^3.$$

Finally following Theorem 2 we calculated all the monomials which are not divisible by any of the leading terms in the above line (for example using the function `OrderId` whose code can be found in Appendix 6). There are 45 terms as expected from Theorem 3

$$\begin{array}{l} r^2l^2e^2, r^2l^2a^2, r^2l^2e, r^2le^2, rl^2e^2, r^2l^2a, r^2la^2, rl^2a^2, \\ r^2l^2, r^2le, rl^2e, r^2e^2, rle^2, l^2e^2, r^2la, rl^2a, r^2a^2, rla^2, l^2a^2, \\ r^2l, rl^2, r^2e, rle, l^2e, re^2, le^2, r^2a, rla, l^2a, ra^2, la^2, \\ r^2, rl, l^2, re, le, e^2, ra, la, a^2, \\ r, l, e, a, 1. \end{array}$$

The symmetry inherent in the 45-run design leads to a simple form of the design ideal. Because of this, whatever initial ordering we consider the reduced Gröbner basis of the design ideal and the leading terms are always the same (in algebraic terms we talk of a total Gröbner basis). In particular the estimable terms do not depend on the choice of the initial ordering. Indeed, the estimable model of 45 terms corresponding to this list could conceivably have been found by hand.

For the actual experiment, indicated by the 23 black dots in Figure 1, the lack of symmetry in the design leads to a more complex design ideal and the Gröbner bases for the design ideal depend on the initial ordering of the factors. Computer algebra then enables the construction of the ideal for the whole experiment. The functions we used are in Appendices 6 and 7. For the initial ordering $r \succ l \succ a \succ e$ and the `DegRevLex` monomial ordering we obtained the following Gröbner basis for the design ideal

$$\begin{aligned}
&ae + a - e - 1,\\
&e^3 - e,\\
&re^2 + re + e^2 + e,\\
&rle + le^2 + rl + le,\\
&r^2e + l^2e - le^2 + r^2 + l^2 - le - e - 1,\\
&a^3 - a,\\
&l^3 - l,\\
&r^3 - r,\\
&r^2la + rla,\\
&r^2l^2 + r^2l + rl^2 + rl,\\
&l^2e^2 + l^2e - le^2 - le,\\
&l^2a^2 - 5/2r^2l - 3/2rl^2 + 1/2r^2a + 1/2rla - l^2a + 1/2ra^2 + 3/2l^2e - 1/2r^2\\
&\qquad -3/2rl - 1/2ra + 1/2la - 1/2a^2 - 3/2le - 1/2l + 1/2,\\
&rla^2 + 2r^2l + rl^2 - rla + la^2 - l^2e + rl - la + le,\\
&r^2a^2 - r^2l - rl^2 + rla + l^2e - r^2 - rl + la - a^2 - le - l + 1,\\
&rl^2a + 2r^2l + rl^2 - rla + l^2a - 2l^2e + rl - l^2 - la + 2le + l.
\end{aligned} \tag{4}$$

The leading terms of the design ideal are

$$ae, e^3, re^2, rle, r^2e, a^3, l^3, r^3, r^2la, r^2l^2, l^2e^2, l^2a^2, rla^2, r^2a^2, rl^2a.$$

The list of 23 estimable terms obtained is

$$r^2l, rl^2, r^2a, rla, l^2a, ra^2, la^2, l^2e, le^2, r^2, rl, l^2, ra, la, a^2, re, le, e^2, r, l, a, e, 1.$$

In modelling situations such as the present one lower order terms are to be preferred to higher order terms, thus we may naturally use the reverse total degree ordering. Then we only need to decide the initial ordering of the factors to determine the total ordering. There are therefore $24 = 4!$ possible initial orderings corresponding to the 24 permutations of the factors.

Repeating the above calculation we find that all 24 initial orderings give broadly similar results, as shown in Table 2. The intercept, all first order terms, all second order terms except ae and all third order interaction terms in a, l and r are identifiable, independent of the ordering. Also two terms from elr, er^2, e^2l, and el^2 will be identifiable depending on the chosen order of the factors. No third order terms involving a and e appear and neither does re^2. In practice the effect of any third order interaction involving e is expected to be so small that the final choice of order is unimportant.

Notice in this example that the terms involving the full factorial and the original fraction actually occur in the Gröbner basis of the design ideal. This does not always happen but does here because of the symmetry of the full factorial and the particular choice we made of the monomial ordering. Fontana, Pistone and Rogantin (1997) explain in detail the application to classical factorial design.

Algebraically speaking the estimable terms we obtain from the procedure described above form a vector space basis for a space of equivalent classes. Each of these classes contains the polynomials that have same remainder under the division by the design ideal Gröbner basis and they can be considered as the set of aliased terms under that design and the monomial ordering used. Thus we can estimate one and only one element from each class. In particular the Gröbner basis is a direct generalisation of the notion of confounding arising in classical factorial design.

For example let us consider the first polynomial of the Gröbner basis calculated above: $ae + a - e - 1$. The remainder when its leading term ae is divided by the Gröbner basis itself is $-a + e + 1$. In other words

$$ae = -a + e + 1 \quad \text{modulo the design ideal}$$

In statistical terms the interaction ae is aliased with $-a + e + 1$ which is another linear combination of estimable terms (see the above discussion).

Thus we can see this as a generalisation of the notion of defining relations and use it to build a generalised alias table in the following sense. Each row of the table is labeled by a different estimable term. In this example we have 23 rows. Some information on the rows is obtainable from the Gröbner bases, for example

$$\begin{aligned} 1 &= -a + e - ae = r^2e + l^2e - le^2 + r^2 + l^2 - le - e \\ &= -(r^2a^2 - r^2l - rl^2 + rla + l^2e - r^2 - rl + la - a^2 - le - l)\ldots \end{aligned}$$

coming from rows $1, 5$ and 22 of the ideal 4. Similarly we have the following

$$\begin{aligned} e &= e^3 = -(re^2 + re + e^2)\ldots \\ a &= a^3\ldots \end{aligned}$$

$$\vdots$$

where the equalities are modulo the design ideal. In each row there is an infinite number of equalities because each equivalence class contains terms of arbitrary high order. Moreover if we multiply a row by a polynomial and divide with respect to the Gröbner bases, then we still get a row of the table since the operation of multiplication by polynomials is compatible with the operations of the quotient space. That is modulo the design ideal the equality

$$r = s \qquad \text{implies} \qquad pr = ps$$

for any $r, s, p \in \mathbb{Q}[x_1, \ldots, x_m]$.

Notice that at the end of Section 2 we used the same fact differently. Since the intended experiment was a full factorial from which a standard quadratic model can be estimated, it may be of interest to determine which model estimable by the design is aliased with the standard quadratic model, i.e. what is the structure of the coefficients of the remainder of the division of the standard quadratic polynomial model by the Gröbner bases for the design ideal, given a monomial ordering. The number of indeterminates involved in this division is 85 (variables plus coefficients) and the remainder, which we omit, is 5 pages long.

3.2 Modelling

We assume the standard Normal linear model, that is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}). \tag{5}$$

Following standard practice, logs of the initial responses were used for modelling exhaust emissions; see Draper et al (1994).

A property of the algebraic procedure for identifiability is that if any monomial term is identifiable then so are all its divisors. A corollary is that every model must include the intercept. This reflects the common practice of including all lower order terms of a given monomial in modelling. If we only consider models for which the divisibility criteria holds then we reduce the number of possible models for any given ordering of the factors to around 7000 compared to $2^{23} - 1$, around 8 million, models if this condition is not applied. These models can be quickly generated and fitted to the data, for any ordering.

When analysing data from an engineering experiment it is desirable, whenever possible, to base the model selection on graphical methods, by using Daniel's plots, for example. The analysis and modelling decisions are then more accessible to the engineers. Due to the non-orthogonality of this experiment a Daniel's plot is not possible and we have adopted a graphical approach, similar to Mallow's C_p.

For any model which provides a good fit to the data, that is one which includes all significant terms, the residual sum of squares (RSS) will be $\sigma^2\chi^2_{df}$ where $df = 23 - p$ and $p = \text{rank}(\boldsymbol{X})$ is the dimension of the fitted model. Note that if we use a submodel from one of the remainder terms r, X has full rank. If the fitted model does not include all significant terms then the RSS will be inflated by the square of the bias introduced by omitting the relevant terms. Therefore if we calculate the RSS for each model and plot them against df then the models which include all the significant terms will be distributed around a line with gradient σ^2, similarly models which do not include important terms will lie above this line.

Figure 2 shows the RSS plotted against the degrees of freedom, df, for the 7,307 estimable models for the ordering $r > l > a > e$. As discussed before the initial ordering makes little difference. The boxes in Figure 2 indicate which linear terms are present in those models and we deduce that any reasonable model must include, at the very least, the linear terms for r, l and a.

Figure 3 plots the residual sum of squares against df for the 6,915 models which include these first-order terms. On this plot we have reduced the number of points as follows. For any value of df for which we have more than 10 models we obtained an estimate $\tilde{\sigma}^2_{df}$ of σ^2, based on the models with $\text{rank}(\boldsymbol{X}) = p$, by dividing the median of the RSS by $\chi^2_{df,.5}$. Then at each value of df we have plotted $\tilde{\sigma}^2_{df}$ as a cross and added whiskers between $\tilde{\sigma}^2_{df}\chi^2_{df,.05}$ and $\tilde{\sigma}^2_{df}\chi^2_{df,.95}$. We have plotted the RSS which lie outside these limits and at values of df where we have no more than 10 distinct models to judge the value of σ^2.

If the first-order terms r, l and a were the only linear significant terms then we would expect the estimates, $\tilde{\sigma}^2_{df}$ to lie on a straight line and for only a few points to be plotted outside the whiskers. Figure 3 shows that the first-order model in r, l and a is not adequate, there is an obvious kink around $df = 8$. This is because some significant low order terms are not included, the majority of smaller models will not include these terms and therefore the RSS will be inflated. However the larger models, lower df, are more likely to include the significant lower terms so the RSS will be lower.

We increase the model sequentially by looking at the model which has smallest RSS among models of a given size. Figure 4, which plots the RSS for the 3570 models including the terms: r, l, a, ra, a^2, l^2, is the first instance of an approximate straight line and therefore a reasonable model has terms

$$1, r, l, a, ra, a^2, l^2.$$

One could argue that a better model might be obtained by looking at the next model, which would include the r^2 term, but in the event this model was not needed because of engineering considerations. The result confirmed the prior engineering information that exhaust gas recycling should be unimportant, air-fuel ratio should be the most important and should have a nonlinear effect although this does not show itself as the range of the air-fuel ratio in our experiment is limited, as is the use of air-fuel ratio in modern vehicles, by the use of catalytic converters. The final model is similar to that obtained by using a two-stage analysis of the completed 3^3 experiment in Holliday (1997) and to that in Draper et al (1994).

4 Discussion: the fan of an experimental design

The set of all models generated by the algebraic method as we range over all monomials is a very rich class of models. It is named the *fan* of an experimental design in Caboara, Pistone, Riccomagno and Wynn (1997). It does not in general cover all polynomial models subject to the divisibility conditions; counterexamples can be found. However it forms the starting point for a characterisation of all such models which is likely to be completed by generalising the theory of Gröbner bases via the use of partial rather than total monomial orderings. In addition, in Caboara, Pistone, Riccomagno and Wynn (1997) it led to the introduction of a maximal fan design in which all saturated models subject to the divisibility condition are estimable.

It can be argued that since designs with a larger fan have a wider portfolio of estimable models they are preferable in some modelling context to classical full factorial designs (whose fans typically have only a single saturated model). Related to this is that the present paper only really discusses the forward problem: given a particular design deduce a class of models that are estimable with that design. The so-called "inverse problem" which is to find a suitable, or optimal design, for a given model is not directly discussed here. However progress has been made using the algebraic method of this paper by Fontana, Pistone and Rogantin (1997) and Caboara and Robbiano (1997).

5 Acknowledgments

The authors wish to thank the CoCoA group in Genova, Italy for their help with the use of the package. The last two authors acknowledge the support of the UK Engineering and Physical Science Research Council. The case study was also supported by EPSRC as part of a CASE award in collaboration with Ford Motor Co. Ltd.

References

[1] Caboara, M. & Riccomagno, E. (1997) An algebraic computational approach to the identifiability of Fourier models. Journal of Symbolic Computation (Under review).

[2] Caboara, M., Pistone, G., Riccomagno, E. & Wynn, H.P. The Fan of an Experimental Design. The Annals of Statistics (Submitted).

[3] Caboara, M. & Robbiano, L. (1997). Families of ideals in statistics. Proceedings of the ISSAC 97 (Maui, Hawaii, July 97) Küchlin Ed., New York, 404–117.

[4] Capani, A., Niesi, G. & Robbiano, L. (1995) CoCoA*, a system for doing Computations in Commutative Algebra.* Available via anonymous ftp from `lancelot.dima.unige.it`. See also `http://cocoa.dima.unige.it/`.

[5] Char, B., Geddes, K., Gonnet, G., Leong, B., Monogan, M & Watt, S. (1991) *MAPLE V Library Reference Manual.* Springer-Verlag New York.

[6] Cox, D., Little, J. & O'Shea, D. (1996) *Ideal, Varieties, and Algorithms*, Springer-Verlag, New York. (Second edition).

[7] Draper, N.T., Pozueta, L., Davis, T.P., & Grove, D.M. (1994) Isolation of degrees of freedom for Box-Behnken designs. Technometrics, Vol.36, No.3, pp.283-291.

[8] Fontana, R., Pistone, G. & Rogantin, M-P. (1997). Algebraic analysis and generation of two-level designs. Statistica Applicata. 9,1 (In press).

[9] Heywood, J.B. (1988) *Internal Combustion Engine Fundamentals.* McGraw-Hill.

[10] Holliday, T. (1996) *The Design and Analysis of Engine Mapping Experiments*, PhD Thesis, University of Birmingham.

[11] Holliday, T., Lawrance, A.J. & Davis, T.P. (1997). Engine mapping: a two-stage regression approach. Technometrics (to appear).

[12] Pistone, G. & Wynn, H.P. (1996) Generalised Confounding with Gröbner Bases, Biometrika 83,3:653–666.

[13] Sturmfels B. *Gröbner Bases and Convex polytopes.* Ed. AMS, Providence R I, 1995.

6 CoCoA code

```
// Input:  a list with as many element as indeterminates.
// Output:  the reduced GB of the single point design ideal.
// N.B.:  the monomial ordering is specified at the start of the session.
Define OnePointIdeal(P)
       If Len(P) <> NumIndets Then
             Return Error('Wrong number of indeterminates')
       End;
       F:=1;
       While F<= NumIndets And P[F]=0 Do
             F:=F+1;
       End;
       L:= [ Indet(K) - P[K] | K In 1..NumIndets];
       Return Ideal(L);
End;

// Input:  a list of list (list of points = design).
// Output:  the reduced GB of the ideal generated by that design.
// N.B.:  the monomial ordering is specified at the start of the session.
Define DesignIdeal(L)
       I:OnePointIdeal(Head(L));
       Foreach P In Tail(L) Do
             Catch J:=OnePointIdeal(P) In E End;
       If Type(E) = ERROR Then Print P, ':'; Return E End;
       I:=Intersection(I,J)
       End;
       Return I
End;

// Input:  a list of monomials (leading terms) and a positive integer.
// Output:  the list of monomials not divisible by any element in the input list.
// N.B.:  M is a positive integer.  If M is too big then the function is slow.
//       If M is too small then not all the identifiable terms are captured.
Define OrderId(Lm,M);
       Model:=[1];
       For A:=1 To M Do
              Model:=Concat( Model, Monomials(DensePoly(A)) ) End;
       C:=NewList(Len(Lm),0);
       O:=[ X In Model | [Div(X,Y) | Y In Lm] = C ];
       Return O;
End;
```

7 Maple code

```
> ## Maple procedure for the computation of design ideals
> ## Input: N is the number of design points;
> ##            d is the dimension;
> ##            A is a list whose elements are the components of the design points
> ##            listed as a long list point after point
> ##            ordering is either plex or tdeg
> ## Output: is the ordering-reduced Grobner basis and the leading terms
> ## Note that the initial ordering is that in which the components of
> the design points are given.
> ## Note also that the actual estimable terms are to be obtained by hand.
> ## For big designs Grobner basis computation in Maple is not very efficient.
```

```
Design :=

    proc(N,d,A,ordering)
    local j,i,Id,l,V,var,GB,DesId,LT;
        V := -N;
        l := 1;
        var := [];
        Id := [];
        for j to d do
            var := [op(var),x[j]];
            for i to N do
                Id := [op(Id),t[j]*(x[i]-A[l])]; l := l+1; V := V+t[i]
            od
        od;
        Id := [op(Id),V];
        for i to N do  var := [t[i],op(var)] od;
        GB := gbasis(Id,var,plex);
        DesId := [seq(GB[i],i = N+1 .. nops(GB))];
        var := [seq(var[i],i = N+1 .. N+d)];
        if ordering = 'tdeg' then
            DesId := gbasis(DesId,var,tdeg);
            LT := [seq(leadmon(DesId[i],var,tdeg)[2],i = 1 .. nops(DesId))];
            RETURN(DesId,LT)
        else
            DesId := gbasis(DesId,var,plex);
            LT := [seq(leadmon(DesId[i],var,plex)[2],i = 1 .. nops(DesId))];
            RETURN(DesId,LT)
        fi
    end
```

r	l	a	e
-1	0	1	1
-1	1	1	1
-1	0	1	0
0	1	1	0
0	-1	1	0
1	0	1	0
1	0	1	-1
0	1	1	-1
0	0	1	-1
-1	-1	1	-1
-1	0	1	-1
-1	1	1	-1
1	-1	0	-1
1	0	0	-1
0	1	0	-1
-1	-1	0	-1
-1	0	0	-1
-1	1	0	-1
1	0	-1	-1
0	-1	-1	-1
0	0	-1	-1
-1	-1	-1	-1
-1	1	-1	-1

Table 1: 23 run experiment

	ea^2	ae^2	al^2	la^2	le^2	el^2	re^2	er^2	ra^2	ar^2	rl^2	lr^2	alr	ael	aer	elr
aelr			*	*				*	*	*	*	*	*			*
aerl			*	*		*			*	*	*	*	*			*
aler			*	*				*	*	*	*	*	*			*
alre			*	*	*			*	*	*	*	*	*			
arel			*	*	*	*			*	*	*	*	*			
arle			*	*	*	*			*	*	*	*	*			
ealr			*	*				*	*	*	*	*	*			*
earl			*	*		*			*	*	*	*	*			*
elar			*	*				*	*	*	*	*	*			*
elra			*	*				*	*	*	*	*	*			*
eral			*	*		*			*	*	*	*	*			*
erla			*	*		*			*	*	*	*	*			*
laer			*	*				*	*	*	*	*	*			*
lare			*	*	*			*	*	*	*	*	*			
lear			*	*				*	*	*	*	*	*			*
lera			*	*				*	*	*	*	*	*			*
lrae			*	*	*			*	*	*	*	*	*			
lrea			*	*	*			*	*	*	*	*	*			
rael			*	*	*	*			*	*	*	*	*			
rale			*	*	*	*			*	*	*	*	*			
real			*	*	*	*			*	*	*	*	*			
rela			*	*	*	*			*	*	*	*	*			
rlae			*	*	*	*			*	*	*	*	*			
rlea			*	*	*	*			*	*	*	*	*			

Table 2: The degree 3 part of the 24 estimable models found by `DegRevLex`

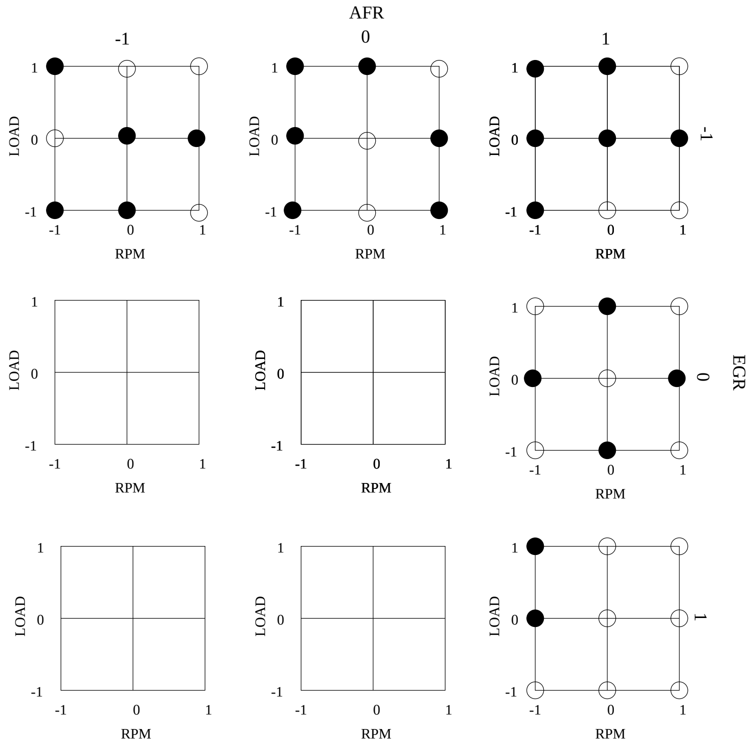

Figure 1: Bidimensional representation of the 3^4 full factorial. The original designed experiment includes the empty circles; completed runs are given by black dots.

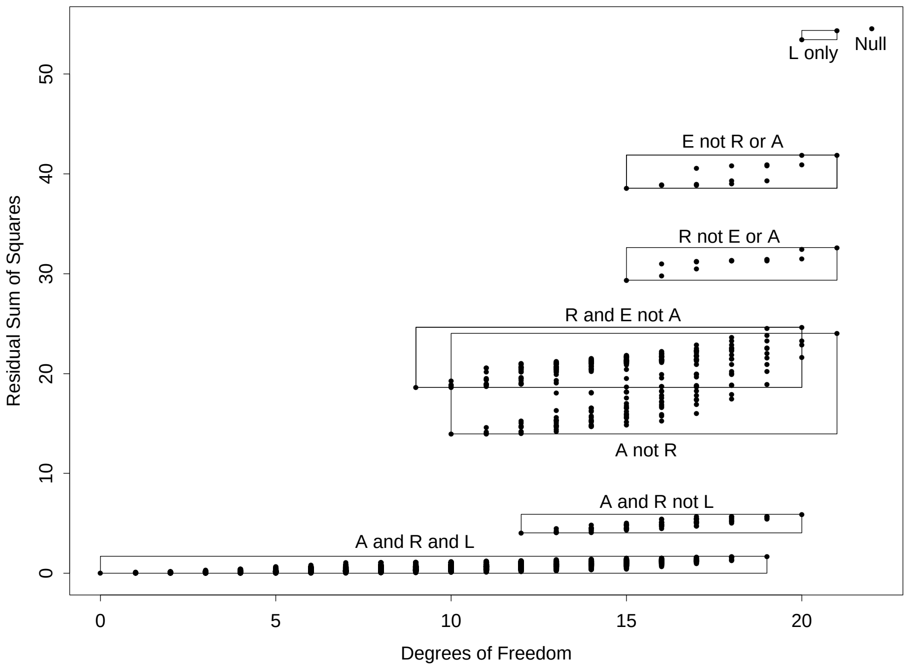

Figure 2: Residual sum of squares against degrees of freedom for all estimable models under the ordering $r > l > a > e$.

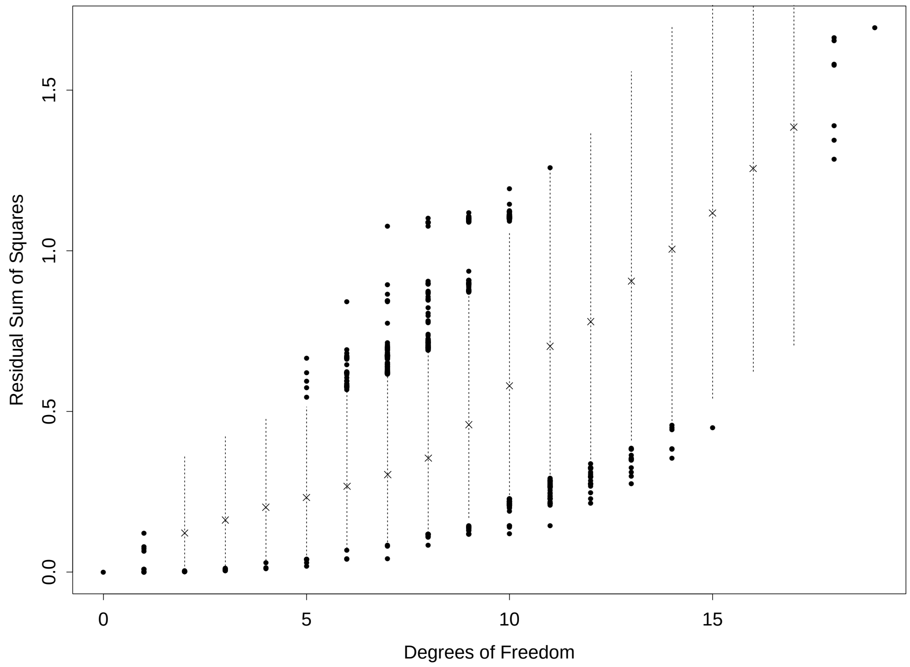

Figure 3: Residual sum of squares against degrees of freedom for all estimable models including the linear terms, r, l and a.

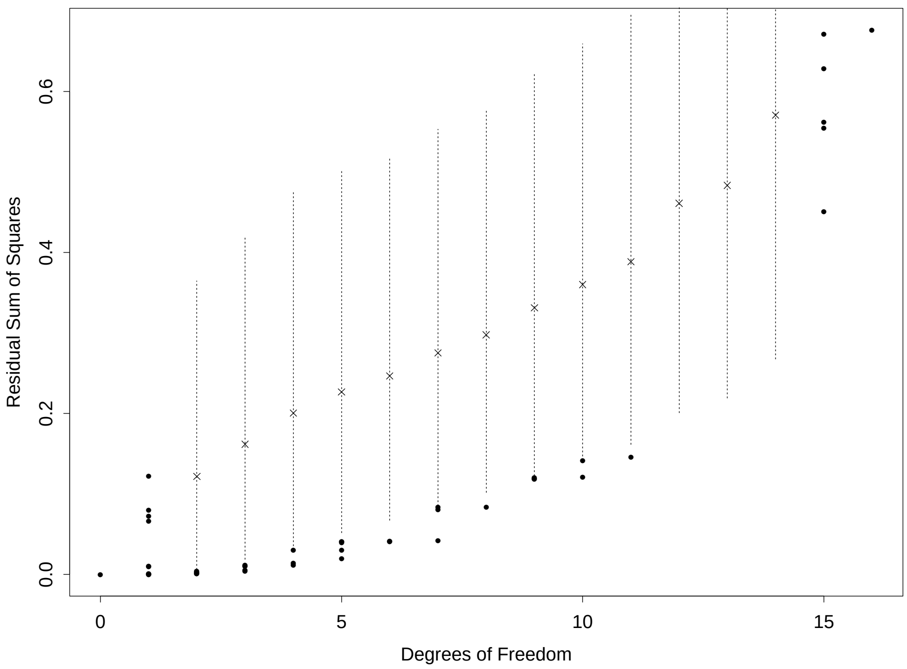

Figure 4: Residual sum of squares against degrees of freedom for all estimable models including the terms: r, l, a, ra, a^2 and l^2.