

A7-5

LMは日本の時系列構造を どうエンコードするか

*佐々木 睦史¹

鴨田 豪¹・高橋 良允¹・Benjamin Heinzerling^{2, 1}・坂口 慶祐^{1, 2}

1 東北大学 2 理化学研究所

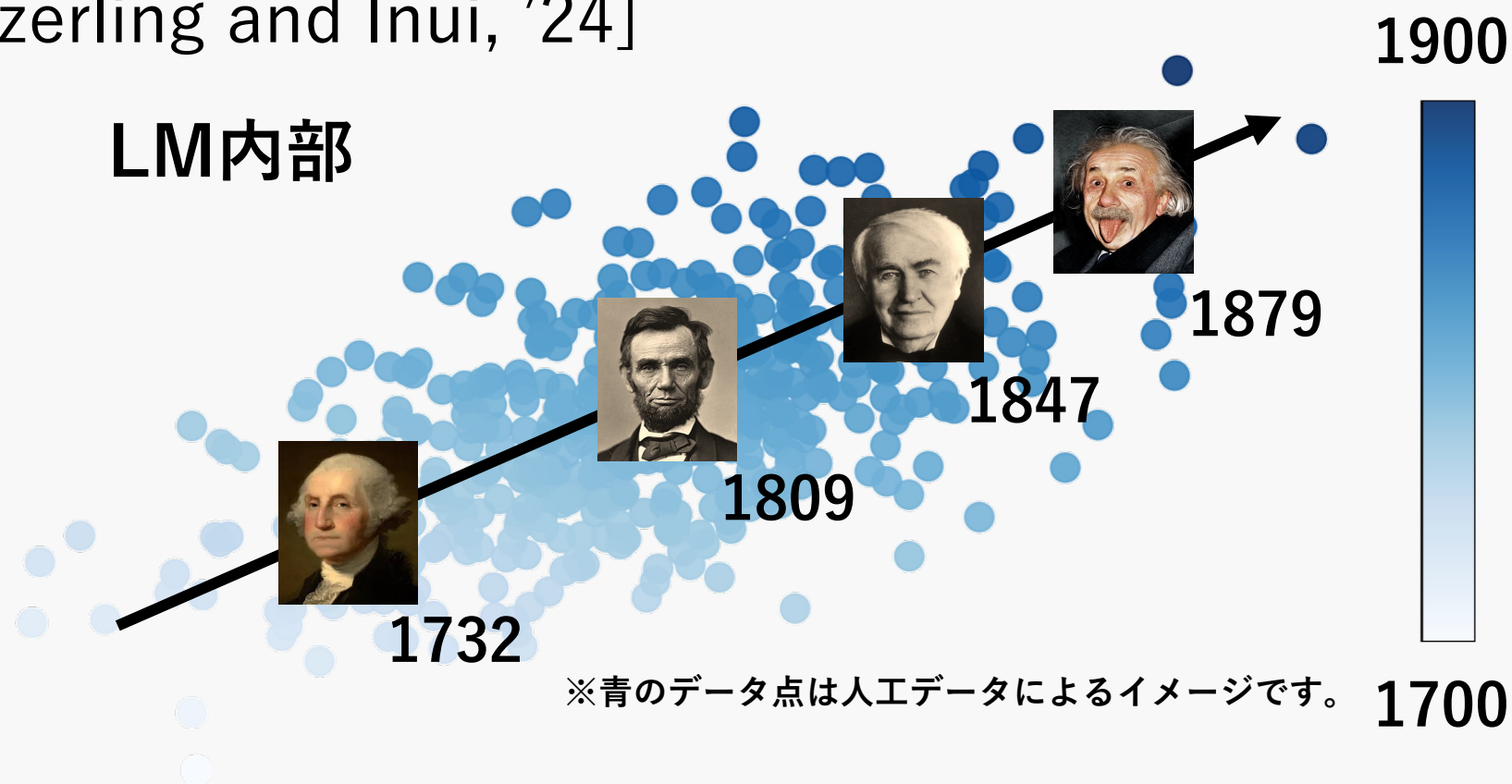
言語処理学会第31回年次大会 / 2025年3月12日

LM内部の知識構造に関する研究

- LMは知識を内部でどう表現しているかが研究されている
 - 数値情報 [Heinzerling and Inui, '24]
 - 地理的情報 [Gurnee and Tegmark, '24]
 - 色 [Patel and Pavlick, '22]
- LM内部の知識構造を調べることは内部機序の理解につながる
 - LMの推論メカニズムの解明、透明性の向上

先行研究：LMが生年の知識を構造に反映する

- 西洋の人物の内部表象から単調な時系列方向を取り出せる
[Heinzerling and Inui, '24]

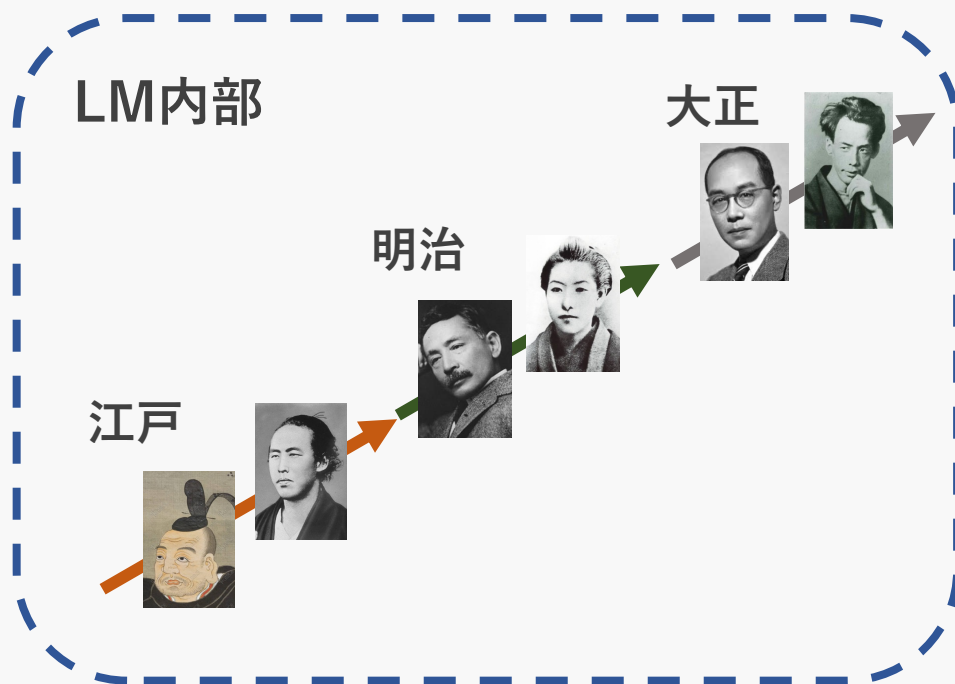


本研究の対象：LMが表現する日本の時系列構造

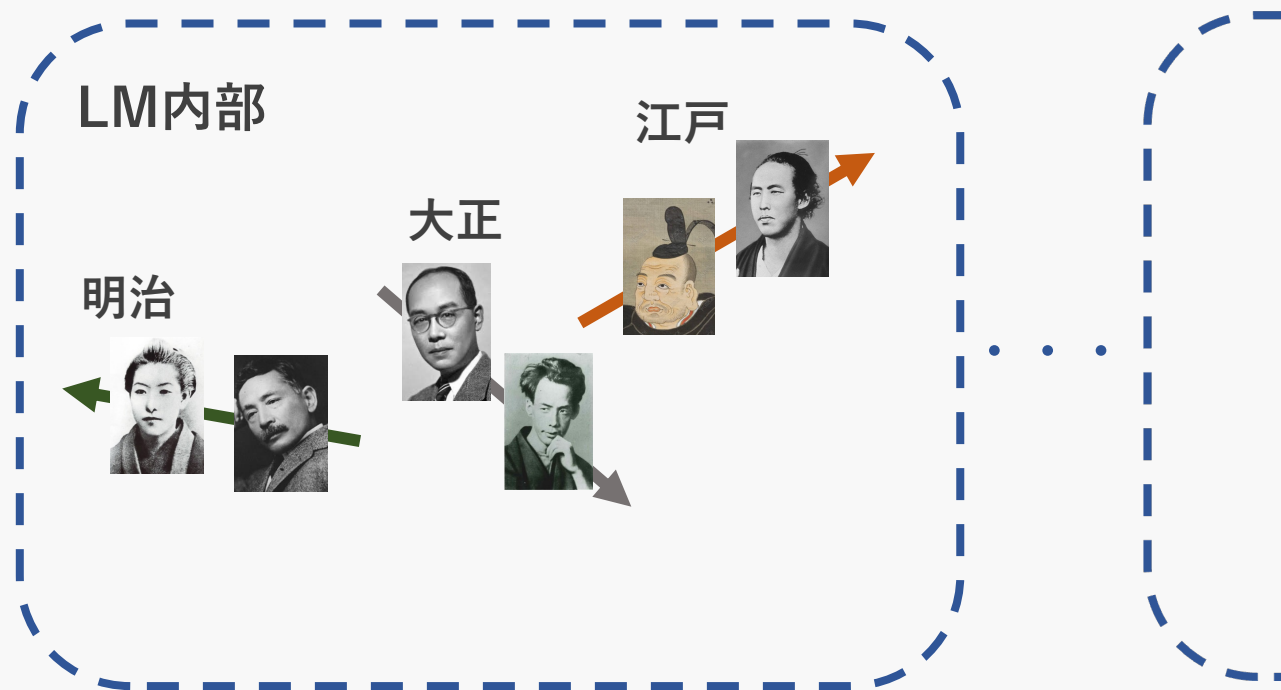
- 日本の時系列の特徴
 - 同じ年を表す複数の表現方法（**西暦**、**和暦**）が存在
 - 歴史が長く、**時代区分が明確**
- 本研究の問い
 - LMの日本の時系列表現はどう取得できるか
 - 取得した時系列表現はLM内部でどう表現されるか

本研究のアプローチ：時代間の表現の違いに着目

可能性1. 時代間でも方向と位置が揃う



可能性2. 時代間では位置と方向がバラバラ



時代ごとに人物と生年のデータを取得

- WikiDataを利用
 - 時代ごとに人物と生年を取得
- 有名な人物のデータに絞る
 - モデル (Swallow-13b) が正しい生年を回答できるデータを抽出
 - 件数は最も少ない大正で555件
- 時系列全体データの作成

人物	生年	時代
葛飾北斎	1760	江戸
松尾芭蕉	1644	江戸
⋮	⋮	江戸
樋口一葉	1872	明治
芥川龍之介	1892	明治
⋮	⋮	明治
⋮	⋮	⋮
桐生祥秀	1995	平成
高橋みなみ	1991	平成
⋮	⋮	平成

時代ごとにLMの隠れ状態を収集

- 時代ごとにLMに生年を問い合わせ、人物の隠れ状態を集める
 - 入力：「何年に（人名）が生まれましたか？」
 - 収集する隠れ状態の位置（40層のSwallow-13bの場合）
 - 第24層（相対位置0.6）
 - トークン位置は人名に続く助詞の「が」の位置

層の位置、トークン位置を探索した結果、この位置に時系列情報が最も含まれる（詳細は付録参照）

PLSを用いて隠れ状態から時代表現を抽出

- PLS (部分的最小二乗法) のステップ
 - 時代の d 次元隠れ状態 X から y (生年) の情報を用いて回帰に重要な r 個の成分に次元圧縮
 - r 個の成分で y との最小二乗法を行い回帰する
- PLSの目標
 - 時代ごとの隠れ状態から生年に回帰する時代ベクトル $w \in \mathbb{R}^{1 \times d}$ を抽出

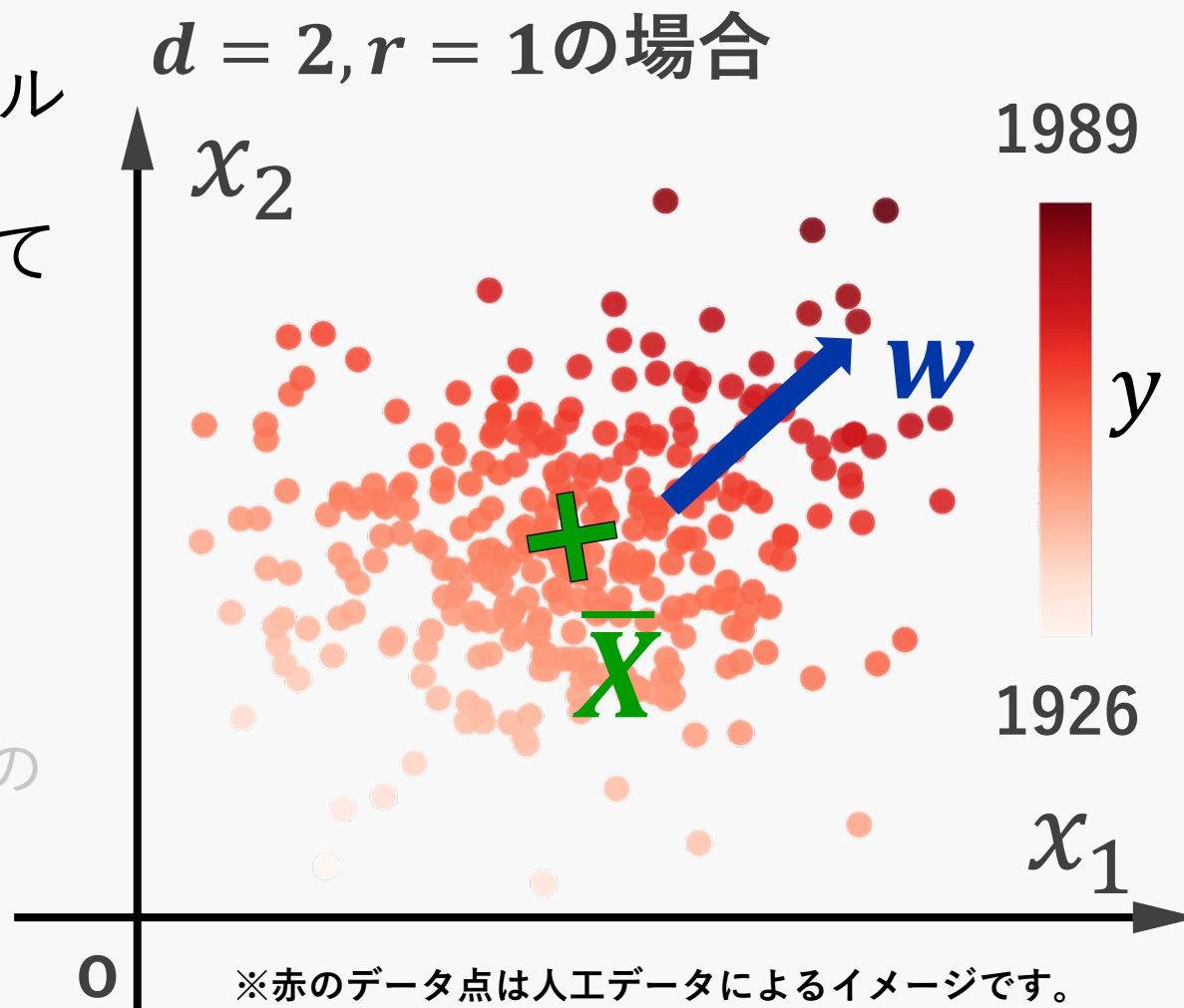
$$y = Xw^T$$

PLSを用いて隠れ状態から時代表現を抽出

- 時代位置 $\bar{X}$ は X の平均ベクトル
- 時代ベクトル w はPLSを用いて抽出した時代方向
- 時代表現 $\hat{X}$ は $\bar{X}$ を中心とする X の w への写像

$\hat{X}$ は X から時系列情報を抽出したもの

➡ 時代間の $\hat{X}$ を同一のLM空間で比較できる

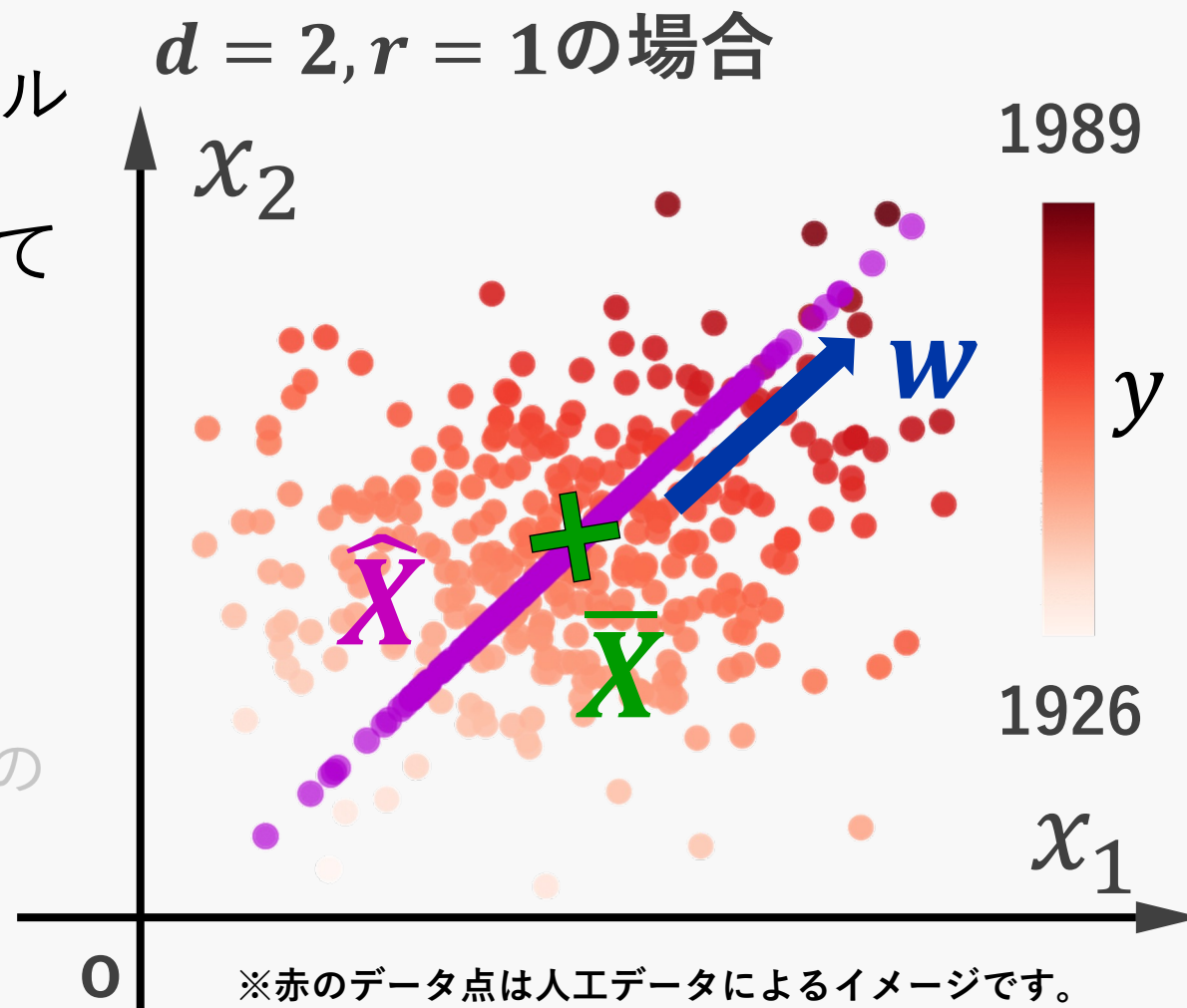


PLSを用いて隠れ状態から時代表現を抽出

- 時代位置 $\bar{X}$ は X の平均ベクトル
- 時代ベクトル w はPLSを用いて抽出した時代方向
- 時代表現 $\hat{X}$ は $\bar{X}$ を中心とする X の w への写像

$\hat{X}$ は X から時系列情報を抽出したもの

→ 時代間の $\hat{X}$ を同一のLM空間で比較できる

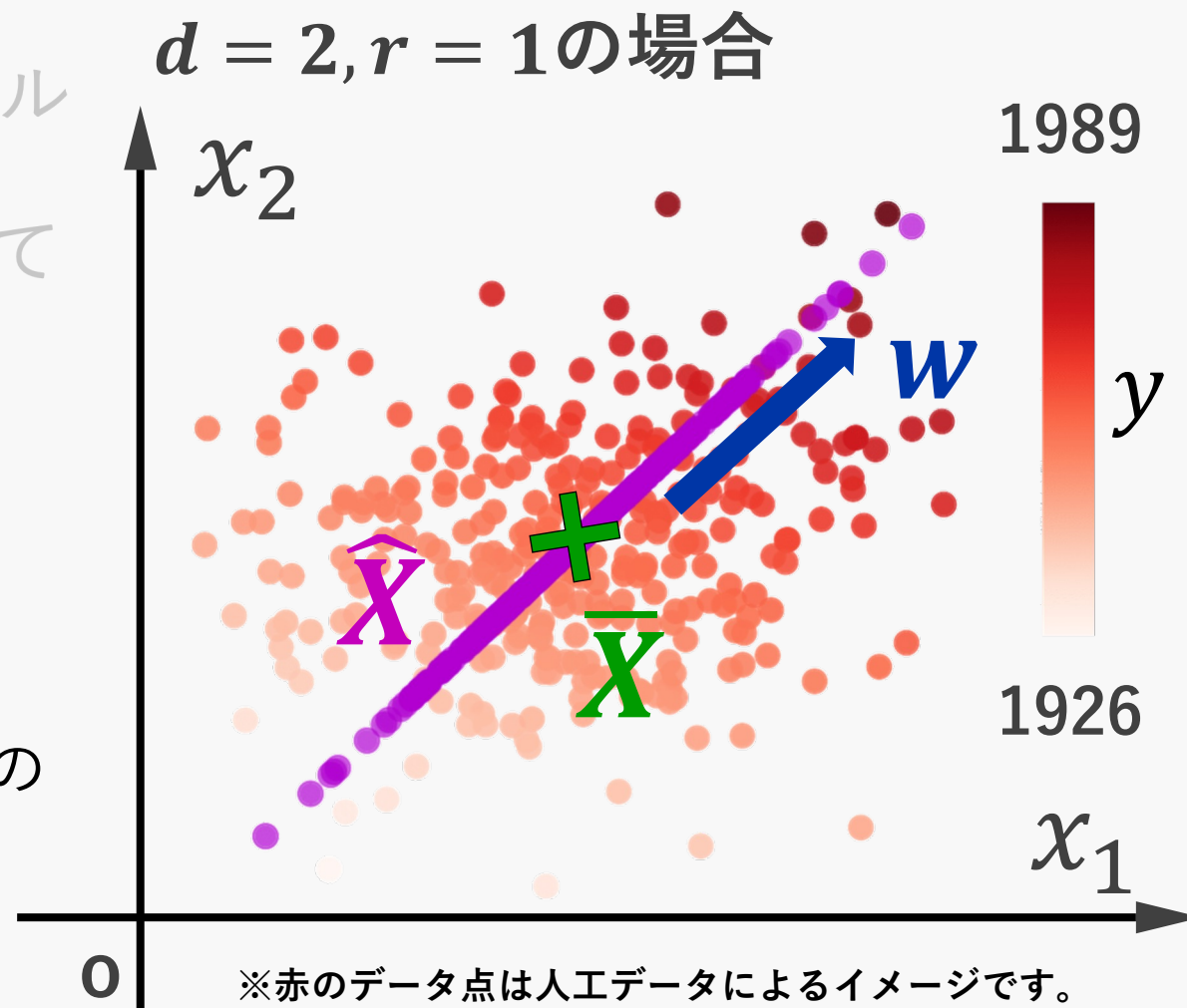


PLSを用いて隠れ状態から時代表現を抽出

- 時代位置 $\bar{X}$ は X の平均ベクトル
- 時代ベクトル w はPLSを用いて抽出した時代方向
- 時代表現 $\hat{X}$ は $\bar{X}$ を中心とする X の w への写像

$\hat{X}$ は X から時系列情報を抽出したもの

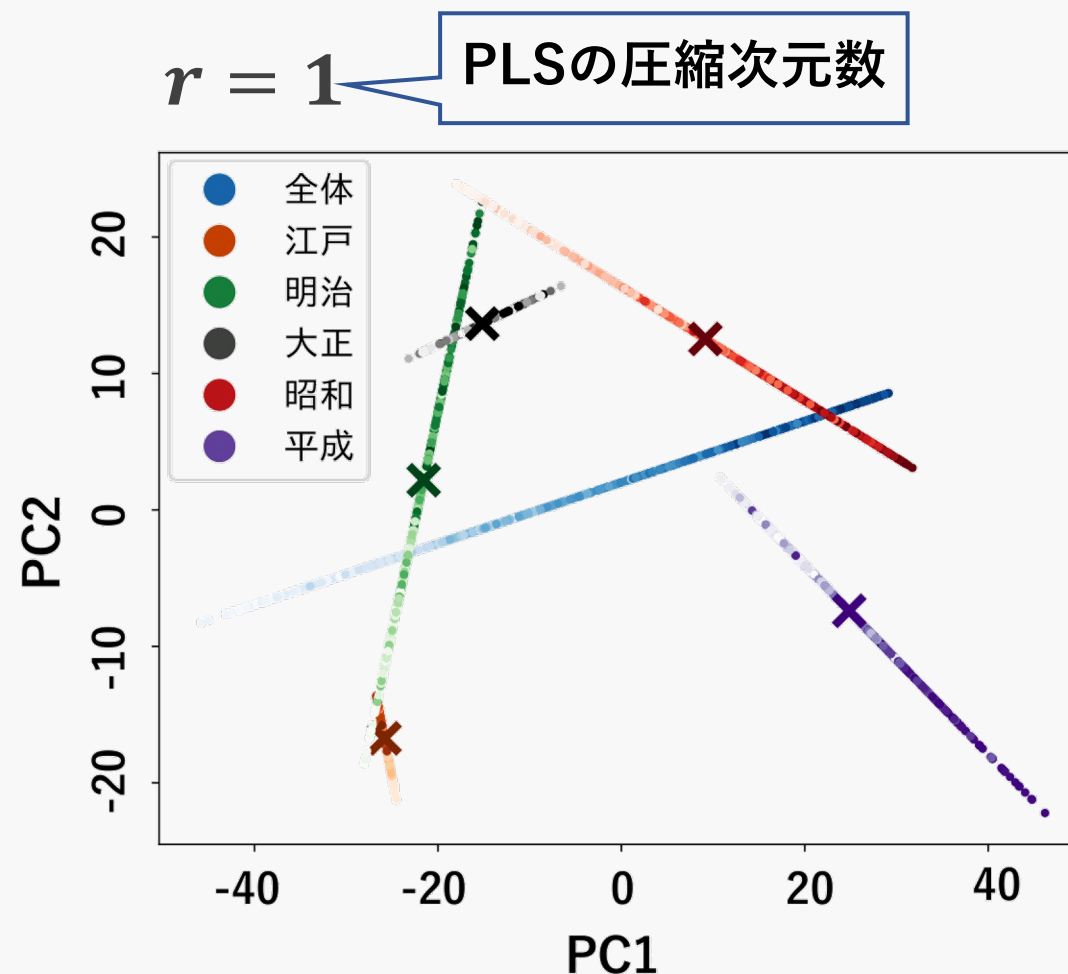
➡ 時代間の $\hat{X}$ を同一のLM空間で比較できる



時代表現 $\hat{X}$: 時系列構造の全体像が示される

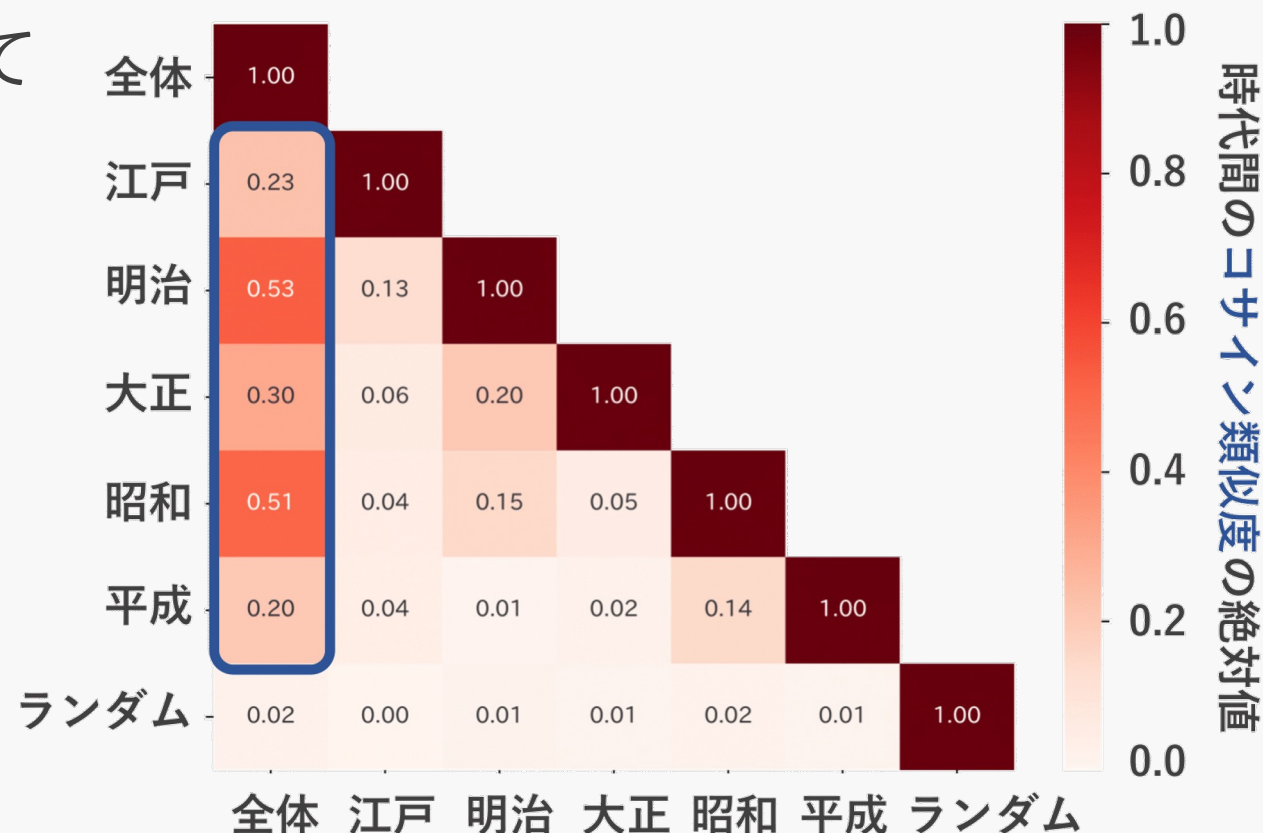
江戸から平成の $\hat{X}$ をPCAで可視化

- 時代位置 $\bar{X}$ は江戸から平成の順
 - 位置はバツ印 $\times$
- 全体の方向に緩やかに沿って
各時代表現 $\hat{X}$ が並ぶ



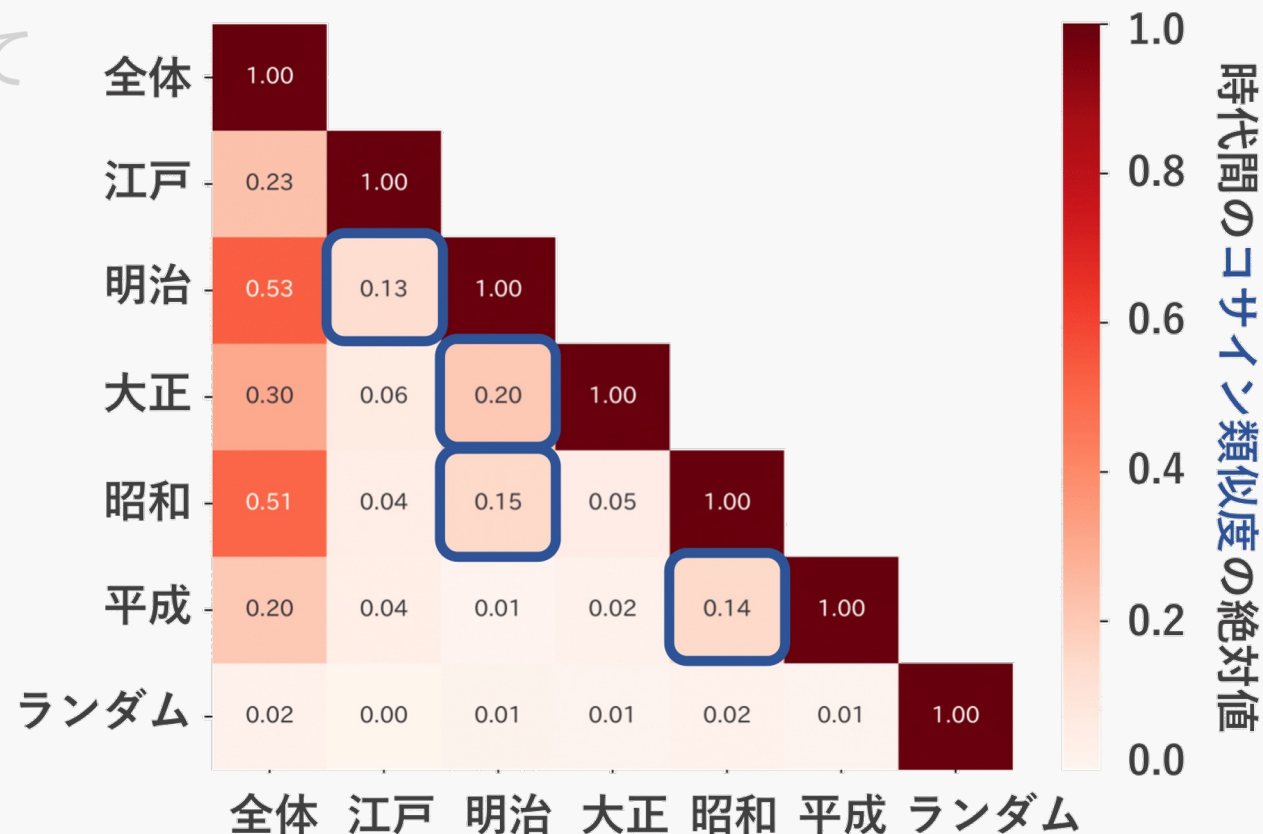
時代ベクトル w : 全体と各時代は類似性あり

- 全体と各時代で**西暦**に関して類似した表現を持つ可能性
- 隣接した時代間では**一部の特徴が共通**している可能性
- 他の時代間は**表現が異なる**



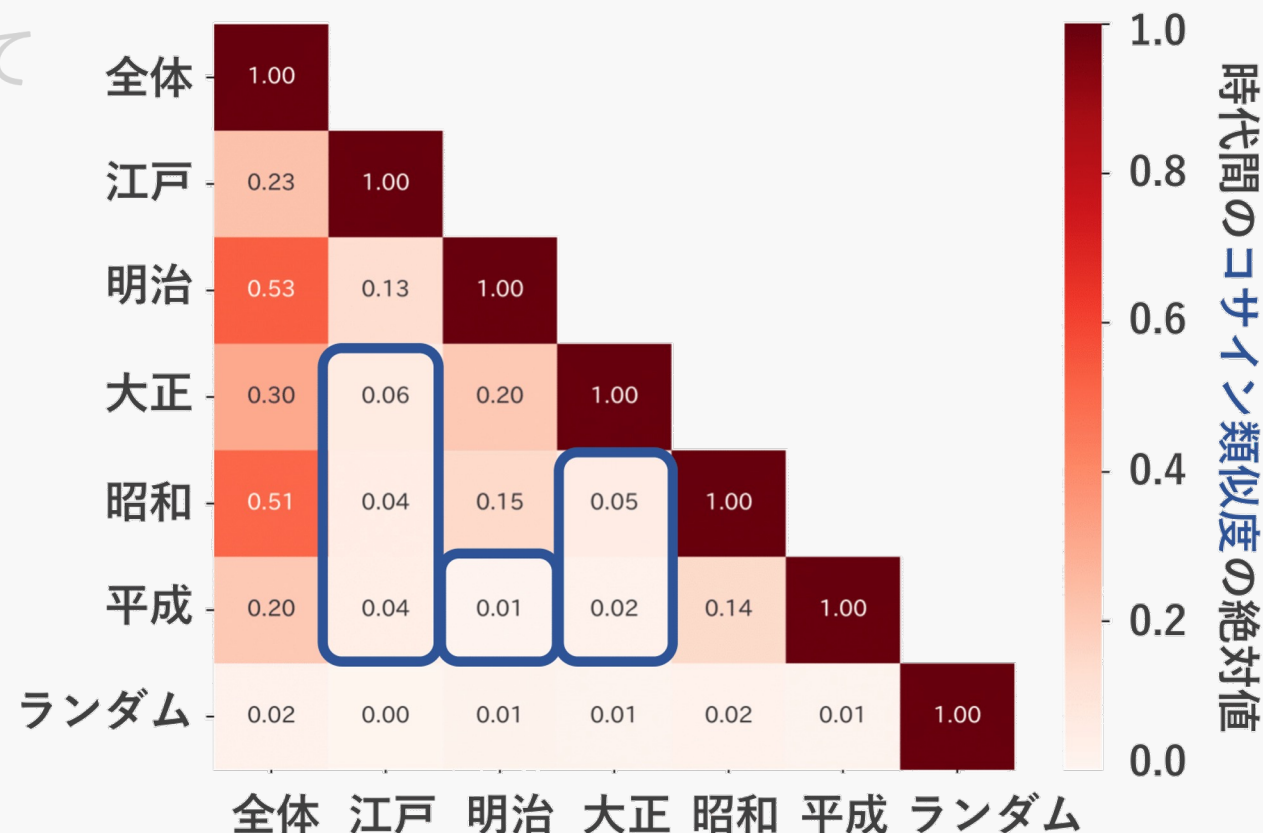
時代ベクトル w : 一部の時代間に僅かな類似性

- 全体と各時代で西暦に関して類似した表現を持つ可能性
- 隣接した時代間では**一部の特徴が共通**している可能性
- 他の時代間は**表現が異なる**



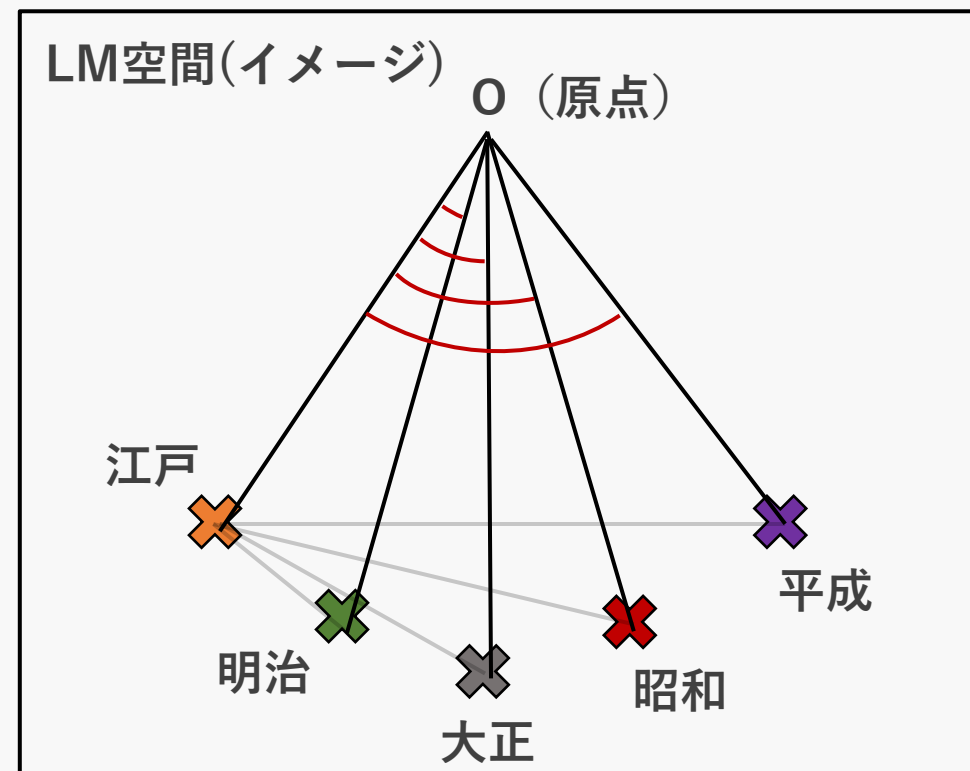
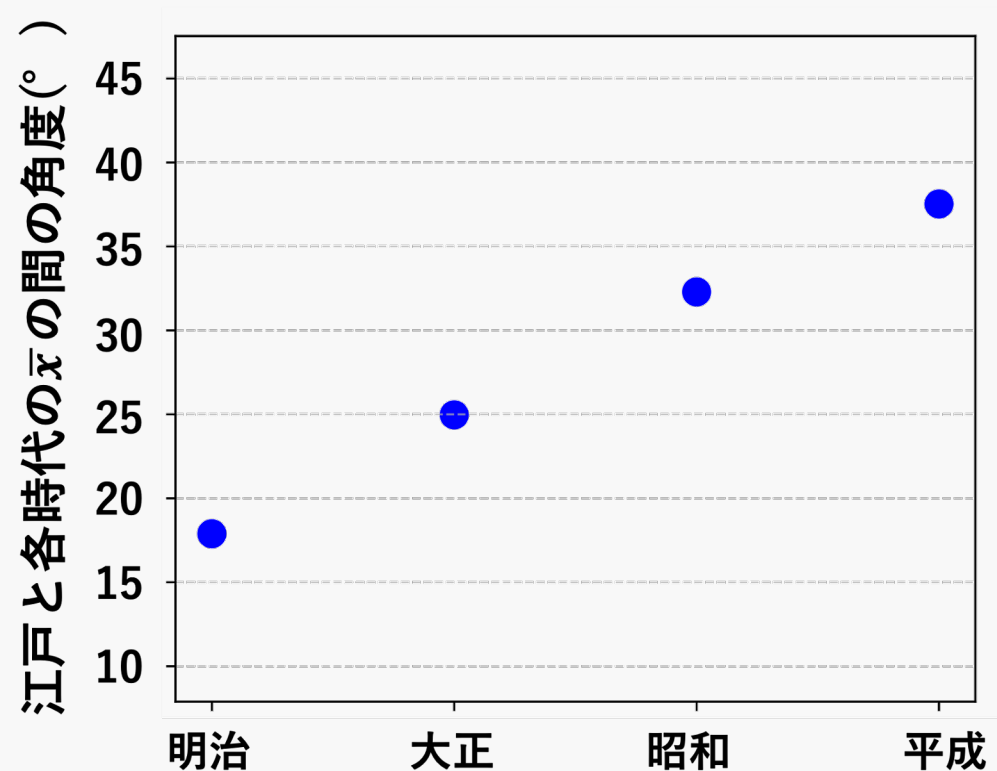
時代ベクトル w : その他の時代間に類似性なし

- 全体と各時代で西暦に関して類似した表現を持つ可能性
- 隣接した時代間では一部の特徴が共通している可能性
- 他の時代間は表現が異なる



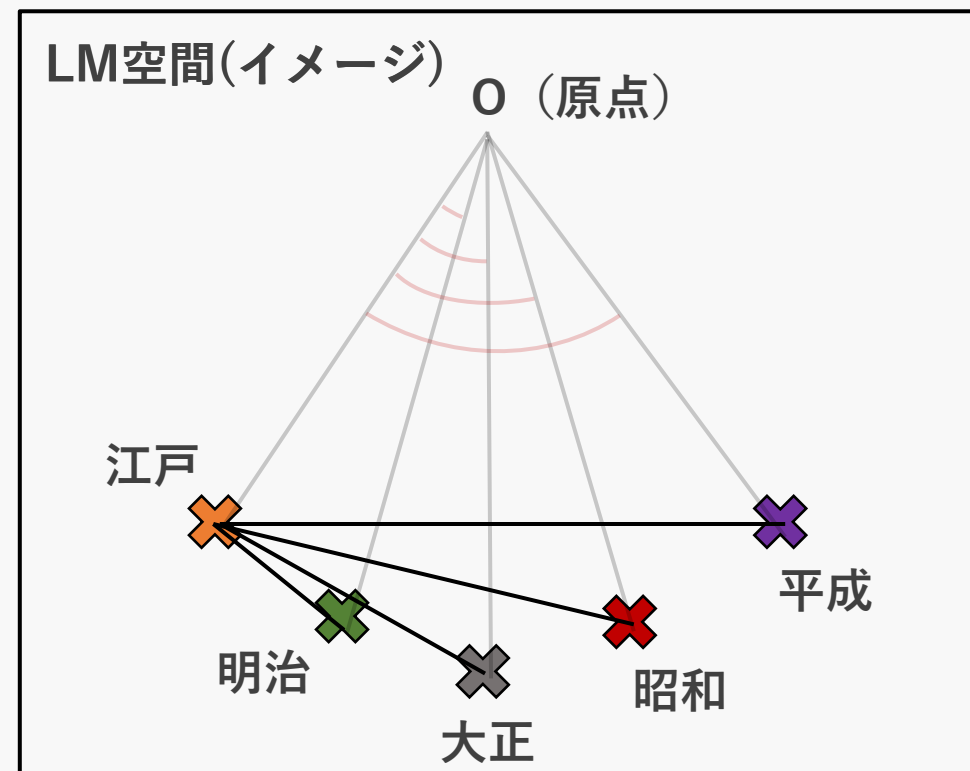
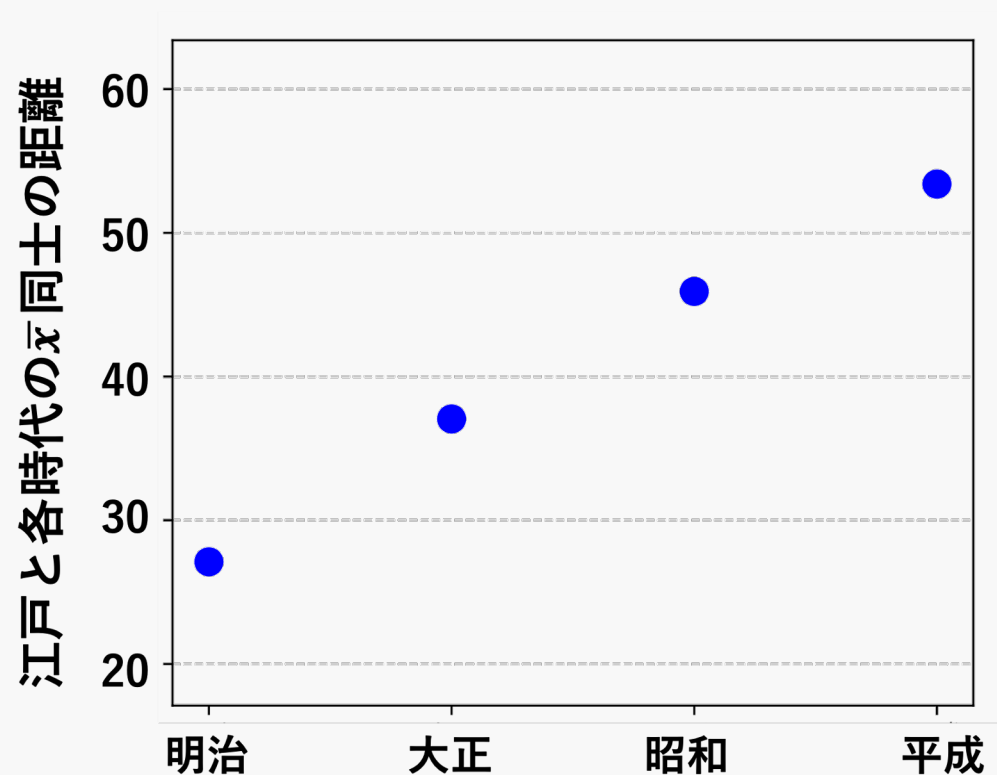
時代位置 $\bar{x}$: 江戸から平成の順に並ぶ

- 時代位置 $\bar{x}$ の時代同士の角度比較



時代位置 $\bar{x}$: 江戸から平成の順に並ぶ

- 時代位置 $\bar{x}$ の時代同士の距離比較



結論

- 日本の時系列表現はどう取得できるか
 - 時代ベクトル w 、時代位置 $\bar{x}$ 、時代表現 $\hat{x}$ により取得できる
- 取得した時系列表現はLM内部でどう表現されるか
 - 時代ベクトルは全体と各時代、隣接時代間は類似し、他は類似しない
 - 時代位置は江戸から平成の順に並ぶ
 - 時代表現は時系列全体の方向に緩やかに沿って並ぶ
- 把握した時系列構造がLMの推論にどう影響するかは今後の展望

| 付録

時系列情報が含まれる隠れ状態はどの位置か？

- 生年に回帰できる隠れ状態を探索

- 回帰手法：PLS
- プロンプト

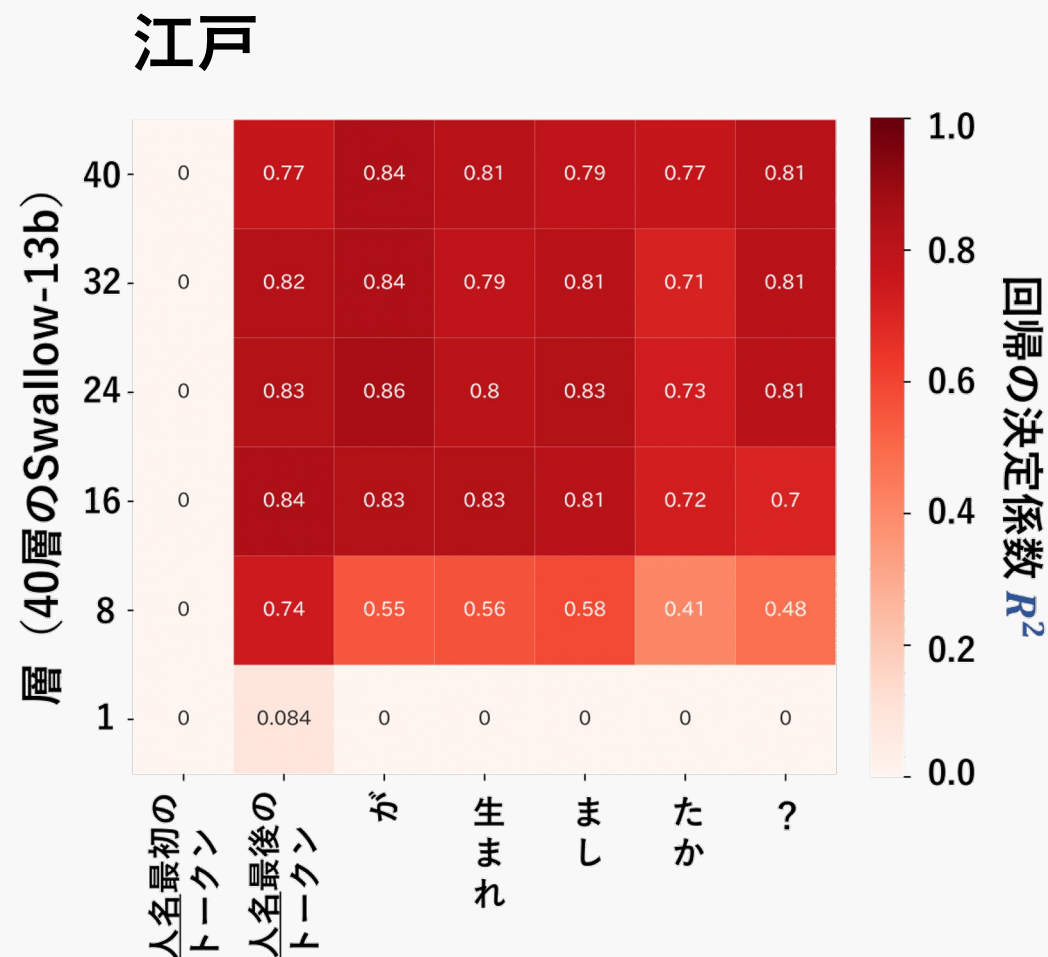
何年に(人名)が生まれましたか？

- 指標：回帰の決定係数 R^2

ある位置の隠れ状態の R^2 が高い

→ 隠れ状態が生年に回帰できる

→ その位置に時系列情報が含まれる

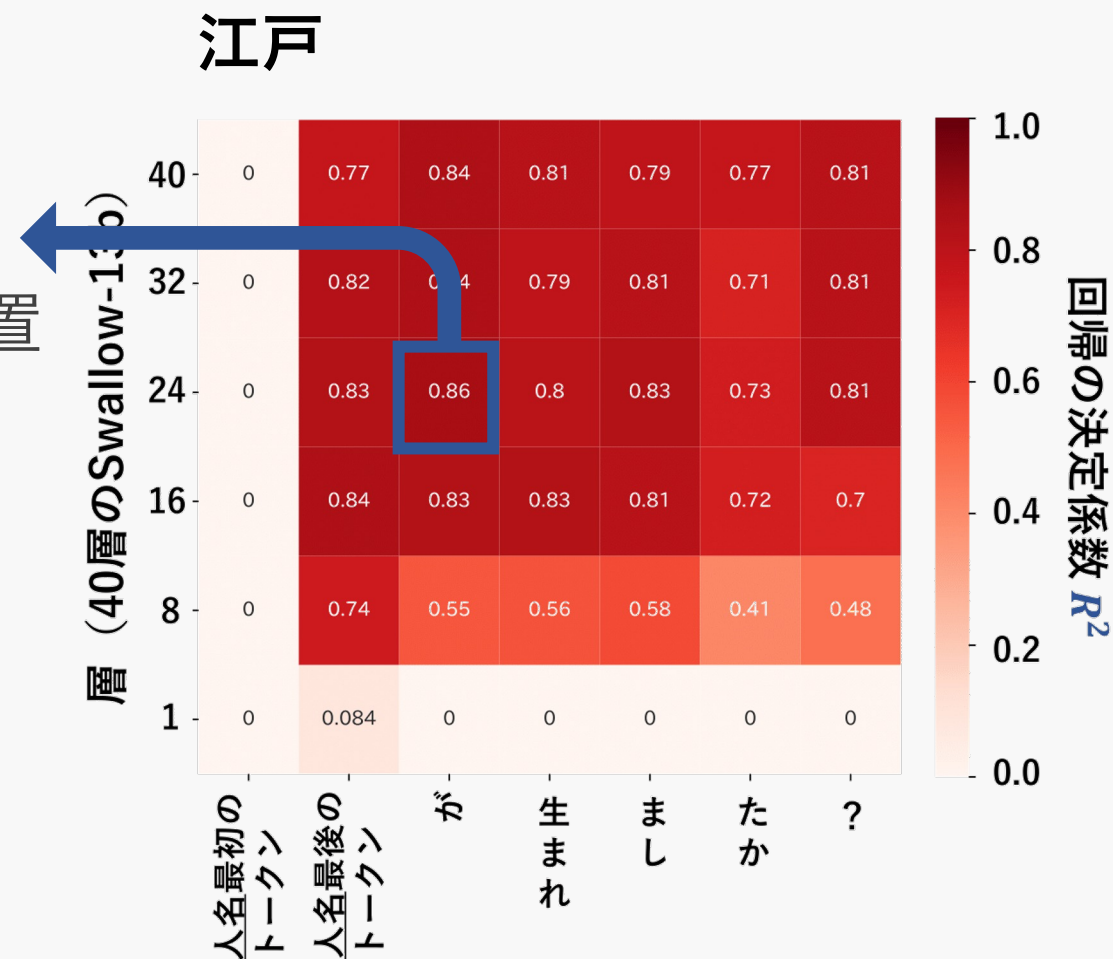


時系列情報が含まれる隠れ状態はどの位置か？

- R^2 が高い隠れ状態の位置
 - 第24層 (40層のSwallow-13b)
 - 人名に続く「が」のトークン位置
 - 江戸以外も同様の傾向

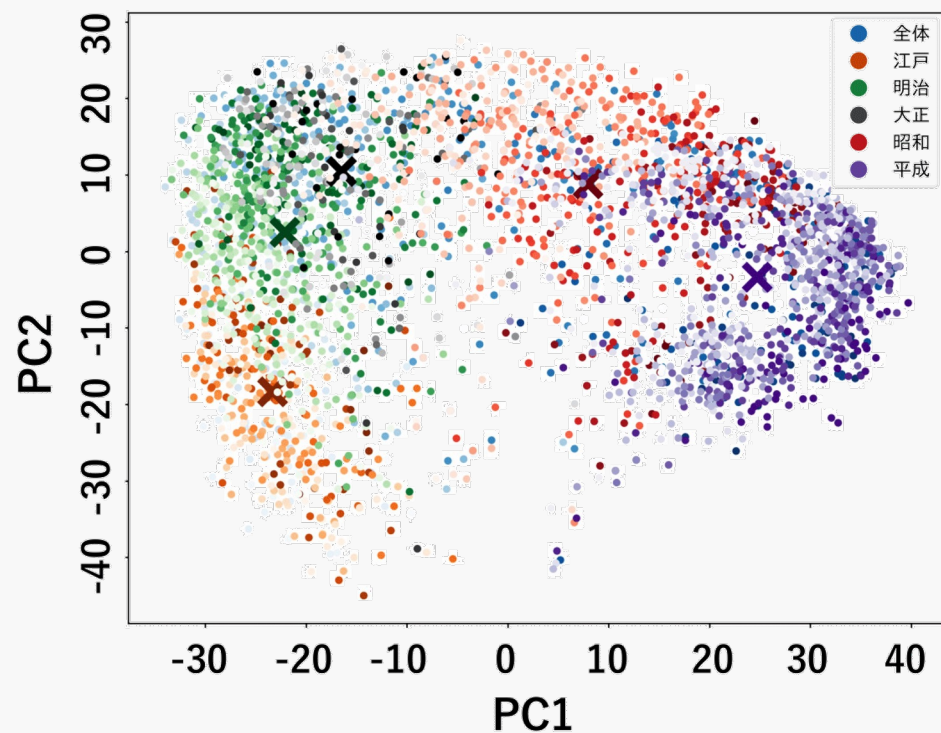


本実験ではこの位置の隠れ状態を用いて時系列情報を抽出した



時代表現抽出の有効性の検証

隠れ状態 X をそのまま

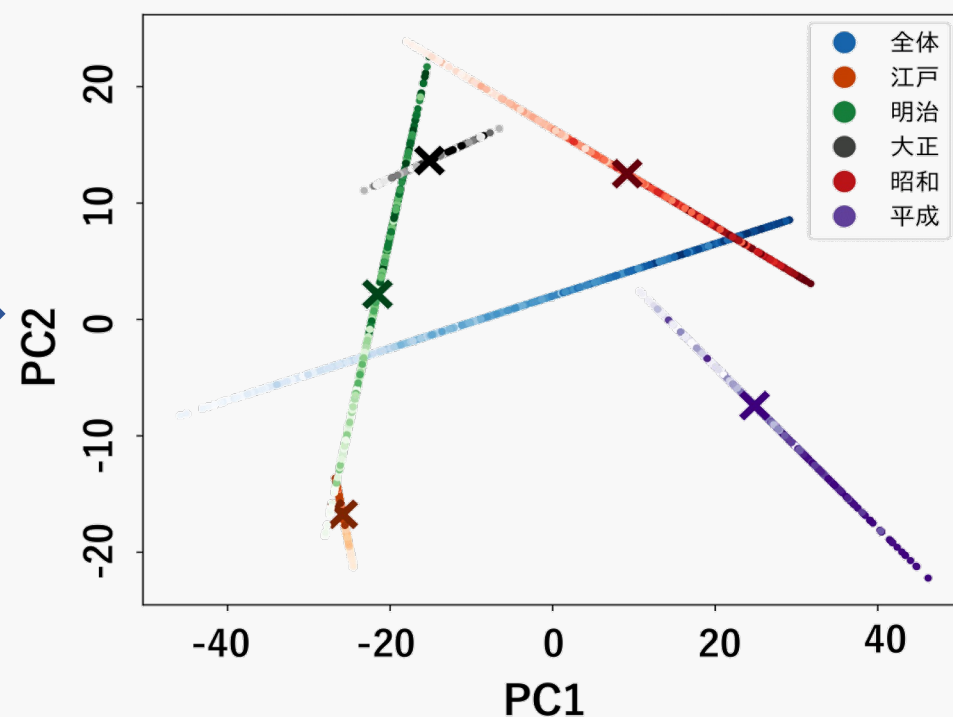


各時代の方が向がわからない

抽出



時代表現 $\hat{X}$ ($r = 1$)

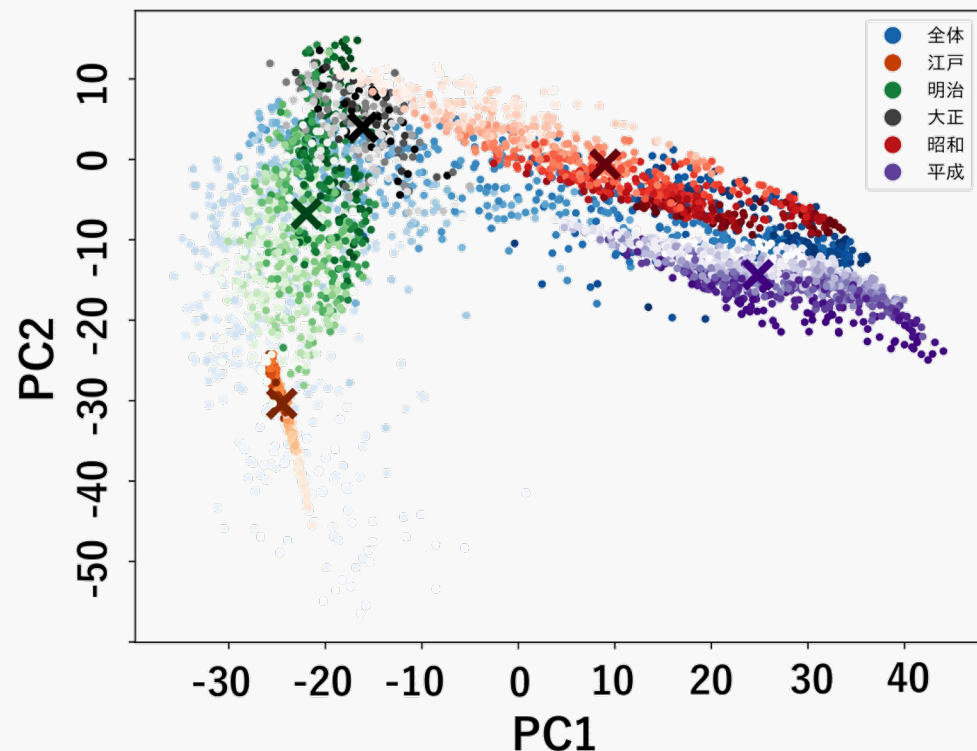


各時代の方が向がはっきりする

時系列構造の全体像の図示（ r を増やすと・・・）

- PLSの圧縮次元数 r を増やして詳細な時代表現 $\hat{\mathbf{X}}$ を図示

$r = 2$



$r = 3$

