Systemic networks for high-dimensional exposures, mediators, and health outcomes

Jul. 25th 2018

Presenter: Jai Woo Lee

Graphs - Systemic Networks

Nodes: elements, metabolites, genes, or sequence positions

Edges: Interactions, regulations, activations or inhibitions

Weight: Strength

Other conditions: Denseness or Sparseness, disconnected components, or walks
Theoretical Problem – Mathematical Games
(With Math & CS Prof. Alan Frieze at CMU)

Rules:
1) Two players take alternative turns
2) Each player can start on any node
3) Each player can move to adjacent nodes of the current node
4) The player who removed final chips in the network wins
5) The upper bound of chips which each player can remove is fixed

Problems:
1) Design the strategy the first player wins.
2) Prove that the strategy is correct

Application Problem – Travelling Salesman Problem
(With Business Prof. John Hooker at CMU)

Problem:
Find the shortest trip not violating time limits for each city

Methods:
1) Linear Programs with AMPL (A Mathematical Programming Language)
2) Parallel computing for a large problem
My Current Research

1) Exposomics (placental elements)
2) Metabolomics (cord blood metabolites)
3) Health outcome (birth-weight)

Informatics (Networks Science, Data integration, and Data Science)

Background

**Placental Elements**: A vital role in biological systems involved in major processes as oxygen transport, or as catalysts like enzymes, in many metabolic reactions

**Metabolomics**: High-throughput analysis of metabolites, Functional end-point of physiology
Aim 1

Partial Correlation

It’s raining outside

My bike in the front yard is wet due to rain

My car in the front yard is wet due to rain

It’s raining outside

My bike in the front yard is wet due to rain

My car in the front yard is wet due to rain
Aim 1: Problem Description

Zn

Cd  ?  Mn

Aim 1: Methods (key word: Weight)

Note: In step 3, weight hubs or nodes of biological interest
Aim1: Definition- hub

Why is “node A” a hub in this network?

The degree (the number of neighbors) of each node should be checked.

\[
\begin{align*}
\text{deg}(A) &= 3 \quad \text{<- Max!} \\
\text{deg}(B) &= 1 \\
\text{deg}(C) &= 1 \\
\text{deg}(D) &= 1 \\
\text{deg}(E) &= 1 \\
\text{deg}(F) &= 1 \\
\text{deg}(G) &= 0
\end{align*}
\]

Aim1: Simulated Data

Define 40 features with 120 samples with following conditions
a) Define 3-5 hub nodes with 8-10 edges
b) Define 10-15 nodes with 1-2 edges
c) All the other nodes have no edges (totally disconnected)
d) Using eigenvalues and eigenvectors, generate the simulated data
Aim 1: Simulation result 1

Why/How does AIC outperform BIC and CV in this problem?

1) High sample size makes BIC penalize harsh
2) Low feature size makes CV less effective (vs. Gene expressions)

Aim 1: Simulation result 2

Scenario 1: Weight only hubs
Scenario 2: Weight hubs and non-hubs
Scenario 3: Weight only non-hubs
Scenario 4: Weight randomly chosen hubs

These scenarios were tested after testing other parameters such as sample size, the number of hubs, density of network and so on.
Aim1: Real Data application result 1

As, Cd, and ZN are 3-weighted

Aim1: Real Data application result 2

3-weighted female babies

3-weighted male babies
Aim 1: Future Direction

1) Can we get a tight bound of weight values for hubs?
   (a more mathematically rigorous bound?)

2) This method can be applied to other observed data for babies at transition?

3) This project was accepted and published as a paper,
   “Penalized Estimation of sparse concentration matrices based on prior knowledge with applications to placenta elemental data”
   Jai Woo Lee, Tracy Punshon, Erika L. Moen, Margaret R. Karagas, and Jiang Gui in Computational Biology and Chemistry journal.

Aim 2
Aim 2: Ideas
Keyword= Covariate

Aim 2: Problem Description
“Placental elements as covariates”
Aim2: Method – step 1
Gaussian Graphical model <-> Ising Model

In this complete sub-graph, correlation values on edge indicate closeness.

To do:
Negate or get reciprocal values of weights and apply travelling salesman problem. Then, we get the shortest path cycle
### Aim 2: Simulation results 1

**p:** features, **d:** degree of covariates, **c:** covariates, **n:** samples

```
<table>
<thead>
<tr>
<th>(p, d, c, n)</th>
<th>Method</th>
<th>Adjustment effect</th>
<th>True Positive</th>
<th>True Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1 (200, 10, 3, 100)</td>
<td>Adjusted Ising</td>
<td>0%</td>
<td>67%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>ANTAC</td>
<td>0%</td>
<td>83%</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>Adjusted GGM</td>
<td>83%</td>
<td>100%</td>
<td>99%</td>
</tr>
<tr>
<td>Model 2 (200, 10, 3, 200)</td>
<td>Adjusted Ising</td>
<td>11%</td>
<td>73%</td>
<td>96%</td>
</tr>
<tr>
<td></td>
<td>ANTAC</td>
<td>0%</td>
<td>93%</td>
<td>99%</td>
</tr>
<tr>
<td></td>
<td>Adjusted GGM</td>
<td>93%</td>
<td>100%</td>
<td>99%</td>
</tr>
<tr>
<td>Model 3 (200, 10, 3, 400)</td>
<td>Adjusted Ising</td>
<td>17%</td>
<td>67%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>ANTAC</td>
<td>0%</td>
<td>99%</td>
<td>96%</td>
</tr>
<tr>
<td></td>
<td>Adjusted GGM</td>
<td>100%</td>
<td>99%</td>
<td>99%</td>
</tr>
<tr>
<td>Model 4 (200, 40, 3, 200)</td>
<td>Adjusted Ising</td>
<td>0%</td>
<td>70%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>ANTAC</td>
<td>0%</td>
<td>88%</td>
<td>92%</td>
</tr>
<tr>
<td></td>
<td>Adjusted GGM</td>
<td>73%</td>
<td>93%</td>
<td>97%</td>
</tr>
<tr>
<td>Model 5 (200, 40, 3, 400)</td>
<td>Adjusted Ising</td>
<td>0%</td>
<td>67%</td>
<td>98%</td>
</tr>
<tr>
<td></td>
<td>ANTAC</td>
<td>0%</td>
<td>99%</td>
<td>97%</td>
</tr>
<tr>
<td></td>
<td>Adjusted GGM</td>
<td>86%</td>
<td>96%</td>
<td>98%</td>
</tr>
</tbody>
</table>
```
Aim2: Simulation results 3

![Diagram](https://via.placeholder.com/150)

\[
\begin{pmatrix}
0.000 & 0.269 & 0.168 & 0.201 & 0.291 \\
0.255 & 0.000 & 0.746 & 0.669 & 0.793 \\
0.178 & 0.743 & 0.000 & 0.826 & 0.327 \\
0.212 & 0.624 & 0.832 & 0.000 & 0.308 \\
0.332 & 0.783 & 0.321 & 0.342 & 0.000
\end{pmatrix}
\]

M1 -> M3 -> M2 -> M5 -> M4 -> M1
(The Shortest tour or the most correlated tour)

Aim2: Real Data application

1) Real data application?
Fecal Metabolomics data by Biocrates
(Data processing is required)

2) Which metal covariates should I choose?
   a) biomedically? As, Cd, and Hg
   b) statistically? Apply Principal Component Analysis on Placental Elements Data
Aim 3: Ideas
Keyword: Group

Dense!

Placental Elements

Birthweight
Aim 3: Problem Description

Can we use dense sub-networks, possibly cliques, to estimate one vector of continuous values?

Aim 3: Method

- Step 1: Construct the network using Lasso to estimate the network after using AIC to pick tuning parameter.

- Step 2: Identify “clique”. We define clique to be a complete sub network.

- Step 3. Consider cliques as groups and apply various group Lasso methods to find the best one

- Step 4. Fit a liner regression model with cross-validation to estimate outcome y.
Aim3: Simulation Results 1

After Fitting a liner regression model with cross-validation to estimate outcome y “on 40 nodes, make three cliques of size 25, 10, 5”

<table>
<thead>
<tr>
<th>Grouping Methods</th>
<th>Error rate ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>grLasso: Group lasso (Yuan and Lin, 2006)</td>
<td>1.0012</td>
</tr>
<tr>
<td>grMCP: Group MCP (minimax concave penalty)</td>
<td>1</td>
</tr>
<tr>
<td>grSCAD: Group SCAD (smoothly clipped absolute deviation)</td>
<td>1.0136</td>
</tr>
<tr>
<td>cMCP: composite minimax concave penalty</td>
<td>1.0103</td>
</tr>
<tr>
<td>gel: Group exponential lasso (Breheny, 2015)</td>
<td>1.0054</td>
</tr>
</tbody>
</table>

Aim3: Simulation Results 2

Using grMCP to fit a liner regression model with cross-validation to estimate outcome y “on 40 nodes, make three cliques of size 25, 10, 5”

Only maximal cliques (error= 1)
Smaller cliques (eight cliques of size 5) (error=1.02)
No cliques (error=1.06)

=> Using maximal cliques gives the best results
Aim3: Real data application

Group 1 (11 elements): Zn, Ca, Cu, Na, Sr, Se, P, K, Ba, Hg, Sb
Group 2 (4 elements): Hg, Na, Mg, Si

Aim3: Future direction – Another Problem

Metagenomics & Disease Status

a) Continuous data => dichotomized data
b) Gaussian Graphical Model => Ising Model
c) linear regression => Logistic Regression
Aim X: Networks, Pathways, and Mediation analysis

![Diagram showing metabolic pathway]

Aim X: Software Development

Codes for generating exposure-outcome, exposure-mediator, and mediator-outcomes were completed.

From this step, I can try two things.
1) Exposure -> Mediator -> health outcome  (Shortest path analysis)

2) Exposure -> outcome
   vs.
   Exposure -> mediators -> outcome
   (Mediation analysis)
Acknowledgements

Dissertation Committee
Dr. Jiang Gui (co-mentor)
Dr. Margaret R. Karagas (co-mentor)
Dr. Megan E. Romano
Dr. Hongzhe Li (Upenn)

Qualifying Examination Committee
Dr. Brock C. Christensen
Dr. Anne G. Hoen
Dr. Zhigang Li