数理統計 補助資料
～主成分分析～

2024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

- i 番目のデータ：$\boldsymbol{x}_i = [x_{i1}, \ldots, x_{iP}]^T$ $(i = 1, \ldots, N)$
  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，その傾向を要約・可視化する手法

- 次元削減（次元圧縮，次元縮約）法の一種
- 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}

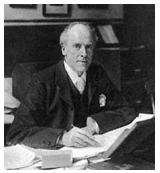
合成指標（主成分）

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

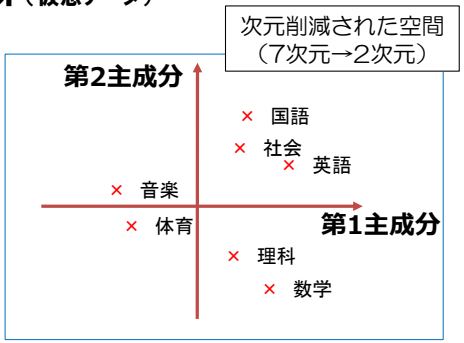
主成分分析による多変量データの可視化

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

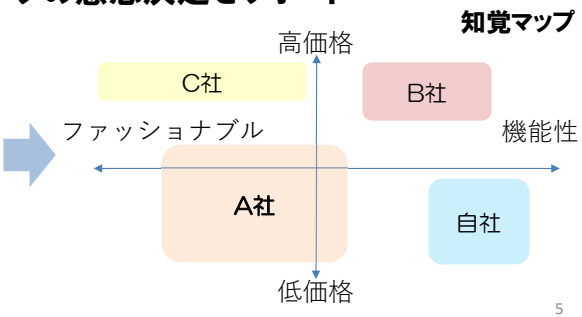
マーケティングでの応用例

知覚マップの作成でポジショニング戦略をサポート

- **ポジショニング戦略：自社の商品・サービスを他社のそれと差別化することで優位性の獲得を目指すマーケティング戦略**
- **消費者アンケートから作成した知覚マップにより，客観的に自社の立ち位置を理解**
- **経営やマーケティングの意思決定をサポート**

消費者1000人の平均点

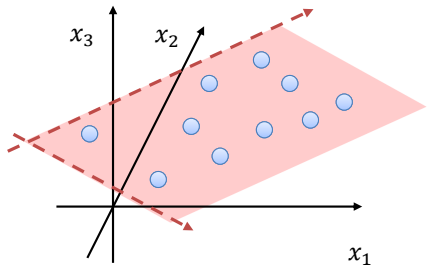
消費者ID	おしゃれ	使いやすい	またに使いたい	値段がお得	質がよい	…
A社	80.2	82.5	85.1	85.9	79.8	…
B社	68.5	86.9	88.3	70.2	90.1	…
C社	93.3	71.7	90.4	53.5	97.4	…
自社	58.4	90.1	92.1	93.2	68.3	…



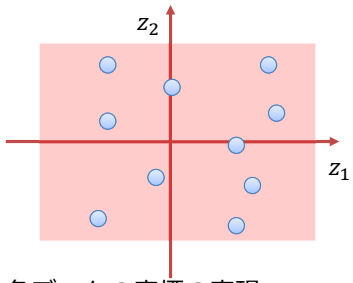
次元削減

高次元データを低次元データへ変換する方法の総称

- **高次元データの持つ情報をできるだけ保持したまま，低い次元の空間でその情報を表現したい**
- **例：3次元空間 (x_1, x_2, x_3) 内のデータがすべて2次元平面上を通る場合，新しい座標系 (z_1, z_2) で元の3次元空間内のデータの持つ情報（分布）と全く同じ情報を表現できる**



各データの座標の表現 (x_1, x_2, x_3)



各データの座標の表現 (z_1, z_2)

主成分分析による次元削減

特別な場合を除き，次元削減で情報の損失は不可避

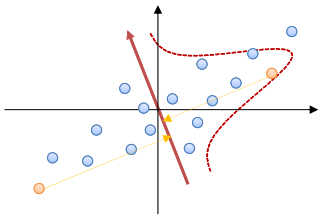
- **その上で，情報の損失が少ない次元削減が望ましい**

主成分分析の次元削減の方針

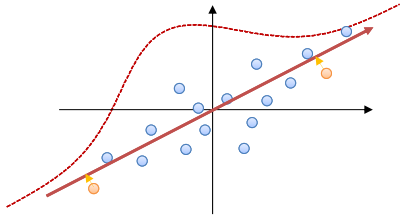
- **正射影したデータの分散が最大となる低次元空間を見つける**

例：2次元データを1次元へ削減

射影後のデータの分散が小さい低次元空間



射影後のデータの分散が大きい低次元空間

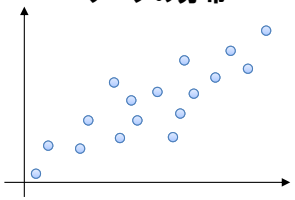


多変量データへの前処理

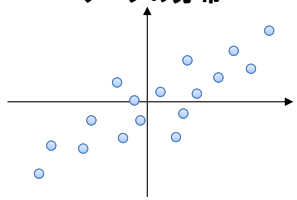
多変量データ X は各変数 p に対して中心化 or 標準化

- **中心化処理は必須**
- **標準化処理（各変数の分散が1となる変換処理）は必要に応じて実施**

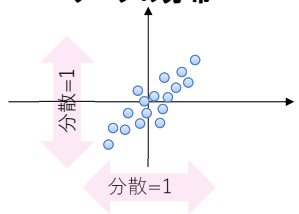
前処理前のデータの分布



中心化処理後のデータの分布



標準化処理後のデータの分布

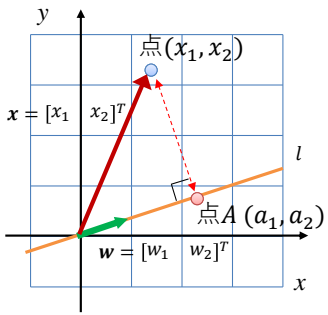


これ以降，データ X は標準化されているとして話を進める

- **R や Python のパッケージの多くでは自動的に中心化される**
- **また，オプションで標準化も容易に可能**

数学的準備

定理：あるベクトル x の指す座標から原点を通る直線 l へ降ろした垂線と直線 l の交点 A を考える(つまり，ベクトル x から直線 l への正射影)。また，ベクトル w は直線 l と同じ向きの単位ベクトル，z をスカラー定数とする。このとき，点 A とベクトル zw の指す点が一致するのは，$z = w^T x$ のときである



左図での説明

垂直に交わる直線の傾き性質から $\frac{a_2-x_2}{a_1-x_1}\frac{w_2}{w_1} = -1$。
整理すると $w_1(a_1 - x_1) + w_2(a_2 - x_2) = 0$。

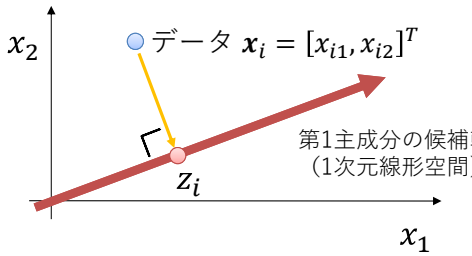
$a_1 = zw_1$ かつ $a_2 = zw_2$ となるとき、
$w_1(zw_1 - x_1) + w_2(zw_2 - x_2) = 0$
$(w_1^2 + w_2^2)z = x_1w_1 + x_2w_2.$

$\boldsymbol{w}$は単位ベクトルなので $w_1^2 + w_2^2 = 1$ より
$z = x_1w_1 + x_2w_2 = \boldsymbol{w}^T\boldsymbol{x}$

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

- **z_i：データ x_i を第1主成分の候補軸へ正射影した点**



1次元空間へのデータ x_i の
z軸への正射影

$$z_i = w_1x_{i1} + w_2x_{i2}$$

$\boldsymbol{w} = [w_1, w_2]^T$ が推定対象

第1主成分は $\{z_1, \dots, z_N\}$ の分散を最大化する w を求める

- **第2主成分：第1主成分と直交する空間で分散を最大化する正射影軸を求める**
- **第3主成分：第1主成分 & 第2主成分と直交する空間で分散を最大化する正射影軸を求める。第4主成分以降も同様**

第1主成分の求め方 #1

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

- **第1主成分を算出する重み係数：$w = [w_1, \dots, w_P]^T$**
- **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}

定理：分散共分散行列は半正定値行列である check!

2次形式の制約付き最適化問題とその最大固有値と固有ベクトルの性質を利用できる

第1主成分の求め方 #2

$\{z_1, \dots, z_N\}$ の標本分散 s_z の最大化

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

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

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

X が中心化 or 標準化されているデータの場合は $\bar{x}_p = 0$。また，$\bar{z} = 0$

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

- **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

2次形式の制約条件付き最大化問題

目的関数： maximize $s_z = \boldsymbol{w}^T S_x \boldsymbol{w}$
制約条件： $\boldsymbol{w}^T \boldsymbol{w} = 1$

check!

- この最適化問題は固有値問題 $S_x \boldsymbol{w} = \lambda \boldsymbol{w}$ に帰着
- 固有値問題の解
  固有値：$\lambda_1 > \lambda_2 > \cdots > \lambda_P \geq 0$（半正定値行列の固有値のため非負）
  各固有値 p に対応する固有ベクトル：$\widehat{\boldsymbol{w}}_1, \widehat{\boldsymbol{w}}_2, \ldots, \widehat{\boldsymbol{w}}_P$
- $S_x \widehat{\boldsymbol{w}}_1 = \lambda_1 \widehat{\boldsymbol{w}}_1 \Rightarrow \widehat{\boldsymbol{w}}_1^T S_x \widehat{\boldsymbol{w}}_1 = \lambda_1$ より固有値 λ_1 は固有ベクトル $\widehat{\boldsymbol{w}}_1$ へ正射影されたデータの分散の値と同値

固有ベクトル $\widehat{\boldsymbol{w}}_1$ が分散 s_z を最大化する重み係数

第2主成分以降の主成分の求め方

第2主成分への射影の重み係数：$\widehat{\boldsymbol{w}}_2$

- 分散共分散行列は対称行列 ⇒ 固有ベクトルは直交
  固有ベクトル：$\widehat{\boldsymbol{w}}_1, \widehat{\boldsymbol{w}}_2, \ldots, \widehat{\boldsymbol{w}}_P$ はすべて直交している
- 第2主成分の定義は，第1主成分と直交する空間で分散を最大化する射影軸。λ_2 に対応する $\widehat{\boldsymbol{w}}_2$ が第1主成分の次に正射影後の分散が大きい
- 第3主成分以降も同様

前頁の固有値問題を解くことで，第1主成分から第P主成分までの正射影に必要な重み係数が全て計算されている

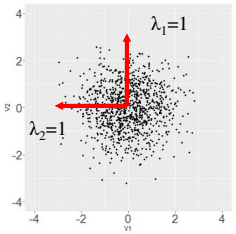
寄与率

第 j 主成分が持つ分散の割合： $\dfrac{\lambda_j}{\lambda_1+\lambda_2+\cdots+\lambda_P}$

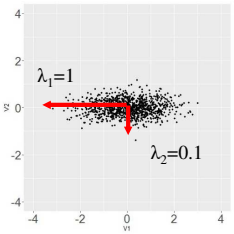
累積寄与率： $\dfrac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\lambda_2+\cdots+\lambda_P}$ $(j < P)$

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

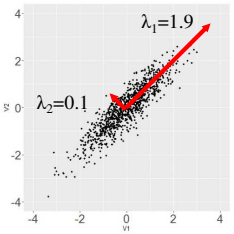
第1主成分の寄与率=0.5
第2主成分の寄与率=0.5



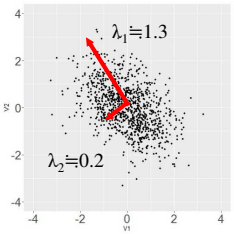
第1主成分の寄与率≒0.9
第2主成分の寄与率≒0.1



第1主成分の寄与率=0.95
第2主成分の寄与率=0.05



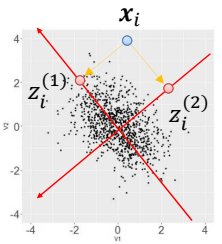
第1主成分の寄与率≒0.87
第2主成分の寄与率≒0.13



可視化のために：主成分得点と主成分負荷量

主成分得点

- 各データ $\boldsymbol{x}_i$ を各主成分へ正射影した値
  データ $\boldsymbol{x}_i$ の第1主成分の得点：$z_i^{(1)} = \widehat{\boldsymbol{w}}_1^T \boldsymbol{x}_i$
  データ $\boldsymbol{x}_i$ の第2主成分の得点：$z_i^{(2)} = \widehat{\boldsymbol{w}}_2^T \boldsymbol{x}_i, \cdots$



主成分負荷量

- 変数 p と第 j 主成分の主成分負荷量：
  変数ベクトル $\boldsymbol{v}_p = [x_{1p}, \ldots, x_{Np}]^T$ とデータ $i = 1, \cdots, N$ の第 j 主成分得点 $\boldsymbol{z}_j = [\widehat{\boldsymbol{w}}_j^T \boldsymbol{x}_{i=1}, \ldots, \widehat{\boldsymbol{w}}_j^T \boldsymbol{x}_{i=N}]^T$ との相関係数

$$r(\boldsymbol{v}_p, \boldsymbol{z}_j) = \frac{\mathrm{Cov}[\boldsymbol{v}_p, \boldsymbol{z}_j]}{\sqrt{V[\boldsymbol{v}_p] V[\boldsymbol{z}_j]}} = \frac{\lambda_j \widehat{w}_{jp}}{\sqrt{1 \times \lambda_j}} = \sqrt{\lambda_j} \widehat{w}_{jp}$$

標準化されている場合

$\mathrm{Cov}[\boldsymbol{v}_p, \boldsymbol{z}_j] = \lambda_j \widehat{w}_{jp}$ の証明は省略。各変数の定義に沿って導出可

主成分分析を用いた知覚マップ

例： Consumer Brand Rating Data (simulated data)

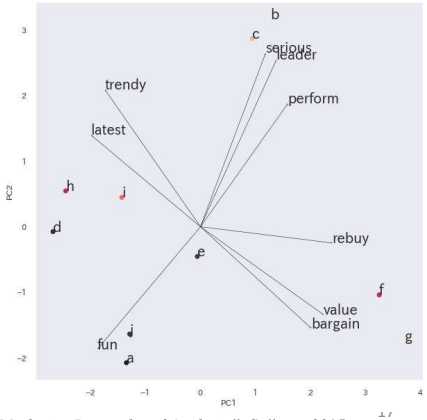
- 10**種のコーヒーブランドの模擬調査データ**
- 100**人が各ブランドについて**9**の観点から評価（**1**点～**10**点）**

  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



この結果はデータを標準化している

C. Chapman, E.M. Feit, "*R for Marketing Research and Analytics*", Splinger 2015

主成分分析による可視化法

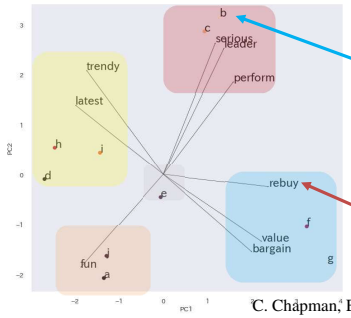
例： Consumer Brand Rating Data (simulated data)

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

→ **ブランド** i **の低次元空間上の座標** $\left(z_i^{(1)}, z_i^{(2)}\right)$

- **変数** p **と第**1**主成分の負荷量：** $\sqrt{\lambda_1}\hat{w}_{1p}$
- **変数** p **と第**2**主成分の負荷量：** $\sqrt{\lambda_2}\hat{w}_{2p}$

→ **変数** j **の低次元空間上の座標** $\left(\sqrt{\lambda_1}\hat{w}_{1p}, \sqrt{\lambda_2}\hat{w}_{2p}\right)$



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

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

C. Chapman, E.M. Feit, "*R for Marketing Research and Analytics*", Splinger 2015

演習問題

主成分分析の結果の解釈：

1. **主成分分析を用いて作成した前頁の知覚マップの第**1**主成分と第**2**主成分の内容をそれぞれ解釈しなさい**
2. **主成分分析を用いて作成した知覚マップ内のブランド** a **から** j **の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**