

LDワークショップ Impromptu Talks

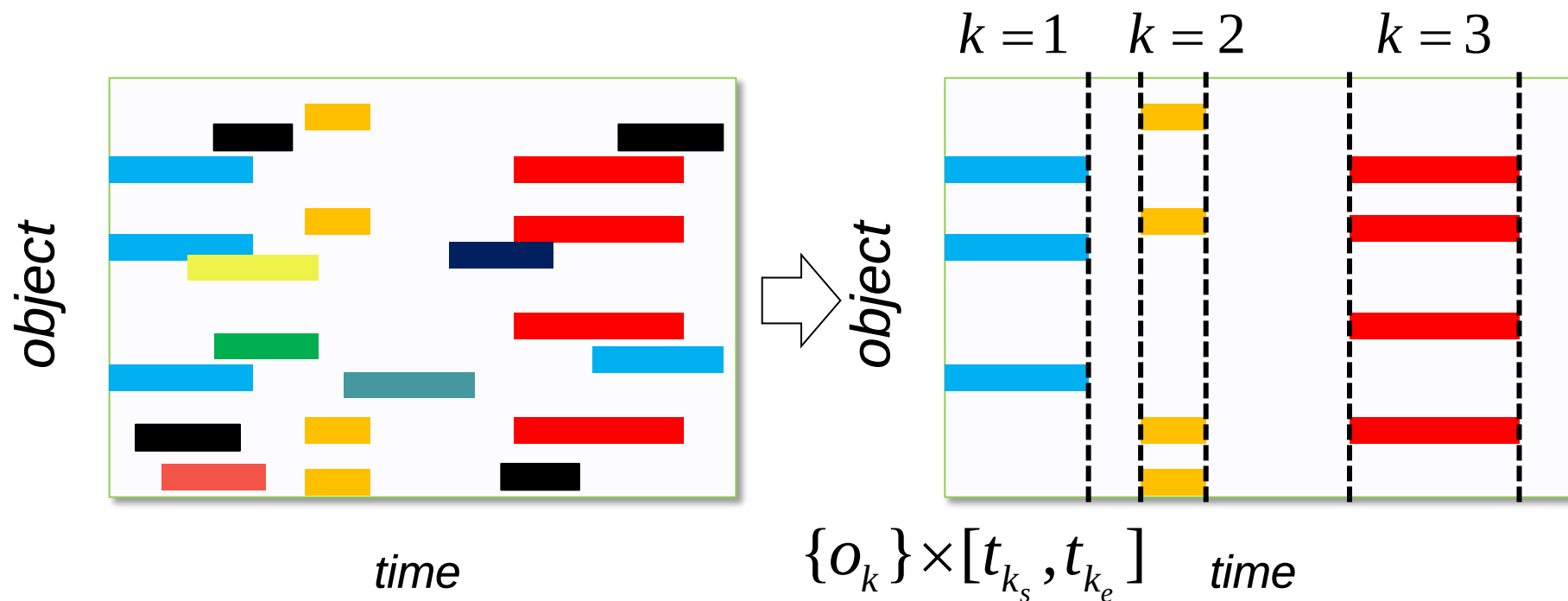
時変多重関係データからの 重要クラスタ抽出

2011年6月22日

上田 修功

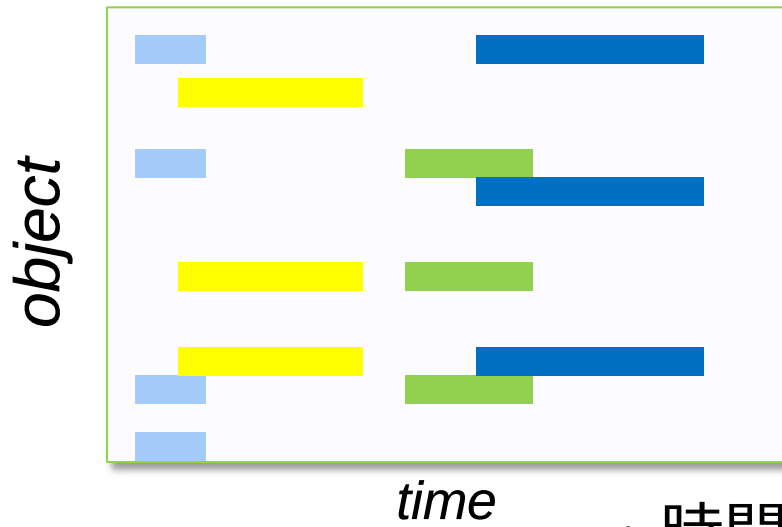


時変多重関係データ分析 (基本LDタスク!?)

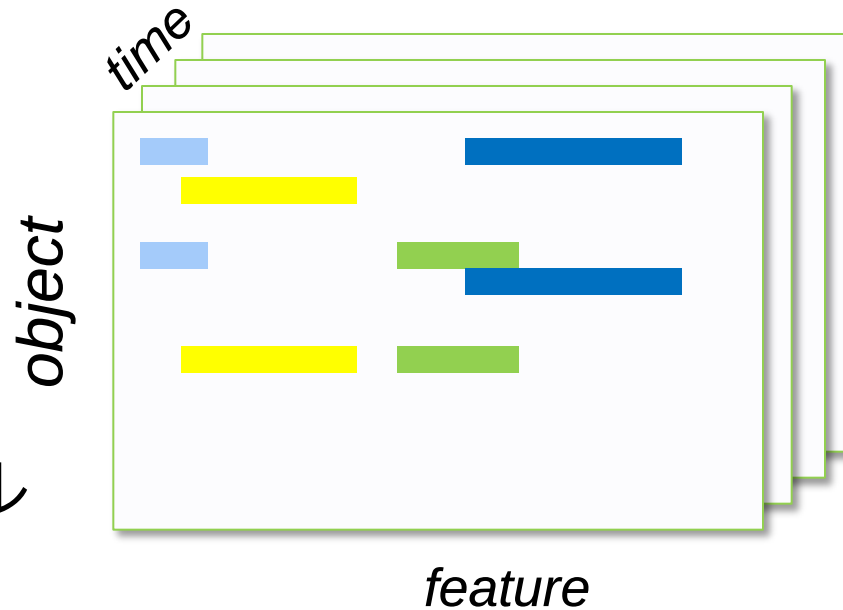
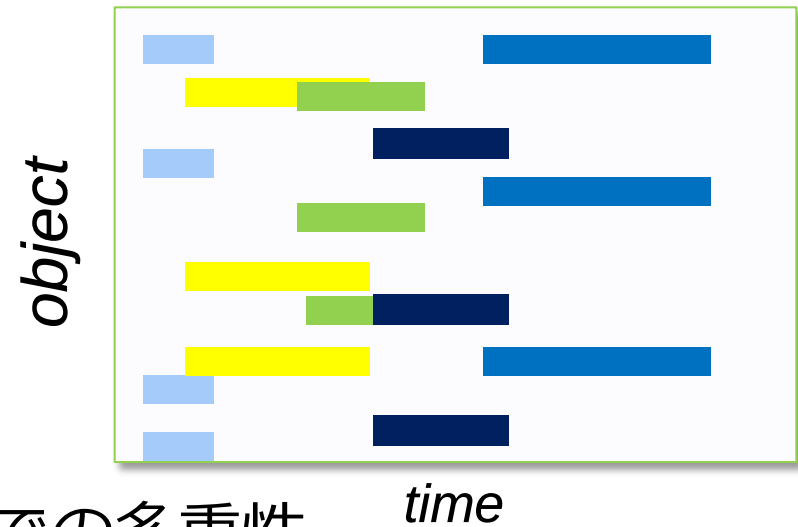


- 全objectをクラスタリングするのではなく、**重要なクラスタを同時に抽出**する問題
- 各objectは、複数の潜在クラスに多重に帰属しても良い (多重メンバーシップ: 視点が異なればメンバーシップも異なる。各objectは多重性を持つ。) ²

Some Variants



・ 時間方向での多重性

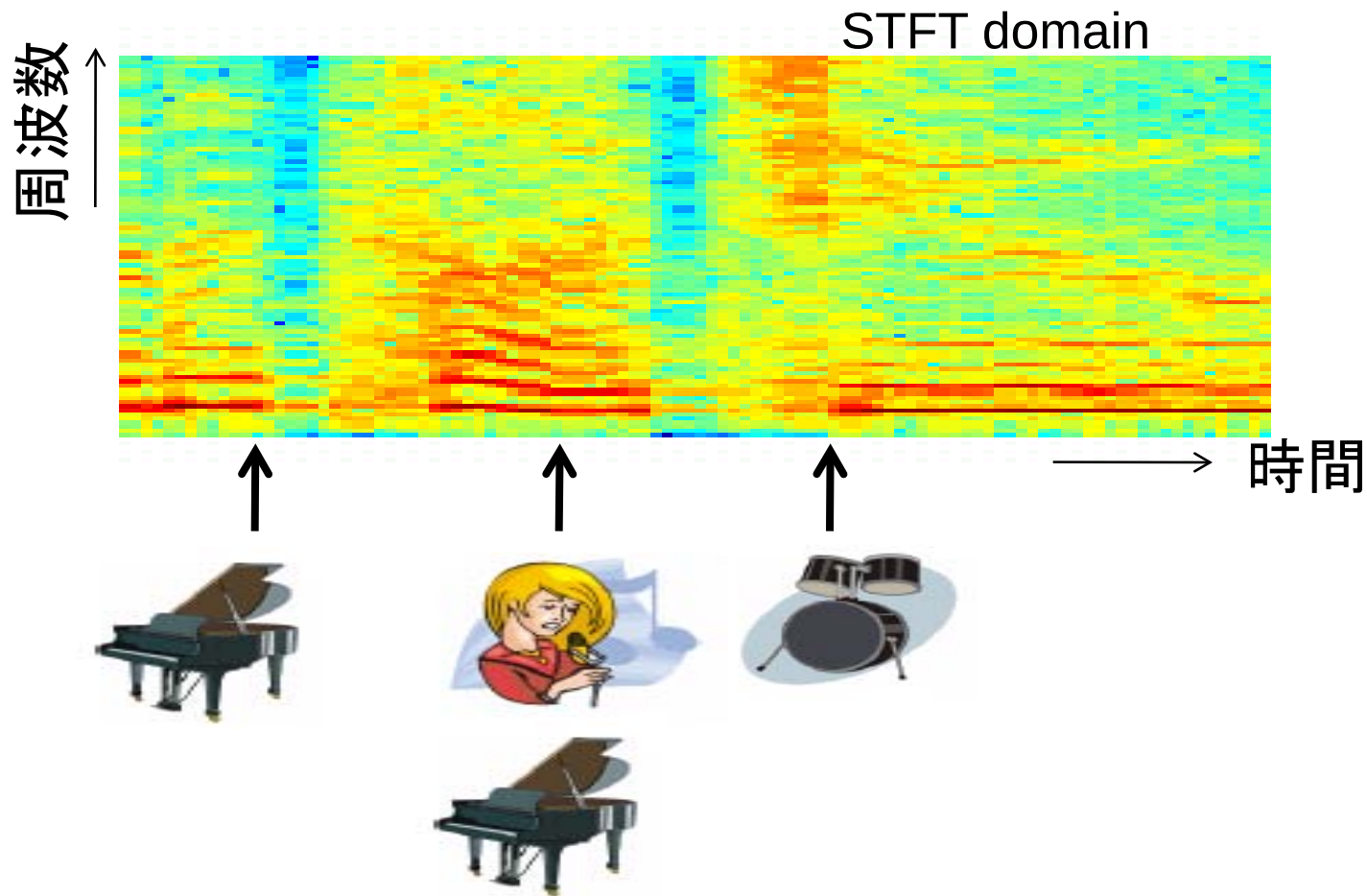


・ テンソル

アプリケーション

- ・**購買データ分析**： 例) ある期間（月/年）で購買パターンの類似したアイテム群を抽出
- ・**大量センサネットワーク分析**： 例) 災害予知
- ・**ネットワーク分析**： 例) SNSの同型群抽出
- ・**遺伝子解析**： 例) 品種とそのDNA多型
- ・**単一マイクからの音源分離**： 例) 複数の楽器の音分離
- ・
- ・
- ・

応用例（音源分離）



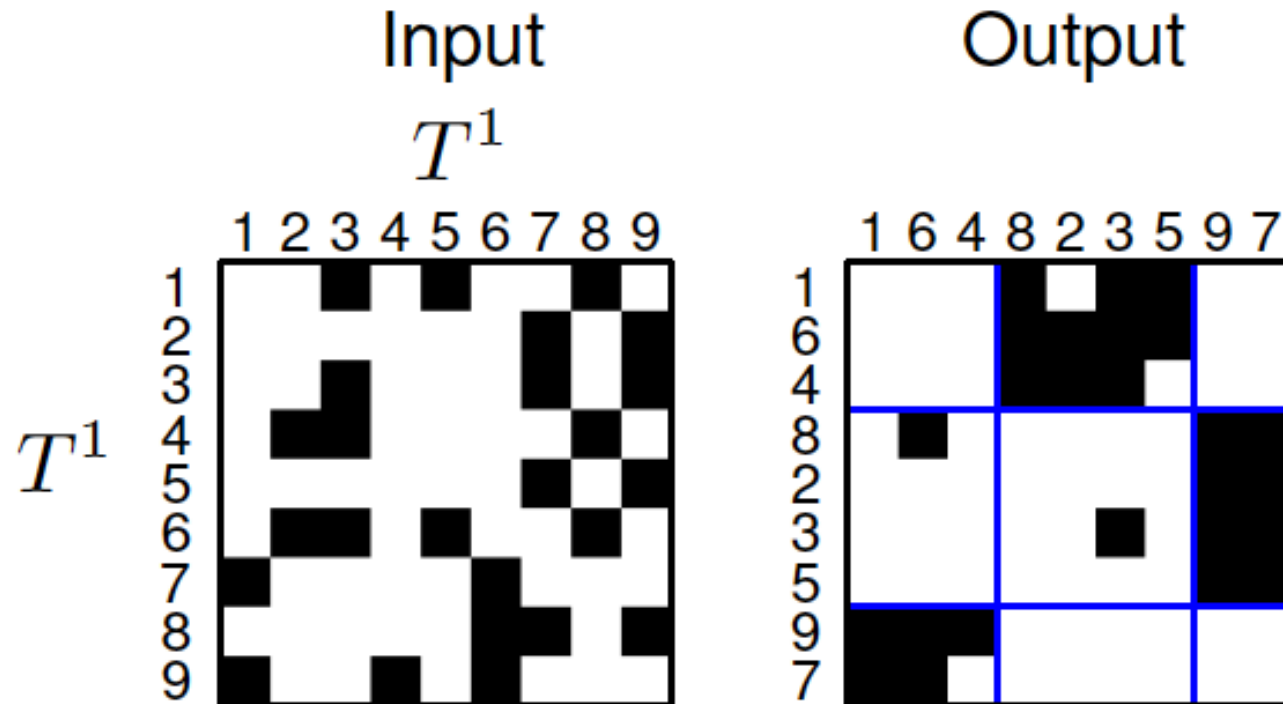
- 各時刻でどの楽器（複数可）が演奏されているかを推定（**単一マイク下**での音源分離）

関連従来研究：関係データ解析

Stochastic Block Models:

Estimation and prediction for stochastic blockstructures

K.Nowicki and T.A.B.Snijders, JASA 96 (455), 2001

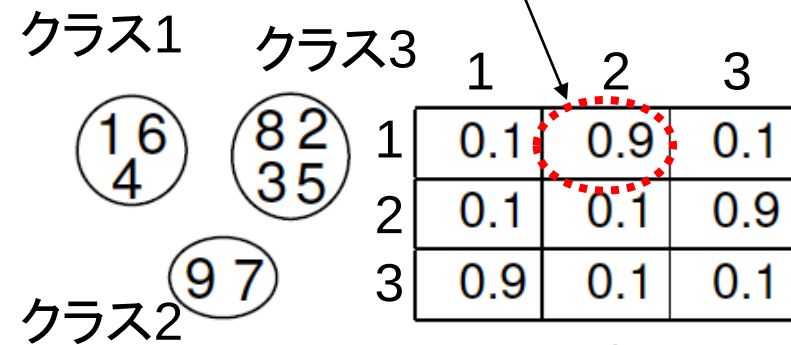


IRMによるデータ生成過程

Learning Systems of Concepts with an Infinite Relational Model,
C.Kemp, J.B.Tenenbaum, T.Yamada, N.Ueda, AAAI 2006 381-388

Stochastic Block Modelsのノンパラベイズversion:

$\eta(1, 2)$: クラス1,2間に"1"を生成する確率

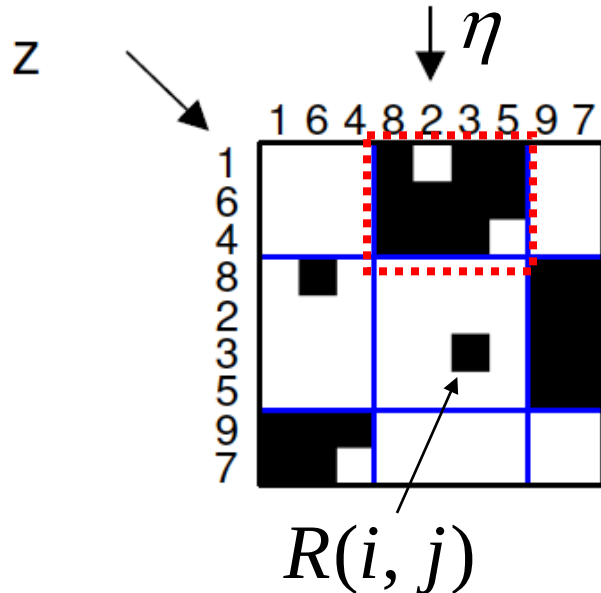


$$z | \gamma \sim \text{CRP}(\gamma)$$

$$\eta(a, b) | \beta \sim \text{Beta}(\alpha, \beta)$$

$$R(i, j) | z, \eta \sim \text{Bernoulli}(\eta(z_i, z_j))$$

object i のクラス
object j のクラス



Note: a, b クラスインデックス
 i, j objectインデックス

MAP学習

$$P(\mathbf{Z} | \mathbf{R}) \propto P(\mathbf{R} | \mathbf{Z})P(\mathbf{Z}) \longrightarrow \text{Max}_{\mathbf{Z}}$$

但し, $P(\mathbf{Z}): \text{CRP}(\gamma)$

$$P(\mathbf{R} | \mathbf{Z}) = \int P(\mathbf{R} | \mathbf{Z}, \eta) p(\eta | \beta) d\eta$$

$$= \prod_{a,b \in I} \frac{\text{Beta}(S(a,b) + \alpha, \bar{S}(a,b) + \beta)}{\text{Beta}(\alpha, \beta)}$$

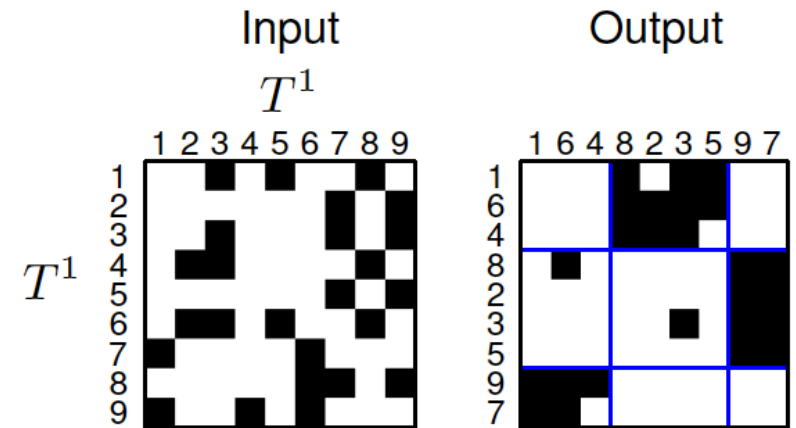
ベータ関数

priorの共役性より解析的に求まる

従来のblockモデルの課題

• 零値の扱い

- 零値=missing?
- Bernoulliの対称性



• 全objectのクラスタリング

- 全objectには興味はない

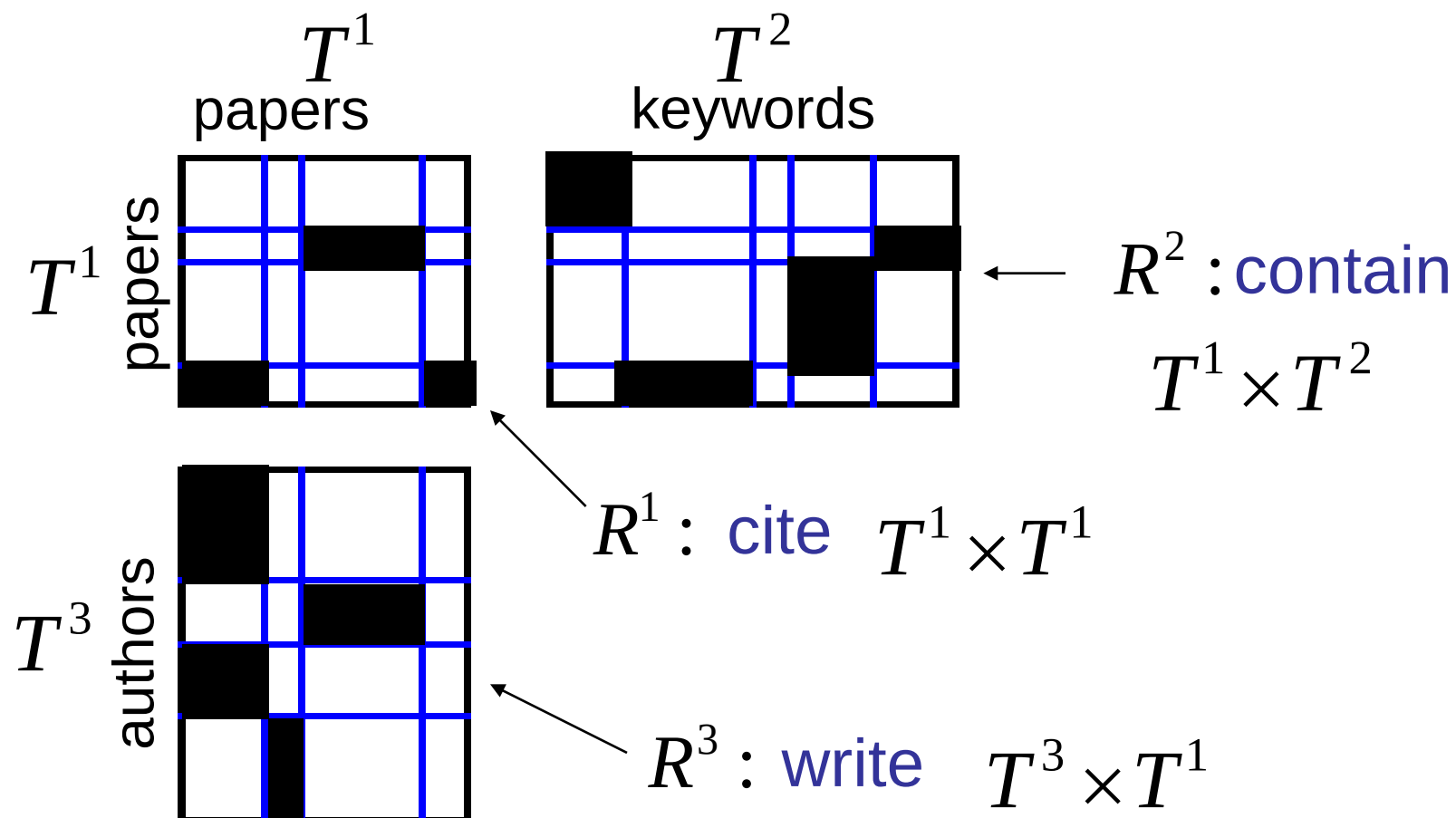


部分集合を選択する潜在変数の導入

参考

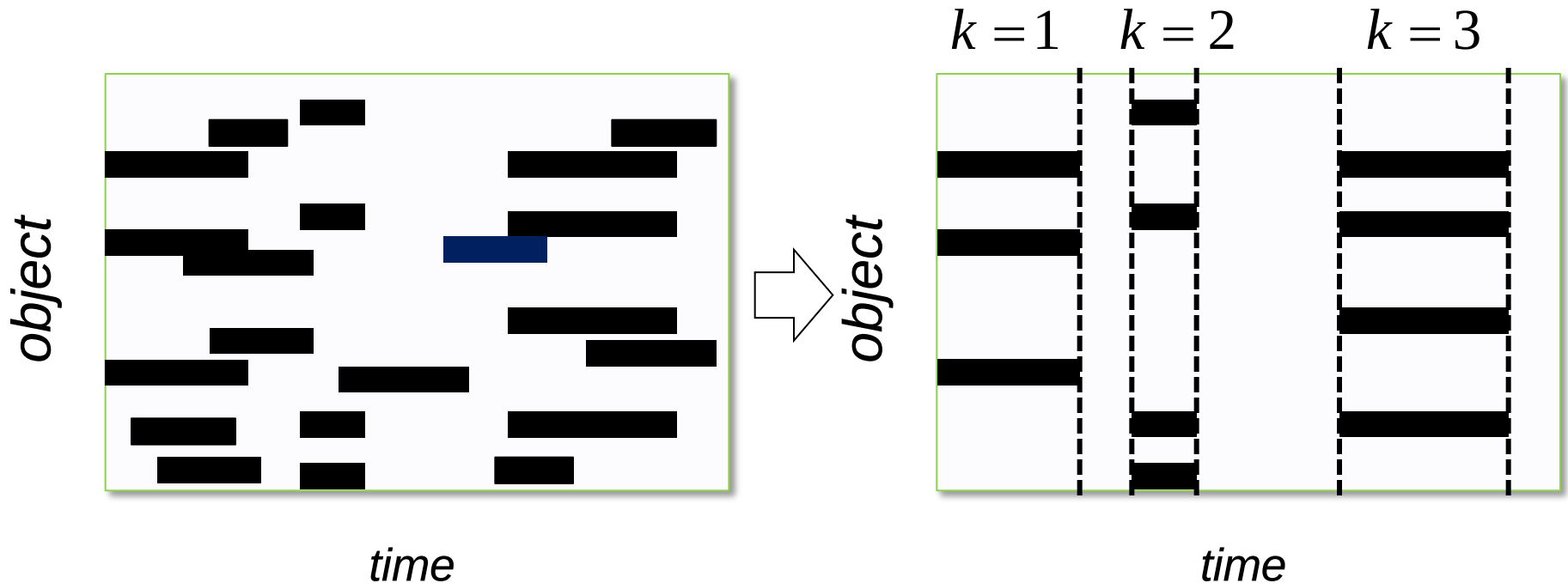
Hoff et al., **Subset clustering** of binary sequences,
Biometrics, vol.61,no.4, 2005

ex) 論文引用データ解析 ($n=3, m=3$)



3重の同時クラスタリング

Non-overlapped & Binary case



提案法の生成モデル(ノンパラベイズ)

i : object k : latent class

$\phi_i | a_i, b_i \sim \text{Beta}(a_i, b_i)$ 全クラス共通のパラメータ

$\theta_{k,i} | c_i, d_i \sim \text{Beta}(c_i, d_i)$ クラス固有のパラメータ

$\lambda_i | e_i, f_i \sim \text{Beta}(e_i, f_i)$

$r_{k,i} | \lambda_i \sim \text{Bernoulli}(\lambda_i)$

Object i がクラス k
固有か否かの潜在変数

$\beta | \gamma \sim \text{Stick}(\gamma)$

$\pi_k | \beta \sim \text{DP}(\alpha, \beta)$

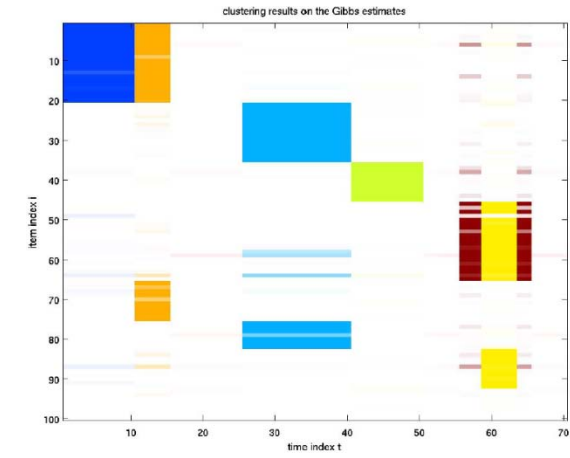
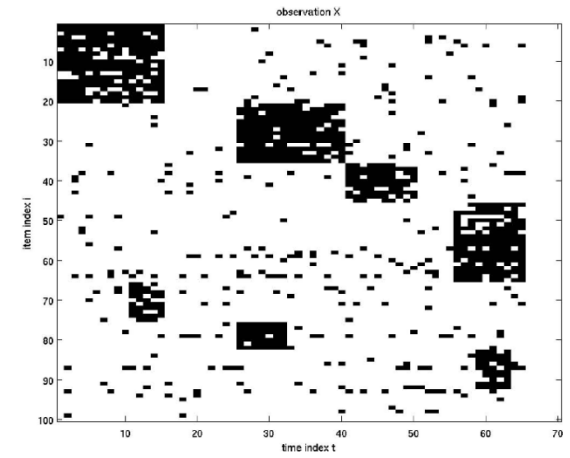
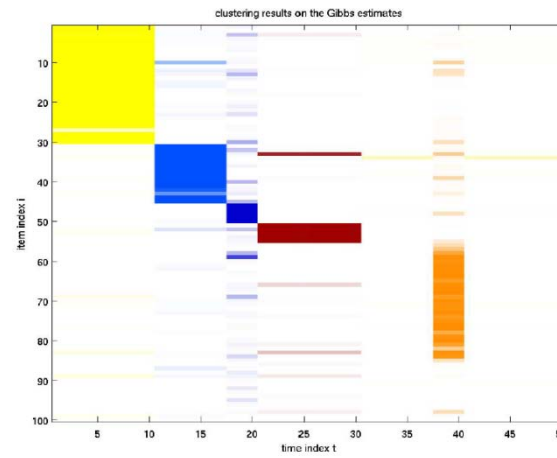
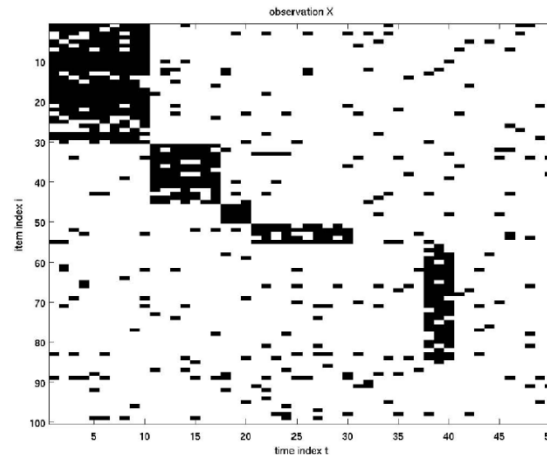
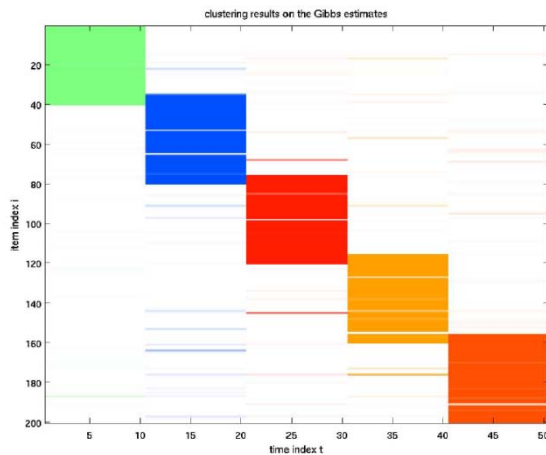
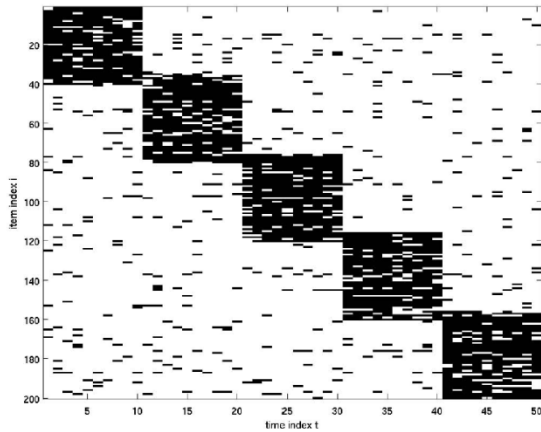
$z_t | z_{t-1}, \{\pi\} \sim \text{Multinomial}(\pi_{z_{t-1}})$

Infinite HMM

$x_{i,t} | z_t, r_{k,i}, \{\phi\}, \{\theta\} \sim \text{Bernoulli}\left(\theta_{z_t,i}^{r_{k,i}} \phi_i^{1-r_{k,i}}\right)$

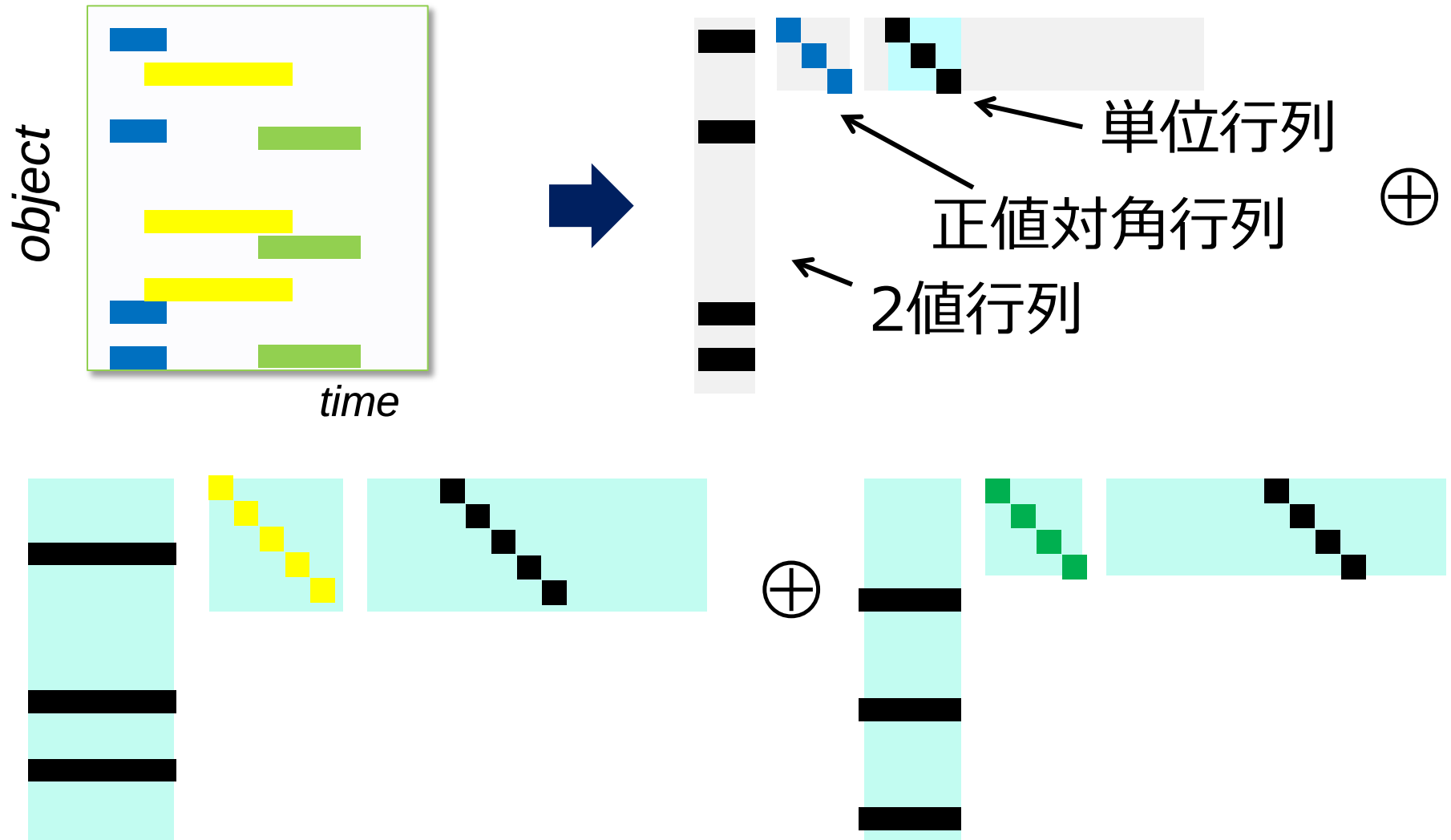
出力分布

ノイズ有り人工データでの結果



各色がクラスに相当し、色の濃さが事後確率に対応

Overlapped case



Note: 色のついたパターンは各要素が同じ値である必要はなく、同じ色のパターンは、単に類似している事を意味する

今後の展開

- Overlapped caseでの推論
- 実データでの検証
- 異種関係データとの統合
- テンソル化への発展