

Posets and algebraic statistics

Nicholas Eriksson

Department of Statistics
Stanford University

July 2, 2007

Outline

- 1 Posets
- 2 Algebra
- 3 Conjunctive Bayesian networks

Structural learning

How do you learn the graph?

- Bayesian MCMC search over graphs — maximize a likelihood function.
- Structural EM algorithm [Friedman-Koller]
- Covariance matrix methods.
- ...

Posets

Definition

A **poset** (P, \leq) is a set P with a binary relation \leq which is

antisymmetric $x \leq y \Leftrightarrow y \not\leq x$

reflexive $x \leq x$

transitive $x \leq y, y \leq z \Rightarrow x \leq z$

Distributive lattices

Definition

A (lower) **order ideal** I in a poset P is a set $I \subset P$ such that if $x \in P$ and $y \in I$ and $x \leq y$, then $x \in I$.

Definition

The **distributive lattice** of P , written $J(P)$ is the set of all order ideals of P , ordered by inclusion.

Theorem

- Every distributive lattice is equal to $J(P)$ for some poset P .
- The meet-irreducible elements of $J(P)$ form a poset isomorphic to P .

Varieties

Definition

A (parametric) **variety** V is (the Zariski closure of) the set of points $(x_1, \dots, x_n) \in \mathbb{R}^n$ such that

$$x_1 = f_1(\theta_1, \dots, \theta_m)$$

$$\vdots$$

$$x_n = f_n(\theta_1, \dots, \theta_m)$$

where $(\theta_1, \dots, \theta_m) \in \mathbb{R}^m$, and f_1, \dots, f_n are polynomials in m unknowns.

Chains

Definition

A **chain** C in a poset P is a totally ordered subset $C \subset P$. "Totally ordered" means that for any two elements $x, y \in C$, either $x \leq y$ or $y \leq x$.

Ideals

Definition

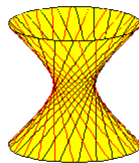
The ideal of a variety $V \subset \mathbb{R}^n$ is the set of all polynomials g in n unknowns x_1, \dots, x_n such that

$$f(x_1, \dots, x_n) = 0 \quad \text{if} \quad (x_1, \dots, x_n) \in V$$

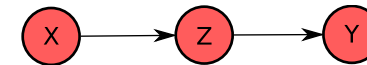
Independence

Two independent binary random variables X, Y . Set $\Pr(X = 0) = \alpha$ and $\Pr(Y = 0) = \beta$.

$$\Pr(X, Y) : \begin{array}{cc} & X = 0 & X = 1 \\ \begin{array}{c} Y = 0 \\ Y = 1 \end{array} & \begin{pmatrix} \alpha\beta & (1-\alpha)\beta \\ \alpha(1-\beta) & (1-\alpha)(1-\beta) \end{pmatrix} \end{array}$$



Graphical model



$$\Pr(X, Z, Y) = \Pr(X) \Pr(Z|X) \Pr(Y|Z)$$

5 parameters:

$$\begin{aligned} \theta^X &= \Pr(X = 0) & \theta_0^Z &= \Pr(Z = 0|X = 0) & \theta_0^Y &= \Pr(Y = 0|Z = 0) \\ \theta_1^Z &= \Pr(Z = 0|X = 1) & \theta_1^Y &= \Pr(Y = 0|Z = 1) \end{aligned}$$

8 probabilities: $p_{ijk} := \Pr(X = i, Y = j, Z = k)$.

Ideal:

$$\begin{aligned} &\langle p_{000} + p_{001} + p_{010} + p_{011} + p_{100} + p_{101} + p_{110} + p_{111} - 1, \\ &\quad p_{011}p_{110} - p_{010}p_{111}, \\ &\quad p_{001}p_{100} + p_{001}p_{101} + p_{010}p_{101} + p_{011}p_{101} + \\ &\quad p_{100}p_{101} + p_{101}^2 + p_{101}p_{110} + p_{101}p_{111} - p_{101} \rangle \end{aligned}$$

Why algebraic statistics?

- For statistics:
 - ▶ Sampling from discrete exponential families [Diaconis–Sturmfels 1998, ...]
 - ▶ Identifiability [Allman–Rhodes 2005, ...]
 - ▶ Likelihood ratio tests and singularities [Drton 2007]
 - ▶ New model selection algorithms [Eriksson-Yao 2007, ...]
- Computational biology:
 - ▶ Parametric inference [Waterman–Eggert–Lander 1992, ...]
 - ▶ Phylogenetic invariants [Cavender–Felsenstein 1987, ...]
 - ▶ ...
- For mathematics: problems in
 - ▶ algebraic geometry
 - ▶ combinatorics
 - ▶ discrete geometry
 - ▶ ...

Conjunctive Bayesian networks

A CBN \mathcal{E} is the directed graphical model with binary random variables $(X_e)_{e \in \mathcal{E}}$ and graph \mathcal{E} and conditional probability tables are

$$[\Pr(X_e = b \mid X_{\text{pa}(e)} = a)]_{a \in \{0,1\}^{\text{pa}(e)}, b \in \{0,1\}} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 - \theta_e & \theta_e \end{bmatrix},$$

where $\text{pa}(e)$ denotes the parents of e in the acyclic directed graph \mathcal{E} .

Joint probability distribution

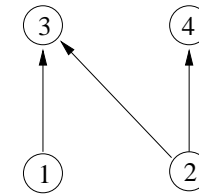
For each event e , let θ_e be the probability that e occurs (and require that the genotypes respect the event orders).

$$P_g(\theta) = \prod_{e \in g} \theta_e \cdot \prod_{e \in \min(g^c)} (1 - \theta_e)$$

Events that have occurred Events that could occur, but haven't

Example

$$\begin{aligned} \theta^1 &= (a \quad 1 - a) \\ \theta^2 &= (b \quad 1 - b) \\ \theta^3 &= \begin{pmatrix} c_{00} & 1 - c_{00} \\ c_{01} & 1 - c_{01} \\ c_{10} & 1 - c_{10} \\ c_{11} & 1 - c_{11} \end{pmatrix} \\ \theta^4 &= \begin{pmatrix} d_0 & 1 - d_0 \\ d_1 & 1 - d_1 \end{pmatrix} \end{aligned}$$



Conjunctive Bayesian network: $c_{00} = c_{01} = c_{10} = d_0 = 0$.

$$\begin{aligned} p_{0000} &= (1 - a)(1 - b), & p_{0100} &= (1 - a)b(1 - d_1) \\ p_{0101} &= (1 - a)bd_1, & p_{1000} &= a(1 - b) \\ p_{1100} &= ab(1 - c_{11})(1 - d_1), & p_{1101} &= ab(1 - c_{11})d_1 \\ p_{1110} &= abc_{11}(1 - d_1), & p_{1111} &= abc_{11}d_1 \end{aligned}$$

A change of variables (Möbius inversion)

The probability of observing genotype g

$$P_g(\theta) = \prod_{e \in g} \theta_e \cdot \prod_{e \in \min(g^c)} (1 - \theta_e).$$

The probability of observing **at least** genotype h

$$Q_h(\theta) = \sum_{g: h \subseteq g} P_g(\theta) = \prod_{e \in h} \theta_e.$$

This suggests we should change coordinates:

$$q_h = \sum_{g: h \subseteq g} p_g.$$

In q_h coordinates, the conjunctive Bayesian network becomes a toric variety with Gröbner basis (Hibi)

$$\langle q_g q_h - q_{g \cup h} q_{g \cap h} \mid g, h \text{ incomparable} \rangle$$

A Gröbner basis

Fix a linear extension of the reverse inclusion order on \mathcal{G} . Let \prec be the degree reverse lexicographic monomial ordering on $\mathbb{R}[p_g : g \in \mathcal{G}]$ induced by the variable ordering given by that linear extension.

Theorem (B-E-S)

The reduced Gröbner basis of the ideal $I_{\mathcal{E}}$ with respect to the monomial ordering \prec consists of the trivial linear invariant $\sum_{g \in \mathcal{G}} p_g - 1$, with leading term p_{\emptyset} , together with one homogeneous quadratic polynomial

$$p_g \cdot p_h - p_{g \cup h} \cdot p_{g \cap h} + \prec\text{-lower terms}$$

for each incomparable pair of genotypes $\{g, h\}$ in the distributive lattice \mathcal{G} .