Combinatorial problems on trees and graphical models

Pratik Misra

North Carolina State University

pmisra@ncsu.edu

Preliminary Exam
October 2nd, 2019

Outline

- Bounds on the expected size of the maximum agreement subtree for a given tree shape
  - Preliminaries
  - Lower bound
  - Upper bound
  - Based on "Bounds on the expected size of the maximum agreement subtree for a given tree shape" (`arXiv 1809.04488`, 2019)
- Vanishing ideal of a Gaussian Graphical Model
  - Conjecture of Sturmfels and Uhler
  - Results obtained so far

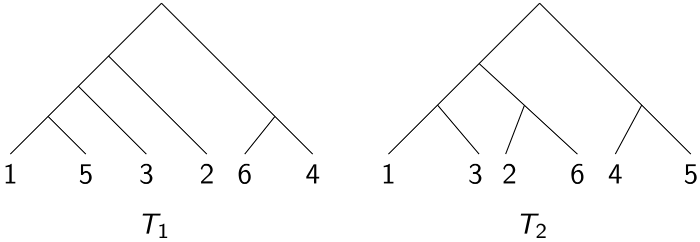
Preliminaries

Definition

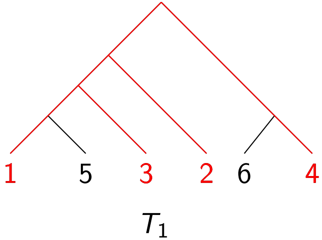
Let T be a rooted binary tree with n leaves, leaf labeled by $[n]$.

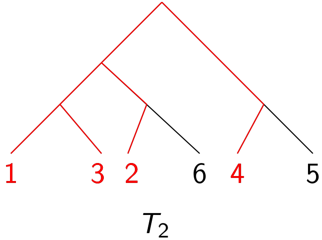
- For any $S \subseteq [n]$, the *binary restriction tree* $T|_S$ is the subtree of T obtained after deleting all the leaves that are not in S and suppressing the internal nodes of degree 2.
- The subtree $T|_S$ is rooted at the most recent common ancestor of S.
- If T_1 and T_2 are two trees leaf labeled by X, then a subset $S \subseteq X$ is an *agreement set* of T_1 and T_2 if $T_1|_S = T_2|_S$.
- A *maximum agreement subtree* is a subtree obtained from an agreement set of T_1 and T_2 of maximal size.
- $\mathrm{MAST}(T_1, T_2)$ denotes the number of leaves in a maximum agreement subtree of T_1 and T_2

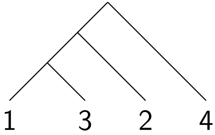
Preliminaries



Preliminaries







Maximum agreement subtree of T_1 and T_2

Main Theorem

Theorem (M.-,Sullivant)

Let T_1 and T_2 be two trees generated from the uniform distribution on rooted binary trees with n leaves with same tree shape (that is, T_2 is a random leaf relabeling of T_1). Then

$$E[\mathrm{MAST}(T_1, T_2)] = \Theta(\sqrt{n}).$$

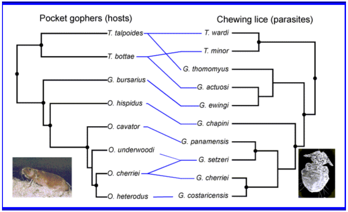
- Lower Bound
  - Divide the trees into *blobs*
  - Blobs help us in constructing an agreement subtree between the two trees
- Upper Bound
  - Generalize a previously known result for random tree distributions that are exchangeable but not necessarily sampling consistent

Motivation

- Rooted binary trees are used in evolutionary biology to represent the evolution of a set of species.
- The leaves denote the existing species and the internal nodes denote the unknown ancestors.
- Different tree reconstruction methods, and different datasets on the same set of species, can lead to the reconstruction of different trees.
- Important to measure the distance between different trees constructed
- The *maximum agreement subtree* is one of the measures of discrepancy between trees.

Motivation: Cospeciation

- Let T_H be a phylogenetic tree of host species, and T_P a phylogenetic tree of parasite species.
- Host and parasites are paired, so T_H and T_P have same label set.
- If $\mathrm{MAST}(T_H, T_P)$ is large, reject hypothesis that T_H and T_P evolved independently. i.e, large$\mathrm{MAST}(T_H, T_P) \implies$ cospeciation.
- Need distribution of $\mathrm{MAST}(T_1, T_2)$ for random trees under null hypothesis of independence to perform hypothesis test.



Hafner, M.S., Nadler, S.A. (1988) Nature 332: 258-259

Previously known results

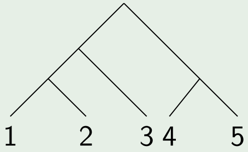
- Martin and Thatte [4] conjectured that if T_1 and T_2 are balanced rooted binary trees with n leaves, then $\mathrm{MAST}(T_1, T_2) \geq \sqrt{n}$.
- Simulations by Bryant, McKenzie, and Steel [2] suggest that under the uniform distribution on the rooted binary trees with n leaves, the expected size of $\mathrm{MAST}(T_1, T_2)$ is of the order $\Theta(n^a)$ with $a \approx 1/2$.
- The main result in this section provides evidence for Martin and Thatte's conjecture.

Lower Bound - Blobs

Definition

- A *cherry blob* is a set of leaves in T consisting of all leaves below a vertex in the tree.
- An *edge blob* is a nonempty set of leaves of the form $C_1 \setminus C_2$ where C_1 and C_2 are two nonempty cherry blobs.
- A *blob* in T is either a cherry blob or an edge blob.

Example

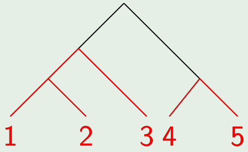


Lower Bound - Blobs

Definition

- A *cherry blob* is a set of leaves in T consisting of all leaves below a vertex in the tree.
- An *edge blob* is a nonempty set of leaves of the form $C_1 \setminus C_2$ where C_1 and C_2 are two nonempty cherry blobs.
- A *blob* in T is either a cherry blob or an edge blob.

Example



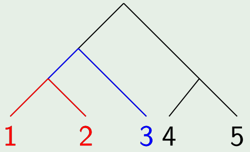
Cherry blobs: $\{1, 2\}, \{1, 2, 3\}, \{4, 5\}$

Lower Bound - Blobs

Definition

- A *cherry blob* is a set of leaves in T consisting of all leaves below a vertex in the tree.
- An *edge blob* is a nonempty set of leaves of the form $C_1 \setminus C_2$ where C_1 and C_2 are two nonempty cherry blobs.
- A *blob* in T is either a cherry blob or an edge blob.

Example



Cherry blobs: $\{1, 2\}, \{1, 2, 3\}, \{4, 5\}$
Edge blobs: $\{3\}$

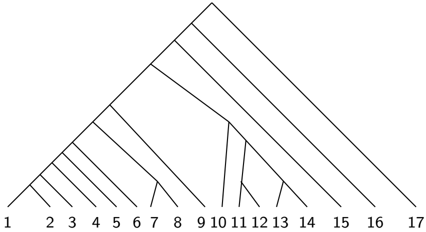
k-Blobification

- Given an integer k and a tree T, a *k-blobification* of T is a collection $\mathcal{B}$ of blobs of T such that,
  - for all distinct blobs $B_1, B_2 \in \mathcal{B}$, $B_1 \cap B_2 = \emptyset$,
  - and for all $B \in \mathcal{B}$, $k \leq |B| \leq 2k - 2$.

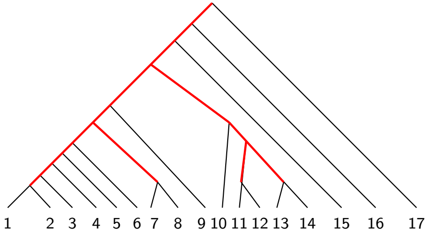
Definition

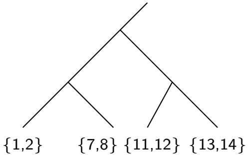
- The *prescaffold tree* is the subtree $T|_S$ where S is a set of leaves containing one leaf from each cherry blob in $\mathcal{B}$
- The *scaffold tree* is the unlabelled version of $T|_S$ where S is a set of leaves containing one leaf from each of the blobs in $\mathcal{B}$

Example: 2 - blobification



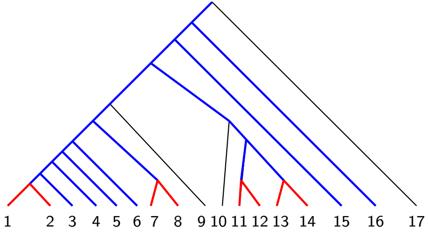
Example: 2 - blobification

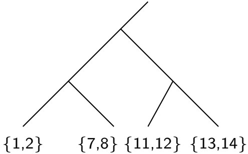




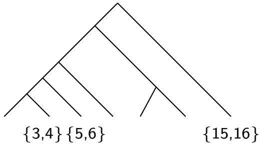
Prescaffold tree

Example: 2 - blobification





Prescaffold tree



Scaffold tree

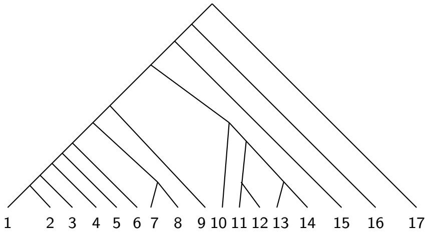
k-blobification using greedy algorithm

Algorithm

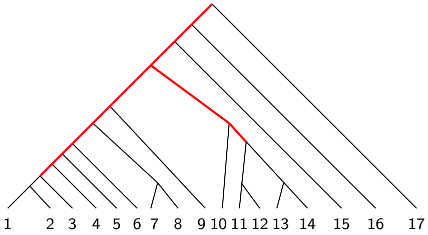
Input: A binary tree T and an integer k.
Output: A scaffold tree whose leaves correspond to blobs of size $\geq k$ and $\leq 2k-2$.

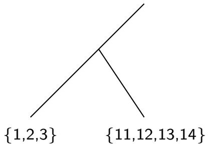
- *Include all the cherry blobs of minimal size in $\mathcal{C}$*
- *Each edge of the prescaffold tree has some smaller blobs hanging off of size $k-1$ or less*
- *Group these blobs together from the bottom edges of the prescaffold until they produce an edge blob of size between k and $2k-2$*
- *Add the edge blobs to the set $\mathcal{E}$*
- *The algorithm stops when each edge of the scaffold tree has at most $k-1$ leaves.*
- *The set $\mathcal{B} = \mathcal{C} \cup \mathcal{E}$ is called the greedy k-blobification*

Example: greedy 3 - blobification



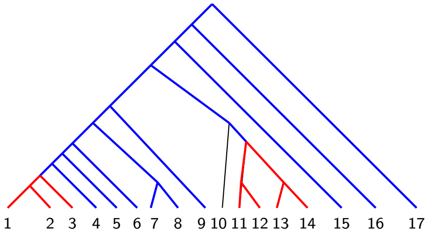
Example: greedy 3 - blobification

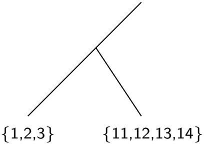




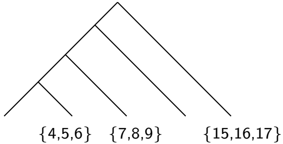
Prescaffold tree

Example: greedy 3 - blobification





Prescaffold tree



Scaffold tree

Some useful results

Lemma

If T is a rooted binary leaf-labeled tree with n leaves, then for all $k \geq 2$, T has a k-blobification with at least $\frac{n}{4k}$ blobs.

Proof idea:

- Applying the *greedy k-blobification* algorithm on T, we get 'a' cherry blobs and 'b' edge blobs.
- There are at most $2a - 1$ edges in the prescaffold tree, each having at most $k - 1$ leaves unassigned to any blob.
- Taking the number of leaves at its most extreme gives us the lower bound.

Some useful results

Lemma

Let S_1 and S_2 be uniformly random subsets of $[n]$, each of size at least $\sqrt{n}$. The probability that $S_1 \cap S_2 \neq \emptyset$ is at least $1 - e^{-1}$.

Proof idea:

- The probability that $S_1 \cap S_2 \neq \emptyset$ is minimized when both S_1 and S_2 have $\sqrt{n}$ elements.
- The probability that $S_1 \cap S_2 = \emptyset$ is $\frac{\binom{n-\sqrt{n}}{\sqrt{n}}}{\binom{n}{\sqrt{n}}} \leq e^{-1}$.

Lower Bound

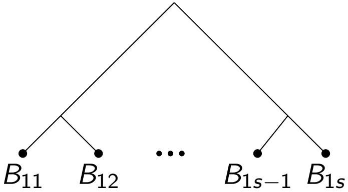
Theorem (M.-,Sullivant)

Let T_1 and T_2 be two uniformly random trees on n leaves among all trees with the same tree shape (i.e. T_2 is a random leaf relabeling of T_1). Then the expected size of $MAST(T_1, T_2)$ is at least $\sqrt{n}(1 - e^{-1})/4$.

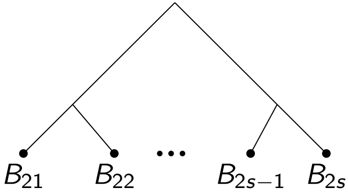
Proof idea:

- The $\sqrt{n}$-blobification of T_1 and T_2 is denoted by $\mathcal{B}_1 = \{B_{11}, \ldots, B_{1s}\}$ and $\mathcal{B}_2 = \{B_{21}, \ldots, B_{2s}\}$ and has at least $\sqrt{n}/4$ blobs.
- Both trees have the same tree shape and so have the same scaffold tree.
- The probability that $B_{1i} \cap B_{2i} \neq \emptyset$ is at least $1 - e^{-1}$.
- Selecting one leaf from each of $B_{1i} \cap B_{2i}$ gives us an agreement subtree of the shape of a subtree of the scaffold tree.

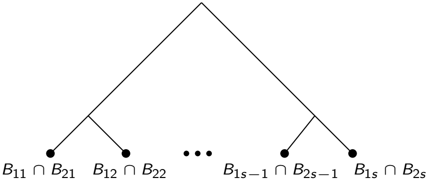
Proof idea (continued):



Scaffold tree of T_1



Scaffold tree of T_2



An agreement subtree of T_1 and T_2

Upper Bound

$RB(n)$: Set of all rooted binary trees with n leaves
$RB(S)$: Set of all rooted binary trees with leaf label set S, where $S \subseteq [n]$
P_n : Any probability distribution on $RB(n)$

Definition

A distribution on $RB(n)$ is *exchangeable* if any two trees which differ only by a permutation of leaves have the same probability.

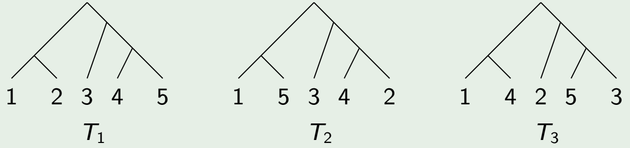
Upper Bound

$RB(n)$: Set of all rooted binary trees with n leaves
$RB(S)$: Set of all rooted binary trees with leaf label set S, where $S \subseteq [n]$
P_n : Any probability distribution on $RB(n)$

Definition

A distribution on $RB(n)$ is *exchangeable* if any two trees which differ only by a permutation of leaves have the same probability.

Example



If P_5 is any exchangeable distribution on $RB(5)$, then the probabilities of all the trees would be the same, i.e, $P_5(T_1) = P_5(T_2) = P_5(T_3)$.

Upper Bound

Definition

A family of distributions on random trees is said to satisfy *sampling consistency* if for all n, all $s < n$, all $S \subseteq [n]$ with $|S| = s$, and all $t \in RB(S)$,

$$P_s[t] = \sum_{T \in RB(n):\, T|_S = t} P_n[T].$$

Theorem (Bernstein, Ho, Long, Steel, John, Sullivant - 2015)

Consider an exchangeable and sampling consistent distribution on rooted binary trees. Then for any $\lambda > e\sqrt{2}$ there is a value m such that, for all $n \geq m$,

$$E[\mathrm{MAST}(T_1, T_2)] \leq \lambda\sqrt{n}$$

where T_1, T_2 are sampled from this distribution.

Upper Bound

- Since we only have a probability distribution P_n on n leaves and do not have a family of distributions P_s for $s < n$, we can not talk about sampling consistency.
- Need to define some new probability distributions on $RB(s)$ for $s < n$
- For any $s < n$, and $t \in RB(s)$ we define

$$P_s[t] = \sum_{T \in RB(n):\, T|_{[s]} = t} P_n[T].$$

Proposition

Let P_n be an exchangeable distribution defined on $RB(n)$. Then for any $s < n$, P_s satisfies exchangeability property on $RB(s)$.

Upper Bound

As we have defined a family of distributions using P_n, we deduce a strengthened version that does not require sampling consistency.

Theorem (M.-,Sullivant)

For any $\lambda > e\sqrt{2}$ there is a value m such that, for all $n \geq m$,

$$E[\mathrm{MAST}(T_1, T_2)] \leq \lambda\sqrt{n},$$

where T_1 and T_2 are sampled from any exchangeable distribution on $RB(n)$.

Proof idea:

- Proved exactly the way as Theorem 4.3 in [1]
- Equalities which followed from sampling consistency now follow from the definition of P_s and the fact that P_s is also exchangeable.

Upper bound

As the uniform distribution on trees with the same shape is exchangeable, we have the following Corollary:

Corollary

Let T_1 and T_2 be generated from the uniform distribution on rooted binary trees with n leaves with same tree shape (that is, T_2 is a random leaf relabeling of T_1). Then for any $\lambda > e\sqrt{2}$ there is a value m such that, for all $n \geq m$,

$$E[\mathrm{MAST}(T_1, T_2)] \leq \lambda\sqrt{n}.$$

Simulations with Blobification

- The blobification idea has the potential to improve the lower bounds on the expected size of the MAST in other contexts
- Construct the scaffold tree as a comb tree
- An agreement subtree could be obtained by comparing blobs matched along the path from the root to the deepest leaf in each scaffold tree
- Apply this technique on uniformly randomly trees
- The current best lower bound for the expected size of the MAST for two uniformly random trees on n leaves is $\Omega(n^{1/8})$ [1]

Simulations

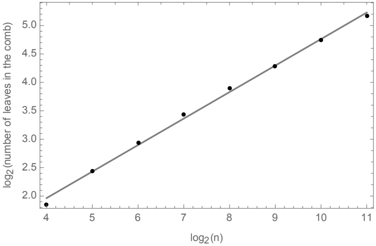


Figure: Log-log plot of the simulated expected size of the greedy comb scaffold

- The greedy comb scaffold algorithm applied to uniformly random binary trees with $k = \sqrt{n}$ on 2^n leaves for $n = 4, \dots, 11$, with 1000 samples for each value of n
- Slope of the line of best fit is approximately 0.466
- A strategy based on blobification could yield an estimate of $\Omega(n^{0.466})$

Summary

- The lower bound of the expected size of MAST of two uniformly random trees on n leaves is $\sqrt{n}(1-e^{-1})/4$, which is obtained using $\sqrt{n}-$ *blobification*
- The upper bound of the expected size of $\mathrm{MAST}(T_1, T_2)$ is $\lambda\sqrt{n}$, where $\lambda > e\sqrt{2}$ and T_1, T_2 are sampled from any exchangeable distribution on $RB(n)$
- The idea of blobification can be used to improve the lower bound of the expected size of MAST

Vanishing ideal of a Gaussian Graphical model

Preliminaries:

Definition

- Any positive definite $n \times n$ matrix Σ can be seen as the covariance matrix of a multivariate normal distribution in $\mathbb{R}^n$.
- The inverse matrix $K = \Sigma^{-1}$ is called the *concentration matrix* of the distribution.
- Statistical models where K can be written as a linear combination of some fixed linearly independent symmetric matrices $K_1, K_2, ..., K_d$ are called *linear concentration* models.
- Let $\mathbb{S}^n$ denote the vector space of real symmetric matrices and let $\mathcal{L}$ be a linear subspace of $\mathbb{S}^n$ generated by $K_1, K_2, ..., K_d$. The set $\mathcal{L}^{-1}$ is defined as

$$\mathcal{L}^{-1} = \{\Sigma \in \mathbb{S}^n : \Sigma^{-1} \in \mathcal{L}\}.$$

Gaussian Graphical models

- *Undirected Gaussian graphical model* is obtained when $\mathcal{L}$ is defined by the vanishing of some off-diagonal entries of K.
- Fix a graph $G = ([n], E)$ with vertex set $[n] = \{1, 2, ..., n\}$ and edge set E, which is assumed to contain all self loops.
- The subspace $\mathcal{L}$ is generated by the set $\{K_{ij} | (i, j) \in E\}$ of matrices K_{ij} with 1-entry at the $(i, j)^{th}$ and $(j, i)^{th}$ position and 0 in all other positions.
- The homogeneous prime ideal of all the polynomials in $\mathbb{R}[\Sigma] = \mathbb{R}[\sigma_{11}, \sigma_{12}, ..., \sigma_{nn}]$ that vanish on $\mathcal{L}^{-1}$ is denoted by P_G.

Computing the vanishing ideal

Problem

For a given graph G, find a generating set of the ideal P_G.

One way to compute P_G is to eliminate the entries of an indeterminate symmetric $n \times n$ matrix K from the following system of equations:

$$\Sigma \cdot K = Id_n, \qquad K \in \mathcal{L},$$

where Id_n is the $n \times n$ identity matrix.

Computing the vanishing ideal

Problem

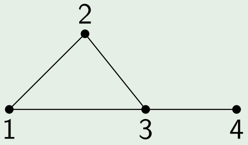
For a given graph G, find a generating set of the ideal P_G.

One way to compute P_G is to eliminate the entries of an indeterminate symmetric $n \times n$ matrix K from the following system of equations:

$$\Sigma \cdot K = Id_n, \qquad K \in \mathcal{L},$$

where Id_n is the $n \times n$ identity matrix.

Example

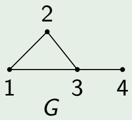
$G =$ 

Computing the vanishing ideal

Example (Continued)

Let $G = ([4], E)$. The matrices Σ and K for this graph are:

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}, K = \begin{bmatrix} k_{11} & k_{12} & k_{13} & 0 \\ k_{12} & k_{22} & k_{23} & 0 \\ k_{13} & k_{23} & k_{33} & k_{34} \\ 0 & 0 & k_{34} & k_{44} \end{bmatrix},$$



The ideal P_G can be calculated by using the equation $\Sigma \cdot K = Id_4$.

$$P_G = \langle \Sigma \cdot K - I \rangle = \langle \sigma_{11}k_{11} + \sigma_{12}k_{12} + \sigma_{13}k_{13} - 1, ..., \sigma_{34}k_{34} + \sigma_{44}k_{44} - 1 \rangle.$$

Eliminating the k_{ij} variables, we get

$$P_G = \langle \sigma_{13}\sigma_{34} - \sigma_{14}\sigma_{33}, \sigma_{23}\sigma_{34} - \sigma_{24}\sigma_{33}, \sigma_{14}\sigma_{23} - \sigma_{13}\sigma_{24} \rangle.$$

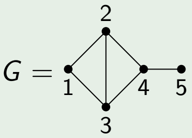
Separation and clique sums

Definition

- Let $G = (V, E)$ be a graph and let A, B, and C be disjoint subsets of the vertex set of G with $A \cup B \cup C = V$.
- The set C *separates* A and B if for any $a \in A$ and $b \in B$, any path from a to b passes through a vertex in C.
- The set C is called a *clique* of G if the subgraph induced by C is a complete graph.
- The graph G is a *c-clique sum* of smaller graphs G_1 and G_2 if there exists a partition (A, B, C) of its vertex set such that
  i) C is a clique with $|C| = c$,
  ii) C separates A and B,
  iii) G_1 and G_2 are the subgraphs induced by $A \cup C$ and $B \cup C$ respectively.

Separation and clique sums

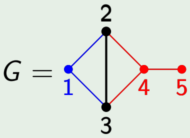
Example



- In the graph G, let $A = \{1\}$, $B = \{4, 5\}$ and $C = \{2, 3\}$.
- As every path from $\{1\}$ to $\{4, 5\}$ passes through $\{2, 3\}$, C *separates* A and B.
- C is a *clique* as the subgraph induced by C is a complete graph.

Separation and clique sums

Example



- In the graph G, let $A = \{1\}$, $B = \{4, 5\}$ and $C = \{2, 3\}$.
- As every path from $\{1\}$ to $\{4, 5\}$ passes through $\{2, 3\}$, C *separates* A and B.
- C is a *clique* as the subgraph induced by C is a complete graph.
- G is a *2-clique sum* of G_1 (blue) and G_2 (red), where G_1 and G_2 are the subgraphs induced by $\{1, 2, 3\}$ and $\{2, 3, 4, 5\}$ respectively.

Conditional independence

Proposition

Let $X = (X_1, X_2, ..., X_n)$ be a Gaussian random vector. If $A, B, C \subseteq [n]$ are pairwise disjoint subsets, then X_A is conditionally independent of X_B given X_C (i.e $A \perp\!\!\!\perp B | C$) if and only if the submatrix $\Sigma_{A\cup C, B\cup C}$ of the covariance matrix Σ has rank $|C|$.

- The *Gaussian conditional independence ideal* for the conditional independence statement $A \perp\!\!\!\perp B | C$ is given by

$$J_{A \perp\!\!\!\perp B|C} = \langle (|C|+1) \times (|C|+1) \text{ minors of } \Sigma_{A\cup C, B\cup C} \rangle.$$

- Let G be an undirected graph and (A, B, C) be a partition with C separating A from B.
  - The conditional independence statement $A \perp\!\!\!\perp B | C$ holds for all multivariate normal distributions where the covariance matrix Σ is obtained from G (by the global Markov property).

Conditional independence ideal

- The *conditional independence ideal* for a graph G is defined by

$$CI_G = \sum_{A \perp\!\!\!\perp B | C \text{ holds for } G} J_{A \perp\!\!\!\perp B | C}.$$

- In terms of a graph G, the conditional independence ideal is the set of determinantal constraints obtained from the zeros of the concentration matrix (non edges of the graph).

Proposition

Let (A, B, C) be any separating partition of G. Then the rank of the submatrix $\Sigma_{A \cup C, B \cup C}$ of the covariance matrix Σ is $|C|$ and hence the generators of CI_G vanish on the matrices in $\mathcal{L}^{-1}$. So for any given graph G,

$$CI_G \subseteq P_G.$$

Conditional independence ideal

Question

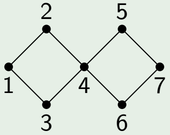
Is $CI_G = P_G$ *for all graphs ?* *No!*

Conditional independence ideal

Question

Is $CI_G = P_G$ for all graphs ? *No!*

Example

$G=$ 

Let $(A, B, C) = (\{1, 2, 3\}, \{5, 6, 7\}, \{4\})$ and let G_1 and G_2 be the subgraphs induced by $A \cup C$ and $B \cup C$ respectively. Then

$$P_G = CI_G + \langle m \rangle \neq CI_G,$$

where m is a homogeneous polynomial of degree 4 which cannot be obtained from any determinantal constraints given by P_{G_1}, P_{G_2} or the 2×2 minors of $\Sigma_{A\cup C, B\cup C}$.

1-clique sums

Question

For which class of graphs is $P_G = CI_G$?

Definition

A graph G is called a *1-clique sum of complete graphs* if there exists a partition (A, B, C) of its vertex set such that

i) $|C| = 1$,

ii) C separates A and B,

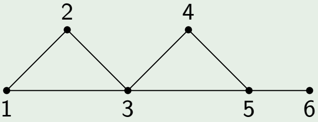
iii) the subgraphs induced by $A \cup C$ and $B \cup C$ are either complete graphs or 1-clique sum of complete graphs.

Definition

A vertex c is called a *central vertex* if there exists a 1-clique partition (A, B, C) with $C = \{c\}$.

1-clique sums

Example

$G =$ 

- Let $G = ([6], E)$. Consider the partition (A, B, C) where $A = \{1, 2\}$, $B = \{4, 5, 6\}$ and $C = \{3\}$.
- The subgraph induced by $A \cup C$ is a complete graph.
- The subgraph induced by $B \cup C$ is a *1-clique sum of complete graphs* (with $(A_1, B_1, C_1) = (\{3, 4\}, \{6\}, \{5\})$).
- Similarly, $(\{1, 2, 3, 4\}, \{6\}, \{5\})$ is also a *1-clique partition*.
- 3 and 5 are the *central vertices* of the graph.

The Conjecture

Conjecture (Sturmfels, Uhler - 2009)

The prime ideal P_G of an undirected Gaussian graphical model is generated in degree ≤ 2 if and only if each connected component of the graph G is a 1-clique sum of complete graphs. In this case, P_G has a Gröbner basis consisting of entries of Σ and 2×2 minors of Σ.

Aim: To show that CI_G is equal to the vanishing ideal P_G when G is a 1-clique sum of complete graphs.
In this case, the conditional independence ideal can be written as

$$CI_G = \langle \bigcup_{(A,B,C) \in C_1(G)} 2 \times 2 \text{ minors of } \Sigma_{A \cup C, B \cup C} \rangle,$$

where $C_1(G)$ denotes the set of all 1-clique partitions of G.

Some useful results

Proposition

If G is a 1-clique sum of complete graphs, then there exists a unique shortest path between any two vertices i and j in G.

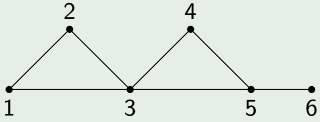
- We denote the unique shortest path between i and j by $i \leftrightarrow j$.
- If (A, B, C) is a 1-clique partition of G with $C = \{c\}$, and if $i \in A, j \in B$ then $i \leftrightarrow j$ decomposes into $i \leftrightarrow c \cup c \leftrightarrow j$.
- Let $F = \{f_{ij} : 1 \leq i \leq j \leq n\} \subseteq \mathbb{R}[k_{11}, k_{12}, ..., k_{nn}]$, where f_{ij} is $\det(K)$ times the $(i, j)^{th}$ coordinate of K^{-1}.
- Shortest path monomial : Each f_{ij} has the monomial

$$\prod_{(i',j') \in i \leftrightarrow j} k_{i'j'} \prod_{t \notin i \leftrightarrow j} k_{tt}$$

as one of its terms.

Some useful results

Example

$G =$ 

For this graph G, the polynomial f_{12} has the monomial $k_{12}k_{33}k_{44}k_{55}k_{66}$ as one of its terms.
Similarly, f_{14} has the monomial $k_{13}k_{34}k_{22}k_{55}k_{66}$ as one of its terms.

Final result

Theorem (M.-,Sullivant)

The conjecture given by Sturmfels and Uhler is true.

Proof Idea:

- Existence of the unique shortest path allows us to define the *shortest path map* ψ
- $\ker\psi = CI_G$
- Construct a partial term order on $\mathbb{R}[F]$ using the shortest path monomial
- Using this term order, define the *initial term map* ϕ
- $\ker\psi = \ker\phi$
- F forms a SAGBI (Subalgebra Analogue to Gröbner Basis for Ideals) basis of $\mathbb{R}[F]$

Future Projects

Question

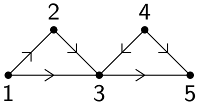
Let G be a 1-clique sum of two smaller graphs G_1 and G_2 attached at the vertex $\{c\}$. Is the following relation true:

$$P_G = \langle P_{G_1} + P_{G_2} + 2 \times 2 \text{ minors of } \Sigma_{A\cup C, B\cup C}\rangle : \langle \sigma_{cc}\rangle^{\infty}$$

- Computations in Macaulay2 suggests that this result might be true as it holds for various examples.

Question

Can we find a generating set of P_G for directed acyclic graphs (DAGs) using similar techniques (especially for directed acyclic analogue of 1-clique sum of complete graphs)?



References

- Daniel Irving Bernstein, Lam Si Tung Ho, Colby Long, Mike Steel, Katherine St. John and Seth Sullivant. Bounds on the expected size of the maximum agreement subtree. *SIAM J. Discrete Math.* **29** (2015), no. 4, 2065–2074.
- David Bryant, Andy McKenzie, Mike Steel. The size of a Maximum agreement subtree for random binary trees. in BioConsensus, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **61**, AMS, Providence, RI, 2003, pp. 55–65.
- Beatrix Jones, Mike West. Covariance decomposition in undirected Gaussian graphical models, Biometrika, Volume 92, Issue 4, December 2005, 779-786,
- Daniel M. Martin and Bhalchandra. D. Thatte: The maximum agreement subtree problem. *Discrete Appl. Math.* **161** (2013):1805-1817.
- Pratik Misra, Seth Sullivant. Bounds on the expected size of the maximum agreement subtree for a given tree shape, `https://arxiv.org/abs/1809.04488`.
- Bernd Sturmfels, Caroline Uhler. Multivariate Gaussians, Semidefinite Matrix Completion and Convex Algebraic Geometry, Annals of the Institute of Statistical Mathematics (2010) 62:603-638
- Bernd Sturmfels. Gröbner Bases and Convex Polytopes, University Lecture Series, Volume 8, AMS, 1996.
- Seth Sullivant. Algebraic Statistics, AMS, 2018.