The Nonparanormal SKEPTIC and Its Application
Outline

• The Nonparanormal SKEPTIC
• inferring biochemical networks
The Nonparanormal SKEPTIC

Han Liu
Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205 USA
HANLIU@CS.JHU.EDU

Fang Han
Johns Hopkins University, 615 N. Wolfe Street, Baltimore, MD 21205 USA
FHAN@JHSPH.EDU

Ming Yuan
Georgia Institute of Technology, 765 Ferst Drive, NW Atlanta, GA 30332 USA
MYUAN@ISYE.GATECH.EDU

John Lafferty
University of Chicago, 5734 S. University Avenue, Chicago, IL 60637 USA
LAFFERTY@UCHICAGO.EDU

Larry Wasserman
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213 USA
LARRY@STAT.CMU.EDU
the precision matrix

- inverse of the covariance matrix
- Θ
- if the data is multivariate normal:

<table>
<thead>
<tr>
<th>node</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>~</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>~</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>~</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>~</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>~</td>
<td>0</td>
<td>0</td>
<td>~</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>~</td>
<td>0</td>
<td>~</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>~</td>
<td>0</td>
</tr>
</tbody>
</table>
2 problems

• dimension >> # observations
• data is not multivariate normal
dimension $>>$ # observations

- log likelihood

$$\log \det \Theta - \text{tr}(S\Theta) - \text{(terms involving the mean)}$$

Max

$$\log \det \Theta - \text{tr}(S\Theta) - \rho \|\Theta\|_1$$
data is not multivariate normal

- trick 1 (the nonparanormal):

\[ f(X) := (f_1(X_1), \ldots, f_d(X_d))^T \sim N(0, \Sigma^0) \]


- trick 2 (nonparametric correlation):

\[
\hat{\rho}_{jk} = \frac{\sum_{i=1}^{n} (r^i_j - \bar{r}_j)(r^i_k - \bar{r}_k)}{\sqrt{\sum_{i=1}^{n} (r^i_j - \bar{r}_j)^2 \cdot \sum_{i=1}^{n} (r^i_k - \bar{r}_k)^2}}
\]

\[
\hat{S}_{jk}^\rho = 2 \sin \left( \frac{\pi}{6} \hat{\rho}_{jk} \right)
\]
\[ \log \det \Theta - \operatorname{tr}(S\Theta) - \rho \|\Theta\|_1 \]

\[ \hat{\Omega} = \arg \min_{\Omega \geq 0} \left\{ \operatorname{tr}(\hat{S}\Omega) - \log |\Omega| + \lambda \sum_{j \neq k} |\Omega_{jk}| \right\} \]
Pristionchus pacificus

- satellite model organism of *C. elegans*
- necromenic association with Scarab beetles
- global distribution
  - diverse habitats
  - diverse but structured genetic background

Collaboration: Ralf J. Sommer, Director, MPI, Tuebingen, Germany
data set

• ~450 strains
• 2 replicates each
• positive and negative ionization high resolution lcms (metabolome)
• restriction site associated dna maker snp calls (genome)
snp data set

<table>
<thead>
<tr>
<th></th>
<th>snp_locus_1</th>
<th>snp_locus_2</th>
<th>snp_locus_3</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample_3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1% genomic coverage

# alleles | count
---|---
1 | 194
2 | 2947
3 | 1

Minor Allele Frequencies

SNP Number

0.0 0.1 0.2 0.3 0.4 0.5
0 500 1000 1500 2000 2500 3000
liquid chromatography coupled mass spectrometry (lcms)

chromatography column

mass spectrometer

total ion chromatogram

column: http://www.waters.com/webassets/cms/category/media/snapshot/ACQUITY_Column.jpg
mass spectrometer: https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcSJGwVjgNgUcS9gVxipz6-wrL5jrVypj09BYwFnlfvHGSfXXdg
<table>
<thead>
<tr>
<th></th>
<th>peak_1 (m,rt)</th>
<th>peak_2 (m,rt)</th>
<th>peak_3 (m,rt)</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sample_3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

~2,000 features
ascaroside centric metabolic network
ascaroside centric metabolic network

<table>
<thead>
<tr>
<th>Start Node</th>
<th>End Node</th>
<th>Shortest Path</th>
<th>Shortest Path To Random Node From Start Node</th>
<th>Shortest Path To Random Node From End Node</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ascr#9</td>
<td>pasc#12</td>
<td>1</td>
<td>9.18</td>
<td>10.74</td>
<td>-0.128061447</td>
</tr>
<tr>
<td>pasc#9</td>
<td>pasc#12</td>
<td>2</td>
<td>13.28</td>
<td>10.88</td>
<td>-0.076858659</td>
</tr>
<tr>
<td>ascr#9</td>
<td>pasc#9</td>
<td>3</td>
<td>9.6</td>
<td>12.52</td>
<td>-0.626094706</td>
</tr>
</tbody>
</table>
advantages of this method

• requires no prior knowledge
• unsupervised
• group wise interference
• generalizable
• efficient
• functional