次元圧縮を用いた重回帰分析による協調フィルタリング

藤井 流華

学籍番号: 1410112

総合情報学科 経営情報学コース

岡本研究室

はじめに

情報推薦システム

アイテムの特徴やユーザの嗜好からユーザが好みそうなアイテムを推薦

協調フィルタリング

購入履歴やスコアのデータから目的のアイテムのスコアを予測する

ユーザ-アイテム行列

代表的な手法

行列因子分解
アソシエーションルール
ベイジアンネットワーク

研究背景

R. Sinha, K. Swearingen: The role of transparency in recommender systems. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems, 830–831, 2002.

推薦の透明性：推薦される理由が明らかであるかどうか

推薦に透明性があるほうが
推薦されたものを好み，推薦を信頼する [R. Sinha, 2002]

	スコア予測	透明性	計算コスト
行列因子分解	○	×	○
アソシエーションルール	×	○	△
ベイジアンネットワーク	○	○	×
重回帰分析	○	○	△

研究目的

P. Peduzzi, J. Concato, A.R. Feinstein, T.R. Holford: Importance of Events Per Independent Variable in Proportional Hazards Regression Analysis Ⅱ. Accuracy and precision of regression estimates, J. of Clinical Epidemiology, 48(12), 1503-1510, 1995

協調フィルタリングに重回帰分析を応用し,
目的のアイテムのスコアを予測するだけでなく,
予測に影響を与えている他のアイテムも推定する技術の開発

協調フィルタリングにおける課題

ユーザ–アイテム行列は巨大でスパース性が強い

→ 解析に使える観測データが少なくなる

重回帰分析は変数の数に対して観測データ数が10倍程度必要

[P. Peduzzi, 1995]

想定される解決法

正則化

次元圧縮

→ **計算コスト高い**

次元圧縮を用いた重回帰分析

学習用ユーザーアイテム行列

m ユーザ

n アイテム

X

y

説明変数　　目的変数

直接法

推定したい予測モデル

$$y = \alpha_0 + X\alpha$$

圧縮空間

k

m X' y

$$\alpha_0 = \beta_0$$
$$\alpha = W\beta$$

$$y = \beta_0 + XW\beta$$

||

圧縮空間: 予測モデル

$$y = \beta_0 + X'\beta$$

word2vec

評価されたスコアのみ学習

↓

高速

n k 次元 W

$$X' = XW$$

実験内容

1. 提案法と直接法の予測精度の比較
2. 提案法と直接法に正則化を適用した場合の比較
3. 提案法と直接法で推定したパラメータの相関係数の比較

予測精度の評価　N：テストデータ数　$\hat{y}_i$：予測値　y_i：真値

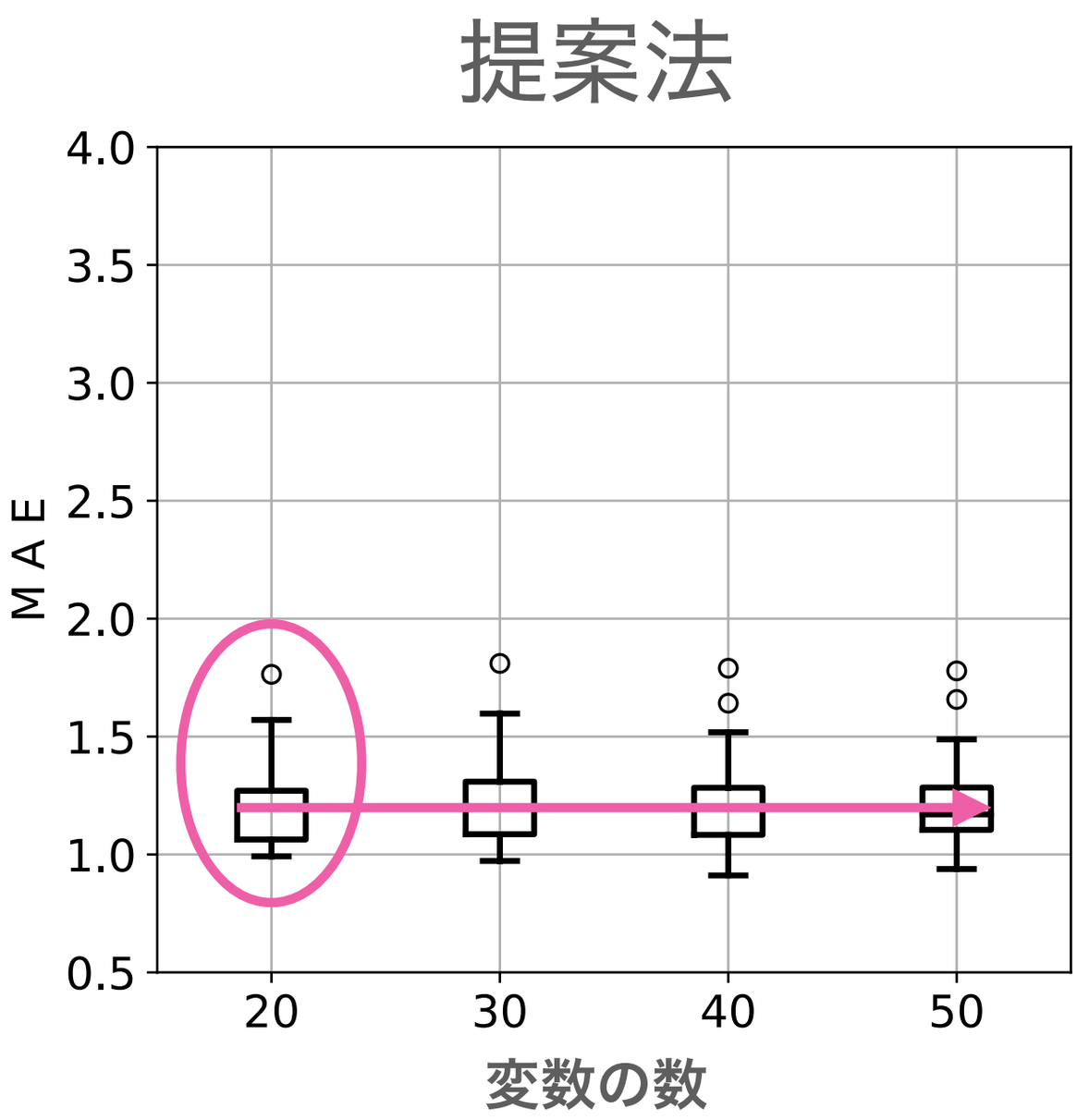
$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

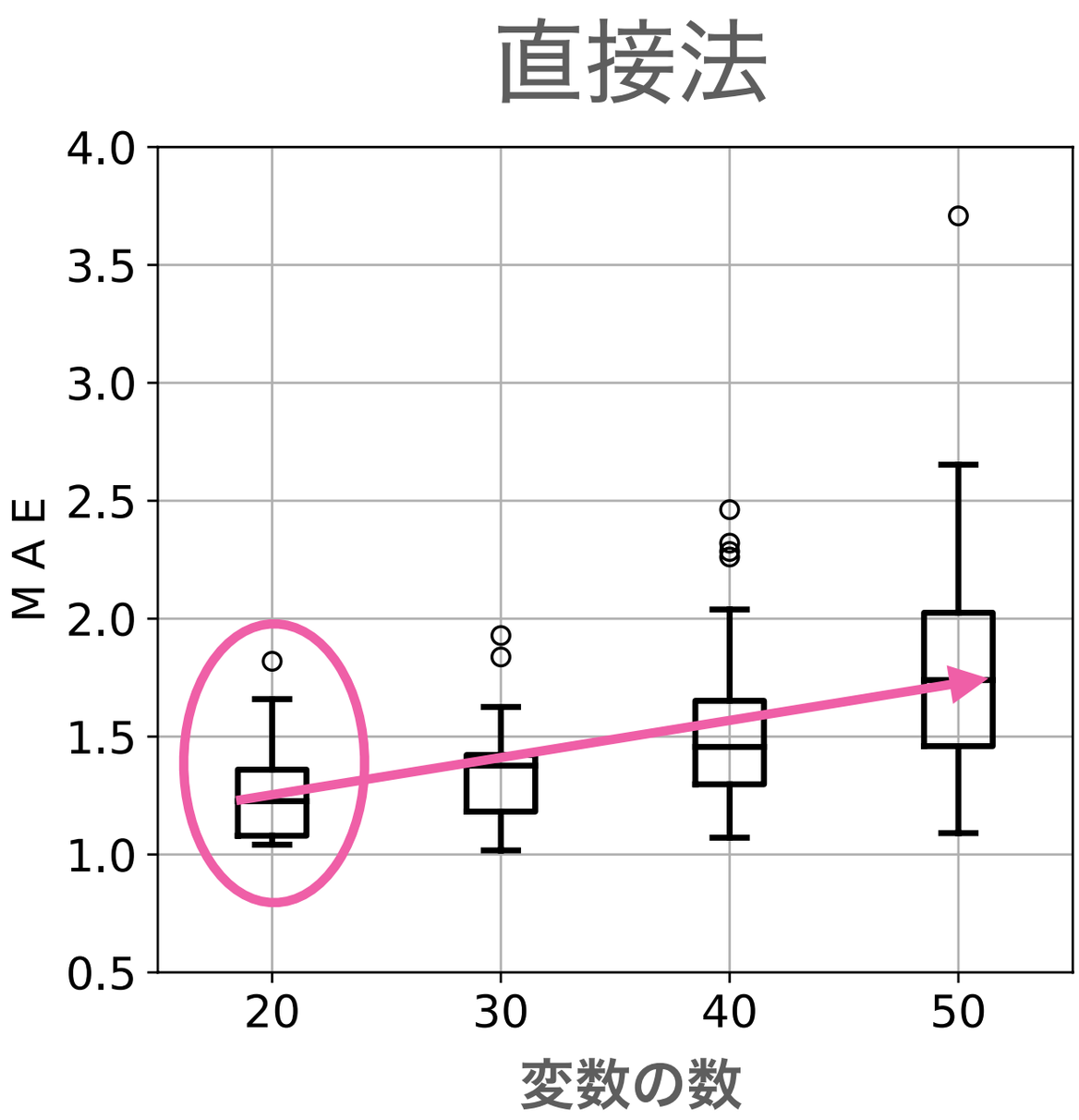
使用データ　Book Crossing データセット

ユーザが書籍につけた１～10までのスコアを集計したもの
ユーザ数: 278,858　書籍数: 271,379　総スコア数: 383,852
収集期間は2014年8月～9月の1週間

提案法と直接法の予測精度

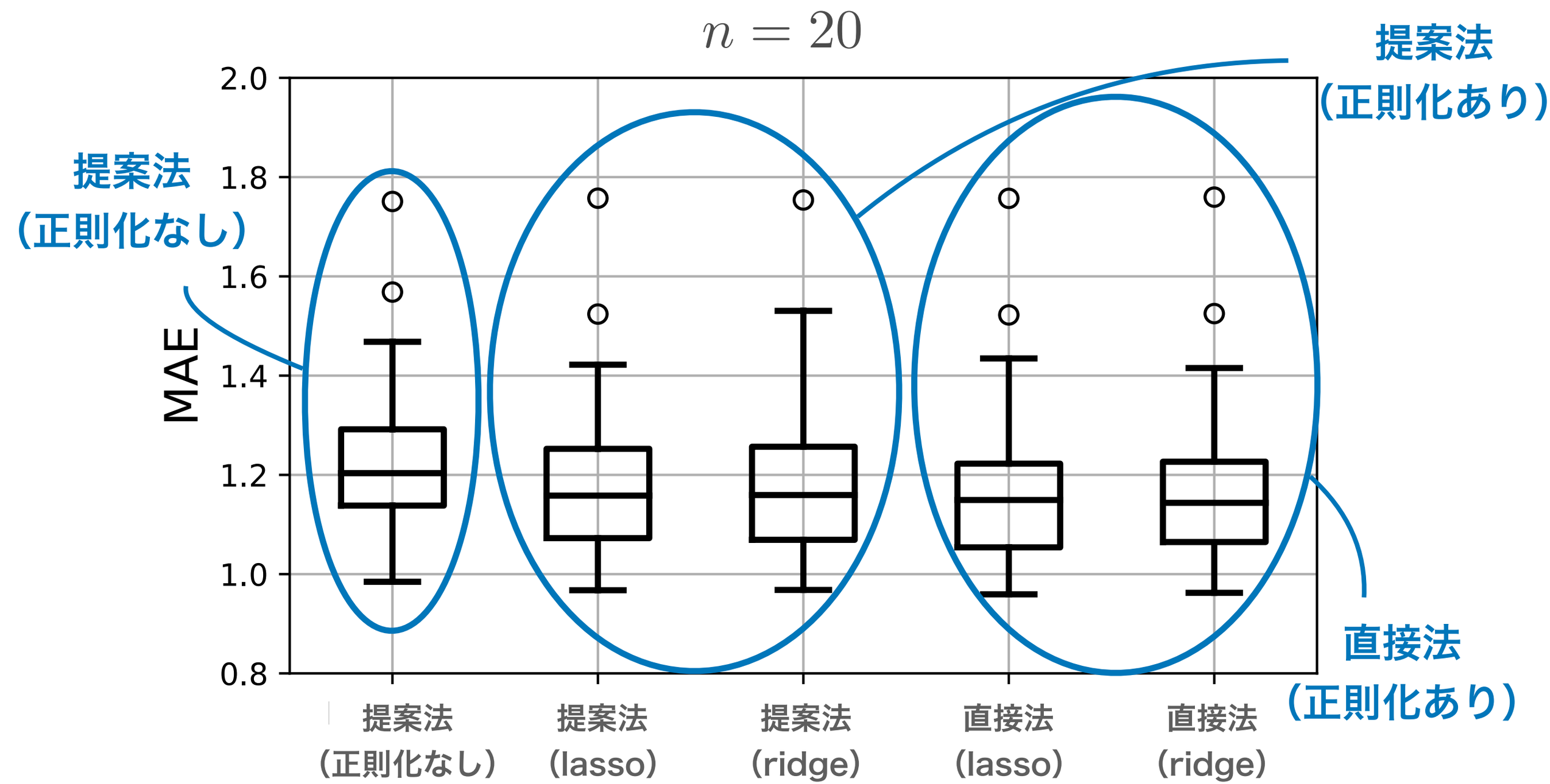
提案法が正しく動作するか予測精度の観点で検証





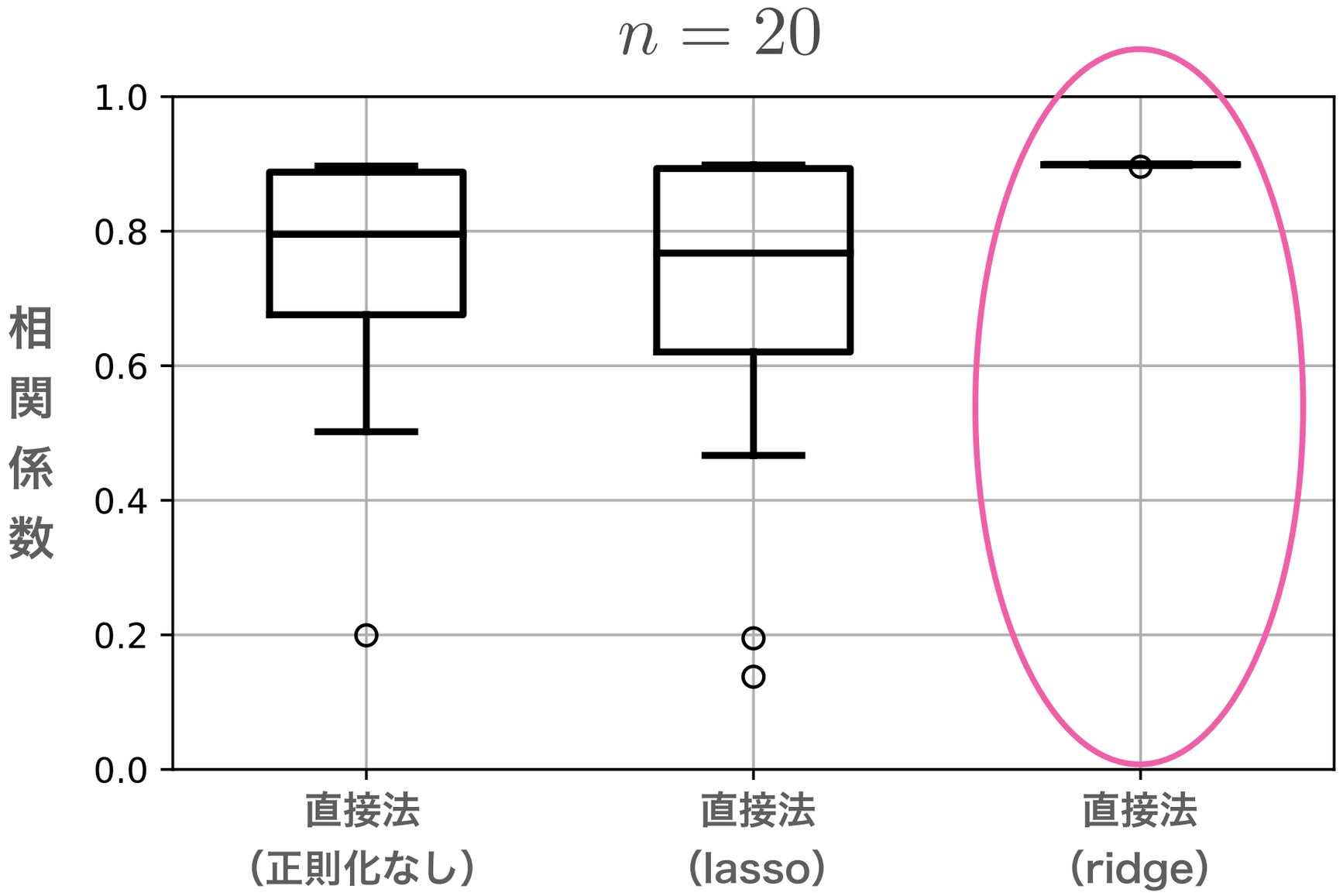
正則化を適用したときの予測精度

提案法と他の手法（正則化）との違いを予測精度の観点で検証



提案法のパラメータとの相関係数

提案法で得られたパラメータと
直接法や他の手法で得られたパラメータの比較



おわりに

本研究のまとめ

目的　スコアを予測するだけでなく，透明性がある推薦技術の開発

提案法　協調フィルタリングに重回帰分析を応用
Word2vecを用いて次元圧縮し，パラメータを推定

実験　予測精度の観点での提案法の有効性を確認
提案法とL2正則化のパラメータが類似

今後の研究の方向性

- 説明変数の数を増やし，提案法と正則化（従来法）の計算コストの観点での違いを明らかにする
- 適切なword2vecのハイパーパラメータの検証

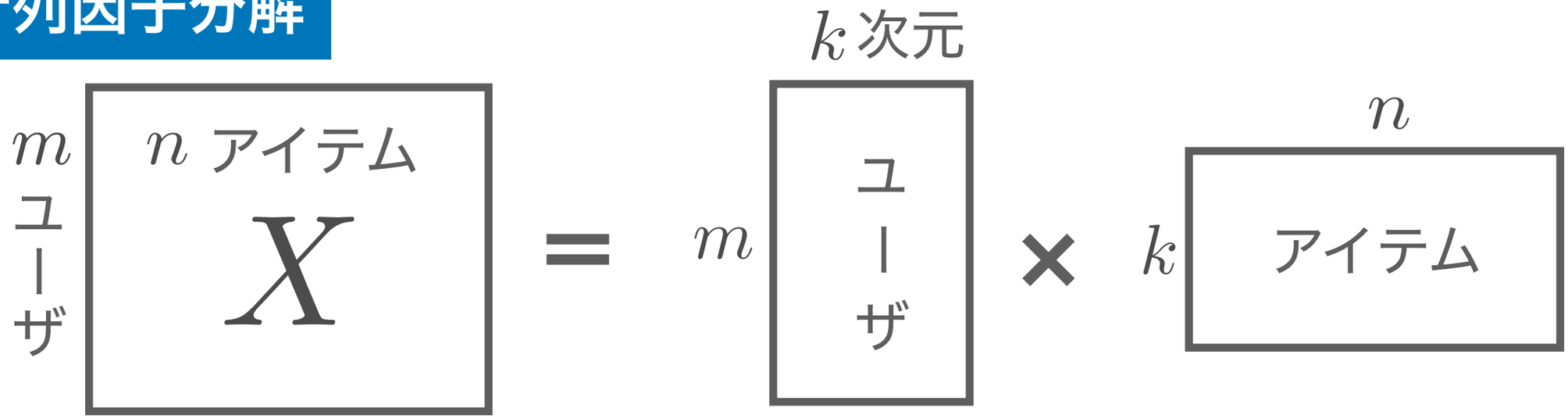
質問用スライド

2017/10/3

2017年度 総合情報学科 卒業論文中間発表

補足：代表的な手法

行列因子分解



アソシエーションルール

「Aが購入されたらBも購入される」といったルールを見つける
アイテム同士の相関関係を発見する

ベイジアンネットワーク

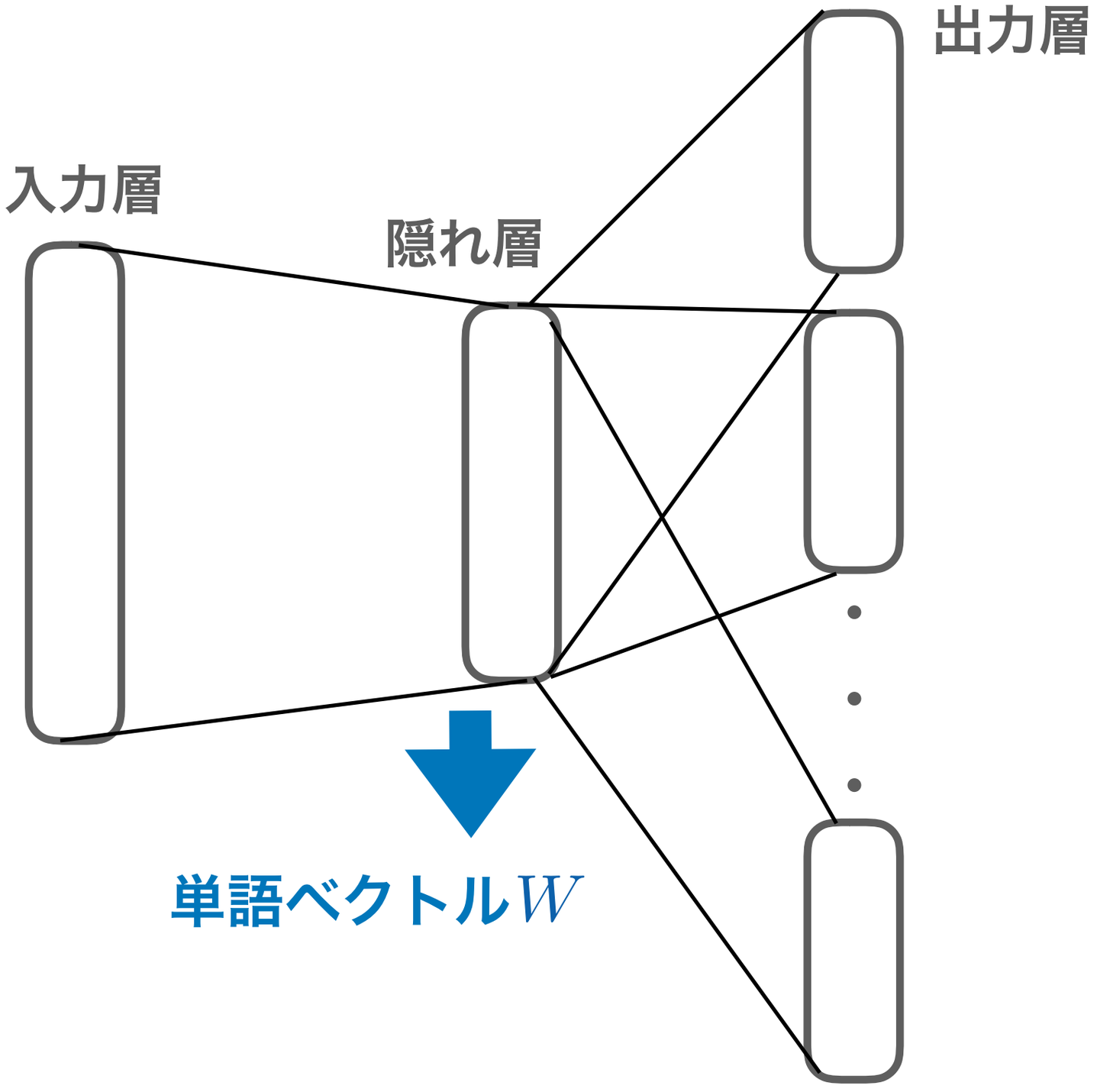
変数の因果関係を
条件付き確率で示す

	以前購入	未購入
男性	85%	10%
女性	4%	1%

補足：word2vec

ユーザーアイテム行列

word2vec用学習データ



補足：word2vec

ハイパーパラメータ

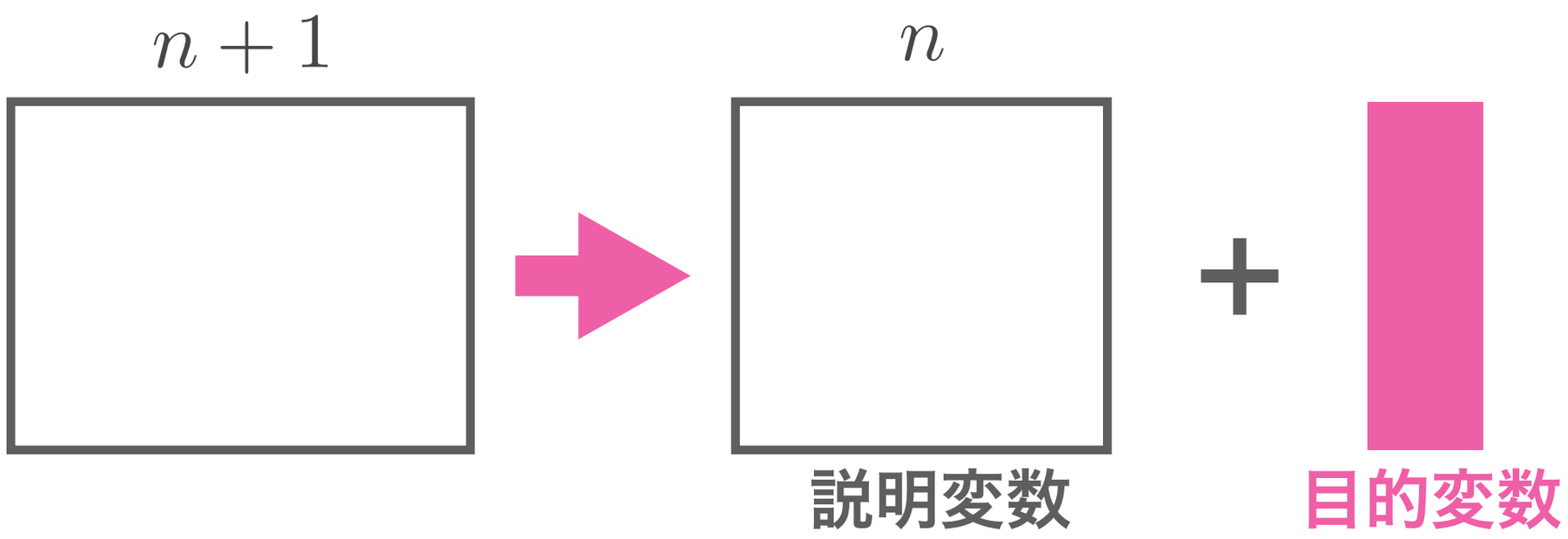
圧縮次元数	10
学習の最大単語数	100
ネガティブサンプリング数	5
トレーニング反復回数	500
最小出現数	0
学習モデル	Skip-gram

補足：変数

目的変数の中に欠損値があるデータは使用しない

→ **全書籍を変数とするとデータ数が足りず,**
パラメータ推定できない可能性がある

ユーザに評価されている書籍上位$n+1$件を変数として選択



欠損値はスコアの中央値である5.5で補完