主成分分析

評価と視覚化

村田 昇

講義の内容

- 第 1 日 : 主成分分析の考え方
- **第 2 日 : 分析の評価と視覚化**

主成分分析の復習

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
  - 変量の間の関係を明らかにする
- 分析の方針
  - データの情報を最大限保持する変量の線形結合を構成
  - データの情報を最大限反映する座標 (方向) を探索
  - **データの情報を保持する = データを区別することができる**

分析の考え方

- 1 変量の特徴量 $\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_1, \dots, \boldsymbol{a}^\mathsf{T}\boldsymbol{x}_n$ を構成
  - 観測データ $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ のもつ情報を最大限保持するベクトル $\boldsymbol{a}$ を **適切に** 選択
  - $\boldsymbol{a}^\mathsf{T}\boldsymbol{x}_1, \dots, \boldsymbol{a}^\mathsf{T}\boldsymbol{x}_n$ の変動 (ばらつき) が最も大きい方向を選択
- **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

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

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

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

固有値問題

- 最適化問題

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

- 解の条件

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

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

  - 未定係数法を用いている

主成分負荷量と主成分得点

- $\boldsymbol{a}$: **主成分負荷量** (principal component loading)
- $\boldsymbol{a}^\mathsf{T} \boldsymbol{x}_i$: **主成分得点** (principal component score)
- 第 1 主成分負荷量

  $X^\mathsf{T} X$ の第 1(最大) 固有値 λ_1 に対応する固有ベクトル $\boldsymbol{a}_1$
- 第 k 主成分負荷量

  $X^\mathsf{T} X$ の第 k 固有値 λ_k に対応する固有ベクトル $\boldsymbol{a}_k$

演習

問題

- 以下の問に答えなさい
  - ベクトル $\boldsymbol{a}$ を $X^\mathsf{T} X$ の単位固有ベクトルとするとき

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

    の値を求めよ
  - 行列 X を中心化したデータ行列，ベクトル $\boldsymbol{a}_k$ を第 k 主成分負荷量とするとき，第 k 主成分得点の平均まわりの平方和

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

    を X と $\boldsymbol{a}_k$ で表せ

解答例

- 固有値・固有ベクトルの性質を利用する

  $X^\mathsf{T} X$ の固有値・固有ベクトルを $\lambda_k, \boldsymbol{a}_k$ とする．$\boldsymbol{a} = \boldsymbol{a}_k$ とすれば

  $$\begin{aligned} f(\boldsymbol{a}_k) &= \boldsymbol{a}_k^\mathsf{T} X^\mathsf{T} X \boldsymbol{a}_k \\ &= \boldsymbol{a}_k^\mathsf{T} \lambda_k \boldsymbol{a}_k && \text{(固有ベクトル)} \\ &= \lambda_k && \text{(単位ベクトル)} \end{aligned}$$

  - 定義に従い計算すればよい (前回の復習)

$$
\begin{aligned}
f(\boldsymbol{a}_k) &= \sum_{i=1}^{n} (\boldsymbol{a}_k^\mathsf{T} \boldsymbol{x}_i - \boldsymbol{a}_k^\mathsf{T} \bar{\boldsymbol{x}})^2 \\
&= \sum_{i=1}^{n} (X\boldsymbol{a}_k)_i^2 \\
&= \sum_{i=1}^{n} (X\boldsymbol{a}_k)_i (X\boldsymbol{a}_k)_i \\
&= (\boldsymbol{a}_k^\mathsf{T} X^\mathsf{T})(X\boldsymbol{a}_k) = \boldsymbol{a}_k^\mathsf{T} X^\mathsf{T} X \boldsymbol{a}_k
\end{aligned}
$$

寄与率

寄与率の考え方

- 回帰分析で考察した寄与率の一般形

$$
(\text{寄与率}) = \frac{(\text{その方法で説明できる変動})}{(\text{データ全体の変動})}
$$

- 主成分分析での定義 (proportion of variance)

$$
(\text{寄与率}) = \frac{(\text{主成分の変動})}{(\text{全体の変動})}
$$

Gram 行列のスペクトル分解

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

$$
X^\mathsf{T} X = \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_k^\mathsf{T}
$$

  - 固有値と固有ベクトルによる行列の表現

- 主成分の変動の評価

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

  - 固有ベクトル (単位ベクトル) の直交性を利用

寄与率の計算

- 主成分と全体の変動

$$
\begin{aligned}
(\text{主成分の変動}) &= \sum_{i=1}^{n} (\boldsymbol{a}_k^\mathsf{T} \boldsymbol{x}_i - \boldsymbol{a}_k^\mathsf{T} \bar{\boldsymbol{x}})^2 = \boldsymbol{a}_k^\mathsf{T} X^\mathsf{T} X \boldsymbol{a}_k = \lambda_k \\
(\text{全体の変動}) &= \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 = \sum_{l=1}^{p} \boldsymbol{a}_l^\mathsf{T} X^\mathsf{T} X \boldsymbol{a}_l = \sum_{l=1}^{p} \lambda_l
\end{aligned}
$$

- 固有値による寄与率の表現

$$
(\text{寄与率}) = \frac{\lambda_k}{\sum_{l=1}^{p} \lambda_l}
$$

全体の変動

- スペクトル分解との関係

$$
\begin{aligned}
(\text{全体の変動}) &= \sum_{i=1}^{n} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}\|^2 \\
&= \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathsf{T}} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \\
&= \mathrm{tr} X X^{\mathsf{T}} = \mathrm{tr} X^{\mathsf{T}} X \\
&= \mathrm{tr} \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_k \boldsymbol{a}_l^{\mathsf{T}} = \sum_{k=1}^{p} \lambda_k \mathrm{tr} \boldsymbol{a}_k \boldsymbol{a}_l^{\mathsf{T}} = \sum_{k=1}^{p} \lambda_k \boldsymbol{a}_l^{\mathsf{T}} \boldsymbol{a}_k \\
&= \sum_{k=1}^{p} \lambda_k
\end{aligned}
$$

累積寄与率

- **累積寄与率** (cumulative proportion)

  第 k 主成分までの変動の累計

$$
(\text{累積寄与率}) = \frac{\sum_{l=1}^{k} \lambda_l}{\sum_{l=1}^{p} \lambda_l}
$$

  - 累積寄与率はいくつの主成分を用いるべきかの基準
  - 一般に累積寄与率が 80%程度までの主成分を用いる

解析の事例

データセット

- 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
  - 総務省 https://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0
    - 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
- データ (の一部) の内容

各変数の分布

- 変数間の散布図

- 変数のばらつきに大きな違いがある

Table 1: 社会生活統計指標

Pref	Forest	Agri	Ratio	Land	Goods	Area
Hokkaido	67.9	1150.6	4.23	96.8	283.3	1
Aomori	63.8	444.7	1.03	186	183	2
Iwate	74.9	334.3	1.01	155.2	179.4	2
Miyagi	55.9	299.9	1.84	125.3	365.9	2
Akita	70.5	268.7	0.81	98.5	153.3	2
Yamagata	68.7	396.3	0.88	174.1	157.5	2
Fukushima	67.9	236.4	1.51	127.1	184.5	2
Ibaraki	31	479	2.3	249.1	204.9	3
Tochigi	53.2	402.6	1.55	199.6	204.3	3
Gumma	63.8	530.6	1.55	321.6	270	3
Saitama	31.9	324.7	5.72	247	244.7	3
Chiba	30.4	565.5	4.9	326.1	219.7	3
Tokyo	34.8	268.5	10.63	404.7	1062.6	3
Kanagawa	38.8	322.8	7.18	396.4	246.1	3
Niigata	63.5	308.6	1.81	141.9	205.5	4
Toyama	56.6	276.1	0.84	98.5	192.4	4
Ishikawa	66	271.3	0.91	112	222.9	4
Fukui	73.9	216.1	0.62	98.5	167.3	4
Yamanashi	77.8	287.4	0.66	325.3	156.2	4
Nagano	75.5	280	1.65	211.3	194.4	4
Gifu	79	283.7	1.6	192.1	167.9	4
Shizuoka	63.1	375.8	2.91	314.5	211.4	4
Aichi	42.2	472.3	5.89	388.9	446.9	4
Mie	64.3	310.6	1.43	174.3	170.1	5
Shiga	50.5	222.8	1.11	104.9	170.7	5
Kyoto	74.2	267.8	2.05	212.5	196.7	5
Osaka	30.1	216.3	6.96	238.8	451.2	5
Hyogo	66.7	261.2	4.35	197.7	212.5	5
Nara	76.8	207	1.07	182.7	147	5
Wakayama	76.4	251.1	0.76	278.4	136.4	5
Tottori	73.3	249.9	0.45	187.6	162.2	6
Shimane	77.5	214.1	0.55	140.8	141.1	6
Okayama	68	254.8	1.51	184.9	207.8	6
Hiroshima	71.8	286.2	2.24	192.2	304.6	6
Yamaguchi	71.6	216.9	1.11	125.8	158.9	6
Tokushima	75.2	315.4	0.59	313.5	134.5	7
Kagawa	46.4	249.5	0.77	242.9	232.9	7
Ehime	70.3	288.5	1.09	231.6	179.4	7
Kochi	83.3	354.2	0.57	339.9	137.9	7
Fukuoka	44.5	381	4.01	255.6	295.7	8
Saga	45.2	468.7	0.66	230.3	137.9	8
Nagasaki	58.4	428.9	1.08	296	154	8
Kumamoto	60.4	456.6	1.41	285.5	172.5	8
Oita	70.7	360.1	0.92	222.8	148.3	8
Miyazaki	75.8	739.1	0.87	487.7	170.6	8
Kagoshima	63.4	736.5	1.3	351.2	169.4	8
Okinawa	46.1	452.4	1.13	232.8	145.4	8

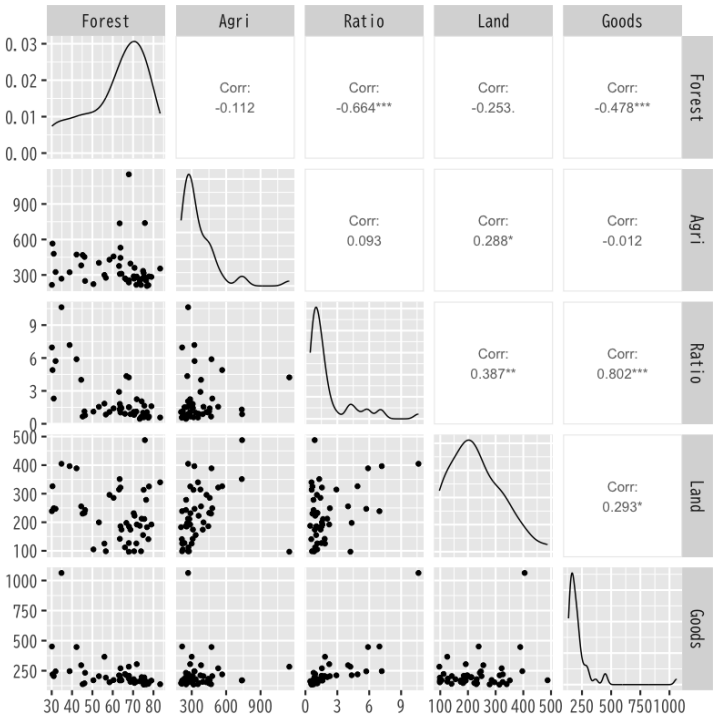

Figure 1: データの散布図

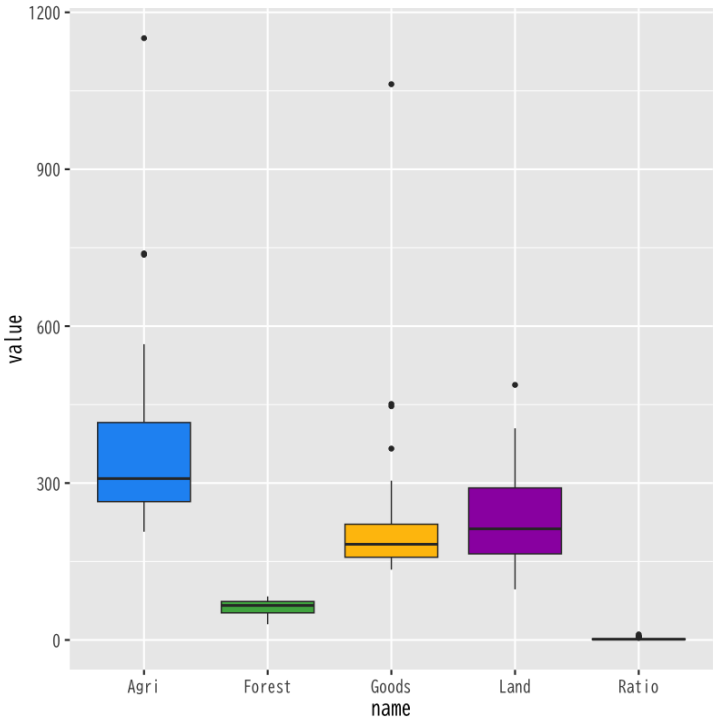

Figure 2: 各変数の箱ひげ図

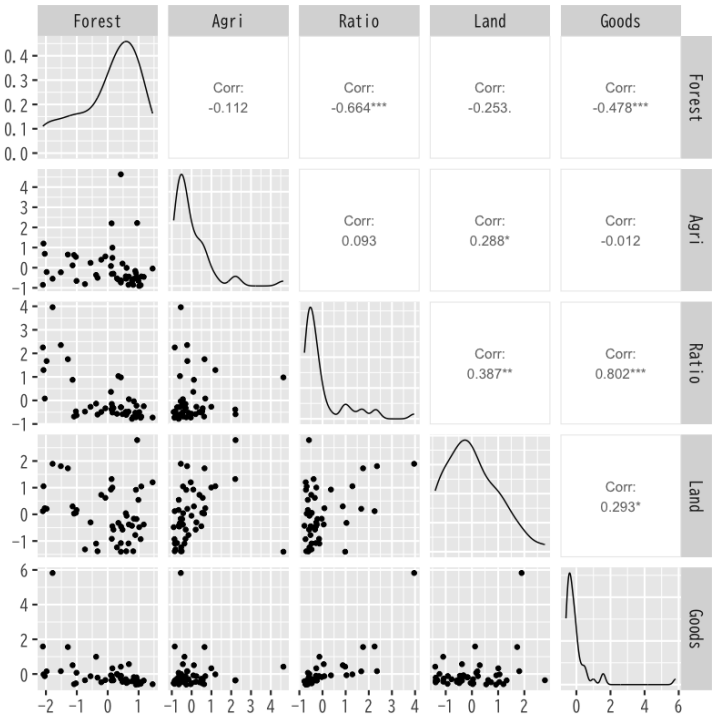

Figure 3: 正規化したデータの散布図

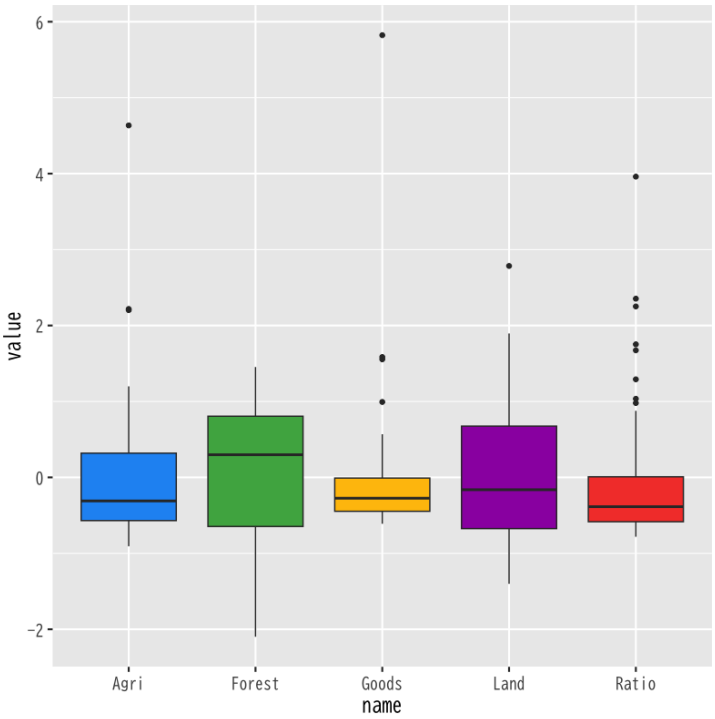

Figure 4: 各変数の箱ひげ図

データの正規化

- 各変数の標本平均を 0，不偏分散を 1 に規格化する
- 変数のばらつきをそろえる

主成分分析

- 主成分負荷量 (正規化なし)

Table 2

	PC1	PC2	PC3	PC4	PC5
Forest	-0.014	0.048	-0.0004	-0.998	-0.049
Agri	0.973	0.121	-0.197	-0.008	0.0004
Ratio	0.002	-0.012	0.00002	0.049	-0.999
Land	0.222	-0.247	0.943	-0.016	0.003
Goods	0.065	-0.960	-0.267	-0.048	0.009

  - 第 1: 分散が大きく関連している Agri と Land が支配的
  - 第 2: 次に分散が大きな Goods が支配的
- 寄与率

Table 3

	PC1	PC2	PC3	PC4	PC5
Standard deviation	173.275	148.037	81.523	12.972	1.052
Proportion of Variance	0.511	0.373	0.113	0.003	0.00002
Cumulative Proportion	0.511	0.884	0.997	1.000	1

- 第 1,2 主成分得点の表示
- 第 3,2 主成分得点の表示
- 主成分負荷量 (正規化あり)

Table 4

	PC1	PC2	PC3	PC4	PC5
Forest	-0.487	0.105	-0.457	0.686	-0.268
Agri	0.134	0.812	0.479	0.305	0.035
Ratio	0.585	-0.151	0.045	0.164	-0.778
Land	0.355	0.485	-0.742	-0.290	0.069
Goods	0.526	-0.269	-0.095	0.571	0.562

  - 第 1: 人の多さに関する成分 (正の向きほど人が多い)
  - 第 2: 農業生産力に関する成分 (正の向きほど高い)
- 寄与率
- 第 1,2 主成分得点の表示
- 第 3,2 主成分得点の表示

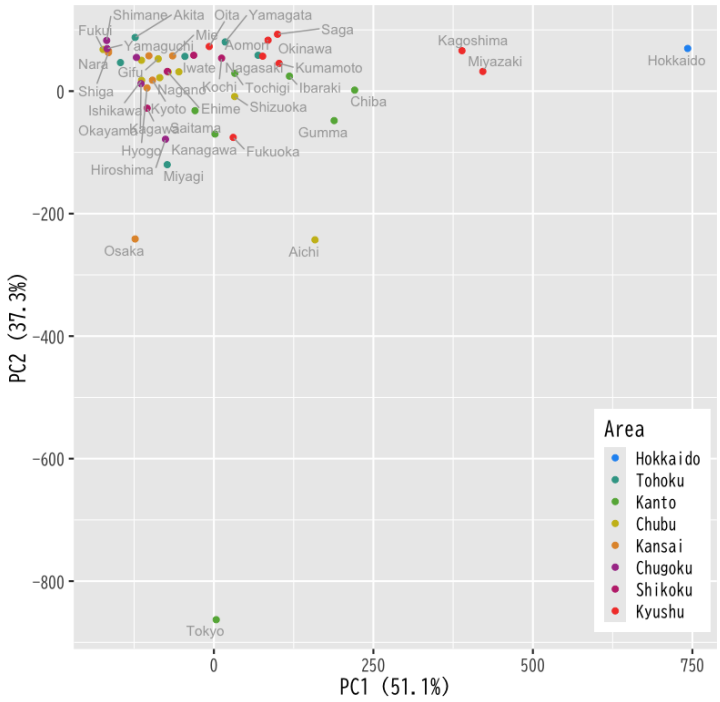

Figure 5: 主成分得点による散布図 (正規化なし)

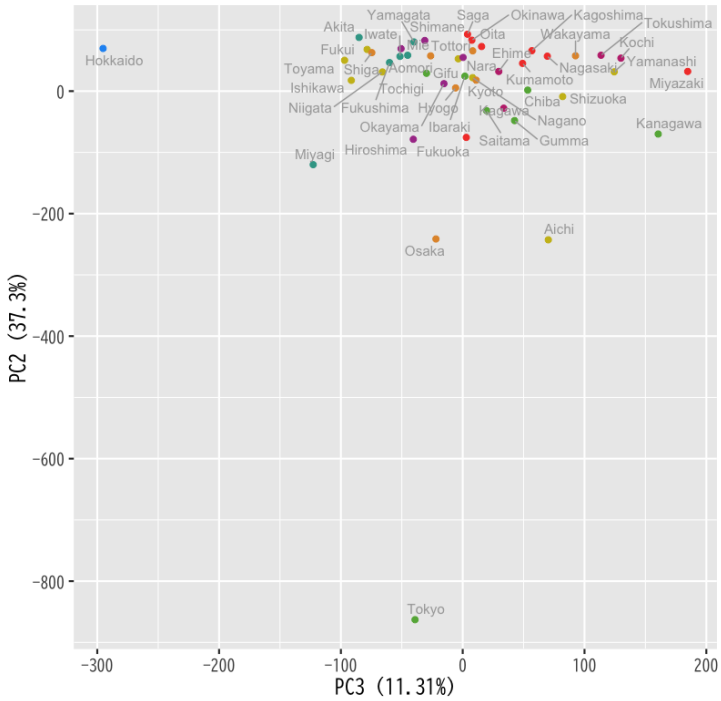

Figure 6: 主成分得点による散布図 (正規化なし)

Table 5

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.590	1.070	0.820	0.708	0.392
Proportion of Variance	0.506	0.229	0.134	0.100	0.031
Cumulative Proportion	0.506	0.735	0.869	0.969	1

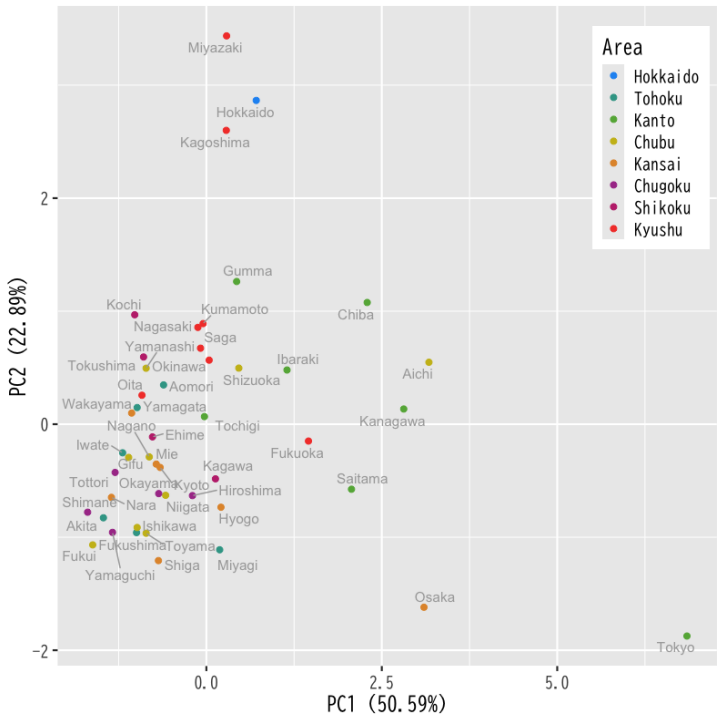

Figure 7: 主成分得点による散布図 (正規化あり)

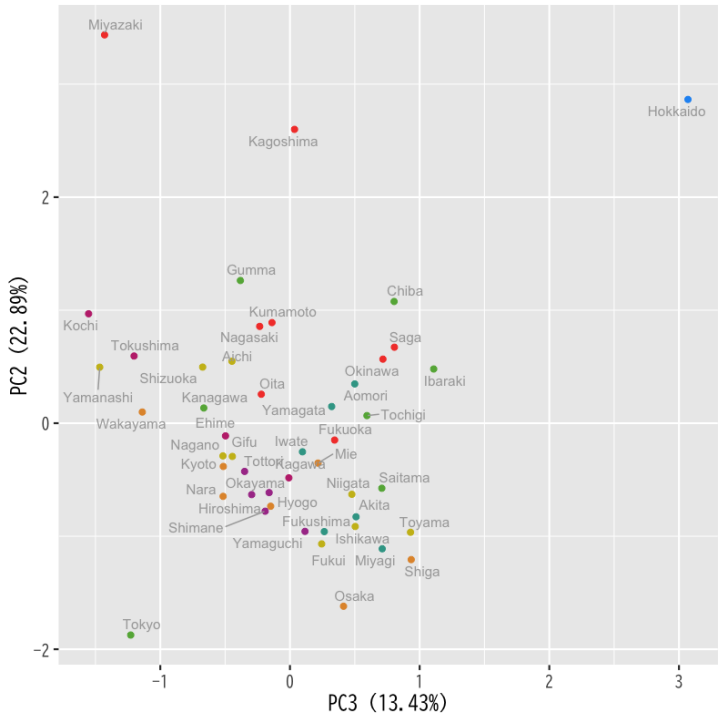

Figure 8: 主成分得点による散布図 (正規化あり)

演習

問題

- 以下の問に答えなさい
  - 正規化条件を満たす線形変換 $x'_{ij} = a_j(x_{ij} - b_j)$ を求めよ

$$\frac{1}{n}\sum_{i=1}^{n} x'_{ij} = 0, \quad \frac{1}{n-1}\sum_{i=1}^{n} (x'_{ij})^2 = 1$$

  - 正規化されたデータ行列を

$$X' = \begin{pmatrix} \boldsymbol{x}'^{\mathsf{T}}_1 \\ \vdots \\ \boldsymbol{x}'^{\mathsf{T}}_n \end{pmatrix} = \begin{pmatrix} x'_{11} & \cdots & x'_{1p} \\ \vdots & & \vdots \\ x'_{n1} & \cdots & x'_{np} \end{pmatrix}$$

    と書くとき，$X'^{\mathsf{T}}X'$ の対角成分を求めよ

解答例

- 標本平均の定義どおりに計算すればよい

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} x'_{ij} &= \frac{1}{n}\sum_{i=1}^{n} \left(a_j(x_{ij} - b_j)\right) \\ &= a_j\left(\frac{1}{n}\sum_{i=1}^{n} x_{ij} - b_j\right) \\ &= 0 \end{aligned}$$

したがって

$$b_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} = \bar{x}_j \quad \text{(元の変数の標本平均)}$$

- 不偏分散も同様に計算すればよい

$$\begin{aligned}\frac{1}{n-1}\sum_{i=1}^{n}(x'_{ij})^2 &= a_j^2 \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2 \\ &= 1\end{aligned}$$

したがって

$$a_j = \left(\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2\right)^{-1/2} \quad \text{(標準偏差の逆数)}$$

- 不偏分散での標準化であることに注意する

$$(X'^{\mathsf{T}}X')_{jj} = \sum_{i=1}^{n}(x'_{ij})^2 = n-1$$

主成分負荷量

主成分負荷量と主成分得点

- 負荷量 (得点係数) の大きさ : 変数の貢献度
- **問題点**
  - 変数のスケールによって係数の大きさは変化する
  - 変数の正規化 (平均 0，分散 1) がいつも妥当とは限らない
- スケールによらない変数と主成分の関係
  相関係数 を考えればよい

相関係数

- $\boldsymbol{e}_j$: 第 j 成分は 1，それ以外は 0 のベクトル
- $X\boldsymbol{e}_j$: 第 j 変数ベクトル
- $X\boldsymbol{a}_k$: 第 k 主成分得点ベクトル
- 主成分と変数の相関係数

$$\begin{aligned}\mathrm{Cor}(X\boldsymbol{a}_k, X\boldsymbol{e}_j) &= \frac{\boldsymbol{a}_k^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{e}_j}{\sqrt{\boldsymbol{a}_k^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{a}_k}\sqrt{\boldsymbol{e}_j^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{e}_j}} \\ &= \frac{\lambda_k \boldsymbol{a}_k^{\mathsf{T}}\boldsymbol{e}_j}{\sqrt{\lambda_k}\sqrt{(X^{\mathsf{T}}X)_{jj}}}\end{aligned}$$

正規化データの場合

- $X^{\mathsf{T}}X$ の対角成分は全て $n-1$ ($(X^{\mathsf{T}}X)_{jj} = n-1$)
  - 第 k 主成分に対する相関係数ベクトル

$$\boldsymbol{r}_k = \sqrt{\lambda_k/(n-1)} \cdot \boldsymbol{a}_k, \quad (\boldsymbol{r}_k)_j = \sqrt{\lambda_k/(n-1)} \cdot (\boldsymbol{a}_k)_j$$

  - **主成分負荷量の比較**
    - 同じ主成分 (k を固定) への各変数の影響は固有ベクトルの成分比
    - 同じ変数 (j を固定) の各主成分への影響は固有値の平方根で重みづけ
- 正規化されていない場合は変数の分散の影響を考慮

データ行列の分解表現

特異値分解

- 階数 r の $n \times p$ 型行列 X の分解

$$X = U\Sigma V^{\mathsf{T}}$$

  - U は $n \times n$ 型直交行列, V は $p \times p$ 型直交行列
  - Σ は $n \times p$ 型行列

$$\Sigma = \begin{pmatrix} D & O_{r,p-r} \\ O_{n-r,r} & O_{n-r,m-r} \end{pmatrix}$$

    - $O_{s,t}$ は $s \times t$ 型零行列
    - D は $\sigma_1 \geq \sigma_2 \geq \sigma_r > 0$ を対角成分とする $r \times r$ 型対角行列

特異値

- 行列 Σ の成分表示

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & O_{r,p-r} \\ & & \sigma_r & \\ & O_{n-r,r} & & O_{n-r,m-r} \end{pmatrix}$$

- D の対角成分 : X の **特異値** (singular value)

特異値分解による Gram 行列の表現

- Gram 行列の展開

$$\begin{aligned} X^{\mathsf{T}}X &= (U\Sigma V^{\mathsf{T}})^{\mathsf{T}}(U\Sigma V^{\mathsf{T}}) \\ &= V\Sigma^{\mathsf{T}}U^{\mathsf{T}}U\Sigma V^{\mathsf{T}} \\ &= V\Sigma^{\mathsf{T}}\Sigma V^{\mathsf{T}} \end{aligned}$$

- 行列 $\Sigma^{\mathsf{T}}\Sigma$ は対角行列

$$\Sigma^{\mathsf{T}}\Sigma = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_r^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

演習

問題

- 行列 X の特異値分解を $U\Sigma V^{\mathsf{T}}$ とし，行列 U の第 k 列ベクトルを $\boldsymbol{u}_k$，行列 V の第 k 列ベクトルを $\boldsymbol{v}_k$ とするとき，以下の問に答えなさい
  - 行列 U, V の列ベクトルを用いて X を展開しなさい
  - Gram 行列 $X^{\mathsf{T}}X$ の固有値を特異値で表しなさい
  - 行列 X の主成分負荷量を求めなさい
  - それぞれの負荷量に対応する主成分得点を求めなさい

解答例

- Σ が対角成分しか持たないことに注意すると以下のように展開される

$$X = U\Sigma V^{\mathsf{T}} = \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^{\mathsf{T}}$$

- 先週の演習問題と特異値分解を比較する

$$X^{\mathsf{T}}X = V\Sigma^{\mathsf{T}}\Sigma V^{\mathsf{T}} = A^{\mathsf{T}}\Lambda A$$

  より

$$\lambda_k = \begin{cases} \sigma_k^2, & k \leq r \\ 0, & k > r \end{cases}$$

- 転置に気をつけて同様に比較すればよい

$$A = \begin{pmatrix} \boldsymbol{a}_1^{\mathsf{T}} \\ \vdots \\ \boldsymbol{a}_p^{\mathsf{T}} \end{pmatrix}$$

  と定義されているので主成分負荷量 (固有ベクトル) は行列 V の列ベクトル

$$\boldsymbol{a}_k = \boldsymbol{v}_k$$

- 主成分得点の定義どおり計算する

$$X\boldsymbol{a}_k = U\Sigma V^{\mathsf{T}}\boldsymbol{v}_k = \sigma_k \boldsymbol{u}_k$$

  - ただし $k > r$ のとき $\sigma_k = 0$ とする
  - V と U は大きさが異なるので注意する

バイプロット

データ行列の分解

- 行列 U の第 k 列ベクトル $\boldsymbol{u}_k$
- 行列 V の第 k 列ベクトル $\boldsymbol{v}_k$
- データ行列の特異値分解 : (Σ の非零値に注意)

$$X = U\Sigma V^{\mathsf{T}} = \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^{\mathsf{T}}$$

データ行列の近似表現

- 第 k 主成分と第 l 主成分を用いた行列 X の近似 X'

$$X \simeq X' = \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^{\mathsf{T}} + \sigma_l \boldsymbol{u}_l \boldsymbol{v}_l^{\mathsf{T}}$$

- 行列の積による表現

$$\begin{aligned} X' =& GH^{\mathsf{T}}, (0 \le s \le 1) \\ & G = \begin{pmatrix} \sigma_k^{1-s} \boldsymbol{u}_k & \sigma_l^{1-s} \boldsymbol{u}_l \end{pmatrix}, \quad H = \begin{pmatrix} \sigma_k^{s} \boldsymbol{v}_k & \sigma_l^{s} \boldsymbol{v}_l \end{pmatrix} \end{aligned}$$

バイプロット

- 関連がある 2 枚の散布図を 1 つの画面に表示する散布図を一般に**バイプロット** (biplot) と呼ぶ
  - 行列 G, H の各行を 2 次元座標と見なす

$$X' = GH^{\mathsf{T}}$$

    - 行列 G の各行は各データの 2 次元座標
    - 行列 H の各行は各変量の 2 次元座標
  - X の変動を最大限保持する近似は $k = 1, l = 2$
  - パラメタ s は 0, 1 または 1/2 が主に用いられる
    - $s = 0$: データの散布図は主成分得点 (G は主成分得点)
    - $s = 1$: データの散布図は正規化される
    - $s = 1/2$: 変量ベクトルが相関ベクトル (正規化されたデータ)

解析の事例

バイプロット

- 主成分負荷量 (正規化あり)
- 寄与率
- 第 1,2 主成分によるバイプロット
- 第 3,2 主成分によるバイプロット
- 中心部の拡大 (第 1,2 主成分)
- 中心部の拡大 (第 3,2 主成分)

Table 6

	PC1	PC2	PC3	PC4	PC5
Forest	-0.487	0.105	-0.457	0.686	-0.268
Agri	0.134	0.812	0.479	0.305	0.035
Ratio	0.585	-0.151	0.045	0.164	-0.778
Land	0.355	0.485	-0.742	-0.290	0.069
Goods	0.526	-0.269	-0.095	0.571	0.562

Table 7

	PC1	PC2	PC3	PC4	PC5
Standard deviation	1.590	1.070	0.820	0.708	0.392
Proportion of Variance	0.506	0.229	0.134	0.100	0.031
Cumulative Proportion	0.506	0.735	0.869	0.969	1

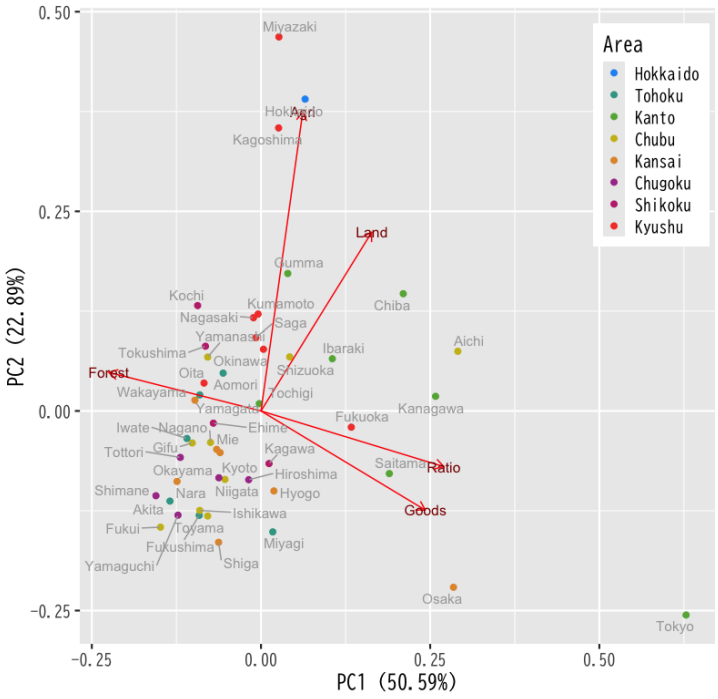

Figure 9: 主成分分析のバイプロット (第 1,2)

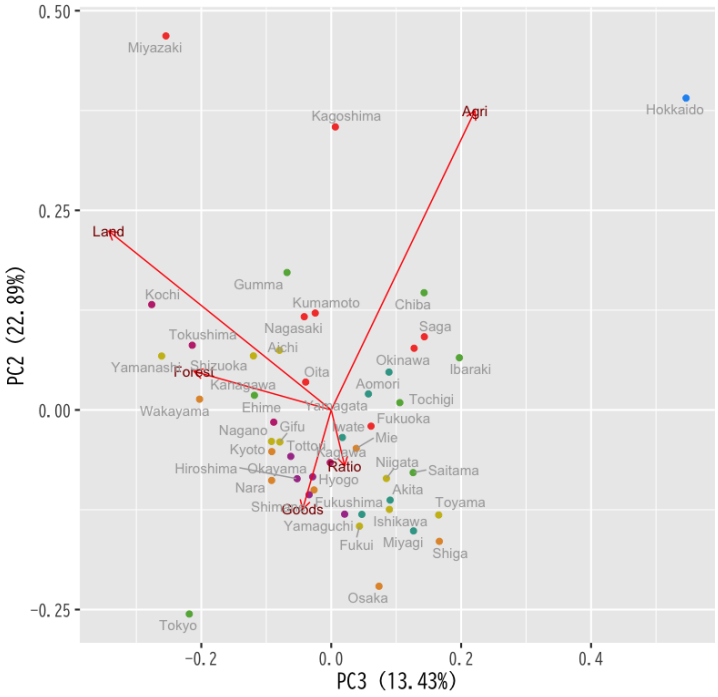

Figure 10: 主成分分析のバイプロット (第 3,2)

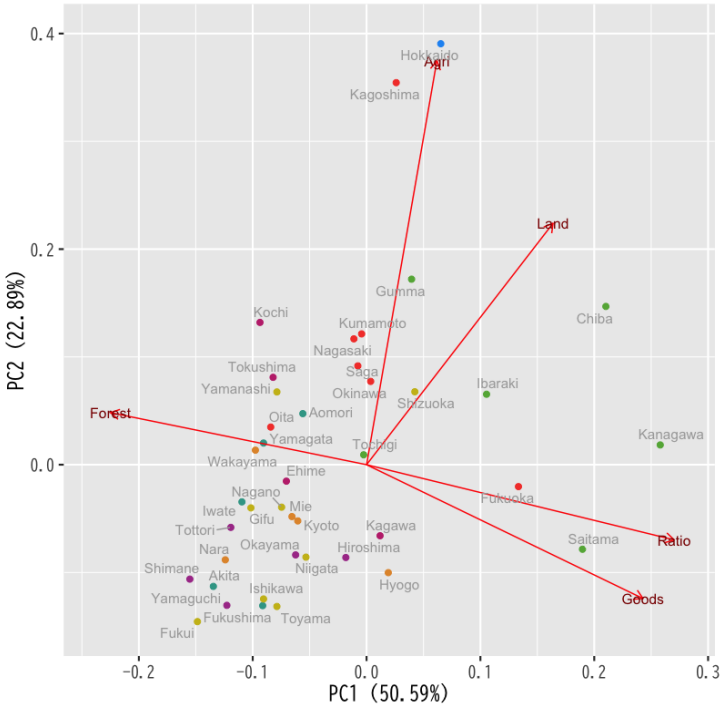

Figure 11: 主成分分析のバイプロット (第 1,2)

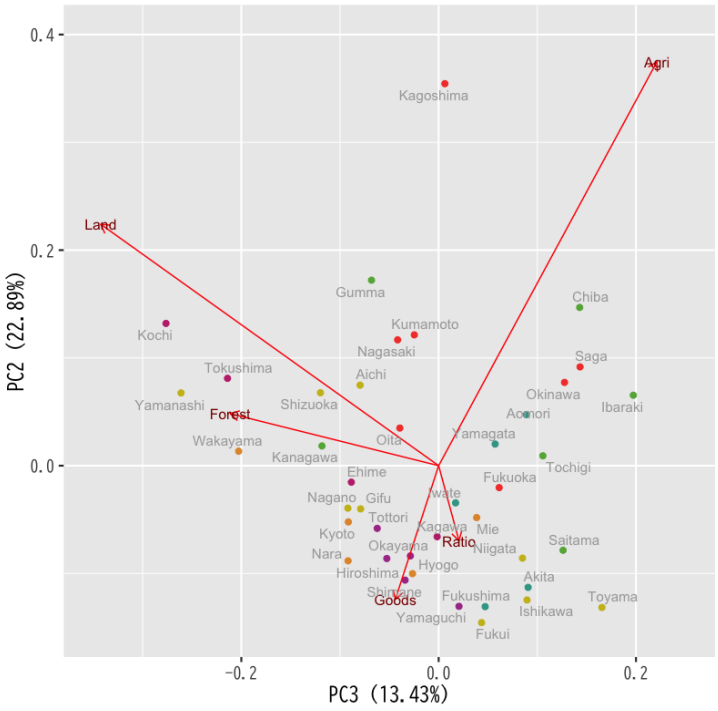

Figure 12: 主成分分析のバイプロット (第 3,2)

次回の予定

- **第 1 日 : 判別分析の考え方**
- 第 2 日 : 分析の評価