Algebraic Statistics

Part 1: **Eva Riccomagno**
Department of Mathematics, Università di Genova
riccomagno@dima.unige.it

Part 2: **Henry P. Wynn**
Department of Statistics, London School of Economics
H.Wynn@lse.ac.uk

... with the support of: **Hugo Maruri-Aguilar**
Department of Statistics, London School of Economics
H.Maruri-Aguilar@lse.ac.uk

Second de Brún Workshop Galway 30 June, 2009

Rationale

Polynomials and ratios of polynomials appear in statistics and probability under various forms, in model representations as well as in inferential procedures.

Algebraic geometry studies (ratios of) polynomials and the zeros set of systems of polynomial equations.

Algebraic statistics uses techniques from (real, computational) algebraic geometry, and commutative algebra, geometric combinatorics, ... to gain insight into the structure and properties of statistical models and to advise in model analysis.

This, in turn, may prompt research in algebraic geometry.

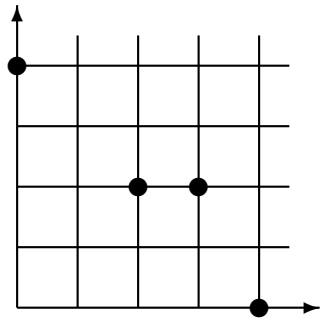
Statistica Sinica: special issue on Algebraic Statistics and Computational Biology, 17:4 (2007).

Journal of Symbolic Computation: special issue on Computational Algebraic Statistics, 41:2 (2006).

* From designs to polynomial ideals

$R = \mathcal{K}[x_1, \ldots, x_k]$

Finite set of points: settings for experiments or support for discrete random vectors.



$$\begin{array}{c} \mathcal{D} \\ \hline (0,4) \\ (2,2) \\ (4,0) \\ (3,2) \end{array}$$

$$\begin{cases} g_1 := x(x-2)(x-4)(x-3) \\ g_2 := (y-4)(y-2)y \\ g_3 := (y-2)(x+y-4) \\ g_4 := (x-3)(x+y-4) \end{cases}$$

$$\text{Ideal}(\mathcal{D}) = \left\{ \sum_{i=1}^{4} f_i(x,y) g_i(x,y) : f_i \text{ are polynomials} \right\}$$

Move the focus from $\mathcal{D}$ to Ideal($\mathcal{D}$).

Finite set of points in k dimensions are zero-dimensional algebraic varieties. A polynomial ideal is associated to an algebraic variety.

Some properties of Ideal($\mathcal{D}$)

Polynomial ideals are generated by a finite number of polynomials (Hilbert basis theorem). Gröbner bases are particular generator sets.

Definition: the radical ideal of a polynomial ideal I is

$$\sqrt{I} = \{f \in R : f^m \in I \text{ for some } m \in \mathbb{Z}_{>0}\}$$

Ideal I is radical if $I = \sqrt{I}$.

Two polynomial ideals generate the same varieties if their radical ideals are equal (cf. Strong Nullstellensatz for algebraically closed fields).

Computation of Ideal($\mathcal{D}$)

Ideal($\mathcal{D}$) is the intersection of ideals of single design points, computable from the coordinates of the points.

Efficient algorithms to determine

1. Ideal($\mathcal{D}$) given the coordinates of the points in $\mathcal{D}$
2. Ideal($\mathcal{D}$) given the structure of $\mathcal{D}$
3. the coordinates of $\mathcal{D}$ given Ideal($\mathcal{D}$)
4. change generator sets
5. operations on designs traduce to operations on ideals

Gröbner bases are a main computational tool.

1. Ideal($\{d\}$) $= \langle x_1 - d_1, \ldots, x_m - d_m \rangle$ and Ideal($\mathcal{D}$) $= \bigcap_{d \in \mathcal{D}}$ Ideal($\{d\}$) or... easier CoCoA with `IdealOfPoints`

Abbott J., Bigatti A., Kreutzer M., and Robbiano L. (2000) Computing ideals of points. J. Symb. Comput. 30, 341356.

2. ...

3. solve systems for polynomial equations $\begin{cases} x^2 - x = 0 \\ y^2 - y = 0 \\ xy = 0 \end{cases} \quad \forall d \in \mathcal{D}$

4. FGLM algorithm

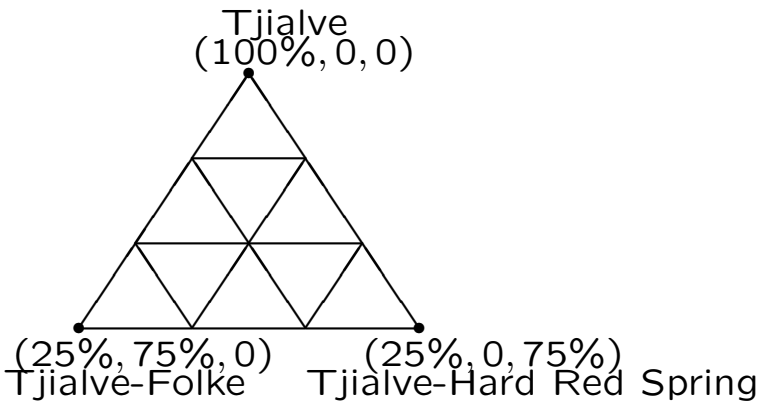
Faugère J.C., Gianni P., Lazard D., and Mora T. (1993) Efficient computation of zero-dimensional Gröbner bases by change of ordering. J. Symbolic Comput. 16, 329344.

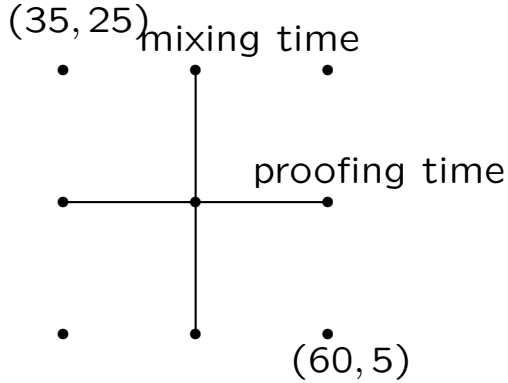
5. add/remove equations to/from Ideal($\mathcal{D}$) and remove/add points to $\mathcal{D}$

2. One of the key parameters of bread quality is loaf volume.

Five main factors affect it: three wheat flours $(x_1, x_2, x_3) \subset \Delta_2 \subset \mathbb{R}^3$, mixing time $z_1 \in \mathbb{R}_{\geq 0}$ and proving time of the dough $z_2 \in \mathbb{R}_{\geq 0}$.

The experiment consists of the cross-product of a simplex lattice designs and a full factorial design.





$$\text{Ideal}(Simplex \times Full factorial) = \langle f_1, f_2 : f_1 \in \text{Ideal}(Simplex), f_2 \in \text{Ideal}(Full factorial) \rangle$$

T Naes, E M Faergestad, J Cornell (1998). A comparison of methods for analyzing data from a three component mixture experiment in the presence of variation created by two process variables. Chemometrics and Intelligent Laboratory Systems 41:221-235.

Gröbner bases

A monomial or term ordering on R, τ, is an ordering relation on the terms in R s.t. • $x^\alpha \succ 1$ for all α and • if $x^\alpha \succ x^\beta$ then $x^{\alpha+\gamma} \succ x^{\beta+\gamma}$ for all α, β, γ.

Definition: $g_1, \ldots, g_t \in R$ form a Gröbner basis w.r.t a τ if

$$\langle \mathsf{LT}_\tau(g_1), \ldots, \mathsf{LT}_\tau(g_t) \rangle = \langle \mathsf{LT}_\tau(\mathsf{Ideal}(g_1, \ldots, g_t) \rangle$$

Properties of Gröbner bases

1. A τ-Gröbner basis $\{g_1, \ldots, g_t\} \subset I \subset R$ is a generator set of I.
2. There are algorithms to compute Gröbner bases from any generator set of I (Buchberger's algorithm, S-polynomials).
3. The set of reduced Gröbner bases obtained by varying τ is finite.
4. Elimination ...
5. For $f \in R$, τ and $G = \{g_1, \ldots, g_t\}$ τ-Gröbner basis, there exist unique $g \in \langle g_1, \ldots, g_t \rangle$ and $r \in R$ such that $f = g + r$ with $g = \sum_{i=1}^{t} f_i g_i$, $f_i \in R$, and no term of r is divisible by $\mathrm{LT}_\tau(g_i)$, $g_i \in G$.
6. The τ-reduced Gröbner basis is unique.

Definition: A Gröbner basis G is reduced if for all $g \in G$ $\mathrm{LC}_\tau(g) = 1$ and no term of g lies in $\langle \mathrm{LT}_\tau(\mathrm{Ideal}(G \setminus \{g\}) \rangle$.

The space of functions over a design

The set of real polynomial functions over $\mathcal{D}$, the coordinate ring of $\mathcal{D}$, is isomorphic to a computable space of polynomials

$$\mathcal{L} = \mathbb{R}[\mathcal{D}] = \{f : \mathcal{D} \longrightarrow \mathbb{R}\} \sim_{\mathbb{R}} \mathbb{R}[x_1, \ldots, x_k]/\,\mathrm{Ideal}(\mathcal{D})$$

Computations are performed via Gröbner bases of Ideal($\mathcal{D}$) `GBasis`

They are finitely many Gröbner bases for a polynomial ideal.

A vector space basis of $\mathcal{L}$ can be easily derived from a Gröbner basis of Ideal($\mathcal{D}$) `QuotientBasis`

* Software

CoCoA is a CCA software developed at the University of Genova.

A number of other CCA systems are available:

Maple `http://www.maplesoft.com`,

Mathematica `http://www.wolfram.com`,

Singular `http://www.singular.uni-kl.de` ....

We use in particular CoCoA and Maple for general purpose algebraic computations. In special cases we use

R `http://www.r-project.org` for computations oriented to statistics;

4ti2 `http://www.4ti2.de` for special computations needed for combinatorial problems.

A CCA software makes exact computations on number fields e.g. $\mathbb{Q}$, $\mathbb{Z}_p$ ..., on rings of polynomials e.g. $k[x,y,z]$, on ideals e.g. $I = \langle x-y, y-z\rangle$, $I+J$, IJ, quotient rings e.g. $k[x,y,z]/\langle x-y, y-z\rangle$.

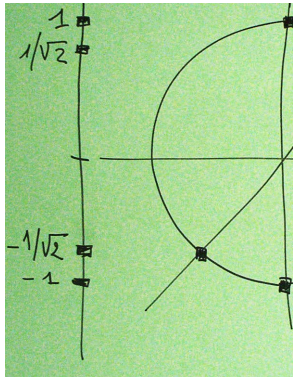
Example*

```
Use R::= Q[x[1..4]], Lex;          -- defines the polynomial ring
Ord(R);                            -- term order
Eqs := [4x[1]+2x[2]-x[3]-x[4] - 1,2x[1]-x[2]-x[3]-4x[4],x[1]+2x[2]-4x[3]+3x[4] +2 ,
        x[1]+2x[2]+3x[3]+4x[4]];   -- generating equations
Idl := Ideal(Eqs); Idl;            -- defines the ideal
GBasis(Idl);                       -- computes a Groebner basis

Mat([
  [1, 0, 0, 0],
  [0, 1, 0, 0],
  [0, 0, 1, 0],
  [0, 0, 0, 1]
])
------------------------------
Ideal(4x[1]+2x[2]-x[3]-x[4]-1, 2x[1]-x[2]-x[3]-4x[4],
x[1]+2x[2]-4x[3]+3x[4]+2, x[1]+2x[2]+3x[3]+4x[4])
------------------------------
[4x[1]+2x[2]-x[3]-x[4]-1, 7x[3]+x[4]-2, 5/2x[2]+11/2x[4]+1, 17/35x[4]+81/140]
------------------------------
```

*Structure, elimination, complex roots of unity

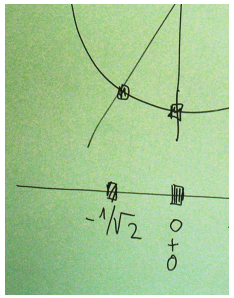
$$\begin{cases} x^2+y^2-1=0, \\ x(x-y)=x^2-xy=0 \end{cases} \Longrightarrow \begin{cases} x+2y^3-2y=x-2y(1-y^2)=0, \\ -2y^4+3y^2-1=(2y^2-1)(y^2-1)=0 \end{cases}$$

```
Use R::= Q[xy], Lex; Eqs := [x^2+y^2-1,x(x-y)];
Idl := Ideal(Eqs); Eqs; Idl; GBasis(Idl);

[x^2 + y^2 - 1, x^2 - xy]
------------------------------
Ideal(x^2 + y^2 - 1, x^2 - xy)
------------------------------
[x + 2y^3 - 2y, -2y^4 + 3y^2 - 1]
------------------------------
```



$$\begin{cases} x^2+y^2-1=0, \\ x(x-y)=x^2-xy=0 \end{cases} \Longrightarrow \begin{cases} -xy+x^2=0, \\ y^2+x^2=1, \\ -2x^3+x=0. \end{cases}$$

```
Use R::= Q[xy], Xel; Eqs := [x^2+y^2-1,x(x-y)];
Idl := Ideal(Eqs); Eqs; Idl; GBasis(Idl);

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



* Experiments with mixtures/compositional data

"In the general mixture problem, the measured response is assumed to depend only on the proportions of the ingredients present in the mixture and not on the amount of the mixture".* A mixture-amount experiment is a mixture experiment performed at two or more levels of total amount. At times process variables are considered (bread loaf).

The sample space for a k factor experiments is (a subset of) the $(k-1)$-simplex

$$\sum_{i=1}^{k} x_i = 1 \qquad 0 \leq L \leq x_i \leq U \leq 1$$

In response surface theory we assume the existence of a continuous function over the simplex which we then approximate

$$\eta = \phi(x_1, \ldots, x_k)$$

Usual assumptions are made: for n trials and with y_u the measured response at trial u, we assume that

$$y_u = \eta_u + \varepsilon_u$$

where the ε_u are uncorrelated and identically distributed with zero mean and equal variance.

* J. A. Cornell. *Experiments with mixtures*. John Wiley & Sons, New York, third edition, 2002. page 4

The functional constraint on the factors has consequences on the structure of the measured response models which we take to be polynomial

1. models can be written in different ways (aliasing/confounding): $k = 2$

$$\alpha_0 + \alpha_1 x_1 = \alpha_0(x_1 + x_2) + \alpha_1 x_1 = \alpha_0 x_2 + (\alpha_0 + \alpha_1)x_1$$

2. in particular trials at $(0, \ldots, 0)$ are not allowed and the usual interpretation of the intercept does not make sense

3. general interpretation of the other coefficients is not as usual e.g. slope,

4. notions of effect and interactions are not obvious

5. ...

Bibliography

J. A. Cornell. *Experiments with mixtures*. John Wiley & Sons], New York, third edition, 2002.

J. Aitchison. *The statistical analysis of compositional data*. Chapman & Hall, London, 1986.
——

H. Scheffé. Experiments with mixtures. *J. Roy. Statist. Soc. Ser. B*, 20:344–360, 1958.

H. Scheffé. The simplex-centroid design for experiments with mixtures. *J. Roy. Statist. Soc. Ser. B*, 25:235–263, 1963.

R.D. Snee and D.W. Marquardt. Screening concepts and designs for experiments with mixtures. *Technometrics*, 18:19–29, 1976.
——

P. J. Claringbold. Use of the simplex design in the study of joint action of related hormones. *Biometrics*, 1955.

B.J. McConkey, P.G. Mezey, D.G. Dixon, and B.M. Grenberg. Fractional simplex designs for interaction screening in complex mixtures. *Biometrics*, 56:824–832, 2000.

D. R. Cox. A note on polynomial response functions for mixtures. *Biometrika*, 1971.

G. F. Piepel, R. D. Hicks, J. M. Szychowski, and J. L. Loeppky. Methods for assessing curvature and interaction in mixture experiments. *Technometrics*, 44(2):161–172, 2002.
——

J. N. Darroch; J. Waller. Additivity and interaction in three-component experiments with mixtures. *Biometrika*, 1985.

N. R. Draper and F. Pukelsheim. Mixture models based on homogeneous polynomials. *J. Statist. Plann. Inference*, 71(1-2):303–311, 1998.

Pairs of (canonical polynomials, designs) - Effect of functional constraints on parameter estimation

Often none of those designs for canonical polynomials can be considered e.g. because

1. a fraction must be considered
   (a) the number of points in the design is large for k and m large (Scheffé 1963, Appendix B)
   (b) there are missing observations
   (c) measuraments are taken only in a subsets of the sample space ...
2. there is a mixed number of levels
3. there are process variables
4. ...

What models are identifiable? What terms are confounded? How to interpret the parameters in terms of e.g. curvature?

Example: simplex centroid design in three factors

$$\mathcal{D} = \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/3 & 1/3 & 1/3 \end{array} \qquad \mathcal{C}_{\mathcal{D}} = \begin{array}{ccc} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \\ d & d & 0 \\ 0 & e & e \\ f & 0 & f \\ g & g & g \end{array} = \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{array}$$

Homogeneous model function of degree n

We choose η s.t. $\boxed{\eta(tx_1,\ldots,tx_k) = t^s\eta(x_1,\ldots,x_k)}$ where $t > 0$ represents the total mixture amount and s is the degree of the homogeneous model. Examples

$$s = 2 \quad ((tx_1)^2 - (tx_2)^2 + 3(tx_3)^2) = t^2(x_1^2 - x_2^2 + 3x_3^2)$$

$$s \qquad \beta_1\tfrac{x_1}{x_2+x_3} + \beta_2\tfrac{x_2}{x_1+x_3} + \beta_3\tfrac{x_3}{x_1+x_2}$$

$$s = 1 \quad \textstyle\sum_i \beta_i x_i + \sum_{i<j} \beta_{ij} \min(x_i, x_j)$$

Moreover we choose η to be polynomial. Various arguments support this choice.

1. When the total mixture amount is t a homogeneous model of degree s ensures that all terms of the model are affected by the same multiple t^s

2. Assume the model $\eta_3 = f(x_1, x_2, x_3)$. A fourth component is introduced which is thought to have an additive effect w.r.t any mixture of the first three components. Cornell (2002, pag 301) represents this effect as $\eta_4 = \beta_4 x_4 + (1 - x_4) f(\frac{x_1}{1-x_4}, \frac{x_2}{1-x_4}, \frac{x_3}{1-x_4})$. Invariance under addition of the addive component implies

$$\begin{aligned} f(x_1, x_2, x_3) &= (1 - x_4) f(\tfrac{x_1}{1-x_4}, \tfrac{x_2}{1-x_4}, \tfrac{x_3}{1-x_4}) \\ &= (x_1 + x_2 + x_3) f(\tfrac{x_1}{x_1+x_2+x_3}, \tfrac{x_2}{x_1+x_2+x_3}, \tfrac{x_3}{x_1+x_2+x_3}) \end{aligned}$$

3. Draper and Pukelsheim (1998) use Kronecker products to define homogeneous polynomial models

4. Projective varieties and homogeneous polynomials are naturally associated: $d_1 = (1, 4, 2)$ and $d_2 = (2, 8, 4)$ represent the same projective point in $\mathbb{P}^2(\mathbb{R})$. Note that $x_2 - x_3^2$ vanishes on d_1 but not on d_2 while $x_2^2 - x_3^2$ vanishes on both d_1 and d_2

If $\mathcal{D} \subset \mathcal{K}^k$ is a compositional data set then

$$x_1 + \ldots, x_k - 1 \in \text{Ideal}(\mathcal{D})$$

In particular retrieve only slack models, where one factor is not present at all/is totally confounded.

The missing factor can be reintroduced by

1. substituting the constant term with the $\sum_{i=1}^{k} x_i = 1$ condition

2. homogenising the slack model

3. use projective geometry and adapting the procedure above to homogeneous polynomial models*

* fairly standard... N. R. Draper and F. Pukelsheim. Mixture models based on homogeneous polynomials. *J. Statist. Plann. Inference*, 71(1-2):303–311, 1998... use Kronecker products

Example... $\mathcal{D} = \{(0,0,1),(0,1,0),(1,0,0),(\frac{1}{3},\frac{1}{3},\frac{1}{3})\}$. Then

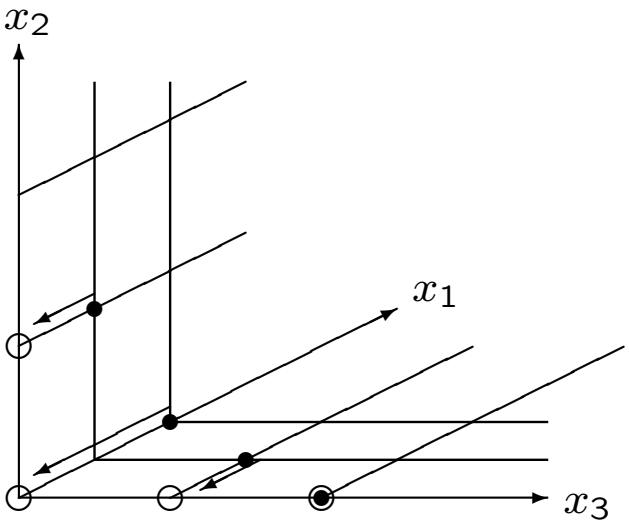
$$\text{Ideal}(\mathcal{D}) = \text{Ideal}(\ \underline{x_1} + x_2 + x_3 - 1,\quad \underline{x_2^2} - x_3^2 - x_2 + x_3,\quad \underline{x_2x_3} + \tfrac{1}{2}x_3^2 - \tfrac{1}{2}x_3,\quad \underline{x_3^3} - \tfrac{4}{3}x_3^2 + \tfrac{1}{3}x_3)$$

For any term ordering s.t. $x_1 \succ x_2 \succ x_3$, obtain $\boxed{1, x_3, x_3^2, x_2}$.

1. $\beta_0 1 + \beta_1 x_3 + \beta_2 x_3^2 + \beta_3 x_2 = \beta_0 x_1 + (\beta_0 + \beta_1)x_3 + \beta_2 x_3^2 + (\beta_0 + \beta_3)x_2$ only a linear term in x_1

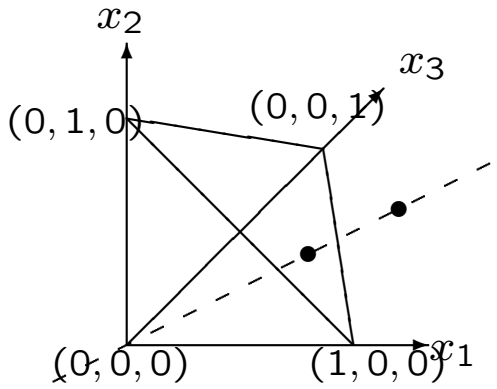
2. $\beta_0 x_1^2 + \beta_1 x_3 x_1 + \beta_2 x_3^2 + \beta_3 x_2 x_1$ or $\beta_0 x_1^3 + \beta_1 x_3 x_1^2 + \beta_2 x_3^2 x_1 + \beta_3 x_2 x_1^2$

Homogeneisation of slack models corresponds to orthogonal projection over $x_3 = 0$



full dots $x_1^2, x_3x_1, x_3^2, x_1x_2$ vs. empty dots $1, x_3, x_3^2, x_2$

... design cone and its ideal



$$\mathcal{D} = \{(1,0,0),(0,1,0),(0,0,1),(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3})\}$$

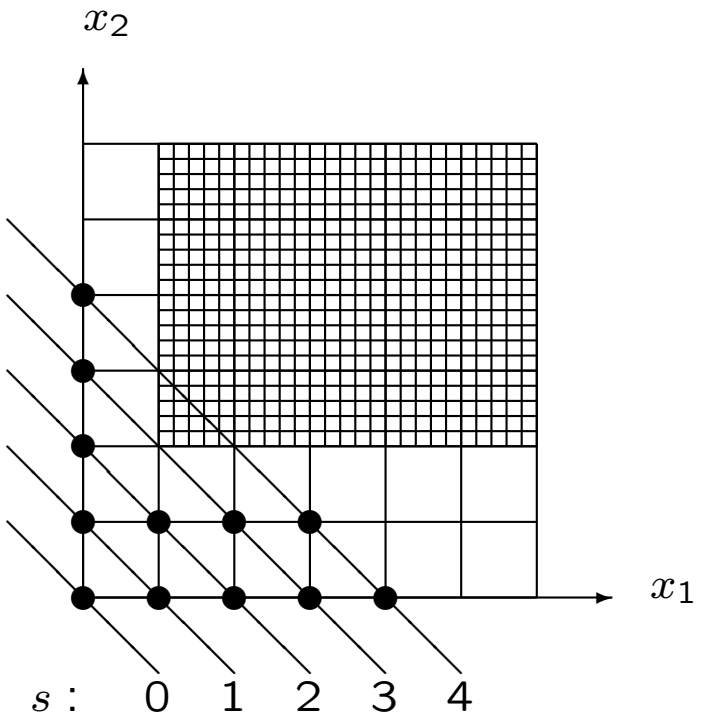
$$\text{Ideal}(\mathcal{C}_{\mathcal{D}}) = \left\langle \underline{x_1x_3} - x_2x_3, \underline{x_1x_2} - x_2x_3, \underline{x_2^2x_3} - x_2x_3^2 \right\rangle$$

design $\longrightarrow$ cones $\longleftrightarrow$ homogeneous ideals

Fix a degree s and obtain a basis of $\mathcal{L}$ formed by monomials of degree s

s	degree s standard monomials
1	x_1, x_2, x_3
2	x_1^2, x_2^2, x_2x_3, x_3^2
$s \geq 3$	x_1^s, x_2^s, $x_2x_3^{s-1}$, x_3^s

$\text{Ideal}(\mathcal{C}_{\{(0,1),(1,0),(1/2,1/2)\}}) = \text{Ideal}(\underline{x_1 x_2^2} - x_1^2 x_2)$



Summary

Theorem 1 *For a mixture design* $\mathcal{D}$
1. $\text{Ideal}(\mathcal{C}_\mathcal{D}) = \langle f \in R] : f$ *is homogeneous and* $f(d) = 0$ *for all* $d \in \mathcal{D}\rangle$*, that is the largest homogeneous ideal in* R *vanishing on all the points of* $\mathcal{D}$*.*
2. $\text{Ideal}(\mathcal{D}) = \text{Ideal}(\mathcal{C}_\mathcal{D}) + \langle \sum_{i=1}^{k} x_i - 1\rangle$*, that is a polynomial vanishing on* $\mathcal{D}$ *can be written as combination of homogeneous components vanishing on* $\mathcal{D}$ *and the sum to one condition.*
3. If G *is a generator set of* $\text{Ideal}(\mathcal{C}_\mathcal{D})$ *then* G *and* $\sum_{i=1}^{k} x_i - 1$ *form a generator set of* $\text{Ideal}(\mathcal{D})$*.*

Theorem 2 *Let* $\mathcal{D}$ *be a mixture design and* $\mathcal{C}_\mathcal{D}$ *its cone. Let* $G = \{l - 1, g_1, \ldots, g_r\}$ *be a Gröbner basis of* $\text{Ideal}(\mathcal{D})$ *with respect to a graded term ordering* τ*. Then* $\{h(g_1), \ldots, h(g_r)\}$ *is a generating set of* $\text{Ideal}(\mathcal{C}_\mathcal{D})$*, where* $h(g)$ *is the homogeneization of* g *with respect to* $l = \sum_{i=1}^{k} x_i$*.*

Theorem 3 *Let* $\mathcal{D}$ *be a mixture design. Then*

$$\dim \mathbb{R}[x_1, \ldots, x_k]_s/\text{Ideal}(\mathcal{C}_\mathcal{D})_s = \dim \mathbb{R}[x_1, \ldots, x_k]_{\leq s}/\text{Ideal}(\mathcal{D})_{\leq s}$$

If moreover $\mathcal{D}$ *has* n *distinct points and* s *is sufficiently large then the dimensions equal* n*.*

N.B. to perform some computations we used the Hilbert function of an ideal and the colon operation.

✻ On repeated measurements and replicated points

	units	$\longrightarrow$	design points	$\longrightarrow$	responses	
	loaf	$\longrightarrow$	setting	$\longrightarrow$	volume	
	Ω	$\overset{d}{\longrightarrow}$	$\mathcal{D} \subset \mathbb{R}^5$	$\overset{y}{\longrightarrow}$	$\mathbb{R}$ or $\mathbb{R}^2$	
A	ω	$\rightarrow$	d	$\rightarrow$	$y(d(\omega))$	distinct points
B	ω_1	$\rightarrow$	d	$\rightarrow$	$y(d(\omega_1)) = y(\omega_1)$	replicated points
	ω_2	$\nearrow$		$\searrow$	$y(d(\omega_2)) = y(\omega_2)$	
C	ω	$\rightarrow$	d	$\rightarrow$	$y^1(d(\omega)) = y^1$	multivariate response
				$\searrow$	$y^2(d(\omega)) = y^2$	
D	ω_i	$\rightarrow$	$d(\omega_i) \sim d$	$\rightarrow$	$y(d(\omega_i))$	error in variables/ random factors

For identifiability B, C and D present the same problem: some rows of the X-matrix for any choice of $f_1, \ldots, f_m \in \mathcal{L}(\mathcal{D})$, e.g. any subset of a standard basis, are identical.

Focus on case D. Namely, consider clouds of points with unknown coordinates. Each cloud is close to a point d whose coordinates are known. The measured responses for each point in a cloud $y_i = y(d(\omega_i))$ are known. We might include non replicated points as well.

I determine algebraic families of points whose given limit points have the correct multiplicity (replicate) $\leftrightarrow$ an analogue of Ideal($\mathcal{D}$)

II determine conditions that ensure the good behaviour of the interpolating polynomial $\leftrightarrow$ the analogue of $R/\,$Ideal($\mathcal{D}$).

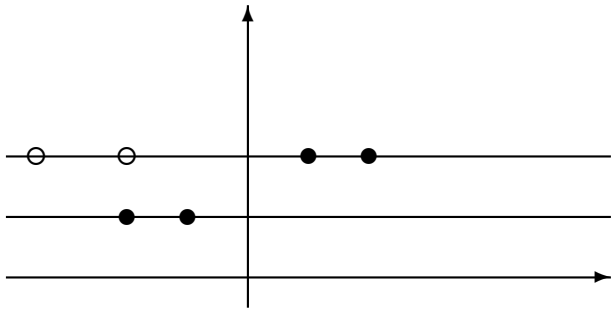
Part I: one replicated point at zero

Consider $d_i = (a_{1i}, \ldots, a_{ki})$, $i = 1, \ldots, r$ unknown distinct points close to 0 and $q_1, \ldots, q_n$ distinct points in $\mathbb{R}^k$.

1. Define $d_i(t) = (ta_{1i}, \ldots, ta_{ki})$ for $t \in \mathbb{R}$. Note $d_i(1) = d_i$ and $d_i(0) = 0$
2. For $t \in \mathbb{R}$ consider the family of distinct points $\mathcal{D}_t = \{d_1(t), \ldots, d_r(t), q_1, \ldots, q_n\}$
3. and the family of ideals Ideal($\mathcal{D}_t$) in $\mathbb{R}[x_1, \ldots, x_k, t] = S$.

For $q_1 = (1, 2), q_2 = (2, 2)$

and $d_1 = (-1, 1), d_2 = (-2, 1)$



Standard bases

Ideal($\mathcal{D}_t$) defines a flat family as for all $t_0 \in \mathbb{R}$:

$$\dim S/\langle \text{Ideal}(\mathcal{D}_t), t - t_0 \rangle = 0$$
$$\deg S/\langle \text{Ideal}(\mathcal{D}_t), t - t_0 \rangle = n + r \; (= \dim_{\mathbb{R}} S/\langle J, t - t_0 \rangle)$$

1. Use the CoCoA commands `Dim` and `Multiplicity` to check them.

2. For $t_0 \neq 0$ it is case A and we are interested in $t_0 = 0$.

For almost all $t_0 \in \mathbb{R}$ including $t_0 = 0$ there exists a monomial ideal $I \subset \mathbb{R}[x_1, \ldots, x_k]$ such that

$$\text{LT}(\text{Ideal}(\mathcal{D}_t), t - t_0) = \langle t, I \rangle$$

1. I does not depend on t_0.

2. I can be computed.

1. In particular $S/\langle J, t - t_0 \rangle \sim_{\mathbb{R}} R/I \sim_{\mathbb{R}} \mathsf{Span}(x^\alpha : x^\alpha \notin \mathsf{LT}(I))$ does not depend on t_0.

2. Furthermore I and the standard bases do not depend on the choice of the d_i's.

3. I is a partial analogue of $\mathsf{Ideal}(\mathcal{D})$ for distinct points.

4. Note that as I is a monomial ideal we have lost information on the aliasing/confounding structure of the design.

5. This construction generalises to more replicated points.

Example for five replicated points

For $i = 1, \ldots, 5$ let A_i be the limit points and X_i the clouds

$$
\begin{array}{ll}
X_1 = \{(0,0),(1,0),(0,1),(-1,0),(0,-1)\} & A_1 = (0,0) \\
X_2 = \{(2,1),(1,2)\}, & A_2 = (1,1) \\
X_3 = \{(-2,1),(-1,2)\}, & A_3 = (-1,1) \\
X_4 = \{(-2,-1),(-1,-2)\}, & A_4 = (-1,-1) \\
X_5 = \{(1,-2),(2,-1)\} & A_5 = (1,-1)
\end{array}
$$

We want to compute the limit ideal when collapsing X_i to A_i, $i = 1, \ldots, 5$ (we assume the collapsing process in independent for each cloud).

1. From Ideal$(X_1) = \langle xy, x^3 - x, y^3 - y\rangle$ obtain $J_1 = \mathrm{LT}(I(X_1)) = \langle xy, x^3, y^3\rangle$

2. Change coordinates and move A_2 to the origin, then A_2 and X_2 become e.g. $(0,0)$ and $\{(1,0),(0,1)\}$. In the new coordinate system Ideal$(X_2) = \langle X + Y - 1, Y^2 - Y\rangle$, giving $\mathrm{LT}(I(X_2)) = \langle X + Y, Y^2\rangle$. In the old system this becomes $J_2 = \langle x + y - 2, y^2 - 2y + 1\rangle$.

3. ...

4. Intersect all these limit ideals.

5. Using degrevlex with $x > y$, the intersection is generated by

$$\underline{x^3y} + xy^3 - 2xy,$$
$$\underline{x^4} + 4x^3y - 2x^2y^2 + 4xy^3 + y^4 - 8xy,$$
$$2\underline{y^5} + x^2y - 3y^3,$$
$$2\underline{xy^4} + x^3 - 3xy^2,$$
$$2\underline{x^2y^3} - x^2y - y^3.$$

We do not use explicitly the extra variable t, that is $\mathcal{D}_t$.

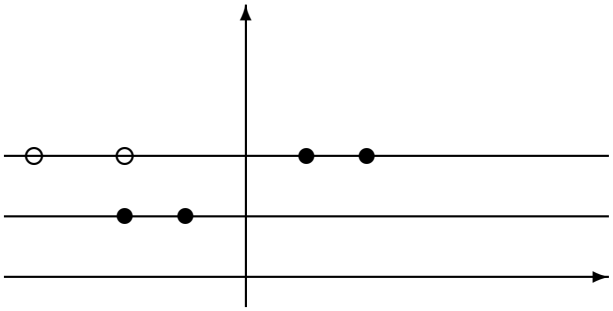
Part II: determine conditions that ensure the good behaviour of the interpolating polynomial $\leftrightarrow$ the analogue of $R/\,\text{Ideal}(\mathcal{D})$.

Consider the family of design/model matrices $X_t, t \in \mathbb{R}$, obtained by evaluating a given monomial basis at $\mathcal{D}_t$.

Example (contd):

For $q_1 = (1,2), q_2 = (2,2)$

and $d_1 = (-1,1), d_2 = (-2,1)$



and the monomial basis $\{1, x, y, xy\}$

$$X_t = \begin{pmatrix} 1 & 1 & 2 & 2 \\ 1 & 2 & 2 & 4 \\ 1 & -t & t & -t^2 \\ 1 & -2t & t & -2t^2 \end{pmatrix}$$

This is full rank for $t \in \mathbb{R}$ except for a set of zero Lebesgue measure.

Let y_i be the value observed at $q_i, i = 1, 2$ and y_3, y_4 the values at 0.

We need to choose 'responses' at the moving points $Y_t = [y_1, y_2, y_3(t), y_4(t)]$ and consider the linear system $Y_t = X_t\theta$ with symbolic solutions (Cramer rule)

$$\theta_i(t) = \frac{\det(X_{t,i})}{\det(X_t)} = (X_t^{-1}Y_t)_i$$

We require that $y_3(t), y_4(t)$ are chosen so that

- $\lim_{t\mapsto 0}\theta_i(t)$ exists finite for $i = 1, \ldots, 4 = n + r$
- $y_3(1) = y_3$, $y_4(1) = y_4$ and $y_3(0) = y_4(0) = a$
- $y_3(t), y_4(t)$ are polynomials of a small as possible degree.

a could be the mean value of the measured responses at the replicated points, the MLE under the mean square cost function.

Example (contd):

$$\begin{aligned}
\theta_1 &= -y_4(t) - 2y_3(t) \\
\theta_2 &= (y_3(t) - y_4(t))/t \\
\theta_3 &= (-2y_3(t) + 7 + y_4(t))/2 \\
\theta_4 &= -1 - (y_3(t) - y_4(t))/2t
\end{aligned}$$

1. The order of infinitesimal in $t = 0$ of $\det(X_t) = -t(t-2)^2$ is 1

2.

$$\begin{aligned}
Y_t(x,y) &= \theta_1(t) + \theta_2(t)x + \theta_3(t)y + \theta_4(t)xy \\
&\downarrow \quad t \mapsto 0 \\
\bar{Y}(x,y) &= a + 0.3x - \tfrac{(a-7)}{2}y + \tfrac{0.7}{2}xy
\end{aligned}$$

Projecting to the support

- Projecting the interpolating polynomial as limiting polynomials from the $\mathcal{D}_t$
- and computing the interpolating polynomial over the un-replicated design

yield the same set of identifiable monomials.

Example (contd): the normal form of $\bar{Y}(x, y)$ in $R/\,\text{Ideal}(q_1, \ldots, q_r, 0)$ corresponds to the interpolation of $q_1, \ldots, q_r, 0$ at $y_1, \ldots, y_r, a$ using a standard basis of $\text{Ideal}(q_1, \ldots, q_r, 0)$.

$$\text{NF}(\bar{Y}) = a + (7 - a)y/2 - 2x$$

The final interpolator depends on the chosen value a.

Comments

- As far as we could we based our proofs on techniques from linear algebra also in order to reduce computational complexity.
- Relax the assumption that the points move towards there central point along lines.
- Properly implement this.
- Retain/obtain information on the aliasing structure?
- Partition the X_t matrices so as to use a part to estimate the model variance?
- Exploit the differential aspects hidden in our construction.
- Maybe use Hilbert scheme theory.

Notari and ER (2009). Replicated measurements and algebraic statistics. In Algebraic and geometric methods in statistics (Gibilisco, ER, Rogantin and Wynn eds.) (Cambridge, 2009?).

* Algebraic theory of sudoku (Fontana, Rogantin)

Sudoku is a logic-based number placement puzzle. The objective is to fill a 9×9 grid so that each column, each row, and each of the nine 3×3 boxes (also called blocks or regions) contains the digits from 1 to 9, only one time each (that is, exclusively). The puzzle setter provides a partially completed grid.

`http://en.wikipedia.org/wiki/Sudoku`

	00	01	02	10	11	12	20	21	22
00	5	3	4	6	7	8	9	1	2
01	6	7	2	1	9	5	3	4	8
02	1	9	8	3	4	2	5	6	7
10	8	5	9	7	6	1	4	2	3
11	4	2	6	8	5	3	7	9	1
12	7	1	3	9	2	4	8	5	6
20	9	6	1	5	3	7	2	8	4
21	2	8	7	4	1	9	6	3	5
22	3	4	5	2	8	6	1	7	9

General case: $p^2 \times p^2$ grid.

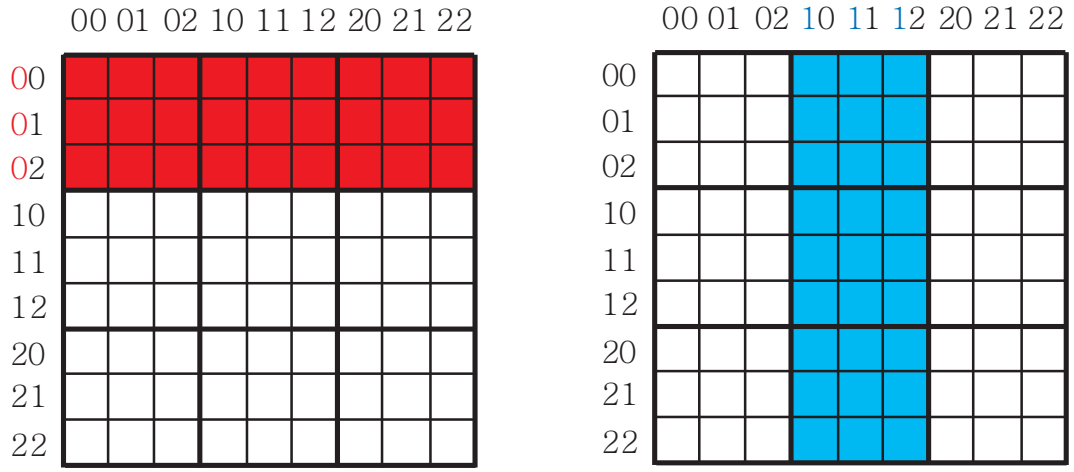
Consider a sudoku as a **fraction of a factorial design** in four factors each with p^2 levels: R, C, B, S for rows, columns, boxes and symbols, resp.

A *treatment combination* denotes a symbol and its position (row, column and box). The three "position" factors are not linearly independent.

Formalization

We split each factor R and C into 2 pseudo-factors (a typical trick in design of experiments): R_1, R_2, C_1, C_2

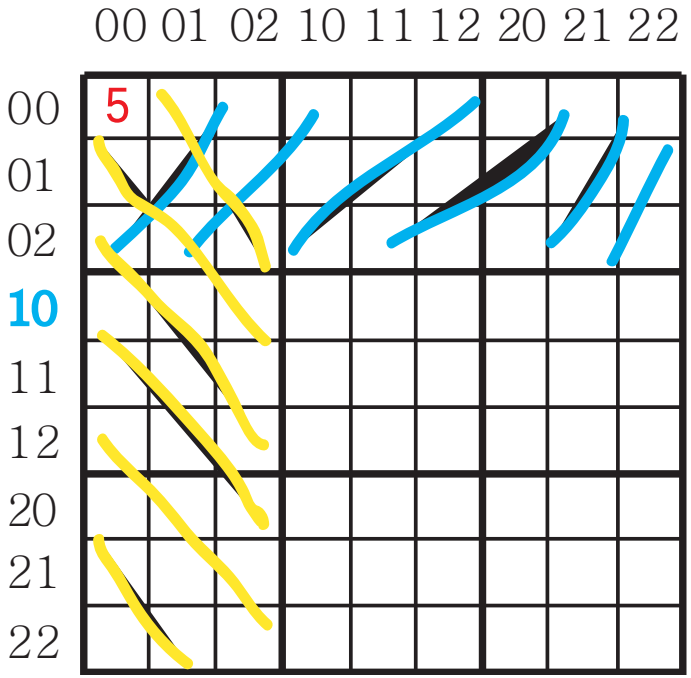
R_1 identifies the "band"
C_1 identifies the "stack"
R_2 identifies a row within a "band"
C_2 identifies a column within a "stack"



The box factor B corresponds to the two pseudo-factors R_1 and C_1.

Encode a sudoku as a fraction of a full factorial design with 5 factors, 4 factors have p levels each and one factor has p^2 levels.

Example: For $p = 3$ the number 5 in the top left corner becomes $(0, 0, 0, 0, 5)$



r_1	r_2	c_1	c_2	s
0	0	0	0	5
0	0	0	1	..
0	0	0	2	..
0	0	1	0	..
..	..	..	..	..
1	0	0	0	..
..	..	..	..	..
2	2	2	0	..
2	2	2	1	..
2	2	2	2	..

The game rules translate into values of the coefficients of the indicator function: $\sum_{\alpha \in L} b_\alpha x^\alpha$. For example

1. the fraction has p^4 points (the number of cells in the grid)

$$b_{00000} = 1/p^2$$

2. each cell appears exactly once

$R_1 \times R_2 \times C_1 \times C_2$ is a full factorial design:

$$b_{r_1 r_2 c_1 c_2 0} = 0$$

3. each symbol appears exactly once

   (a) in each row $R_1 \times R_2 \times S$ is a full factorial design:

$$b_{r_1 r_2 00 s} = 0$$

   (b) in each column $C_1 \times C_2 \times S$ is a full factorial design:

$$b_{00 c_1 c_2 s} = 0$$

   (c) in each box $R_1 \times C_1 \times S$ is a full factorial design:

$$b_{r_1 0 c_1 0 s} = 0$$

Let M be the set of indices corresponding to these constraints.

Generation of sudoku grids

Sudoku fractions are all and only the solution of the system of polynomial equations

$$\begin{cases} b_\alpha = \sum_{\beta \in L} b_\beta \; b_{[\alpha-\beta]} & \text{with } \alpha \in L \\ b_\alpha = 0 & \text{with } \alpha \in M \subset L \end{cases}$$

Full solution for $p = 2$, using CoCoA.

* LINK DESIGN CONTINGENCY TABLE TAKY+ROG

* Markov Bases for exact conditional inference in contingency tables

"We construct Markov chain algorithms for sampling from discrete exponential families conditional on a sufficient statistic... The algorithms involve computations in polynomial rings using Gröbner bases."

Let $\mathcal{D} \subset \mathbb{R}^k$ be the set of cells of a contingency tables (structural zeros).

Assume the log-linear model

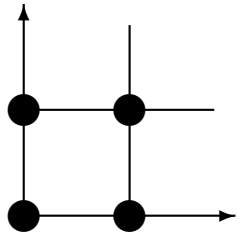
$$p(x;\theta) = Z(\theta)\exp(\sum_{i=1}^{d}\theta_i T_i(x)) \qquad \text{for } x \in \mathcal{D}$$

with parameter vector $\theta^t = [\theta_1, \ldots, \theta_d] \in \mathbb{R}^d$ and non-negative integer valued sufficient statistics $T : \mathcal{D} \longrightarrow \mathbb{Z}^d_{\geq 0} \setminus \{0\}$.

For N independent draws from $p(\cdot;\theta)$, the statistics $T = \sum_{i=1}^{N} T(x_i)$ is sufficient for θ.

Example

For $\mathcal{D} = \{(0,0),(0,1),(1,0),(1,1)\} \in \mathbb{R}^2$, a vector-space basis of the set of functions over the design is $E = \{1, xz, y, xy\}$



Note 1 The saturated exponential model on $\mathcal{D}$ is

$$\exp\{\theta_{00} + \theta_{10}x + \theta_{01}y + \theta_{11}xy\}$$

and the submodel with sufficient statistics (x, y) is

$$\begin{aligned} p(x,y) &= \exp\{\psi_{10}x + \psi_{01}y - K(\psi_{01},\psi_{10})\} \\ &= \zeta_{00}\zeta_{10}^{x}\zeta_{01}^{y} \end{aligned}$$

where $\zeta_{00} = \exp\{-K(\psi_{01},\psi_{10})\}$, $\zeta_{10} = \exp\{\psi_{10}\}$ and $\zeta_{01} = \exp\{\psi_{01}\}$

Note 2 The integer vector $[\log p(x,y)]_{(x,y)\in\mathcal{D}}$ belongs to the span of

$$Z_1 = \begin{array}{c|ccc|c} & 1 & x & y & \\ \hline (0,0) & 1 & 0 & 0 & 1 \\ (0,1) & 1 & 0 & 1 & -1 \\ (1,0) & 1 & 1 & 0 & -1 \\ (1,1) & 1 & 1 & 1 & 1 \end{array} = Z_2$$

namely

$$Z_2^t[\log p(x,y)]_{(x,y)\in\mathcal{D}} = 0$$

Now $Z_2 = \underbrace{[1,0,0,1]^t}_{Z_2^+} - \underbrace{[0,1,1,0]^t}_{Z_2^-}$ and thus

$$\begin{aligned} (Z_2^+)^t[\log p(x,y)]_{(x,y)\in\mathcal{D}} &= (Z_2^-)^t[\log p(x,y)]_{(x,y)\in\mathcal{D}} \\ p(0,0)p(1,1) &= p(1,0)p(0,1) \end{aligned}$$

Markov Bases

Let $\mathcal{D} \subset \mathbb{R}^k$ be a finite set
$T : \mathcal{D} \longrightarrow \mathbb{Z}^d_{\geq 0} \setminus \{0\}$ a non-zero, positive, integer valued function
$\mathcal{F}_t = \left\{ f : \mathcal{D} \to \mathbb{Z}_{>0} : \sum_x f(x)T(x) = t \right\} \subset \mathbb{R}[\mathcal{D}]$, a level curve
$\mathcal{Y}_t = \left\{ (x_1, \ldots, x_N) \in \mathcal{D}^N : T(x_1) + \ldots + T(x_N) = t \right\}$, set of samples with fixed values of the sufficient statistics.

Aim: enumerate $\mathcal{Y}_t$ or sample from the uniform distribution over $\mathcal{Y}_t$. This is computationally difficult for reasonable size problems. Instead sample from the hyper-geometric distribution over $\mathcal{F}_t$ by considering the map which associates to a sample its contingency table,

$$\psi : \mathcal{Y}_t \longrightarrow \mathcal{F}_t$$
$$\psi(x_1, \ldots, x_N) = \sum_{x \in \mathcal{D}} e_x \sum_{k=1}^{N} \mathbf{1}_x(x_i)$$

where $(e_x)_{x \in \mathcal{D}}$ is the canonical basis in $\mathbb{R}^{\mathcal{D}}$ and $\mathbf{1}_x$ the indicator function of $x \in \mathcal{D}$.

A Markov basis is a set of functions $f_1, \ldots, f_m : \mathcal{D} \longrightarrow \mathbb{Z}$ such that
a) $\sum_x f_i(x)T(x) = 0$ for all $i = 1, \ldots, m$ and
b) for $f, f' \in \mathcal{F}_t$ $f' = f + \sum_{i=1}^{A} e_j f_{i_j}$ with $e_j = \pm 1$ and $f + \sum_{i=1}^{a} e_j f_{i_j} \geq 0$
(there is a path from f to f' which preserves $\mathcal{F}_t$)

From this construct a stationary Markov chain on $\mathcal{F}_t$ with transition matrix

$$\begin{array}{ll} \pi(f, f + f_i) = 1/(2m) & \text{if } f + f_i \geq 0 \\ \pi(f, f - f_i) = 1/(2m) & \text{if } f - f_i \geq 0 \end{array}$$

Ex. $\begin{bmatrix} 2 & 4 \\ 3 & 1 \end{bmatrix} + \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ keeps the margin.

The algebraic formalism and toric ideals

To $x \in \mathcal{D}$ associate an indeterminate p_x then

$$(f : \mathcal{D} \to \mathbb{N}) \iff \prod_{x \in \mathcal{D}} p_x^{f(x)} =: \mathbf{p}^{f(x)} \quad \text{Ex.} \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ 2 & 4 & 3 & 1 \end{pmatrix} \qquad p_1^2 p_2^4 p_3^3 p_4$$

$$(f : \mathcal{D} \to \mathbb{Z}) \iff \mathbf{p}^{f^+(x)} - \mathbf{p}^{f^-(x)} \quad \text{Ex.} \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ 1 & -1 & -1 & 1 \end{pmatrix} \quad p_2 p_3 - p_1 p_4$$

Let $t_1, \ldots, t_d$ be other indeterminates then

$$\begin{pmatrix} T : \mathcal{D} & \longrightarrow \mathbb{N}^d \setminus \{0\} \\ x & \longmapsto (T_1(x), \ldots, T_d(x)) \end{pmatrix} \iff \begin{pmatrix} \phi_T : & \mathbb{R}[\mathcal{D}] & \longrightarrow & \mathbb{R}[t_1, \ldots, t_d] \\ & \mathbf{1}_x & \longmapsto & t_1^{T_1(x)} \ldots t_d^{T_d(x)} \end{pmatrix}$$

and ϕ_T is a ring homomorphism.

Write $\mathbf{t}^{T(x)}$ for the monomial $t_1^{T_1(x)} \ldots t_d^{T_d(x)}$

The link

Let I_T be the kernel of ϕ_T, namely $I_T = \{f \in \mathbb{R}[\mathcal{D}] : \phi_T(f) = 0\}$.

Note that

$$\sum_x f(x)T(x) = 0 \iff \left(\mathbf{p}^{f^+(x)} - \mathbf{p}^{f^-(x)} \in I_T\right)$$

and that I_T is the set of polynomials in the $(p_x, x \in \mathcal{D})$ indeterminates that vanish on the set of monomials $\{\mathbf{t}^{T(x)} : x \in \mathcal{D}\}$.

$\{f_1, \ldots, f_m\}$ is a Markov basis $\iff \left\langle \mathbf{p}^{f_i^+(x)} - \mathbf{p}^{f_i^-(x)} : i = 1, \ldots, m \right\rangle = I_T$