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 $\boldsymbol{a}^\top\boldsymbol{x}_1,\dots,\boldsymbol{a}^\top\boldsymbol{x}_n$ 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
  - $\boldsymbol{a}^\top\boldsymbol{x}_1,\dots,\boldsymbol{a}^\top\boldsymbol{x}_n$ 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

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

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 $X^\top X$ の二次形式

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

固有値問題

- 最適化問題

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

- 解の条件
  $f(\boldsymbol{a})$ の極大値を与える $\boldsymbol{a}$ は $X^\top X$ の固有ベクトルである
  $X^\top X \boldsymbol{a} = \lambda \boldsymbol{a}$
  - 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)
- $\boldsymbol{a}^\top \boldsymbol{x}_i$: **主成分得点** (principal component score)
- 第 1 主成分負荷量
  $X^\top X$ 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
  $X^\top X$ 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 $X^\top X$ の単位固有ベクトルとするとき

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

    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

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

    を X と $\boldsymbol{a}_k$ で表せ

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 $X^\top X$ (非負定値対称行列) のスペクトル分解

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

  - 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

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

  - 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

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

- 固有値による寄与率の表現

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

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
- データ (の一部) の内容

各変数の分布

- 変数間の散布図

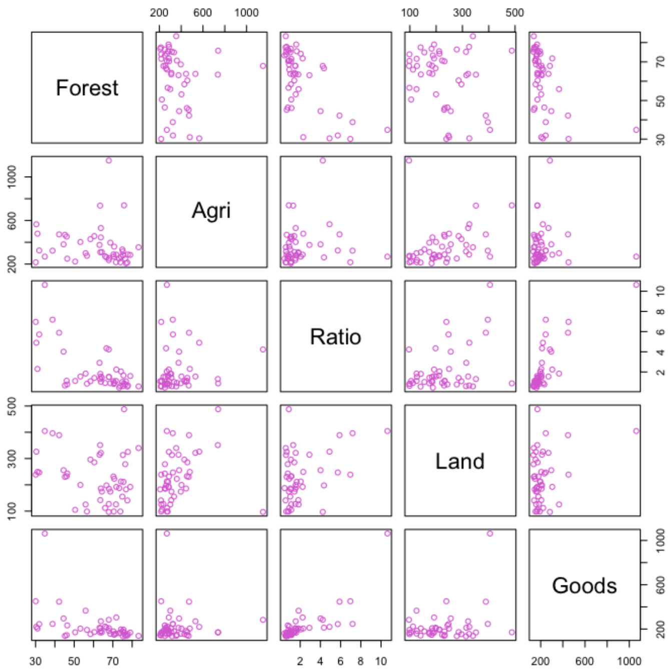

図 1: データの散布図

- 変数のばらつきに大きな違いがある

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)
  - 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
- 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)
  - 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)

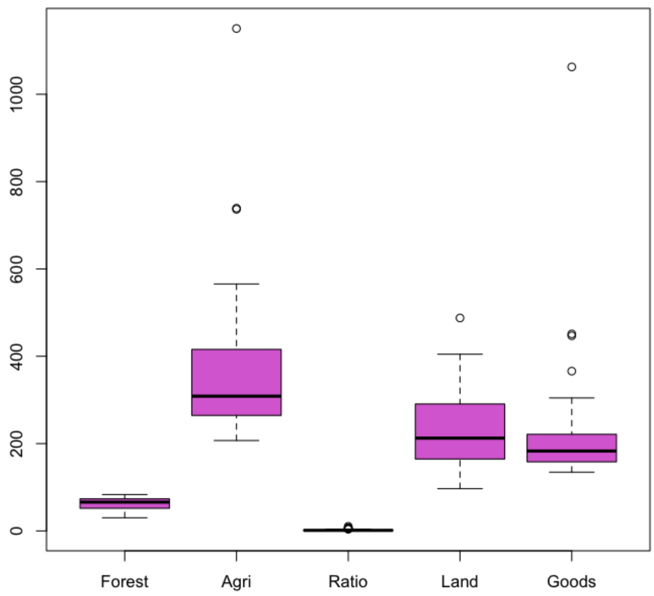

図 2: 各変数の箱ひげ図

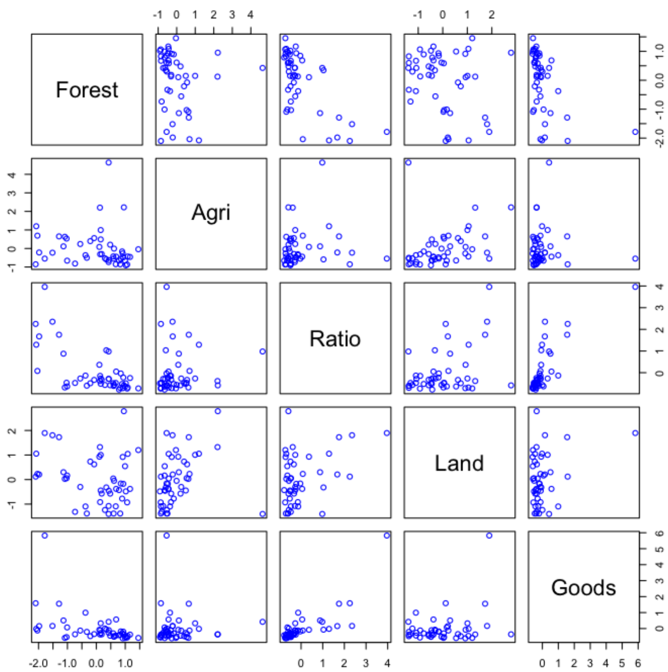

図 3: 正規化したデータの散布図

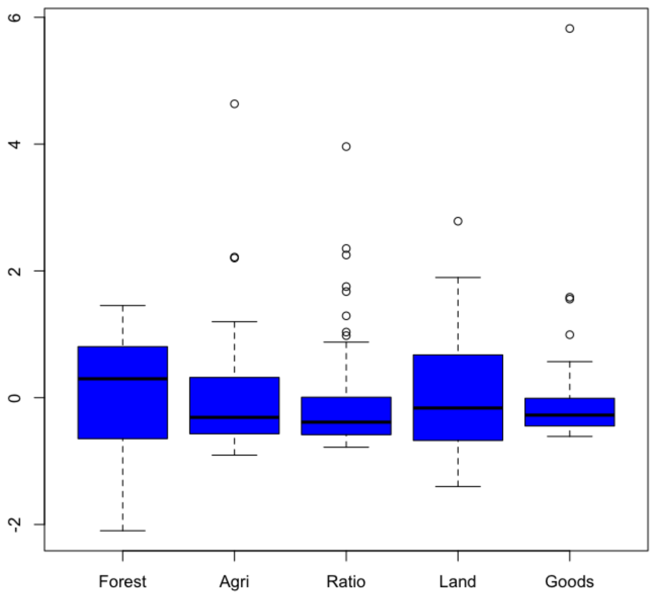

図 4: 各変数の箱ひげ図

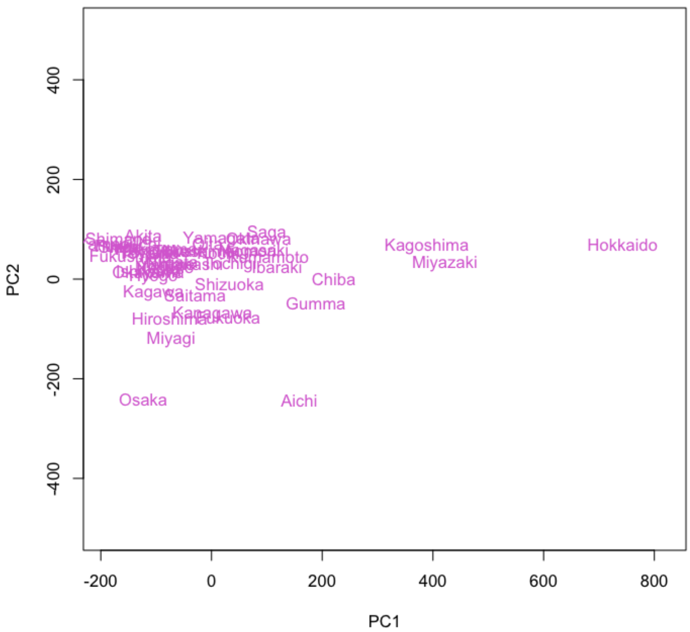

図 5: 主成分得点による散布図 (正規化なし)

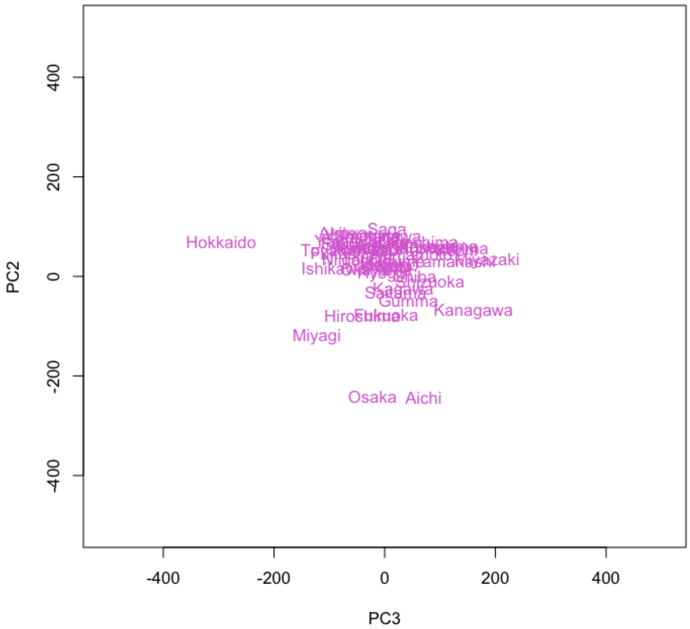

図 6: 主成分得点による散布図 (正規化なし)

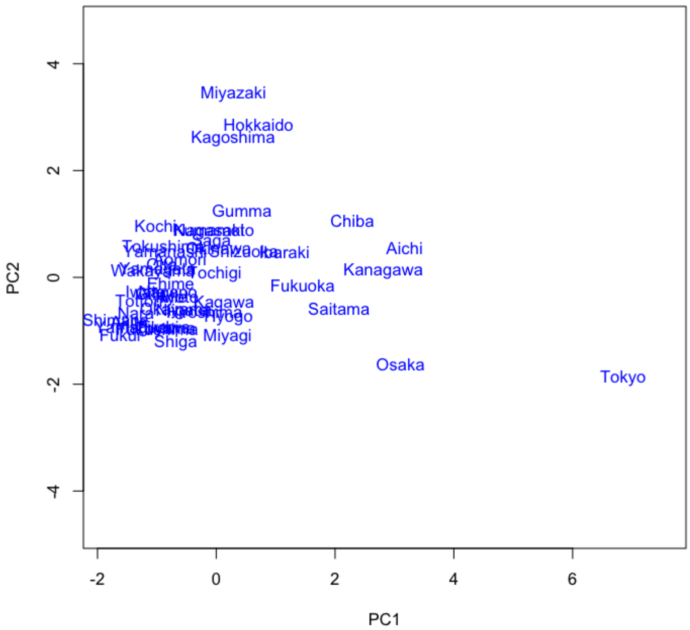

図 7: 主成分得点による散布図 (正規化あり)

- 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

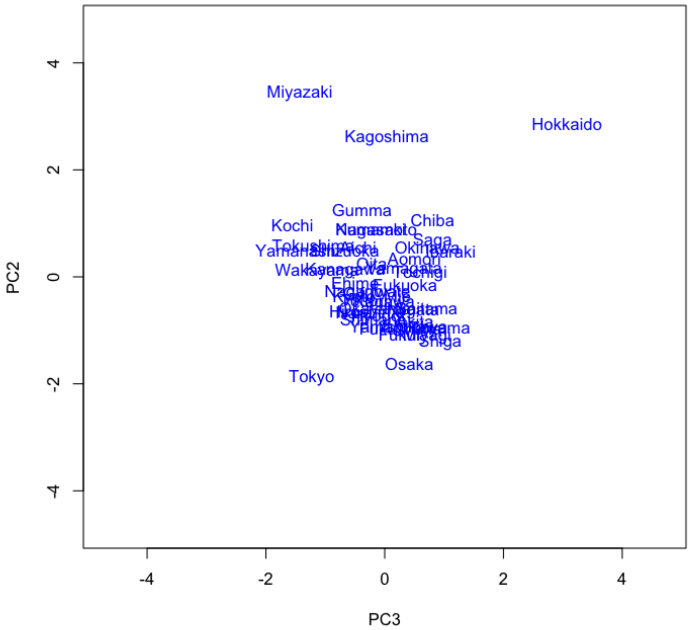

図 8: 主成分得点による散布図 (正規化あり)

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

$$X' = \begin{pmatrix} \boldsymbol{x}'^{\top}_1 \\ \vdots \\ \boldsymbol{x}'^{\top}_n \end{pmatrix} = \begin{pmatrix} x'_{11} & \cdots & x'_{1p} \\ \vdots & & \vdots \\ x'_{n1} & \cdots & x'_{np} \end{pmatrix}$$

  と書くとき，$X'^{\top}X'$ の対角成分を求めよ

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:

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

正規化データの場合

- $X^\top X$ の対角成分は全て $n-1$ ($(X^\top X)_{jj} = n-1$)
  - 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:

$$X = U\Sigma V^\top$$

  - 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

$$\Sigma = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & O_{r,p-r} \\ & & \sigma_r & \\ & O_{n-r,r} & & O_{n-r,m-r} \end{array}\right)$$

- 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

バイプロット

特異値と固有値の関係

- 行列 V の第 k 列ベクトル $\boldsymbol{v}_k$
- 特異値の平方

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

- Gram 行列の固有値問題

$$X^\top X \boldsymbol{v}_k = V\Sigma^\top \Sigma V^\top \boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k$$

  - $X^\top X$ の固有値は行列 X の特異値の平方
  - 固有ベクトルは行列 V の列ベクトル $\boldsymbol{a}_k = \boldsymbol{v}_k$

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)

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

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'

$$X \simeq X' = \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top + \sigma_l \boldsymbol{u}_l \boldsymbol{v}_l^\top$$

- 行列の積による表現

$$\begin{aligned} X' =& GH^\top, (0 \leq s \leq 1) \\ G =& \begin{pmatrix} \sigma_k^{1-s} \boldsymbol{u}_k & \sigma_l^{1-s} \boldsymbol{u}_l \end{pmatrix}, \quad H = \begin{pmatrix} \sigma_k^{s} \boldsymbol{v}_k & \sigma_l^{s} \boldsymbol{v}_l \end{pmatrix} \end{aligned}$$

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

$$X' = GH^\top$$

  - 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

- パラメタ s は 0, 1 または 1/2 が主に用いられる
- X の変動を最大限保持する近似は $k = 1, l = 2$

解析の事例

バイプロット

- 主成分負荷量
- 寄与率
- 第 1,2 主成分によるバイプロット

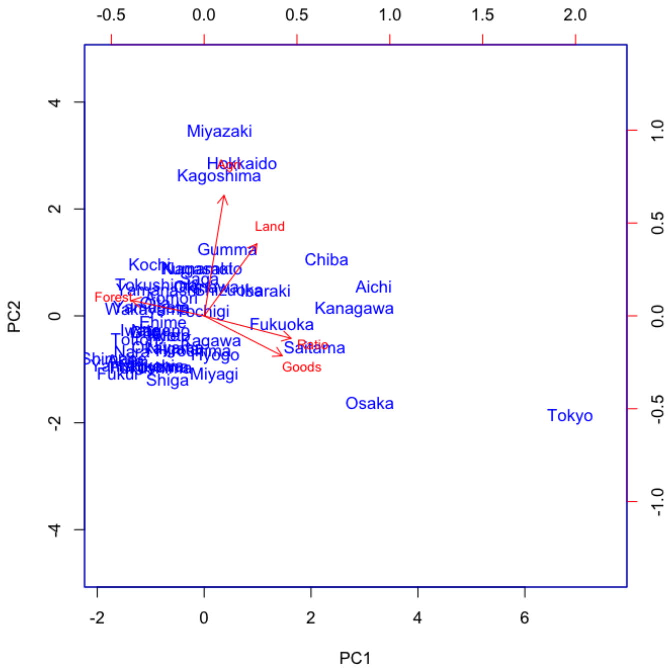

図 9: 主成分分析のバイプロット (第 1,2)

- 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)

次回の予定

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

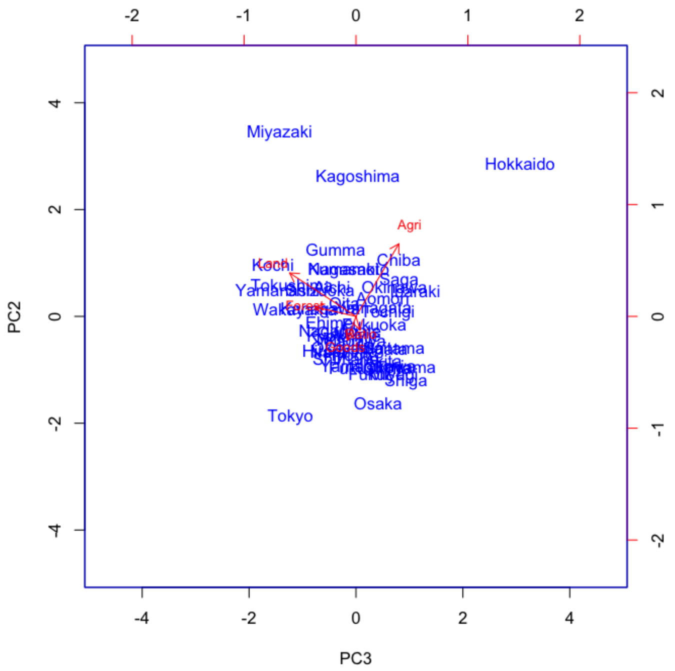

図 10: 主成分分析のバイプロット (第 3,2)

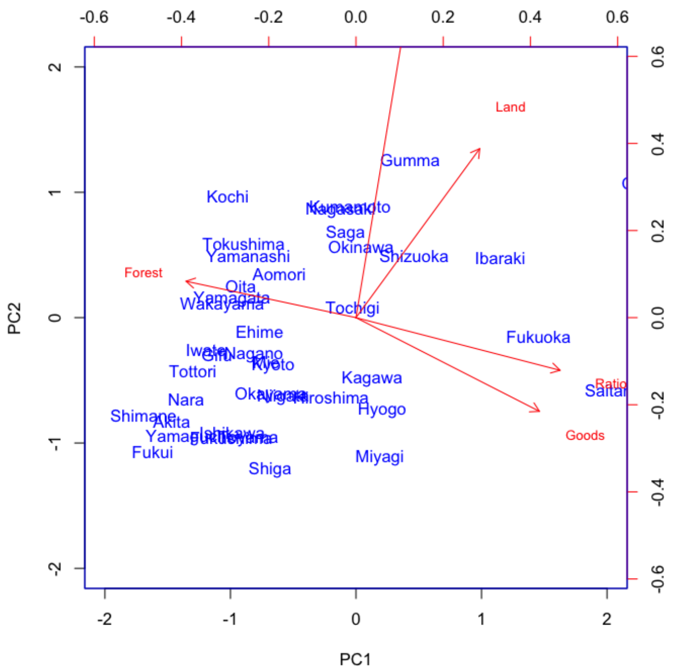

図 11: 主成分分析のバイプロット (第 1,2)

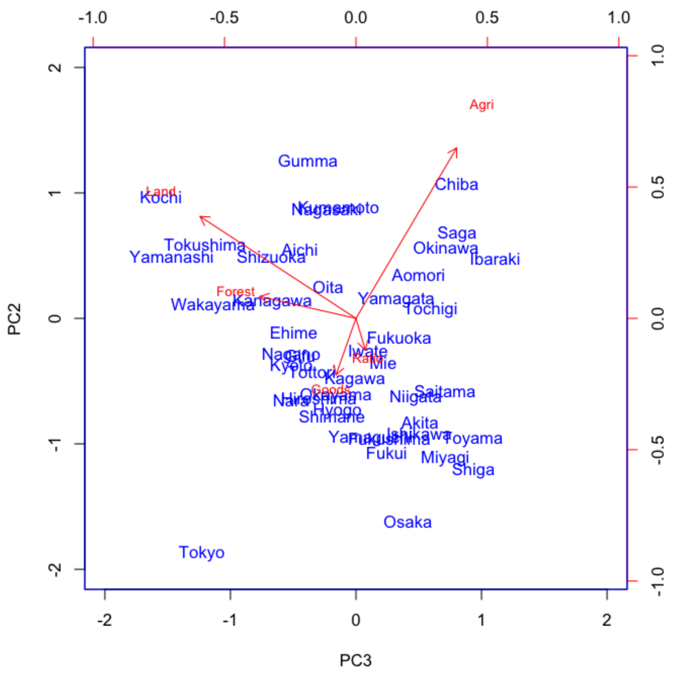

図 12: 主成分分析のバイプロット (第 3,2)