Finiteness results in statistics using algebra

Jan Draisma
TU Eindhoven

Leiden, 4 November 2010

Sampling contingency tables

Month of birth	Month of death: Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
March	1	0	0	0	2	1	0	0	0	0	0	1	5
April	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
June	2	0	0	0	1	0	0	0	0	0	0	0	3
July	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sept	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

Diaconis-Sturmfels, 1995

Independent? Compare χ^2-statistic.

Sampling tables with prescribed marginals

Repeatedly add or subtract $\begin{matrix} +1 & -1 \\ -1 & +1 \end{matrix}$ (*Metropolis-Hastings*).

Theorem (classical)
Matrices in $\mathbb{N}^{r\times c}$ with prescribed marginals are connected by such moves (*Markov basis*).

Markov bases

Definition
$A : \mathbb{N}^n \longrightarrow \mathbb{N}^m$ additive
$M \subseteq \ker A$ *Markov basis for* A if
$Au = Av \Rightarrow \exists u_0, u_1, \ldots, u_k \in \mathbb{N}^m$
with $u_0 = u$, $u_{i+1} - u_i \in \pm M$, $u_k = v$.

Example
$n = r \times c$, $m = r + c$
A : matrix $\mapsto$ its row and column sums
$M = \{E_{ij} + E_{kl} - E_{il} - E_{kj} \mid i, j, k, l\}$

Theorem (Diaconis-Sturmfels, 1995)
Every A has a *finite* Markov basis.

$\rightsquigarrow$ Sampling algorithm for many conditional distributions!

Markov bases, continued

Proof sketch

$A : \mathbb{N}^n \to \mathbb{N}^m$

$x = (x_1, \ldots, x_m)$ source variables

$y = (y_1, \ldots, y_n)$ target variables

$S := \{y_j - x^{Ae_j} \mid j\}$ binomial equations

Hilbert's Basis Theorem:

$\mathrm{Ideal}(S) \cap \mathbb{C}[y_1, \ldots, y_n]$ finitely generated

in fact, by binomial generators $y^u - y^v$

take corresponding $u - v$. $\square$

Example

$n = 2 \times 3, m = 2 + 3$

x_1, x_2, z_1, z_2, z_3 source variables

$y_{11}, \ldots, y_{23}$ target variables

$S = \{y_{ij} - x_i z_j \mid i, j\}$

binomial generators $\{y_{ij} y_{kl} - y_{il} y_{kj} \mid i, j, k, l\}$

Markov basis $\{E_{ij} + E_{kl} - E_{il} - E_{kl} \mid i, j, k, l\}$

Finite up to symmetry?

Observation (Hoşten-Sullivant,...)
Markov basis $\{E_{ij} + E_{kl} - E_{il} - E_{kl} \mid i, j, k, l\}$
finite up to permutations for $r, c \to \infty$.

No three-way interaction
$A : \mathbb{N}^{n_1 \times n_2 \times n_3} \longrightarrow \mathbb{N}^{n_1 \times n_2} \times \mathbb{N}^{n_1 \times n_3} \times \mathbb{N}^{n_2 \times n_3}$
Minimal Markov basis for $n_3 = 2$ contains

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 1 & 0 \end{bmatrix} \text{ on top of } \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & -1 & 0 \end{bmatrix}$$

$\rightsquigarrow$ *not* finite up to $\mathrm{Sym}(n_1) \times \mathrm{Sym}(n_2)$ as $n_1, n_2 \to \infty$.

Independent set theorem

Set-up
d: number of discrete random variables
(e.g. month of birth, month of death, gender $\rightsquigarrow d = 3$)
n_i: number of values of variable i
Γ: collection of subsets of $\{1, \ldots, d\} \leftrightarrow$ prescribed marginals
(e.g. $\Gamma = \{12, 13, 23\}$)
$A_\Gamma : n_1 \times \ldots \times n_d$-table $\mapsto$ (its F-marginals$)_{F \in \Gamma}$

Theorem (Hillar-Sullivant, 2009)
$T \subseteq \{1, \ldots, d\}$ s.t. $|T \cap F| \leq 1$ for all $F \in \Gamma$
$\rightsquigarrow A_\Gamma$ has a finite Markov basis
up to $\prod_{i \in T} \mathrm{Sym}(n_i)$ as $n_i \to \infty$ for all $i \in T$.

Remarks
satisfied for $d = 2, \Gamma = \{1, 2\}, T = \{1, 2\}$
not for $d = 3, \Gamma = \{12, 13, 23\}, T = \{1, 2\}$
not necessary, e.g. $\Gamma = \{12 \ldots d\}$ and $T = \{1, \ldots, d\}$.

Gaussian factor analysis

Model

$Z_1, \ldots, Z_k \sim \mathcal{N}(0,1)$ independent factors
$X_1, \ldots, X_n$ observed
$X_i = \sum_{j=1}^{k} s_{ij} Z_j + \epsilon_i$
$\epsilon_i \sim \mathcal{N}(0, v_i)$ independent noise
$\mathcal{M}_{k,n} = \{\Sigma = SS^T + \mathrm{diag}(v) \mid S, v\}$

Raymond Cattell (1971): **fluid** vs **crystallised** intelligence

Proposal (Drton,Sturmfels,Sullivant 07)
Use polynomial relations among entries of $\Sigma \in \mathcal{M}_{k,n}$ to test model against data.

Example ($k = 2, n = 5$**)**

$\frac{1}{10} \sum_{\pi \in \mathrm{Sym}(5)} \mathrm{sgn}\,(\pi) \sigma_{\pi(1)\pi(2)} \sigma_{\pi(2)\pi(3)} \sigma_{\pi(3)\pi(4)} \sigma_{\pi(4)\pi(5)} \sigma_{\pi(5)\pi(1)} = 0;$
Pentad (Kelley, 1928, *Cross-roads in the mind of man*)
$\mathrm{codim}\, \mathcal{M}_{2,5} = 1$

Gaussian factor analysis, continued

Observation
$\Sigma \in \mathcal{M}_{k,n} \Rightarrow \Sigma[I] \in \mathcal{M}_{k,|I|}$

Question (Drton-Sturmfels-Sullivant)
equations for $\mathcal{M}_{k,n}$ as $n \longrightarrow \infty$?
$\exists n_0 \, \forall n \geq n_0$ all equations for $\mathcal{M}_{k,n}$ are generated by those for $\mathcal{M}_{k,n_0}$?

Theorem (de Loera-Sturmfels-Thomas 1995)
Yes for $k = 1$: $n_0 = 4$, off-diagonal 2×2-subdeterminants.

Theorem (Brouwer-D, 2010)
Yes for $k = 2$: $n_0 = 6$, pentads and off-diagonal 3×3-subdeterminants.

Theorem (D, 2009)
Set-theoretically yes for all k.

Fundamental tool

Increasing maps
$\mathrm{Inc}(\mathbb{N}) = \{\pi : \mathbb{N} \to \mathbb{N} \mid \pi(1) < \pi(2) < \ldots\}$
$\mathbb{C}[x_1, x_2, \ldots]$
$\pi x_i = x_{\pi(i)}$

Cohen (1967), Aschenbrenner-Hillar (2007)
I ideal in $\mathbb{C}[x_1, x_2, \ldots]$
$\mathrm{Inc}(\mathbb{N})$-stable
$\Rightarrow I$ generated by finitely many $\mathrm{Inc}(\mathbb{N})$-orbits
$\mathrm{Inc}(\mathbb{N})$-*Noetherian*

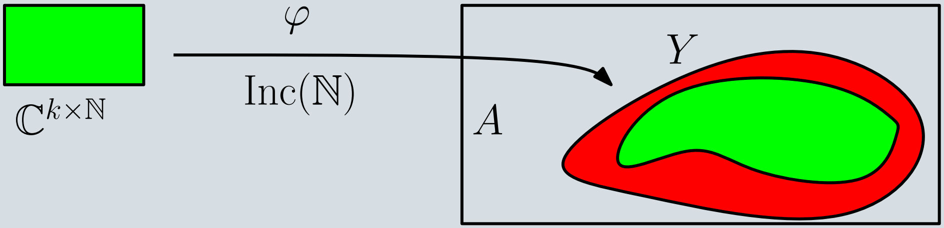
Cohen (1987), Hilllar-Sullivant (2009)

$$\mathbb{C}\begin{bmatrix} x_{11} & x_{12} & \cdots \\ \vdots & \vdots & \\ x_{k1} & x_{k2} & \cdots \end{bmatrix} \text{ is } \mathrm{Inc}(\mathbb{N})\text{-Noetherian.}$$

Replaces Hilbert's basis theorem in proofs of *finite up to symmetry*.

Further issues

1. infinite-dimensional Buchberger algorithm (Cohen 1987, La Scala-Levandovskyy 2010, Brouwer-D 2010)
2. Other monoids and well-quasi-orders? (Higman!, Kruskal?, Robertson-Seymour??)
3. Over real numbers, Euclidean closure?
4. Bottom line



(A, φ) "reasonable" $\Rightarrow$ image defined by finitely many $\mathrm{Inc}(\mathbb{N})$-orbits of equations.

Tensors of bounded rank

$V_1, \ldots, V_p$ vector spaces
$V_1 \otimes \cdots \otimes V_p$ tensor product

Rank ≤ 1
$\mu = v_1 \otimes \cdots \otimes v_p$

Rank $\leq k$
$\omega = \mu_1 + \ldots + \mu_k,\ \mu_i$ rank ≤ 1

Border rank $\leq k$
Zariski closure

Theorem (D-Kuttler, 2010)
$\forall k\, \exists d\, \forall p\, \forall V_1, \ldots, V_p$:
$\{$tensors of border rank $\leq k\}$
defined in degree $\leq d$.