

主成分分析

基本的な考え方

村田 昇

講義概要

- 第1日：主成分分析の考え方
- 第2日：分析の評価と視覚化

主成分分析の考え方

主成分分析

- 多数の変量のもつ情報の分析・視覚化
 - 変量を効率的に縮約して少数の特徴量を構成する
 - 特徴量に關与する変量間の関係を明らかにする
- PCA (Principal Component Analysis)
 - 構成する特徴量：主成分 (principal component)

主成分分析の例

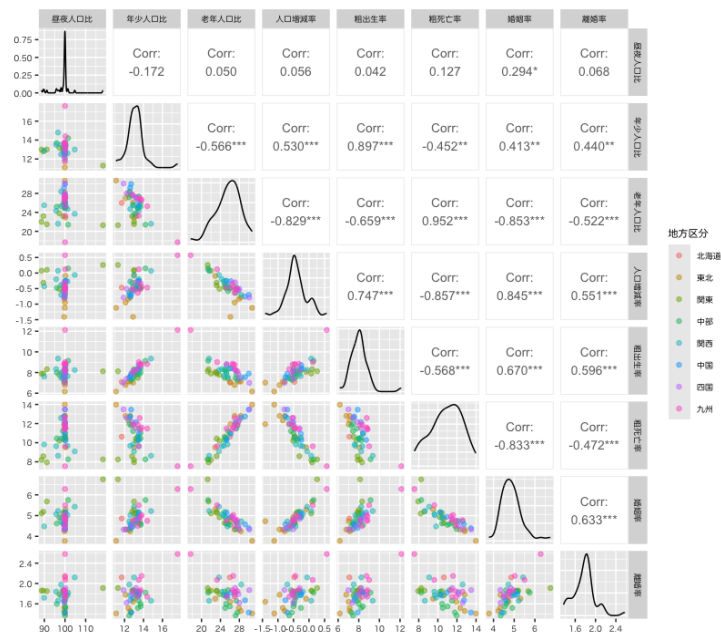


Figure 1: 人口関連データの散布図

都道府県名	地方区分	昼夜人口比	年少人口比	老年人口比	人口増減率	粗出生率	粗死亡率	婚姻率	離婚率
北海道	北海道	100.0	11.7	26.0	-0.47	7.09	10.63	4.86	2.12
青森県	東北	100.0	12.1	27.0	-0.95	6.79	12.81	4.33	1.78
岩手県	東北	99.7	12.4	27.9	-0.84	7.12	12.33	4.32	1.52
宮城県	東北	100.2	13.0	22.9	-0.09	8.05	9.51	5.30	1.70
秋田県	東北	99.9	11.1	30.7	-1.12	6.16	13.98	3.78	1.41
山形県	東北	99.8	12.6	28.3	-0.78	7.13	12.81	4.24	1.46
福島県	東北	99.6	12.9	26.1	-1.41	7.02	11.94	4.73	1.64
茨城県	関東	97.2	13.2	23.8	-0.51	7.78	10.20	4.92	1.79
栃木県	関東	99.1	13.2	23.2	-0.40	8.02	10.43	5.13	1.85
群馬県	関東	99.9	13.4	24.9	-0.45	7.49	10.63	4.64	1.77
埼玉県	関東	88.6	13.0	22.0	0.07	7.90	8.20	5.10	1.86
千葉県	関東	89.5	12.8	23.2	-0.31	7.89	8.59	5.19	1.86
東京都	関東	118.4	11.3	21.3	0.26	8.12	8.25	6.75	1.91
神奈川県	関東	91.2	13.0	21.5	0.10	8.32	7.94	5.68	1.85
新潟県	中部	100.0	12.5	27.2	-0.64	7.45	11.97	4.35	1.37
富山県	中部	99.8	12.7	27.6	-0.55	7.28	11.79	4.50	1.43
石川県	中部	100.2	13.4	25.0	-0.26	8.21	10.51	4.91	1.52
福井県	中部	100.1	13.7	26.0	-0.50	8.40	11.01	4.55	1.55
山梨県	中部	99.0	12.9	25.6	-0.58	7.44	11.21	4.60	1.87
長野県	中部	99.9	13.5	27.4	-0.47	7.81	11.48	4.67	1.66
岐阜県	中部	96.0	13.7	25.2	-0.48	8.00	10.45	4.62	1.60
静岡県	中部	99.9	13.4	24.9	-0.37	8.25	10.23	5.17	1.84
愛知県	中部	101.5	14.2	21.4	0.15	9.14	8.26	5.75	1.82
三重県	関西	98.1	13.5	25.3	-0.38	8.00	10.44	4.89	1.76
滋賀県	関西	96.6	14.8	21.6	0.07	9.35	8.64	5.22	1.66
京都府	関西	101.2	12.6	24.7	-0.27	7.66	9.68	5.02	1.77
大阪府	関西	104.7	13.0	23.7	-0.06	8.24	9.09	5.43	2.12
兵庫県	関西	95.7	13.5	24.3	-0.20	8.34	9.63	5.07	1.84
奈良県	関西	89.9	12.9	25.5	-0.43	7.60	9.82	4.48	1.72
和歌山県	関西	98.1	12.5	28.4	-0.70	7.51	12.59	4.72	1.98
鳥取県	中国	100.0	13.2	27.2	-0.51	8.20	12.15	4.74	1.83
島根県	中国	100.0	12.7	30.0	-0.70	7.90	13.46	4.40	1.43
岡山県	中国	99.9	13.5	26.2	-0.26	8.41	10.94	4.94	1.82
広島県	中国	100.3	13.5	25.3	-0.25	8.72	10.28	5.15	1.78
山口県	中国	99.5	12.6	29.2	-0.76	7.55	12.74	4.58	1.67
徳島県	四国	99.7	12.2	28.0	-0.51	7.40	12.60	4.34	1.62
香川県	四国	100.2	13.2	27.1	-0.30	8.25	11.50	4.84	1.91
愛媛県	四国	100.1	12.8	27.8	-0.56	7.87	12.17	4.51	1.79
高知県	四国	99.9	11.9	30.1	-0.79	7.00	13.49	4.33	1.87
福岡県	九州	100.1	13.5	23.3	0.12	9.01	9.63	5.50	2.07
佐賀県	九州	100.2	14.4	25.3	-0.47	8.83	11.48	4.75	1.74
長崎県	九州	99.8	13.4	27.0	-0.64	8.33	11.92	4.50	1.74
熊本県	九州	99.6	13.7	26.5	-0.33	8.85	11.38	4.96	1.87
大分県	九州	100.0	12.9	27.6	-0.50	8.14	11.86	4.77	1.85
宮崎県	九州	100.0	13.8	26.7	-0.44	8.75	11.59	5.03	2.15
鹿児島県	九州	99.9	13.6	27.0	-0.53	8.78	12.59	4.78	1.84
沖縄県	九州	100.0	17.6	17.7	0.57	12.12	7.54	6.28	2.58

分析の枠組み

- x_1, \dots, x_p : 変数
- z_1, \dots, z_d : 特徴量 ($d \leq p$)
- 変数と特徴量の関係 (線形結合)

$$z_k = a_{1k}x_1 + \dots + a_{pk}x_p \quad (k = 1, \dots, d)$$

- 特徴量は定数倍の任意性があるので以下を仮定

$$\|a_k\|^2 = \sum_{j=1}^p a_{jk}^2 = 1$$

主成分分析の用語

- 特徴量 z_k
 - 第 k 主成分得点 (principal component score)
 - 第 k 主成分

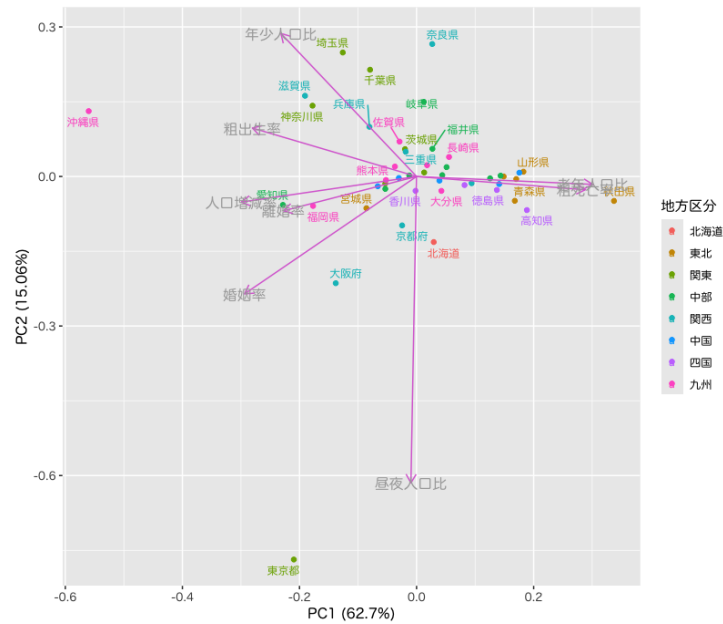


Figure 2: 主成分得点のバイプロット

- 係数ベクトル \mathbf{a}_k
 - 第 k 主成分負荷量 (principal component loading)
 - 第 k 主成分方向 (principal component direction)

分析の目的

- 目的

主成分得点 z_1, \dots, z_d が変数 x_1, \dots, x_p の情報を効率よく反映するように主成分負荷量 $\mathbf{a}_1, \dots, \mathbf{a}_d$ を観測データから決定する
- 分析の方針 (以下は同値)
 - データの情報を最も保持する変量の **線形結合を構成**
 - データの情報を最も反映する **座標軸を探索**
- 教師なし学習 の代表的手法の 1 つ
 - 特徴抽出: 情報処理に重要な特性を変数に凝集
 - 次元縮約: 入力をできるだけ少ない変数で表現

実習

第 1 主成分の計算

記号の準備

- 変数: x_1, \dots, x_p (p 次元)
- 観測データ: n 個の (x_1, \dots, x_p) の組

$$\{(x_{i1}, \dots, x_{ip})\}_{i=1}^n$$

- ベクトル表現

- $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$: i 番目の観測データ (p 次元空間内の 1 点)
- $\mathbf{a} = (a_1, \dots, a_p)^\top$: 長さ 1 の p 次元ベクトル

係数ベクトルによる射影

- データ \mathbf{x}_i の \mathbf{a} 方向成分の長さ

$$\mathbf{a}^\top \mathbf{x}_i \quad (\text{スカラー})$$

- 方向ベクトル \mathbf{a} をもつ直線上への点 \mathbf{x}_i の直交射影

$$(\mathbf{a}^\top \mathbf{x}_i) \mathbf{a} \quad (\text{スカラー} \times \text{ベクトル})$$

幾何学的描像

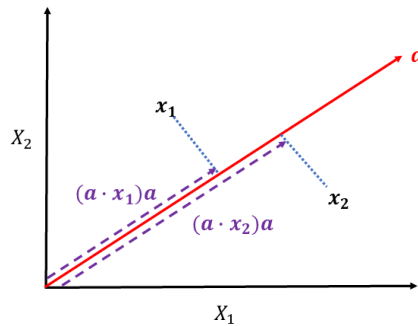


Figure 3: 観測データの直交射影 ($p = 2, n = 2$ の場合)

ベクトル \mathbf{a} の選択の指針

- 射影による特徴量の構成

ベクトル \mathbf{a} を **うまく** 選んで観測データ $\mathbf{x}_1, \dots, \mathbf{x}_n$ の情報を最も保持する 1 変量データ z_1, \dots, z_n を構成

$$z_1 = \mathbf{a}^\top \mathbf{x}_1, z_2 = \mathbf{a}^\top \mathbf{x}_2, \dots, z_n = \mathbf{a}^\top \mathbf{x}_n$$

- 特徴量のばらつきの最大化

観測データの **ばらつき** を最も反映するベクトル \mathbf{a} を選択

$$\arg \max_{\mathbf{a}} \sum_{i=1}^n (\mathbf{a}^\top \mathbf{x}_i - \mathbf{a}^\top \bar{\mathbf{x}})^2, \quad \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i,$$

ベクトル \mathbf{a} の最適化

- 最適化問題

制約条件 $\|\mathbf{a}\| = 1$ の下で以下の関数を最大化せよ

$$f(\mathbf{a}) = \sum_{i=1}^n (\mathbf{a}^\top \mathbf{x}_i - \mathbf{a}^\top \bar{\mathbf{x}})^2$$

- この最大化問題は必ず解をもつ
 - $f(\mathbf{a})$ は連続関数
 - 集合 $\{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\| = 1\}$ はコンパクト (有界閉集合)

第1主成分の解

行列による表現

- 中心化したデータ行列

$$X = \begin{pmatrix} \mathbf{x}_1^\top - \bar{\mathbf{x}}^\top \\ \vdots \\ \mathbf{x}_n^\top - \bar{\mathbf{x}}^\top \end{pmatrix} = \begin{pmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1p} - \bar{x}_p \\ \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{np} - \bar{x}_p \end{pmatrix}$$

- 評価関数 $f(\mathbf{a})$ は行列 $X^\top X$ の二次形式

$$f(\mathbf{a}) = \mathbf{a}^\top X^\top X \mathbf{a}$$

ベクトル \mathbf{a} の解

- 最適化問題

$$\text{maximize } f(\mathbf{a}) = \mathbf{a}^\top X^\top X \mathbf{a} \quad \text{s.t. } \mathbf{a}^\top \mathbf{a} = 1$$

- 制約付き最適化なので未定係数法を用いればよい

$$L(\mathbf{a}, \lambda) = f(\mathbf{a}) + \lambda(1 - \mathbf{a}^\top \mathbf{a})$$

の鞍点

$$\frac{\partial}{\partial \mathbf{a}} L(\mathbf{a}, \lambda) = 0$$

を求めればよいので

$$2X^\top X \mathbf{a} - 2\lambda \mathbf{a} = 0$$

$$X^\top X \mathbf{a} = \lambda \mathbf{a} \quad (\text{固有値問題})$$

- 解の条件

$f(\mathbf{a})$ の極大値を与える \mathbf{a} は $X^\top X$ の固有ベクトルとなる

$$X^\top X \mathbf{a} = \lambda \mathbf{a}$$

第1主成分

- 固有ベクトル \mathbf{a} に対する $f(\mathbf{a})$ は行列 $X^\top X$ の固有値

$$f(\mathbf{a}) = \mathbf{a}^\top X^\top X \mathbf{a} = \mathbf{a}^\top \lambda \mathbf{a} = \lambda$$

- 求める \mathbf{a} は行列 $X^\top X$ の最大固有ベクトル (長さ 1)
- **第1主成分負荷量**: 最大 (第一) 固有ベクトル \mathbf{a}
- **第1主成分得点**

$$z_{i1} = a_1 x_{i1} + \cdots + a_p x_{ip} = \mathbf{a}^\top \mathbf{x}_i, \quad (i = 1, \dots, n)$$

実習

Gram 行列の性質

Gram 行列の固有値

- $X^T X$ は半正定値行列
- $X^T X$ の固有値は 0 以上の実数
 - 固有値を重複を許して降順に並べる

$$\lambda_1 \geq \cdots \geq \lambda_p \quad (\geq 0)$$

- 固有値 λ_k に対する固有ベクトルを \mathbf{a}_k (長さ 1) とする

$$\|\mathbf{a}_k\| = 1, \quad (k = 1, \dots, p)$$

Gram 行列のスペクトル分解

- $\mathbf{a}_1, \dots, \mathbf{a}_p$ は **互いに直交** するようとりとることができる

$$j \neq k \quad \Rightarrow \quad \mathbf{a}_j^T \mathbf{a}_k = 0$$

- 行列 $X^T X$ (半正定値行列) のスペクトル分解

$$\begin{aligned} X^T X &= \lambda_1 \mathbf{a}_1 \mathbf{a}_1^T + \lambda_2 \mathbf{a}_2 \mathbf{a}_2^T + \cdots + \lambda_p \mathbf{a}_p \mathbf{a}_p^T \\ &= \sum_{k=1}^p \lambda_k \mathbf{a}_k \mathbf{a}_k^T \end{aligned}$$

- 固有値と固有ベクトルによる行列の表現

第2主成分以降の計算

第2主成分の考え方

- 第1主成分
 - 主成分負荷量: ベクトル \mathbf{a}_1
 - 主成分得点: $\mathbf{a}_1^T \mathbf{x}_i$ ($i = 1, \dots, n$)
- 第1主成分負荷量に関してデータが有する情報

$$(\mathbf{a}_1^T \mathbf{x}_i) \mathbf{a}_1 \quad (i = 1, \dots, n)$$

- 第1主成分を取り除いた観測データ (分析対象)

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i - (\mathbf{a}_1^T \mathbf{x}_i) \mathbf{a}_1 \quad (i = 1, \dots, n)$$

第2主成分の最適化

- 最適化問題

制約条件 $\|\mathbf{a}\| = 1$ の下で以下の関数を最大化せよ

$$\tilde{f}(\mathbf{a}) = \sum_{i=1}^n (\mathbf{a}^T \tilde{\mathbf{x}}_i - \mathbf{a}^T \bar{\mathbf{x}})^2 \quad \text{ただし} \quad \bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{x}}_i$$

第2主成分以降の解

行列による表現

- 中心化したデータ行列

$$\tilde{X} = \begin{pmatrix} \tilde{\mathbf{x}}_1^T - \bar{\mathbf{x}}^T \\ \vdots \\ \tilde{\mathbf{x}}_n^T - \bar{\mathbf{x}}^T \end{pmatrix} = X - X\mathbf{a}_1\mathbf{a}_1^T$$

- Gram 行列

$$\begin{aligned} \tilde{X}^T \tilde{X} &= (X - X\mathbf{a}_1\mathbf{a}_1^T)^T (X - X\mathbf{a}_1\mathbf{a}_1^T) \\ &= X^T X - X^T X\mathbf{a}_1\mathbf{a}_1^T - \mathbf{a}_1\mathbf{a}_1^T X^T X + \mathbf{a}_1\mathbf{a}_1^T X^T X\mathbf{a}_1\mathbf{a}_1^T \\ &= X^T X - \lambda_1 \mathbf{a}_1\mathbf{a}_1^T - \lambda_1 \mathbf{a}_1\mathbf{a}_1^T + \lambda_1 \mathbf{a}_1\mathbf{a}_1^T \mathbf{a}_1\mathbf{a}_1^T \\ &= X^T X - \lambda_1 \mathbf{a}_1\mathbf{a}_1^T \\ &= \sum_{k=2}^p \lambda_k \mathbf{a}_k \mathbf{a}_k^T \end{aligned}$$

第2主成分

- Gram 行列 $\tilde{X}^T \tilde{X}$ の固有ベクトル \mathbf{a}_1 の固有値は 0

$$\tilde{X}^T \tilde{X} \mathbf{a}_1 = 0$$

- Gram 行列 $\tilde{X}^T \tilde{X}$ の最大固有値は λ_2
- 解は第2固有値 λ_2 に対応する固有ベクトル \mathbf{a}_2

-
- 以下同様に第 k 主成分負荷量は $X^T X$ の第 k 固有値 λ_k に対応する固有ベクトル \mathbf{a}_k

実習

次回の予定

- 第1日：主成分分析の考え方
- 第2日：分析の評価と視覚化