Thammasat University, Bangkok

Algebraic Statistics of Design of Experiments

Giovanni Pistone

Collegio Carlo Alberto

Bangkok, March 14 2012

The talk

1. A short history of AS from a personal perspective. See a longer account by E. Riccomagno in
   - E. Riccomagno, Metrika **69**(2-3), 397 (2009), ISSN 0026-1335, `http://dx.doi.org/10.1007/s00184-008-0222-3`.
2. The algebraic description of a designed experiment. I use the tutorial by M.P. Rogantin:
   - `http://www.dima.unige.it/~rogantin/AS-DOE.pdf`
3. Topics from the theory, cfr. the course teached by E. Riccomagno, H. Wynn and Hugo Maruri-Aguilar in 2009 at the Second de Brun Workshop in Galway:
   - `http://hamilton.nuigalway.ie/DeBrunCentre/SecondWorkshop/online.html`.

CIMPA Ecole "Statistique", 1980 à Nice (France)

Sumate Sompakdee, GP, and Chantaluck Na Pombejra

First paper

The Annals of Statistics
1998, Vol. 26, No. 1, 363–397

ALGEBRAIC ALGORITHMS FOR SAMPLING FROM CONDITIONAL DISTRIBUTIONS

BY PERSI DIACONIS[1] AND BERND STURMFELS[2]

Cornell University and University of California, Berkeley

We construct Markov chain algorithms for sampling from discrete exponential families conditional on a sufficient statistic. Examples include contingency tables, logistic regression, and spectral analysis of permutation data. The algorithms involve computations in polynomial rings using Gröbner bases.

1. Introduction. This paper describes new algorithms for sampling from the conditional distribution, given a sufficient statistic, for discrete exponential families. Such distributions arise in carrying out versions of Fisher's exact test for independence and goodness of fit. They also arise in constructing uniformly most powerful tests and accurate confidence intervals via Rao–Blackwellization. These and other applications are described in Section 2. As shown below, the new algorithms are a useful supplement to traditional asymptotic theory, which is useful for large data sets, and exact enumeration, which is useful for very small data sets.

People

- Roberto Fontana, DISMA Politecnico di Torino.
- Fabio Rapallo, Università del Piemonte Orientale.
- Eva Riccomagno DIMA Università di Genova.
- Maria Piera Rogantin, DIMA, Università di Genova.
- Henry P. Wynn, LSE London.

AS and DoE: old biblio and state of the art

Beginning of AS in DoE

- G. Pistone, H.P. Wynn, Biometrika **83**(3), 653 (1996), ISSN 0006-3444
- L. Robbiano, *Gröbner Bases and Statistics*, in *Gröbner Bases and Applications (Proc. of the Conf. 33 Years of Gröbner Bases)*, edited by B. Buchberger, F. Winkler (Cambridge University Press, 1998), Vol. 251 of *London Mathematical Society Lecture Notes*, pp. 179–204
- R. Fontana, G. Pistone, M. Rogantin, Journal of Statistical Planning and Inference **87**(1), 149 (2000), ISSN 0378-3758
- G. Pistone, E. Riccomagno, H.P. Wynn, *Algebraic statistics: Computational commutative algebra in statistics*, Vol. 89 of *Monographs on Statistics and Applied Probability* (Chapman & Hall/CRC, Boca Raton, FL, 2001), ISBN 1-58488-204-2

All topics: state of the art

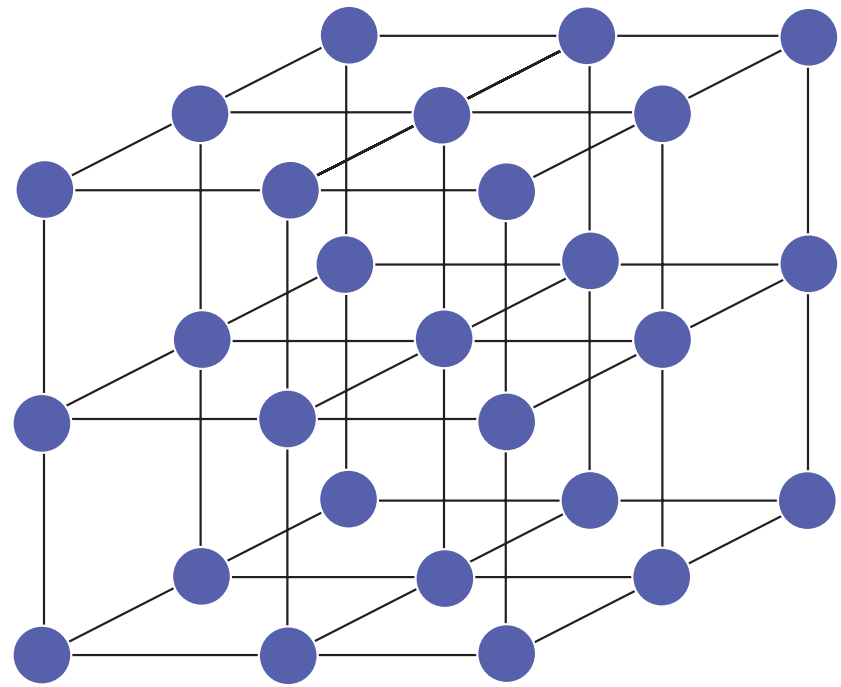
- Algebraic Statistics in the Alleghenies at the Pennsylvania State University, June 8 to June 15, 2012 `http://www.math.psu.edu/morton/aspsu2012/index.html`.

Designed Experiment?

Factorial design

To evaluate the reliability of an electrical engine, study the dependence on three factors: - diameter of rotor - number of windings - type of cooling liquid of the push-off force of a spark control valve. Example on p. 247 in C.F.J. Wu, M. Hamada, *Experiments. Planning, analysis, and parameter design optimization, A Wiley-Interscience Publication* (John Wiley & Sons Inc., New York, 2000), ISBN 0-471-25511-4.

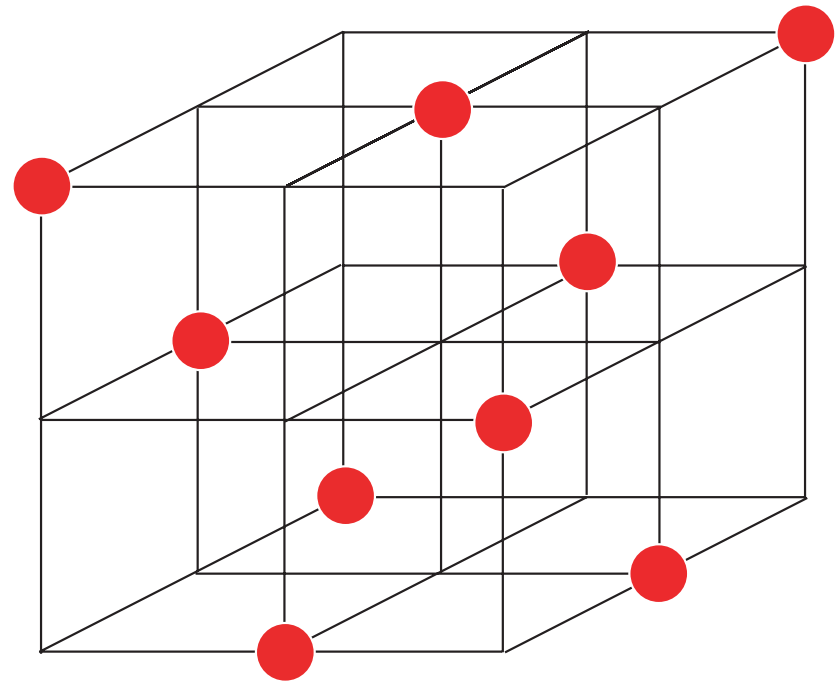
- Each factor has three levels,
- treated as ordinal levels (low - medium - high),
- coded by integer numbers $-1, 0, 1$.
- Each *treatment* is a point in $\mathbb{Z}^3$.



Fractional factorial design

The experiment is realized on fewer treatments because of problems with costs, times, practical constraints, setting the factors and measuring the responses . . . How to choose the treatments? Which **fraction** for "the best" study of the responses?

time	*pressure*	*moisture*
-1	-1	-1
-1	0	0
-1	1	1
0	-1	0
0	0	1
0	1	-1
1	-1	1
1	0	-1
1	1	0



Design and functions on the design

Response	Factors		
Push-off force (lb)	Time	Pressure	Moisture
111.1	−1	−1	−1
131.0	−1	0	0
65.4	−1	1	1
125.5	0	−1	0
46.9	0	0	1
113.7	0	1	−1
72.5	1	−1	1
141.1	1	0	−1
134.2	1	1	0

$$\text{Force} = \theta_0 + \theta_1\ \text{T} + \theta_2\ \text{P} + \theta_3\ \text{M} + \theta_{12}\ \text{T} \cdot \text{P} + \theta_{13}\ \text{T} \cdot \text{M} + \ldots$$

If enought terms are included, the first column of the table is exactly encoded by the θ's. We are going to discuss which terms should be included.

Full factorial designs and fractions: notations

Definition

- $A_i = \{a_{ij} \ : \ j = 1, \ldots, n_i\}$ **factors**

  a_{ij} **levels** coded by rational numbers $\mathbb{Q}^m$ or complex numbers $\mathbb{C}^m$
- $\mathcal{D} = A_1 \times \ldots \times A_m \subset \mathbb{Q}^m$ (or $\mathcal{D} \subset \mathbb{C}^m$) with $N = \prod_{j=1}^{m} n_j$ points
  full factorial design
- A **fraction** is a subset $\mathcal{F} \subset \mathcal{D}$;

Responses on a design

$f: \quad \mathcal{D} \mapsto \mathbb{R}$ **(functions defined on $\mathcal{D}$)** "Design" indicates either "Full factorial design" or "Fractional factorial design" or ...

Definition

- $X_i : \mathcal{D} \ni (d_1, \ldots, d_m) \mapsto d_i$ **projection**, frequently called **factor**
- $X^\alpha = X_1^{\alpha_1} \cdots X_m^{\alpha_m}, \qquad \alpha_i < n_i, \ i = 1, \ldots, m$
  $\alpha = (\alpha_1, \ldots, \alpha_m)$
  monomial responses or terms or interactions

  The term X^α has *order k* if in α there are k non-null values:
  X^α is an *interaction of order k* (binary case: order = degree)
- **Mean value of** f on $\mathcal{D}$, $E_\mathcal{D}(f)$: $\quad E_\mathcal{D}(f) = \frac{1}{\#\mathcal{D}} \sum_{d \in \mathcal{D}} f(d)$
- A **contrast** is a response f such that $\quad E_\mathcal{D}(f) = 0.$
- Two responses f and g are **orthogonal on** $\mathcal{D}$ if $E_\mathcal{D}(f\ g) = 0.$

The saturated polynomial regression model

- $L = \{(\alpha_1, \ldots, \alpha_m) \ : \ \alpha_i < n_i, \ i = 1, \ldots, m\}$ **exponents (or logarithms) of all the interactions**
- **saturated regression model**:
  For all $d_i \in \mathcal{D}$ and for the observed value y_i

$$y_i = \sum_{\alpha \in L} \theta_\alpha \, X^\alpha(d_i)$$

  with $\theta_\alpha \in \mathbb{R}$ or $\theta_\alpha \in \mathbb{C}$.
  In vector notation, $Y = (y_1, \ldots, y_n)$ response measured on the points of $\mathcal{D}$:

$$Y = \sum_{\alpha \in L} \theta_\alpha \, X^\alpha$$

- $Z = [X^\alpha(d)]_{d \in \mathcal{D}, \alpha \in L}$ **matrix of the saturated regression model**
- $\theta = (\theta_\alpha)_{\alpha \in L}$ **vector of the coefficients Matrix notation of the complete regression model**

$$Y = Z\theta$$

Confounding

On the full factorial design:
the complete regression model **is identifiable**, i.e. there is a unique solution w.r.t. θ:

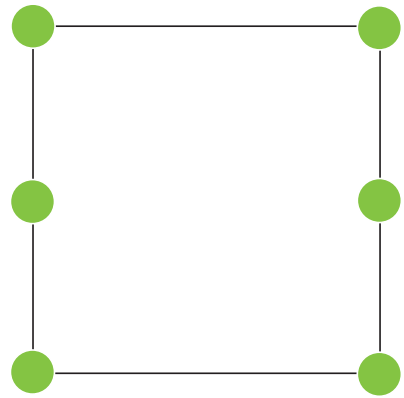
$$\hat{\theta} = Z^{-1}\ Y$$

(the matrix Z is a square full rank matrix) **On a fraction**:
there is not a unique solution w.r.t. θ
($Z = [X^{\alpha}(d)]_{d \in \mathcal{F}, \alpha \in L}$ has fewer rows than columns)

$$\hat{\theta} = (Z'Z)^{-} Z'\ Y$$

2×3 full factorial design

$$\begin{cases} A_1 = \{-1, +1\}, & n_1 = 1; \\ A_2 = \{-1, 0, 1\}, & n_2 = 3. \end{cases} \quad \mathcal{D} = \begin{array}{|r|r|} \hline -1 & -1 \\ -1 & 0 \\ -1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ \hline \end{array}$$



$$Y = \theta_{00} + \theta_{10}X_1 + \theta_{01}X_2 + \theta_{02}X_2^2 + \theta_{11}X_1X_2 + \theta_{12}X_1X_2^2$$

monomial responses:

$1, \quad X_1, \quad X_2, X_2^2, \quad X_1X_2, X_1X_2^2$

$L =$

$\{(0,0), (1,0), (0,1), (0,2), (1,1), (1,2)\}$

1	X_1	X_2	X_2^2	X_1X_2	$X_1X_2^2$
1	-1	-1	1	1	-1
1	-1	0	0	0	0
1	-1	1	1	-1	-1
1	1	-1	1	-1	1
1	1	0	0	0	0
1	1	1	1	1	1

Design and design ideal

- Each finite set of points $\mathcal{D} \subseteq \mathbb{Q}^m$ is the set of the solutions of a system of polynomial equations.
- Each real response on $\mathcal{D}$ is (identified with) a polynomial function with real coefficients.
- Different polynomials can be aliased on $\mathcal{D}$, but there are *minimal* representation.

- The solutions of the equation $x^3 - x = 0$ are $\mathcal{D} = \{-1, 0, +1\}$.
- The response

x	-1	0	$+1$
y	-2	1	2

  is represented on the design points by

$$f(x) = 1 + 2x - x^2$$

- The polynomial

$$g(x) = 1 + x - x^2 + x^3$$

  is aliased with f on $\mathcal{D}$ because $g(x) - f(x) = x^3 - x = 0$ on $\mathcal{D}$.

A language from Commutative Algebra

- Consider a number field of real numbers, here the field of rational numbers $\mathbb{Q}$, and the ring of polynomials with indeterminates $x_1, \ldots, x_d$ and coefficients in the field $\mathbb{Q}[x_1, \ldots, x_d]$.
- A *design* $\mathcal{D}$ is a finite set of distinct points in the vector space $\mathbb{Q}^d$.

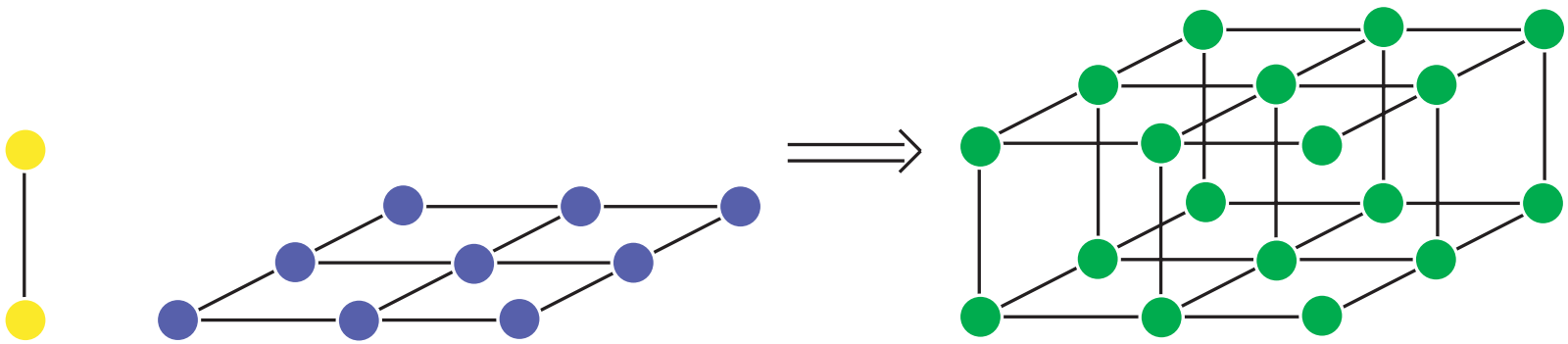
- The *design ideal* is the set of all polynomials in the ring that are zero on all points of the design.
- The design ideal has a finite number of generators. A set of generators is called a set of *generating equations* of the design

- If the polynomial $f \in \mathbb{Q}[x]$ is zero at $-1, 0, +1$, it is divided by $(x+1)x(x-1) = x^3 - x$. The design ideal is the set of all polynomials of the form $q(x)(x^3 - x)$.
- The generating polynomial is $x^3 - x$.

Operations on designs 1

Product of designs

$\mathcal{D}_1 \subset \mathbb{Q}^{m_1}$ $\mathcal{D}_2 \subset \mathbb{Q}^{m_2}$, $\mathcal{D}_1 \times \mathcal{D}_2 \subset \mathbb{Q}^{m_1+m_2}$.



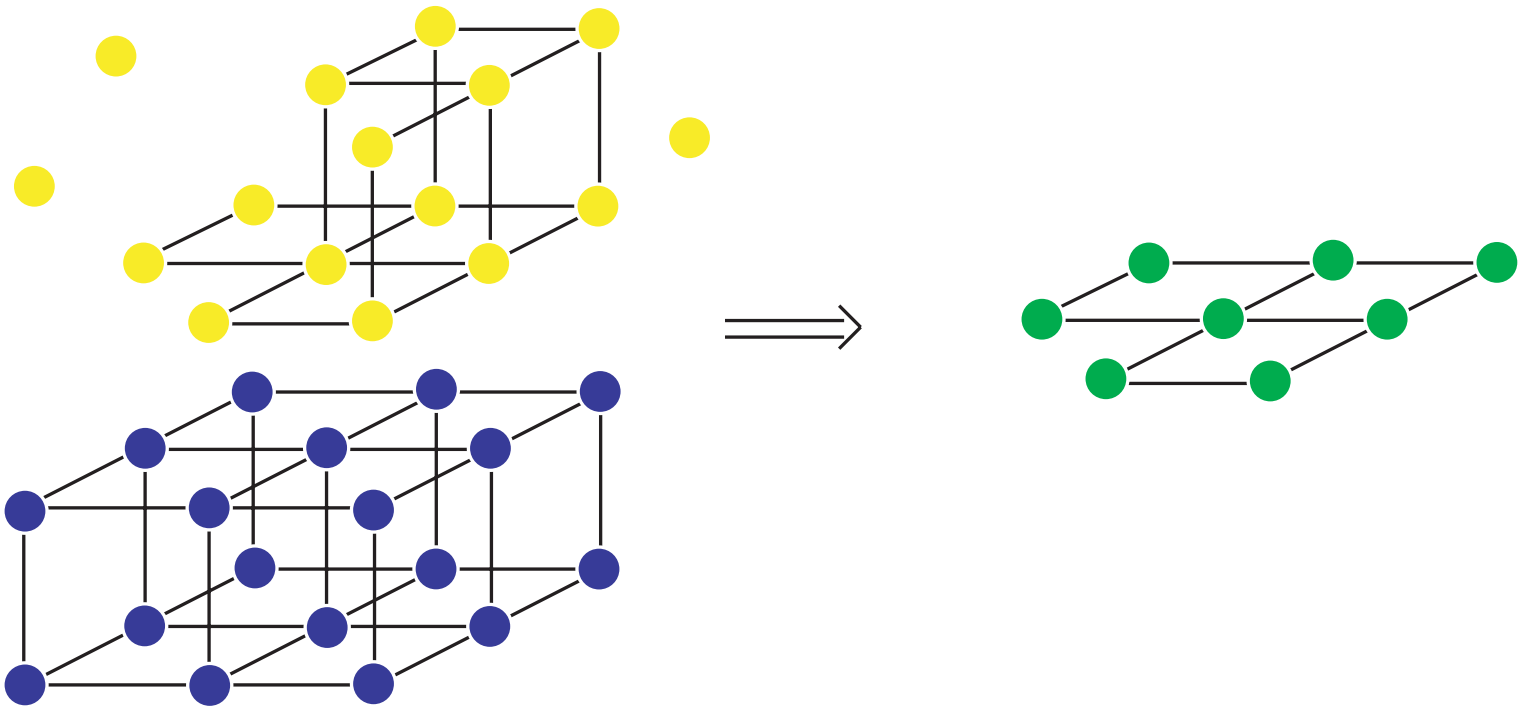
$\mathsf{Ideal}\,(\mathcal{D}_i) = I_i$, $\mathsf{Ideal}\,(\mathcal{D}_1 \times \mathcal{D}_2) = \mathsf{Ideal}\,(I_1, I_2)$

Operations on designs 2

Intersection

$\mathcal{D} \subset \mathbb{Q}^m$ $I = I(\mathcal{D})$

J ideal in $\mathbb{Q}[x_1, \ldots, x_m]$



$I + J$ ideal in $k[x_1, \ldots, x_m]$

$$I + J = \{f + g \mid f \in I \; g \in J\}$$

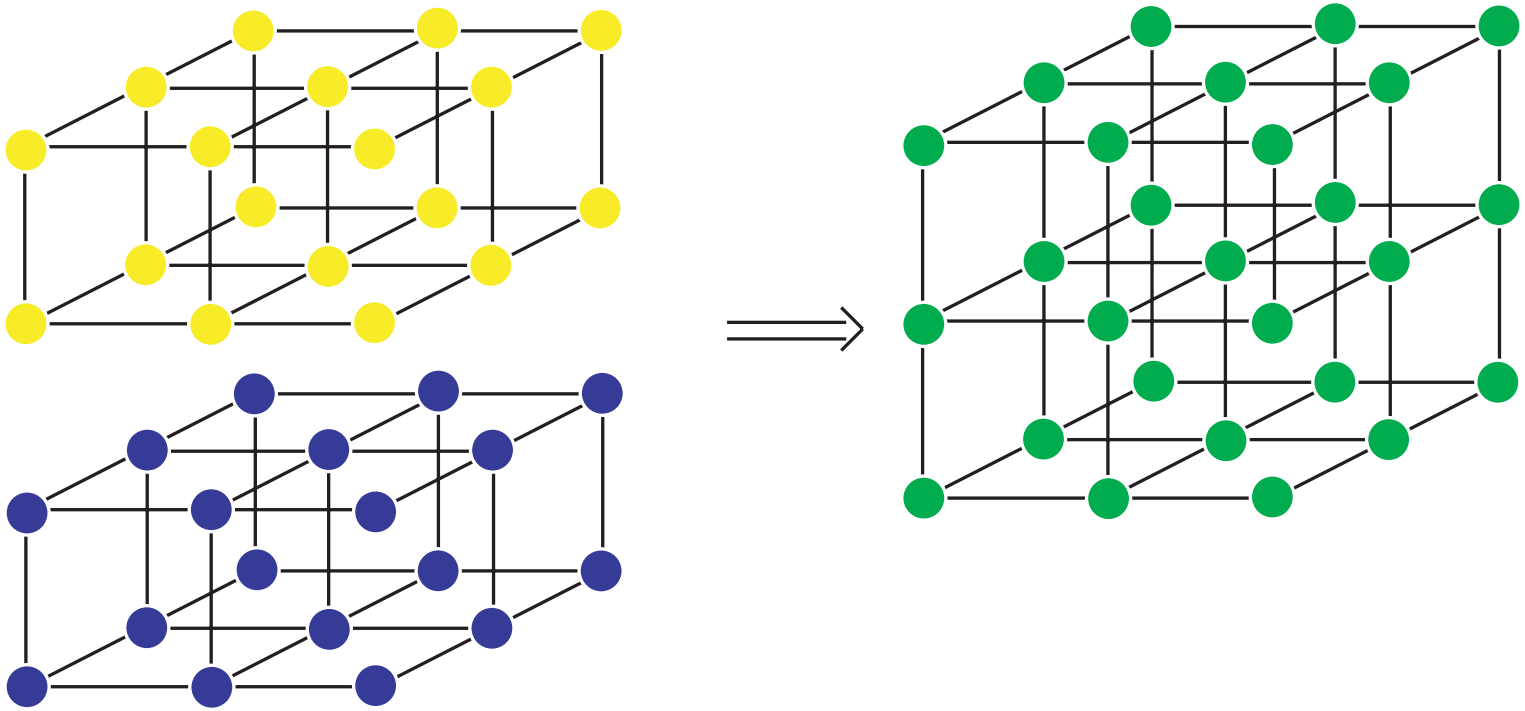
$$\text{Variety}(I + J) = \mathcal{D} \cap \text{Variety}(J)$$

Operations on designs 3

Union

$\mathcal{D}_1, \mathcal{D}_2 \subset \mathbb{Q}^m$

$\mathcal{D}_1 \cup \mathcal{D}_2 \subset \mathbb{Q}^m$



Basic theory

Definition

A *term order* is

- a total order $\prec$ on monomials x^α, s.t.
- $1 \prec x^\alpha$,
- $x^\alpha \prec x^\beta$ iff $x^{\alpha+\gamma} \prec x^{\beta+\gamma}$.

Definition

- Given a term order, the *leading term* $\mathrm{LT}(f)$ of a polynomial $f \in \mathbb{Q}[x_1, \ldots, x_d]$ is identified.
- A generating set $\mathcal{G} = \{g_1, \ldots, g_k\}$ of the design ideal I is a *Gröbner basis* if the set of leading terms of the design ideal I is generated by $\{\mathrm{LT}(g_1), \ldots, \mathrm{LT}(g_k)\}$

Theorem

- *Many term orders exists.*
- *There is a finite test for Gröbner basis.*
- *There is a finite algorithm that produces a G-basis $\mathcal{G}$ from any generating set.*
- *The set of all monomials that are not divided by a leading term in the G-basis form a linear basis of the space of responses.*

2^{3-1}

The fraction

$$\mathcal{F} = \begin{array}{ccc} +1 & +1 & +1 \\ -1 & -1 & +1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \end{array}$$

has design ideal I generated by

$$\mathcal{B} = \begin{cases} x^2 - 1, \\ y^2 - 1, \\ z^2 - 1, \\ xyz - 1. \end{cases}$$

which is not a G-basis.

In fact, the polynomial $xy - z$ belongs to the design ideal I, but

$$\mathrm{LT}\,(xy - z) = xy$$

cannot be obtained from lhe LT's of $\mathcal{B}$. The G-basis is

$$\mathcal{G} = \begin{cases} x^2 - 1, \\ y^2 - 1, \\ z^2 - 1, \\ xy - z, \\ xz - y, \\ yz - x. \end{cases}$$

The linear basis is $1, x, y, z$.

CoCoA 2^{3-1}

```
Use R::=Q[x,y,z];                          --- Defines the ring
List:=[x^2-1,y^2-1,z^2-1,xyz-1];           --- polynomials in a basis
I:=Ideal(List);                            --- computes the ideal
G:=GBasis(I);G;                            --- computes the G-basis

-------------------------------------------------------
---            ___/       ___/        \              ---
--            /     _ \  /     _ \    , \              --
--            \    |   | \    |   |  ___ \             --
---           ____, __/  ____, __/ _/    _\           ---
-------------------------------------------------------
--     Version       : 4.7.3                          --
--     Online Help   : type  ?   or   ?keyword        --
--     Web site      : http://cocoa.dima.unige.it     --
-------------------------------------------------------


--------------------------------
-- The current ring is R ::= Q[x,y,z];
--------------------------------
[z^2 - 1, y^2 - 1, x^2 - 1, -xy + z, yz - x, xz - y]
--------------------------------
```

CoCoA 1FAT

```
Use R::=Q[x,y];              --- Defines the ring
List:=[xy,x^2+y^2-1];        --- polynomials in a basis
I:=Ideal(List);              --- computes the ideal
G:=GBasis(I);G;              --- computes the G-basis

...

-------------------------------
[x^2 + y^2 - 1, xy, y^3 - y]
-------------------------------
```

- The monomial basis is
$$1, x, y, y^2,$$
- the interaction is not estimable as $xy = 0$.

CoCoA I

CoCoA program by M.P. Rogantin.

```
-------------------
-- Given a fraction (PointsF) and the full factorial design (D)
--   > the ideal of the fraction: IdF
--   > a list of estimable terms: Est
--   > the model matrix of a given order:
--        ModelMat(Order,PointsF)
--   > the change of basis matrix:
--        MatNF(Order,Est,IdF)
--   > the Normal Forms of the interaction terms of a
--   >     given order: NormForm(Order,Idea);
--   > a test for pair resolution:
--        TestResol(R,Est,Id)
--   > the indicator function of the fraction:
--         InFu(PointsF,D) and Indicat(PointsF)
-------------------
```

CoCoA II

```
Use R::=Q[x[1..9]];
D:=[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1
PointsF:= [
 [ 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,-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, -1, -1, 1,-1,-1, 1, 1] ,
```

CoCoA III

```
 [-1,-1, -1, -1,-1, 1, 1,-1,-1]];

IdF:=IdealOfPoints(PointsF);
Est:=QuotientBasis(IdF);Est;

-- INTERACTIONS
-- Interaction terms of a given order
Define Interac(Order);
  L := Subsets(Indets(),Order);
  P := [];
  For I := 1 To Len(L) Do
    Append(P, Product([L[I,J]| J In 1..Len(L[I])]))
  End;
  Return P;
End;

-- Interaction terms up to a given order
Define Model(Order);
```

CoCoA IV

```
 Inte:=[1];
   For I:= 1 To Order Do
    Inte:=Concat(Inte,Interac(I));
   End;
Return Inte;
End;

-- Model matrix of a given order
Define ModelMat(Order,Points);
Inte:=Model(Order);
LMod:=Len(Inte);
LF:=Len(Points);
W:=NewList(LF,LMod);
W:=[[Eval(Inte[J],Points[I]) | J In 1..LMod]
                | I In 1..LF];
UnSet Indentation;
Return Mat(W);
End;
```

CoCoA V

```
-- Change of bases matrix
Define MatNF(Order,Est,Idea);
 Inte:=Model(Order);
 LMod:=Len(Inte);
 LF:=Len(Est);
 NFor:=NewList(LMod);
 NFor:=[NF(P,Idea) | P In Inte];
 D:=NewList(LF,LMod);
 D:=[[CoeffOfTerm(Est[I],NFor[J])
           | J In 1..LMod] | I In 1..LF];
 Unset Indentation;
 Return Mat(D);
End;

-- Normal Forms of the terms of a given order
Define NormForm(Order,Idea);
```

CoCoA VI

```
 Inter:=Interac(Order);
 LIn:=Len(Inter);
 NFor:=NewList(LIn);
 NFor:=[NF(P,Idea) | P In Inter];
 Return NFor;
End;

-- Test for pair resolution
Define TestResol(R,Est,Idea);
 Order:=R/2;
 NFor:=NormForm(Order,Idea);
 St:=Order-1;
 L:=Sum([Bin(NumIndets(),I) | I In 0..St]);
 LIn:=Bin(NumIndets(),Order);
 CNF:=NewList(L,LIn);
 CNF:=[[CoeffOfTerm(Est[I],NFor[J])
                | J In 1..LIn] | I In 1..L];
        -- The NForms are in column
```

CoCoA VII

```
 K:=FALSE;
 For J:=1 To LIn Do
  For I:=1 To L Do
   IF CNF[I,J]<> 0 Then Return K  End;
  End;
 End;
 K:=TRUE;
 Return K;
End;

--Indicator Funtion (First way)
Define InFu(Points,D);
ND:=Len(D);
PA:=NewList(ND,0);
P:=NewList(ND);
 For H:=1 To Len(Points) Do
  For K:=1 To Len(PA) Do If Points[H]=D[K] Then PA[K]:=1  End; E
 End;
```

CoCoA VIII

```
IdD:=IdealAndSeparatorsOfPoints(D);
  For K:=1 To Len(PA) Do  P[K]:=PA[K]*IdD.Separators[K]  End;
F:=Sum(P);
Return F;
End;


--Indicator Funtion (Second way)
Define Indicat(Points);
 LF:=Len(Points);
 N:=NumIndets();
 LMod:=2^N;
 W:=ModelMat(N,Points);
 Ter:=Model(N);
 F:=Sum([Sum([W[I,J] | I In 1..LF])/LMod *Ter[J]| J In 1..LMod])
 Return  F;
End;
```

CoCoA IX

```
---


-- Orthogonalities

Define Delta(L,M);
D:=Diff(Set(Concat(L,M)),Intersection(L,M));
Sort(D);
Return D;
End;

Define Orthog(IndFunction);
UnSet Indentation;
H:=1..NumIndets();
Li:=Support(IndFunction);
T:=NewList(Len(Li),0);
LL:=NewList(Len(Li),0);
For I:=1 To Len(Li) Do
```

CoCoA X

```
  T[I]:=Log(Li[I]);
LL[I]:=NonZero([T[I,J]*H[J]|J IN 1..9]);
End;
LiInd:=Subsets(H);
Li2:=LiInd;
Len2:=Len(LL)-1;
For I:=1 To Len2 Do
  LiInd:=Li2;
  LenL:=Len(LiInd);
  For J:=I+1 To LenL Do
   D:=Delta(LiInd[I],LiInd[J]);
   If D IsIn LL Then
         Li2:=Diff(Li2,[LiInd[J]]);
         J:=J-1;
         End;
  End;
End;
Return LiInd;
```

CoCoA XI

```
End;


-----


Est;
InFun1:=InFu(PointsF,D);
InFun2:=Indicat(PointsF);
LO:=Orthog(InFun2);LO;


LD:=[x[I]^2-1 | I In 1..NumIndets()];
Set Indentation;
L:=Concat(LD,[InFun1-1]);
IdF:=Ideal(L);
NFor2:=NormForm(3,IdF);NFor2;
```