Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling

Simón Lunagómez\textsuperscript{1}, Sayan Mukherjee\textsuperscript{1,2,3,4}, Robert L. Wolpert\textsuperscript{1,5}

\textsuperscript{1}Department of Statistical Science
\textsuperscript{2}Department of Computer Science
\textsuperscript{3}Institute for Genome Sciences & Policy
\textsuperscript{4}Department of Mathematics
\textsuperscript{5}Nicholas School of the Environment

Duke University

April 27, 2010
Main idea

1. Variable interactions can be modeled as conditional dependence
Main idea

1. Variable interactions can be modeled as conditional dependence
2. Conditional independence can be modeled as a hyper-graph
Main idea

1. Variable interactions can be modeled as conditional dependence
2. Conditional independence can be modeled as a hyper-graph
3. Hyper-graphs with $d$ nodes can be visualized using convex sets centered at $d$ points in $\mathbb{R}^m$
Main idea

1. Variable interactions can be modeled as conditional dependence
2. Conditional independence can be modeled as a hyper-graph
3. Hyper-graphs with \( d \) nodes can be visualized using convex sets centered at \( d \) points in \( \mathbb{R}^m \)
4. Computational geometry and topology have many modeling tools for statistical inference
The problem

Given iid draws $X_1, \ldots, X_n$ from $f(X)$ where $X_i \in \mathbb{R}^d$ infer the factorized density

$$f(x) = \prod_{\alpha \in \mathcal{C}(G)} \eta_\alpha(x)$$

$\mathcal{C}(G)$ are complete sets of $G$ and $\eta_\alpha(x)$ are potentials.
An example and notation

Given variables $X, Y, Z$ with joint density

$$X \perp Y \mid Z \iff f(x, y, z) = f_{XZ}(x, z)f_{YZ}(y, z)f_Z(z)$$

then

$$X \perp Y \mid Z$$
Example
Two parts to the problem

1. Infer the graph structure: $G$. This is the focus of this talk.
The problem setup

Two parts to the problem

1. Infer the graph structure: $G$. This is the focus of this talk.
2. Infer the hyper Markov law: $\eta_\alpha(x \mid \theta)$. This will not be the focus, will use standard methods.
   Example: for a Gaussian the parameters are the mean and covariance $\theta = \{\mu, \Sigma\}$. 
Going beyond graphs

Often we may be interested in high-order interactions that a graph cannot capture. For example

\[ f(x) = f(x_2, x_6)f(x_1, x_2)f(x_1, x_6)f(x_3, x_4, x_5). \]
Example
Example
Junction trees

A very common and useful decomposition of joint densities is

$$f(x) = \frac{\prod_{a \in \mathcal{P}(G)} \psi_a(x | \theta)}{\prod_{b \in \mathcal{I}(G)} \psi_b(x | \theta)}$$

where $\mathcal{P}(G)$ and $\mathcal{I}(G)$ are the prime components and separators of $G$ and $\psi_b(x | \theta)$ are marginal densities.
Example
Potential functions

Another very common and useful decomposition of joint densities is

\[ f(x) = \prod_{a \in \mathcal{C}(G)} \phi_a(x_a | \theta_a) \]

where \( \mathcal{C}(G) \) are complete sets of the graph and \( \phi_a(x_a | \theta_a) \) are potential functions.
Likelihood

Recall

\[ f(x) = \frac{\prod_{a \in \mathcal{P}(G)} \psi_a(x | \theta)}{\prod_{b \in \mathcal{S}(G)} \psi_b(x | \theta)}, \quad f(x) = \prod_{a \in \mathcal{C}(G)} \phi_a(x_a | \theta_a). \]
Likelihood

Recall

\[ f(x) = \frac{\prod_{a \in \mathcal{P}(G)} \psi_a(x \mid \theta)}{\prod_{b \in \mathcal{S}(G)} \psi_b(x \mid \theta)}, \quad f(x) = \prod_{a \in \mathcal{C}(G)} \phi_a(x_a \mid \theta_a). \]

Two types of parameters \((\psi \in \Psi), (\phi \in \Phi)\) and \(\theta \in \Theta\)

\[ f(x \mid \psi, \theta) = \frac{\prod_{a \in \mathcal{P}(G(\psi))} \psi_a(x \mid \theta)}{\prod_{b \in \mathcal{S}(G(\psi))} \psi_b(x \mid \theta)}, \quad f(x \mid \phi, \theta) = \prod_{a \in \mathcal{C}(G)} \phi_a(x_a \mid \theta_a) \]
Likelihood

Recall

\[ f(x) = \frac{\prod_{a \in \mathcal{P}(G)} \psi_a(x | \theta)}{\prod_{b \in \mathcal{L}(G)} \psi_b(x | \theta)}, \quad f(x) = \prod_{a \in \mathcal{E}(G)} \phi_a(x_a | \theta_a). \]

Two types of parameters \((\psi \in \Psi), (\phi \in \Phi)\) and \(\theta \in \Theta\)

\[ f(x | \psi, \theta) = \frac{\prod_{a \in \mathcal{P}(G(\psi))} \psi_a(x | \theta)}{\prod_{b \in \mathcal{L}(G(\psi))} \psi_b(x | \theta)}, \quad f(x | \phi, \theta) = \prod_{a \in \mathcal{E}(G)} \phi_a(x_a | \theta_a). \]

The likelihood is

\[ \text{Lik}(X_1, \ldots, X_n) \propto \prod_{i=1}^{n} f(X_i | \psi, \theta), \quad \text{Lik}(X_1, \ldots, X_n) \propto \prod_{i=1}^{n} f(X_i | \phi, \theta). \]
Marginal likelihood

If we only care about the graph structure we look at the marginal likelihood

$$\Pr\{G \mid X_1, \ldots, X_n\} \propto \int_{\Theta_G} f(x \mid \theta, G) p(G) p(\theta \mid G) d\theta.$$
We would like to sample from

\[ \text{Post} \left( \theta, \psi(G) \mid (X_i)_{i=1}^n \right) \propto \left[ \prod_{i=1}^n f(X_i \mid \psi, \theta) \right] \pi(\theta, \psi(G)). \]

Typically this is hard.
Markov chain Monte Carlo

Consider the posterior distribution

$$\pi := \text{Post} \left( \theta, \psi(G) \mid (X_i)_{i=1}^n \right).$$
Consider the posterior distribution

\[ \pi := \text{Post} \left( \theta, \psi(G) \mid (X_i)_{i=1}^n \right). \]

Construct a Markov chain with transition kernel

\[ Q(\theta_i, \psi_i(G) \mid \theta_j, \psi_j(G)), \]

such that the stationary distribution is \( \pi \).
Main idea

Priors on $d$ points in $\mathbb{R}^m$ can be induced to place priors on graphs with $d$ nodes. The device used to do this is an abstract simplicial complex – intersections of convex sets in $\mathbb{R}^m$. 
Main idea

Priors on $d$ points in $\mathbb{R}^m$ can be induced to place priors on graphs with $d$ nodes. The device used to do this is an abstract simplicial complex – intersections of convex sets in $\mathbb{R}^m$.

This is in contrast to Erdös-Rényi random graphs.
Proximity graph
Graphs from two different complexes
Nerves

Definition

Let $F = \{A_j, \ j \in I\}$ be a finite collection of distinct nonempty convex sets. The nerve of $F$ is given by

$$Nrv(F) = \{\sigma \subseteq I : \bigcap_{j \in \sigma} A_j \neq \emptyset\}.$$
**Čech Complex**

**Definition**

Let \( \mathcal{V} \) be a finite set of points in \( \mathbb{R}^d \) and \( r > 0 \). Denote by \( \mathbb{B}_d \) the closed unit ball in \( \mathbb{R}^d \). The Čech complex corresponding to \( \mathcal{V} \) and \( r \) is the nerve of the sets \( B_{v,r} = v + r\mathbb{B}_d \), \( v \in \mathcal{V} \). This is denoted by \( Nrv(\mathcal{V}, r, Čech) \).
Delaunay Triangulation

Definition
Let $\mathcal{V}$ be a finite set of points in $\mathbb{R}^d$. The Delaunay triangulation corresponding to $\mathcal{V}$ is the nerve of the sets $C_v = \{x \in \mathbb{R}^d : \|x - v\| \leq \|x - u\|, \ u \in \mathcal{V}\}$ for $v \in \mathcal{V}$. This is denoted by $Nrv(\mathcal{V}, \text{Delaunay})$, and the sets $C_v$ are called Voronoi cells.
Alpha Complex

Definition

Let $\mathcal{V}$ be a finite set of points in $\mathbb{R}^d$ and $r > 0$. The Alpha complex corresponding to $\mathcal{V}$ and $r$ is the nerve of the sets $B_{v,r} \cap C_v$, $v \in \mathcal{V}$. This is denoted by $Nrv(\mathcal{V}, r, Alpha)$. 
Abstract simplicial complex

Definition
Let $\mathcal{V}$ be a finite set. A simplicial complex with base set $\mathcal{V}$ is a family $\mathcal{K}$ of subsets of $\mathcal{V}$ such that $\tau \in \mathcal{K}$ and $\sigma \subseteq \tau$ implies $\sigma \in \mathcal{K}$. The elements of $\mathcal{K}$ are called simplices, and the number of connected components of $\mathcal{K}$ is denoted $\#(\mathcal{K})$. 
**Definition**

Let $\mathcal{K}$ be a simplicial complex, and denote by $|\tau|$ the cardinality of a simplex $\tau \in \mathcal{K}$. The $p$-skeleton of $\mathcal{K}$ is the collection of all $\tau \in \mathcal{K}$ such that $|\tau| \leq p + 1$. The elements of the $p$-skeleton are called $p$-simplices and the 1-skeleton is just a graph (more precisely, it is $\mathcal{V} \cup \mathcal{E}$ for a uniquely determined graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$).
Points, nerves, and skeletons
Alpha example

\[
f(x) = \frac{\psi_{12}(x_1, x_2)\psi_{235}(x_2, x_3, x_5)\psi_{4}(x_4)}{\psi_{2}(x_2)}.\]

\[ f(x) = \frac{\psi_{1235}(x_1, x_2, x_3, x_5) \psi_{14}(x_1, x_4)}{\psi_1(x_1)}. \]
Observation

Different families of convex sets induce different restrictions in graph space.
Observation

Different families of convex sets induce different restrictions in graph space.

- Delaunay triangulation: Clique size cannot exceed \( d + 1 \).
Observation

Different families of convex sets induce different restrictions in graph space.

- Delaunay triangulation: Clique size cannot exceed $d + 1$.
- Čech complex: No star graph with central node of degree higher than the “kissing number” – 6 for $d = 2$, 12 for $d = 3$. 
Good things about the point set representation

1. Number of parameters to specify graphs or hyper-graphs depends on the number of vertices, not the edge set.
Good things about the point set representation

1. Number of parameters to specify graphs or hyper-graphs depends on the number of vertices, not the edge set.
2. The parameter space of factorizations $\Psi$ will be a subset of $\mathbb{R}^d$ which is easy to sample.
Definition

Fix integers $d, m \in \mathbb{N}$ and let $V = (V_1, \ldots, V_d)$ be drawn from a probability distribution $Q$ on $(\mathbb{R}^m)^d$. For any class $A$ of convex sets in $\mathbb{R}^m$ and radius $r > 0$, the graph $G(V, r, A)$ is said to be a Random Geometric Graph (RGG).
Isomorphic

Definition
Write \( G_1 \cong G_2 \) for two graphs \( G_i = (V_i, E_i) \) and call the graphs isomorphic if there is a 1:1 mapping \( \chi : V_1 \to V_2 \) such that \( (v_i, v_j) \in E_1 \iff (\chi(v_i), \chi(v_j)) \in E_2 \) for all \( v_i, v_j \in V_1 \).
Feasible graph

Definition

Fix numbers \(d, n \in \mathbb{N}\), a class \(\mathcal{A}\) of convex sets in \(\mathbb{R}^d\), and a distribution \(Q\) on the random vectors \(V\) in \((\mathbb{R}^d)^n\). A graph \(\Gamma\) is said to be feasible if for some number \(r > 0\),

\[
\Pr \{ \mathcal{G}(V, r, \mathcal{A}) \cong \Gamma \} > 0.
\]
Choice of distribution

We use two distributional models for $Q$.

1. Uniform iid.

Random geometric graphs

Uniform
Mattérv III
Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling

Random geometric graphs

Erdös-Rényi
## Comparison

| Graph | $|V|$ | Edges | 25% | 50% | 75% | 3-Cliques | 25% | 50% | 75% |
|-------|-----|-------|-----|-----|-----|-----------|-----|-----|-----|
| Uniform | 75 | 161 | 171 | 182 | 134 | 160 | 190 |
| Matérn (0.035) | 75 | 154 | 161 | 170 | 110 | 124 | 144 |
| ER (0.050) | 75 | 130 | 138 | 146 | 6 | 8 | 11 |
| ER (0.065) | 75 | 172 | 181 | 189 | 14 | 18 | 22 |
| Uniform | 50 | 69 | 75 | 81 | 34 | 43 | 57 |
| Matérn (0.035) | 50 | 66 | 71 | 76 | 27 | 35 | 43 |
| Matérn (0.050) | 50 | 62 | 67 | 71 | 22 | 27 | 33 |
| ER (0.050) | 50 | 56 | 61 | 67 | 1 | 2 | 4 |
| ER (0.065) | 50 | 74 | 79 | 85 | 3 | 5 | 7 |
| Uniform | 20 | 9 | 12 | 14 | 1 | 2 | 4 |
| Matérn (0.035) | 20 | 9 | 11 | 13 | 1 | 1 | 3 |
| Matérn (0.050) | 20 | 8 | 10 | 12 | 0 | 1 | 2 |
| ER (0.050) | 20 | 8 | 9 | 11 | 0 | 0 | 0 |
| ER (0.065) | 20 | 10 | 12 | 15 | 0 | 0 | 1 |
General setting

Likelihood

\[ f(x) = \prod_{a \in \mathcal{C}(G)} \phi_a(x_a | \theta_a) \quad \text{or} \quad f(x) = \frac{\prod_{a \in \mathcal{P}(G)} \psi_a(x_a | \theta_a)}{\prod_{b \in \mathcal{S}(G)} \psi_b(x_b | \theta_b)}. \]

Prior specification

\[ p(\theta, G) = p(\theta | G) p(G) \]
Proposition

Every feasible graph in $\mathbb{R}^d$ may be represented in the form $\mathcal{G}(\mathcal{V}, r, A)$ for a collection $\mathcal{V}$ of $n$ points in the unit ball $\mathbb{B}^d$ and for $r = \frac{1}{n}$. 
Proposition

Every feasible graph in $\mathbb{R}^d$ may be represented in the form $G(\mathcal{V}, r, A)$ for a collection $\mathcal{V}$ of $n$ points in the unit ball $B^d$ and for $r = \frac{1}{n}$.

Implication: $r = \frac{1}{n}$ and write $G(\mathcal{V}, A)$ instead of $G(\mathcal{V}, r, A)$ or simply $G(\mathcal{V})$ if $A$ is understood.
Regimes of interest

\[ V = (V_1, \ldots, V_n) \overset{iid}{\sim} \mathbb{B}^d \quad \text{for } G(V, r, A): \mathbb{E}[\# \mathcal{E}] \leq \binom{n}{2} (2r)^d; \text{ for } r = \frac{1}{n} \text{ in dimension } d = 2 \text{ note } \mathbb{E}[\# \mathcal{E}] < 2. \]
Regimes of interest

\[ V = (V_1, \ldots, V_n) \overset{iid}{\sim} \mathbb{B}^d \] for \( G(V, r, A) \): \( \mathbb{E}[\#E] \leq \binom{n}{2}(2r)^d \); for \( r = \frac{1}{n} \) in dimension \( d = 2 \) note \( \mathbb{E}[\#E] < 2 \).

**Proposition**

The empty graph on \( n \) vertices cannot be expressed as \( G(V, r, \check{\text{Cech}}) \) for any \( V \subset \mathbb{B}^d \) with \( r \geq (n^{1/d} - 1)^{-1} \).
Objective

Given a random sample $x = \{x_1, ..., x_n\}$ our objective is to sample from the marginal likelihood

$$\Pr\{G \mid x\} \propto \int_{\Theta_G} f(x \mid \theta, G) p(G) p(\theta \mid G) \, d\theta.$$
Objective

Given a random sample $x = \{x_1, \ldots, x_n\}$ our objective is to sample from the marginal likelihood

$$\Pr\{G \mid x\} \propto \int_{\Theta_G} f(x \mid \theta, G) p(G) p(\theta \mid G) d\theta.$$ 

We will do this by using a Metropolis/Hastings approach. We will define a random walk on the configuration space $\mathcal{V} = \{V_1, \ldots, V_d\}$ that has the marginal likelihood as its stationary distribution.
Proposal

(1) Local moves – Given a configuration $\mathcal{V}(t)$ we propose to move to $\mathcal{V}(\ast)$ by a random walk that is informally reflecting Brownian motion on the unit interval. The walk is parameterized in spherical coordinates, radius and Euler angles.
Proposal

(1) Local moves – Given a configuration $\mathcal{V}^{(t)}$ we propose to move to $\mathcal{V}^{(*)}$ by a random walk that is informally reflecting Brownian motion on the unit interval. The walk is parameterized in spherical coordinates, radius and Euler angles.

(2) Global moves – Propose $\mathcal{V}^{(*)} \sim Q$
Local move
Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling

Model specification

Sampling from the posterior distribution

### Metropolis/Hastings

Given the likelihood and marginal likelihood

\[
 f(x \mid \theta, G) = \prod_{i=1}^{N} f(x_i \mid \theta, G), \quad \mathcal{M}(G) = \int_{\Theta_G} f(x \mid \theta, G) \ p(\theta \mid G) \ d\theta.
\]

A proposed move from \( V^{(t)} \) to \( V^* \) is accepted with probability

\[
 1 \wedge H^{(t)} \quad \text{for}
\]

\[
 H^{(t)} = \frac{\mathcal{M}(G^*) \ p(V^*) \ q(V^{(t)} \mid V^*)}{\mathcal{M}(G^{(t)}) \ p(V^{(t)}) \ q(V^* \mid V^{(t)})}.
\]
Denote by $\hat{\mathcal{G}}(n, d, \mathcal{A})$ the finite set of feasible graphs with $n$ vertices in $\mathbb{R}^d$. 
Denote by $\mathcal{G}(n, d, \mathcal{A})$ the finite set of feasible graphs with $n$ vertices in $\mathbb{R}^d$.

For each $\mathcal{G} \in \mathcal{G}(n, d, \mathcal{A})$ let $V_{\mathcal{G}} \subset (\mathbb{B}^d)^n$ denote the set of all points $\mathbf{V} = \{V_1, \ldots, V_n\} \in (\mathbb{B}^d)^n$ for which $\mathcal{G} \cong \mathcal{G}(\mathbf{V}, \frac{1}{n}, \mathcal{A})$, and set $\mu(\mathcal{G}) = Q(V_{\mathcal{G}})$. 

Definitions
Proposition

The sequence $G^(t) = G(V^(t), \frac{1}{n}, A)$ induced by the prior MCMC procedure described above samples each feasible graph $G \in \hat{G}(n, d, A)$ with asymptotic frequency $\mu(G)$. The posterior procedure described above samples each feasible graph with asymptotic frequency $\mu(G | x)$, the posterior distribution of $G$ given the data $x$ and hyper Markov prior $p(\theta | G)$. 

Convergence
The model

$$f_\theta(x) = \frac{\psi_\theta(x_1, x_4, x_{10}) \psi_\theta(x_1, x_8, x_{10}) \psi_\theta(x_4, x_5) \psi_\theta(x_8, x_9) \psi_\theta(x_2, x_3, x_9) \psi_\theta(x_6) \psi_\theta(x_7)}{\psi_\theta(x_4) \psi_\theta(x_8) \psi_\theta(x_9) \psi_\theta(x_1, x_{10})}$$
The marginals

Distribution function

$$\Psi_\theta(x_I) = \left(1 - n_I + \sum_{i \in I} x_i^{-\theta}\right)^{-1/\theta}$$

and density function

$$\psi_\theta(x_I) = \theta^{n_I} \frac{\Gamma(n_I + 1/\theta)}{\Gamma(1/\theta)} \left(1 - n_I + \sum_{i \in I} x_i^{-\theta}\right)^{-n_I-1/\theta} \left(\prod_{i \in I} x_i\right)^{-1-\theta}$$

on $[0, 1]^{n_I}$ for some $\theta \in \Theta = (0, \infty)$, for each clique $\{v_i : i \in I\}$ of size $n_I$. 
Three simulation examples

Example 1: $G$ is in the space generated by $A$

### Posterior inference

<table>
<thead>
<tr>
<th>Graph Topology</th>
<th>Posterior prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1, 4, 10][1, 8, 10][4, 5][8, 9][2, 3, 9][6][7]</td>
<td>0.963</td>
</tr>
<tr>
<td>[1, 4, 10][1, 8, 10][4, 5][8, 9][2, 3, 9][6][5, 7]</td>
<td>0.021</td>
</tr>
<tr>
<td>[1, 4, 10][1, 8][4, 5][8, 9][2, 3, 9][6][7]</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Inferred graphs
Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling

Three simulation examples

Example 2: Factorization based on nerves

The model

\[ f(x \mid G, \theta) = c_G \phi_\theta(x_1, x_2)\phi_\theta(x_1, x_6)\phi_\theta(x_2, x_6)\phi_\theta(x_3, x_4, x_5). \]
Three simulation examples

Example 2: Factorization based on nerves

Posterior inference – uniform

<table>
<thead>
<tr>
<th>Maximal simplices</th>
<th>Posterior prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{3, 4, 5} {1, 2} {2, 6} {1, 6}</td>
<td>0.609</td>
</tr>
<tr>
<td>{1, 2, 6} {3, 4} {4, 5} {3, 5}</td>
<td>0.161</td>
</tr>
<tr>
<td>{3, 5} {1, 6} {3, 4} {1, 2} {2, 6}</td>
<td>0.137</td>
</tr>
</tbody>
</table>
Posterior inference – repulsive

<table>
<thead>
<tr>
<th>Maximal simplices</th>
<th>Posterior prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{3, 4, 5} {1, 2} {2, 6} {1, 6}</td>
<td>0.824</td>
</tr>
<tr>
<td>{1, 2, 6} {3, 4, 5}</td>
<td>0.111</td>
</tr>
<tr>
<td>{1, 2, 6} {3, 4} {3, 5} {4, 5}</td>
<td>0.002</td>
</tr>
</tbody>
</table>
Proposed complices

Three filtrations

1. $\alpha$-complex in $\mathbb{R}^2$
2. $\alpha$-complex in $\mathbb{R}^3$
3. Čech-complex in $\mathbb{R}^2$
The model

\[ f_\theta(x) = \frac{\psi_\theta(x_2, x_3, x_4)\psi_\theta(x_1, x_3)\psi_\theta(x_5)}{\psi_\theta(x_3)} \]
Posterior inference for $\mathcal{M}_1$

<table>
<thead>
<tr>
<th>Model</th>
<th>Filtration</th>
<th>HPP Models</th>
<th>PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}_1$</td>
<td>$\alpha$ in $\mathbb{R}^2$</td>
<td>$[2, 3, 4][1, 3][5]$</td>
<td>0.964</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[2, 3, 4][1, 2, 3][5]$</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[2, 3, 4][1, 2, 3][3, 5]$</td>
<td>0.012</td>
</tr>
<tr>
<td>$\mathcal{M}_1$</td>
<td>$\alpha$ in $\mathbb{R}^3$</td>
<td>$[2, 3, 4][1, 3][5]$</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 3][2, 3][3, 4][5]$</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 3][2, 3][2, 4][5]$</td>
<td>0.003</td>
</tr>
<tr>
<td>$\mathcal{M}_1$</td>
<td>Čech in $\mathbb{R}^2$</td>
<td>$[2, 3, 4][1, 3][5]$</td>
<td>0.595</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 2, 3, 4][5]$</td>
<td>0.179</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 2, 3][2, 3, 4][5]$</td>
<td>0.168</td>
</tr>
</tbody>
</table>
Model 2

\[ f_\theta(x) = \frac{\psi_\theta(x_1, x_2, x_4)\psi_\theta(x_1, x_3, x_4)\psi_\theta(x_1, x_4, x_5)}{(\psi_\theta(x_1, x_4))^2} \]
### Posterior inference for $\mathcal{M}_2$

<table>
<thead>
<tr>
<th>Model</th>
<th>Filtration</th>
<th>HPP Models</th>
<th>PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{M}_2$</td>
<td>$\alpha$ in $\mathbb{R}^2$</td>
<td>$[1, 2][1, 3, 4][1, 4, 5]$</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[2, 4][1, 3, 4][3, 4, 5]$</td>
<td>0.115</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 2, 4][1, 3, 4][3, 4, 5]$</td>
<td>0.112</td>
</tr>
<tr>
<td>$\mathcal{M}_2$</td>
<td>$\alpha$ in $\mathbb{R}^3$</td>
<td>$[1, 2, 4][1, 3, 4][1, 4, 5]$</td>
<td>0.976</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 4, 5][1, 2, 3, 4]$</td>
<td>0.016</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 2, 4][1, 3][1, 4, 5]$</td>
<td>0.009</td>
</tr>
<tr>
<td>$\mathcal{M}_2$</td>
<td>Čech in $\mathbb{R}^2$</td>
<td>$[1, 2, 4][1, 3, 4][1, 4, 5]$</td>
<td>0.758</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 2, 4][1, 3, 4][1, 3, 5]$</td>
<td>0.177</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$[1, 2, 4][1, 3, 4][4, 5]$</td>
<td>0.148</td>
</tr>
</tbody>
</table>
Challenges

(1) RGG theory for repulsive processes.
Open problems

Challenges

(1) RGG theory for repulsive processes.
(2) Characterization of feasible graph space.
Challenges

(1) RGG theory for repulsive processes.
(2) Characterization of feasible graph space.
(3) Scaling to large graphs – computational challenges, other complexes.
Challenges

(1) RGG theory for repulsive processes.
(2) Characterization of feasible graph space.
(3) Scaling to large graphs – computational challenges, other complexes.
(4) Directed graphs – Forman/Morse theory?
Challenges

(1) RGG theory for repulsive processes.
(2) Characterization of feasible graph space.
(3) Scaling to large graphs – computational challenges, other complexes.
(4) Directed graphs – Forman/Morse theory?
(5) Random walks on simplicial complexes.
Open problems

Challenges

(1) RGG theory for repulsive processes.
(2) Characterization of feasible graph space.
(3) Scaling to large graphs – computational challenges, other complexes.
(4) Directed graphs – Forman/Morse theory?
(5) Random walks on simplicial complexes.
(6) Social networks.
Acknowledgements

Funding:

- Center for Systems Biology at Duke
- NSF
- NIH
Dimension reduction on huge data sets

Stoyan Georgiev$^{3,5}$, Sayan Mukherjee$^{1,2,3,4}$, Nick Patterson$^6$

Department of Statistical Science$^1$
Institute for Genome Sciences & Policy$^3$
Department of Computer Science$^2$
Department of Mathematics$^4$
Computational Biology and Bioinformatics$^5$
Duke University
Broad Institute$^6$
MIT

April 27, 2010
### Wishart

<table>
<thead>
<tr>
<th>matrix size</th>
<th>exact PCA</th>
<th>randomized PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>30 mins</td>
<td>1 min</td>
</tr>
<tr>
<td>8,000</td>
<td>3 hr 30 min</td>
<td>3 min</td>
</tr>
<tr>
<td>12,000</td>
<td>8 hr 30 min</td>
<td>6 min</td>
</tr>
</tbody>
</table>
## Gaussian

<table>
<thead>
<tr>
<th>Matrix Size</th>
<th>Randomized PCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000-by-50,000</td>
<td>15 min</td>
</tr>
<tr>
<td>20,000-by-60,000</td>
<td>1 hr 30 min</td>
</tr>
<tr>
<td>30,000-by-60,000</td>
<td>2 hr 30 min</td>
</tr>
<tr>
<td>50,000-by-80,000</td>
<td>6 hr</td>
</tr>
<tr>
<td>10,000-by-400,000</td>
<td>4 hr 25 min</td>
</tr>
<tr>
<td>20,000-by-400,000</td>
<td>11 hr 45 min</td>
</tr>
</tbody>
</table>