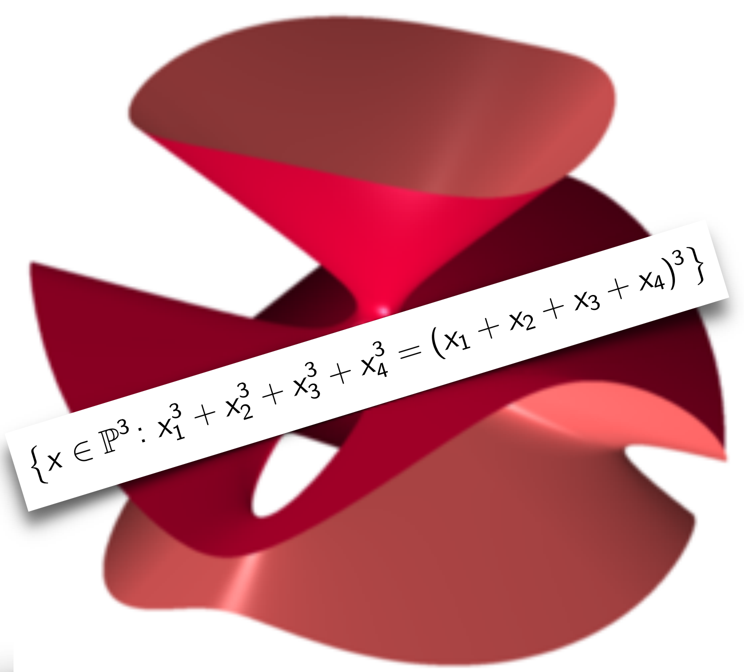
Algebraic Statistics & Quantifier Elimination

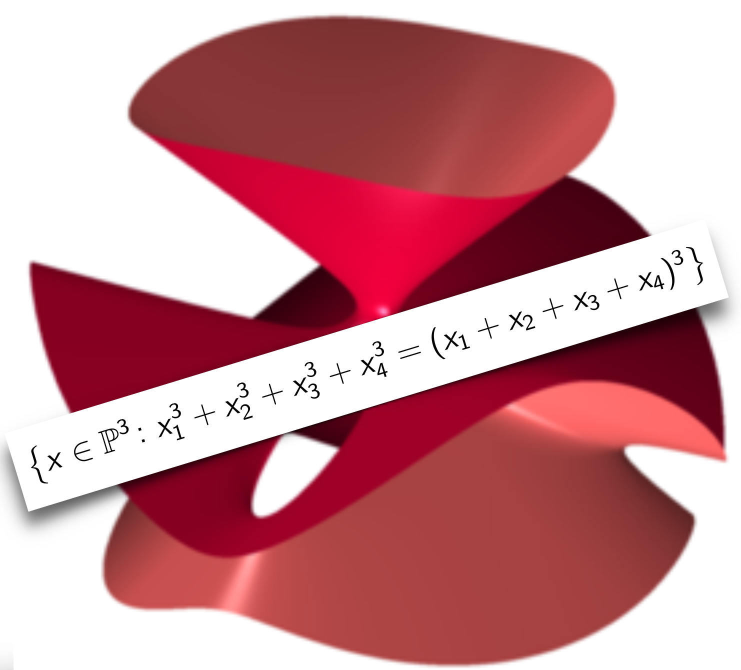
Daniel Suess

What is Algebraic Statistics?



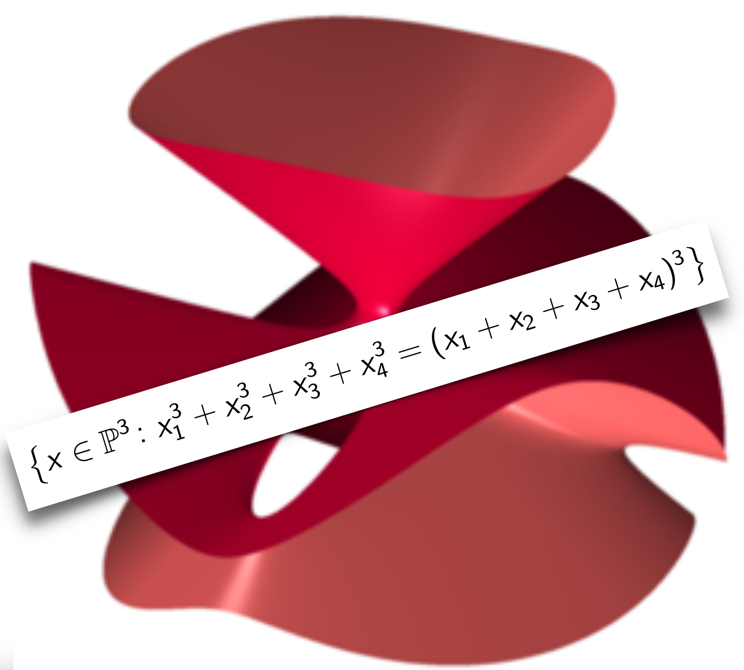
What is Algebraic Statistics?

understand algebraic
structure of statistical models



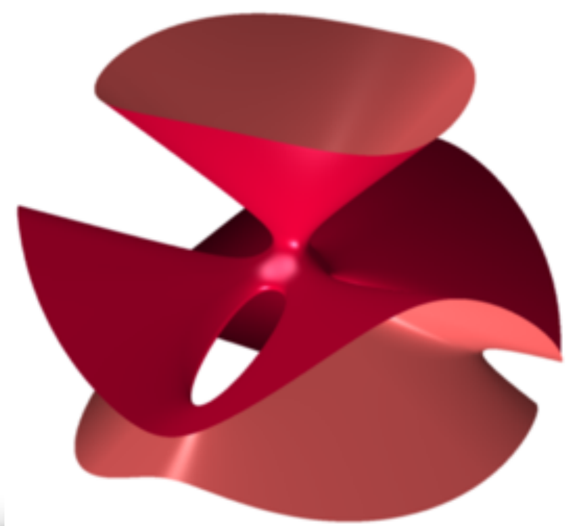
What is Algebraic Statistics?

understand algebraic
structure of statistical models



application of techniques from
computational algebra,
algebraic geometry, …

Algebraic Structure in Statistics

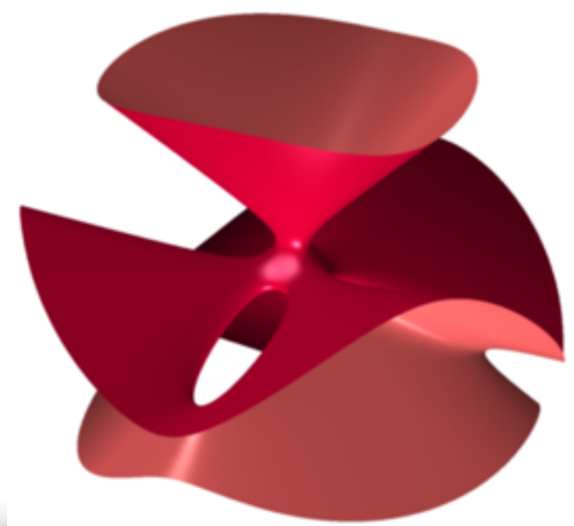


Semi-algebraic set $S \subset \mathbb{R}^n$

$$S = \Big\{x \in \mathbb{R}^n \colon\ p_i(x_1, \ldots, x_n) = 0,$$
$$q_j(x_1, \ldots, x_n) \leq 0\Big\}$$

for finitely many polynomials p_i, q_i with coefficients in $\mathbb{Q}$

Algebraic Structure in Statistics



Semi-algebraic set $S \subset \mathbb{R}^n$

$$S = \Big\{ x \in \mathbb{R}^n \colon\ p_i(x_1, \dots, x_n) = 0, \\ q_j(x_1, \dots, x_n) \le 0 \Big\}$$

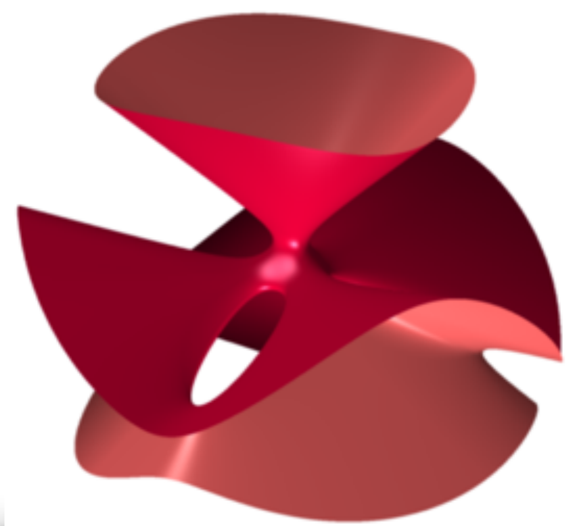
for finitely many polynomials p_i, q_i with coefficients in $\mathbb{Q}$

- probability simplex

  $$\Delta_{n-1} = \Big\{ x \in \mathbb{R}^n \colon\ \sum_i x_i = 1, x_i \ge 0 \Big\}$$

- joint prob. of discrete, indep. RVs
- max. likelihood estimates
- …

Algebraic Structure in Statistics



Semi-algebraic set $S \subset \mathbb{R}^n$

$$S = \Big\{x \in \mathbb{R}^n \colon\ p_i(x_1, \dots, x_n) = 0,\quad q_j(x_1, \dots, x_n) \leq 0\Big\}$$

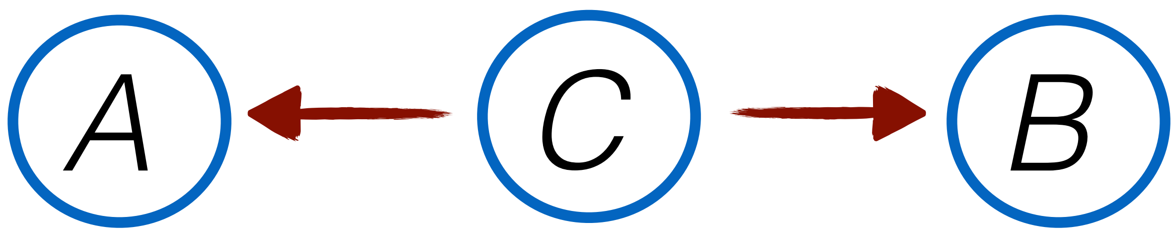
for finitely many polynomials p_i, q_i with coefficients in $\mathbb{Q}$

- probability simplex

  $$\Delta_{n-1} = \Big\{x \in \mathbb{R}^n \colon\ \sum_i x_i = 1, x_i \geq 0\Big\}$$

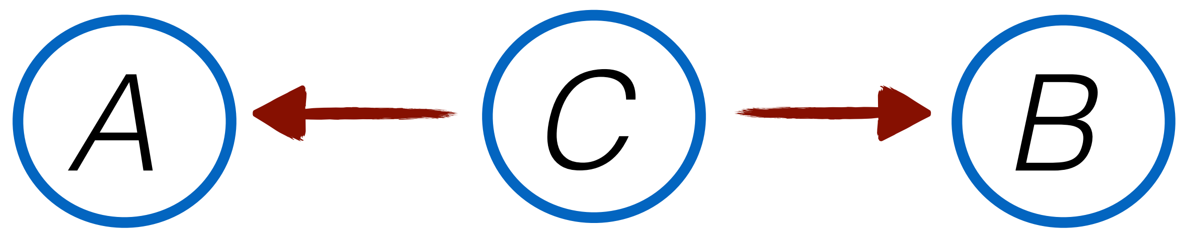
- joint prob. of discrete, indep. RVs
- max. likelihood estimates
- …

Conditional Independence: Explicit & Implicit Form



$$A \perp\!\!\!\perp B \mid C$$

Conditional Independence: Explicit & Implicit Form



$$A \perp\!\!\!\perp B \mid C$$

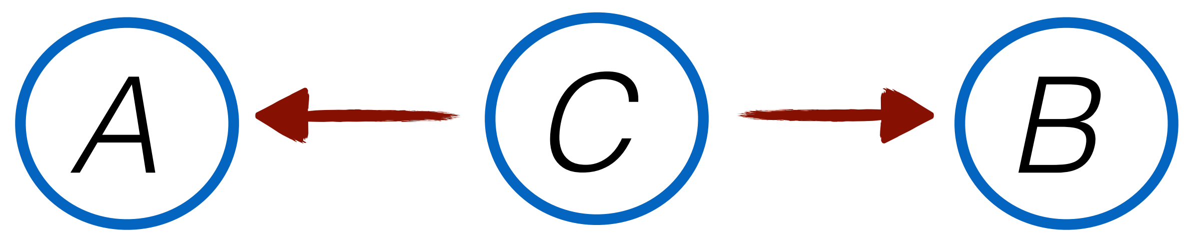
explicit

$$P(A, B, C) = P(A|C)\, P(B|C)\, P(C)$$

$$\Longleftrightarrow$$

$$p_{abc} = q_{ac}\, r_{bc}\, s_c$$

Conditional Independence: Explicit & Implicit Form



$$A \perp\!\!\!\perp B \mid C$$

$$P(A, B, C) = P(A|C)\, P(B|C)\, P(C)$$

$$\Longleftrightarrow$$

$$p_{abc} = q_{ac}\, r_{bc}\, s_c$$

$$p_{abc}\, p_{a'b'c} - p_{ab'c}\, p_{a'bc} = 0$$

Additional constraint for the semi-algebraic set of probabilities consistent with DAG.

Conditional Independence: Explicit & Implicit Form

$$A \leftarrow C \rightarrow B$$

$$A \perp\!\!\!\perp B \mid C$$

explicit

$$P(A, B, C) = P(A|C)\, P(B|C)\, P(C)$$

$$\Longleftrightarrow$$

$$p_{abc} = q_{ac}\, r_{bc}\, s_c$$

implicit

$$p_{abc}\, p_{a'b'c} - p_{ab'c}\, p_{a'bc} = 0$$

Additional constraint for the semi-algebraic set of probabilities consistent with DAG.

$$p_{abc} = q_{ac}\, r_{bc}\, s_c$$

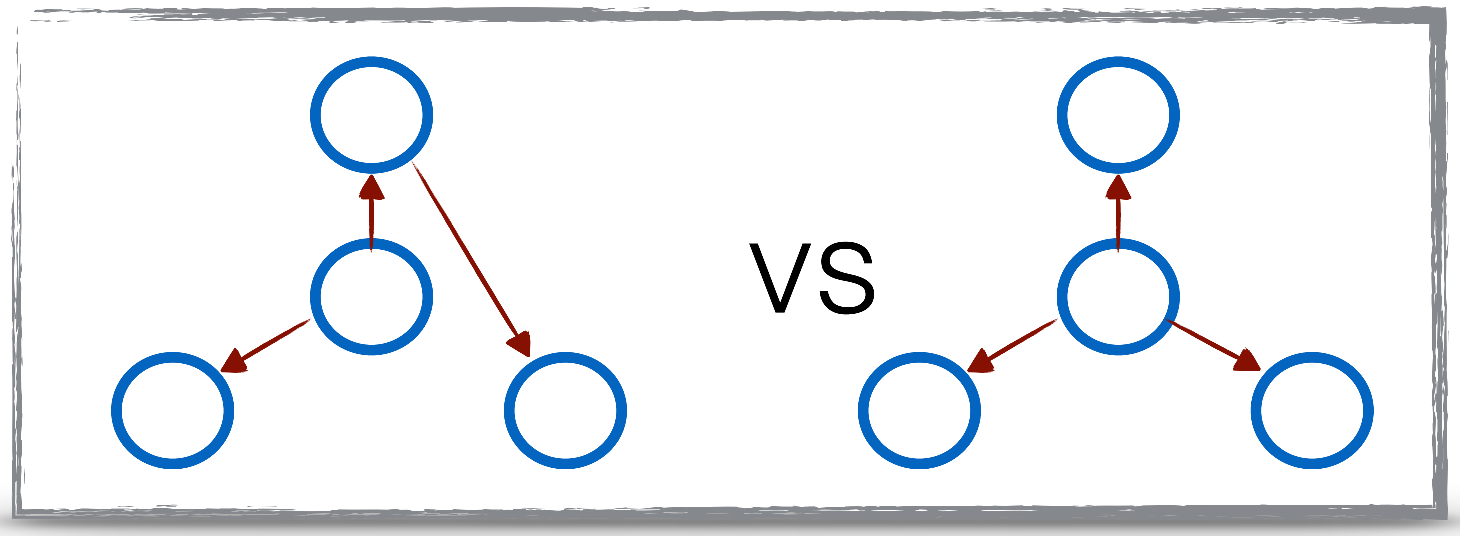
$\Longleftrightarrow$ for all c, $(p_{abc})_{ab}$ is rank 1

$\Longleftrightarrow$ $(p_{abc})_{ab}$ has determinantal rank 1

$\Longleftrightarrow$ determinant of all 2×2 minors $= 0$

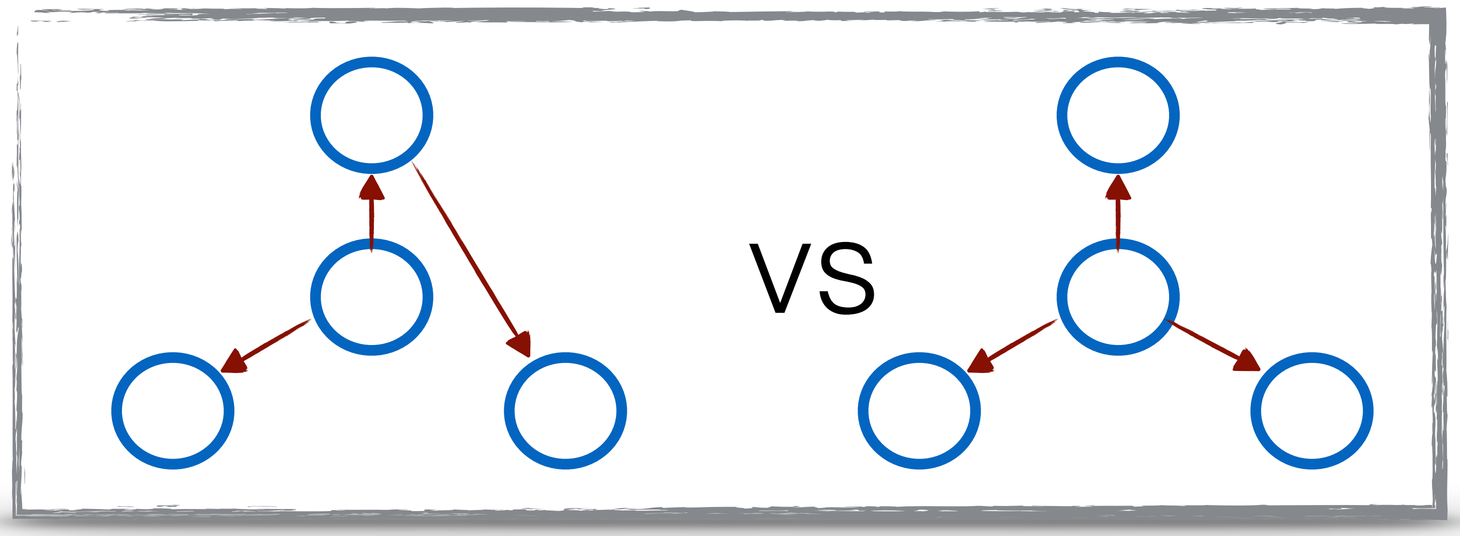
$\Longleftrightarrow$ $\forall a, a', b, b', c:\ p_{abc}\, p_{a'b'c} - p_{ab'c}\, p_{a'bc} = 0$

Possible Applications



distinguishing Bayesian networks from observational data

Possible Applications

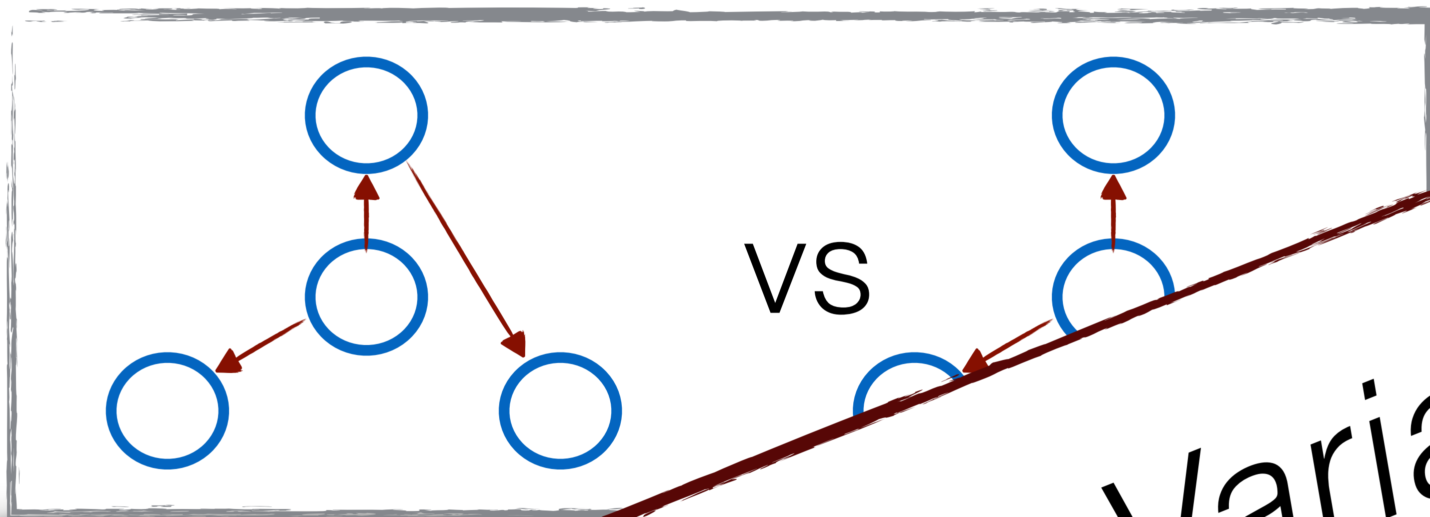


distinguishing Bayesian networks from observational data

Bell inequalities / quantum non locality

Image: http://www.belfasttelegraph.co.uk/news/northern-ireland/john-stewart-bell-belfast-city-council-to-bend-rules-on-the-naming-of-streets-to-honour-the-scientist-who-corrected-albert-einstein-30586279.html

Possible Applications



Bayesian

ata

Hidden Variables?

Bell inequa /
quantum non locality

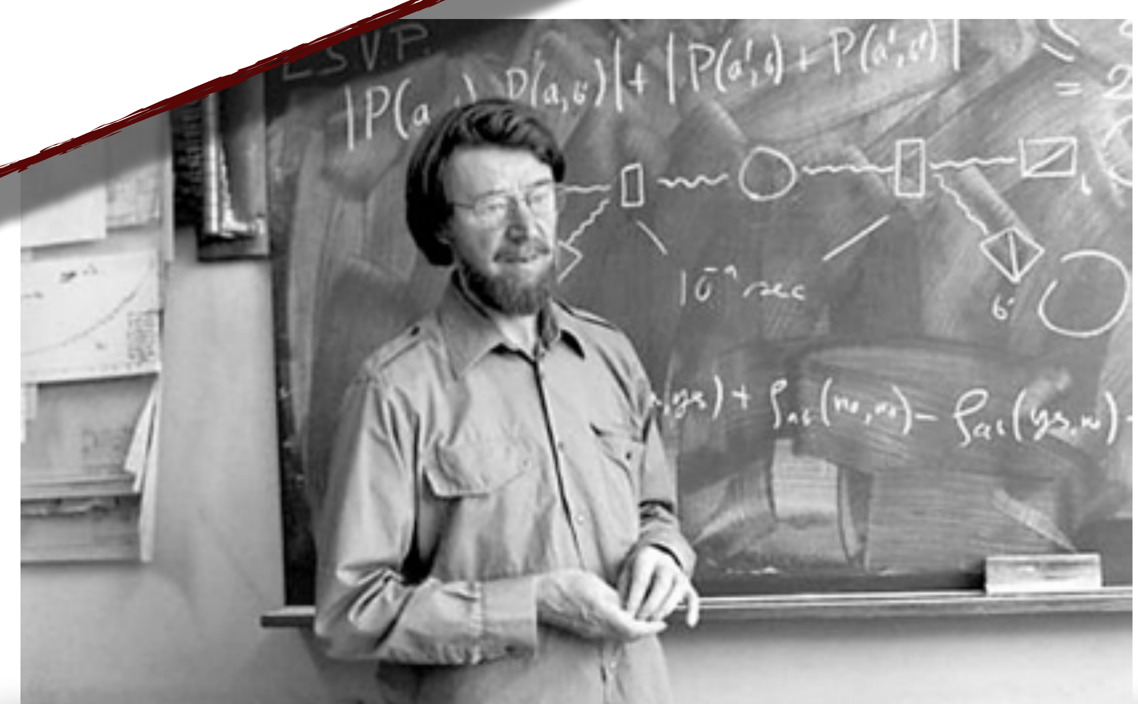


Image: http://www.belfasttelegraph.co.uk/news/northern-ireland/john-stewart-bell-belfast-city-council-to-bend-rules-on-the-naming-of-streets-to-honour-the-scientist-who-corrected-albert-einstein-30586279.html

Quantifier Elimination: Conditional Independence

p_{abc} compatible with $A \leftarrow C \rightarrow B$

$$\left\{ \begin{array}{ll} (p_{abc}): & \exists (q_{ac}) \exists (r_{bc}) \exists (s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

Quantifier Elimination: Conditional Independence

p_{abc} compatible with $A \leftarrow C \rightarrow B$

$$\left\{ (p_{abc}) : \begin{array}{l} \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ \underline{(\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c)} \end{array} \right\}$$

$$\boxed{P(A, B, C) = P(A|C)\, P(B|C)\, P(C)}$$

Quantifier Elimination: Conditional Independence

p_{abc} compatible with $A \leftarrow C \rightarrow B$

$$\left\{ \begin{array}{ll} (p_{abc}): & \exists (q_{ac}) \exists (r_{bc}) \exists (s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

$$= \pi \Big(\{ (p_{abc}, q_{ab}, r_{bc}, s_c) : \; \cdots \} \Big)$$

with $\pi : \mathbb{R}^{n_1 + n_2} \to \mathbb{R}^{n_1}, (p_{abc}, q_{ab}, r_{bc}, s_c) \mapsto (p_{abc})$

Quantifier Elimination: Conditional Independence

p_{abc} compatible with $A \leftarrow C \rightarrow B$

$$\left\{ (p_{abc}) : \quad \begin{array}{l} \exists (q_{ac}) \exists (r_{bc}) \exists (s_c) \\ (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

$$= \pi \Big(\{ (p_{abc}, q_{ab}, r_{bc}, s_c) : \ \cdots \} \Big)$$

with $\pi : \mathbb{R}^{n_1+n_2} \to \mathbb{R}^{n_1}, (p_{abc}, q_{ab}, r_{bc}, s_c) \mapsto (p_{abc})$

Theorem (Tarski-Seidenberg) The image of a semi-algebraic set under a projection map π is a semi-algebraic set.

Quantifier Elimination: Conditional Independence

p_{abc} compatible with $A \leftarrow C \rightarrow B$

$$\left\{ \begin{array}{ll} (p_{abc}) : & \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

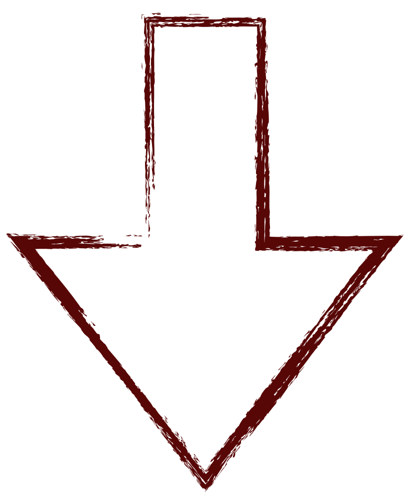
$$= \pi\Big(\{(p_{abc}, q_{ab}, r_{bc}, s_c) : \ \cdots \} \Big)$$

with $\pi : \mathbb{R}^{n_1+n_2} \to \mathbb{R}^{n_1}, (p_{abc}, q_{ab}, r_{bc}, s_c) \mapsto (p_{abc})$

$$\boxed{\begin{array}{c} \{(p, q) \in \mathbb{R}^2 : (\exists x \in \mathbb{R})\, x^2 + px + q = 0\} \\ = \{(p, q) \in \mathbb{R}^2 : p^2 \geq 4q\} \end{array}}$$

Quantifier Elimination: Algorithms

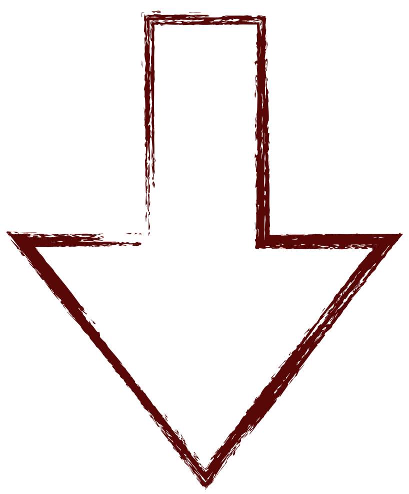
Input: formula Ψ



Output: equivalent,
quantifier-free formula Ψ'

Quantifier Elimination: Algorithms

Input: formula Ψ



Output: equivalent, quantifier-free formula Ψ'

- real valued variables
- rational constants
- operations $(+,-,\times)$
- binary relations $(=,\neq,<,\leq)$
- logical connectives $(\wedge,\vee,\neg,\Rightarrow,\Leftrightarrow)$
- quantifiers $(\forall,\exists)$

Quantifier Elimination: Algorithms

Input: formula ψ

Output: equivalent, quantifier-free formula ψ'

- real valued variables
- rational constants
- operations $(+,-,\times)$
- binary relations $(=,\neq,<,\leq)$
- logical connectives $(\wedge,\vee,\neg,\Rightarrow,\Leftrightarrow)$
- quantifiers $(\forall,\exists)$

- Tarski's algorithm
- Cylinder Algebraic decomposition

Quantifier Elimination: Hidden Variables

p_{abc} compatible with $A \leftarrow C \rightarrow B$

$$\left\{ \begin{array}{ll} (p_{abc}): & \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

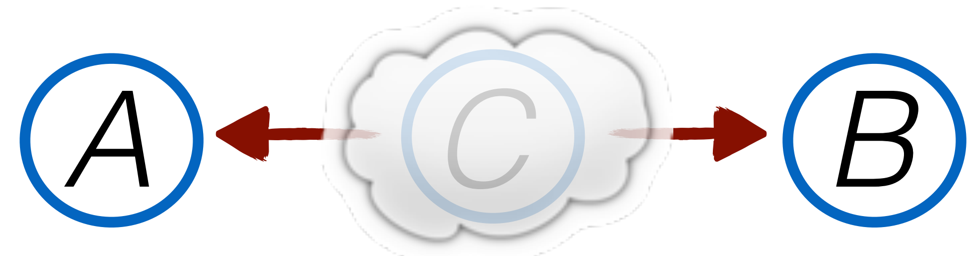
Quantifier Elimination: Hidden Variables

p_{abc} compatible with 

$$\left\{ (p_{abc}) : \begin{array}{l} \exists (q_{ac}) \exists (r_{bc}) \exists (s_c) \\ (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

Quantifier Elimination: Hidden Variables

p_{ab}

$\cancel{p_{abc}}$ compatible with 

$$\left\{ \begin{array}{ll} (p_{abc}): & \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

Quantifier Elimination: Hidden Variables

p_{ab}

$\cancel{p_{abc}}$ compatible with

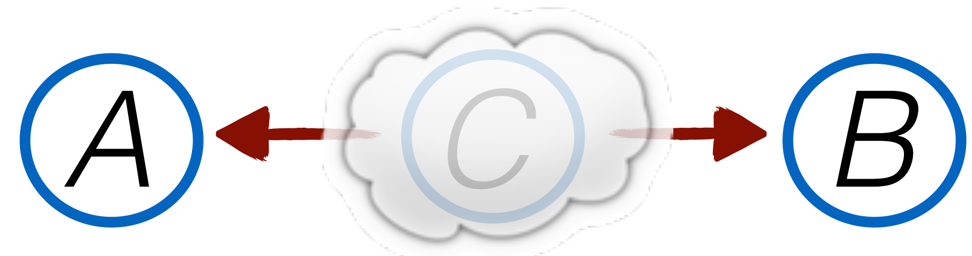
(p_{ab})

$$\left\{ \begin{array}{ll} \cancel{(p_{abc})}: & \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & (\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c) \end{array} \right\}$$

Quantifier Elimination: Hidden Variables

p_{ab}

$\cancel{p_{abc}}$ compatible with



(p_{ab})

$$\left\{ \begin{array}{ll} \cancel{(p_{abc})}: & \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & \cancel{(\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c)} \end{array} \right\}$$

$$\bigwedge_{ab} p_{ab} = \sum_c q_{ac} \times r_{bc} \times s_c$$

Quantifier Elimination: Hidden Variables

p_{ab}

$\cancel{p_{abc}}$ compatible with

$$(p_{ab})$$

$$\left\{ \begin{array}{ll} \cancel{(p_{abc})}: & \exists(q_{ac})\exists(r_{bc})\exists(s_c) \\ & (\bigwedge_c \sum_a q_{ac} = 1) \wedge (\bigwedge_c \sum_b r_{bc} = 1) \wedge (\sum_c s_c = 1) \wedge \\ & (\bigwedge_{ac} q_{ac} \geq 0) \wedge (\bigwedge_{bc} r_{bc} \geq 0) \wedge (\bigwedge_c s_c \geq 0) \wedge \\ & \cancel{(\bigwedge_{abc} p_{abc} = q_{ac} \times r_{bc} \times s_c)} \end{array} \right\}$$

$$\bigwedge_{ab} p_{ab} = \sum_c q_{ac} \times r_{bc} \times s_c$$

- no additional constraints on p_{ab} for $N_A = N_B = N_C = 2$
- $N_A = N_B = 3$ did not finish

Quantifier Elimination: Other Applications

Membership Problem

$$(A \perp C) \wedge (A \perp B \mid C) \Rightarrow (A \perp B)\ ?$$

Quantifier Elimination: Other Applications

Membership Problem

$$\underbrace{(A \perp C) \wedge (A \perp B \mid C)}_{\psi(p)} \Rightarrow \underbrace{(A \perp B)}_{\psi'(p)}\ ?$$

Quantifier Elimination: Other Applications

Membership Problem

$$\underbrace{(A \perp C) \wedge (A \perp B \mid C)}_{\psi(p)} \Rightarrow \underbrace{(A \perp B)}_{\psi'(p)}\ ?$$

$$\forall p \colon \psi(p) \implies \psi'(p)$$

Quantifier Elimination: Other Applications

Membership Problem

$$\underbrace{(A \perp C) \wedge (A \perp B \mid C)}_{\psi(p)} \Rightarrow \underbrace{(A \perp B)}_{\psi'(p)}\ ?$$

$$\forall p \colon \psi(p) \implies \psi'(p)$$

Identifiability Problem

Can we uniquely identify the parameter θ in a parametric model from given observation?

Quantifier Elimination: Other Applications

Membership Problem	Identifiability Problem
$\underbrace{(A \perp C) \wedge (A \perp B \mid C)}_{\psi(p)} \Rightarrow \underbrace{(A \perp B)}_{\psi'(p)}$?	*Can we uniquely identify the parameter θ in a parametric model from given observation?*
$\forall p \colon \psi(p) \implies \psi'(p)$	$\forall \theta, \theta' \colon (\mathrm{O}(\theta) = \mathrm{O}(\theta')) \implies (\theta = \theta')$

Conclusion

Conclusion

- probabilities compatible with directed Bayesian network constitute semi-algebraic set:

Conclusion

- probabilities compatible with directed Bayesian network constitute semi-algebraic set:
  - no hidden variables → quadratic equations

Conclusion

- probabilities compatible with directed Bayesian network constitute semi-algebraic set:
  - no hidden variables → quadratic equations
  - hidden variables → possibly QE

Conclusion

- probabilities compatible with directed Bayesian network constitute semi-algebraic set:
  - no hidden variables → quadratic equations
  - hidden variables → possibly QE
- CAD: best known algorithm for QE, but too slow beyond toy models

Conclusion

- probabilities compatible with directed Bayesian network constitute semi-algebraic set:
  - no hidden variables → quadratic equations
  - hidden variables → possibly QE
- CAD: best known algorithm for QE, but too slow beyond toy models
- also for continuous RV (exponential families)

Conclusion

- probabilities compatible with directed Bayesian network constitute semi-algebraic set:
  - no hidden variables → quadratic equations
  - hidden variables → possibly QE
- CAD: best known algorithm for QE, but too slow beyond toy models
- also for continuous RV (exponential families)
- https://github.com/dseuss/algstat.git

Theorem (Tarski) Every Tarski sentence (formula without free variables) is equivalent to a quantifier-free formula.

Theorem (Tarski-Seidenberg) Let $f\colon X \to Y$ be a semi-algebraic map. Then, the image $f(X) \subseteq Y$ is a semi-algebraic set.

ALGORITHM ComputeCAD(F,j)

Input: $\mathrm{F} \subset Q[x_1, \ldots, x_j]$.

Output: (K_j, α_j) where K_j is an F-sign-invariant *CAD* of R^j and
α_j is a set of algebraic sample points, one per cell in K_j.

Recurse: If $j > 1$, then do
$\Phi(\mathrm{F}) := \Phi_1(\mathrm{F}) \cup \Phi_2(\mathrm{F}) \cup \Phi_3(\mathrm{F})$
$(K_{j-1}, \alpha_{j-1}) := ComputeCAD(\Phi(\mathrm{F}), j-1)$,
else find the roots $r_1, \ldots, r_m$ of all polynomials in F and do
$K_1 := \{[-\infty, r_1), [r_1, r_1], (r_1, r_2), \ldots, (r_m, +\infty]\}$
$\alpha_1 := \{r_1 - 1, r_1, (r_1 + r_2)/2, \ldots, r_m, r_m + 1\}$
Return (K_1, α_1)

Lift: For every cell $C_i \in K_{j-1}$ do

1. Compute the product of all polynomials in F that do not vanish at the sample point α_i of C_i and call the resulting polynomial $\pi(\alpha_i, x)$

2. Find the roots $r_1, \ldots, r_m$ of $\pi(\alpha_i, x)$

3. Set $K_{j,i} := \{\{C_i \times [-\infty, r_1)\}, \{C_i \times [r_1, r_1]\}, \{C_i \times (r_1, r_2)\}, \ldots, \{C_i \times (r_m, +\infty]\}\}$
Comment: $K_{j,i}$ are the cylinders over C_i.

4. Set $\alpha_{j,i} := \{(\alpha_i, r_1 - 1), (\alpha_i, r_1), (\alpha_i, (r_1 + r_2)/2), \ldots, (\alpha_i, r_m + 1)\ \}$
Comment: $\alpha_{j,i}$ are the algebraic sample points for the cylinders over C_i.

$K_j := \bigcup_i K_{j,i}$; $\alpha_j := \bigcup_i \alpha_{j,i}$

Return (K_j, α_j)