数理統計 補助資料
～主成分分析～

2023年度2学期：　月曜1限, 水曜3限
担当教員：　石垣　司

多変量データと多変量解析

多変量データの表記法

- データ i：$\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iP}]^T$ $(i = 1, \ldots, N)$
  N：データ数，P：変数の数（データの次元）
- 全データ：$X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$

ID	変数1	変数2	…	変数 P	
1	x_{11}	x_{12}	…	x_{1P}	$\boldsymbol{x}_1^T$
2	x_{21}	x_{21}	…	x_{2P}	$\boldsymbol{x}_2^T$
⋮	⋮	⋮	⋱	⋮	⋮
i	x_{i1}	x_{i2}	…	x_{iP}	$\boldsymbol{x}_i^T$
⋮	⋮	⋮	⋱	⋮	⋮
N	x_{N1}	x_{N2}	…	x_{NP}	$\boldsymbol{x}_P^T$

可視化

予測・判別

仮説検証

発見

主成分分析

多変量データの傾向を説明する合成指標を作成し，その傾向を可視化や要約するための手法

- 次元削減法（次元圧縮，次元縮約）の一種
- 多変量データの傾向を代表する低次元の主成分を見つける

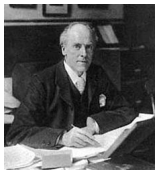
多変量データ

ID	変数1	変数2	…	変数 P
1	x_{11}	x_{12}	…	x_{1P}
2	x_{21}	x_{21}	…	x_{2P}
⋮	⋮	⋮	⋱	⋮
i	x_{i1}	x_{i2}	…	x_{iP}
⋮	⋮	⋮	⋱	⋮
N	x_{N1}	x_{N2}	…	x_{NP}

合成指標
（主成分）

第1主成分
⋮
第M主成分

$M < P$



Karl Pearson 1857-1936

多変量データをよく説明できる
少数の軸を作り出す

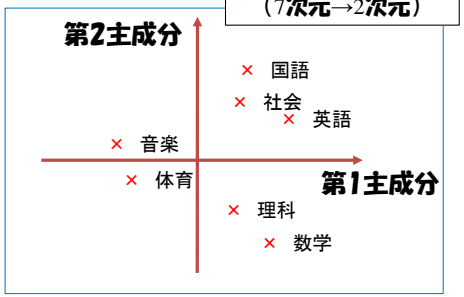
主成分分析による多変量データの可視化

多変量データの傾向を直感的に把握したい

- 例：300人分の中学校のテスト（仮想データ）

生徒 No.	国語	数学	理科	社会	英語	音楽	体育
1	83	60	55	81	90	50	93
2	70	80	78	80	55	44	59
3	50	90	95	70	80	80	49
4	60	44	44	99	78	73	30
5	57	80	80	50	67	64	59
6	55	65	70	65	67	30	70
7	80	73	66	46	55	58	88
8	98	40	50	88	99	93	54
9	55	77	88	40	89	88	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

次元削減された空間
（7次元→2次元）



可視化の場合は2次元に次元削減

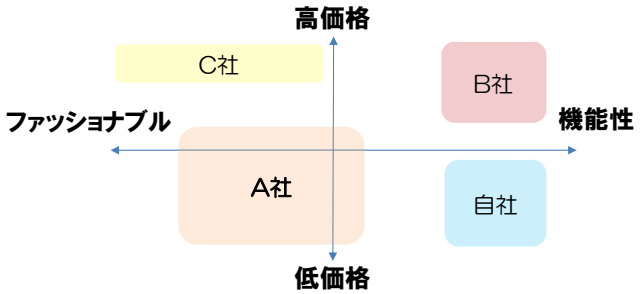
主成分の内容を結果から解釈

- 第1主成分：受験用学力
- 第2主成分：文系・理系能力

マーケティングにおける応用例

商品・サービスの差別化 ⇒ 優位性獲得のポジション
- 顧客ニーズの理解と価値の提供
- そのポジションを実現するマーケティング・ミックスの策定

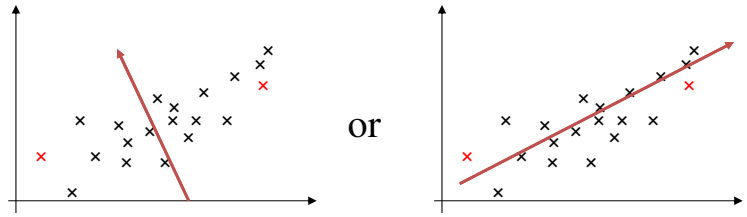
知覚マップ（←本日は知覚マップをデータから作成）
- 消費者の目線での各ブランドの市場内での位置づけを可視化
- 新製品のポジショニングや既存製品の再ポジショニングに利用



次元削減

高次元データを低次元データへ変換する方法
- 本日の講義内容は可視化を目的とした2次元への次元削減を主に説明するが，3次元以上の空間への次元削減も可能

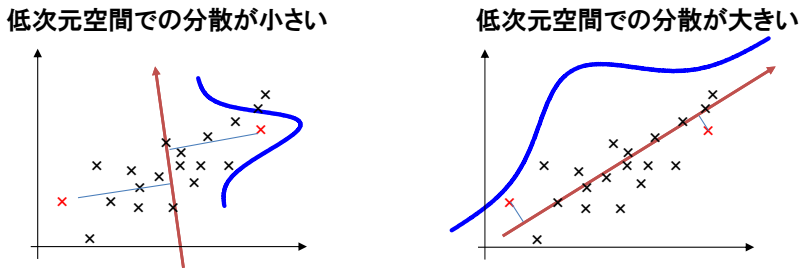
主成分分析は高次元のデータを低次元空間へ射影する次元削減の手法
- 下図の2次元から1次元への次元削減はどちらが良い？
  ⇒ 情報の損失を最小化する次元削減が望ましい



主成分分析の指標

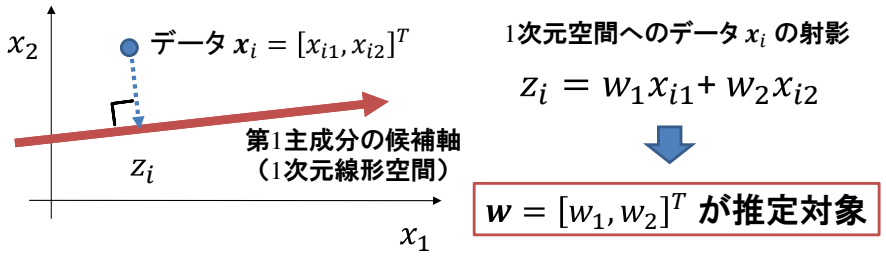
次元削減による情報損失は必至
- その上で，次元削減後もデータ間の違いが分かりやすい情報損失が少ない次元削減が望ましい

主成分分析での情報損失の測り方
⇒ 次元削減後の低次元空間でのデータの分散



主成分の定義

第1主成分：次元削減後の分散を最大化する射影軸
- 第 i 主成分：第 $1, \dots, i-1$ 主成分と直交する空間で分散を最大化する射影軸
- z_i： データ x_i を第1主成分の候補軸へ射影した点



第1主成分の求め方
- $\{z_1, \dots, z_N\}$ の分散を最大化する w_1, w_2 を推定

第1主成分の求め方 #1

$\{z_1, \ldots, z_N\}$ の標本分散を最大化する w を決める

Notation

- 第1主成分の重み: $\boldsymbol{w} = [w_1, \ldots, w_P]^T$
- データ i の第1主成分: $z_i = \boldsymbol{w}^T \boldsymbol{x}_i$
- 変数 p の標本平均: $\bar{x}_p = \frac{1}{N}\sum_{i=1}^{N} x_{ip}$
- データの標本分散共分散行列: $S_x = \begin{bmatrix} s_{11} & \cdots & s_{1P} \\ \vdots & \ddots & \vdots \\ s_{P1} & \cdots & s_{PP} \end{bmatrix}$
- i 番目と j 番目の変数の標本共分散: s_{ij}

第1主成分の求め方 #2

$\{z_1, \ldots, z_N\}$ の標本分散の最大化

- $\{z_1, \ldots, z_N\}$ の標本平均 $\bar{z}$ と標本分散 s_z check!

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i = \sum_{p=1}^{P} w_p \bar{x}_p$$

$$s_z = \frac{1}{N}\sum_{i=1}^{N} (z_i - \bar{z})^2 = \boldsymbol{w}^T S_x \boldsymbol{w}$$

制約条件: $\|\boldsymbol{w}\|^2 = \boldsymbol{w}^T \boldsymbol{w} = 1$

- w の要素を大きくすると s_z も大きくなってしまうために制約条件を導入。知りたいのは射影軸の方向のみ

第1主成分の求め方 #3

<u>制約条件付き2次形式の最大化問題</u>

- 目的関数: $\boldsymbol{w}^T S_x \boldsymbol{w}$,
- 制約条件: $\boldsymbol{w}^T \boldsymbol{w} = 1$
- この最適化は固有値分解: $S_x \boldsymbol{w} = \lambda \boldsymbol{w}$ に帰結
  - 固有値: $\lambda_1 > \lambda_2 > \cdots > \lambda_P \geq 0$
  - 対応する固有ベクトル: $\widehat{\boldsymbol{w}}_1, \widehat{\boldsymbol{w}}_2, \ldots, \widehat{\boldsymbol{w}}_P$
  - ⇒ 最大固有値に対応する固有ベクトルが分散 s_z を最大化
- 最適解: $\boldsymbol{w}^* = \widehat{\boldsymbol{w}}_1 = [\widehat{w}_{11}, \ldots, \widehat{w}_{1P}]^T$

前回の講義内容(分散共分散行列の固有値分解)を思い出して、この結果をイメージしてみよう

第2以降の主成分以降の求め方

求め方: $\begin{cases} \text{目的関数: } \boldsymbol{w}^T S_x \boldsymbol{w} \\ \text{制約条件: } \boldsymbol{w}^T \boldsymbol{w} = 1 \text{ and } \boldsymbol{w}^T \widehat{\boldsymbol{w}}_1 = 0 \end{cases}$

第2主成分の定義は第1主成分と直交する射影軸

- 実は, 最適解は前頁で求めた $\boldsymbol{w}_2^* = \widehat{\boldsymbol{w}}_2 = [\widehat{w}_{21}, \ldots, \widehat{w}_{2P}]^T$

累積寄与率: $\frac{\lambda_1 + \cdots + \lambda_M}{\lambda_1 + \lambda_2 + \cdots + \lambda_P}$ $(M < P)$

- 多変量データがもつ情報(分散)を、第 M 主成分までで どのくらい保持できているかの指標

主成分分析を用いた知覚マップ #1

例： Consumer Brand Rating Data (simulated data)

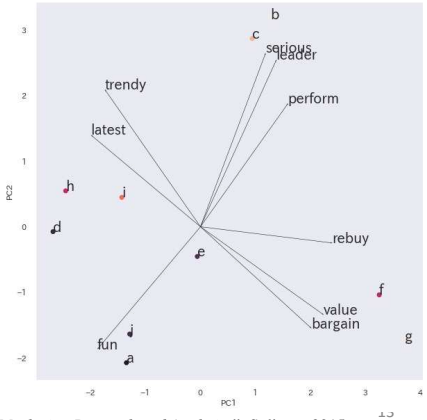
- 10**種のコーヒーブランドの模擬調査データ**
- 100**人が各ブランドについて**9**の観点から評価（**1**点～**10**点）**

ブランド： a ~ j

観点： perform, leader, latest, fun, serious, bargain, value, trendy, rebuy

ID	perform	leader	⋯	rebuy
a	x_{11}	x_{12}	⋯	x_{1P}
b	x_{21}	x_{21}	⋯	x_{2P}
⋮	⋮	⋮	⋱	⋮
j	x_{i1}	x_{i2}	⋯	x_{iP}

PCA



主成分分析を用いた知覚マップ #2

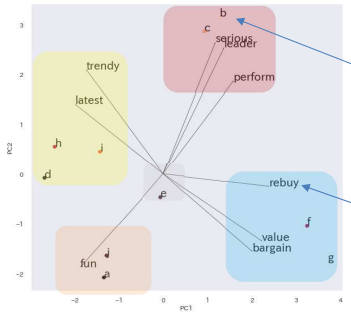
例：Consumer Brand Rating Data (simulated data)

- **ブランド** i **の第**1**主成分(主成分得点)：** $z_i^{(1)} = \hat{\boldsymbol{w}}_1^T \boldsymbol{x}_i$
- **ブランド** i **の第**2**主成分(主成分得点)：** $z_i^{(2)} = \hat{\boldsymbol{w}}_2^T \boldsymbol{x}_i$

ブランド i **の圧縮次元上の座標** $\left(z_i^{(1)}, z_i^{(2)}\right)$

- **変数** j **と第**1**&第**2**主成分の相関係数(因子負荷量)：** $\sqrt{\lambda_1}\hat{w}_{1j}, \sqrt{\lambda_2}\hat{w}_{2j}$

変数 j **の圧縮次元上の因子負荷量** $\left(\sqrt{\lambda_1}\hat{w}_{1j}, \sqrt{\lambda_2}\hat{w}_{2j}\right)$



ブランド b **の圧縮次元上の座標** $\left(z_b^{(1)}, z_b^{(2)}\right)$

変数 "rebuy" **の圧縮次元上の因子負荷量** $\left(\sqrt{\lambda_1}\hat{w}_{1,\text{rebuy}}, \sqrt{\lambda_2}\hat{w}_{2,\text{rebuy}}\right)$

演習問題

次の主成分分析の結果から読み取れる各ブランドの特徴についてまとめなさい。また, ブランド a **から** j **の内一つを自由に選択し, そのブランドが取り得る戦略について議論しなさい。ただし, 各ブランドを販売する企業の社会的意義, 経営戦略, 経営状況, 強みや弱みなどの状況は自由に設定してよい**