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	年少人口比	老年人口比	婚姻率	離婚率	高校数／面積	交通事故	犯罪件数	食料費	住居費	貯蓄率
北海道	11.7	26.0	4.86	2.12	1.34	274.2	8.98	22.5	5.9	19.4
青森県	12.1	27.0	4.33	1.78	2.63	386.7	6.12	24.6	5.4	16.2
岩手県	12.4	27.9	4.32	1.52	2.19	261.6	4.83	24.7	7.1	19.1
宮城県	13.0	22.9	5.30	1.70	3.18	447.7	8.85	23.7	5.0	9.5
秋田県	11.1	30.7	3.78	1.41	1.85	266.2	4.12	22.9	7.9	15.2
山形県	12.6	28.3	4.24	1.46	2.24	614.9	5.54	22.2	6.0	5.1
福島県	12.9	26.1	4.73	1.64	2.65	498.9	8.13	22.7	5.7	25.1
茨城県	13.2	23.8	4.92	1.79	3.09	500.6	13.00	21.9	6.8	20.0
栃木県	13.2	23.2	5.13	1.85	2.68	404.3	11.53	20.4	6.9	17.4
群馬県	13.4	24.9	4.64	1.77	3.56	925.2	10.49	23.9	4.4	1.6
埼玉県	13.0	22.0	5.10	1.86	7.81	493.6	13.91	23.0	7.0	21.6
千葉県	12.8	23.2	5.19	1.86	5.24	370.2	13.36	26.2	5.2	8.9
東京都	11.3	21.3	6.75	1.91	31.03	358.5	14.13	25.1	7.7	18.0
神奈川県	13.0	21.5	5.68	1.85	16.09	408.6	9.46	24.5	7.4	15.9
新潟県	12.5	27.2	4.35	1.37	2.35	357.2	8.71	23.5	5.7	13.0
富山県	12.7	27.6	4.50	1.43	2.86	459.6	6.14	22.7	7.2	36.1
石川県	13.4	25.0	4.91	1.52	4.03	443.3	6.93	23.3	4.6	22.3
福井県	13.7	26.0	4.55	1.55	3.72	394.0	7.07	24.1	4.7	29.2
山梨県	12.9	25.6	4.60	1.87	4.62	706.0	8.61	25.1	5.7	16.7
長野県	13.5	27.4	4.67	1.66	3.14	487.9	8.27	21.4	8.2	16.0
岐阜県	13.7	25.2	4.62	1.60	3.68	502.3	12.18	24.0	5.1	25.5
静岡県	13.4	24.9	5.17	1.84	5.23	989.2	9.58	23.8	5.6	10.7
愛知県	14.2	21.4	5.75	1.82	7.39	668.5	16.04	27.2	5.7	27.1
三重県	13.5	25.3	4.89	1.76	3.52	551.9	12.03	22.0	4.9	6.7
滋賀県	14.8	21.6	5.22	1.66	4.47	570.4	9.73	25.8	4.6	18.7
京都府	12.6	24.7	5.02	1.77	8.83	471.3	14.37	26.9	6.7	18.8
大阪府	13.0	23.7	5.43	2.12	19.77	544.4	17.52	25.4	7.8	15.1
兵庫県	13.5	24.3	5.07	1.84	7.67	611.3	13.71	25.7	7.2	6.5
奈良県	12.9	25.5	4.48	1.72	6.22	395.6	9.55	22.5	6.7	11.8
和歌山県	12.5	28.4	4.72	1.98	4.65	547.6	11.01	25.8	4.4	22.9
鳥取県	13.2	27.2	4.74	1.83	3.40	238.7	8.45	23.7	6.3	10.8
島根県	12.7	30.0	4.40	1.43	3.88	244.0	6.27	23.4	7.8	22.6
岡山県	13.5	26.2	4.94	1.82	4.04	775.9	12.30	22.7	8.4	21.1
広島県	13.5	25.3	5.15	1.78	5.63	521.4	9.08	23.6	7.3	23.2
山口県	12.6	29.2	4.58	1.67	4.95	501.5	7.94	21.3	7.5	16.3
徳島県	12.2	28.0	4.34	1.62	3.81	645.9	8.32	21.3	7.2	15.0

香川県	13.2	27.1	4.84	1.91	4.19	1075.5	9.27	21.2	5.9	23.6
愛媛県	12.8	27.8	4.51	1.79	4.02	502.3	11.35	23.1	9.0	20.6
高知県	11.9	30.1	4.33	1.87	4.05	435.6	10.56	22.6	6.9	20.0
福岡県	13.5	23.3	5.50	2.07	5.94	849.1	14.46	22.4	7.4	11.9
佐賀県	14.4	25.3	4.75	1.74	3.38	1078.3	9.62	21.6	6.6	20.6
長崎県	13.4	27.0	4.50	1.74	4.83	499.4	5.99	22.8	6.7	8.3
熊本県	13.7	26.5	4.96	1.87	3.00	543.3	7.75	21.4	7.5	13.2
大分県	12.9	27.6	4.77	1.85	3.67	511.3	6.88	20.6	6.4	17.1
宮崎県	13.8	26.7	5.03	2.15	2.93	957.3	8.39	22.6	6.7	8.0
鹿児島県	13.6	27.0	4.78	1.84	2.81	565.3	6.24	20.8	7.6	13.7
沖縄県	17.6	17.7	6.28	2.58	5.48	475.3	8.85	24.3	10.4	24.6

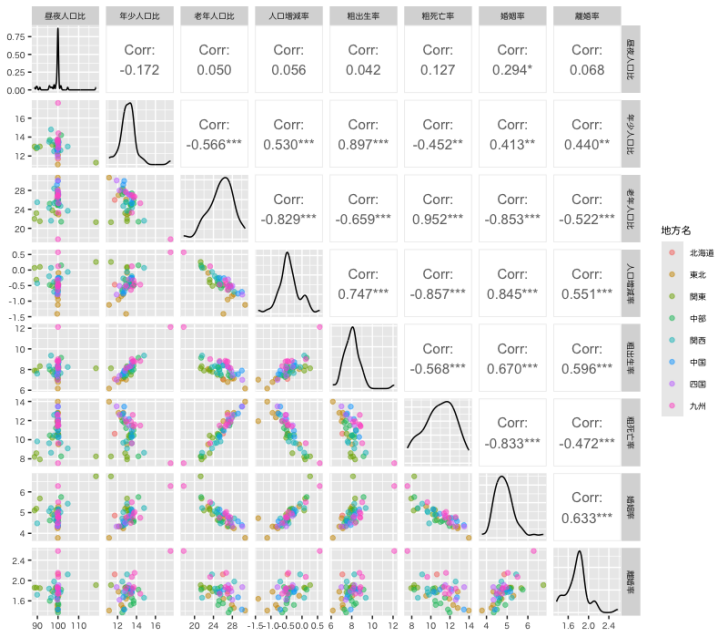

Figure 1: 県別の生活環境 (教育・労働などに関連する項目)

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)

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

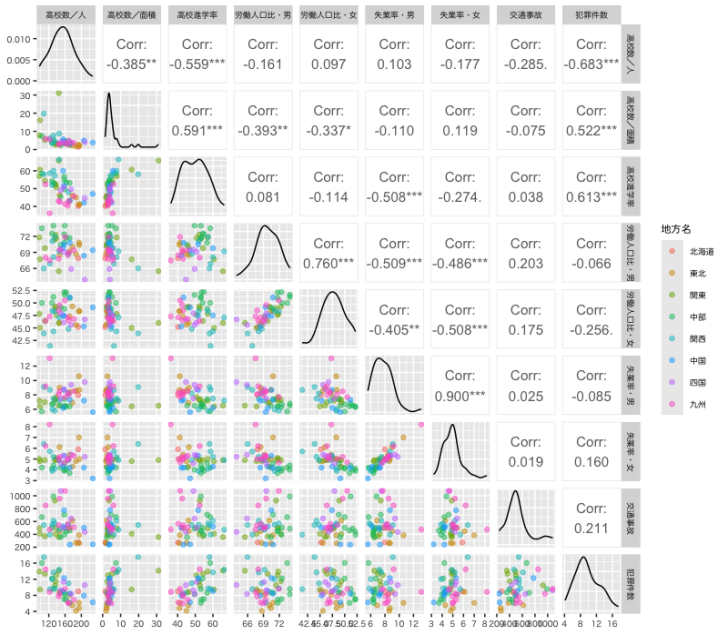

Figure 2: 県別の生活環境 (教育・労働などに関連する項目)

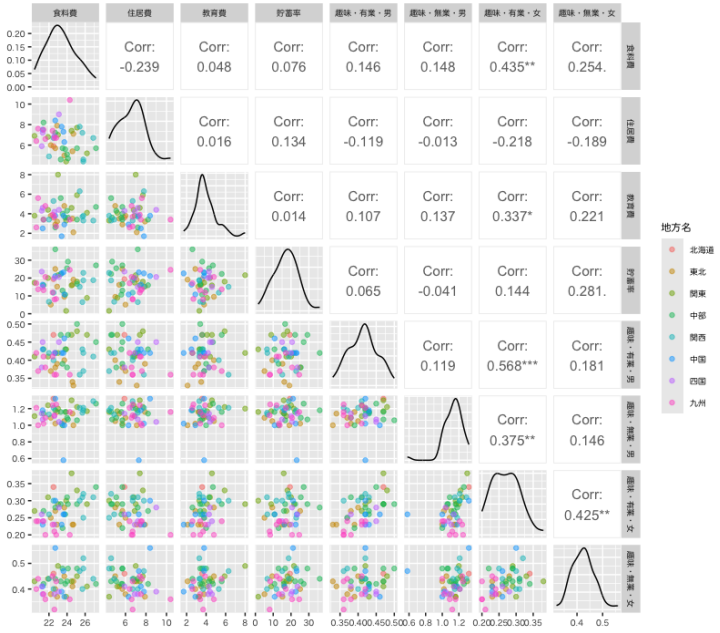

Figure 3: 県別の生活環境 (教育・労働などに関連する項目)

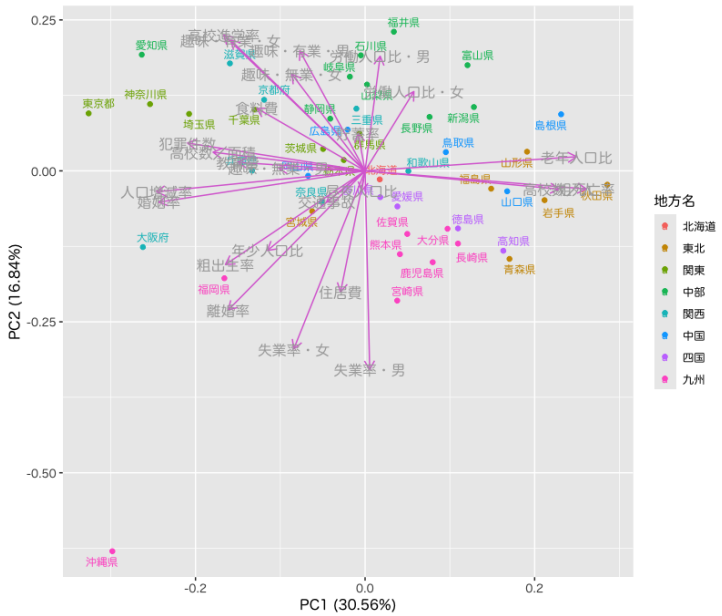

Figure 4: 県別の生活環境の主成分分析

- 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：分析結果を取得する関数

- 主成分得点を計算する

```
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

```
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

```
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

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

幾何学的描像

ベクトル $\boldsymbol{a}$ 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

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

- 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

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

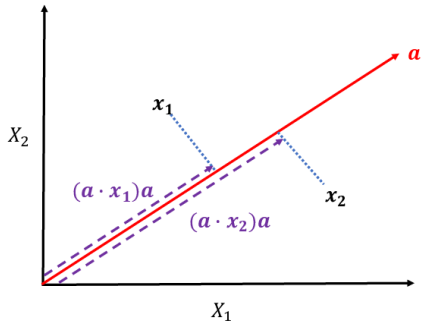

Figure 5: 観測データの直交射影 ($p = 2, n = 2$ の場合)

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

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

- 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

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

- 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 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

```
#' 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

- $X^\mathsf{T}X$ は非負定値対称行列
- $X^\mathsf{T}X$ 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 $\boldsymbol{a}_k$(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

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

- 行列 $X^\mathsf{T}X$ (非負定値対称行列) のスペクトル分解

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

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

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

- 第 1 主成分を取り除いた観測データ (分析対象)

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

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

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

- Gram 行列

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

第 2 主成分

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

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

- Gram 行列 $\tilde{X}^\top\tilde{X}$ 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 $X^\top X$ 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**