# Design of Experiments and Polynomial Algebra

John B. Little

Department of Mathematics and Computer Science, College of the Holy Cross

little@mathcs.holycross.edu

ADM Seminar

Clemson University

February 5, 2004

# Outline of Talk

- Some problems in design of statistical experiments

- How Gröbner bases give a partial solution

- Border bases and multiplication matrices

- A general solution

reference: Caboara, M. and Robbiano, L. "Families of Estimable Terms," Proceedings of ISSAC 2001.

## §1. The Problems

Suppose you were an auto manufacturer and you wanted to assess customer reactions to different combinations of 5 features in a new car model you were planning:

1. $x_1 =$ engine type $\in \{0, 1, \ldots, \ell_1 - 1\}$

2. $x_2 =$ transmission $\in \{0, 1, \ldots, \ell_2 - 1\}$

3. $x_3 =$ exterior color $\in \{0, 1, \ldots, \ell_3 - 1\}$

4. $x_4 =$ upholstery type $\in \{0, 1, \ldots, \ell_4 - 1\}$

5. $x_5 =$ instrument package $\in \{0, 1, \ldots, \ell_5 - 1\}$

# The Problems, cont.

- Survey some random sample of customers to see how they rate each of the $\ell_1 \cdot \ell_2 \cdot \cdots \cdot \ell_5$ combinations.

- Gives a *preference function* $p : D \to \mathbf{Q}$, where

$$D = \prod_{i=1}^{5} \{0, 1, \ldots, \ell_i - 1\},$$

called the *full factorial* $\ell_1 \times \cdots \times \ell_5$ *design*.

- Since $D$ is a finite set, every $p : D \to \mathbf{Q}$ can be represented by a *polynomial in 5 variables* (multivariate Lagrange interpolation)

# The Problems, cont.

- Moreover, $D \subset \mathbf{R}^5$ is defined by:

$$\begin{aligned} x_1(x_1-1)(x_1-2)\cdots(x_1-\ell_1+1) &= 0 \\ x_2(x_2-1)(x_2-2)\cdots(x_2-\ell_2+1) &= 0 \\ &\vdots \\ x_5(x_5-1)(x_5-2)\cdots(x_5-\ell_5+1) &= 0 \end{aligned}$$

  The LHS's generate an ideal in $\mathbf{R}[x_1, \ldots, x_5]$ we'll call $I(D)$ – polynomials vanishing on $D$.

- $\Rightarrow$ the *distinct* polynomial functions $p : D \to \mathbf{Q}$ are in 1-1 correspondence with elements of the quotient ring $\mathbf{Q}[x_1, \ldots, x_5]/I(D)$.

- Remainders on polynomial division $\Rightarrow$ the monomials $x_1^{\alpha_1} \cdots x_5^{\alpha_5}$, $0 \le \alpha_i \le \ell_i - 1$ give a vector space basis for $\mathbf{Q}[x_1, \ldots, x_5]/I(D)$.

# The Problems, cont.

- Realistically, though, it would not be possible to get preference data from each customer for all $\ell_1 \times \cdots \times \ell_5$ combinations if $\ell_i$ were realistic numbers, say 5 - 10 (too many combinations!!)

- On the other hand, could get data on a subset $F \subset D$, called a *fractional design*, or *fraction*.

- But then we're back to non-unique representatives for polynomial functions, *unless* we also restrict the collection of monomials $x^\alpha = x_1^{\alpha_1} \cdots x_5^{\alpha_5}$ appearing in the preference function.

# The Problems, cont.

- Choice of a collection of monomials $\mathcal{B} \Leftrightarrow$ choice of a *statistical model* for form of preference function, and

- **Proposition 1.** TFAE for $|\mathcal{B}| = |F|$:

  1. There exists a unique polynomial in span($\mathcal{B}$) interpolating any polynomial function on $F$

  2. For $F = \{P_i\}$, $\mathcal{B} = \{x^{\alpha(j)}\}$,
  $$\det(P_i^{\alpha(j)}) \neq 0$$

  3. $\mathcal{B}$ gives a basis for $\mathbf{Q}[x_1, \ldots, x_5]/I(F)$.

# Some Examples

For simplicity, stick to the full factorial $3^2$ design: $D = \{0,1,2\} \times \{0,1,2\}$. We have

$$I(D) = \langle x^3 - 3x^2 + 2x, y^3 - 3y^2 + 2y \rangle$$

1. Say $F = \{(0,0),(1,1),(2,2),(1,2)\}$ and $\mathcal{B}_1 = \{1, x, y, x^2\}$ Then

$$\det(P_i^{\alpha(j)}) = \det \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 4 \\ 1 & 1 & 2 & 1 \end{pmatrix} = -2$$

   $\mathcal{B}_1$ does satisfy Proposition 1.

2. Same $F$, $\mathcal{B}_2 = \{1, x^2, y^2, xy\}$

$$\det(P_i^{\alpha(j)}) = \det \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 4 & 4 & 4 \\ 1 & 1 & 4 & 2 \end{pmatrix} = 0$$

   $\mathcal{B}_2$ does not satisfy Proposition 1.

# Estimable Terms

- From now on, we will make the following assumption on the structure of the set $\mathcal{B}$:

$$x^\alpha \in \mathcal{B} \text{ and } x^\beta | x^\alpha \Rightarrow x^\beta \in \mathcal{B}.$$

- This says $\mathcal{B}$ is what is known as a "delta-set," or a "standard set," or a *complete set of estimable terms* (following Caboara and Robbiano).

- Examples: $\mathcal{B}_1 = \{1, x, y, x^2\}$ from before is a complete set of estimable terms; $\mathcal{B}_2 = \{1, x^2, y^2, xy\}$ is *not*.

- Motivation for this comes primarily from the theory of Gröbner bases – we'll say more about this presently.

# The Key Problems

1. Given the fraction $F$, what complete sets of estimable terms $\mathcal{B}$ exist for that $F$? $\Leftrightarrow$ What models can be identified by considering the fraction $F$?

2. (*) Given a complete set of estimable terms $\mathcal{B}$, what fractions $F$ satisfy Proposition 1? $\Leftrightarrow$ Given a model, which fractions can be used to identify that model?

# §2. "Review" of Gröbner bases

- Want a constructive way to answer questions about polynomial ideals (such as ideal membership problem).

- Start from a *monomial order*: a well-ordering $>$ on set of monomials compatible with ring multiplication: $x^\alpha > x^\beta \Rightarrow x^\alpha \cdot x^\gamma > x^\beta \cdot x^\gamma$.

- $>$ selects a *leading term* $LT(f)$ from each polynomial

- Given an ideal $I \subset \mathbf{Q}[x_1, \ldots, x_n]$, a Gröbner basis for $I$ is a set $G \subset I$ such that for all $f \in I$, there exist some $g \in G$ such that $LT(g)|LT(f)$.

# Gröbner bases, cont.

- Example: $I = \langle xy - 1, x^2 + y^2 - 2\rangle$, lexicographic order with $x > y$ in $\mathbf{Q}[x, y]$.

- $LT(xy - 1) = xy$ and $LT(x^2 + y^2 - 2) = x^2$.

- Note $-x(xy - 1) + y(x^2 + y^2 - 2)$
$$= x + y^3 - 2y \in I,$$
but $x = LT(x + y^3 - 2y) \notin \langle xy, x^2\rangle$.

- A lex Gröbner basis for this ideal is
$$G = \{y^4 - 2y^2 + 1, x + y^3 - 2y\}$$

- A generalization of polynomial division exists with any set of divisors. If we divide by a Gröbner basis $G$, then we get $f \in I = \langle G\rangle \Leftrightarrow$ remainder is zero.

# Zero-dimensional ideals

- *Buchberger's algorithm* computes a Gröbner basis $G$ from any generating set for $I$ as input; *Buchberger-Möller algorithm* gives $G$ for $I(F)$ for any finite set of points. (Also recent work of Gao-Farr here!)

- If $G = \{g_1, \ldots, g_t\}$ is a Gröbner basis for $I$, then the set $\mathcal{B}$ of *monomials in the complement* of $\langle LT(G) \rangle = \langle LT(g_1), \ldots, LT(g_t) \rangle$ gives a basis for $\mathbf{Q}[x_1, \ldots, x_n]/I$.

- Example: $G = \{\underline{y^4} - 2y^2 + 1, \underline{x} + y^3 - 2y\} \Rightarrow$
$$\mathcal{B} = \{1, y, y^2, y^3\}$$

- Terminology: $I$ is "zero-dimensional" $\Leftrightarrow$ set of common zeroes over $\mathbf{C}$ is finite $\Leftrightarrow$ $\mathcal{B}$ is finite.

# Connections

- The monomial basis $\mathcal{B}$ of $\mathbf{Q}[x_1, \ldots, x_n]/I$ consisting of the *monomials in the complement* of $\langle LT(G) \rangle$ for a Gröbner basis $G$ for $I$ is *always* a complete set of estimable terms (i.e. "delta-set") as above.

- Given a fraction $F \subset D$, let $I(F)$ be the ideal of all polynomials vanishing on $F$.

- $\Rightarrow$ One *partial solution* of first problem is obtained like this: pick any monomial order $>$, find $G =$ Gröbner basis of $I(F)$ with respect to $>$. The set $\mathcal{B}$ of monomials in the complement of $\langle LT(G) \rangle$ is one complete set of estimable terms for $F$.

# Examples

Take $D$ = full factorial $3^2$ design, and $F = \{(0,0),(1,1),(2,2),(1,2)\}$ as above.

With $>$ = lex order $y > x$, we get a GB of $I(F)$:

$$\begin{aligned} G \;=\; \{&x^3 - 3x^2 + 2x, \\ &(x-1)y - (x^2 - x), \\ &y^2 - 3y - x^2 + 3x\} \end{aligned}$$

The set of monomials in the complement of $\langle LT(G)\rangle$ is $\mathcal{B}_1 = \{1, x, y, x^2\}$, since $\langle LT(G)\rangle = \langle x^3, xy, y^2\rangle$. (Same as a previous example!)

Can check that $\mathcal{B}_2 = \{1, x, y, xy\}$ is also a complete set of estimable terms for this $F$, *but no Gröbner basis* $G$ for $I(F)$ has $\langle LT(G)\rangle = \langle x^2, y^2\rangle$ (since three of the points collinear, $F$ not a complete intersection of two conics).

# §3. Border Bases

Suppose $\mathcal{B}$ is a basis of $\mathbf{Q}[x_1, \ldots, x_n]/I$.

- The *border* of $\mathcal{B}$ is:

$$\mathcal{B}^+ = \{x_i \cdot x^\alpha \notin \mathcal{B} : x^\alpha \in \mathcal{B}\}$$

- $x^\beta \in \mathcal{B}^+ \Rightarrow$ there exist $a_{\beta\alpha} \in \mathbf{Q}$ such that

$$g_\beta = x^\beta - \sum_{x^\alpha \in \mathcal{B}} a_{\beta\alpha} x^\alpha \in I$$

- Call the collection of all these $g_\beta$ a *border basis* for $I$.

- **Proposition 2.** The $g_\beta$ generate $I$.

# Multiplication maps

The $g_\beta$ in a border basis are closely related to the multiplication maps

$$\begin{array}{rcl} m_{x_i} : \mathbf{Q}[x_1,\ldots,x_n]/I & \to & \mathbf{Q}[x_1,\ldots,x_n]/I \\ f & \mapsto & x_i \cdot f \end{array}$$

In fact, the *matrix* of $m_{x_i}$ with respect to $\mathcal{B}$ can be "read off" the $g_\beta$ directly:

**Proposition 3.** If $x^\alpha \in \mathcal{B}$ and $x_i \cdot x^\alpha \in \mathcal{B}$ get a column in $M_i = [m_{x_i}]$ with one 1 in row corresponding to $x_i \cdot x^\alpha \in \mathcal{B}$ all other entries zero. If $x_i x^\alpha = x^\beta \in \mathcal{B}^+$, get a column in $M_i$ consisting of the vector of coefficients from $g_\beta$.

Note: By construction, $M_i M_j = M_j M_i$ for all $i, j$.

# Example, continued

With $F = \{(0,0),(1,1),(2,2),(1,2)\}$, and estimable terms $\mathcal{B} = \{1, x, y, xy\}$ as above, get $\mathcal{B}^+ = \{x^2, x^2y, xy^2, y^2\}$ and border basis:

$$\begin{aligned} g_{(2,0)} &= x^2 - x + y - xy \\ g_{(2,1)} &= x^2y + 2y - 3xy \\ g_{(1,2)} &= xy^2 + 2x - 3xy \\ g_{(0,2)} &= y^2 + 2x - 2y - xy \end{aligned}$$

So for instance

$$M_1 = [m_x] = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & -1 & 0 & -2 \\ 0 & 1 & 1 & 3 \end{pmatrix}$$

# Other "angles"

- The matrices $M_i$ have been much studied in polynomial system solving (work of Stetter, Mourrain, ... )

- Interesting fact: eigenvalues of $M_i$ are the $x_i$-coordinates of the points in $V(I)$.

- $\Rightarrow$ hybrid symbolic-numeric techniques for polynomial system solving based on eigenvalue methods.

- Border bases and the multiplication matrices $M_i$ can replace Gröbner bases for many other operations on zero-dimensional ideals. For instance, can test for ideal membership:

$$f \in I \Leftrightarrow f(M_1, \ldots, M_n) = 0$$

  (0 matrix).

# Problem 2

For the rest of the talk, we will consider the second general problem stated above:

Suppose $D$ is the full factorial $\ell_1 \times \cdots \times \ell_n$ design and $\mathcal{B}$ is a complete set of estimable terms. What are all the fractions $F$ such that $\mathcal{B}$ forms a basis of $\mathbf{Q}[x_1, \ldots, x_n]/I(F)$?

Caboara and Robbiano formulate a general technique for solving this problem using border bases and the multiplication matrices $M_i$.

The idea is as follows.

# The Method

1. Form $\mathcal{B}^+$ as before and for each $\beta \in \mathcal{B}^+$ define $g_\beta = x^\beta - \sum_{x^\alpha \in \mathcal{B}} a_{\beta\alpha} x^\alpha$ as above, but with $a_{\beta\alpha}$ *indeterminates*.

2. From the $g_\beta$, form matrices $M_i = [m_{x_i}]$ as in Proposition 3.

3. Impose the equations $M_i M_j = M_j M_i$ all $i, j$, and $f_k(M_1, \ldots, M_n) = 0$ where
$$f_k = x_k(x_k - 1) \cdots (x_k - \ell_k + 1) = 0$$
$1 \leq k \leq n$ are the defining equations of $D$ (a big system (*) of polynomial equations in $a_{\beta\alpha}$)

4. **Theorem 4.** (C-R) There is a 1-1 correspondence
$$\{solutions\ of\ (*)\} \leftrightarrow \{F \subset D\}$$
which have $\mathcal{B}$ as basis.

# Final Example

We continue with $D$ = full factorial $3^2$ design, but in form

$$D = \{-1, 0, 1\}^2$$

(instead of $\{0, 1, 2\}^2$ – simplifies equations somewhat!)

Take $\mathcal{B} = \{1, x, y, xy\}$ as above. We have $\mathcal{B}^+ = \{x^2, x^2y, xy^2, y^2\}$ and

$$\begin{aligned}
g_{(2,0)} &= x^2 - a_{200} - a_{201}x - a_{202}y - a_{203}xy \\
g_{(2,1)} &= x^2y - a_{210} - a_{211}x - a_{212}y - a_{213}xy \\
g_{(1,2)} &= xy^2 - a_{120} - a_{121}x - a_{122}y - a_{123}xy \\
g_{(0,2)} &= y^2 - a_{020} - a_{021}x - a_{022}y - a_{023}xy
\end{aligned}$$

Then

$$M_1 = [m_x] = \begin{pmatrix} 0 & a_{200} & 0 & a_{210} \\ 1 & a_{201} & 0 & a_{211} \\ 0 & a_{202} & 0 & a_{212} \\ 0 & a_{203} & 1 & a_{213} \end{pmatrix}$$

# Example, cont.

and

$$M_2 = [m_y] = \begin{pmatrix} 0 & 0 & a_{020} & a_{120} \\ 0 & 0 & a_{021} & a_{121} \\ 1 & 0 & a_{022} & a_{122} \\ 0 & 1 & a_{023} & a_{123} \end{pmatrix}$$

The equations

$$\begin{aligned} M_1 M_2 - M_2 M_1 &= 0 \\ f_1(M_1) = f_2(M_2) &= 0 \end{aligned}$$

give 44 nontrivial equations in 16 variables $a_{ijk}$. Want to determine all solutions.

Somewhat surprisingly(?), with a good Gröbner basis package like *Singular*, even using a lexicographic order, this computation is extremely fast ($<$ 1 sec. on my 2.5 year old 900MHz Pentium III/Windows laptop).

# Results

- There are 81 solutions; each corresponding to a fraction $F$ which can be used to estimate preference function modeled by $\mathcal{B}$ by Theorem 4. Note: $\binom{9}{4} = 126$ fractions of size 4.

- For instance, one solution gives mult. matrices:

$$[m_x] = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & -1 & 1 & 0 \end{pmatrix}$$

$$[m_y] = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \end{pmatrix}$$

# Results, cont.

- The corresponding border basis is

$$\begin{aligned} g_{(2,0)} &= x^2 + x + y + xy \\ g_{(2,1)} &= x^2 y - y \\ g_{(1,2)} &= xy^2 + y \\ g_{(0,2)} &= y^2 + xy \end{aligned}$$

- The corresponding fraction is

$$F = \{(0,0), (-1,0), (1,-1), (-1,1)\}$$

- As in one of our earlier examples, no Gröbner basis for $I(F)$ would give this complete set of estimable terms $\mathcal{B} = \{1, x, y, xy\}$ because three of the points are collinear.

- Similarly for the other 80 solutions.

# Some Final Comments

One question that might have occurred to you: Is this method necessary? Why not just enumerate all fractions $F$ of the desired size and test each on $\mathcal{B}$ using the criterion of Proposition 1:

**Proposition 1.** TFAE for $|\mathcal{B}| = |F|$:

1. There exists a unique polynomial in span($\mathcal{B}$) interpolating any polynomial function on $F$

2. For $F = \{P_i\}$, $\mathcal{B} = \{x^{\alpha(j)}\}$,
$$\det(P_i^{\alpha(j)}) \neq 0$$

3. $\mathcal{B}$ gives a basis for $\mathbf{Q}[x_1, \ldots, x_5]/I(F)$.

On small examples like the one we presented that would be easy enough.

# Final Comments, cont.

But, let's consider a larger problem.

- Say $D$ is the full factorial $5^5$ design, and say we have some $\mathcal{B}$ containing 10 monomials (some class of statistical models).

- How many possible fractions are there?
$$\binom{5^5}{10} = \binom{3125}{10} \doteq 2.4 \times 10^{28}$$
of them (!)

- So we would need to compute that many $10 \times 10$ determinants. (Good luck!!)

# Final Comments, concluded

- On the other hand, Caboara and Robbiano's method would involve considering the 5 $10 \times 10$ multiplication matrices.

- Typically, something like 1/2 of the entries might come from border basis elements, so $\sim 250$ variables.

- Would have $\binom{5}{2} = 10$ matrix equations
$$M_i M_j - M_j M_i = 0$$
and the 5 equations
$$f_k(M_1, \ldots, M_5) = 0$$
$\Rightarrow \sim 1000$ equations in $\sim 250$ variables to solve.

- Still a *very large problem*, but more tractable!