Sparse Gaussian Chain Graph Models

Calvin McCarter & Seyoung Kim

Background on Gaussian Chain Graph Models

Chain graph models
- Given partition \( \{x_1, \ldots, x_d\} \), where \( x_d \in \mathbb{R}^{|x_d|} \)
- \( p(x) = \prod_{d=1}^{d} p(x_d | \text{pa}(x_d)) \)
- CRF as chain component model:
  - Directed edges: \( x_{\text{pa}}(x) \rightarrow x_{x} \)
  - Undirected edges: \( x_{x} \)

- Conditional independencies from moralized graph:

Non-adjacent variables in moralized graph are conditionally independent given all other variables.

Sparse Two-Layer Gaussian Chain Graph Models

\[ p(y, x) = p(y|x)p(x) \]

\[ = \left( \exp\left(-\frac{1}{2} x^T \Theta_{xy} x - x^T \Theta_{xy} y / A(x_\text{pa}(y)) \right) \right) \left( \exp\left(-\frac{1}{2} x^T \Theta_{xx} x / A_2 \right) \right) \]

\[ = N \left( -\Theta_{xy}^{-1} \Theta_{xy} x, \Theta_{xx} \right) \left( \Theta_{xy}^{-1} \right) \]

inference

Functional mapping between modules

Sparse Multi-Layer Gaussian Chain Graph Models for Integrative Genomic Data Analysis

Learn cascades of networks with multiple data types instead of a single network from gene expression data

Advantages of CGGMs as Chain Component Models

<table>
<thead>
<tr>
<th>Chain Component Model</th>
<th>Sparse Multivariate Regression</th>
<th>Sparse CGGM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization</td>
<td>Bi-convex</td>
<td>Convex</td>
</tr>
<tr>
<td>Computation time</td>
<td>Slow</td>
<td>Fast</td>
</tr>
<tr>
<td>Structured sparsity</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Leverage model structure for semi-supervised learning?</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Simulation Results

Better graph structure recovery and prediction accuracy, regardless of true component model!

- Problem size: 500 x's, 100 y's, 50 o's
- 400 training samples with 200 samples missing y's

Linear Regression-based True Component Model

Precision/recall curves for graph structure recovery

Recovery of structured sparsity using CGGM-based component model

Integrative Genomic Data Analysis

- 3 layer chain graph model.
- 1000 SNPs, 200 gene expressions, and 100 phenotypes
- From pancreatic islets study for diabetically mice.
- 306 training samples, 100 validation samples, 100 test samples
- Gene expression data missing for 150 mice.