Partial Order Structure and Information Geometry
(順序構造と情報幾何)

Mahito Sugiyama (ISIR, Osaka University, PRESTO)
(杉山 麗人; 大阪大学産業科学研究所, JSTさきがけ)
Today’s Model on Poset \((S, \leq)\)

\[
\log p(x) = \sum_{s \in S} \zeta(s, x) \theta(s)
\]

\[
p(x) = \sum_{s \in S} \mu(x, s) \eta(s)
\]
Today’s Model on Poset \((S, \leq)\)

\[
\log p(x) = \sum_{s \in S} \zeta(s, x) \theta(s)
\]

\[
p(x) = \sum_{s \in S} \mu(x, s) \eta(s)
\]

- **Probability**
- **Zeta function**
- **Möbius function**
- **Expectation**
  - (Frequency in pattern mining)
  - (Sufficient statistics in exponential family)

Coefficient of log-linear model
(Bias/weight in Boltzmann machines)
(Natural parameter of exponential family)
Outcome

• Given a poset \((S, \leq)\) and consider distributions on \(S\)
  – The least element \(\bot \in S\) is assumed

1. KL divergence decomposition:

\[
\]

with \(Q\) s.t. \(\theta_Q(x) = \theta_R(x)\) or \(\eta_Q(x) = \eta_P(x)\) for all \(x \in S \setminus \{\bot\}\)

2. The set of probability distributions on \((S, \leq)\) is a dually flat manifold w.r.t. \(\theta\) and \(\eta\)
  – \(p, \theta,\) and \(\eta\) are coordinate systems
  – \(\theta\) and \(\eta\) are orthogonal
  – \(\theta\) introduces the structure of exponential family
  – \(\eta\) introduces the structure of mixture family
Partially Ordered Sets

\{x, y, z\} \quad \text{Power set}

\{x, y\} \quad \{x, z\} \quad \{y, z\}

\{x\} \quad \{y\} \quad \{z\}

\emptyset

Power set
Partially Ordered Sets

\[ \{x, y, z\} \]
\[ \{x, y\} \]
\[ \{x, z\} \]
\[ \{y, z\} \]
\[ \{x\} \]
\[ \{y\} \]
\[ \{z\} \]

Power set

Positive integers

\[ \emptyset \]
Partially Ordered Sets

\[
\{x, y, z\} \\
\{x, y\} \quad \{x, z\} \quad \{y, z\} \\
\{x\} \quad \{y\} \quad \{z\} \\
\emptyset
\]

Power set

Positive integers

\[
0 \quad 1 \quad \lambda
\]

Prefixes

\[
000 \quad 001 \quad 010 \quad 011 \quad 100 \quad 101 \quad 110 \quad 111
\]

3/39
Partially Ordered Sets

{\{x, y, z\}\n\{x, y\}\n\{x, z\}\n\{y, z\}\n\{x\}\n\{y\}\n\{z\}\n\{\emptyset\}\n
Power set

Directed Acyclic Graph

Predecessor graph

Positive integers

0
1
2
3

\lambda

000
001
010
011
100
101
110
111

0
1

00
01
10
11

λ

3/39
Probability distribution on posets (partially ordered sets)
Posets with Probability Distribution

Probability distribution on posets (partially ordered sets)

Information geometry

Decomposition in the log-linear model

\[ \log p(x) = \sum \zeta(s, x)\theta(s) \]
Probability distribution on posets (partially ordered sets)

Decomposition in the log-linear model

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

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

\[
\begin{array}{ccc}
0 & 0 & 1 \\
1 & 0 & 0 \\
1 & 1 & 1 \\
0 & 0 & 0 \\
1 & 1 & 0 \\
0 & 1 & 1 \\
1 & 0 & 1 \\
1 & 0 & 1 \\
1 & 1 & 0 \\
\end{array}
\]
Binary vectors (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

Poset (itemset lattice)
Binary vectors (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
ID10: 0 1 0

Poset (itemset lattice)

Frequency = 0.3
Binary vectors (Transaction database)

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>2</th>
<th>3</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>

Poset (itemset lattice)

Frequency = 0.3
Binary vectors (Transaction database)

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<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>

Poset (itemset lattice)

Frequency = 0.3
Probability = 0.3
Upward = Pattern mining

\[ \eta \text{: Frequency} \]
\[ p \text{: Probability} \]

\[ \eta(\{\bullet, \bullet\}) = p(\{\bullet, \bullet\}) + p(\{\bullet, \bullet, \bullet\}) \]
Upward = Pattern mining
Downward = Log-linear analysis

$\eta$: Frequency
$p$: Probability
$\theta$: Coefficient of log-linear model

$$\eta(\{\bullet, \circ\}) = p(\{\bullet, \circ\}) + p(\{\bullet, \circ, \star\})$$

$$\log p(\{\bullet, \circ\}) = \theta(\{\bullet, \circ\}) + \theta(\{\bullet\}) + \theta(\{\circ\}) + \theta(\emptyset)$$
\[ \log p(x) = \sum \zeta(s, x) \theta(s) \]
\[
\log p(x) = \sum \zeta(s, x) \theta(s)
\]

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

Natural parameter

e.g. Gaussian
\( \eta(x) = \sum \zeta(x, s)p(s) \)

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

Sufficient statistics of exponential family

\( \log p(x) = \sum \zeta(s, x)\theta(s) \)

Natural parameter

Exponential family:

\( p(x) = \exp(\sum \theta(s)F_s(x) - \psi(\theta)) \)

e.g. Gaussian
Möbius Inversion on Posets

- **Zeta function** $\zeta: S \times S \rightarrow \{0, 1\}$:

  $$\zeta(s, x) = \begin{cases} 
  1 & \text{if } s \leq x, \\ 
  0 & \text{otherwise}
  \end{cases}$$

- **Möbius function** $\mu: S \times S \rightarrow \mathbb{Z}$, defined as $\mu = \zeta^{-1}$:

  $$\mu(x, y) = \begin{cases} 
  1, & \text{if } x = y, \\
  -\sum_{x \leq s < y} \mu(x, s) & \text{if } x < y, \\
  0 & \text{otherwise}
  \end{cases}$$

- **The Möbius inversion formula** [Rota (1964)]:

  $$g(x) = \sum_{s \in S} \zeta(s, x)f(s) \iff f(x) = \sum_{s \in S} \mu(s, x)g(s)$$
Möbius Function Is Generalization of Inclusion-Exclusion Principle

- For sets $A, B, C$,
  \[
  |A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |B \cap C| - |A \cap C| + |A \cap B \cap C|
  \]

- In general, for $A_1, A_2, \ldots, A_n$,
  \[
  \left| \bigcup_{i} A_i \right| = \sum_{J \subseteq \{1, \ldots, n\}, J \neq \emptyset} (-1)^{|J|-1} \left| \bigcap_{j \in J} A_j \right|
  \]

- The Möbius function $\mu$ is the generalization of $"(-1)^{|J|-1}"$
Log-linear Model with Möbius Inversion

- Log-linear model and its sufficient statistics:

$$
\log p(x) = \sum_{s \in S} \zeta(s, x) \theta(s) = \sum_{s \leq x} \theta(s),
$$

$$
\eta(x) = \sum_{s \in S} \zeta(x, s) p(s) = \sum_{s \geq x} p(s)
$$

- Generalization of the log-linear model on binary vectors:

$$
\log p(x) = \sum_i \theta^i x^i + \sum_{i<j} \theta^{ij} x^i x^j + \cdots + \theta^{1\ldots n} x^1 x^2 \cdots x^n,
$$

- From the Möbius inversion formula,

$$
\theta(x) = \sum_{s \in S} \mu(s, x) \log p(s), \quad p(x) = \sum_{s \in S} \mu(x, s) \eta(s)
$$
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

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

\[
\{ \bullet, \bullet \}, \quad 0.2
\]

\[
\{ \bullet \}, \quad 0.2
\]

\[
\{ \bullet \}, \quad -1.79
\]

\[
\varnothing, \quad 0.1
\]

\[
0.3 \quad 0.4
\]

\[
0.5 \quad 0.6
\]

\[
1.10 \quad 1.39
\]

\[
-2.30
\]

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

\[
\eta(x) = \sum_{s \geq x} p(s)
\]
Triple for each node

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

Probability distribution is a "point" in 3D space
Triple for each node

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

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

Probability distribution is a “point” in 3D space
Triple for each node

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

Probability distribution is a “point” in 3D space
Triple for each node

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

\begin{align*}
\{ \bullet, \circ \} \\
0.3 & 0.2 & -1.79 \\
0.5 & 0.2 & \\
1.10 & 0.1 & 0.2 \quad \eta(x) = \sum_{s \geq x} p(s) \\
\emptyset & 0.1 & 0.4 \\
1.0 & 0.6 & 0.6 \quad \log p(x) = \sum_{s \leq x} \theta(s) \\
-2.30 & 1.39 & 1.39
\end{align*}

one-to-one

\[ p \, \eta \, \theta \]
Triple for each node

\[ p, \eta, \theta \]

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

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

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

One-to-one

\( \eta \leftrightarrow \theta \)
Orthogonality of $\theta$ and $\eta$

- From Möbius inversion,

$$\sum_{s \in S} \zeta(x, s) \mu(s, y) = \delta_{x,y}, \quad \delta_{x,y} = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise} \end{cases}$$

- $\theta$ and $\eta$ are dually orthogonal:

$$\mathbb{E} \left[ \frac{\partial}{\partial \theta(x)} \log p(s) \frac{\partial}{\partial \eta(y)} \log p(s) \right] = \sum_{s \in S} \zeta(x, s) \mu(s, y) = \delta_{x,y}$$

- Partial order structure leads to the same dually flat structure with the exponential family
Existing Approach Limited To Power Set

\[
\begin{align*}
\{x, y, z\} & \rightarrow \text{Power set} \\
\{x, y\} & \rightarrow \{x, z\} \rightarrow \{y, z\} \\
\{x\} & \rightarrow \{y\} \rightarrow \{z\} \\
\emptyset & \rightarrow \{\emptyset\}
\end{align*}
\]
Our Approach Applies To Any Posets

Subset of power set

{x, y, z}
{x, y}
{x}
{y, z}
{y}
∅
Our Approach Applies To Any Posets

Positive integers

Subset of power set

Directed Acyclic Graph

Prefixes
KL Divergence Decomposition

- KL divergence decomposition:


with \( Q \) s.t. \( \theta_Q(x) = \theta_R(x) \) or \( \eta_Q(x) = \eta_P(x) \) for all \( x \in S \setminus \{\perp\} \)

- \( Q \) is called the mixed distribution of \( (P, R) \)
- It is known as the (generalized) Pythagoras theorem in Information Geometry

- We can derive from Möbius inversion:

\[
D_{KL}[P, Q] + D_{KL}[Q, R] - D_{KL}[P, R] = \sum_{s \in S} (\eta_Q(s) - \eta_P(s)) (\theta_Q(s) - \theta_R(s))
\]
Dist. P

Mixed distribution Q

choose $\eta$

Dist. R

choose $\theta$
Mixed distribution $Q$
Mixed distribution $Q$

Mixed distribution $Q$ is decomposed nonnegatively.

Choose $\eta$ and $\theta$.
Nonnegative decomposition of the KL divergence

0.4390 = 0.3946 + 0.0444
Mixed distribution $Q$

Log-linear model

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

Dist. $P$

Uniform dist. $P_0$

choose $\eta$

choose $\theta$

(KNOCK DOWN)
Dist. $P$

Uniform dist. $P_0$

Contribution of the node
\[ = KL[P, Q] = 0.086 \]

Mixed distribution $R$

Log-linear model
\[ \log p(x) = \sum_{s \leq x} \theta(s) \]
Dist. \( P \)

\[ p \]

\[ \eta \]

\[ \theta \]

Uniform dist. \( P_0 \)

\[ 0.25 \]

\[ 0.25 \]

\[ 0.0 \]

\[ 0.25 \]

\[ 0.25 \]

\[ 0.0 \]

\[ 0.3 \]

\[ 0.3 \]

\[ 0.0 \]

The statistics \( \lambda \):

\[ \lambda = 2 \cdot [\text{sample size}] \cdot KL[P, Q] \]

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

\( \Rightarrow \) \( p \)-value can be obtained!

Log-linear model

\[ \log p(x) = \sum_{s \leq x} \theta(s) \]
Poset of Subgraphs
Log-Linear Model on Subgraphs

Log-linear model:
\[
\log p(x) = \sum_{s \subseteq x} \theta(s)
\]

Natural parameter of exponential family
Sufficient statistics of exponential family

\[
\eta(x) = \sum_{s \subseteq x} p(s)
\]
Information of Each Subgraph

\[ \eta_0 \eta_1 \eta_2 \eta_3 \eta_4 \eta_5 \eta_k \]

\[ \theta_0 \theta_1 \theta_2 \theta_3 \theta_4 \theta_5 \theta_k \]

\( P: \text{Empirical distribution} \)

Freq. \[
\begin{bmatrix}
\eta_0 & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5 & \ldots & \eta_k \\
\theta_0 & \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \ldots & \theta_k 
\end{bmatrix}
\]

\( \theta_i \rightarrow \eta_j \)
Information of Each Subgraph

\[ \eta_0 \eta_1 \eta_2 \eta_3 \eta_4 \eta_5 \eta_k \]

\[ \theta_0 \theta_1 \theta_2 \theta_3 \theta_4 \theta_5 \theta_k \]

\[ P: \text{Empirical distribution} \]
Freq. \[\eta_0 \eta_1 \eta_2 \eta_3 \eta_4 \eta_5 \eta_k\]
Coef. \[\theta_0 \theta_1 \theta_2 \theta_3 \theta_4 \theta_5 \theta_k\]

\[ Q: \text{Null distribution} \]
Freq. \[\eta_0 \eta_1 \eta_2 \eta_3 \eta_5 \eta_k\]
Coef. \[? ? ? ? 0 ? ?\]

\[ \text{KL}(P, Q) \]
Information of Each Subgraph

\[ P: \text{Empirical distribution} \]
\[
\begin{array}{cccccc}
\eta_0 & \eta_1 & \eta_2 & \eta_3 & \eta_4 & \eta_5 & \ldots & \eta_k \\
\theta_0 & \theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5 & \ldots & \theta_k \\
\end{array}
\]

\[ Q: \text{Null distribution} \]
\[
\begin{array}{cccccc}
\eta_0 & \eta_1 & \eta_2 & \eta_3 & ? & \eta_5 & \ldots & \eta_k \\
\end{array}
\]

\[ R: \text{Uniform distribution} \]
\[
\begin{array}{cccccc}
\theta'_0 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\
\end{array}
\]

\[ \text{KL}(P, Q) = \text{KL}(P, R) + \text{KL}(Q, R) \]
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

Number of nodes = 2

# features

⇒ combinatorial explosion!
Make a Poset from Data

Dataset

<table>
<thead>
<tr>
<th>ID</th>
<th>1</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<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>

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>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 1 0</td>
</tr>
<tr>
<td>2</td>
<td>1 1 1</td>
</tr>
<tr>
<td>3</td>
<td>1 1 0</td>
</tr>
<tr>
<td>4</td>
<td>1 1 1</td>
</tr>
<tr>
<td>5</td>
<td>1 1 0</td>
</tr>
<tr>
<td>6</td>
<td>1 0 1</td>
</tr>
<tr>
<td>7</td>
<td>1 0 1</td>
</tr>
<tr>
<td>8</td>
<td>1 1 1</td>
</tr>
<tr>
<td>9</td>
<td>1 0 0</td>
</tr>
<tr>
<td>10</td>
<td>0 1 0</td>
</tr>
</tbody>
</table>

\[ p \eta \theta = 1 - \sum p(s) \]
Example on Real Data (kosarak)

- # features: 41,270
- # nodes: 3,253 (Threshold: $10^{-5}$)
- Sample size: 990,002
- # significant interactions: 583
  - Single feature: 537
  - Pairwise interactions: 41
  - Triple interactions: 5

Total runtime: 4.95 seconds
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: $5 \times 10^{-6}$)
# significant interactions: 280
# features in each interaction is between 26 to 41

Total runtime: 4.95 seconds

# features: 468
Conclusion

• A close connection between the partial order structure and information geometry
  - Möbius inversion leads to the dually flat manifolds

• We can decompose the KL divergence and assess the significance on any posets