文学作品のテキストデータを題材としたデータサイエンス演習

お茶の水女子大学 文理融合AI・データサイエンスセンター　土山 玄

発表構成

- **文理融合データサイエンスⅠおよびⅡの概要**
  - 受講生の傾向
  - 授業内容

- **文学作品の計量分析**
  - 主要な研究テーマ
  - データの取得と加工
  - 授業で採り上げた例

- **模擬授業**
  - 主成分分析
  - 文学作品を対象とした分析事例

文理融合データサイエンスⅠおよびⅡ

- **お茶の水女子大学について**
  - 文教育学部、理学部、生活科学部の3学部
  - 1学年500人弱

- **文理融合データサイエンスⅠ**
  - 2019年度：61人
    - 文教育学部：8人
    - 生活科学部：25人
    - 理　学　部：28人
  - 2020年度：71人
    - 文教育学部：31人
    - 生活科学部：22人
    - 理　学　部：18人

- **文理融合データサイエンスⅡ**
  - 2020年度：21人
    - 文教育学部：3人
    - 生活科学部：5人
    - 理　学　部：13人

文理融合データサイエンスⅠおよびⅡの概要

- **授業の流れ**
  - 前週の課題の復習（15分）
  - 講義（50分）
  - 例題（5分）
  - 演習（20分）

    ➡演習ではRを使用

- **成績評価**
  - 中間レポート
  - 期末レポート

  - 人文系データを対象
  - 受講生自ら課題を設定し、分析を行い、考察をレポートにまとめる

文理融合データサイエンスⅠおよびⅡの概要

- **授業内容**
  - 文理融合データサイエンスⅠ
    - 推定・検定
    - 回帰分析・判別分析・主成分分析・クラスター分析

    ★ データサイエンスの倫理

  - 文理融合データサイエンスⅡ
    - 決定木・ランダムフォレスト
    - ネットワーク分析
    - アソシエーション分析
    - サポートベクターマシン
    - ニューラルネットワーク

    ★ データサイエンスの倫理

文学作品の計量分析

- **主要な研究テーマ**
  - 著者の識別
  - 執筆順序の推定

- **データの取得と加工**
  - データの取得
    1. 青空文庫（https://www.aozora.gr.jp/）
    2. 日本古典籍データセット（http://codh.rois.ac.jp/pmjt/）
  - データの加工（形態素解析）
    - web茶まめ（https://chamame.ninjal.ac.jp/）

- **授業で採り上げた例**
  - 夏目漱石の小説における文末表現の変化

模擬授業：主成分分析

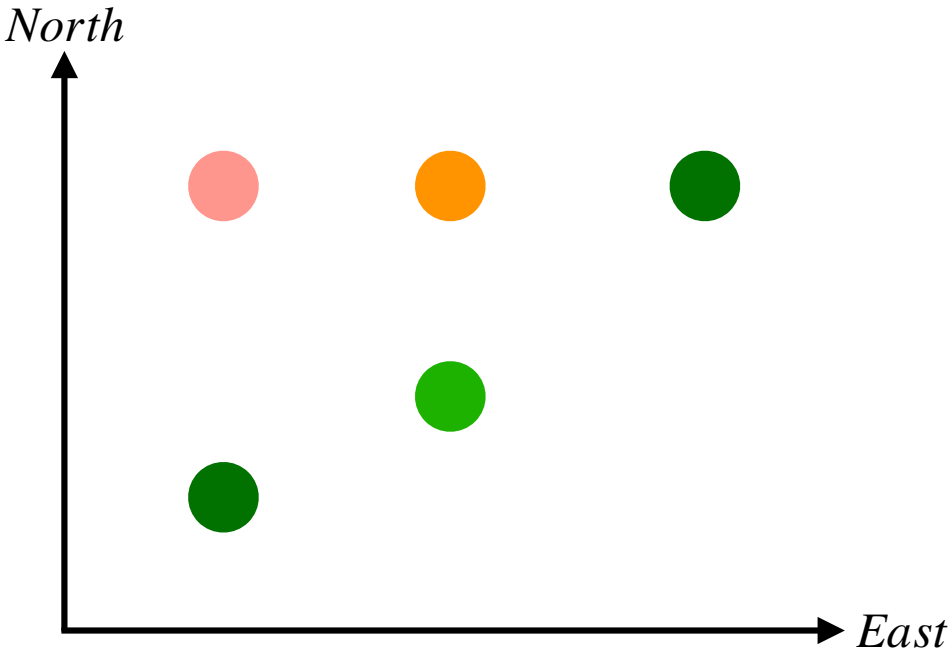
- **データの可視化**
  - 1変数：ヒストグラムや箱ひげ図
  - 2変数：散布図
  - 3変数：3次元散布図
  - 4変数以上：**可視化できない**
    - ➡データの構造を読み取ることができない

- **主成分分析の目的**
  - 多次元データを**少ない変数に縮約**する
    - ➡これを**次元縮約**と言う
    - ➡元のデータから**主成分**と呼ばれる合成変数を作成する

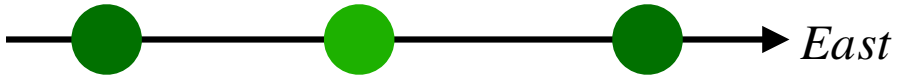
★2つの変数に縮約できれば、散布図としてデータの構造を**可視化できる**

模擬授業：主成分分析

- **主成分分析の考え方**
  - イメージ：公園の上から見下ろしたときの木の配置



① 南から北を見ると



➡3本の木しか見えない

➡縦軸の情報が損失

② 西から東を見ると

➡3本の木しか見えない

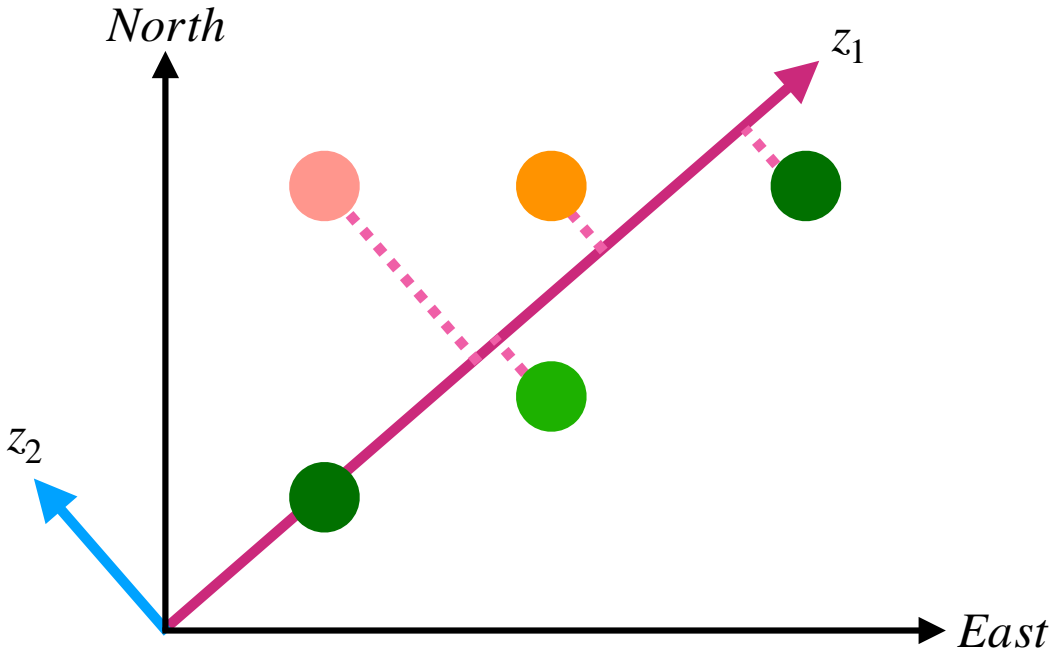
➡横軸の情報が損失

- ①と②のどちらから見ても  は見えない

➡すべての木を見える方角を探す

模擬授業：主成分分析

- **主成分分析の考え方**
  - イメージ：公園の上から見下ろしたときの木の配置



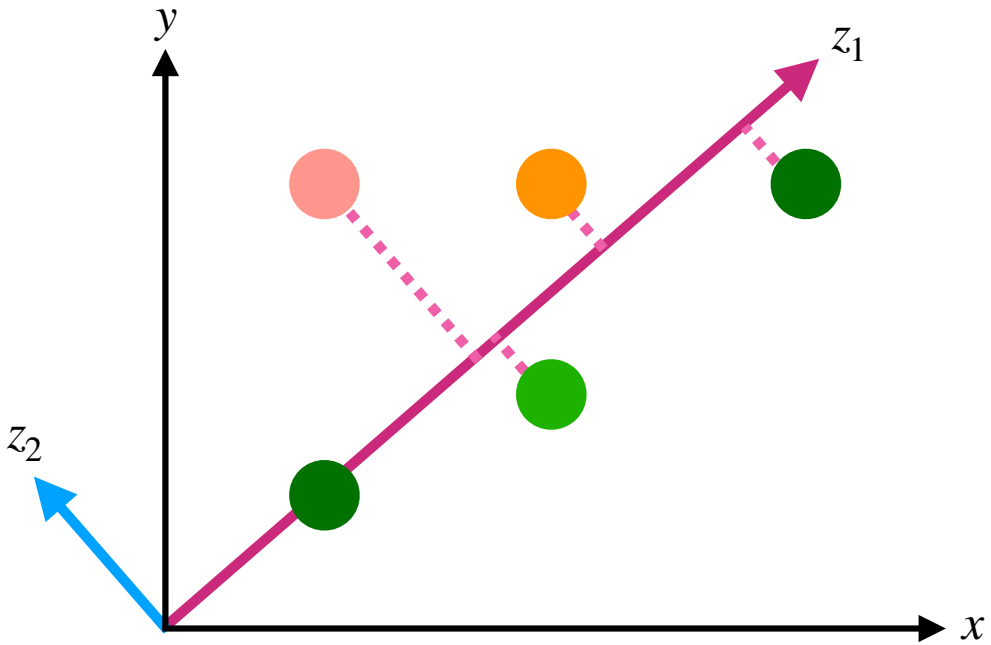
- 個体がもっともよく見えるような軸を探す

- これを統計的に表現すると
  - ➡射影したデータの **分散が最大**になる軸を探す
  - ➡分散が大きいほど情報が多い

- ——→ を第1主成分と言う
- ——→ を第2主成分と言う

模擬授業：主成分分析

- **主成分分析の考え方**
  - イメージ：公園の上から見下ろしたときの木の配置



- 数学的に考えると

  z_1の分散が最大となるa_1とa_2を求める

  $$z_1 = a_1 x + b_1 y$$

  ただし $a_1^2 + b_1^2 = 1$

  ➡固有ベクトルを求める
  - 固有値は主成分の分散となる

- z_2 は z_1に直交する軸の中で分散が最大となる軸

模擬授業：主成分分析

- **主成分得点と主成分負荷量**

$$z_i = a_i x + b_i y$$

主成分得点　　固有ベクトル

  - 主成分得点：合成変数によって求められた新たな値
  - 主成分負荷量：元のデータの変数の主成分に寄与している度合

    ➡固有ベクトルから求められる

- **寄与率**
  - 主成分が持っている元データの情報の割合

$$\text{寄与率} = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}$$

λ_i は各主成分の分散（固有値）

模擬授業：主成分分析

- **主成分分析の手順**
  1. データの取得
  2. データの**分散共分散行列** or **相関係数行列**を求める
  3. 固有値および固有ベクトルを求める
  4. 各主成分の主成分得点を求める
  5. 各主成分の寄与率と**累積寄与率**を求める
  6. 考察

- **分散共分散行列と相関係数行列の相違点**
  - 相関係数行列を用いる場合
    ➡データを標準化した主成分分析

- **累積寄与率**
  - 主成分の寄与率の和

模擬授業：主成分分析

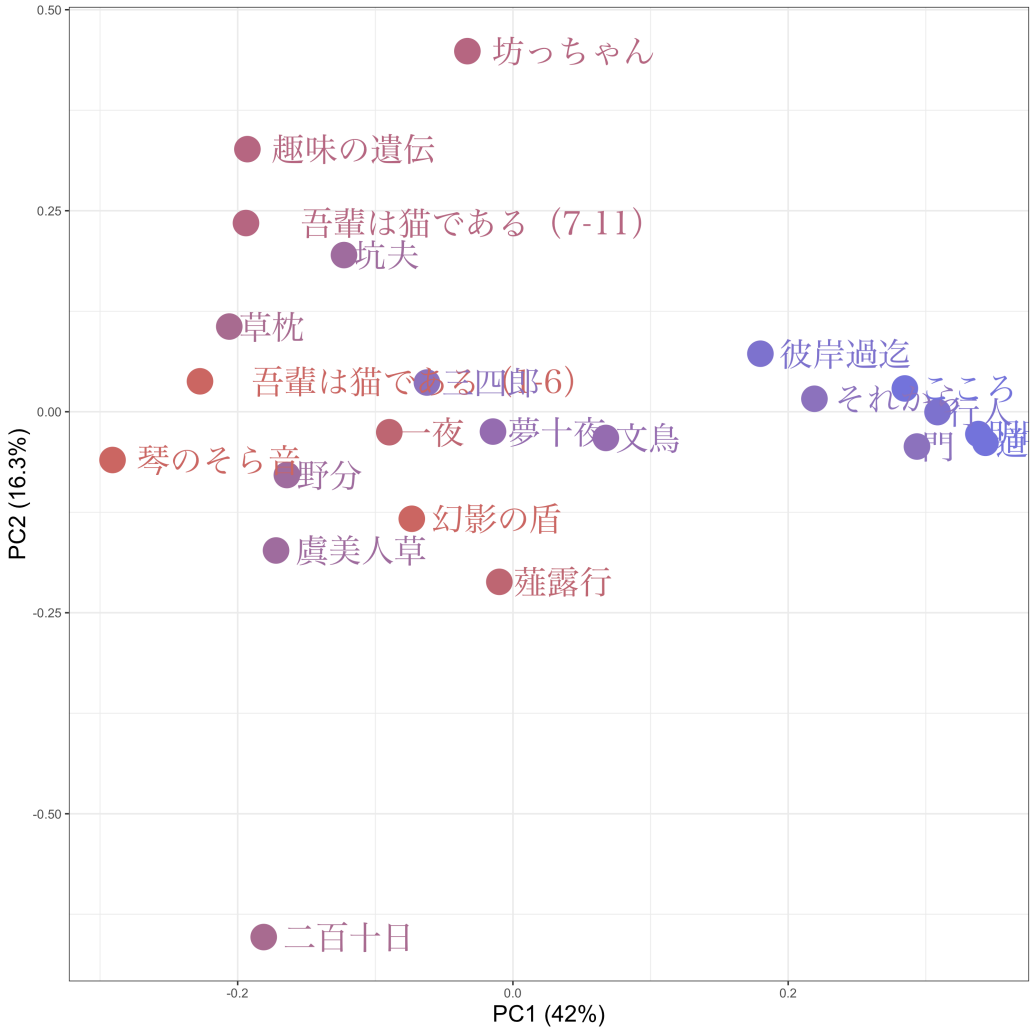
- **文体の変化を計量的に検討する**
  - 作家の**文体的特徴**の出現傾向は変化するのだろうか？
    ➡文体的特徴：文章にあらわれる書き手の習慣的な表現形式

  - 分析対象
    - 夏目漱石の小説22作品
      地の文（会話文を削除）
    - 青空文庫から取得

タイトル	発表時期
吾輩は猫である（一〜六）	1905年1月
幻影の盾	1905年4月
琴のそら音	1905年5月
一夜	1905年9月
薤露行	1905年11月
吾輩は猫である（七〜十一）	1906年1月
趣味の遺伝	1906年1月
坊っちゃん	1906年4月
草枕	1906年9月
二百十日	1906年10月
野分	1907年1月
虞美人草	1907年6月23日〜1907年10月29日
坑夫	1908年1月1日〜1908年4月6日
文鳥	1908年6月
夢十夜	1908年7月
三四郎	1908年9月1日〜1908年12月29日
それから	1909年5月31日〜1909年8月14日
門	1910年3月1日〜1910年6月12日
彼岸過迄	1912年1月1日〜1912年4月29日
行人	1912年12月6日〜1913年11月15日
こころ	1914年4月20日〜1914年8月11日
道草	1915年6月3日〜1915年9月14日
明暗	1916年5月26日〜1916年12月14日

模擬授業：主成分分析

- **文末に出現する単語上位16語の分析結果**
  - 第1主成分と第2主成分の主成分得点の散布図

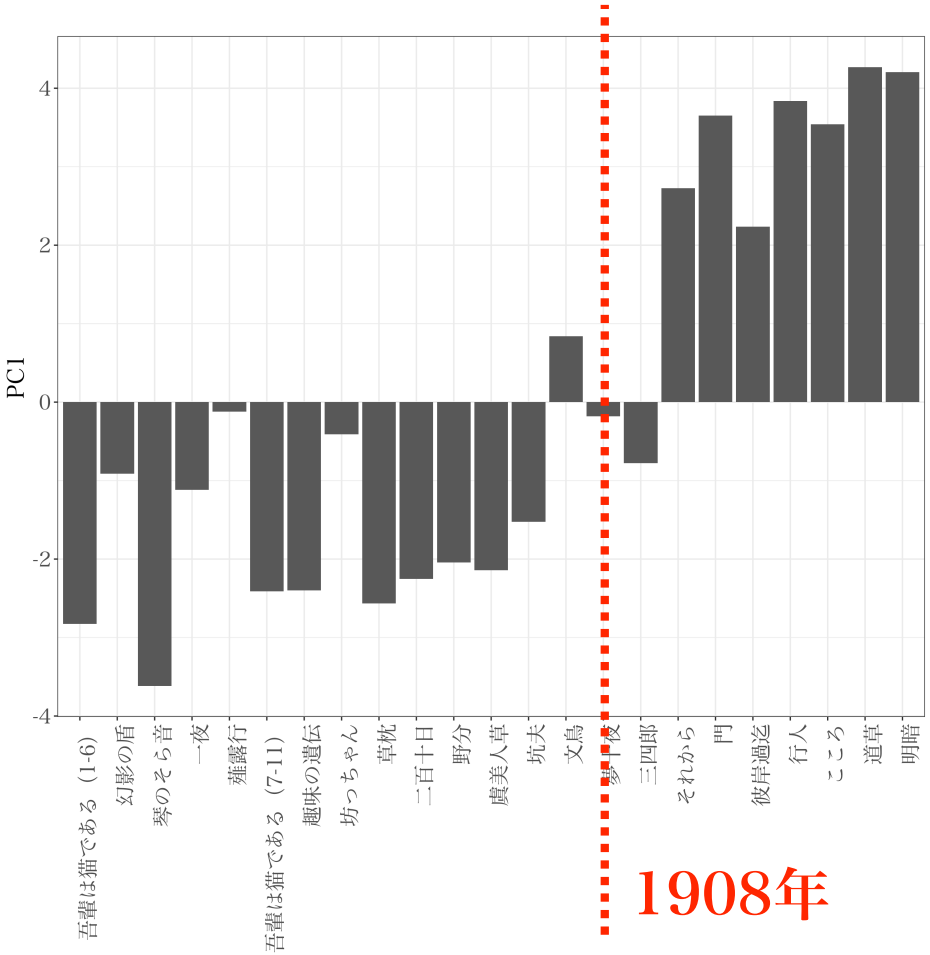


★ 色のグラデーション
  - 1905年発表の作品は赤
  - 1916年発表の作品は青

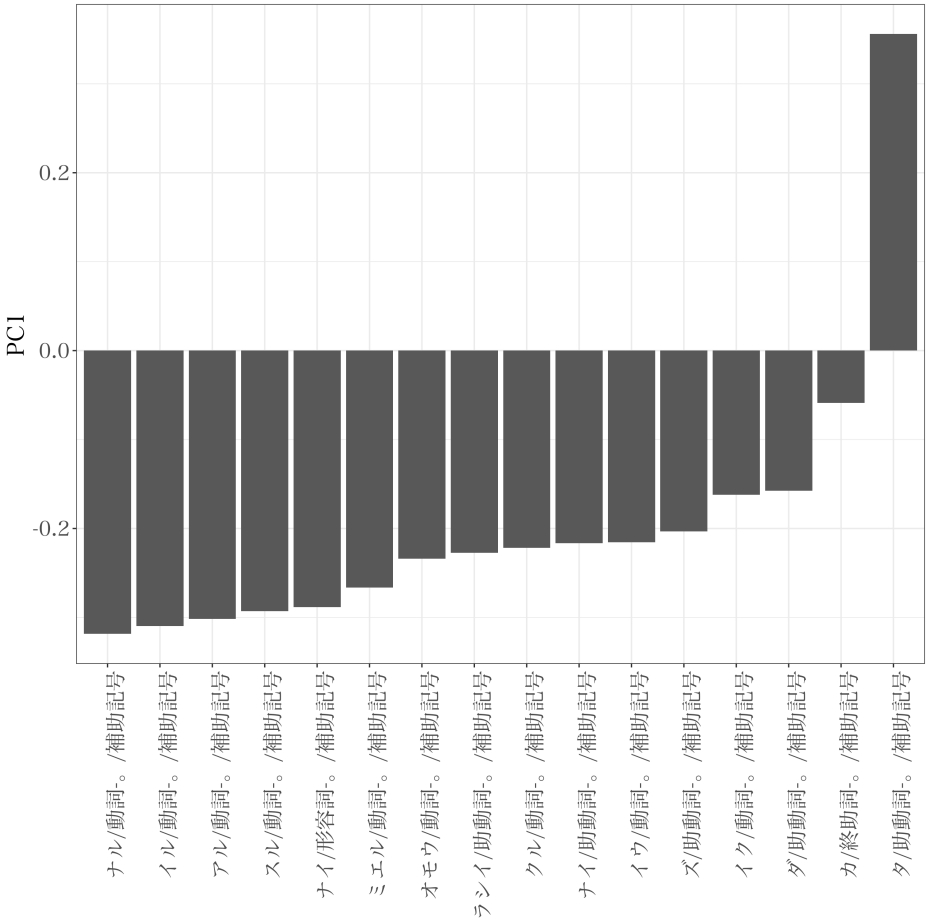
- 横軸が出版年を表していると考えられる

文学模擬授業：主成分分析

- **文末に出現する単語上位16語の分析結果**
  - 第1主成分の解釈



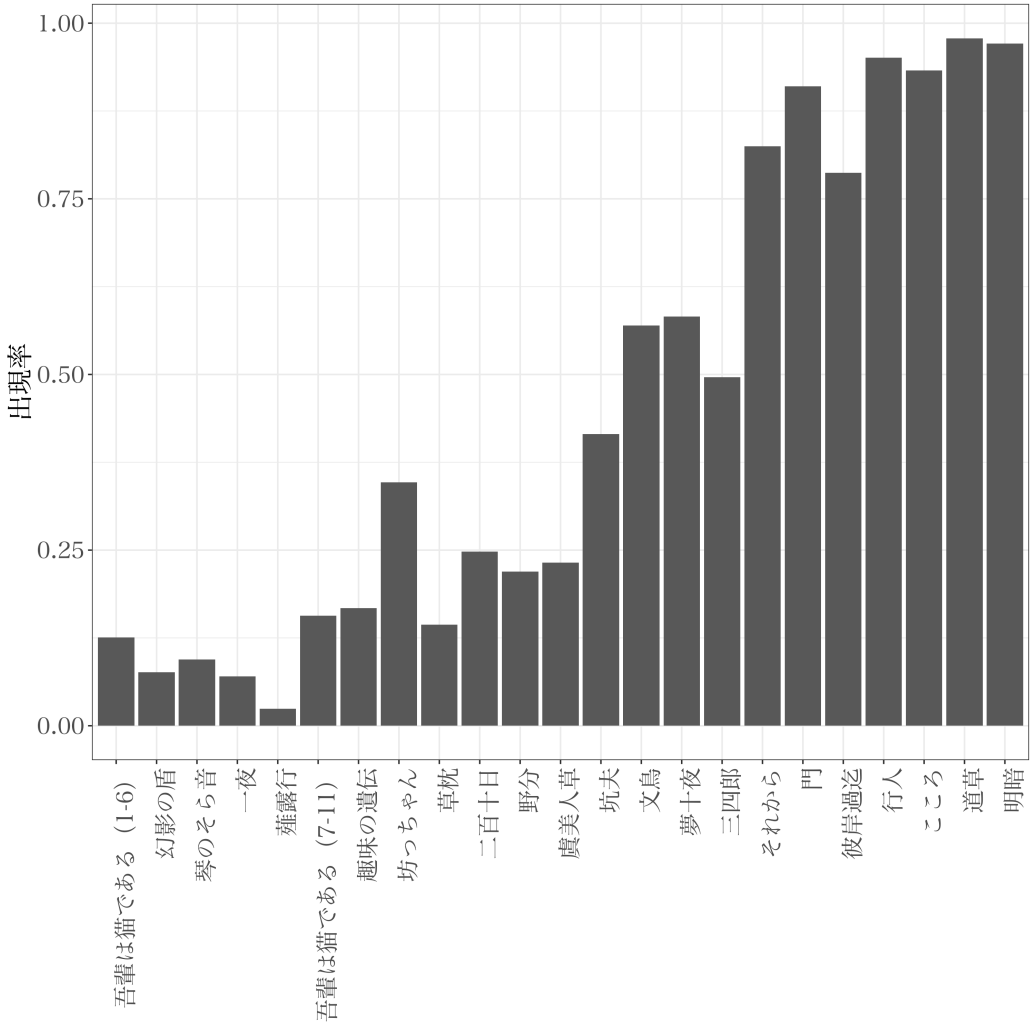
主成分得点（PC1）



主成分負荷量（PC1）

模擬授業：主成分分析

・文末における助動詞「た」の出現率



- 文末における助動詞の出現率が増加している
  ➡特に助動詞の「た」
  - 1905年の出現率は10%程度
  - 1916年の出現率は90%以上
- 1908年頃から文体に変化が認められると考えられる

受講生の反応

- **文学作品を分析することについて**
  - 文学作品の計量分析を知るために受講した学生も少数ながらいる
  - 美術などの計量的な研究事例も知りたい

- **分析手法について**
  - 多変量解析は理解できるが、推定と検定が難しい
  - 初めて聞く単語が多くて混乱する

- **Rの演習について**
  - 演習で実際に分析することで理解が深まったと感じる
  - 関数が多くて覚えきれない
  - 講義資料と同じように分析できない