たくさんの情報を効率的に解析

欠損値を含むデータからの高速パターン抽出

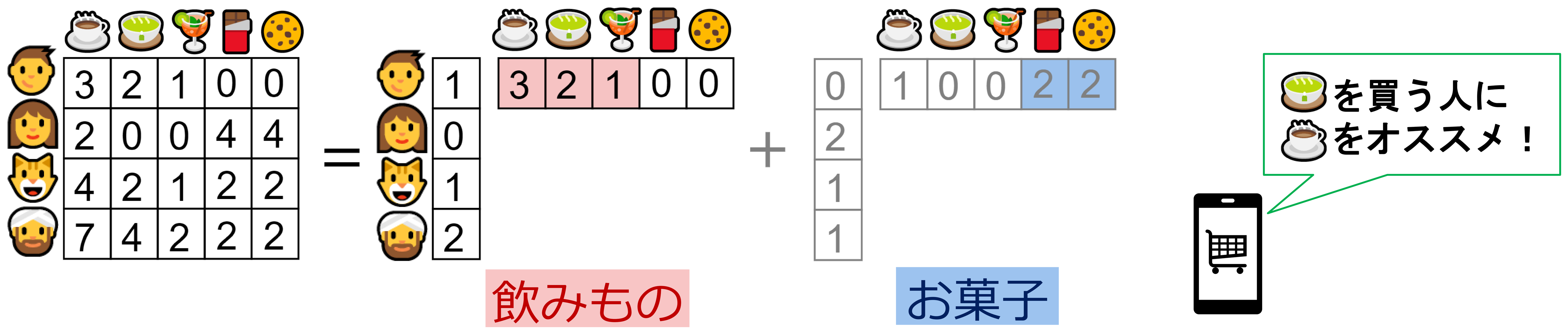
杉山研 ガラムカリ和

NII

■複数の行列から，最も支配的なパターンを高速に見つける公式を発見！
■この公式を応用して，欠損を含むデータから高速にパターンを発見！

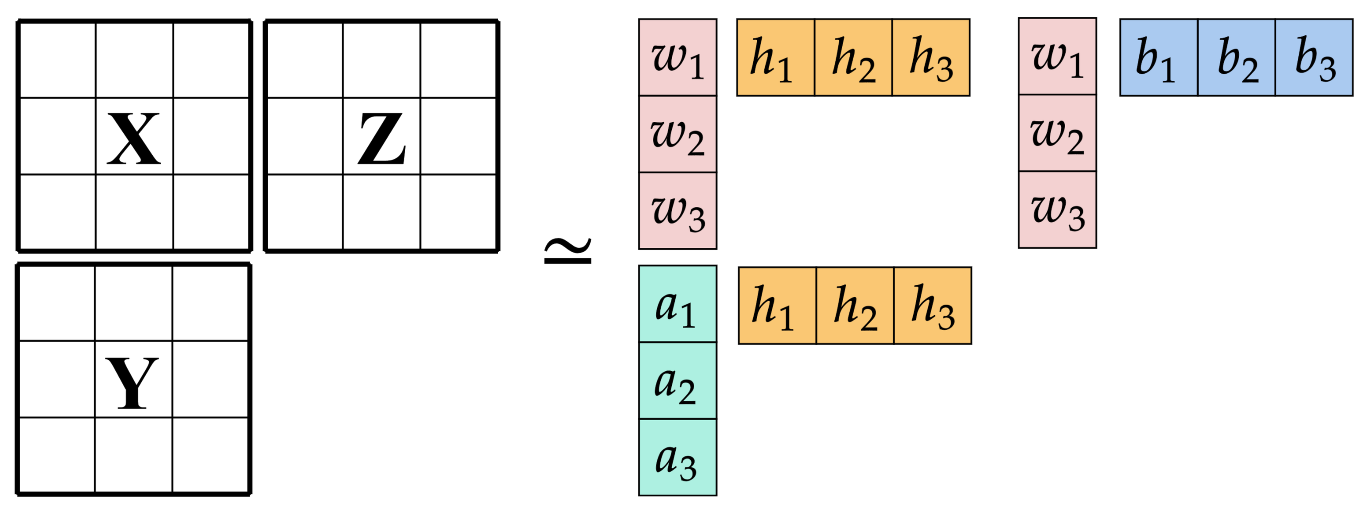
行列分解

データを**非負行列**として捉え，**分解**して**パターン**を取り出す．



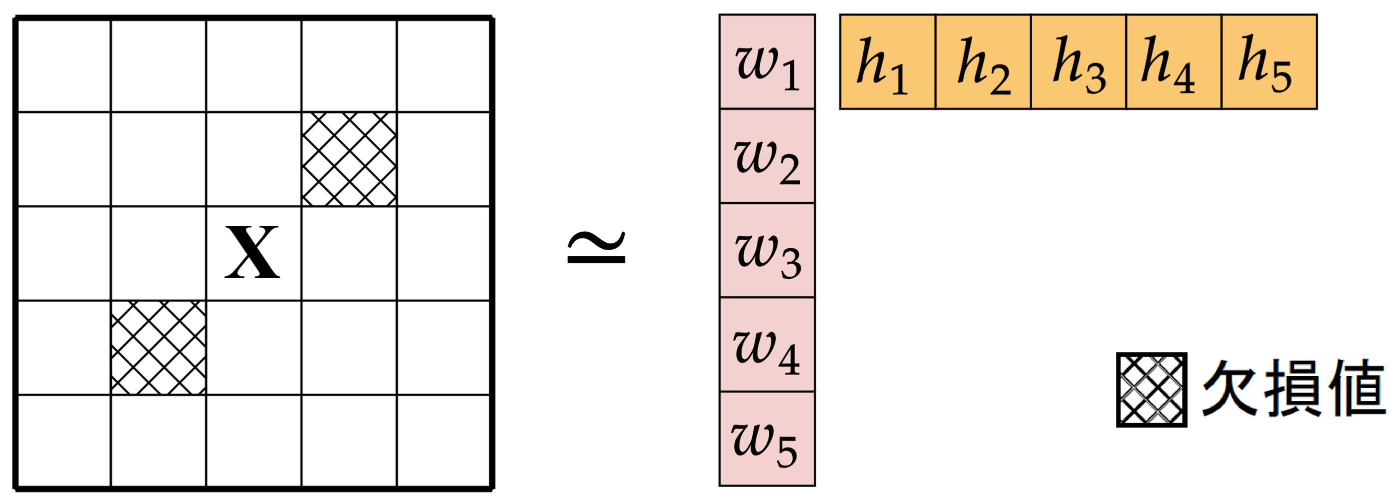
■複合行列分解

複数の行列を一度に分解



■欠損値を含む行列分解

データに含まれる欠損値に対処



複合行列の最良ランク１分解の公式

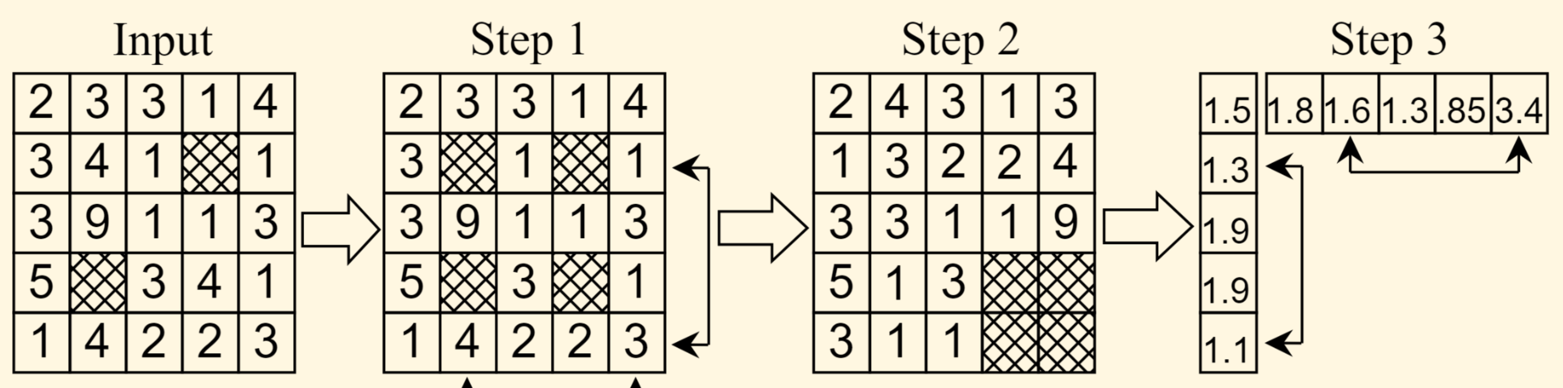
$$w_i = \frac{\sqrt{S(X)}}{S(X)+\beta S(Z)}\left\{\sum_{j=1}^{J} X_{ij} + \beta\sum_{m=1}^{M} Z_{im}\right\} \qquad a_n = \frac{\sum_{j=1}^{J} Y_{nj}}{\sqrt{S(X)}}$$

$$h_j = \frac{\sqrt{S(X)}}{S(X)+\alpha S(Y)}\left\{\sum_{i=1}^{I} X_{ij} + \alpha\sum_{n=1}^{N} Y_{nj}\right\} \qquad b_m = \frac{\sum_{i=1}^{I} Z_{im}}{\sqrt{S(X)}}$$

厳密な解が一発で求まる公式*

* 再構成誤差をKL情報量で定義した場合に最良ランク１分解を実現

欠損値を含む行列の高速ランク１分解法



欠損を増やして複合行列分解に帰着して公式を使用

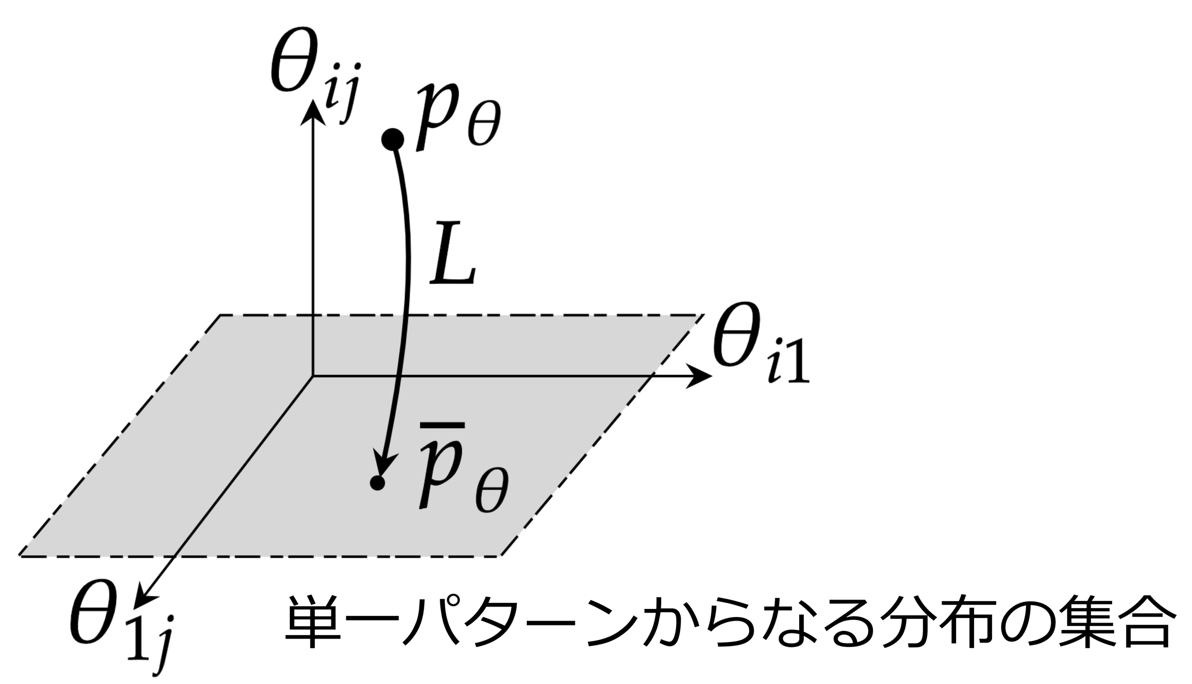
研究のアイデア

行列の番地が確率変数の分布を導入

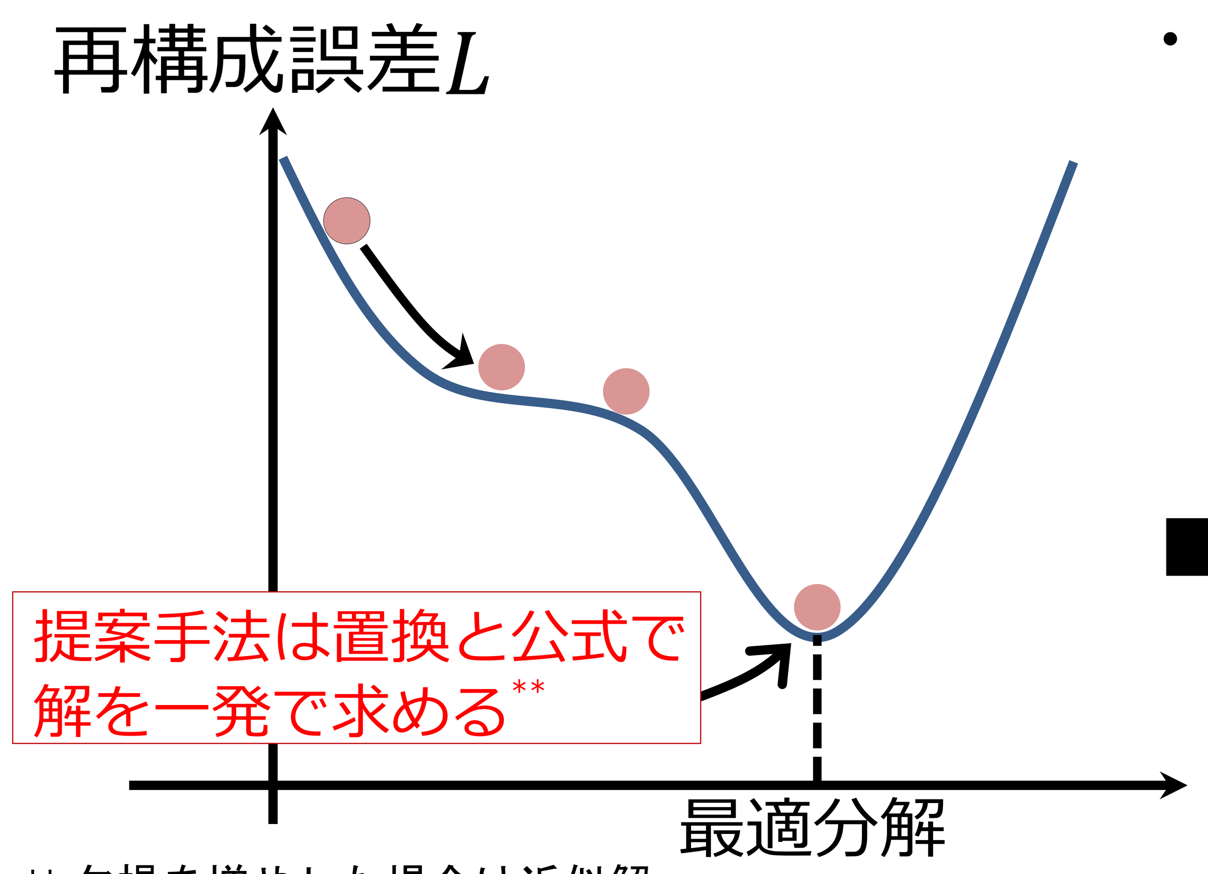
$$p_\theta(i,j) = \exp\left[\sum_{i'=1}^{i}\sum_{j'=1}^{j}\theta_{i'j'}\right]$$

例：$X_{22} = p_\theta(2,2)$
$= \exp[\theta_{11}+\theta_{21}+\theta_{12}+\theta_{22}]$

確率分布の幾何学の理論で公式を導出



何がすごい？



** 欠損を増やした場合は近似解

■従来手法

・坂の傾きを求めながら最適な分解表現を探索．
- 始点を適切に選択する必要がある．
- 時間がかかる．
- 1stepでどれくらい進むかを決める必要がある．
- 停止条件を決める必要がある．

■提案手法

- **置換と解の公式だけ**で完結！
- **高速**に解が求まる．

実験結果

提案手法と従来手法KL-WNMFの比較実験

5〜10倍の高速化

データセット名	行列のサイズ	欠損数	欠損増加率	相対誤差	相対実行時間
DailySunSpot	(73718, 9)	3247	1	1	0.12845
CaliforniaHousing	(20640, 9)	207	1	1	0.11821
MTSLibrary	(1533078, 4)	1247722	1	1	0.18327
BigMartSaleForecas	(8522, 5)	1463	1	1	0.12699
CreditCardApproval	(590, 7)	25	1.92	1.0018	0.12212
HumanResourceAnaly	(14999, 7)	519	1.96146	1.0168	0.11858
PerthHousePrice	(33656, 14)	16585	2.61345	1.0004	0.15382
SleepData	(62, 8)	12	2.75	1.0211	0.18208
Bostonhousing	(506, 14)	120	5.6	1.003	0.1097

詳細

[1] Ghalamkari, K., Sugiyama, M., **Fast Rank-1 NMF for Missing Data with KL Divergence**, AISTATS2022
[2] ガラムカリ和, 杉山麿人, **欠損を含む非負行列の高速なランク１分解** 第120回人工知能基本問題研究会(SIG-FPAI 2022)