Workshop on Geometric and Algebraic Statistics 3

CRiSM Warwick April 5-7, 2010

Algebraic Statistics . an overview

Eva Riccomagno
riccomagno@dima.unige.it

and C. Fassino, H. Maruri-Aguilar, V. Pirino, G. Pistone, F. Rapallo, M.P. Rogantin, H. Wynn...

Some citations

"Algebraic statistics is concerned with the development of techniques in algebraic geometry, commutative algebra, and combinatorics, to address problems in statistics and its applications. On the one hand, algebra provides a powerful tool set for addressing statistical problems. On the other hand, it is rarely the case that algebraic techniques are ready-made to address statistical challenges [...] This way the dialogue between algebra and statistics benefits both disciplines." **Lectures on Algebraic Statistics by Drton, Sturmfels, Sullivant, Birkhäuser 2009**

"Algebraic statistics is the use of algebra to advance statistics. Algebra has been useful for experimental design, parameter estimation, and hypothesis testing." **Wikipedia**

"It might seem natural that where a statistical model can be defined in algebraic terms it would be useful to use the full power of modern algebra to help with the description of the model and the associated statistical analysis." **Algebraic and geometric methods in statistics, Gibilisco, Riccomagno, Rogantin Wynn (eds), Cambridge 2010**

"[...] build a bridge between the approximate data of the real world and the exact structures of commutative algebra" **Approximate commutative algebra, Robbiano and Abbott (eds), Springer 2009**

[...] experimental design in the language of contingency table can be taken as the study of tables with prohibited cells and Markov bases for models over such designs can be better developed;

the links between optimal experimental design and the algebraic method in experimental design has not yet been established, although optimal designs often exhibit symmetries and formal invariant theory might be use;

computational algebra in main-stream probability theory is ripe for more development and we should be particularly interested when there are applications in statistics.

The foundations in areas like semi-group theory in Markov chains and algebraic combinatorics for counting special congurations using generating function techniques, together with asymptotics, may prove to be fruitful leads to statistical applications.
Algebraic methods in statistics and probability II, Viana and Wynn (eds), AMS 2009

Design

$\mathcal{D}$ finite set of points in $\mathbb{R}^k$ — Ideal($\mathcal{D}$)

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

Saturated hierarchical models — L order ideals

Products in $\mathcal{L}$ — normal form

Example: Plackett-Burman (PB8) design with eight runs, seven factors and generator + - - + - + +

```
Use R::=Q[x[1..7]], Lex;
PB8:= [   [ 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]   ] ;

I:=IdealOfPoints(PB8);  I;
Ideal( x[7]^2 - 1, x[6]^2 - 1, x[5]^2 - 1,
              x[4] + x[5]x[7], x[3] - x[5]x[6]x[7],
              x[2] + x[6]x[7], x[1] + x[5]x[6] )
-------------------------------


QuotientBasis(I);
[1, x[7], x[6], x[5], x[6]x[7], x[5]x[7], x[5]x[6], x[5]x[6]x[7]]
-------------------------------


NF( x[1]x[2]x[3]x[4]x[5]x[6]x[7] , I );
-1
-------------------------------
```

```
Use R::=Q[x[1..7]];
D:=[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1]><[-1,1];
PB8:= [ ... ];
IdealOfPoints(PB8);
[ 1, x[7], x[6], x[5], x[4], x[3], x[2], x[1]]
-------------------------------


InFun1:=Fu(PB8,D);InFun1;
- 1/16x[1]x[2]x[3]x[4]x[5]x[6]x[7]
+ 1/16x[1]x[3]x[4]x[5] + 1/16x[1]x[2]x[3]x[6]
+ 1/16x[2]x[4]x[5]x[6] + 1/16x[2]x[3]x[4]x[7]
+ 1/16x[1]x[2]x[5]x[7] + 1/16x[1]x[4]x[6]x[7]
+ 1/16x[3]x[5]x[6]x[7]
- 1/16x[1]x[2]x[4] - 1/16x[2]x[3]x[5] - 1/16x[3]x[4]x[6]
- 1/16x[1]x[5]x[6] - 1/16x[1]x[3]x[7] - 1/16x[4]x[5]x[7]
- 1/16x[2]x[6]x[7]
+ 8/128
-------------------------------
```

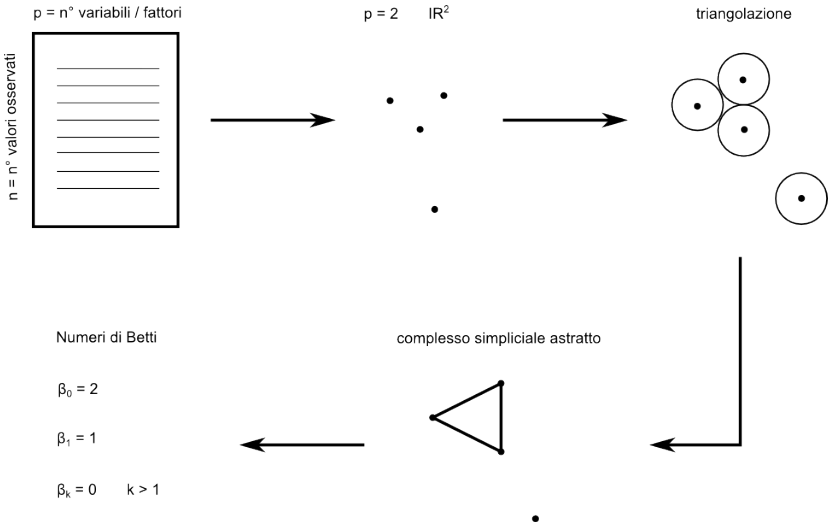
If $\mathcal{D} \subset 2^d$, then L is of square free monomials. It corresponds to an abstract simplicial complex. Its Betti numbers give information on the "connectiveness" property of the identifiable model.

```
Use R::=Q[x[1..7]],             DegLex;                    Lex;
PB8:= [ ... ]; I:=IdealOfPoints(PB8); QuotientBasis(I);

HilbertSeries(R/I);

[1, x[7], x[6], x[5], x[4], x[3], x[2], x[1]]
---------------------------------
[1, x[7], x[6], x[5], x[6]x[7], x[5]x[7], x[5]x[6], x[5]x[6]x[7]]
---------------------------------
(1 + 7x[1])                        (1 + 3x[1] + 3x[1]^2 + x[1]^3)
---------------------------------
```

In DegLex $\beta_0 = 1$ and $\beta_1 = 7$ in Lex $\beta = (1, 3, 3, 1)$

Betti and design

Carlsson '09, Pirino '11



Note

```
Use R::=Q[x[1..7]], Lex; PB8:= [ ... ] ;
I:=Ideal(x[1]*x[2]*x[3]*x[4]*x[5]*x[6]*x[7]);
Foreach T In PB8 Do
       V:=[ K+T | K In PB8 ];
       W:=[ [Abs(K)/2 | K In A] | A In V ];
       I:=I+Cast( [LogToTerm(A) | A In W], IDEAL);
EndForeach;
I:=I+Ideal([ X^2 | X In Indets()]);

GBasis(I);     HilbertSeries(R/I);

[ x[1]x[3]x[7], x[1]x[5]x[6], x[4]x[5]x[7], x[1]x[2]x[4],
  x[3]x[4]x[6], x[2]x[6]x[7], x[2]x[3]x[5],
  x[1]^2, x[2]^2, x[3]^2, x[4]^2, x[5]^2, x[6]^2, x[7]^2 ]
-------------------------------

(1 + 7x[1] + 21x[1]^2 + 28x[1]^3 + 7x[1]^4)
-------------------------------
```

For $\mathcal{D}$, a term-ordering, G a τ-G-basis of $I(\mathcal{D})$, and a polynomial p

$$\begin{aligned} p(x) &= \sum_{g \in G} s_g(x) g(x) + r(x) \\ &= \sum_{g \in G} s_g(x) g(x) + \sum_{d \in \mathcal{D}} p(d) l_d(x) \end{aligned}$$

where l_d is the Lagrange polynomial for $d \in \mathcal{D}$.
Let μ be a measure which admits all moments and $X \sim \mu$. Then

$$\mathrm{E}_\mu\left(p(X)\right) = \mathrm{E}_\mu\left(r(X)\right) = \sum_{d \in \mathcal{D}} p(d)\, \mathrm{E}_\mu\left(l_d(X)\right)$$

for all p s.t. $\mathrm{E}_\mu\left(p(X) - r(X)\right) = 0$

Hermite case

Let μ be the standard Gaussian distribution over the real and let H_i $(i = 0, 1, \dots)$ be the Hermite polynomials. Let $\mathcal{D} = \{x : H_n(x) = 0\}$. Then

- $p(x) = q(x)H_n(x) + r(x)$ with $\deg_x r < n$
- $\mathrm{E}_\mu(p(X)) = \sum_{d\in\mathcal{D}} p(d)\,\mathrm{E}_\mu(l_d(X))$ if and only if $c_n(q) = 0$

  where $q(x) = \sum_{i=0}^{+\infty} c_i(q)H_i(x)$.

Let $\mathcal{D} = \{(x_1, \dots, x_k) : H_{n_i}(x_i) = 0 \;\; i = 1, \dots, k\}$. Then

- $p(x) = \sum_{i=1}^{k} q_i(x)H_{n_i}(x_i) + r(x)$ with $\deg_{x_i} r < n_i$
- $\mathrm{E}_\mu(p(X)) = \mathrm{E}_\mu(r(X))$ if and only if $c_{n_i}(q_i(x)) = 0$ for $i = 1, \dots, k$.

Note that the c_n are linear combinations of the coefficients of p.

Generalisations

To a fractions of the zeros of the Hermite polynomials.

To $\mathcal{D}$ any set of points, in particular sparse grids.

For Hermite we are finalising macros in cocoa whose indeterminates are the Hermite polynomials. Generalise them to other classes of polynomials.

The contingency table of a design Aoki, Takemura '06, Rapallo, Rogantin '10

Let $\mathcal{D}$ be a fraction of a full factorial design possibly with replicated values

	$B1$	$B2$	$B3$
$A1$	4	1	0
$A2$	$N.A.$	2	2

it can be read as a contingency table whose entries are the number of replicates: $n_{ij} = f_{ij}$.

The counting polynomial, a straightforward generalisation of the indicator function, can be computed to give information on the design structure.

More interestingly, this opens the way to the applicability of Markov bases in the analysis and design of experiments.

Let $\mathcal{D} \subset \mathbb{R}^k$ be the set of cells of a contingency tables, $T : \mathcal{D} \longrightarrow \mathbb{N}^d \setminus \{0\}$ a function, $\mathcal{F}_t = \{f : \mathcal{D} \to \mathbb{N} : \sum_x f(x)T(x) = t\}$, the level curve of T at t.

A Markov basis is a set of functions $f_1, \ldots, f_m : \mathcal{D} \longrightarrow \mathbb{Z}$ such that

1. $\sum_x f_i(x)T(x) = 0$ for all $i = 1, \ldots, m$ and
2. for $f, f' \in \mathcal{F}_t$ $f' = f + \sum_{j=1}^{A} e_j f_{i_j}$ with $e_j = \pm 1$ and $f + \sum_{i=1}^{a} e_j f_{i_j} \geq 0$, $0 \leq a \leq A \leq m$ (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

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

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

Polynomials and integer valued functions

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

- To the non-negative integer valued function $f : \mathcal{D} \to \mathbb{N}$ associate $\mathbf{p}^{f(x)} := \prod_{x \in \mathcal{D}} p_x^{f(x)}$ $(2,4,3,1) \leftrightarrow p_{x_1}^2 p_{x_2}^4 p_{x_3}^3 p_{x_4}^1$
- To the integer valued function $f : \mathcal{D} \to \mathbb{Z}$ associate $\mathbf{p}^{f^+(x)} - \mathbf{p}^{f^-(x)}$ $(1,-1,-1,1) \leftrightarrow p_{x_2}p_{x_3} - p_{x_1}p_{x_4}$
- To the multivalued integer function $T : \mathcal{D} \longrightarrow \mathbb{N}^d \setminus \{0\}$ associate the ring homomorphism

$$\begin{array}{rccc} \phi_T : & \mathbb{R}[\mathcal{D}] & \longrightarrow & \mathbb{R}[t_1, \ldots, t_d] \\ & \mathbf{1}_x & \longmapsto & t_1^{T_1(x)} \ldots t_d^{T_d(x)} \end{array}$$

$$\left(\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ \hline 1 & 2 & 3 & 4 \\ 1 & 3 & 2 & 4 \end{array} \right) \leftrightarrow \left(t_1 t_2,\, t_1^2 t_2^3,\, t_1^3 t_2^2,\, t_1^4 t_2^4 \right)$$

Markov bases and toric models

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

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

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 $\Longleftrightarrow \left\langle \mathbf{p}^{f_i^+(x)} - \mathbf{p}^{f_i^-(x)} : i = 1, \ldots, m \right\rangle = I_T$

From this algebraic MCMC and exact test for contingency tables, model selection, p-value computation for sparse data, ...

Note that I_T is a toric ideal, i.e. generated by binomials.

Algebraic statistical models

If a family of probability distributions on a measurable space can be described through equalities (and inequalities) of (ratios of) polynomials, then it is a (semi)-algebraic statistical model.

Example (two-way tables)

Let $\Delta = \{P \in \mathbb{R}^{I\times J} : \sum_{i,j} P_{ij} = 1, P_{ij} \geq 0\}$ and $f_1, \ldots, f_n$ be polynomials in the P_{ij}. Then if $\{P \in \mathbb{R}^{I\times J} : f_1((P_{ij})_{i,j}) = \ldots = f_n(P) = 0\} \cap \Delta \neq \emptyset$ it is an algebraic statistical model.

The independence model is toric and its defining polynomials are

$$P_{i,j}P_{k,h} - P_{i,h}P_{k,j} \text{ for } 1 \leq i < k \leq I; 1 \leq j < h \leq J$$

Also the independence model is

$$\{P : P = cr^t\} \cap \Delta$$

with c, r probability distributions.

Mixture of independence models   Rapallo, Carlini '10

The mixture of k-independence model is

$$\{P : P = \alpha_1 c_1 r_1^t + \ldots + \alpha_k c_k r_k^t\} \cap \Delta$$

with $c_i, r_i, (\alpha_1, \ldots, \alpha_k)$ probability distributions, namely $c_{ij}, r_{ij} \geq 0$ and $\sum_j c_{ij} = 1 = \sum_j r_{ij}$.

That is the model does not contain all matrices of rank $\leq k$.

The non-negative rank of an $I \times J$ matrix P, denoted with $rank_+(P)$ is the smallest integer k such that there exist non-negative vectors $c_1, \ldots, c_k$ and $r_1, \ldots, r_k$ and the decomposition $P = c_1 r_1^t + \ldots + c_k r_k^t$ holds.

There is no algorithm for the computation of the non-negative rank.

On the importance of inequalities see also Settimi, Smith '00, Zwiernik '10.

Algebraic independence models for multi-way tables

Let X, Y, Z be binary random variables with $X \perp Y|Z$ namely

$$\text{on } z=0 \;\; P(X=i, Y=j|Z=0) = P(X=i|Z=0)P(Y=j|Z=0)$$
$$\text{on } z=1 \;\; P(X=i, Y=j|Z=1) = P(X=i|Z=1)P(Y=j|Z=1)$$

Applying the previous result to both conditions and intersecting

$Z=0$	$X=0$	$X=1$
$Y=0$	p_{000}	p_{010}
$Y=1$	p_{100}	p_{110}
	1	2

$Z=1$	$X=0$	$X=1$
$Y=0$	p_{001}	p_{011}
$Y=1$	p_{101}	p_{111}
	3	4

$$I_{X\perp Y|Z} = \langle p_{000}p_{110} - p_{010}p_{100}, \;\; p_{001}p_{111} - p_{011}p_{101} \rangle$$

$$M_{X\perp Y|Z} = \{P \in \mathbb{R}^{2^3} : p_{000}p_{110} - p_{010}p_{100} = 0 = p_{001}p_{111} - p_{011}p_{101}\} \cap \Delta$$

Do the distributions in $M_{X \perp Y|Z}$ satisfy the condition

$$\frac{P(Y = Z = 0|X = 0)}{P(Y = Z = 1|X = 0)} = \frac{P(Y = Z = 0|X = 1)}{P(Y = Z = 1|X = 1)} \qquad ?$$

(Almost) equivalently does the minor "14" belong to the model?

$with(PolynomialIdeals);$

$$[\, <,>, Add, Contract, EliminationIdeal, EquidimensionalDecomposition, Generators, HilbertDimension, IdealContainment, IdealInfo, IdealMembership, Intersect, IsMaximal, IsPrimary, IsPrime, IsProper, IsRadical, IsZeroDimensional, MaximalIndependentSet, Multiply, NumberOfSolutions, Operators, PolynomialIdeal, PrimaryDecomposition, PrimeDecomposition, Quotient, Radical, RadicalMembership, Saturate, Simplify, UnivariatePolynomial, VanishingIdeal, ZeroDimensionalDecomposition, in, subset\,] \tag{1}$$

$T1 := p000 \cdot p110 - p010 \cdot p100, p001 \cdot p111 - p011 \cdot p101;$

$$p000\, p110 - p010\, p100,\ p001\, p111 - p011\, p101 \tag{2}$$

$M := \langle T1 \rangle;$

$$\langle p000\, p110 - p010\, p100,\ p001\, p111 - p011\, p101 \rangle \tag{3}$$

$T2 := p000 \cdot p111 - p100 \cdot p011;$

$$p000\, p111 - p100\, p011 \tag{4}$$

$IdealMembership(T2, M);$

$$false \tag{5}$$

$IdealMembership(p000 \cdot p110 - p010 \cdot p100, M);$

$$true \tag{6}$$

How many distributions satisfy the model?
We started with 8 parameters. How many free parameters are there?
$X \perp Y | Z$ and $X \perp Z | Y$

$with(PolynomialIdeals)$:
$T1 := p000 \cdot p110 - p010 \cdot p100, p001p111 - p011 \cdot p101$:
$T2 := p000 \cdot p101 - p100 \cdot p001, p010 \cdot p111 - p110 \cdot p011$:
$M := \langle T1, T2 \rangle$:
$IsProper(M)$; # *M is not the saturated model nor the empty model*

$$true$$

$IsZeroDimensional(M)$; # *M is not a finite set*

$$false$$

$NumberOfSolutions(M)$;

$$\infty$$

$MaximalIndependentSet(M)$;

$$\{p001p111, p110, p010, p100, p011\}$$

$HilbertDimension(M)$;

$$5$$

This suggests a non-standard set of 5 parameters.
Now M should still be intersected with the simplex.

On the many parametrizations Pistone, Wynn, Kobayashi, Smith, Zwiernik ...

On a finite set $\mathcal{D} \subset \mathbb{N}^k \setminus \{0\}$ consider an exponential model

$$p(x;\psi) = \exp(\psi_{0000} + \psi_{0100}x_2 + \psi_{0001}x_4)$$
$$\exp(\psi_{1000}x_1 + \psi_{1100}x_1x_2 + \psi_{1001}x_1x_4)\exp(\psi_{0010}x_3 + \psi_{0110}x_2x_3 + \psi_{0011}x_3x_4)$$

- raw probabilities $(p(d), d \in \mathcal{D}) \in \Delta$
- vector space representation $\begin{cases} p_\theta = \theta_{0000} + \sum_{\alpha \in L_0} \theta_\alpha x^\alpha \\ \theta_{0000} = 1 - \sum_{\alpha \in L_0} \theta_\alpha m_\alpha \end{cases}$

  where $m_\alpha = \mathrm{E}_0(X^\alpha)$
- (toric) $\begin{cases} p(x;\psi) & = \exp\left(\sum_{\alpha \in M} \psi_\alpha x^\alpha\right) = \prod_{\alpha \in M} \exp\left(\psi_\alpha x^\alpha\right) \\ & = \zeta_0 \prod_{\alpha \in M_0} \zeta_\alpha^{x^\alpha} = p(x;\zeta) \end{cases}$

  where $\zeta_\alpha = \exp(\psi_\alpha)$

Use elimination theory to change parametrization

$$\zeta \leftrightarrow p \leftrightarrow \theta$$

For example, elimination of ζ from the p-ζ equations gets an implicit representation of the model, which for graphical model consists of a set of binomials.

From H.P. Wynn's talk at Wogas 2

x: control (or input) variables
θ: a basic parameter vector
η: a parameter vector that may be considered as depending on x (e.g. a mean).

An algebraic statistical model is a statement that (x, θ, η) lie on an affine algebraic variety:

$$h(x, \theta, \eta) = 0,$$

together with a statement that the joint distribution of outputs $Y_1, \ldots Y_n$ depends on

$$\theta, (x_i, \eta_i), \qquad i = 1, \ldots, n$$

- Regression: if η is a mean $\eta = f(x, \theta)$ and f is a polynomial, then $\eta - f(x, \theta) = 0$. Eliminate θ to get an implicit description of the relation between x and η.

- Variance components: if $\Sigma_{ij} = \mathrm{Cov}(Y_i, Y_j)$ then $(\Sigma^{-1})_{ij} = 0$ implies algebraic conditions on the entries of Σ.

  Gaussian independence models (Drton et al '08, Massa in progress).

  For $X \sim \mathcal{N}_k(0, \Sigma)$ the condition $X_3 \perp X_2 | X_1$ corresponds to $\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13} = 0$ together with Σ being semi-definite positive which is a semi-algebraic condition.

  As in the discrete case some operations with models can be performed by manipulating the model ideals.

Bibliography

- D. A. Cox, J. B. Little, D. O'Shea, *Ideals, varieties, and algorithms, An introduction to computational algebraic geometry and commutative algebra*. Third edition. Undergraduate Texts in Mathematics. Springer, New York, 2007.
- G. Pistone, E. Riccomagno, H. P. Wynn. *Algebraic Statistics. Computational commutative algebra in statistics*. Monographs on Statistics and Applied Probability, 89. Chapman & Hall/CRC, Boca Raton, FL, 2001. CRC Press, 2001.
- Algebraic and geometric methods in statistics (Gibilisco, Riccomagno, Rogantin and Wynn, eds) Cambridge University Press, Cambridge, 2010.
- Algebraic methods in statistics and probability II, Viana and Wynn (eds), AMS, 2009.
- Approximate commutative algebra, Robbiano and Abbott (eds), Springer 2009.
- Lectures on Algebraic Statistics by Drton, Sturmfels, Sullivant, Birkhäuser 2009.
- G. Carlsson, *Topology and Data* Bulletin of the American Mathematical Society, Vol. 46, pp.255-308, 2009.
- G. Pistone and H.P. Wynn. *Generalised confounding with Grbner bases*. Biometrika 83 (1996), no. 3, 653–666.
- K. Ye, *Indicator function and its application in two-level factorial designs*, The Annals of Statistics 31,3:984994, 2003.

- Y. Berstein, H. Maruri-Aguilar et al., *Minimal aberration and the state polytope for experimental design* AISM, 2010.
- I. Dinwoodie (1998). The Diaconis-Sturmfels algorithm and rules of succession. Bernuolli 4,3:401-410.
- F. Rapallo (2003). Algebraic Markov Bases and MCMC for two-way contingency tables. Scand. J. of Stats 30:385-397.
- A. Krampe, S. Kuhnt (2010). Model Selection for Contingency Tables with Algebraic Statistics. In *Algebraic and Geometric Methods in Statistics*, Chapter 4.
- R. Settimi, J. Q. Smith (2000). Geometry, moments and conditional independence trees with hidden variables. The Annals of Statistics, 28(4), 11791205.
- P. Zwiernik, Ph.D. Thesis, 2010.