Information Decomposition on Structured Space

Mahito Sugiyama (Osaka Univ.)
Hiroyuki Nakahara (RIKEN), Koji Tsuda (UTokyo)
Contributions

• We build information geometry for posets (partially ordered sets)
  – Decomposition of KL divergence

• Key observations:
  – \( \theta \)-coordinate \( \rightarrow \) principal ideals (lower sets) \( \rightarrow \) \( p \)-coordinate
    ○ \( \theta \)-coordinate: coefficients of a log-linear model
    ○ \( p \)-coordinate: probabilities
  – \( p \)-coordinate \( \rightarrow \) principal filters (upper sets) \( \rightarrow \) \( \eta \)-coordinate
    ○ \( \eta \)-coordinate: frequencies (sufficient statistics)

• Code: https://git.io/decomp
Summary

Probability distribution on posets (partially ordered sets)

Decomposition in the log-linear model

Information geometry

\[ \log p(x) = \sum \theta(s) \]
Summary

Probability distribution on posets (partially ordered sets)

\[ \log p(x) = \sum \theta(s) \]

Decomposition in the log-linear model

Numerical score (KL divergence) and the \( p \)-value for higher-order interactions

\[
\begin{align*}
0 & 0 & 1 \\
1 & 0 & 0 \\
1 & 1 & 1 \\
0 & 0 & 0 \\
1 & 1 & 0 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 1 & 0
\end{align*}
\]
Transaction database

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0
ID 4: 1 1 1
ID 5: 1 1 0
ID 6: 1 0 1
ID 7: 1 0 1
ID 8: 1 1 1
ID 9: 1 0 0
ID 10: 0 1 0

Itemset lattice

Frequency = 0.3
**Transaction database**

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ID10</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Itemset lattice**

- Frequency = 0.3
- Probability = 0.3
Upward = Pattern mining

Itemset lattice

η: Frequency
p: Probability

η(\{\text{●}, \text{●}\}) = p(\{\text{●}, \text{●}\}) + p(\{\text{●}, \text{●}, \text{●}\})
Upward = Pattern mining
Downward = Log-linear analysis

η: Frequency
p: Probability
θ: Coefficient of log-linear model

η({
  }{ 
  }) = p({
  }{ 
  }) + p({
  }{ 
  }{ 
  })

log p({
  }{ 
  }) = θ({
  }{ 
  }) + θ({
  }) + θ({
  }) + θ(∅)
\[ \log p(x) = \sum_{s \leq x} \theta(s) \]
\[ \log p(x) = \sum_{s \leq x} \theta(s) \]  

Exponential family: \[ p(x) = \exp\left( \sum \theta(s)F_s(x) - \psi(\theta) \right) \]
\[ \eta(x) = \sum_{s \geq x} p(s) \]

\[ \eta(x) = \mathbb{E}[F_x(s)] \]

Sufficient statistics of exponential family

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]

Natural parameter

Exponential family:

\[ p(x) = \exp\left( \sum \theta(s) F_s(x) - \psi(\theta) \right) \]
Triple for each node

\[ \eta(x) = \sum_{s \geq x} p(s) \]

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]
Triple for each node

\[ \begin{align*}
\eta(x) &= \sum_{s \geq x} p(s) \\
\log p(x) &= \sum_{s \leq x} \theta(s)
\end{align*} \]
Triple for each node

\[ p \]
\[ \eta \]
\[ \theta \]

\{\textcolor{blue}{\bigcirc}, \textcolor{red}{\bigcirc}\}

\textcolor{blue}{\bigcirc} 0.2

\textcolor{red}{\bigcirc} -1.79

\textcolor{blue}{\bigcirc} 0.2

\\{\textcolor{blue}{\bigcirc}\} 0.3

\\{\textcolor{red}{\bigcirc}\} 0.5

\textcolor{blue}{\bigcirc} 0.1

\textcolor{red}{\bigcirc} 1.0

\emptyset 1.10

\\{\textcolor{blue}{\bigcirc}\} 0.1

\\{\textcolor{red}{\bigcirc}\} 0.2

\\{\textcolor{blue}{\bigcirc}, \textcolor{red}{\bigcirc}\} -2.30

\{\textcolor{blue}{\bigcirc}\} 0.4

\{\textcolor{red}{\bigcirc}\} 0.6

\{\textcolor{blue}{\bigcirc}, \textcolor{red}{\bigcirc}\} 1.39

Probability distribution is a “point” in 3D space

\[ p(\{\textcolor{red}{\bigcirc}\}) \]

Probability distribution
is a “point” in 3D space

\[ p(\{\textcolor{blue}{\bigcirc}\}) \]

\begin{align*}
\eta(x) &= \sum_{s \geq x} p(s) \\
\log p(x) &= \sum_{s \leq x} \theta(s)
\end{align*}
Triple for each node

\[ p, \eta, \theta \]

\[
\begin{align*}
\{ \bullet, \circ \} & : 0.2 \quad 0.2 \\
\{ \bullet \} & : 0.3 \quad 0.5 \\
\{ \circ \} & : 0.4 \quad 0.6 \\
\emptyset & : 0.1 \quad 1.0 \\
\end{align*}
\]

\[
\eta(\{ \bullet \}) = 0.5 \\
\eta(\{ \bullet, \circ \}) = 0.6 \\
\eta(\{ \circ \}) = 0.2
\]

Probability distribution is a “point” in 3D space

\[
\begin{align*}
\eta(x) &= \sum_{s \geq x} p(s) \\
\log p(x) &= \sum_{s \leq x} \theta(s)
\end{align*}
\]
Triple for each node

\[
\begin{align*}
\eta(x) &= \sum_{s \geq x} p(s) \\
\log p(x) &= \sum_{s \leq x} \theta(s)
\end{align*}
\]
Triple for each node

\[ p \quad \eta \quad \theta \]

\{ \bullet, \bullet \} \quad \{ \bullet \} \quad \{ \bullet \}

0.2 \quad 0.2 \quad -1.79

0.3 \quad 0.5 \quad 1.10

0.2 \quad -1.79 \quad 0.1

0.4 \quad 0.6 \quad 1.39

0.1 \quad 1.0 \quad -2.30

\log p(x) = \sum_{s \leq x} \theta(s)

\eta(x) = \sum_{s \geq x} p(s)

one-to-one
Triple for each node

\[
\begin{align*}
\eta(x) &= \sum_{s \geq x} p(s) \\
\log p(x) &= \sum_{s \leq x} \theta(s)
\end{align*}
\]

\(\theta\) and \(\eta\) are dually orthogonal.

One-to-one

\(p\)

\(\eta\)

\(\theta\)
Dist. $P$

\[
\begin{array}{c}
p \\
\eta \\
\theta \\
\end{array}
\]

\[
\begin{array}{ccc}
0.3 & 0.5 & 1.10 \\
0.2 & 0.2 & -1.79 \\
0.4 & 0.6 & 1.39 \\
\end{array}
\]

Dist. $Q$

\[
\begin{array}{c}
p \\
\eta \\
\theta \\
\end{array}
\]

\[
\begin{array}{ccc}
0.7 & 0.8 & 1.95 \\
0.1 & 0.1 & -1.95 \\
0.1 & 0.2 & 0.0 \\
\end{array}
\]
Mixed distribution $R$

Dist. $P$

Dist. $Q$

$\rho$

$\eta$

$\theta$

Choose $\eta$

Choose $\theta$

Dist. $P$

Dist. $Q$
KL(P, Q) = KL(P, R) + KL(R, Q)

Nonnegative decomposition of the KL divergence
**Log-linear model**

\[
\log p(x) = \sum_{s \leq x} \theta(s)
\]
Dist. $P$

---

Uniform dist. $P_0$

---

Contribution of the node $= KL(P, R) = 0.086$

---

Log-linear model

$$
\log p(x) = \sum_{s \leq x} \theta(s)
$$
The statistics $\lambda$:

$$\lambda = 2 \cdot [\text{sample size}] \cdot KL(P, R)$$

follows $\chi^2$-distribution with d.f. $[\#\text{nodes} - 1]$

$\Rightarrow p$-value can be obtained!
Make a Poset from Data

Dataset

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0
ID 4: 1 1 1
ID 5: 1 1 0
ID 6: 1 0 1
ID 7: 1 0 1
ID 8: 1 1 1
ID 9: 1 0 0
ID10: 0 1 0

Frequency = 0.3
Probability = 0.3

Number of nodes = 2

⇒ combinatorial explosion!
Make a Poset from Data

Dataset

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0
ID 4: 1 1 1
ID 5: 1 1 0
ID 6: 1 0 1
ID 7: 1 0 1
ID 8: 1 1 1
ID 9: 1 0 0
ID10: 0 1 0

Frequency = 0.3

Probability = 0.3

Probability ≥ 0.2
(user specified threshold)
Remove Nodes with Probability 0

Dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ID 6</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID 7</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ID 8</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ID 9</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ID10</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
\emptyset = 1 - \sum p(s)
\]

\[
\begin{align*}
p &= 0.2 \\
\eta &= 0.3 \\
\theta &= 0.6 \\
p^2 &= 0.0 \\
\end{align*}
\]

\[
\begin{align*}
0.405 &= 0.3 + 0.3 + 0.0 + 0.6 \\
-1.61 &= 0.3 + 0.3 + 0.0 + 0.6 + 0.0 + 1.0
\end{align*}
\]
Example on Real Data (kosarak)

- **Sample size:** 990,002
- **# nodes:** 3,253 (Threshold: $10^{-5}$)
- **# features:** 41,270
- **# nodes:** 3,253
- **Total runtime:** 4.95 seconds
  - Single feature: 537
  - Pairwise interactions: 41
  - Triple interactions: 5

```plaintext
ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0 ...
ID 4: 1 1 1
ID 5: 1 1 0
...:
```

# significant interactions: 583
Example on Real Data (accidents)

ID 1: 1 1 0
ID 2: 1 1 1
ID 3: 1 1 0 ...
ID 4: 1 1 1
ID 5: 1 1 0

Sample size: 340,183
# nodes: 281
(Threshold: $5 \times 10^{-6}$)

# features: 468
# nodes: 281
# significant interactions: 280
# features in each interaction is between 26 to 41

Total runtime: 4.95 seconds
Conclusion

• We build information geometry for posets (partially ordered sets)
  – Natural connection between the information geometric dual coordinates and the partial order structure
  – Code: https://git.io/decomp

• We can decompose a probability distribution and assess the significance of any-order interactions

• Related papers: