Replications with Gröbner Bases

A.M. Cohen
A. Di Bucchianico
E. Riccomagno

ABSTRACT: We present an extension of the Gröbner basis method for experimental design introduced in Pistone and Wynn (1996) to designs with replicates. This extension is presented in an abstract regression analysis framework, based on direct computations with functions and inner products. Explicit examples are presented to illustrate our approach.

KEYWORDS: Gröbner basis; replicates; orthonormalisation; projection

1 Introduction

Recently tools from algebraic geometry have been introduced in experimental design. See Pistone and Wynn (1996), Pistone, Riccomagno and Wynn (2000) and Riccomagno (1997). They are particularly useful in the analysis of complex experiments where there is a large number of factors and runs and the structure of the design is not regular, for example there are missing observations from a standard full factorial experiment. Confounding relations among factors and interactions are encoded in the Gröbner bases associated with a design allowing us to interpret confounding relations of the kind $I = AB$ (where A and B are factors and I is the constant term) for a large class of designs and models.

A major requirement for the application of this technique is that the design has no replicates. However, there are several practical situations where replicates are useful. In the present work we extend the algebraic methods to designs with replicates.

The main idea is to introduce a new variable that counts how many times a point appears in the design. For example, the one-dimensional design with five points $\mathcal{D}^* = \{0,0,1,1,2\}$ becomes the two-dimensional object $\mathcal{D} = \{(0,1),(0,2),(1,1),(1,2),(2,1)\}$. This is encoded in the following set of polynomials in two indeterminates

$$x^3 - 3x^2 + 2x, \quad x^2t - xt - x^2 + x, \quad t^2 - 3t + 2,$$

where x represents the design factor and t counts the number of replicates.

The polynomials above can be used to construct a polynomial system of equations whose zeros are the points in $\mathcal{D}$. The zeros of the first polynomial, involving only the x indeterminates, are the distinct points in $\mathcal{D}^*$.
Least squares models are fitted to the data with replications as polynomial interpolators using the Gröbner basis method. In order to accommodate this process, we present a vector space setting for regression analysis in terms of functions on the design points. We suggest to perform estimation after orthonormalisation of the model terms (see also Giglio et al., 2000). The traditional sums of squares appear naturally as the lengths of the terms in the orthonormalised model. The coefficients from the non-orthonormalised model are obtained simply by comparing coefficients. A pleasant feature from the computational point of view is that to compute regression coefficients, we do not need to perform matrix inversion as in the standard matrix way of computing regression coefficients. We present several explicit examples to illustrate our method.

2 Basic setup

We start by fixing notation. A design without replicates is a finite subset of $\mathbf{R}^d$. The main idea behind the algebraic geometry approach to experimental design is to view a design as a variety, i.e. the set of common zeroes of a finite set of polynomials. Statistical analysis of data starts with finding a polynomial that interpolates the data at the design points. If points of a design are replicated, then strictly speaking we are dealing with multi-sets rather than ordinary sets. This causes problems for the algebraic geometric approach. Namely there is no polynomial (function) that takes different values at the same point. We overcome this difficulty by introducing an extra variable that counts how many times a point appears in the design

Definition 2.1 *A design $\mathcal{D}$ with replicates is a finite set of points in $\mathbf{R}^d \times \mathcal{L}$, where $\mathcal{L}$ is a finite ordered set (the label set). The associated unreplicated design $\mathcal{D}^*$ is defined by $\mathcal{D}^* = \{a^* \in \mathbf{R}^d \mid \exists \ell \in \mathcal{L} \text{ such that} (a^*, \ell) \in \mathcal{D}\}$. Each element a of $\mathcal{D}$ is of the form $a = (a^*, \ell)$. Thus we may alternatively define $\mathcal{D}^*$ as $\mathcal{D}^* := \{a^* \mid a \in \mathcal{D}\}$.*

Designs without replicates are special designs such that for each $a^* \in \mathcal{D}^*$ there is exactly one $\ell \in \mathcal{L}$ such that $(a^*, \ell) \in \mathcal{D}$. Two designs are isomorphic if their associated unreplicated designs coincide and there is a bijection between the two designs. In general, the unreplicated design $\mathcal{D}^*$ is obtained by projecting $\mathcal{D}$ onto the first d factors. The operation of projection does not take into account the number of replicates. It has a nice algebraic counterpart (see Theorem 4.3 below).

Notation 2.2 *Let $\mathcal{D} \subset \mathbf{R}^d \times \mathcal{L}$ be a design. The set of real-valued functions on $\mathcal{D}$ is denoted by $\mathcal{L}(\mathcal{D})$.*

The inner product on $\mathcal{L}(\mathcal{D})$ given in Definition 2.3 below is directly related to least squares estimation.

Definition 2.3 *If $\mathcal{D}$ is a design, then for all $f, g \in \mathcal{L}(\mathcal{D})$ we define an inner product by $\langle f\,,\,g\rangle_{\mathcal{D}} := \sum_{a\in\mathcal{D}} f(a)g(a)$. A norm is defined on $\mathcal{L}(\mathcal{D})$ by $\|f\|_{\mathcal{D}} = \sqrt{\langle f\,,\,f\rangle_{\mathcal{D}}}$.*

Note that weighted least squares is easily incorporated in this setup by slightly changing the definition of the inner product $\langle .\,,\,.\rangle_{\mathcal{D}}$.
Let $\mathcal{D}$ be a design. Suppose our statistical model is

$$Y(x) = f(x,\theta) + \varepsilon(x), \tag{1.1}$$

where $\theta \in \mathbf{R}^p$ and $\varepsilon(x)$ is a real-valued random variable for all $x \in \mathcal{D}^*$ with $\mathbf{E}\varepsilon(x) = 0$ and $\mathbf{V}\varepsilon(x) = \sigma^2$. Suppose we have observations $Y_1, \ldots, Y_N$ from this model, where Y_i is $Y(a_i^*)$ for $a_i^* \in \mathcal{D}^*$ and replications are allowed, i.e. a_i^* may be equal to a_j^* for $i \neq j$. Then the **least squares estimator** for the parameter vector θ is given by

$$\widehat{\theta} = \min_{\theta\in\Theta} \sum_{i=1}^{N} \mid Y_i - f(a_i^*, \theta) \mid^2 . \tag{1.2}$$

Let $g \in \mathcal{L}(\mathcal{D})$ be the unique function in $\mathcal{L}(\mathcal{D})$ such that $g(a_i) = Y_i$ for all $i = 1, \ldots, N$. Since

$$\widehat{\theta} = \min_{\theta\in\Theta} \sum_{i=1}^{N} \mid Y_i - f(a_i^*, \theta) \mid^2 = \min_{\theta\in\Theta} \|g - f(.,\theta)\|_{\mathcal{D}}^2 , \tag{1.3}$$

we see that least squares estimation corresponds to a minimum distance problem in $\mathcal{L}(\mathcal{D})$ with the inner product in Definition (2.3). Note that a function $f(x)$ for $x = (x_1, \ldots, x_d) \in \mathcal{D}^*$ can be naturally extended for $x = (x_1, \ldots, x_d, x_{d+1}) \in \mathcal{D}$ by $(x_1, \ldots, x_d, x_{d+1}) \longmapsto f(x_1, \ldots, x_d)$.

3 Identifiability of linear models

In the sequel we restrict ourselves to linear models, i.e. models such that $f(x, \theta)$ is a linear function of the components of the parameter vector θ

$$Y(x) = \sum_{\alpha\in\mathcal{M}} \theta_\alpha \, p_\alpha(x) + \varepsilon(x), \tag{1.4}$$

where p_α $(\alpha \in \mathcal{M})$ is an element of $\mathcal{L}(\mathcal{D}^*)$. Clearly the p_α's can be viewed as elements of $\mathcal{L}(\mathcal{D})$. Recall that $\mathcal{L}(\mathcal{D})$ is a vector space over the real numbers and $\mathcal{L}(\mathcal{D})$ is isomorphic to $\mathbf{R}^N$, where N is the number of points in $\mathcal{D}$.

Definition 3.1 *A linear model (1.4) is identifiable by a design $\mathcal{D}$ if the functions p_α ($\alpha \in \mathcal{M}$) are linearly independent elements of $\mathcal{L}(\mathcal{D})$.*

The classical notion of identifiability is equivalent to our definition. Indeed, let $Y = X\theta + \varepsilon$ be a linear model where X is a matrix with p columns. If the design matrix X has rank less than p, then θ is not identifiable since different values of θ yield the same value of $X\theta$. This actually means that the model coincides for different parameter values when restricted to the design points. In other words, the functions on $\mathcal{D}$ that take as values the components of the columns of X are linearly dependent.
For linear models, least squares estimation is the orthogonal projection of g onto span $\{p_\alpha \mid \alpha \in \mathcal{M}\}$. Note that if $\{p_\alpha \mid \alpha \in \mathcal{M}\}$ is an orthogonal subset of $\mathcal{L}(\mathcal{D})$, then elementary linear algebra arguments yield that

$$\widehat{\theta}_\alpha = \frac{\langle g\,,\, p_\alpha\rangle_{\mathcal{D}}}{\langle p_\alpha\,,\, p_\alpha\rangle_{\mathcal{D}}}. \tag{1.5}$$

The functional description of least squares estimation has some advantages over the usual vector space description. It is more natural in our opinion since the model description is also at a functional level. A numerical advantage is that we do not need matrix inversion to compute the coefficient estimates. Indeed, orthogonalisation by the Gram-Schmidt procedure becomes a simple recursive procedure. Note that contrary to the classical use of Gram-Schmidt in the case of $\mathbf{R}^N$, we use Gram-Schmidt in a symbolic way in the space of polynomials. In this polynomial setting rewriting the estimated orthogonalised model in terms of the original model corresponds simply to collect coefficients.
The functional description given here differs from the abstract setting to linear models initiated by Kruskal (1961). See Drygas (1970) for a self-contained treatment. Specifically we extensively use computations with polynomials in the next sections. A paper which is closer in spirit to our paper is Neumaier and Seidel (1992), where a design is seen as a normalized measure and optimal designs are derived using arguments in $\mathcal{L}(\mathcal{D})$.

4 A polynomial algebraic representation of $\mathcal{L}(D)$

The set of real-valued functions over a finite set of distinct points can be described using particular classes of polynomials. More precisely let $\mathcal{D}$ be a design in $\mathbf{R}^d \times \mathcal{L}$, let $\mathbf{R}[x_1, \ldots, x_{d+1}]$ be the polynomial ring in $d+1$ indeterminates with real coefficients and let Ideal$(\mathcal{D}) \subset \mathbf{R}[x_1, \ldots, x_{d+1}]$ be the set of all polynomials whose zeros include the design points. Then the quotient space $\mathbf{R}[x_1, \ldots, x_{d+1}]/\text{Ideal}(\mathcal{D})$ is a description or representation of $\mathcal{L}(D)$. Moreover vector space bases of $\mathbf{R}[x_1, \ldots, x_{d+1}]/\text{Ideal}(\mathcal{D})$ made of monomials can be determined with Gröbner basis methods. We require the definition of a term ordering.

Definition 4.1 *A term ordering τ on the monomials of $\mathbf{R}[x_1, \ldots, x_d]$ is a total well-ordering such that $x^\alpha \prec_\tau x^\beta$ implies $x^\alpha x^\gamma \prec_\tau x^\beta x^\gamma$ for all $\gamma \neq 0$.*

Theorem 4.2 *Given a design $\mathcal{D} \subset \mathbf{R}^d \times \mathcal{L}$, a term ordering τ and a Gröbner basis $G \subset \mathbf{R}[x_1, \ldots, x_{d+1}]$ for $\mathcal{D}$ with respect to τ, then a vector space basis of $\mathbf{R}[x_1, \ldots, x_{d+1}]/Ideal(\mathcal{D})$ is given by*

$$\begin{aligned} Est_{\mathcal{D},\tau} &:= \{x^\alpha \mid x^\alpha \textit{ is not divisible} \\ &\qquad \textit{by any of the leading terms of the elements of } G\} \\ &= \{x^\alpha \mid \alpha \in L_{\mathcal{D},\tau}\}\,. \end{aligned}$$

Moreover, if the set $\{p_\alpha \mid \alpha \in \mathcal{M}\}$ in Model (1.4) is a subset of $Est_{\mathcal{D},\tau}$, then Model (1.4) is identifiable. The set $Est_{\mathcal{D},\tau}$ has exactly N elements where N is the cardinality of $\mathcal{D}$.

Proof. For the first part see for example Cox et al. (1996) and for the second and third parts see Pistone, Riccomagno and Wynn (2000). ■

Note that Theorem 4.2 applies to any design with no replicates, namely to a set of distinct points. Designs defined according to Definition 2.1 are particular examples of sets of distinct points where the "label indeterminate", x_{d+1} distinguishes replicated points. For designs with replicates the trick here is to consider in Model (1.4) only terms of Est not involving x_{d+1}.

For statistical inference we need a design, a model, and observations. In a classical screening setup a model is chosen first. However, we may also choose the model after seeing the design (for example the planned design was not completed and there are missing points, see Holliday et al., 1999). In this case, Theorem 4.2 provides a powerful tool in the choice of a regression vector for a linear model of the type in (1.4).

In general different term orderings give different Est sets and also typically Est includes monomials involving the label indeterminate x_{d+1} which clearly should not be included in Model (1.4). This suggests to partition Est, equivalently L, in three disjoint parts

$$L_{\mathcal{D},\tau} = L_x^* \cup L_{x_{d+1}} \cup L_{x,x_{d+1}}$$

where L_x^* includes all the elements of $L_{\mathcal{D},\tau}$ that do not involve x_{d+1}, $L_{x_{d+1}}$ includes all the monomials in $Est_{\mathcal{D},\tau}$ which involve only x_{d+1} and $L_{x,x_{d+1}}$ includes the remaining terms. The set $\mathcal{M}$ in Model (1.4) can be chosen to be a subset of L_x^*. The combination of the choice of the term ordering and of the structure of the design determines these three parts.

A reasonable choice for the term ordering is one that eliminates the x_{d+1} variable. For elimination term ordering we refer to Cox et al. (1996) and here simply observe that an effect of eliminating x_{d+1} is that the number of monomials in $L_{x_{d+1}}$ is as small as possible.

We conclude this section by showing with polynomial algebra techniques that identifiability is not affected by replications. The elimination of x_{d+1}

from Ideal($\mathcal{D}$) corresponds to projecting $\mathcal{D} \subset \mathbf{R}^d \times \mathcal{L}$ onto $\mathbf{R}^d$. For some term orderings the Gröbner basis of Ideal($\mathcal{D}^*$) and $Est_{\mathcal{D}^*}$ can be easily deduced from the Gröbner basis of Ideal($\mathcal{D}$) and $Est_{\mathcal{D}}$.

Theorem 4.3 *1) Ideal($\mathcal{D}$) $\cap \mathbf{R}[x_1, \ldots, x_d]$ = Ideal($\mathcal{D}^*$). 2) Let G be the Gröbner basis of Ideal($\mathcal{D}$) with respect to a term ordering eliminating x_{d+1}. The Gröbner basis of Ideal($\mathcal{D}^*$) is $G \cap \mathbf{R}[x_1, \ldots, x_d]$.*

Proof. 1) Assume $f \in$ Ideal($\mathcal{D}$) $\cap \mathbf{R}[x_1, \ldots, x_d]$. Then for all $a = (a^*, \ell) \in \mathcal{D}$, we have that $0 = f(a) = f(a^*, \ell) = f(a^*)$ as $f \in \mathbf{R}[x_1, \ldots, x_d]$. This implies $f \in$ Ideal($\mathcal{D}^*$). The converse is obvious. 2) See Cox et al. (1996). ∎

Clearly Theorem 4.3 applies when instead of x_{d+1} we need to eliminate some other variable. The projection is now on the remaining variables and replications may appear. For example the projection of the 2^2 design at levels ± 1 over the first factor gives ± 1 replicated twice.

5 Examples

The analysis of observations suggested in the paper proceeds as follows. Given a design $\mathcal{D}$, compute $Est_{\mathcal{D},\tau}$ where τ is a term ordering that eliminates the extra variable t. Orthonormalise the terms of $Est_{\mathcal{D},\tau}$ that do not involve t. Collect coefficients to determine the parameters of the wanted model from the estimated coefficients in the orthonormalised model.

Example 5.1 (2^2 full factorial design with centre points)
Consider the 2^2 design at levels ± 1. The standard model associated with it is

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_{12} x_1 x_2.$$

Clearly 1, x_1^2 and x_2^2 are confounded on $\mathcal{D}$. Suppose that we want to test linearity by adding quadratic terms to the model. The simplest way to extend this design such that quadratic terms become identifiable, is to add centre points. We add four observations at $(0,0)$ and the design becomes

$$\begin{aligned} \mathcal{D} &= \{(-1,-1,1),(-1,1,1),(1,-1,1),(1,1,1), \\ &\quad\ (0,0,1),(0,0,2),(0,0,3),(0,0,4)\} \\ &= \{a_i \mid i = 1,\ldots,8\}. \end{aligned}$$

We use a term ordering σ that eliminates the variable t and is a degree reverse lexicographic ordering on x_1 and x_2 (Cox et al., 1996). We obtain

$$Est_{\mathcal{D}} = \{1, x_1, x_2, x_1x_2, x_2^2, t, t^2, t^3\}.$$

An orthonormal basis for the linear span of the terms not involving t is

$$\left\{ \frac{1}{\sqrt{8}}, \frac{x_1}{2}, \frac{x_2}{2}, \frac{x_1x_2}{2}, \frac{x_2^2 - \frac{1}{2}}{\sqrt{2}} \right\}.$$

Let g be the polynomial that interpolates the observations $Y_1, \ldots, Y_8$ at the design points, $g(a_i) = Y_i$, $i = 1, \ldots, 8$. The traditional sums of squares correspond to the squares of the inner products. In particular, using Equation (1.5) the average over the centre points minus the average over the full factorial is computed as

$$SS_{\text{pure quadratic}} = \frac{1}{8}\left(Y_1 + \ldots + Y_4 - (Y_5 + \ldots + Y_8)\right)^2 = \left\langle g\,,\, \frac{x_2^2 - \frac{1}{2}}{\sqrt{2}} \right\rangle_{\mathcal{D}}^2 .$$

By simple comparison of terms, we read off the coefficients of the wanted model from the coefficients of the orthonormalised model.

Example 5.2 (Star composite design with centre points)
If the analysis of the 2^2 full factorial design with centre points indicates that there is curvature, then it is practice to study the quadratic terms. We choose to break the aliasing by augmenting the design with four axial points at $(0, \pm 2)$ and $(\pm 2, 0)$. The new design $\mathcal{D}$ is given below

$$\begin{aligned}\mathcal{D} \;=\; & \{(-1,-1,1),(-1,1,1),(1,-1,1),(1,1,1),(-2,0,1),(2,0,1),\\ & (0,2,1),(0,-2,1),(0,0,1),(0,0,2),(0,0,3),(0,0,4)\}.\end{aligned}$$

We use again the term ordering σ used in Example 5.1. The monomials in $Est_{\mathcal{D}}$ not involving the counting variable are

$$\left\{1, x_1, x_2, x_1^2, x_2^2, x_1x_2, x_2^3, x_1x_2^2, x_2^4\right\}$$

An orthonormal basis for the linear span of these terms by applying the Gram-Schmidt procedure to $Est_{\mathcal{D}}$ in the order above, yields

$$\left\{ \frac{\sqrt{3}}{6}, \frac{\sqrt{3}}{6}x_1, \frac{\sqrt{3}}{6}x_2, \frac{\sqrt{6}}{12}\left(x_1^2 - 1\right), \frac{\sqrt{3}}{24}\left(x_1^2 + 3x_2^2 - 4\right), \right.$$
$$\left. \frac{x_1x_2}{2}, \frac{\sqrt{6}}{12}x_1\left(3x_2^2 - 1\right), \frac{\sqrt{6}}{12}x_2\left(x_2^2 - 3\right) \right\}.$$

Example 5.3 (2^{3-1} fractional factorial design with centre points)
Consider the standard 2^{3-1} design with generator $I = ABC$ and four additional centre points. The design is

$$\begin{aligned}\mathcal{D} \;=\; & \{(1,-1,-1,1),(-1,1,-1,1),(-1,-1,1,1),(1,1,1,1),\\ & (0,0,0,1),(0,0,0,2),(0,0,0,3),(0,0,0,4)\}.\end{aligned}$$

Again using the term ordering σ, we obtain $Est_{\mathcal{D}} = \left\{1, x_1, x_2, x_3, x_3^2, t, t^2, t^3\right\}$. An orthonormal basis for the linear span of the terms not involving t is

$$\left\{ \frac{1}{\sqrt{8}}, \frac{x_1}{2}, \frac{x_2}{2}, \frac{x_3}{2}, \frac{x_3^2 - \frac{1}{2}}{\sqrt{2}} \right\}.$$

References

Caboara, M. and Robbiano, L. (1997). Families of ideals in statistics. In: Küchlin, W. (ed). *ISSAC'97, Proceedings of the International Symposium on Symbolic and Algebraic Computation, Hawaii*, pp. 404–417. ACM Press, New York.

Cox, D., Little, J. and O'Shea, D. (1996). *Ideals, Varieties, and Algorithms.* Springer, New York. Second edition.

Drygas, H. (1970). *The Coordinate-Free approach to Gauss-Markov Estimation.* Springer-Verlag, Berlin.

Giglio, B., Riccomagno, E. and Wynn, H. P. (2000). Gröbner bases in regression. *Journal of Applied Statistics*, **27**, 923-928.

Holliday, T., Pistone, G., Riccomagno, E. and Wynn, H. P. (1999). The application of computational algebraic geometry to the analysis of designed experiments: a case study. *Computational Statistics*, **14**, 213-231.

Kruskal, W. (1961). The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations. In: Neyman, J. (ed). *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol. I*, pp. 435–451. University of California Press, Berkeley.

Neumaier, A. and Seidel, J.J. (1992). Measures of strength $2e$ and optimal designs of degree e. *Sankhyā*, **54**, 299-309.

Pistone, G., Riccomagno, E. and Wynn, H.P. (2000). *Algebraic Statistics: Computational Commutative Algebra in Statistics.* Chapman & Hall / CRC Press, London.

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

Riccomagno, E. (1997). Algebraic identifiability in experimental design and related topics. Ph.D. thesis, University of Warwick.

Authors

A.M. Cohen
Eindhoven University of Technology
Department of Mathematics
P.O. Box 513
5600 MB Eindhoven
The Netherlands
amc@win.tue.nl
http://www.win.tue.nl/math/dw/personalpages/amc

A. Di Bucchianico
EURANDOM
and
Eindhoven University of Technology
Faculty of Technology Management
Section Quality of Products and Processes
P.O. Box 513
5600 MB Eindhoven
The Netherlands
A.d.Bucchianico@tm.tue.nl
http://www.tm.tue.nl/vakgr/ppk/bucchianico.htm

E. Riccomagno
EURANDOM
P.O. Box 513
5600 MB Eindhoven
The Netherlands
riccomagno@eurandom.tue.nl
http://euridice.tue.nl/∼ericcoma/