重回帰分析による
推薦の透明性を有した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（商品や店舗，記事など）の内容や特徴，
好みの度合いなどからユーザが好みそうなアイテムを予測するシステム

- **内容ベースフィルタリング**
  アイテムの内容に基づいて推薦を行う手法

- **協調フィルタリング**
  購入履歴やアイテムに付与されたスコアなどから，
  類似ユーザ・アイテムの発見や付与されるスコアを予測する手法

  ├ メモリベース法：スコアの類似関係から予測
  └ モデルベース法：事前に学習したモデルを用いて予測

推薦処理の計算コスト

メモリベース法 ＞ モデルベース法

推薦システムにおける説明に関する説明

推薦システムの説明 [Tintarev, 2007]

- 信頼性：ユーザのシステムに対する信頼を向上させる情報を提供
- 有効性：ユーザが良い決定を行うのを助ける情報を提供
- 透明性：なぜそのアイテムが推薦されたかの情報を提供　　etc.

推薦に透明性があると，ユーザは推薦されたアイテムを好みやすい

[Herlocker+, 2000], [Shinha+, 2002], [Gedilki+, 2014]

→ 協調フィルタリングではメモリベース法が主流

モデルベース協調フィルタリングで推薦の透明性を実現

Tintarev, N.: Explaining Recommendedions, *Proc. of Int. Conf. on User Modeling*, pp. 470-474, 2007

Herlocker, J. L., Konstan, J. A., and Riedl, J.: Explaining Collaborative Filtering Recommendations, *Proc. of the 2000 ACM Conf. on Computer Supported Cooperative Work*, pp. 241–250, 2000

Sinha, R. and Swearingen, K.: The Role of Transparency in Recommender Systems, *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*, pp. 830-831, 2002

Gedikli, F., Jannach, D., and MouzhiGe: How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems, *Int. J. of Human-Computer Studies*, Vol. 72, No. 4, pp. 367–382, 2014

モデルベース協調フィルタリング

手法	スコア予測	推薦の透明性
行列因子分解	○	△
非負値行列因子分解	○	△
ベイジアンネットワーク	○	○
重回帰分析	○	○

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

- 多数の変数を扱う場合,ベイジアンネットワークは高計算コスト
- 並列化の実装は重回帰分析が容易
- 重回帰分析では,偏回帰係数のより推薦の透明性の実現が可能

研究課題

協調フィルタリングにおける課題

目的変数とするアイテムを評価しているユーザのデータのみを使用
→ 学習に使用できるデータ数が少なくなる傾向

↓

過学習が発生する可能性

過学習：学習データに過剰に適合し,未知のデータに対する予測精度が低下すること

- **正則化**　→ L1正則化とL2正則化を適用
- **次元圧縮**　→ 評価されているところのみを学習する
  Word2Vecを用いることで高速に学習

重回帰分析の協調フィルタリングへの応用

学習用ユーザ–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

次元圧縮
（Word2Vec）

n

k 次元

W

$X' = XW$

圧縮空間

k

m

X'

y

正則化による方法

回帰式（正則化法）

$$\boldsymbol{y} = \alpha_0 + \boldsymbol{X}\boldsymbol{\alpha} + R(\alpha)$$

正則化項

予測モデル（正則化法）

$$\hat{y} = \alpha_0 + \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\alpha}$$

$\boldsymbol{x} \in \mathbb{R}^{n \times 1}$

予測モデル（次元圧縮法）

$$\hat{y} = \beta_0 + \boldsymbol{x}^{\mathrm{T}}\underbrace{\boldsymbol{W}\boldsymbol{\beta}}_{\fallingdotseq \alpha}$$

回帰式（次元圧縮法）

$$\boldsymbol{y} = \beta_0 + \boldsymbol{X'}\boldsymbol{\beta} + R(\beta)$$

正則化項

Word2Vecによる線形写像の学習

	アイテム1	アイテム2	アイテム3
ユーザ1	★★★★☆	★★☆☆☆	
ユーザ2	★★★★★		★★☆☆☆
ユーザ3		★★★★☆	★☆☆☆☆

2値化

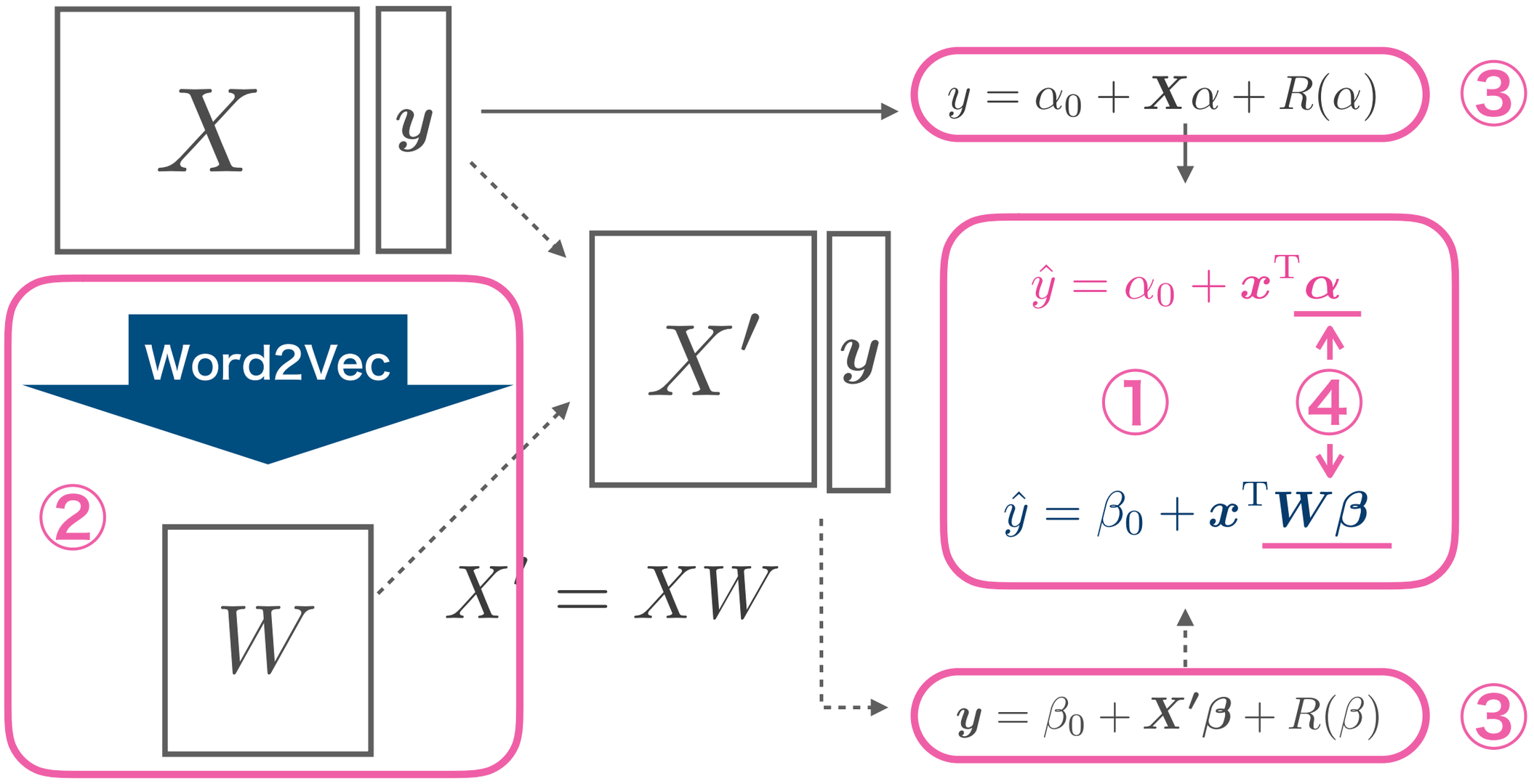
	アイテム1	アイテム2	アイテム3
ユーザ1	1	1	0
ユーザ2	1	0	1
ユーザ3	0	1	1

1: 評価している
0: 評価していない

ユーザ1: {アイテム2,アイテム1}
ユーザ2: {アイテム1,アイテム3}
ユーザ3: {アイテム3,アイテム2}

線形写像の算出

実験内容



①スコア予測能力を検証

②適切なWord2Vecのハイパーパラメータの検証

③回帰式の学習時間の計測

④偏回帰係数の類似度の検証

実験環境（1/2）

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
- ユーザ数: 278,858　書籍数: 271,379　総スコア数: 383,852

Word2Vecのハイパーパラメータ

- 圧縮次元数：20〜200まで20次元ずつ変化させた10種類
- ウィンドウサイズ：2〜10まで2ずつ変化させた5種類

回帰式の構築

目的変数：評価しているユーザの多い書籍上位100件
説明変数：目的変数を除く全書籍　（目的変数毎に回帰式を構築）

実験環境（2/2）

予測精度の評価指標　**MAE(Mean Absolute Error)**

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

y_i：真値　$\hat{y}_i$：予測値　T：テストデータ数

汎化誤差の検証　**ブートストラップ法**

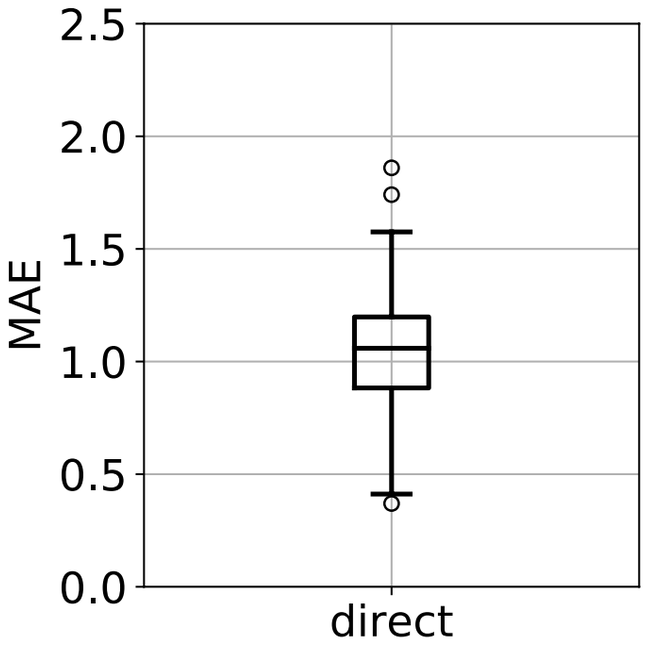
N：ブートストラップ数（N=20）　$\boldsymbol{X}$：学習用ユーザ-アイテム行列

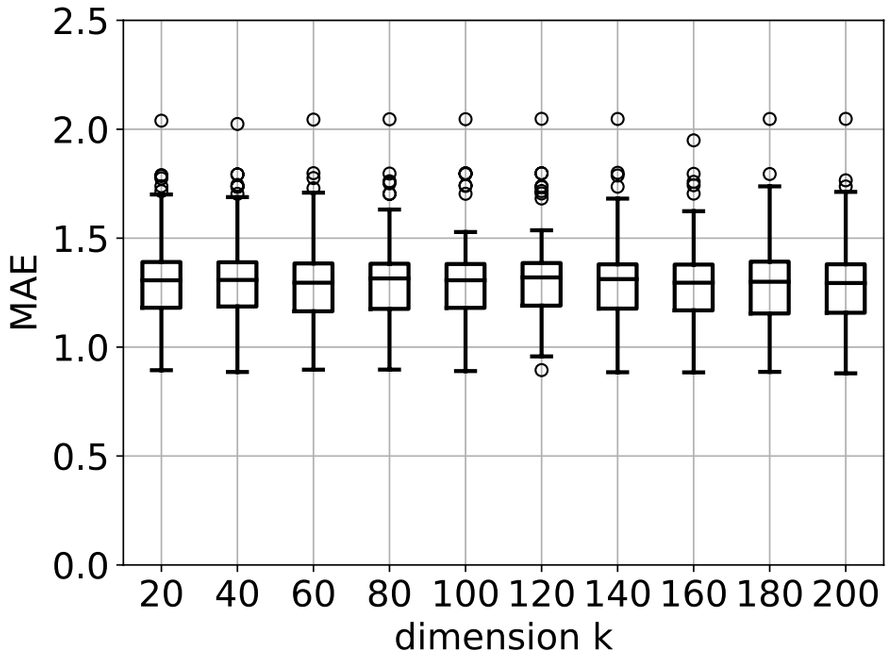
ブートストラップサンプリング集合 $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_i, \cdots, \boldsymbol{X}_N\}$

$$\text{汎化誤差} = E(\boldsymbol{X}, \boldsymbol{X}) + \frac{1}{N}\sum_{i=1}^{N}\{\underbrace{E(\boldsymbol{X}_i, \boldsymbol{X})}_{\text{汎化誤差}} - \underbrace{E(\boldsymbol{X}_i, \boldsymbol{X}_i)}_{\text{経験誤差}}\}$$

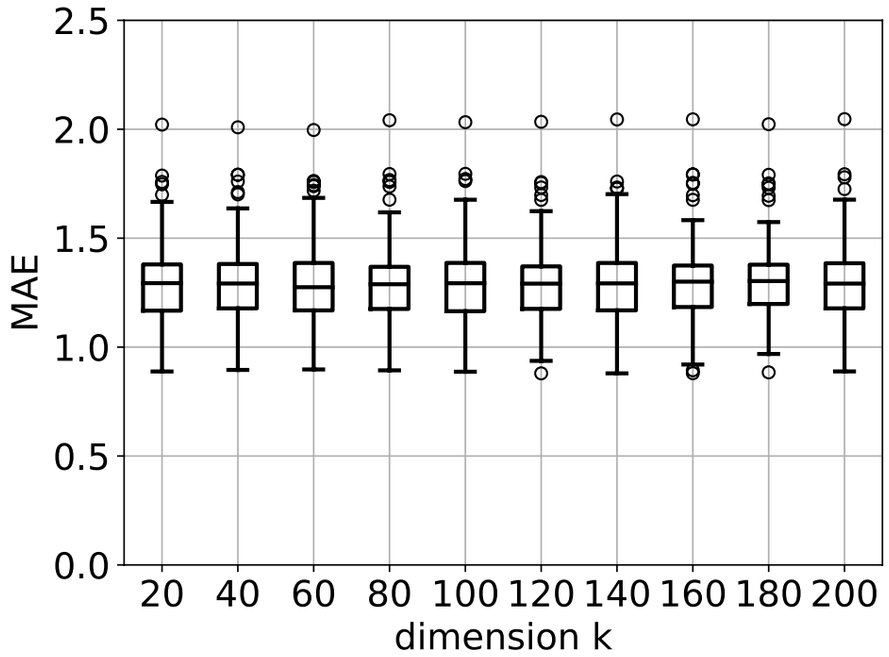
$E(\boldsymbol{X}_i, \boldsymbol{X})$：訓練集合に$\boldsymbol{X}_i$を，テスト集合に$\boldsymbol{X}$を使って推定した予測誤差

L1正則化を適用した正則化法と次元圧縮法のスコア予測精度





ウィンドウサイズ2



ウィンドウサイズ10

正則化法

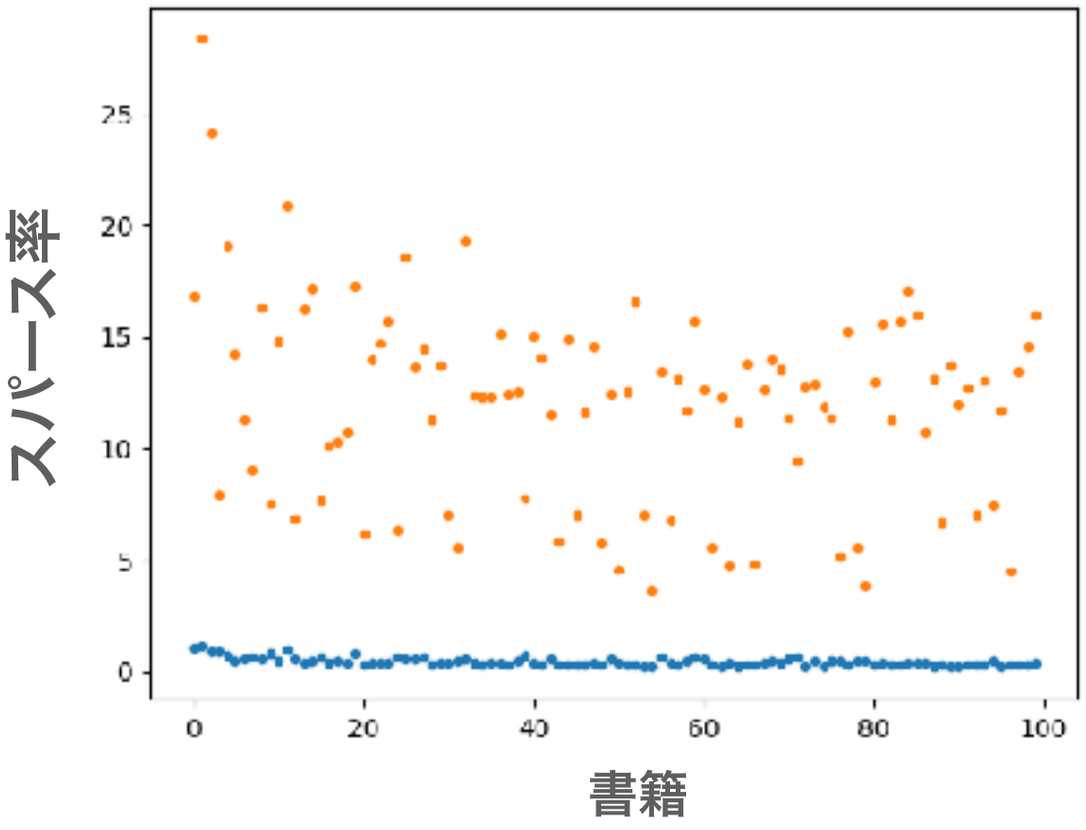
次元圧縮法

- 予測誤差の中央値について,
  - 正則化法は1.08, 次元圧縮法は1.27〜1.32
  - 正則化法のほうが次元圧縮法より0.19~0.24予測誤差が小さい
- 次元圧縮法について, Word2Vecのハイパーパラメータは

  スコア予測に大きく影響を与えていない
- L2正則化も同様の結果を得ている

L1正則化とL2正則化の偏回帰係数の違い

偏回帰係数のスパース率



回帰式

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$$

スパース率の平均

L2正則化 ⟵ 12%

L1正則化 ⟵ 0.44%

スコア予測の計算コストが低くなるため
L1正則化のほうが適している

正則化法と次元圧縮法のモデル学習時間

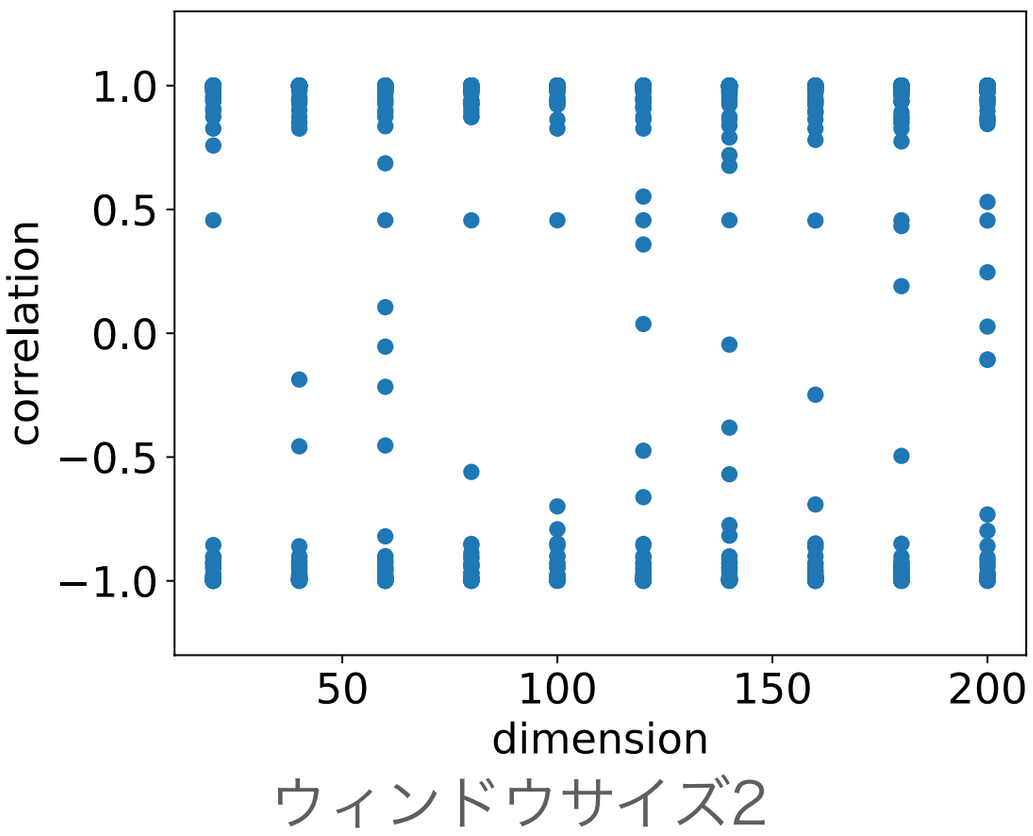
Book-Crossingデータセットを使用した場合

100個の目的変数それぞれの
モデル学習時間の合計

	正則化法	次元圧縮法
L1正則化	34.2 秒	23.6 秒
L2正則化	77.4 秒	23.4 秒

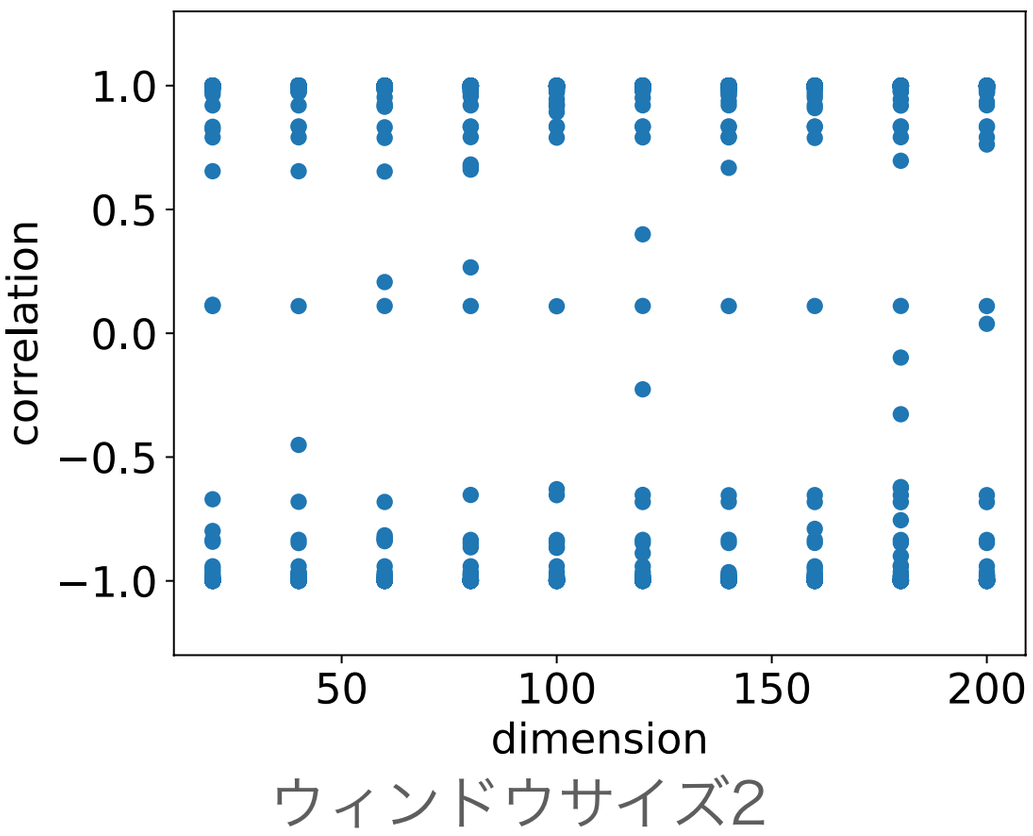
使用計算機
- OS: Ubuntu 16.04.3
- CPU: Intel(R) Xeon(R) CPU E5-2697 v3@ 2.60GHz
- メモリ: 128GB
- ストレージ: 160GB SSD

正則化法と次元圧縮法の偏回帰係数の相関



ウィンドウサイズ2

L1正則化



ウィンドウサイズ2

L2正則化

- 高い正の相関があると同時に高い負の相関もある
  → 必ずしも類似しているとはいえない
- 正則化法の偏回帰係数が正しいと仮定すると,
  次元圧縮法の偏回帰係数は正しいとはいえない

おわりに

目的：モデルベース協調フィルタリングで推薦の透明性を実現

→　一般線形モデルのひとつである重回帰分析の適用

実験結果：

①正則化法と次元圧縮法のスコア予測能力は真値から約1点離れる程度

②Word2Vecのハイパーパラメータはスコア予測に影響を与えない

③本実験の環境下では次元圧縮法のほうがモデル学習時間が短い

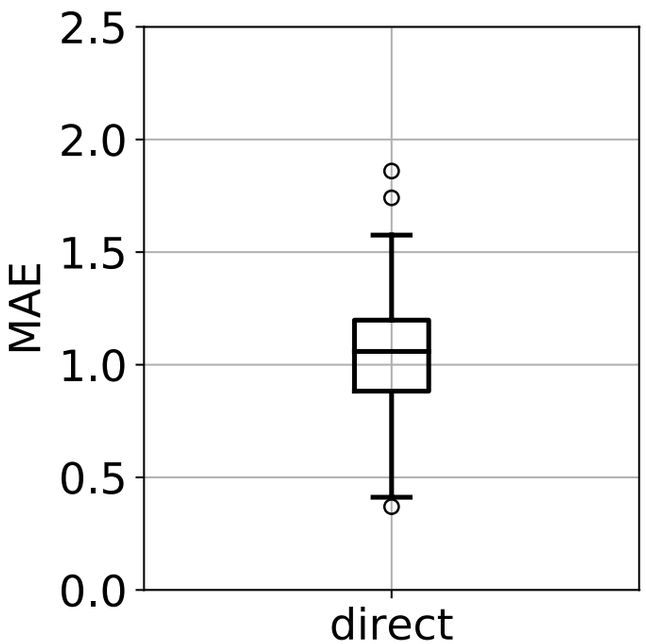
④正則化法と次元圧縮法の偏回帰係数は必ずしも類似しているとはいえない

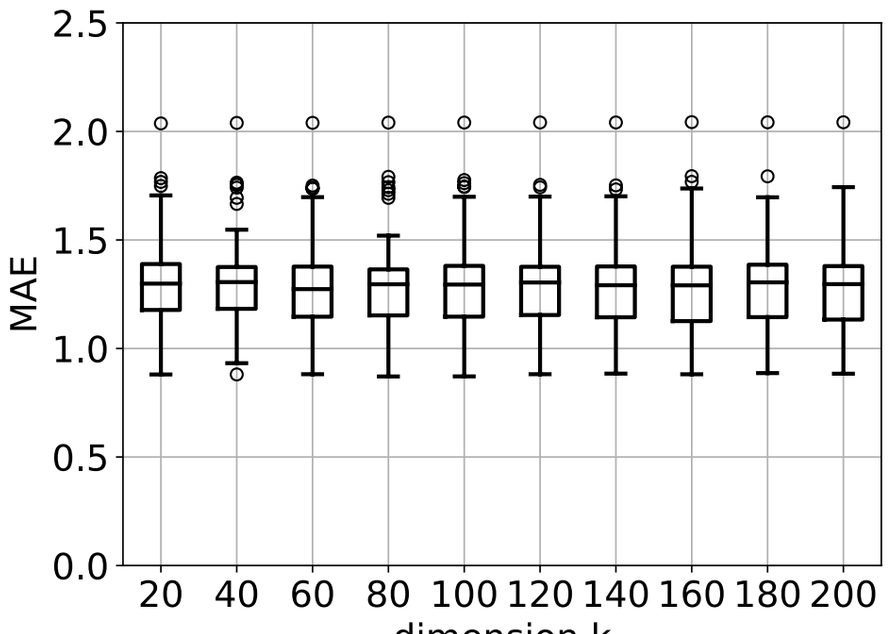
今後の展望

- 他の手法とスコア予測能力の比較
- 偏回帰係数の解釈の仕方を明らかにする

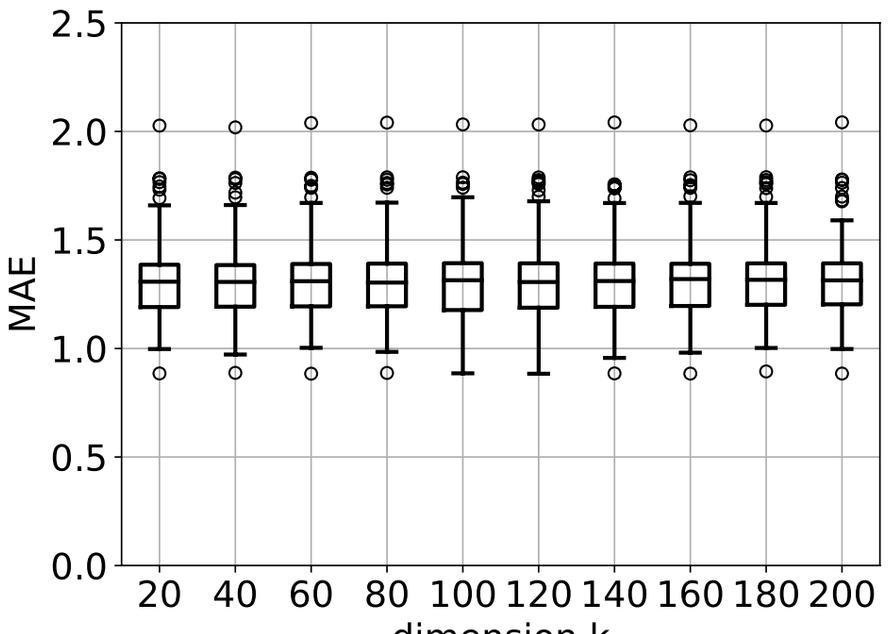
質問用スライド

L2正則化を適用した正則化法と次元圧縮法のスコア予測精度





ウィンドウサイズ2



ウィンドウサイズ10

正則化法

次元圧縮法

- 予測誤差の中央値について,
  - 正則化法は1.09, 次元圧縮法は1.27～1.32
  - 正則化法のほうが次元圧縮法より0.18~0.23予測誤差が小さい
- 次元圧縮法について, Word2Vecのハイパーパラメータは
  スコア予測に大きく影響を与えていない

L1正則化とL2正則化

誤差関数: $R(\boldsymbol{\alpha}) = (\boldsymbol{y} - \hat{\boldsymbol{y}})^2$

- L1正則化

$$R(\boldsymbol{\alpha}) = (\hat{\boldsymbol{y}} - \boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha})^2 + \lambda \underline{||\boldsymbol{\alpha}||_1}$$

L1ノルム

- L2正則化

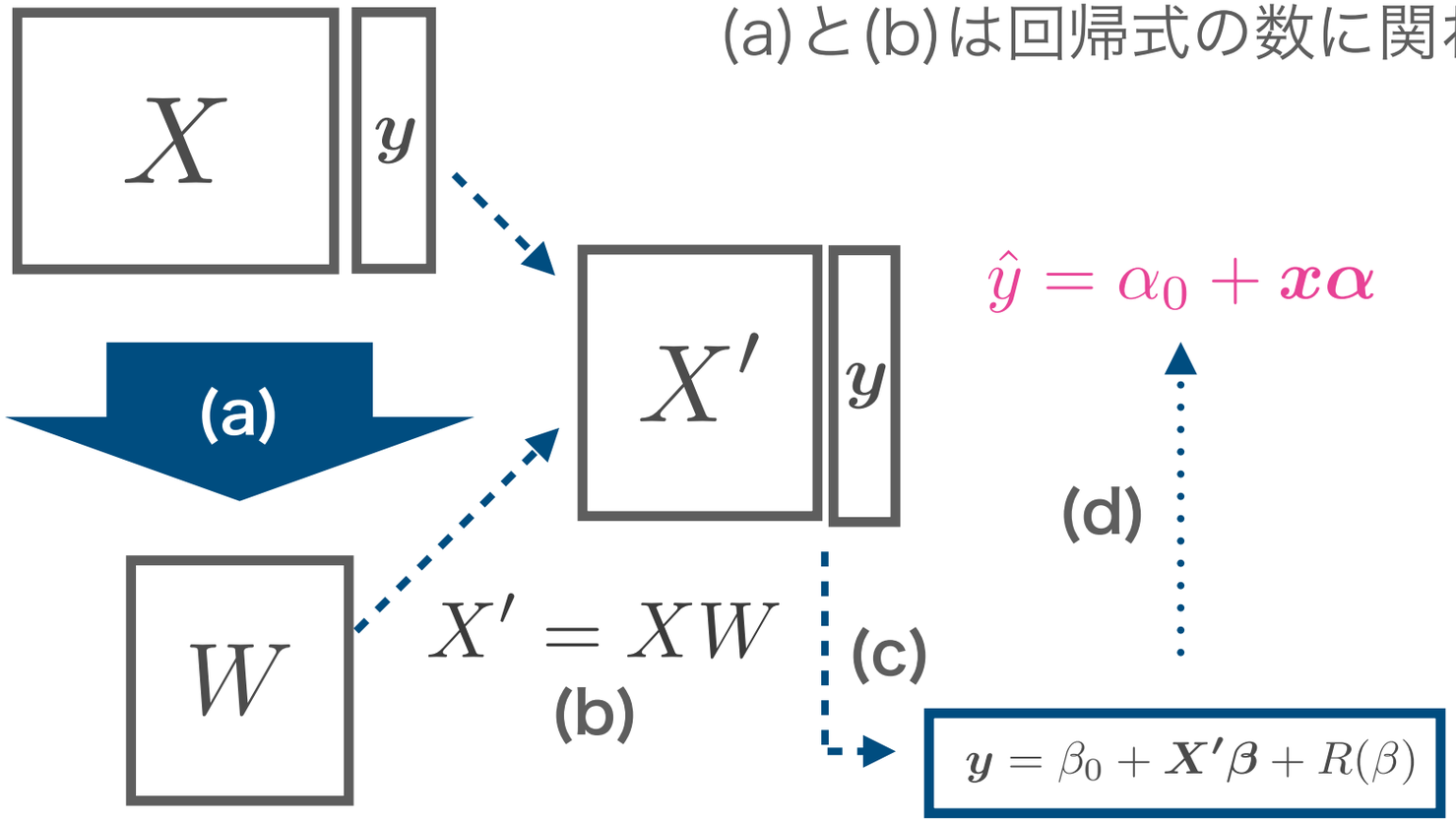
$$R(\boldsymbol{\alpha}) = (\hat{\boldsymbol{y}} - \boldsymbol{\alpha}_0 - \boldsymbol{X}\boldsymbol{\alpha})^2 + \lambda \underline{||\boldsymbol{\alpha}||_2}$$

L2ノルム

正則化法と次元圧縮法のモデル学習時間

	正則化法	次元圧縮法				
		(a)	(b)	(c)	(d)	合計
L1正則化	34.2	0.347	20.0	0.800	2.50	23.6
L2正則化	77.4	0.347	20.0	0.800	2.30	23.4

(a)と(b)は回帰式の数に関わらず共通



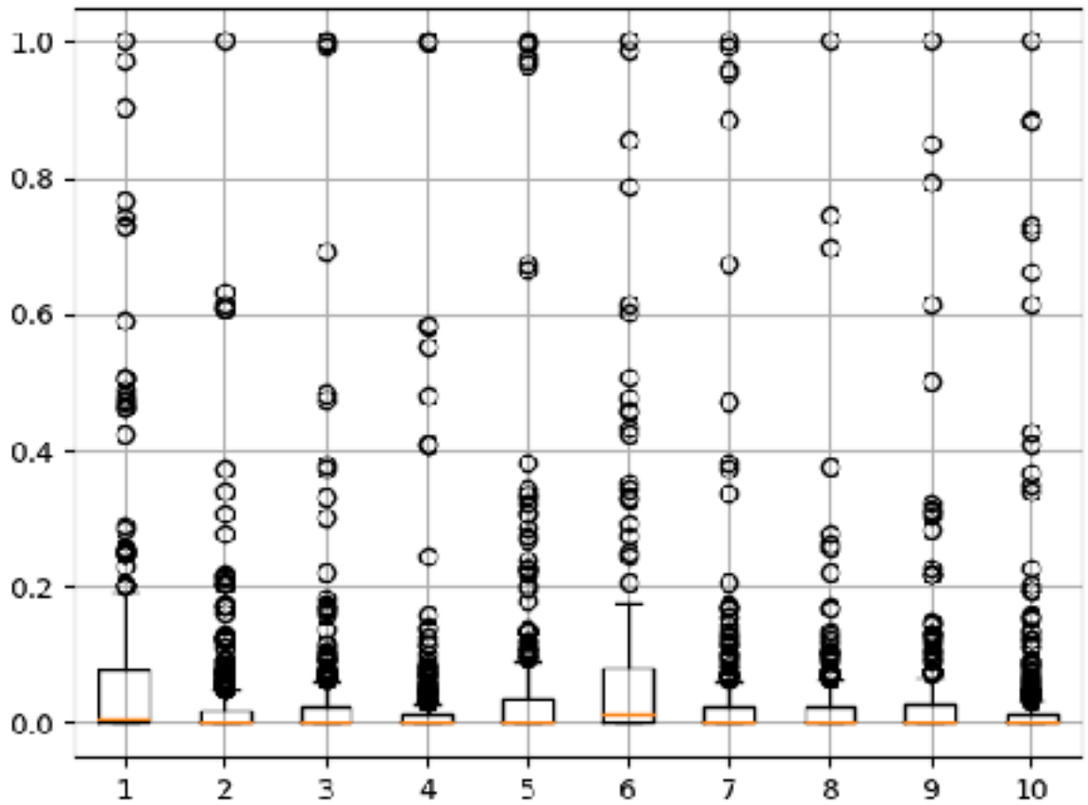
正則化法

$$t = 0.342n$$

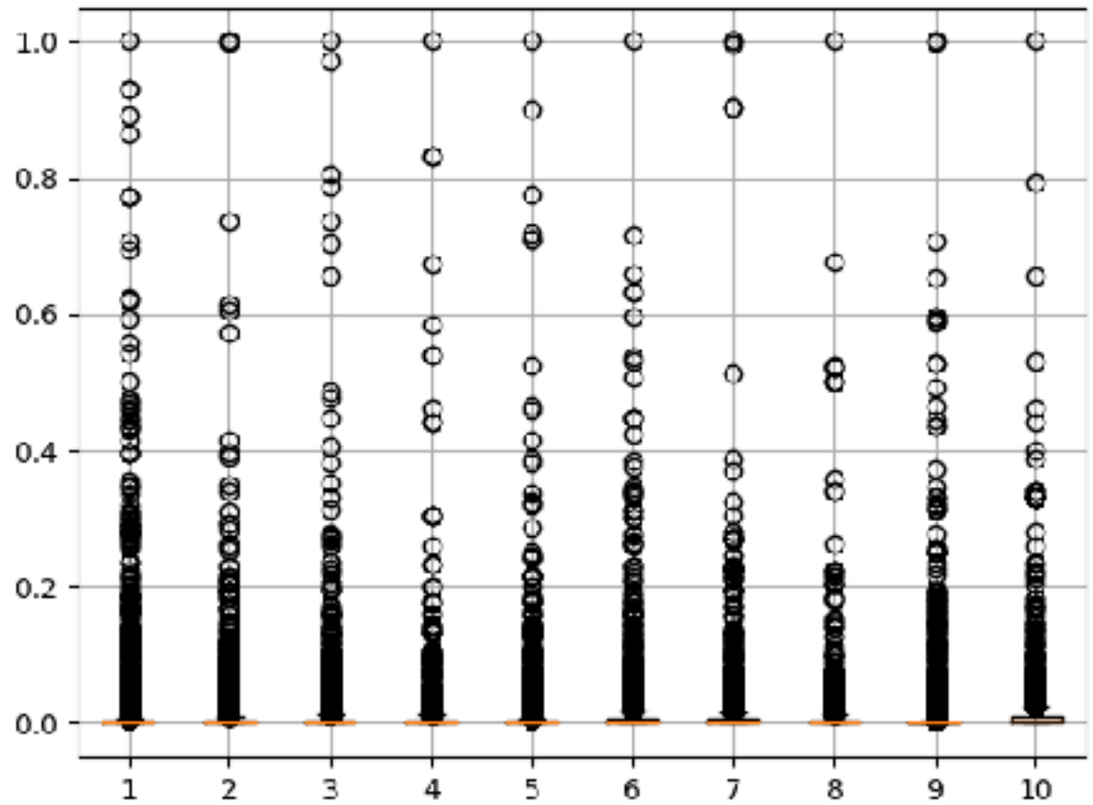
次元圧縮法

$$t = 20.3 + 0.0330n$$

L1正則化とL2正則化の偏回帰係数



L1正則化



L2正則化

正則化法と次元圧縮法の偏回帰係数

正則化法と次元圧縮法で検証

偏回帰係数の相関の良し悪しに規則性がない

同じ回帰式で異なる学習データ・テストデータを与えた場合

- **正則化法で検証**
  - → 偏回帰係数は安定して高い正の相関がある
  - → **ブートストラップサンプリング毎の予測モデルに違いはない**

- **次元圧縮法で検証**
  - → 偏回帰係数の一部には高い正の相関,一部には高い負の相関
  - → **次元圧縮法の予測モデルは安定していない**