Example: Hardy-Weinberg Equilibrium

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

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

The model of Hardy-Weinberg equilibrium is the set

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

$$
I(M) = p_{aa} + p_{aA} + p_{AA} - 1, p_{aa}^2 - 4p_{aa}p_{Aa}
$$

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: ...ACGTGCAAGTGAACGA...
- Chimp: ...ACCTGGAAGGTAAACGA...
- 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: ACCTGCAAGTGAACGA
Chimp: ACCTGCAAGTGAACGA
Gorilla: ACCTGCAACGTAAACTA

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 $X_v \in \{A, C, G, T\}$
  - All random variables corresponding to internal nodes are latent

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

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

\[ p_{ij} = \sum_h \tau_h a_{ij} b_{ihj} c_{hl} d_{lk} \]

Algebraic Perspective on Phylogenetic Models

- Once we fix a tree \( T \) and model structure, we get a map \( \phi^T : \Theta \to \mathbb{R}^{d_T} \).
- \( \Theta \subseteq \mathbb{R}^D \) is a parameter space of numerical parameters (transition matrices associated to each edge).
- The map \( \phi^T \) is given by polynomial functions of the parameters.
- For each \( i_1 \cdots i_n \in \{A, C, G, T\}^n \), \( \phi^T(p) \) gives the probability of the column \( (i_1, \ldots, i_n)^T \) in the alignment for the particular parameter choice \( \theta \).

\[ \phi^T(p, \theta) = \sum_{j=1}^{T} \sum_{k=1}^{R} \tau_j a_{ij} b_{jk} c_{kl} d_{lk} \]

- The phylogenetic model is the set \( M_T = \phi^T(\Theta) \subseteq \mathbb{R}^{d_T} \).

Phylogenetic Varieties and Phylogenetic Invariants

- Let \( R^p := R[p_{ij}] : i, j \in \{A, C, G, T\} \)

**Definition**

Let

\[ I_T := \{ f \in R^p : f(p) = 0 \text{ for all } p \in M_T \} \subseteq R^p \]

\( I_T \) is the ideal of phylogenetic invariants of \( T \).

Let

\[ V_T := \{ p \in R^{4n} : f(p) = 0 \text{ for all } f \in I_T \} \]

\( V_T \) is the phylogenetic variety of \( T \).

- Note that \( M_T \subseteq V_T \).
- Since \( M_T \) is image of a polynomial map \( \dim M_T = \dim V_T \).

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
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 $f \notin I_{T'}$.

  Evaluate phylogenetically informative phylogenetic invariants at empirical distribution $\hat{p}$ to reconstruct phylogenetic trees

Proposition
For each $n$-leaf trivalent tree $T$, let $F_T \subseteq I_T$ be a set of phylogenetic invariants such that, for each $T' \neq T$, there is an $f \in F_T$ such that $f \notin I_{T'}$.

Let $f_T := \sum_{f \in F_T} |f|$.

Then for generic $p \in \cup M_T$, $f_T(p) = 0$ if and only if $p \in M_T$.

Performance of Invariants Methods in Simulations

- 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.

$F_T := \sum_{f \in F_T} f$

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 $p \in M_T$

there does not exist another $T' \neq T$ such that $p \in M_{T'}$.

Generic Identifiability

Definition
The tree parameter in a phylogenetic model is generically identifiable if for all $n$-leaf trees with $T \neq T'$,

$$\dim(M_T \cap M_{T'}) < \min(\dim(M_T), \dim(M_{T'})).$$
**Proposition**

Let $M_0$ and $M_1$ be two algebraic models. If there exist phylogenetically informative invariants $f_0$ and $f_1$ such that

$$f_i(p) = 0 \text{ for all } p \in M_i, \text{ and } f_i(q) \neq 0 \text{ for some } q \in M_1 - M_i,$$

then

$$\dim(M_0 \cap M_1) < \min(\dim M_0, \dim M_1).$$

**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 $M_T$.
- 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
- $\lambda(T, r)$ 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, 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 $r < 4^{n/4}$.

**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

**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
  - $Sec(\mathbb{P}^3 \times \mathbb{P}^3)$
Group-based models

\[
\begin{pmatrix}
\alpha & \beta \\
\beta & \alpha
\end{pmatrix} \begin{pmatrix}
\alpha & \beta & \beta \\
\beta & \alpha & \beta
\end{pmatrix} \begin{pmatrix}
\alpha & \beta & \gamma \\
\beta & \alpha & \beta
\end{pmatrix} \begin{pmatrix}
\alpha & \beta & \gamma & \delta \\
\beta & \alpha & \delta & \gamma
\end{pmatrix}
\]

- Random variables in finite abelian group \( G \).
- Transitions probabilities satisfy \( \text{Prob}(X = g | Y = h) = f(g + h) \).
- This means that the formula for \( \text{Prob}(X_i = g_1, \ldots, X_n = g_n) \) 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:**

\[
q_{0000} = \sum_{r,s,t,u} \prod_{i=1}^n (1 - (-1)^{r+s+t+u}) \text{Prob}(X_{2i-1} = r, X_{2i} = s | \text{root})
\]

\( I_T \) generated by 2 x 2 minors of:

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

---

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

\[
\bigcup_{A,B \in \Sigma(T)} \{3 \times 3 \text{ minors of } \text{Flat}_{A,B}(P)\}
\]

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 \( \text{Flat}_{A,B}(P) \) that are needed in the previous theorem.
References


A. Engström, T. Kahle, S. Sullivant. Multigraded commutative algebras of graph decompositions. 1102.2601


Seth Sullivant (NCSU) Algebraic Statistics July 22, 2012 31 / 32

Seth Sullivant (NCSU) Algebraic Statistics July 22, 2012 32 / 32
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.

Fisher’s Exact Test, Missing Data Problems

Connecting Lattice Points in Polytopes

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

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

Problem
Given $A$ and $b$, find finite $B \subseteq \ker_k A$ such that $A^{-1}[b]_B$ is connected.

Definition
If $B \subseteq \ker_k A$ is a set such that $A^{-1}[b]_B$ is connected for all $b$, then $B$ is a Markov basis for $A$.

Example: 2-way tables

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

$A(u) = \left( \sum_{j=1}^{k_1} u_{ij} \right)_{i=1}^{k_1}$

vector of row and column sums of $u$

$\ker(A) = \{ u \in \mathbb{Z}^{k_1 \times k_2} : \text{row and columns sums of } u \text{ are 0} \}$

Markov basis consists of the $2\binom{k}{2}$ moves like:

\[
\begin{pmatrix}
0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 \\
-1 & 0 & 1 & 0
\end{pmatrix}
\]

3-way tables

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

$A(u) = \left( \sum_{i=1}^{k_1} u_{ijk} \right)_{i=1}^{k_1}$

all 2-way margins of 3-way table $u$

$= \text{all “line sums” of } u$

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

\[
\begin{pmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{pmatrix}
\]

but also non-obvious moves like:

\[
\begin{pmatrix}
1 & -1 & 1 \\
1 & 1 & -1 \\
1 & 1 & 1
\end{pmatrix}
\]
**Fundamental Theorem of Markov Bases**

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

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

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

**Theorem (Diaconis-Sturmfels 1998)**
The set of moves $B \subseteq \ker Z_A$ is a Markov basis for $A$ if and only if the set of binomials

$$\left\langle p^b - p^{\overline{b}} : b \in B \right\rangle$$

generates $I_A$.

**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 $\mathcal{M}_A$?

**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

**Which Fibers are Connected?**

**Problem**
Let $B \subseteq \ker Z_A$. For which $b$ is $A^{-1}[b]_0$ connected? When do $u, v \in A^{-1}[b]$ belong to the same component of $A^{-1}[b]_0$?

**Example (2 × 3)**

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

**“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 × 4 × 6-tables)**
- For $3 \times 4 \times 6$ tables, minimal Markov basis has 355950 elements.
- Largest element has 1-norm 28.
Enter Commutative Algebra

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

$$p^m - p^m' \in K[p]$$

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

Proposition

Let $B \subseteq \ker A$. Then $u, v \in A^{-1}[B]$ are in the same component of $A^{-1}[B]$ if and only if

$$p^u - p^v \in I_B := (p^m - p^m' : m \in B).$$

Theorem (Diaconis-Sturmfels (1998))

A set of moves $B \subseteq \ker A$ is a Markov basis if and only if

$$I_B = I_A := (p^u - p^v : u, v \in \mathbb{N}^n, Au = Av).$$

2 $\times$ 3 tables

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

$$I_B = \left\langle \begin{pmatrix} p_{12} & p_{13} \\ p_{21} & p_{22} \end{pmatrix} \right\rangle = \left\langle \begin{pmatrix} p_{12} & p_{13} \\ p_{21} & p_{22} \end{pmatrix} \right\rangle \cap \left\langle \begin{pmatrix} p_{12} & p_{22} \end{pmatrix} \right\rangle$$

Let $u_1, u_2, u_3, v_1, v_2, v_3$, connected by $B$ if and only if

- they have the same row and column sums and
- $u_1 + u_2 = v_1 + v_2 > 0$.

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

- Decompose ideal $I_B = \bigcap I_i$.
- $p^u - p^v \in I_B \Rightarrow p^u - p^v \in I_i$ 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-Matushevich-Miller, Kahle-Miller (Mesoprimary decomposition)
- Algorithms implemented in binomials.m2 (Kahle 2010)

Graphical Models

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

Definition

Let

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

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

Example (Row and Column Sums)

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

$$u_{ij} \mapsto (\sum_k u_{ik} ; \sum_j u_{jk})(u_{ij})$$

Example (Path)

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

$$u_{ij} \mapsto (\sum_k u_{ik} ; \sum_j u_{jk})(u_{ij})$$

Example (4-cycle)

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

$\mathcal{C}(G) = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}$
Separating Moves (Conditional Independence)

- Let A, B, C partition V(G) such that C separates A and B in G.
- Get moves $e_{u,v} + e_{u',v'} - e_{u,v'} - e_{u',v}$
  where $u,v \in E[G(d)]$, $u,v' \in E[G(d)]$, $u' \in E[G,\bar{A}]$ and $v' \in E[G,\bar{A}]$.
- These moves naturally generalize $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ for 2-way tables.
- $\text{CI}(G)$ is set of all separating moves.

Example (4-cycle)

\[
\begin{align*}
e_{a,b} + e_{a',b'} - e_{a',b} - e_{a,b'} \\
e_{a,b} + e_{a',b} - e_{a,b'} - e_{a',b'}
\end{align*}
\]

Which Fibers Do CI(G) Moves Connect?

Proposition (Hammersley-Clifford, Besag (1974))

$\text{CI}(G)$ spans $\ker\mathbb{Z}_2 A_{G,J}$ for all G.

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

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

Problem

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

Computational Results

Theorem (Kahle-Rauh-S (2012))

Let $\#V(G) = n \leq 5$, $d_i = 2$ for all $i$. Then
- $I_{\text{CI}(G)}$ is radical.
- $A_{G,J}^{-1}[b]_{\text{CI}(G)}$ is connected if $b$ is in the interior of the marginal cone.
- $A_{G,J}^{-1}[b]_{\text{CI}(G)}$ is connected if $b$ is positive (except for $G = K_{2,3}$).
- Every prime component $I_p$ of the form $P_S = \langle p_i : i \in S \rangle + I_{\mathbb{Z}^n}$.
  - Form vector $u_p^n := \sum_{i\in S} p_i$.
  - Check if $A_{G,J}$ spans $I_p$ on boundary of marginal cone for all prime components.
  - If so $B$ has interior point property.

2 x 3 tables

\[
B = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & -1 \\ 0 & -1 & 1 \end{pmatrix}
\]

\[
l_B = \left\langle \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \end{pmatrix} \right\rangle = l_A \cap (p_{21}, p_{22})
\]

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

Proof Ideas

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

Question

- Is \( I_{\mathcal{C}(G)} \) 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.
- Radically fails for \( K_{3,3} \) and \( d = (2, 2, 2, 2, 2) \).

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 \( 16 \times 16 \) matrix \( A_{C_4, d} \).
2. List the elements of \( I_{\mathcal{C}(C_4)} \).
3. Use 4ti2, Macaulay2, or Singular to compute the Markov basis of \( C_4 \).

References

- A. Dicken, S. Eisenbud, B. Sturmfels. Combinatorics of binomial primary decomposition. 0803.3846
- T. Kahle, E. Miller. Decompositions of binomial and monomial congruences and binomial ideals. arxiv:1107.4408