Algebraic Statistics Tutorial I

Seth Sullivant

North Carolina State University

July 22, 2012

Example: Hardy-Weinberg Equilibrium

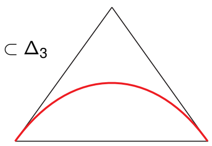
Suppose a gene has two alleles, a and A. If allele a occurs in the population with frequency θ (and A with frequency $1-\theta$) and these alleles are in Hardy-Weinberg equilibrium, the genotype frequencies are

$$\mathrm{P}(X = aa) = \theta^2, \mathrm{P}(X = aA) = 2\theta(1-\theta), \mathrm{P}(X = AA) = (1-\theta)^2$$

The model of Hardy-Weinberg equilibrium is the set

$$\mathcal{M} = \left\{ \left(\theta^2, 2\theta(1-\theta), (1-\theta)^2\right) \mid \theta \in [0,1] \right\} \subset \Delta_3$$

$$\mathcal{I}(\mathcal{M}) = \langle p_{aa} + p_{aA} + p_{AA} - 1, p_{aA}^2 - 4p_{aa}p_{AA} \rangle$$



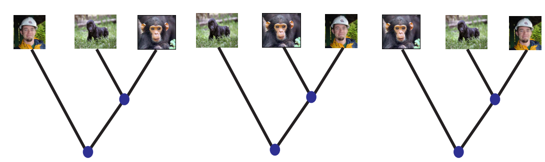
Main Point of This Tutorial

- Many statistical models are described by (semi)-algebraic constraints on a natural parameter space.
  - Generators of the vanishing ideal can be useful for constructing algorithms or analyzing properties of statistical model.
- Two Examples
  - Phylogenetic Algebraic Geometry
  - Sampling Contingency Tables

Phylogenetics

Problem

Given a collection of species, find the tree that explains their history.



- Data consists of aligned DNA sequences from homologous genes

```
Human:   ...ACCGTGCAACGTGAACGA...
Chimp:   ...ACCTTGGAAGGTAAACGA...
Gorilla: ...ACCGTGCAACGTAAACTA...
```


Model-Based Phylogenetics

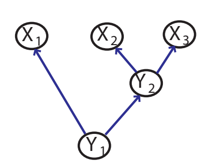
- Use a probabilistic model of mutations
- Parameters for the model are the combinatorial tree T, and rate parameters for mutations on each edge
- Models give a probability for observing a particular aligned collection of DNA sequences

```
Human:   ACCGTGCAACGTGAACGA
Chimp:   ACGTTGCAAGGTAAACGA
Gorilla: ACCGTGCAACGTAAACTA
```

- Assuming site independence, data is summarized by empirical distribution of columns in the alignment.
- e.g. $\hat{p}(AAA) = \frac{6}{18}$, $\hat{p}(CGC) = \frac{2}{18}$, etc.
- Use empirical distribution and test statistic to find tree best explaining data

Phylogenetic Models

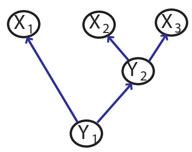
- Assuming site independence:
- Phylogenetic Model is a latent class graphical model
- Vertex $v \in T$ gives a random variable $X_v \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- All random variables corresponding to internal nodes are latent



$$P(x_1, x_2, x_3) = \sum_{y_1} \sum_{y_2} P(y_1)P(y_2|y_1)P(x_1|y_1)P(x_2|y_2)P(x_3|y_2)$$

Phylogenetic Models

- Assuming site independence:
- Phylogenetic Model is a latent class graphical model
- Vertex $v \in T$ gives a random variable $X_v \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$
- All random variables corresponding to internal nodes are latent



$$p_{i_1 i_2 i_3} = \sum_{j_1} \sum_{j_2} \pi_{j_1} a_{j_2, j_1} b_{i_1, j_1} c_{i_2, j_2} d_{i_3, j_2}$$

Algebraic Perspective on Phylogenetic Models

- Once we fix a tree T and model structure, we get a map $\phi^T : \Theta \to \mathbb{R}^{4^n}$.
- $\Theta \subseteq \mathbb{R}^d$ is a parameter space of numerical parameters (transition matrices associated to each edge).
- The map ϕ^T is given by polynomial functions of the parameters.
- For each $i_1 \cdots i_n \in \{A, C, G, T\}^n$, $\phi^T_{i_1 \cdots i_n}(\theta)$ gives the probability of the column $(i_1, \ldots, i_n)'$ in the alignment for the particular parameter choice θ.

$$\phi^T_{i_1 i_2 i_3}(\pi, a, b, c, d) = \sum_{j_1} \sum_{j_2} \pi_{j_1} a_{j_2, j_1} b_{i_1, j_1} c_{i_2, j_2} d_{i_3, j_2}$$

- The phylogenetic model is the set $\mathcal{M}_T = \phi^T(\Theta) \subseteq \mathbb{R}^{4^n}$.

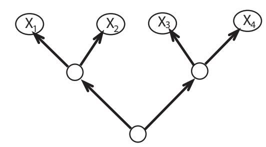
Phylogenetic Varieties and Phylogenetic Invariants

- Let $\mathbb{R}[p] := \mathbb{R}[p_{i_1 \cdots i_n} : i_1 \cdots i_n \in \{A, C, G, T\}^n]$

Definition

Let
$$I_T := \langle f \in \mathbb{R}[p] : f(p) = 0 \text{ for all } p \in \mathcal{M}_T \rangle \subseteq \mathbb{R}[p].$$
I_T is the ideal of phylogenetic invariants of T.
Let
$$V_T := \{p \in \mathbb{R}^{4^n} : f(p) = 0 \text{ for all } f \in I_T\}.$$
V_T is the phylogenetic variety of T.

- Note that $\mathcal{M}_T \subset V_T$.
- Since $\mathcal{M}_T$ is image of a polynomial map $\dim \mathcal{M}_T = \dim V_T$.



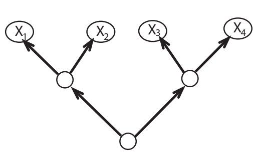
$$\begin{aligned} p_{lmno} &= \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \pi_i a_{ij} b_{ik} c_{jl} d_{jm} e_{kn} f_{ko} \\ &= \sum_{i=1}^{4} \pi_i \left(\left(\sum_{j=1}^{4} a_{ij} c_{jl} d_{jm} \right) \cdot \left(\sum_{k=1}^{4} b_{ik} e_{kn} f_{ko} \right) \right) \end{aligned}$$

$$\Longrightarrow \quad \operatorname{rank} \begin{pmatrix} p_{1111} & p_{1112} & \cdots & p_{1144} \\ p_{1211} & p_{1212} & \cdots & p_{1244} \\ \vdots & \vdots & \ddots & \vdots \\ p_{4411} & p_{4412} & \cdots & p_{4444} \end{pmatrix} \leq 4$$

Splits and Phylogenetic Invariants

Definition

A split of a set is a bipartition $A|B$. A split $A|B$ of the leaves of a tree T is valid for T if the induced trees $T|_A$ and $T|_B$ do not intersect.



- Valid: 12|34
- Not Valid: 13|24

2-way Flattenings and Matrix Ranks

$$p_{ijkl} = \mathrm{P}(X_1 = i, X_2 = j, X_3 = k, X_4 = l)$$

$$\mathrm{Flat}_{12|34}(P) = \begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & \cdots & p_{AATT} \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & \cdots & p_{ACTT} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{TTAA} & p_{TTAC} & p_{TTAG} & \cdots & p_{TTTT} \end{pmatrix}$$

Proposition

Let $P \in \mathcal{M}_T$.
- *If $A|B$ is a valid split for T, then $\mathrm{rank}(\mathrm{Flat}_{A|B}(P)) \leq 4$. Invariants in I_T are subdeterminants of $\mathrm{Flat}_{A|B}(P)$.*
- *If $C|D$ is not a valid split for T, then generically $\mathrm{rank}(\mathrm{Flat}_{C|D}(P)) > 4$.*

Phylogenetic Algebraic Geometry

Phylogenetic Algebraic Geometry is the study of the phylogenetic varieties and ideals V_T and I_T.

- Using Phylogenetic Invariants to Reconstruct Trees
- Identifiability of Phylogenetic Models
- Interesting Math– Useful in Other Problems

Using Phylogenetic Invariants to Reconstruct Trees

Definition

A phylogenetic invariant $f \in I_T$ is phylogenetically informative if there is some other tree T' such that $f \notin I_{T'}$.

- Idea of Cavender-Felsenstein (1987), Lake (1987): Evaluate phylogenetically informative phylogenetic invariants at empirical distribution $\hat{p}$ to reconstruct phylogenetic trees

Proposition

For each n-leaf trivalent tree T, let $\mathcal{F}_T \subseteq I_T$ be a set of phylogenetic invariants such that, for each $T' \neq T$, there is an $f \in \mathcal{F}_T$, such that $f' \notin I_{T'}$.
Let $f_T := \sum_{f \in \mathcal{F}_T} |f|$.
Then for generic $p \in \cup \mathcal{M}_T$, $f_T(p) = 0$ if and only if $p \in \mathcal{M}_T$.

Performance of Invariants Methods in Simulations

- Huelsenbeck (1995) did a systematic simulation comparison of 26 different methods of constructing a phylogenetic tree on 4 leaf trees. Invariant-based methods did poorly.
- HOWEVER... Huelsenbeck only used linear invariants.
- Casanellas, Fernandez-Sanchez (2006) redid these simulations using a generating set of the phylogenetic ideal I_T. Phylogenetic invariants become comparable to other methods.
- For the particular model studied in Casanellas, Fernandez-Sanchez (2006) for a tree with 4 leaves, the ideal I_T has 8002 generators.
$$f_T := \sum_{f \in \mathcal{F}_T} |f|$$
is a sum of 8002 terms.
- Major work to overcome combinatorial explosion for larger trees.

Identifiability of Phylogenetic Models

Definition

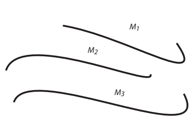
A parametric statistical model is identifiable if it gives 1-to-1 map from parameters to probability distributions.

- "Is it possible to infer the parameters of the model from data?"
- Identifiability guarantees consistency of statistical methods (ML)
- Two types of parameters to consider for phylogenetic models:
  - Numerical parameters (transition matrices)
  - Tree parameter (combinatorial type of tree)

Geometric Perspective on Identifiability

Definition

The unrooted tree parameter T in a phylogenetic model is identifiable if for all
$$p \in \mathcal{M}_T$$
there does not exist another $T' \neq T$ such that
$$p \in \mathcal{M}_{T'}.$$



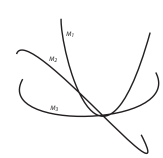
Identifiable

Not Identifiable

Generic Identifiability

Definition

The tree parameter in a phylogenetic model is generically identifiable if for all n-leaf trees with $T \neq T'$,
$$\dim(\mathcal{M}_T \cap \mathcal{M}_{T'}) < \min(\dim(\mathcal{M}_T), \dim(\mathcal{M}_{T'})).$$



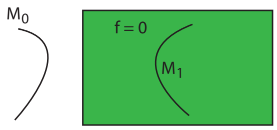
Proving Identifiability with Algebraic Geometry

Proposition

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two algebraic models. If there exist *phylogenetically informative invariants* *f_0 and f_1 such that*

$$f_i(p) = 0 \text{ for all } p \in \mathcal{M}_i, \text{ and } f_i(q) \neq 0 \text{ for some } q \in \mathcal{M}_{1-i}, \text{ then}$$

$$\dim(\mathcal{M}_0 \cap \mathcal{M}_1) < \min(\dim \mathcal{M}_0, \dim \mathcal{M}_1).$$



Phylogenetic Models are Identifiable

Theorem

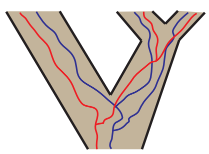
The unrooted tree parameter of phylogenetic models is generically identifiable.

Proof.

- Edge flattening invariants can detect which splits are implied by a specific distribution in $\mathcal{M}_T$.
- The splits in T uniquely determine T. □

Phylogenetic Mixture Models

- Basic phylogenetic model assume same parameters at every site
- This assumption is not accurate within a single gene
  - Some sites more important: unlikely to change
- Tree structure may vary across genes



- Leads to mixture models for different classes of sites
- $\mathcal{M}(T, r)$ denotes a same tree mixture model with underlying tree T and r classes of sites

Identifiability Questions for Mixture Models

Question

For fixed number of trees r, are the tree parameters $T_1, \ldots, T_r$, and rate parameters of each tree (generically) identified in phylogenetic mixture models?

- $r = 1$ (Ordinary phylogenetic models)
  Most models are identifiable on $\geq 2, 3, 4$ leaves. (Rogers, Chang, Steel, Hendy, Penny, Székely, Allman, Rhodes, Housworth, ...)
- $r > 1$ $T_1 = T_2 = \cdots = T_r$ but no restriction on number of trees
  Not identifiable (Matsen-Steel, Stefankovic-Vigoda)
- $r > 1$, T_i arbitrary
  Not identifiable (Mossel-Vigoda)

Theorem (Rhodes-Sullivant 2011)

The unrooted tree and numerical parameters in a r-class, same tree phylogenetic mixture model on n-leaf trivalent trees are *generically identifiable*, *if $r < 4^{\lceil n/4 \rceil}$.*

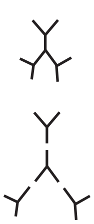
Proof Ideas.

- Phylogenetic invariants from flattenings
- Tensor rank (Kruskal's Theorem) [Allman-Matias-Rhodes 2009]
- Elementary tree combinatorics
- Solving tree and numerical parameter identifiability at the same time □

How to Construct Phylogenetic Invariants?

Theorem (Sturmfels-S, Allman-Rhodes, Casanellas-S, Draisma-Kuttler)

Consider "nice" algebraic phylogenetic model. The problem of computing phylogenetic invariants for any tree T can be reduced to the same problem for star trees $K_{1,k}$.



- The ideal I_T generated by local contributions from each $K_{1,k}$, plus flattening invariants from edges.
- The varieties $V_{K_{1,k}}$ are interesting classical algebraic varieties:
  - toric varieties
  - secant varieties
  - $Sec^4(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)$

Group-based models

$$\begin{pmatrix}\alpha & \beta \\ \beta & \alpha\end{pmatrix} \quad \begin{pmatrix}\alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha\end{pmatrix} \quad \begin{pmatrix}\alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha\end{pmatrix} \quad \begin{pmatrix}\alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha\end{pmatrix}$$

- Random variables in finite abelian group G.
- Transitions probabilities satisfy $Prob(X = g|Y = h) = f(g + h)$.
- This means that the formula for $Prob(X_1 = g_1, \ldots, X_n = g_n)$ is a convolution (over G^n).
- Apply discrete Fourier transform to turn convolution into a product.

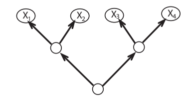
Theorem (Hendy-Penny 1993, Evans-Speed 1993)

In the Fourier coordinates, a group-based model is parametrized by monomial functions in terms of the Fourier parameters.
In particular, the CFN model is a toric variety.

Equations for the CFN Model

Theorem (Sturmfels-S 2005)

For any tree T, the toric ideal I_T for the CFN model is generated by degree 2 determinantal equations.

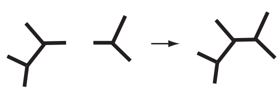


Fourier coordinates:
$q_{lmno} = \sum_{r,s,t,u \in \{0,1\}} (-1)^{rl+sm+tn+uo} p_{rstu}$
I_T generated by 2×2 minors of:

$$\begin{pmatrix} q_{0000} & q_{0001} & q_{0010} & q_{0011} \\ q_{1100} & q_{1101} & q_{1110} & q_{1111} \end{pmatrix} \quad \begin{pmatrix} q_{0000} & q_{0011} \\ q_{0100} & q_{0111} \\ q_{1000} & q_{1011} \\ q_{1100} & q_{1111} \end{pmatrix} \quad \begin{pmatrix} q_{0001} & q_{0010} \\ q_{0101} & q_{0110} \\ q_{1001} & q_{1010} \\ q_{1101} & q_{1110} \end{pmatrix}$$

$$\begin{pmatrix} q_{0100} & q_{0101} & q_{0110} & q_{0111} \\ q_{1000} & q_{1001} & q_{1010} & q_{1011} \end{pmatrix}$$

Gluing Two Trees at a Leaf



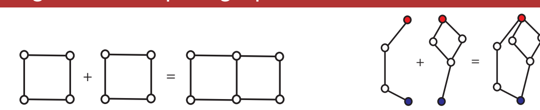
- Let $T = T_1 \# T_2$, tree obtained by joining two trees at a leaf.
- Each ring $\mathbb{C}[p]/I_{T_1}$, $\mathbb{C}[p]/I_{T_2}$ is invariant under action of group $\mathcal{G} = \mathrm{Gl}_r(\mathbb{C})^k$ acting on the glue leaves.

Theorem (Draisma-Kuttler)

- $\mathbb{C}[p]/I_T \cong (\mathbb{C}[p]/I_{T_1} \otimes_{\mathbb{C}} \mathbb{C}[p]/I_{T_2})^{\mathcal{G}}$
- $V_T = (V_{T_1} \times V_{T_2})//\mathcal{G}$ *(GIT quotient)*

- Actions of individual factors $(\mathrm{Gl}_r(\mathbb{C}))$ do no interact.
- Use Reynolds operator and first fundamental theorem of CIT.

Gluing more complex graphs



- Still a group action $(\mathrm{Gl}_r(\mathbb{C})^k)$.
- But factors are not acting independently.
- $\mathbb{C}[p]/I_G \ncong (\mathbb{C}[p]/I_{G_1} \otimes_{\mathbb{C}} \mathbb{C}[p]/I_{G_2})^{\mathcal{G}}$
- $\mathbb{C}[p]/I_G$ generated by degree 1 part of $(\mathbb{C}[p]/I_{G_1} \otimes_{\mathbb{C}} \mathbb{C}[p]/I_{G_2})^{\mathcal{G}}$ (toric fiber product if $r = 1$)

Theorem (Engström-Kahle-S)

Can determine generators of I_G from I_{G_1} and I_{G_2} if the TFP has "low codimension".

- Useful for other problems in algebraic statistics.

Summary: Phylogenetic Algebraic Geometry

- Phylogenetic models are fundamentally algebraic-geometric objects.
- Algebraic perspective is useful for:
  - Developing new construction algorithms
  - Proving theorems about identifiability (currently best available for mixture models)
- Leads to interesting new mathematics, useful for other problems
- Long way to go: Your Help Needed!

Problems

Theorem (Allman-Rhodes 2006)

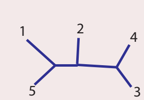
Let T be a trivalent tree with n leaves, and consider the general Markov model on binary characters. The phylogenetic ideal I_T has generating set

$$\bigcup_{A|B \in \Sigma(T)} \{3 \times 3 \text{ minors of } \mathrm{Flat}_{A|B}(P)\}$$

where $\Sigma(T)$ is the set of all valid splits on T. Note that P is a $2 \times 2 \times \cdots \times 2$, n-way tensor.

Problem

For the 5 leaf tree at the right and write down all the matrices $\mathrm{Flat}_{A|B}(P)$ that are needed in the previous theorem.



References

- E. Allman, C. Matias, J. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, **37** no.6A (2009) 3099-3132.
- M. Casanellas, J. Fernandez-Sanchez. Performance of a new invariants method on homogeneous and non-homogeneous quartet trees, *Molecular Biology and Evolution*, **24**(1):288-293, 2006.
- J. Cavender, J. Felsenstein. Invariants of phylogenies: a simple case with discrete states. *Journal of Classification* **4** (1987) 57-71.
- J. Draisma, J Kuttler. On the ideals of equivariant tree models, *Mathematische Annalen* **344**(3):619-644, 2009.
- A. Engström, T. Kahle, S. Sullivant. Multigraded commutative algebra of graph decompositions. `1102.2601`
- S. Evans and T. Speed. Invariants of some probability models used in phylogenetic inference. *Annals of Statistics* **21** (1993) 355-377.
- M. Hendy, D. Penny. Spectral analysis of phylogenetic data. *J. Classification* **10** (1993) 5–24.
- J. Huelsenbeck. Performance of phylogenetic methods in simulation. *Systematic Biology* **42** no.1 (1995) 17–48.
- J. Lake. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution* **4** (1987) 167-191.

- F.A. Matsen and M. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 2007.
- E. Mossel and E. Vigoda Phylogenetic MCMC Are Misleading on Mixtures of Trees. *Science* **309**, 2207–2209 (2005)
- J. Rhodes, S. Sullivant. Identifiability of large phylogenetic mixture models. To appear *Bulletin of Mathematical Biology*, 2011. `1011.4134`
- D. Stefankovic and E. Vigoda. Pitfalls of Heterogeneous Processes for Phylogenetic Reconstruction *Systematic Biology* **56**(1): 113-124, 2007.
- B. Sturmfels, S. Sullivant. Toric ideals of phylogenetic invariants. *Journal of Computational Biology* **12** (2005) 204-228.

Algebraic Statistics Tutorial II

Seth Sullivant

North Carolina State University

June 10, 2012

Generating Random Tables

Problem

Generate a random table from the set of all nonnegative $k_1 \times k_2$ integer tables with given row and column sums.

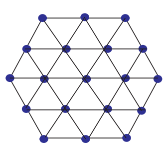
				r_1
				r_2
				r_3
c_1	c_2	c_3	c_4	

Fisher's Exact Test, Missing Data Problems

Random Walk

$$\begin{array}{ccc|c} 2 & 2 & 2 & 6 \\ 2 & 2 & 2 & 6 \\ \hline 4 & 4 & 4 & \end{array} \;+\; \begin{array}{ccc|c} 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & \end{array} \;=\; \begin{array}{ccc|c} 3 & 2 & 1 & 6 \\ 1 & 2 & 3 & 6 \\ \hline 4 & 4 & 4 & \end{array}$$

$$\begin{array}{ccc|c} 3 & 2 & 1 & 6 \\ 1 & 2 & 3 & 6 \\ \hline 4 & 4 & 4 & \end{array} \;+\; \begin{array}{ccc|c} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & \end{array} \;=\; \begin{array}{ccc|c} 4 & 1 & 1 & 6 \\ 0 & 3 & 3 & 6 \\ \hline 4 & 4 & 4 & \end{array}$$



$$\left\{ \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \right\}$$

allow for a connected random walk over these contingency tables.

Connecting Lattice Points in Polytopes

Definition

- Let $A : \mathbb{Z}^n \to \mathbb{Z}^d$ a linear transformation, $b \in \mathbb{Z}^d$.
- $A^{-1}[b] := \{x \in \mathbb{N}^n : Ax = b\}$ (fiber)
- $\mathcal{B} \subset \ker_{\mathbb{Z}} A$

Let $A^{-1}[b]_{\mathcal{B}}$ be the graph with vertex set $A^{-1}[b]$ and $u - -v$ an edge if and only $u - v \in \pm\mathcal{B}$.

Problem

Given A and b, find finite $\mathcal{B} \subseteq \ker_{\mathbb{Z}} A$ such that $A^{-1}[b]_{\mathcal{B}}$ is connected.

Definition

If $\mathcal{B} \subseteq \ker_{\mathbb{Z}} A$ is a set such that $A^{-1}[b]_{\mathcal{B}}$ is connected for all b, then $\mathcal{B}$ is a Markov basis for A.

Example: 2-way tables

Let $A : \mathbb{Z}^{k_1 \times k_2} \to \mathbb{Z}^{k_1 + k_2}$ such that

$$\begin{aligned} A(u) &= \left(\sum_{j=1}^{m} u_{1j}, \ldots, \sum_{j=1}^{m} u_{k_1 j}; \sum_{i=1}^{k} u_{i1}, \ldots, \sum_{i=1}^{k} u_{ik_2} \right) \\ &= \text{vector of row and column sums of } u \end{aligned}$$

$\ker_{\mathbb{Z}}(A) = \{u \in \mathbb{Z}^{k_1 \times k_2} :$ row and columns sums of u are $0\}$
Markov basis consists of the $2\binom{k_1}{2}\binom{k_2}{2}$ moves like:

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \end{pmatrix}$$

3-way tables

Let $A : \mathbb{Z}^{k_1 \times k_2 \times k_3} \to \mathbb{Z}^{k_1 \times k_2 + k_1 \times k_3 + k_2 \times k_3}$ be the linear transformation such that

$$\begin{aligned} A(u) &= \left((\sum_{i_3} u_{i_1 i_2 i_3})_{i_1, i_2}; (\sum_{i_2} u_{i_1 i_2 i_3})_{i_1 i_3}; (\sum_{i_1} u_{i_1 i_2 i_3})_{i_2, i_3} \right) \\ &= \text{all 2-way margins of 3-way table } u \\ &= \text{all “line sums” of } u. \end{aligned}$$

Markov basis depends on k_1, k_2, k_3, contains moves like:

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

but also non-obvious moves like:

$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

Fundamental Theorem of Markov Bases

Definition

Let $A : \mathbb{Z}^n \to \mathbb{Z}^d$. The toric ideal I_A is the ideal

$$\langle p^u - p^v : u, v \in \mathbb{N}^n, Au = Av \rangle \subset \mathbb{K}[p_1, \ldots, p_n],$$

where $p^u = p_1^{u_1} p_2^{u_2} \cdots p_n^{u_n}$.

Theorem (Diaconis-Sturmfels 1998)

The set of moves $\mathcal{B} \subseteq \ker_{\mathbb{Z}} A$ is a Markov basis for A if and only if the set of binomials $\{p^{b^+} - p^{b^-} : b \in \mathcal{B}\}$ generates I_A.

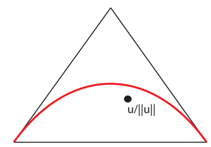
$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \end{pmatrix} \longrightarrow p_{21}p_{33} - p_{23}p_{31}$$

Toric Varieties = Log-linear Models

Definition

The variety $V_A = V(I_A)$ is a toric variety. The statistical model $\mathcal{M}_A = V(I_A) \cap \Delta_m$ is a log-linear model.

- $\mathcal{M}_A = \{p \in \Delta_m : \log p \in \text{rowspan } A\}$.
- Fisher's exact test: Does the data **u** fit the model $\mathcal{M}_A$?



2-way tables: Independence

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \end{pmatrix} \longrightarrow p_{21}p_{33} - p_{23}p_{31} = \begin{vmatrix} p_{21} & p_{23} \\ p_{31} & p_{33} \end{vmatrix}$$

$$I_A = \langle 2 \times 2 \text{ minors of } \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k_2} \\ p_{21} & p_{22} & \cdots & p_{2k_2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k_1 1} & p_{k_1 2} & \cdots & p_{k_1 k_2} \end{pmatrix} \rangle$$

$$V_A = V(I_A) = \{P \in \mathbb{R}^{k_1 \times k_2} : \text{rank } P \leq 1\}$$

$$\mathcal{M}_A = V_A \cap \Delta_{k_1 k_2} = \mathcal{M}_{X_1 \perp\!\!\!\perp X_2}$$

Computing Markov Bases

- Software
  - 4ti2 www.4ti2.de
  - Macaulay2 (4ti2 interface) http://www.math.uiuc.edu/Macaulay2/
  - Singular (toric package) http://www.singular.uni-kl.de/
- Theory
  - Gluing Results
  - Finiteness Theorems
  - Special Configurations

"No Hope" Theorem

Theorem (De Loera-Onn (2006))

- *Every integer vector appears as part of a minimal Markov basis element for $3 \times k_2 \times k_3$ tables (with fixed 2-way margins).*
- *In particular, minimal Markov basis elements for 3-way tables can have arbitrarily large entries and arbitrarily large 1-norm.*

Example ($3 \times 4 \times 6$-tables)

- For $3 \times 4 \times 6$ tables, minimal Markov basis has 355950 elements.
- Largest element has 1-norm 28.

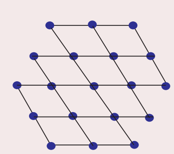
Which Fibers are Connected?

Problem

Let $\mathcal{B} \subseteq \ker_{\mathbb{Z}} A$. For which b is $A^{-1}[b]_{\mathcal{B}}$ connected? When do $u, v \in A^{-1}[b]$ belong to the same component of $A^{-1}[b]_{\mathcal{B}}$?

Example (2×3)

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \right\}$$



			6
			6
4	4	4	

Enter Commutative Algebra

Let $\mathbb{K}[p] := \mathbb{K}[p_1, \ldots, p_n]$. To each $m \in \mathcal{B}$ associate a binomial

$$p^{m^+} - p^{m^-} \in \mathbb{K}[p]$$

where $m = m^+ - m^-$, $p^m = p_1^{m_1} \cdots p_n^{m_n}$.

Proposition

Let $\mathcal{B} \subseteq \ker_{\mathbb{Z}} A$. Then $u, v \in A^{-1}[b]$ are in the same component of $A^{-1}[b]_{\mathcal{B}}$ if and only if

$$p^u - p^v \in I_{\mathcal{B}} := \langle p^{m^+} - p^{m^-} : m \in \mathcal{B} \rangle.$$

Theorem (Diaconis-Sturmfels (1998))

A set of moves $\mathcal{B} \subseteq \ker_{\mathbb{Z}} A$ is a Markov basis if and only if

$$I_{\mathcal{B}} = I_A := \langle p^u - p^v : u, v \in \mathbb{N}^n, Au = Av \rangle.$$

Lattice Walks and Primary Decomposition (Diaconis-Eisenbud-Sturmfels 1998)

- Decompose ideal $I_{\mathcal{B}} = \cap_i I_i$.
- $p^u - p^v \in I_{\mathcal{B}} \Leftrightarrow p^u - p^v \in I_i$ for all i.
- Hope that ideal I_i are easier to analyze.

Theorem (Eisenbud-Sturmfels 1996)

Every binomial ideal has a binomial primary decomposition.

- Dickenstein-Matusevich-Miller, Kahle-Miller (Mesoprimary decomposition)
- Algorithms implemented in `binomials.m2` (Kahle 2010)

2×3 tables

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \right\}$$

$$\begin{aligned} I_{\mathcal{B}} &= \left\langle \begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}, \begin{vmatrix} p_{12} & p_{13} \\ p_{22} & p_{23} \end{vmatrix} \right\rangle \\ &= \left\langle \begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}, \begin{vmatrix} p_{12} & p_{13} \\ p_{22} & p_{23} \end{vmatrix}, \begin{vmatrix} p_{11} & p_{13} \\ p_{21} & p_{23} \end{vmatrix} \right\rangle \cap \langle p_{21}, p_{22} \rangle \\ &= I_A \cap \langle p_{21}, p_{22} \rangle \end{aligned}$$

$\begin{pmatrix} u_{11} & u_{12} & u_{13} \\ u_{21} & u_{22} & u_{23} \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & v_{13} \\ v_{21} & v_{22} & v_{23} \end{pmatrix}$ connected by $\mathcal{B}$ if and only if

- they have the same row and column sums and
- $u_{12} + u_{22} = v_{12} + v_{22} > 0$.

Graphical Models

- G a graph, N-vertices.
- $d \in \mathbb{Z}^N$, $d_i \geq 2$.
- Gives set of margins of $d_1 \times d_2 \times \cdots \times d_n$ array.
- $\mathcal{C}(G)$ = set of maximal cliques in G.

Definition

Let

$$A_{G,d} : \mathbb{Z}^{d_1 \times \cdots \times d_n} \to \mathbb{Z}^k$$

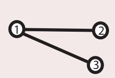
be the linear map that computes the margins associated to all $C \in \mathcal{C}(G)$, of a $d_1 \times \cdots \times d_n$ array.

Example (Row and Column Sums)

$$A_{G,d} : \mathbb{Z}^{d_1 \times d_2} \to \mathbb{Z}^{d_1 + d_2}$$

$$(u_{ij})_{i,j} \mapsto ((\sum_j u_{ij})_i, (\sum_i u_{ij})_j)$$

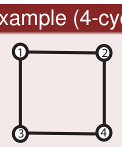
Example (Path)



$$A_{G,d} : \mathbb{Z}^{d_1 \times d_2 \times d_3} \to \mathbb{Z}^{d_1 \times d_2 + d_1 \times d_3}$$

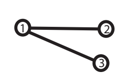
$$(u_{ijk})_{i,j,k} \mapsto ((\sum_k u_{ijk})_{i,j}, (\sum_j u_{ijk})_{i,k})$$

Example (4-cycle)



$$A_{G,d} : \mathbb{Z}^{d_1 \times d_2 \times d_3 \times d_4} \to \mathbb{Z}^{d_1 \times d_2 + d_1 \times d_3 + d_2 \times d_4 + d_3 \times d_4}$$

$$\mathcal{C}(G) = \{\{1,2\}, \{1,3\}, \{2,4\}, \{3,4\}\}$$



$$A_{G,d} : \mathbb{Z}^{d_1 \times d_2 \times d_3} \to \mathbb{Z}^{d_1 \times d_2 + d_1 \times d_3}$$

$$(u_{ijk})_{i,j,k} \mapsto ((\sum_k u_{ijk})_{i,j}, (\sum_j u_{ijk})_{i,k})$$

$$d = (2, 2, 3)$$

$$A_{G,d} = \left(\begin{array}{cccccccccccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{array} \right)$$

$$u = (u_{111}, u_{112}, u_{113}, u_{121}, u_{122}, u_{123}, u_{211}, u_{212}, u_{213}, u_{221}, u_{222}, u_{223})$$

Separating Moves (Conditional Independence)

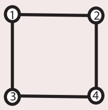
- Let A, B, C partition $V(G)$ such that C separates A and B in G.
- Get moves

$$e_{i_A i_B i_C} + e_{j_A j_B i_C} - e_{i_A j_B i_C} - e_{j_A i_B i_C}$$

where $i_A, j_A \in \prod_{t \in A}[d_t]$, $i_B, j_B \in \prod_{t \in B}[d_t]$, $i_C \in \prod_{t \in C}[d_t]$ in $\ker_{\mathbb{Z}} A_{G,d}$.

- These moves naturally generalize $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ for 2-way tables.
- $CI(G)$ is set of all separating moves.

Example (4-cycle)



$$e_{i_1 i_2 i_3 i_4} + e_{j_1 i_2 i_3 j_4} - e_{i_1 i_2 i_3 j_4} - e_{j_1 i_2 i_3 i_4}$$

$$e_{i_1 i_2 i_3 i_4} + e_{i_1 j_2 j_3 i_4} - e_{i_1 i_2 j_3 i_4} - e_{i_1 j_2 i_3 i_4}$$

Which Fibers Do $CI(G)$ Moves Connect?

Proposition (Hammersley-Clifford, Besag (1974))

$CI(G)$ spans $\ker_{\mathbb{Z}} A_{G,d}$ for all G.

Theorem (Dobra (2002), Geiger, Meek, Sturmfels (2006))

Separating moves $CI(G)$ are a Markov basis for $A_{G,d}$ if and only if G is a chordal graph.

Problem

1. Which fibers $A_{G,d}^{-1}[b]$ are connected by $CI(G)$ for other graphs?
2. What is the primary decomposition of $I_{CI(G)}$?

Computational Results

Theorem (Kahle-Rauh-S (2012))

Let $\#V(G) = n \leq 5$, $d_i = 2$ for all i. Then

- *$I_{CI(G)}$ is radical.*
- *$A_{G,d}^{-1}[b]_{CI(G)}$ is connected if b is in the interior of the marginal cone.*
- *$A_{G,d}^{-1}[b]_{CI(G)}$ is connected if b is positive (except for $G = K_{2,3}$).*

- Every prime component $I_{\mathcal{B}}$ of the form $P_S = \langle p_i : i \in S \rangle + I_{A_S}$.
- Form vector $u_{\bar{S}} := \sum_{i \notin S} e_i$.
- Check if $Au_{\bar{S}}$ is on boundary of marginal cone for all prime components.
- If so $\mathcal{B}$ has interior point property.

2×3 tables

$$\mathcal{B} = \left\{ \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix} \right\}$$

$$\begin{aligned} I_{\mathcal{B}} &= \left\langle \begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}, \begin{vmatrix} p_{12} & p_{13} \\ p_{22} & p_{23} \end{vmatrix} \right\rangle \\ &= I_A \cap \langle p_{21}, p_{22} \rangle \end{aligned}$$

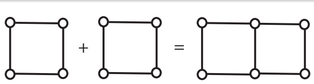
- Analyze monomial ideal $P_S = \langle p_{21}, p_{22} \rangle$
- $u_{\bar{S}} = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}$
- $u_{\bar{S}}$ has a zero column sum
- $\Rightarrow$ all fibers with positive margins (row and column sums) are connected.

Theoretical Results

Proposition (Kahle-Rauh-S (2012))

If $G = G_1 \# G_2$ is a clique sum, then

- *If $I_{CI(G_1)}$ and $I_{CI(G_2)}$ radical, so is $I_{CI(G)}$.*
- *If G_1 and G_2 satisfy interior point property, so does G.*
- *If G_1 and G_2 satisfy positive margins property, so does G.*



Theorem (Kahle-Rauh-S (2012))

1. *For cycle C_n, $I_{CI(C_n)}$ is radical, when $d_i = 2$ for all i.*
2. *For $K_{2,n}$ with $d_1 = d_2 = 2$, $I_{CI(K_{2,n})}$ is radical.*
3. *Interior point property holds in both situations.*

Proof Ideas

- Find minimal primes for $I_{CI(G)}$. All binomial ideals.
- Let $J = \sqrt{I_{CI(G)}} = I_{A_{G,d}} \cap \cap_{i=1}^{k} P_i$.
- Let u, v such that $A_{G,d}u = A_{G,d}v$, so $p^u - p^v \in I_A$.
- Connect u and v using Markov basis moves of $A_{G,d}$.
- Show that $p^u - p^v \in P_i$ for all i, implies we can shortcut moves with $CI(G)$ moves.
- Deduce that $J = I_{CI(G)}$.
- Depends on having Markov basis of $A_{G,d}$, which is obtained in these cases via toric fiber product. (Engström, Kahle, S 2011)

Questions

Question

- Is $I_{CI(G)}$ radical for all G, d?
- Does interior point property hold for all G, d?

Theorem

If there are $n-2$ mutually orthogonal $d' \times d'$ latin squares, then for any 2-connected, triangle free graph on G nodes, and $d_i = d'$ for all i, the interior point property does not hold for (G, d).

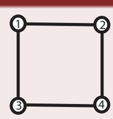
- For C_4 and $d = (3,3,3,3)$ gives failure of interior point property.
- Radicality fails for $K_{3,3}$ and $d = (2,2,2,2,2,2)$.

Summary

- Many statistical problems require the construction of random walks over the lattice points in a polytope.
- A Markov basis provides connectivity for all b.
- If Markov basis too hard to compute, can ask: Which fibers are connected by a "natural" set of moves?
- Binomial primary decomposition gives information about connectivity of fibers with subset of Markov basis.
- Computational and theoretical advances allow us to make progress on graphical models.

Problems

Problem



1. Let $d = (2,2,2,2)$. Construct the 16×16 matrix $A_{C_4,d}$.
2. List the elements of $CI(C_4)$
3. Use 4ti2, Macaulay2, or Singular to compute the Markov basis of C_4.

References

- J. Besag. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society*, Series B, 36 (2), 192Ð236.
- J. De Loera, S. Onn. Markov bases of three-way tables are arbitrarily complicated. *Journal of Symbolic Computation*, **41**:173–181, 2006.
- P. Diaconis, D. Eisenbud, B. Sturmfels. Lattice walks and primary decomposition, Mathematical Essays in Honor of Gian-Carlo Rota, eds. B. Sagan and R. Stanley, Progress in Mathematics, Vol. 161, Birkhauser, Boston, 1998, pp. 173-193.
- P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions, *Annals of Statistics* **26** (1998) 363-397
- A. Dickenstein, L. Matusevich, E. Miller. Combinatorics of binomial primary decomposition. `0803.3846`
- A. Dobra. Markov bases for decomposable graphical models. *Bernoulli* **9** No. 6,(2003) 1-16.
- D. Eisenbud, B. Sturmfels. Binomial ideals, *Duke Mathematical Journal* **84** (1996) 1-45.
- A. Engström, T. Kahle, S. Sullivant. Multigraded commutative algebra of graph decompositions. (2011) `1102.2601`
- D. Geiger, C. Meek, B. Sturmfels. On the toric algebra of graphical models, *Annals of Statistics* **34** (2006) 1463-1492
- T. Kahle, E. Miller. Decompositions of commutative monoid congruences and binomial ideals. `arxiv:1107.4699`
- T. Kahle, J. Rauh, S. Sullivant. Positive margins and primary decomposition. (2012) `1201.2591`