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

主成分分析の例

県毎の生活環境の違いの分析

県名	地方名	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

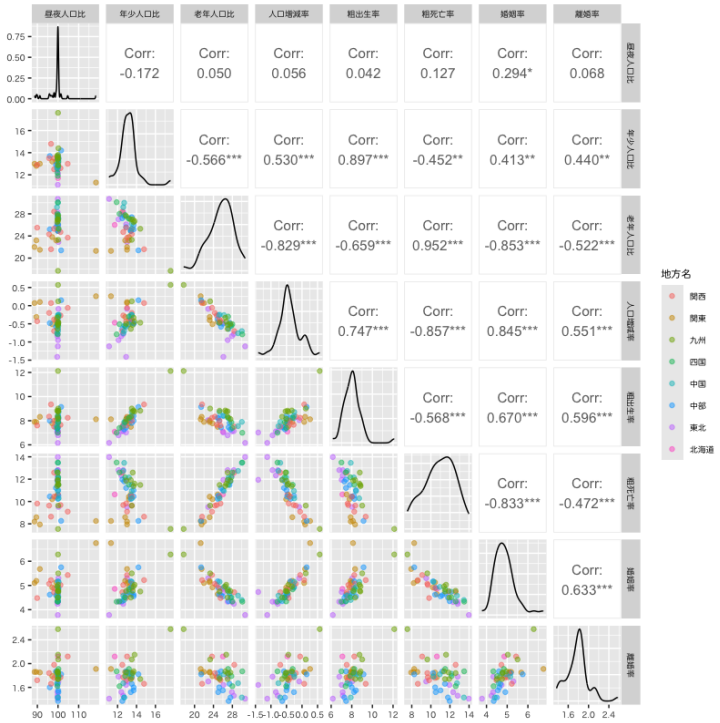

Figure 1: 県別の生活環境 (教育・労働などに関連する項目)

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
  - 特徴量に関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** (princial component)

分析の枠組み

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

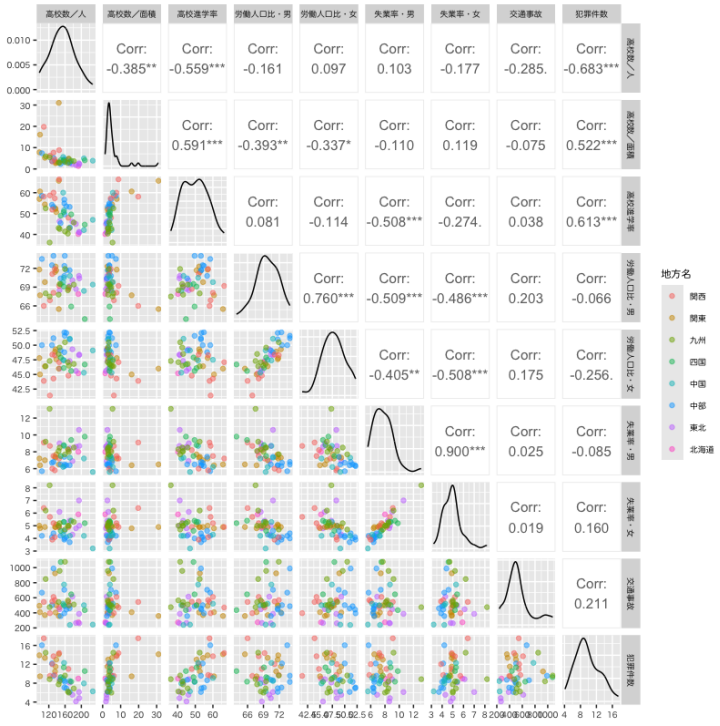

Figure 2: 県別の生活環境 (教育・労働などに関連する項目)

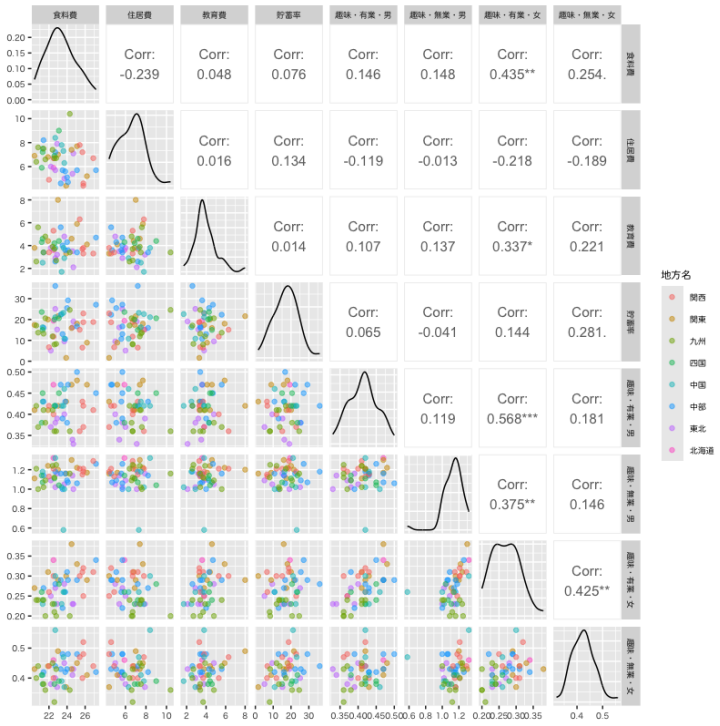

Figure 3: 県別の生活環境 (教育・労働などに関連する項目)

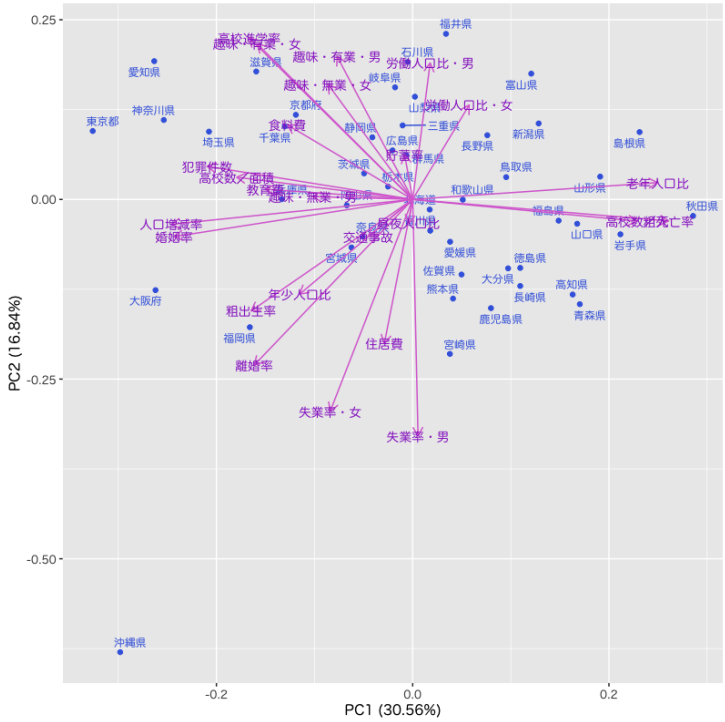

Figure 4: 県別の生活環境の主成分分析

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

- 特徴量は定数倍の任意性があるので以下を仮定

$$\|\boldsymbol{a}_k\|^2 = \sum_{j=1}^{p} a_{jk}^2 = 1$$

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**
- 係数ベクトル $\boldsymbol{a}_k$
  - 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 $z_1, \ldots, z_d$ が変数 $x_1, \ldots, x_p$ の情報を効率よく反映するように主成分負荷量 $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_d$ 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

R : 主成分分析を実行する関数

- R の標準的な関数
  - `stats::prcomp()`
  - `stats::princomp()`
- 計算法に若干の違いがある
  - 数値計算の観点からみると `prcomp()` が優位
  - `princomp()` は S 言語 (商用) との互換性を重視した実装
- 本講義では `prcomp()` を利用

R : 関数 `prcomp()` の使い方

- データフレームの全ての列を用いる場合

```
prcomp(x, retx = TRUE, center = TRUE, scale. = FALSE,
       tol = NULL, rank. = NULL, ...)
#' x: 必要な変数のみからなるデータフレーム
#' center: 中心化 (平均 0) を行って処理するか否か
#' scale.: 規格化 (分散 1) を行って処理するか否か
```

- 列名を指定する (formula を用いる) 場合

```
prcomp(formula, data = NULL, subset, na.action, ...)
#' formula: ~ 変数名 (解析の対象を + で並べる) 左辺はないので注意
#' data: 必要な変数を含むデータフレーム
#' 詳細は '?stats::prcomp' を参照
```

R : 関数 `predict()` の使い方

- 主成分得点を計算する関数

```
predict(object, newdata, ...)
#' object: prcomp が出力したオブジェクト
#' newdata: 主成分得点を計算するデータフレーム
#' 詳細は '?stats::prcomp' または '?stats::predict.prcomp' を参照
```

  - 'newdata' を省略すると分析に用いたデータフレームの得点が計算される
- 主成分分析の結果を取得する関数

```
tidy(x, matrix = "u", ...)
#' x: prcomp が出力したオブジェクト
#' matrix: 結果として取り出す行列 u:scores, v:loadings, d:eigenvalues
#' 詳細は '?broom::tidy.prcomp' を参照
```

- 主成分得点を計算する関数

```
augment(x, data = NULL, newdata, ...)
#' x: prcomp が出力したオブジェクト
#' data: 元のデータ (通常不要)
```

```
#' newdata: 主成分得点を計算するデータフレーム
#' 詳細は '?broom::augment.prcomp' を参照
```

練習問題

- 数値実験により主成分分析の考え方を確認しなさい
  - 以下のモデルに従う人工データを生成する

```
#' 観測データ（2次元）の作成（a のスカラー倍に正規乱数を重畳）
a <- c(1, 2)/sqrt(5) # 主成分負荷量（単位ベクトル）
n <- 100 # データ数
toy_data <- tibble(runif(n, -1, 1) %o% a + rnorm(2*n, sd = 0.3))
```

  - 観測データの散布図を作成
  - 観測データから第 1 主成分負荷量を推定

```
prcomp(toy_data) # 全ての主成分を計算する
a_hat <- prcomp(toy_data)$rotation[,1] # 負荷量 (rotation) の 1 列目が第 1 主成分
```

  - 散布図上に主成分負荷量を描画

```
geom_abline(slope = 傾き, intercept = 切片) # 指定の直線を追加できる
```

第 1 主成分の計算

記号の準備

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

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

- ベクトル表現
  - $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T}$: i 番目の観測データ (p 次元空間内の 1 点)
  - $\boldsymbol{a} = (a_1, \ldots, a_p)^\mathsf{T}$: 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 $\boldsymbol{x}_i$ の $\boldsymbol{a}$ 方向成分の長さ

$$\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_i \quad (\text{スカラー})$$

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

$$(\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_i)\,\boldsymbol{a} \quad (\text{スカラー} \times \text{ベクトル})$$

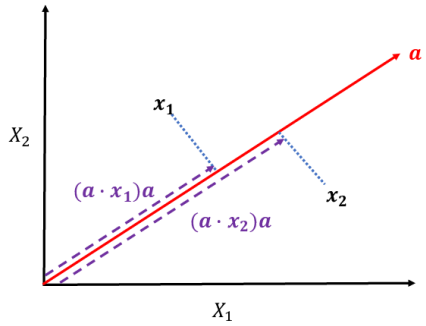

Figure 5: 観測データの直交射影 ($p = 2, n = 2$ の場合)

幾何学的描像

ベクトル a の選択の指針

- 射影による特徴量の構成

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

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

- 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 $\boldsymbol{a}$ を選択

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

ベクトル a の最適化

- 最適化問題

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

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

- この最大化問題は必ず解をもつ
  - $f(\boldsymbol{a})$ は連続関数
  - 集合 $\{\boldsymbol{a} \in \mathbb{R}^p : \|\boldsymbol{a}\| = 1\}$ 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

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

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

$$f(\boldsymbol{a}) = \boldsymbol{a}^\mathsf{T}X^\mathsf{T}X\boldsymbol{a}$$

ベクトル $\boldsymbol{a}$ の解

- 最適化問題

$$\text{maximize} \quad f(\boldsymbol{a}) = \boldsymbol{a}^\mathsf{T}X^\mathsf{T}X\boldsymbol{a} \quad \text{s.t.} \quad \boldsymbol{a}^\mathsf{T}\boldsymbol{a} = 1$$

- 制約付き最適化なので未定係数法を用いればよい

$$L(\boldsymbol{a}, \lambda) = f(\boldsymbol{a}) + \lambda(1 - \boldsymbol{a}^\mathsf{T}\boldsymbol{a})$$

の鞍点

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

を求めればよいので

$$2X^\mathsf{T}X\boldsymbol{a} - 2\lambda\boldsymbol{a} = 0$$
$$X^\mathsf{T}X\boldsymbol{a} = \lambda\boldsymbol{a} \quad \text{(固有値問題)}$$

- 解の条件

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

$$X^\mathsf{T}X\boldsymbol{a} = \lambda\boldsymbol{a}$$

第 1 主成分

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

$$f(\boldsymbol{a}) = \boldsymbol{a}^\mathsf{T}X^\mathsf{T}X\boldsymbol{a} = \boldsymbol{a}^\mathsf{T}\lambda\boldsymbol{a} = \lambda$$

- 求める $\boldsymbol{a}$ は行列 $X^\mathsf{T}X$ 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 $\boldsymbol{a}$
- **第 1 主成分得点**

$$z_{i1} = a_1x_{i1} + \cdots + a_px_{ip} = \boldsymbol{a}^\mathsf{T}\boldsymbol{x}_i, \quad (i = 1, \ldots, n)$$

実習

練習問題

- 第 1 主成分と Gram 行列の固有ベクトルの関係を調べなさい
  - 人工データを生成する
  - 主成分分析を実行する
  - Gram 行列を計算し固有値・固有ベクトルを求める

```r
#' 中心化を行う
X <- scale(toy_data, scale = FALSE)
#' 詳細は '?base::scale' を参照
#' Gram 行列を計算する
G <- crossprod(X)
#' 固有値・固有ベクトルを求める
eigen(G) # 返り値 'values, vectors' を確認
#' 詳細は '?base::eigen' を参照
```

Gram 行列の性質

Gram 行列の固有値

- $X^\top X$ は非負定値対称行列
- $X^\top X$ 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 $\boldsymbol{a}_k$(長さ 1) とする

$$\|\boldsymbol{a}_k\| = 1, \quad (k = 1, \ldots, p)$$

Gram 行列のスペクトル分解

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

$$j \neq k \quad \Rightarrow \quad \boldsymbol{a}_j^\top \boldsymbol{a}_k = 0$$

- 行列 $X^\top X$ (非負定値対称行列) のスペクトル分解

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

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

$$(\boldsymbol{a}_1^\top \boldsymbol{x}_i)\, \boldsymbol{a}_1 \quad (i = 1, \ldots, n)$$

- 第 1 主成分を取り除いた観測データ (分析対象)

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - (\boldsymbol{a}_1^\top \boldsymbol{x}_i)\, \boldsymbol{a}_1 \quad (i = 1, \ldots, n)$$

第 2 主成分の最適化

- 最適化問題

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

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

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

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

- Gram 行列

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

第 2 主成分

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

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

- Gram 行列 $\tilde{X}^\mathsf{T}\tilde{X}$ 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 $\boldsymbol{a}_2$

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

実習

データセットの準備

- 主成分分析では以下のデータセットを使用する
  - `japan_social.csv` (配付)

    総務省統計局より取得した都道府県別の社会生活統計指標の一部
    - Pref : 都道府県名
    - Forest : 森林面積割合 (%) 2014 年

    - Agri : 就業者 1 人当たり農業産出額 (販売農家）(万円) 2014 年
    - Ratio : 全国総人口に占める人口割合 (%) 2015 年
    - Land : 土地生産性（耕地面積 1 ヘクタール当たり）(万円) 2014 年
    - Goods : 商業年間商品販売額［卸売業＋小売業］（事業所当たり）(百万円) 2013 年
    - Area : 地方区分
  - 参考 : https://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0

練習問題

- 前掲のデータを用いて主成分分析を行いなさい
  - 都道府県名を行名としてデータを読み込む

```
js_data <- read_csv("data/japan_social.csv")
```

  - データの散布図行列を描く
  - 各データの箱ひげ図を描き，変数の大きさを確認する
  - 主成分負荷量を計算する

```
js_pca <- prcomp(js_data[-c(1,7)], scale. = TRUE)
#' '-c(1,7)' は都道府県名・地方区分を除く．関数 select() を利用することもできる
#' 'scale.=TRUE' とすると変数を正規化してから解析する
```

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**