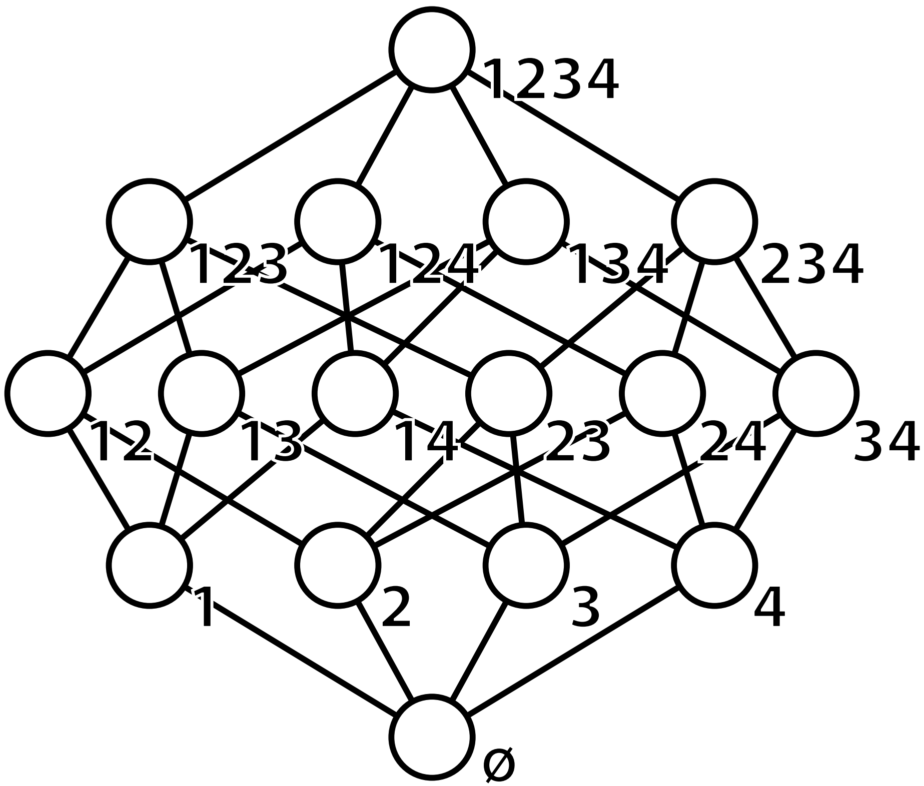
March 17, 2016
@RIKEN BSI

Statistical Analysis on Order Structures

Appendix

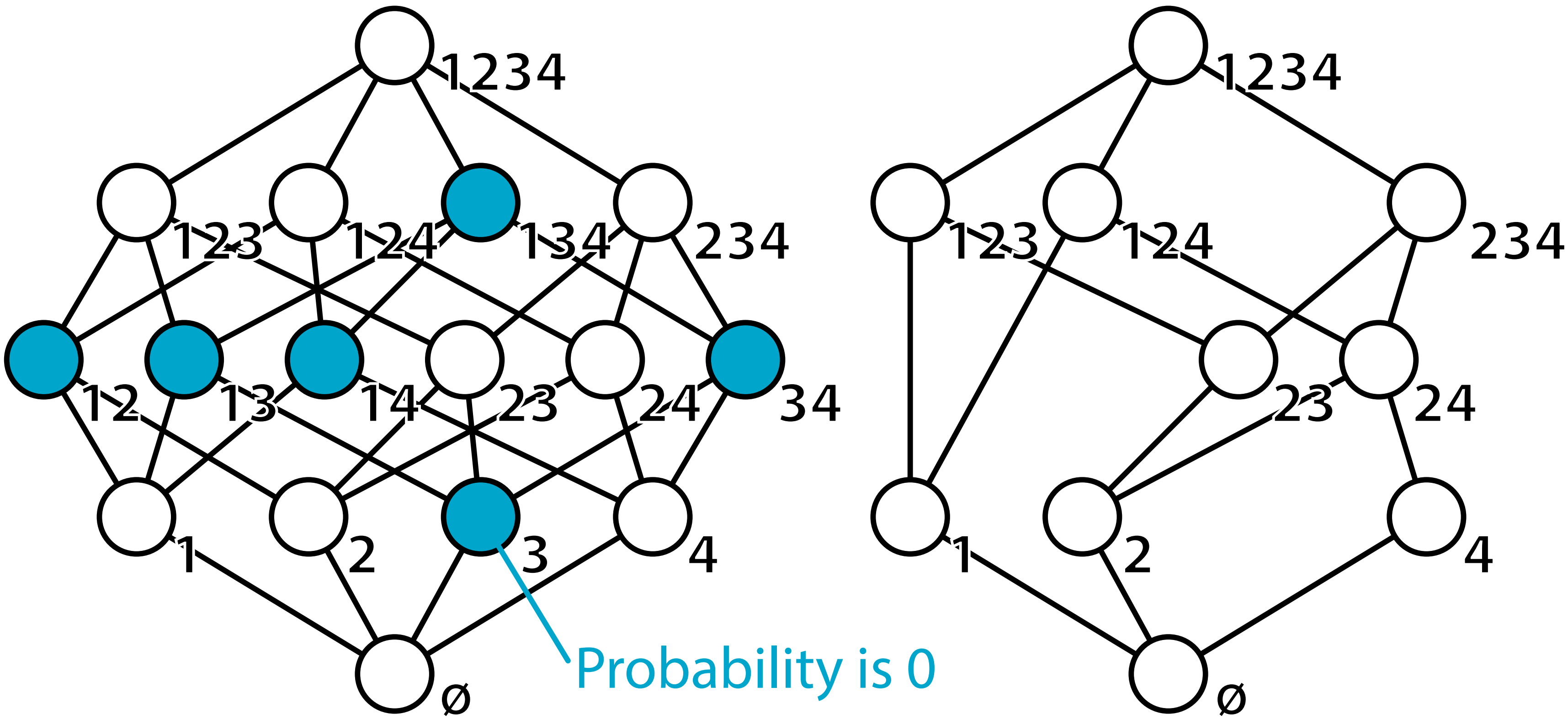
Mahito Sugiyama (ISIR, Osaka University)
（杉山 麿人; 大阪大学産業科学研究所）

Event Combinations of 4 Events



$$\log p(\boldsymbol{x}) = \sum_{i=1}^{4} \theta^{i} x_i + \sum_{i<j} \theta^{ij} x_i x_j + \cdots + \theta^{1234} x_1 x_2 x_3 x_4 - \psi$$

Event Combinations of 4 Events



$$\log p(\boldsymbol{x}) = \sum_{i=1}^{4} \theta^{i} x_i + \sum_{i<j} \theta^{ij} x_i x_j + \cdots + \theta^{1234} x_1 x_2 x_3 x_4 - \psi$$

Background

- Amari's orthogonal decomposition of probability distributions on the complete hierarchy of events
  - Theoretical basis for analyzing higher-order interactions
    - e.g. firing patterns of neurons, gene interactions, word associations in documents, ...
- **Problem**: The hierarchy is often incomplete
  - Some combinations might never occur
    - Combination of a person being male and a person having ovarian cancer can never occur
  - Lack of data; Estimating probabilities for 2^n combinations for n events is almost impossible

Main Results

- We build information geometry for a poset (partially ordered set) of variables

- Natural connection between the information geometric dual coordinates and the partial order structure
  - θ-coordinates → (principal ideal) → p-coordinates
  - θ-coordinates → (principal filter) → η-coordinates

- An efficient algorithm to decompose KL divergence and entropy on an incomplete hierarchy
  - For arbitrary probability distributions p and q on a poset,

    $$D_{\mathrm{KL}}(p, q) = D_{\mathrm{KL}}(p, r) + D_{\mathrm{KL}}(r, q)$$

    for a mixed distribution r of (p, q)

p-coordinate system

- Let $S = \{x_0, x_1, \ldots, x_n\}$
  - Assume that x_0 is the least element $\perp$ and $S^+ = S \setminus \{\perp\}$

- A discrete probability distribution p on S can be viewed as a vector:

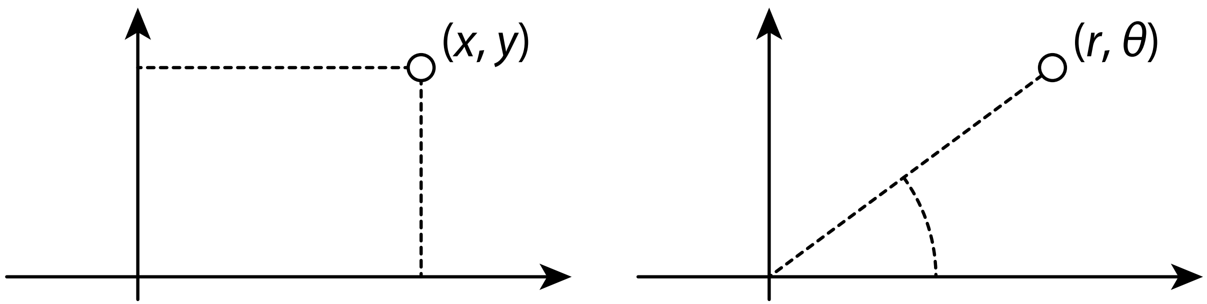
  $\boldsymbol{p} = (\,p(x_1), p(x_2), \ldots, p(x_n)\,)$ (p-coordinate system)

  - This corresponds to a "point" on n-dimensional space
    - There is a condition $\sum_{x \in S} p(x) = 1$
  - A probability distribution forms an n-dimensional manifold

$$\boldsymbol{\mathcal{S}} = \left\{ \boldsymbol{p} \;\middle|\; \forall x \in S.\, p(x) > 0, \sum_{x \in S} p(x) = 1 \right\}$$

Dual Coordinates on $\mathcal{S}$

- In information geometry, dual coordinate systems: θ-coordinate and η-coordinate, are known
  - They are realized as mappings $\theta : S \to \mathbb{R}$, $\eta : S \to \mathbb{R}$
  - θ (η) determines p, and vice versa
    - Analog to 2-dimensional Euclidean space:



- θ and η are dually orthogonal on $\mathcal{S}$

$$\mathbb{E}\left[\frac{\partial}{\partial\theta(s)}\log p(x,\theta)\frac{\partial}{\partial\eta(s')}\log p(x,\eta)\right] = \delta(s,s')$$

Exponential Family

- For a mapping $\theta{:}S \to \mathbb{R}$, the exponential family is

$$p(x;\theta) = \exp\left(\sum_{s \in S^+} \theta(s) F_s(x) - \psi(\theta) \right)$$

  - In Gaussian distribution, $\theta^1 = -\frac{1}{2\sigma^2}$, $\theta^2 = \frac{\mu^2}{\sigma^2}$

- Given a poset S, we propose to define $F_s(x)$ as

$$F_s(x) = \begin{cases} 1 & \text{if } s \le x, \\ 0 & \text{otherwise} \end{cases} \quad \text{and } \psi(\theta) = -\log p(\bot).$$

- We obtain the following log-linear model:

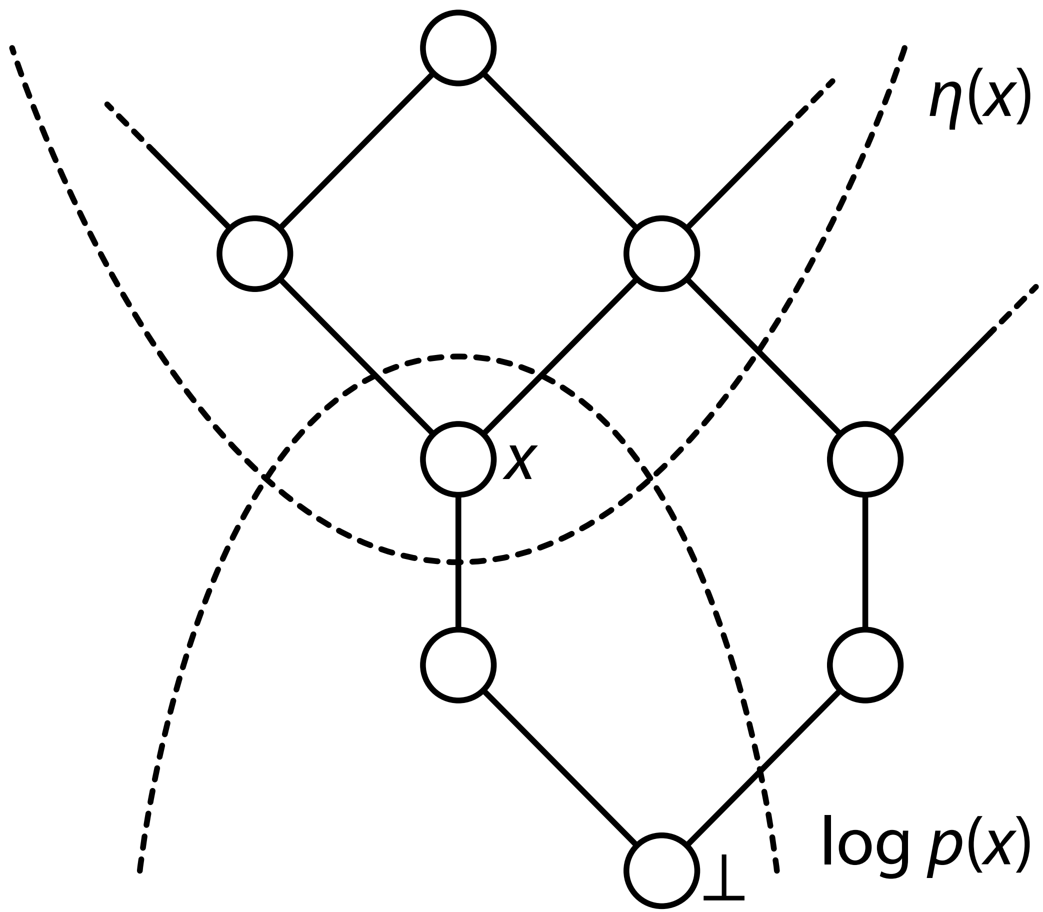
$$\log p(x) = \sum_{s \le x} \theta(s)$$

θ- and η-coordinate systems

- Given a probability distribution $p \in \mathcal{S}$, the θ-coordinate system is recursively computed as
  $$\theta(x) = \log p(x) - \sum_{s<x} \theta(s)$$
  starting from the bottom $\theta(\perp) = \log p(\perp)$

- η is given as the expectation of $F_s(x)$:
  $$\eta(s) = \mathbb{E}[F_s(x)] = \sum_{s \le x} p(x) = \Pr(X \ge s)$$
  - $\eta(x)$ is the support of x in pattern mining!

θ- and η-coordinate systems



$$\eta(x) = \sum_{x \le s} p(s)$$

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

Mixed Coordinate System

- A mixed coordinate system of θ and η
  - The key to decomposition of the KL divergence and entropy

- A mixed distribution $r \in \mathcal{S}$ of (p, q) w.r.t. $I \subseteq S^+$:

$$\begin{cases} \eta_r(x) = \eta_p(x) & \text{if } x \in S^+ \setminus I, \\ \theta_r(x) = \theta_q(x) & \text{if } x \in I, \end{cases}$$

and $r(\perp) = 1 - \sum_{s \in S^+} r(x)$

  - θ_p and η_p are θ- and η-coordinates of p, resp.
  - e.g.: $S^+ = \{1, 2, 3\}$, $I = \{1, 2\}$, then

$$\eta_p = (\, \eta_p(1), \eta_p(2), \eta_p(3) \,),$$

$$\theta_q = (\, \theta_q(1), \theta_q(2), \theta_q(3) \,),$$

mixed coordinate $r = (\, \theta_q(1), \theta_q(2), \eta_p(3) \,)$

KL divergence decomposition

- **[Theorem]** For two distributions p, q and any $I \subseteq S^+$,

  $D_{\mathrm{KL}}(p, q) = D_{\mathrm{KL}}(p, r) + D_{\mathrm{KL}}(r, q)$

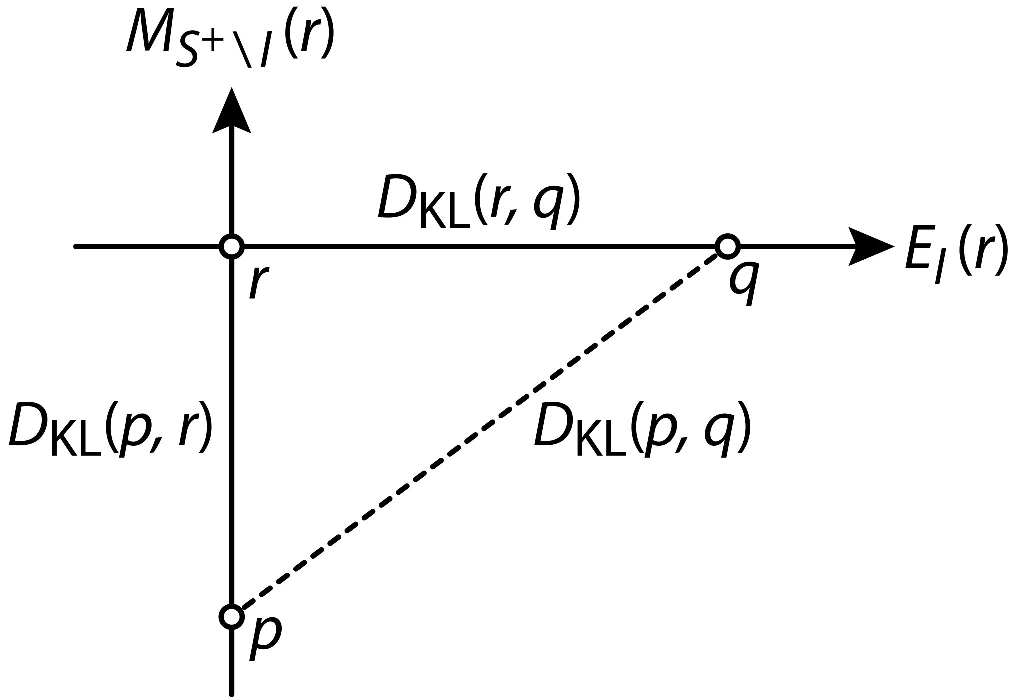
  for the mixed distribution r of (p, q) w.r.t. I

- **[corollary]** A hierarchical set $\{I_0, I_1, \ldots, I_k\}$ with$\emptyset = I_0 \subseteq I_1 \subseteq \cdots \subseteq I_k = S^+$,

  $D_{\mathrm{KL}}(p, q) = D_{\mathrm{KL}}(r_0, r_1) + D_{\mathrm{KL}}(r_1, r_2) + \cdots + D_{\mathrm{KL}}(r_{k-1}, r_k)$

  - r_i is the mixed dist. of (p, q) w.r.t I_i
  - $r_0 = q, r_k = q$

KL divergence decomposition



$$E_I(r) := \{\, v \in \mathcal{S} \mid \forall x \in I.\ \theta_v(x) = \theta_r(x) \,\}$$

$$M_{S^+\setminus I}(r) := \left\{\, v \in \mathcal{S} \mid \forall x \in S^+ \setminus I.\ \eta_v(x) = \eta_r(x) \,\right\}$$

Entropy Decomposition

- Let p_o be the uniform distribution
  - The origin of θ-coordinate ($\forall x \in S.\ \theta(x) = o$)
- The entropy $H(X)$ of X is

$$H(X) = -\sum_{x \in S} p(x) \log p(x) = -D_{\mathrm{KL}}(p, p_o) + \log |S|$$

- If we apply the KL divergence decomposition:

$$H(X) = -\left(D_{\mathrm{KL}}(p, r) + D_{\mathrm{KL}}(r, p_o) \right) + \log |S|$$

  - r is the mixed dist. of (p, p_o) w.r.t. I

The Statistical Significance of θ

- θ is coefficients of the log-linear model:
  $\log p(x) = \sum_{s \le x} \theta(s)$
  - We can assess the statistical significance of each $\theta(x)$

- Null and alternative hypotheses are

  $H_0 : \theta_p(x) = 0, \forall x \in I, \quad H_1 : \theta_p(x) \neq 0, \forall x \in I,$
  - This corresponds to knocking down elements in I

- The statistics $\lambda = 2N D_{\mathrm{KL}}(p, r)$
  - N: Sample size
  - r: The mixed dist. of (p, p_0) w.r.t. I
  - λ follows the χ^2 dist. with the degree of freedom $|S| - 1$, thus we can compute the p-value

Orthogonal Decomp. of Interactions

- Given n events $e_1, e_2, \dots, e_n$
- $p(x)$: the probability of the combination (pattern) $\bigcap_{i \in x} e_i$ for each subset $x \subseteq [n] = \{1, 2, \dots, n\}$
- **Objective**: Decompose $\log p(x)$ to the sum of $\theta(s)$ $(s \subseteq x)$
  - $\theta(s)$ shows the "pure" contribution of interactions $\bigcap_{j \in s} e_j$
    - They are independent of their frequencies $\eta(s)$
  - The order $\leq$ is given according to the inclusion relationship: $x \leq s$ if $x \subseteq s$

Constructing S from Data

- Given N samples $t_1, t_2, \ldots, t_N$
  - Each t_i is a set of events

- Estimate $p(x)$ by the natural estimator

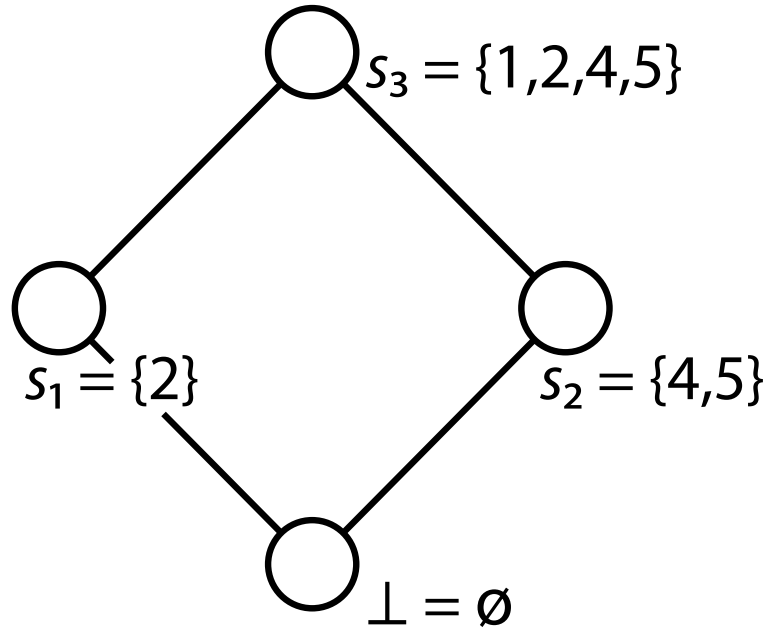
  $\hat{p}(x) = |\{i \in [n] \mid t_i = x\}|/N$
  - For $\perp$, $\hat{p}(\perp) = 1 - \sum_{x \in S^+} \hat{p}(x)$

- We exclude combinations that do not frequently appear in the dataset and set S as

  $S^+ = \{\, x \subseteq [n] \mid \hat{p}(x) \geq \sigma \,\}$
  - σ is a real-valued threshold

Example (1/2)

	Events
t_1	e_2
t_2	e_2
t_3	e_4, e_5
t_4	e_1, e_2, e_4, e_5
t_5	e_1, e_2, e_4, e_5
t_6	e_3
t_7	e_1, e_2, e_4, e_5
t_8	e_4, e_5
t_9	e_1, e_2, e_4, e_5
t_{10}	e_2



$\sigma = 0.2$

Example (2/2)

- $\theta_{\hat{p}}(\perp) = -2.303, \quad \theta_{\hat{p}}(\{2\}) = 1.099,$
  $\theta_{\hat{p}}(\{4, 5\}) = 0.693, \quad \theta_{\hat{p}}(\{1, 2, 4, 5\}) = -0.405$
  - $p(x) = 1.099x_2 + 0.693x_4x_5 - 0.405x_1x_2x_4x_5 - 2.303$
- Let r_x be the mixed distribution of (p, p_0) with $\{x\} \in S$:
  $D_{\mathrm{KL}}(\hat{p}, \hat{r}_{x_1}) = 0.0523,$ p-value of $x_1 = 0.79,$
  $D_{\mathrm{KL}}(\hat{p}, \hat{r}_{x_2}) = 0.0170,$ p-value of $x_2 = 0.95,$
  $D_{\mathrm{KL}}(\hat{p}, \hat{r}_{x_3}) = 0.0040,$ p-value of $x_3 = 0.99.$
  - Note that these large p-values are due to small $N = 10$
  - If $N = 100$, for example, the p-value of x_1 becomes 0.015 and it is significant under the significance level $\alpha = 0.05$

Conclusion & Current Progress

- Theoretical results on information decomposition
  - Can be applied to measuring importance of patterns
- **Future work**:
  - Apply significant pattern mining to other data (e.g. large-scale graphs)
  - Further analyze IG and posets from theory to practice
  - FS project; analyzing brain MRI data (with Dr. Morishima)