数値計算法・数値解析

固有値問題の応用例：主成分分析

宮崎大学 工学部

第 11 回

固有値問題の応用例：主成分分析

複数の特性値を総合評価する指標を作りたい

【例】

- 身長と体重から成長の度合いを総合評価したい
- 物理と数学のテストの点数から成績を総合評価したい
- 幅跳びとボール投げの結果から運動能力を総合評価したい

【主成分分析】

元のデータの持つ情報をより多く含む指標（合成変量）を作る

cf. B3「パターン認識」「データ解析」

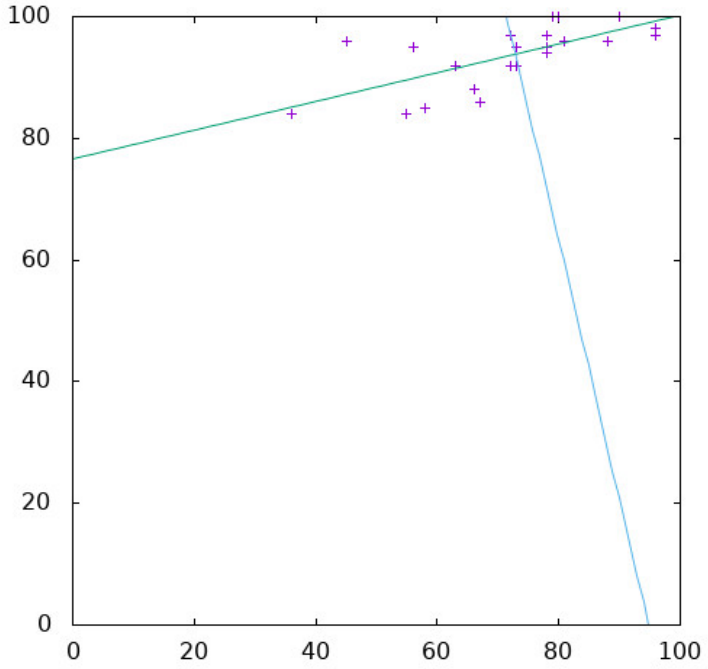
主成分分析の例

N 人の数学と英語のテストの点数 $\{(x_{1i}, x_{2i})|i = 1 \ldots N\}$ が与えられる．これを総合する指標を考える．

- 数学の点数と英語の点数を重み付けて足し合わせて，総合指標を作成．
- なるべく，元の情報を残す

- データの分散が最大となる指標を採用
- データとの二乗誤差の和が最小となる指標を採用

※ どちらも同じ結果になる



主成分分析の導出：分散最大化

新たな指標を $z = a_1 x_1 + a_2 x_2$ とする．

x_1, x_2：数学，英語の点数を表す変量

a_1, a_2：それぞれの重み（指標の単位方向ベクトル）

(1) z の分散

$$v = \frac{1}{N} \sum_i (z_i - \bar{z})^2$$

$$z_i = \sum_j a_j x_{ji}, \quad \bar{z} = \frac{1}{N} \sum_i z_i \quad j = 1, 2$$

(2) (a_1, a_2) は単位方向ベクトル

$$a_1^2 + a_2^2 = 1$$

(2) の条件下で (1) を最大化する

$\Rightarrow$ ラグランジュの未定乗数法（cf. B3「最適化理論」）

主成分分析の導出：分散最大化

以下の式を最大化すればよい

$$F = v - \lambda(a_1^2 + a_2^2 - 1)$$

極大値では，微分は 0

$$\begin{cases} \frac{\partial F}{\partial a_1} = 0 \\ \frac{\partial F}{\partial a_2} = 0 \end{cases}$$

数学，英語の平均点： $\bar{x_j} = \frac{1}{N}\sum_i x_{ji}$

数学，英語の分散共分散： $s_{kl} = \frac{1}{N}\sum_i (x_{ki} - \bar{x_k})(x_{li} - \bar{x_l})$

と置くと

$$v = a_1^2 s_{11} + a_1 a_2 s_{12} + a_1 a_2 s_{21} + a_2^2 s_{22}$$

$$\begin{cases} s_{11} a_1 + s_{12} a_2 = \lambda a_1 \\ s_{21} a_1 + s_{22} a_2 = \lambda a_2 \end{cases}$$

主成分分析の導出：分散最大化

$$\begin{cases} s_{11}a_1 + s_{12}a_2 = \lambda a_1 \\ s_{21}a_1 + s_{22}a_2 = \lambda a_2 \end{cases}$$

$$\begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

$$Va = \lambda a$$

固有値問題！！

主成分分析の導出：分散最大化

$$V = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}$$

は元のデータ $\{(x_{1i}, x_{2i})|i = 1 \ldots N\}$ の分散共分散行列

$$\begin{cases} s_{11}a_1 + s_{12}a_2 = \lambda a_1 \\ s_{21}a_1 + s_{22}a_2 = \lambda a_2 \end{cases}$$

$$\begin{cases} s_{11}a_1^2 + s_{12}a_2a_1 = \lambda a_1^2 \\ s_{21}a_1a_2 + s_{22}a_2^2 = \lambda a_2^2 \end{cases}$$

$$s_{11}a_1^2 + s_{12}a_1a_2 + s_{21}a_2a_1 + s_{22}a_2^2 = \lambda(a_1^2 + a_2^2) = \lambda$$

$$\lambda = v$$

固有値 λ が z の分散になっている．

主成分分析の手順

N 個のデータ（ベクトル）の集合

$$\{(x_{1i}, x_{2i})|i = 1 \ldots N\}$$

が与えられる．

1. データの分散共分散行列 V を求める
2. V の固有値・固有ベクトルを求める
3. 大きい方の固有値に対応する固有ベクトルが求めたい総合指標の単位方向ベクトル

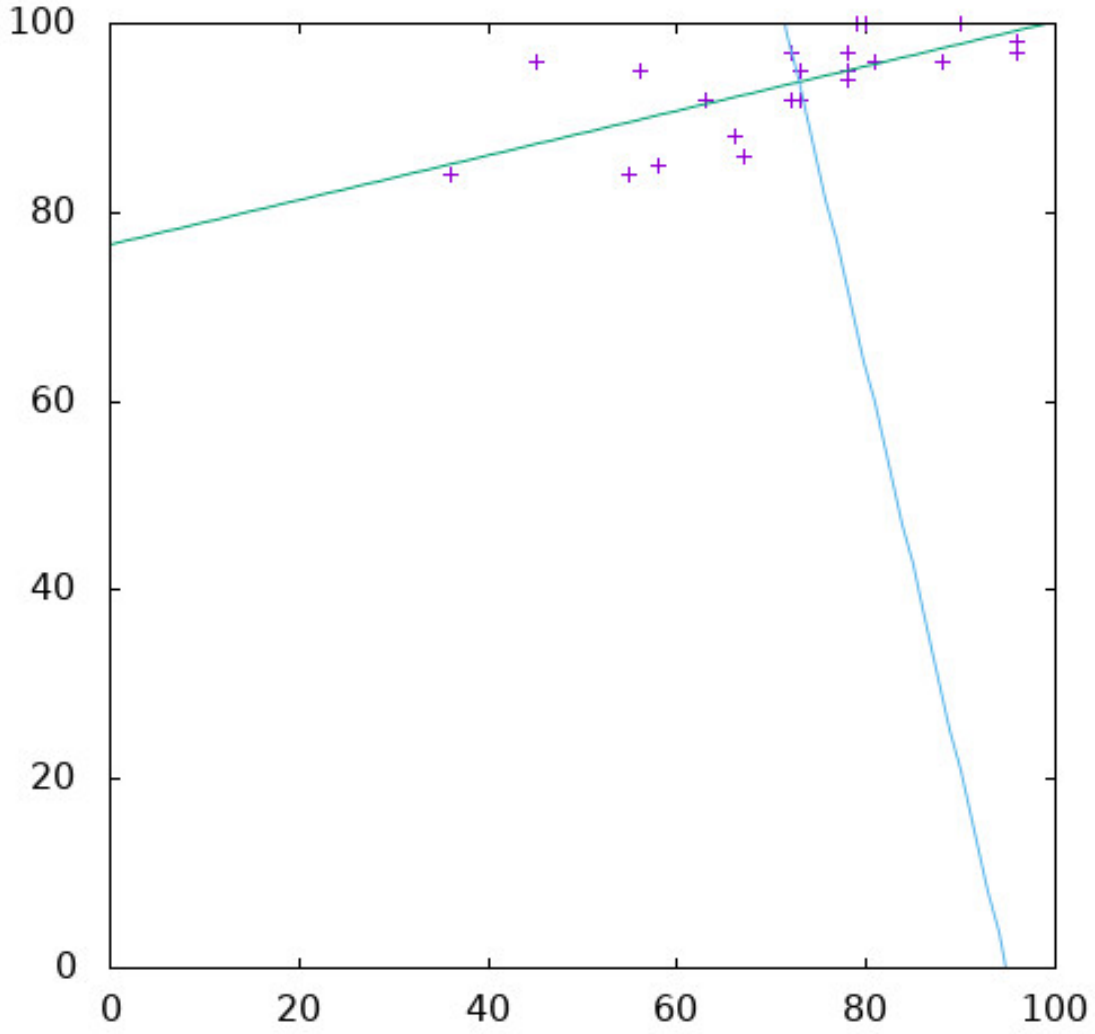
例題

以下の数学と英語の成績を代表する指標を主成分分析で求めよ

ID i	数学 x_1	英語 x_2	ID i	数学 x_1	英語 x_2
1	55	84	13	96	98
2	72	97	14	67	86
3	79	100	15	78	97
4	58	85	16	78	94
5	96	98	17	45	96
6	66	88	18	78	95
7	73	95	19	80	100
8	72	92	20	56	95
9	81	96	21	88	96
10	73	92	22	63	92
11	96	97	23	36	84
12	90	100			

例題：総合指標

$$z = \_\_\_\_\_(x_1 - \bar{x}_1) + \_\_\_\_\_(x_2 - \bar{x}_2)$$



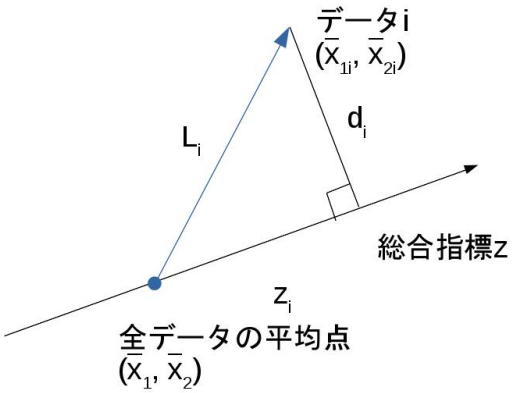
主成分分析での指標の基準

- データの分散が最大となる指標を採用
- データとの二乗誤差の和が最小となる指標を採用

※ どちらも同じ結果になる

$$L_i^2 = z_i^2 + d_i^2$$

$$\sum_i L_i^2 = \sum_i z_i^2 + \sum_i d_i^2$$

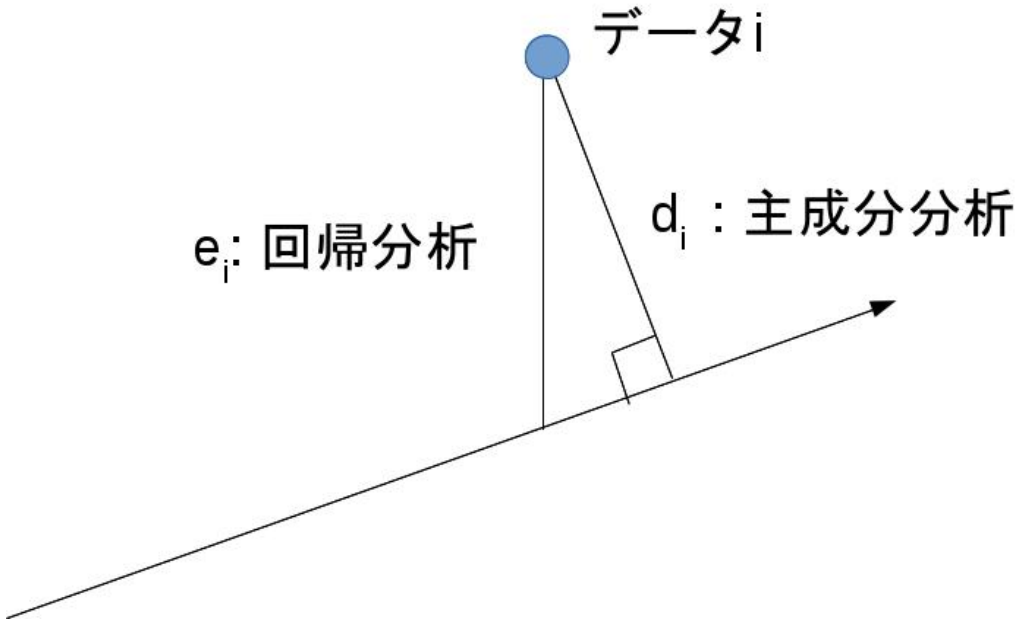


- データ集合が与えられると，$\sum_i L_i^2$ は一定
- 主成分分析は $\sum_i z_i^2$ を最大化
- この時 $\sum_i d_i^2$ は最小

$\sum_i d_i^2$: データと総合指標 z との二乗誤差の和

主成分分析と最小二乗法（回帰分析）

主成分分析：　点と直線の距離を最小化
最小二乗法：　説明変数と目的変数が存在
(回帰分析)　　目的変数と直線とのズレを最小化



p変量の主成分分析

特性値がp個の場合

- データはp次元ベクトル
- 上位k(≥ 1)個の主成分が考えられる

【例】

- ビッグデータからデータの主要な要素を取り出したい
- 多次元の観測データを２次元上に表示したい
- 多次元のデータを少数の次元で表現して圧縮したい

p変量の主成分分析

特性値がp個の場合

- データはp次元ベクトル
- 上位 $k(\geq 1)$ 個の主成分が考えられる

- 第１主成分
  分散が最も大きい方向の主成分
- 第２主成分
  第１主成分と無相関の（直交する）中で分散が最も大きい方向の主成分
- 第３主成分
  第１，２主成分と無相関の（直交する）中で分散が最も大きい方向の主成分

以下同様.

p変量の主成分分析の求め方

分散共分散行列 V は，実対称行列

- 固有ベクトルは互いに直交する
- 固有値がその方向での分散に対応

- 第１主成分
  最大固有値に対応する固有ベクトル
- 第２主成分
  ２番目に大きい固有値に対応する固有ベクトル
- 第３主成分
  ３番目に大きい固有値に対応する固有ベクトル

以下同様．

「べき乗法」で上位 k 個の固有値・固有ベクトルを求めればよい．

寄与率

第 k 主成分が，総合指標としてどの程度重要か？

- p 変量であれば，第 p 主成分まで考えれば元のデータを完全に再現できる
- 全データの分散は，各主成分の分散の和で評価できる

第 k 主成分の寄与率：$R_k = \lambda_k / \sum_j \lambda_j$

寄与率が小さい主成分は無視してもデータの誤差は小さい

主成分分析のプログラム

```
#define P         5
#define N         23
int main(int argc, char *argv[])
{ double      lambda;
  double      data[][P] = { ... };
  double      A[P][P];
  double      x[P] = {1, 0, 0, 0, 0};

  calc_cov(data, A);

  poweigen(A, x, &lambda);
  printf("***␣第１主成分␣:␣");
  print_vec(x);

  return 0;
}
```

主成分分析のプログラム

```
// 分散共分散行列を求める
void calc_cov(double data[N][P], double A[P][P])
{
  double      mean[P] = {0};
  for(int j = 0; j < P; j++) {
    for(int i = 0; i < N; i++) {
        mean[j] += data[i][j];
    }
    mean[j] /= N;
  }

  for(int k = 0; k < P; k++) {
    for(int l = 0; l < P; l++) {
      A[k][l] = 0.0;
      for(int i = 0; i < N; i++) {
        A[k][l] += (data[i][k] - mean[k]) * (data[i][l] - mean[l]);
      }
      A[k][l] /= N;
    }
  }
  return;
}
```