2ND CARLO ALBERTO STOCHASTICS WORKSHOP ALGEBRAIC STATISTICS IN ESTIMABILITY Mar 23, 2012

Collegio Carlo Alberto

DE CASTRO STATISTICS

12:00

Giovanni Pistone (Collegio Carlo Alberto)

Title: Introduction

12:15

Fabio Rapallo (Università del Piemonte Orientale)

Title: From Markov moves in contingency tables to linear model estimability I

12:45 - 14.00 Break

14:00

Roberto Fontana (Politecnico di Torino) and **Maria Piera Rogantin** (Università di Genova)

Title: From Markov moves in contingency tables to linear model estimability II

14:30

Enrico Carlini (Politecnico di Torino)

Title: Perturbation of matrices and non-negative rank

15:00 Discussion

2ND CARLO ALBERTO STOCHASTICS WORKSHOP

DE CASTRO STATISTICS

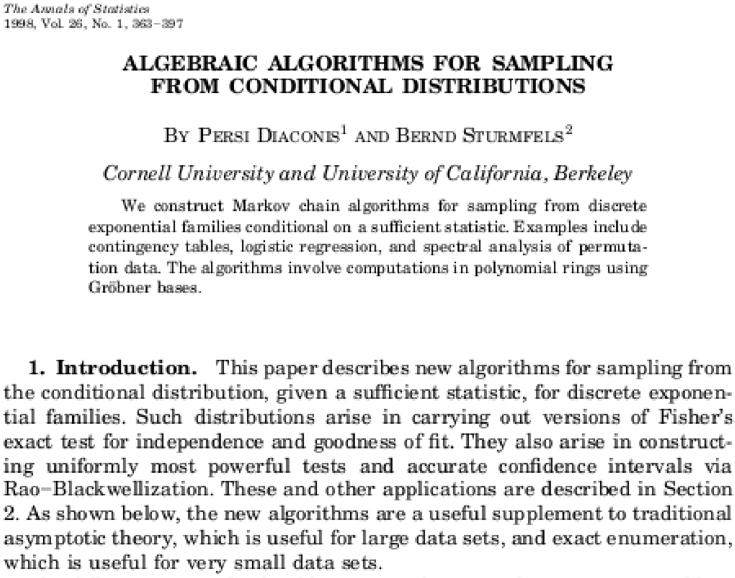
Collegio Carlo Alberto

Introduction:
An 10min course in Algebraic Statistics

Giovanni Pistone

Moncalieri, March 23 2012

1st paper in AS (1993)



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

ALGEBRAIC ALGORITHMS FOR SAMPLING FROM CONDITIONAL DISTRIBUTIONS

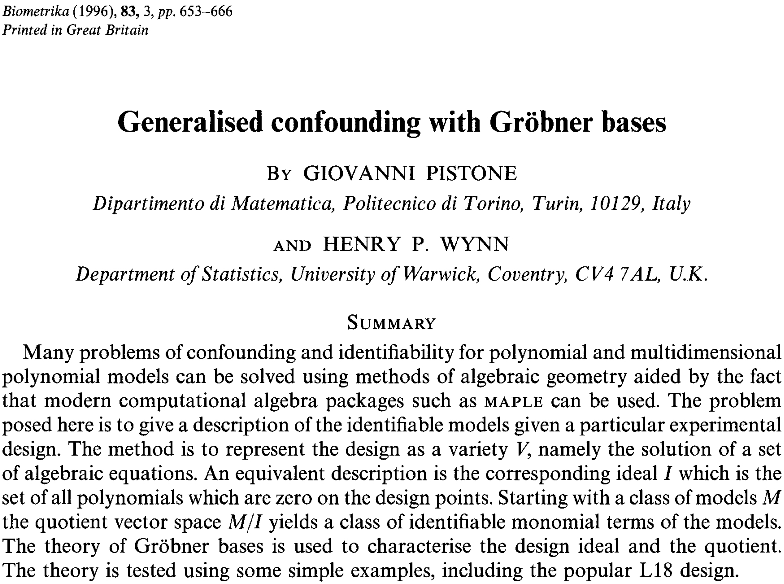
BY PERSI DIACONIS[1] AND BERND STURMFELS[2]

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.

2nd paper in AS



Biometrika (1996), **83,** 3, *pp.* 653–666
Printed in Great Britain

Generalised confounding with Gröbner bases

By GIOVANNI PISTONE
Dipartimento di Matematica, Politecnico di Torino, Turin, 10129, Italy

AND HENRY P. WYNN
Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.

Summary

Many problems of confounding and identifiability for polynomial and multidimensional polynomial models can be solved using methods of algebraic geometry aided by the fact that modern computational algebra packages such as MAPLE can be used. The problem posed here is to give a description of the identifiable models given a particular experimental design. The method is to represent the design as a variety V, namely the solution of a set of algebraic equations. An equivalent description is the corresponding ideal I which is the set of all polynomials which are zero on the design points. Starting with a class of models M the quotient vector space M/I yields a class of identifiable monomial terms of the models. The theory of Gröbner bases is used to characterise the design ideal and the quotient. The theory is tested using some simple examples, including the popular L18 design.

Which algebra?

- B. Sturmfels, *Gröbner bases and convex polytopes* (American Mathematical Society, Providence, RI, 1996), ISBN 0-8218-0487-1

- D. Cox, J. Little, D. O'Shea, *Ideals, varieties, and algorithms*, Undergraduate Texts in Mathematics (Springer-Verlag, New York, 1992), ISBN 0-387-97847-X, an introduction to computational algebraic geometry and commutative algebra

- M. Kreuzer, L. Robbiano, *Computational commutative algebra. 1* (Springer-Verlag, Berlin, 2000), ISBN 3-540-67733-X

- CoCoATeam, CoCoA*: a system for doing Computations in Commutative Algebra*, Available at `http://cocoa.dima.unige.it` (online)

- 4ti2 team, *4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces*, Available at www.4ti2.de (online)

Algebraic Statistics

- 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
- L. Pachter, B. Sturmfels, eds., *Algebraic Statistics for Computational Biology* (Cambridge University Press, 2005)
- M. Drton, B. Sturmfels, S. Sullivant, *Lectures on algebraic statistics*, Vol. 39 of *Oberwolfach Seminars* (Birkhäuser Verlag, Basel, 2009), ISBN 978-3-7643-8904-8, `http://dx.doi.org/10.1007/978-3-7643-8905-5`
- P. Gibilisco, E. Riccomagno, M.P. Rogantin, H.P. Wynn, eds., *Algebraic and geometric methods in statistics* (Cambridge University Press, Cambridge, 2010), ISBN 978-0-521-89619-1

Lingo: Ideals

Definitions

- $R = \mathbb{R}[x_1, \ldots, x_d]$ is the *ring* of polynomials with real coefficients and d indeterminates.
- The *ideal* of the set $\mathcal{D} \subset \mathbb{R}^d$ is

$$\text{Ideal}\,(\mathcal{D}) = \{f \in R\colon\, f(x) = 0 \text{ if } x \in \mathcal{D}\}\,.$$

- A *generating set* or *basis* of $\text{Ideal}\,(\mathcal{D})$ is a set of polynomials $\mathcal{B} \subset \text{Ideal}\,(\mathcal{D})$ such that every $g \in \text{Ideal}\,(\mathcal{D})$ is of the form

$$f(x) = \sum_{j=1}^{n} f_j(x) g_j(x), \quad f_j(x) \in R, g_j \in \mathcal{B}.$$

Theorem (Hilbert 1890)

Every ideal is finitely generated.

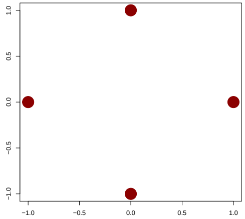
Examples

Positive x-axis in the plane

- $\mathcal{D} = \left\{(x, y) \in \mathbb{R}^2 \colon x \geq 0, y = 0\right\}$
- If $f \in \mathbb{R}[x, y]$ is zero on $\mathcal{D}$, then f has no term with x only, i.e. $f(x, y) = f_1(x, y)y$.
- $\mathcal{B} = \{y\}$ is a generating set of Ideal$(\mathcal{D})$.

1FAT

- $\mathcal{D} = \{(1, 0), (0, 1), (-1, 0), (0, -1)\} \subset \mathbb{R}^2$
- Both $x^2 + y^2 - 1$ and xy belong to Ideal$(\mathcal{D})$.



Lingo: Gröbner basis

Definition

- A *term order* is a total order $\prec$ on monomials $x^\alpha = x_1^{\alpha_1 \cdots x_d^{\alpha_d}}$, $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{Z}^d_{\geq}$, s.t. $1 \prec x^\alpha$ and

$$x^\alpha \prec x^\beta \Longleftrightarrow x^{\alpha+\gamma} \prec x^{\beta+\gamma}.$$

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

Theorem (Buchberger 1965)

- *There is a finite test for Gröbner basis.*
- *There is a finite algorithm that produces a G-basis $\mathcal{G}$ from any basis.*

Classical DoE: 2^{3-1}

The design

$$\mathcal{D} = \begin{array}{ccc} x & y & z \\ \hline +1 & +1 & +1 \\ -1 & -1 & +1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \end{array}$$

has design ideal $\operatorname{Ideal}(\mathcal{D})$ 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, but

$$\operatorname{LT}(xy - z) = xy$$

cannot be obtained from the 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 monomials $1, x, y, z$ are not *aliased*.

CoCoA solves 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]
-------------------------------
```

Lingo: *A*-model

- $\mathcal{X}$ is a finite sample space with reference measure μ.
- A is an integer matrix $A \in \mathbb{Z}_{\geq}^{m+1,\mathcal{X}}$.
- The elements of matrix A are $A_i(x)$, $i = 0 \dots m, x \in \mathcal{X}$, $A_0(x) = 1$.
- The x-column of A, say $A(x)$, is the multi-exponent of a *monomial term* denoted
$$t^{A(x)} = t_0 t_1^{A_1(x)} \cdots t_m^{A_m(x)}$$
- Matrix A defines a statistical model on $(\mathcal{X}, \mu)$ whose *unnormalized probability densities* are
$$q(x; t) = t^{A(x)}, \quad x \in \mathcal{X},$$
for all $t \in \mathbb{R}_{\geq}^{m+1}$ such that $q(\cdot, t)$ is not identically zero.
- The probability densities wrt μ in the A-model are
$$p(x; t) = q(x; t)/Z(t), \quad Z(t) = \sum_{x \in \mathcal{X}} q(x; t)\mu(x).$$
- *If* $t > 0$, $t = \log \theta$ *and* $q(x; \theta) = \exp(\theta \cdot A(x))$.

Example of A-model

Binomial(n, p)

- $\mathcal{X} = \{0, 1, 2, 3, \ldots, n\}$, $\mu(x) = \binom{n}{x}$.
- $A = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cccccc} 0 & 1 & 2 & 3 & \cdots & n \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & 3 & \cdots & n \end{bmatrix} \end{array}$.
- $q(x; t_0, t_1) = t_0 t_1^x$.
- $p(x; t) = \frac{t^x}{\sum_{x=0}^n t^x \binom{n}{x}} = \frac{t^x}{(1+t)^n}$, $x \in \mathcal{X}$, $t \geq 0$.
- $t = p/(1-p)$ are the *odds*:

$$\binom{n}{x} \frac{t^x}{(1+t)^n} = \binom{n}{x} p^x (1-p)^{n-x}.$$

Lingo: Toric ideal

Definition

The ker of the ring homomorphism

$$\mathbb{R}[q(x)\colon x \in \mathcal{X}] \ni q(x) \mapsto t^{A(x)} \in \mathbb{R}[t_0, \dots, t_m]$$

is the *toric ideal* of A.

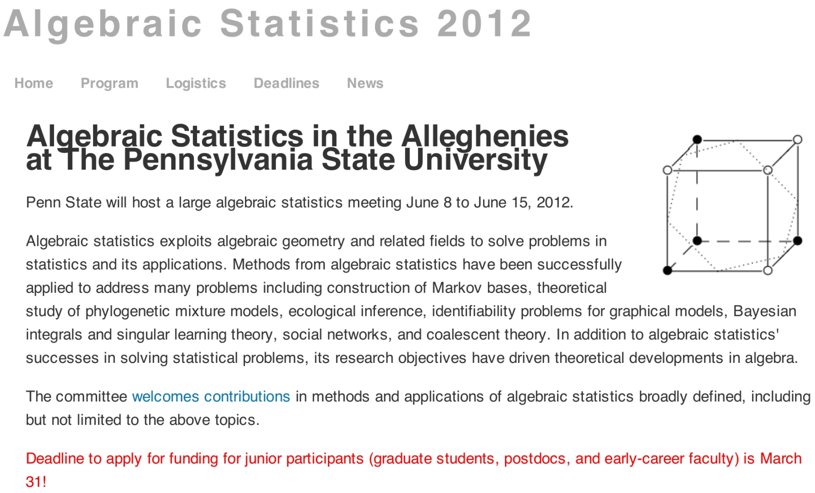
Theorem (Sturmfels 1996)

- Ideal (A) *has a finite basis made of binomials of the form*

  $$\prod_{x\colon u(x)>0} q(x)^{u_+(x)} - \prod_{x\colon u(x)<0} q(x)^{u_-(x)}$$

  with $u \in \mathbb{Z}^{\mathcal{X}}$, $Au = 0$.
- *There exists a Markov basis of* A.

State of the art



Algebraic Statistics 2012

Algebraic Statistics in the Alleghenies at The Pennsylvania State University

Penn State will host a large algebraic statistics meeting June 8 to June 15, 2012.

Algebraic statistics exploits algebraic geometry and related fields to solve problems in statistics and its applications. Methods from algebraic statistics have been successfully applied to address many problems including construction of Markov bases, theoretical study of phylogenetic mixture models, ecological inference, identifiability problems for graphical models, Bayesian integrals and singular learning theory, social networks, and coalescent theory. In addition to algebraic statistics' successes in solving statistical problems, its research objectives have driven theoretical developments in algebra.

The committee welcomes contributions in methods and applications of algebraic statistics broadly defined, including but not limited to the above topics.

Deadline to apply for funding for junior participants (graduate students, postdocs, and early-career faculty) is March 31!