主成分分析 (PCA: Principal Component Analysis)

Ver.4

1. はじめに

英語と数学の試験の点数をもとに学生を成績のよい順に並べることを考えてみよう。その場合, 単純に 2 科目の平均点を用いる場合や, 英語に 0.7, 数学に 0.3 という重み (weight) をつけて合計する場合がある。単純に平均するのは, 0.5 と 0.5 の重みをつけたことに等しい。むずかしいのは, そうした重みを客観的に定めるにはどうすればよいかということである。

主成分分析(Principal Component Analysis)とは, いくつかの指標に重みをつけて総合指標を計算しなければならないとき, 総合指標の分散 (情報量) を最大化する」という考え方で重みを定める方法である。

2. 2 変数による主成分分析の例

2.1 データの準備

ここでは, 学生の成績の問題ではなく, 地域ごとの産業の集積度を決める問題を考えよう。使用するデータは, 次の産業データ"datB_jigyosho2.txt'"である。

表 2.1 都道府県別の産業別従業者数(2001 年, 事業所統計) datB_jigyosho2.txt

	x01 農林魚鉱	x02 インフラ	x03 製造	x04 流通	x05 小売飲食	x06 金融保険	x07 公的機関	x08 生活医療	x09 宿泊	x10 娯楽	x11 整備賃貸	x12 情報広告	x13 専門サ	x14 事業所サ	x15 教育学術	pop 人口
北海道	41859	300611	249095	331124	608941	111549	212308	239282	54048	47125	45020	28094	69277	137538	106621	5675309
青森県	8535	79361	77788	71715	148111	22034	60278	60659	11323	8969	10317	3698	12674	28697	28642	1497036
岩手県	10772	75648	118652	65428	130384	20280	47272	55194	13252	10050	9148	4950	12406	28643	26911	1421796
宮城県	8442	119617	156820	160994	265738	42396	67776	82966	18088	14377	19675	15675	33930	48908	50032	2347166
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	
:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	:	
佐賀県	3235	40356	68419	40539	88033	11828	32675	38635	6172	7435	5379	2010	8161	17363	17176	882639
長崎県	8990	65762	77518	71652	147187	21787	61378	69889	13716	10994	7998	4281	14637	25352	28991	1527398
熊本県	7579	75710	112343	83166	180593	26057	65008	88612	14597	12811	11927	7060	19126	30781	32509	1870416
大分県	5598	61607	77516	55921	125746	20214	44685	55276	12546	8504	8808	4326	13221	24406	23688	1234429
宮崎県	7821	55477	68885	50337	116893	17156	41941	51975	7408	8525	8810	3630	12715	20949	22149	1184535
鹿児島県	12088	80113	100303	83388	175090	24055	63891	81661	13764	11106	11479	4164	18852	28708	35987	1783231
沖縄県	2052	54413	30314	64207	140294	24225	53312	50814	14041	8739	10948	6591	20836	23002	28112	1334122

15 の産業があるが, それらを 6 つの産業にまとめ直したものを用いることにしよう。具体的には, 15 の産業のそれぞれの従業者数を 6 つの産業の従業者数にまとめて, さらに人口 10,000 人当りの産業別従業者数 (x1-x6) を求めておくことにする (表 2.2)。

表 2.2 6 つの産業分類

```
x1=(x05 + x08 + x09 + x10 + x15)/pop * 10000;          /* 1.対個人サ   */
x2=(x04 + x06 + x11 + x12 + x13 + x14)/pop * 10000;    /* 2.対事業所サ*/
x3=x03/pop * 10000;                                    /* 3.製造       */
x4=x07/pop * 10000;                                    /* 4.公的機関   */
x5=x02/pop * 10000;                                    /* 5.インフラ   */
x6=x01/pop * 10000;                                    /* 6.農林魚鉱   */
```

注)

x1　対個人サ　小売業, 飲食業。洗濯 ･理容 ･浴場業, 駐車場業, その他の生活
関連サービス業, 医療業, 保健衛生。旅館, その他の宿泊所。
娯楽業 (映画 ･ビデオ制作業を除く)。教育, 学術研究機関。
x2　対事業所サ　運輸業, 通信業, 卸売業。金融業, 保険業, 不動産業。
自動車整備業, 機械 ･家具等修理業, 物品賃貸業。
映画 ･ビデオ制作業, 放送業, 情報サービス ･調査 ･広告業。
専門サービス業 (他に分類なし)。
協同組合 (他に分類なし), その他の事業所サービス業, 廃棄物処理業。
x3　製造　製造業。
x4　公的機関　公務, 社会保険, 社会福祉, 宗教, 政治 ･経済 ･文化団体。
x5　インフラ　建設業, 電気 ･ガス ･水道 ･熱供給業。
x6　農林魚鉱　農業, 林業, 漁業, 鉱業。事業所の形態をとらない自営業者を除く。
鉱業には石油関連会社が含まれる。

課題09　2変数による主成分分析の準備
都道府県別の産業データを用いて，表2.2のように6つの産業分類で人口10000人当りの産業別従業者数を求めなさい。

1．使用するデータ: datB_popNen.txt

2．出力する項目
(1) 表のタイトル: 「表　人口10000人当りの産業別従業者数(6分類，2001年)」
(2) 6分類ごとの人口10000人当りの産業別従業者数(小数1桁まで，3桁区切りのカンマ不要)

3．提出物
次のものを電子メールの本文に書いて高辻あてに送信しなさい。
(1) 課題09，学籍番号，氏名，ユーザID
(2) SASプログラム
(3) 出力結果の表

4．提出先: tak@reitaku-u.ac.jp

5．件名: datB09 注）必ず半角とすること，大文字と小文字の区別をすること

6．提出物の例

課題09 学籍番号　氏名　ユーザID

SASプログラム

```
options  nocenter nodate  nonumber  linesize=100   pagesize=500;
data indus;
  infile 'f:\datB\datB_jigyosho2.txt';
  length  pref $ 8;
  input   pref $  x01-x15 pop;
        :
        :
proc print  .........
  title  ....
  var    ....
run;
```

表　人口10000人当りの産業別従業者数(6分類，2001年)

OBS	pref	x1	x2	x3	x4	x5	x6
1	北海道	1860.7	1273.2	438.9	374.1	529.7	73.8
2	青森県	1721.4	996.2	519.6	402.6	530.1	57.0
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:

以上

このように，通常はもとのデータをそのまま使用して分析を行うことは少ない。むしろ分析に適したようにもとの変数を加工することが多い。SAS プログラム言語は変数の加工に威力を発揮する。

2.2 2 変数の主成分分析

(1) 考え方

6 つの産業のうち，x_1（対個人サービス）とx_2（対事業所サービス）の 2 変数だけを用いて産業の集積度を表す指標を作成することを考えてみよう。どちらも値が大きければ産業の集積度が高いとみなせる。しかし，2 変数の値の組合せはさまざまであるから，そのままでは集積度の大小を直ちに判断できるとは限らない。集積度を表す何か 1 つの指標が必要である。これを総合指標と呼ぶことにしよう。問題は，x_1 とx_2 とからどのようにして 1 つの総合指標z を合成すればよいかということである。

主成分分析では，総合指標z は次のように x_1 とx_2 との線形結合で構成されると考える。

$$z = l_1 x_1 + l_2 x_2 \quad \text{.....(2.1)}$$

l_1，l_2 は未知の重み係数である。この係数を，次の 2 つの基準を満たすように求めるのが主成分分析である。

基準 A: 総合指標z の分散が最大になること.....(2.2)

基準 B: $l_1^2 + l_2^2 = 1$ であること.....(2.3)

分散に着目するのは，分散は個体を分類（区別）するための情報量を表していると考えられるからである。分散が大きいほど個体を分類しやすく情報量が大きいと考える。よって基準 A は，情報量を最大化する基準でもある。

こうして求めた総合指標z のことを主成分(Principal Component)という。

(2) 解き方

まず z の分散v_z は次のように表される。

$$\begin{aligned} v_z &= \frac{1}{n}\sum (z - \bar{z})^2 \\ &= \frac{1}{n}\sum \left((l_1 x_1 + l_2 x_2) - (l_1 \bar{x}_1 + l_2 \bar{x}_2)\right)^2 \\ &= \frac{1}{n}\sum \left(l_1 (x_1 - \bar{x}_1) + l_2 (x_2 - \bar{x}_2)\right)^2 \\ &= l_1^2 \frac{1}{n}\sum (x_1 - \bar{x}_1)^2 + l_2^2 \frac{1}{n}\sum (x_2 - \bar{x}_2)^2 + 2 l_1 l_2 \frac{1}{n}\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \end{aligned}$$

$$= l_1^2 v_1 + l_2^2 v_2 + 2 l_1 l_2 v_{12} \quad \text{.....(2.4)}$$

ここで, v_1 は x_1 の分散, v_2 は x_2 の分散, v_{12} は x_1 と x_2 の共分散で, 次のように表される。

$$v_1 = \frac{1}{n}\sum (x_1 - \bar{x}_1)^2 \ , \ v_2 = \frac{1}{n}\sum (x_2 - \bar{x}_2)^2 \ , \ v_{12} = \frac{1}{n}\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2) \quad \text{.....(2.5)}$$

z の分散 v_z (2.4)を $l_1^2 + l_2^2 = 1$ (2.3)の制約の下で最大化するような l_1, l_2 を求めるためラグランジェ関数を,

$$W = (l_1^2 v_1 + l_2^2 v_2 + 2 l_1 l_2 v_{12}) - l\,(l_1^2 + l_2^2 - 1)$$

とする。l はラグランジェ乗数である。ここから1 階の条件を求めると,

$$\frac{\partial W}{\partial l_1} = 2(v_1 - l) l_1 + 2 v_{12} l_2 = 0 \quad \text{.....(2.6)}$$

$$\frac{\partial W}{\partial l_2} = 2 v_{12} l_1 + 2(v_2 - l) l_2 = 0 \quad \text{.....(2.7)}$$

$$\frac{\partial W}{\partial l} = -\left(l_1^2 + l_2^2 - 1\right) = 0 \quad \text{.....(2.8)}$$

のように 3 つの条件式が得られる。

ここでまず, (2.6), (2.7)は次のように整理できる。

$$v_1 l_1 + v_{12} l_2 = l l_1 \quad \text{.....(2.9)}$$

$$v_{12} l_1 + v_1 l_2 = l l_2 \quad \text{.....(2.10)}$$

行列を用いて書き改めると

$$\begin{pmatrix} v_1 & v_{12} \\ v_{12} & v_2 \end{pmatrix} \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} = l \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \quad \text{.....(2.11)}$$

と表される。さらに,

$$\boldsymbol{V} = \begin{pmatrix} v_1 & v_{12} \\ v_{12} & v_2 \end{pmatrix}, \ \boldsymbol{l} = \begin{pmatrix} l_1 \\ l_2 \end{pmatrix} \quad \text{.....(2.12)}$$

とおけば,

$$\boldsymbol{V}\boldsymbol{l} = l\boldsymbol{l} \quad \text{.....(2.13)}$$

である。これは数学的には次のことを意味している。

① 行列 $\boldsymbol{V}$ は x_1 と x_2 との分散共分散行列と呼ばれる。

② l は行列 $\boldsymbol{V}$ の固有値である。一般に行列の次数だけ解がある。この場合は 2 つ。

③ ベクトル $\boldsymbol{l}$ は一つの l に対応する固有ベクトルである。

④ 条件(2.8)を加えれば, 一つの l に対応する固有ベクトル $\boldsymbol{l}$ はただ一つに定まる。

さて, ベクトル $\boldsymbol{l}$ は重み係数を表しているが, 固有値 l は何を表すのだろうか。それは(2.13)式の両辺に左から行ベクトル $\boldsymbol{l}' = (l_1, l_2)$ を掛けることで分る。

$$\boldsymbol{l}'\boldsymbol{V}\boldsymbol{l} = \boldsymbol{l}'\, l\boldsymbol{l} \quad \text{.....(2.14)}$$

$$\text{左辺} = \boldsymbol{l}'\boldsymbol{V}\boldsymbol{l} = l_1^2 v_1 + l_2^2 v_2 + 2 l_1 l_2 v_{12} = v_z \text{ (主成分 } z \text{ の分散)} \quad \text{.....(2.15)}$$

$$\text{右辺} = \boldsymbol{l}'\, l\boldsymbol{l} = l\boldsymbol{l}'\boldsymbol{l} = l \quad (\because \boldsymbol{l}'\boldsymbol{l} = 1) \quad \text{.....(2.16)}$$

$$\because \quad l = v_z \quad \text{.....(2.17)}$$

つまり, 固有値 l は主成分 z の分散 v_z を表している。

以上のことを整理すると次のことがいえる。

(3) 解き方のまとめ

① 変数 x_1 と x_2 とによる主成分は,

$$z = l_1 x_1 + l_2 x_2 \quad \text{.....(2.1)}$$

と表される。

② 重み係数 l_1, l_2 は次の基準によって求められる。

基準 A: 総合指標(主成分) z の分散が最大になること.....(2.2)

基準 B: $l_1^2 + l_2^2 = 1$ であること.....(2.3)

③ これを解くには, まず x_1 と x_2 との分散共分散行列を計算して $\boldsymbol{V}$ とする。

④ 行列 $\boldsymbol{V}$ の固有値 l は, 主成分 z の分散 v_z に等しい。

⑤ よって基準 A を満たすには, そのうちの「最大の固有値 l 」を採択すればよい。

⑥ 最大の固有値 l に対応する固有ベクトル $\boldsymbol{l}$ で, かつ基準 B ($l_1^2 + l_2^2 = 1$) を満たすものが重み係数である。この重み係数で合成される主成分を第 1 主成分という。

⑦ (2.1)式により個体ごとの総合指標の値(主成分スコアという)を計算する。

⑧ なお, 数学的に固有値 l は, 行列 $\boldsymbol{V}$ の次数(もとの変数の数)だけある。この例では 2 つである。大きい順に l_1, l_2 のように表す。もし最大の固有値 l_1 の値(主成分の分散つまり情報量)が十分でなければ, 2 番目の大きさの固有値 l_2 も採択する。それぞれについて異なる固有ベクトル(重み係数)が定まる。それらで合成される主成分を, 第1主成分, 第2主成分と呼ぶ。その場合, 総合指標とはいってもただ一つに定めることはできない。

これらの計算を実行するには, SAS の princomp プロシジャを使用する。

(4) princomp による実行

SAS を用いて上の 2 変数の主成分分析を実行してみよう。SAS プログラムは表 2.3 のようになる。またその実行結果は表 2.4 のようになる。

表 2.3 SAS プログラム： 2 変数による主成分分析 (分散共分散行列による)

```
/*********************************************************************
#                                                                    #
#  datB_prin01.sas: 2変数による主成分分析 ( 分散共分散行列による )    #
#                                                                    #
*********************************************************************/
options  nocenter nodate  nonumber  linesize=100   pagesize=500;

/*---------------------------------+
| ... データセットindusを準備する   |
+---------------------------------*/
data indus;
  infile 'f:\datB\datB_jigyosho2.txt';
  length  pref $ 8;
  input   pref $  x01-x15 pop;
  x1=(x05 + x08 + x09 + x10 + x15)/pop * 10000;          /* 1.対個人サ  */
  x2=(x04 + x06 + x11 + x12 + x13 + x14)/pop * 10000;    /* 2.対事業所サ*/
  x3=x03/pop * 10000;                                    /* 3.製造      */
  x4=x07/pop * 10000;                                    /* 4.公的機関  */
  x5=x02/pop * 10000;                                    /* 5.インフラ  */
  x6=x01/pop * 10000;                                    /* 6.農林魚鉱  */
  label
  x1='1.対個人サ '
  x2='2.対事業サ '
  x3='3.製造     '
  x4='4.公的機関 '
  x5='5.インフラ '
  x6='6.農林魚鉱 ';
run;

/*---------------------------------+
| ... 2変数の主成分分析を行う      |
+---------------------------------*/
proc princomp  data=indus  cov out=prin;
  title    ' ';
  title2   '表  2変数による主成分分析 ( 分散共分散行列による )';
  title3   '      (1) 対個人サービス × 対事業所サービス';
  title4   '      (2) 人口10000人当り従業者数による(2001年)';
  var    x1 x2;
run;

/*---------------------------------+
| ... 主成分スコアをプリントする   |
+---------------------------------*/
proc print data=prin;
  title   ' ';
  title2  '表  主成分スコア';
  format prin1-prin2 8.3;
  var pref prin1-prin2;
run;
```

データセットindusを作成する。

1. princompプロシジャ(主成分分析)を実行する。
2. data=indus: データセットindusを使用する。
3. cov: 分散共分散行列から固有値を求める。相関行列から固有値を求める場合は，covオプションをつけない。後の解説を参照。
4. out=prin: 分析結果の主成分スコアをデータセットprinに出力する。第1主成分，第2主成分には自動的にprin1，prin2という変数名がつく。

変数x1とx2の2変数を用いて主成分分析を行うことを指定する

1. printプロシジャを実行する。
2. data=prin: データセットprinを使用する。主成分スコアが記録されている。

出力のフォーマットを8.3に設定

第1主成分スコア(prin1)，第2主成分スコア(prin2)をプリントするように指定。

① princomp プロシジャで主成分分析を実行する。

② data=入力データセット名： 分析に用いるSAS データセット名を指定する。ここではindus。

③ cov： 分散共分散行列から固有値を求めるときに指定する。相関行列から固有値を求める場合は, cov オプションをつけない。後の解説を参照。

④ out=出力データセット名: 分析結果の主成分スコアを出力するSAS データセット名を指定する。ここでは prin。第 1 主成分, 第 2 主成分には prin1, prin2 という変数名がつく。

⑤ var： 主成分分析に用いる変数名を指定する。ここでは, x1 と x2。

表 2.4 SAS 実行結果：2 変数による主成分分析 (分散共分散行列による)

```
表  2変数による主成分分析 ( 分散共分散行列による )
      (1) 対個人サービス × 対事業所サービス
      (2) 人口10000人当り従業者数による(2001年)

The PRINCOMP Procedure

Observations          47        <--- 観測値の件数
Variables              2        <--- 使用した変数の数

           Simple Statistics                    <--- 基礎統計
                   x1                  x2       <--- 使用した変数
Mean        1768.121330         1115.747001     <--- 平均値
StD          139.543957          358.192036     <--- 標準偏差

                 Covariance Matrix              <--- 分散共分散行列

                                  x1                  x2
x1      1.対個人サ          19472.5161          40034.0280
x2      2.対事業サ          40034.0280         128301.5345

Total Variance      147774.05052     <--- x1の分散とx2の分散との合計(全分散): v1+v2

          Eigenvalues of the Covariance Matrix  <--- 分散共分散行列の固有値

        Eigenvalue     Difference     Proportion     Cumulative
          固有値          差            寄与率         累積寄与率
   1    141441.908     135109.766        0.9571         0.9571
   2      6332.142                       0.0429         1.0000

                Eigenvectors      <--- 固有ベクトル

                                  Prin1          Prin2     <--- 主成分1,2の変数名 (自動的にPrinが付く)

x1      1.対個人サ              0.311861       0.950128
x2      2.対事業サ              0.950128       -.311861

表  主成分スコア

OBS      pref          Prin1         Prin2

  1     北海道         178.515        38.867
  2     青森県        -128.145        -7.083
  3     岩手県        -153.043       -65.245
  :       :              :             :
  :       :              :             :
  :       :              :             :
 45     宮崎県        -155.485        28.909
 46     鹿児島県      -146.837        61.837
 47     沖縄県          21.084        41.290
```

x1の分散(v1)

x1とx2の共分散(v12)

x1とx2の共分散(v12)

x2の分散(v2)

第1番目に大きい固有値λ1(第1主成分の分散)
第2番目に大きい固有値λ2(第2主成分の分散)
(注) 固有値の合計 = 全分散 = v1 + v2
(もとの変数の分散の合計)

全分散に占める第1主成分の割合
= λ1 / (v1 + v2)
全分散に占める第2主成分の割合
= λ2 / (v1 + v2)

第2番目の固有値に対応する固有ベクトル
(第2主成分を求めるための重み係数)

第1番目の固有値に対応する固有ベクトル
(第1主成分を求めるための重み係数)

本文中の(2.1)式によって求めた第1主成分のスコア(値)と第2主成分のスコア。ただし，それぞれの平均がゼロになるようにSASが自動的に基準化している。

(5) 分析結果の読み方

① Observations：観測値の件数。予定した通りの件数が入力されたかチェックする。

② Variables： 使用した変数の数。

③ Simple Statistics： 基礎統計。使用した変数の平均値（Maen）と標準偏差（STD）。

④ Covariance Matrix： 分散共分散行列。

$$\boldsymbol{V} = \begin{pmatrix} v_1 & v_{12} \\ v_{12} & v_2 \end{pmatrix}$$

⑤ Total Variance： x_1 の分散と x_2 の分散との合計（全分散）$= v_1 + v_2$。分散は情報量を表すから，全分散とは使用したすべての変数が持っている総情報量を意味する。このうちできるだけ多くの情報量を，数少ない指標にまとめなおそうとする（合成する）のが主成分分析である。この例では，もとの変数の数が 2 つであるから，それより少ない 1 つの指標にまとまるのが望ましい。

⑥ Eigenvalue： 行列 $\boldsymbol{V}$ の固有値。(2.13)式の l に当る。これは主成分の分散 v_z に等しい（2.17 式）。大きい順にたてにプリントされる。例えば第 1 主成分の分散（固有値）は，141441.9 である。数学的に固有値は，もとの変数の数（行列 $\boldsymbol{V}$ の次数）だけある。ここでは 2 つである。どこまでを総合指標として採択するかは次の寄与率を参考に決める。

⑦ なお，すべての固有値の合計は，もとの変数の全分散に等しい。

$$l_1 + l_2 = v_1 + v_2 \quad \text{.....(2.18)}$$

⑧ Proportion： 寄与率。それぞれの主成分の分散が，全分散の何割を占めるかを表す。

第 1 主成分の寄与率＝ $l_1 /(v_1 + v_2)$ .....(2.19)

第 2 主成分の寄与率＝ $l_2 /(v_1 + v_2)$ .....(2.20)

一般にこれに 100%を掛けて%表示で表す。

⑨ 寄与率が大きいほど，その主成分が多くの情報量を集めて合成されたことを意味する。ここでは，第 1 主成分が 95.7%の寄与率を持っている。よって，もとの 2 変数を用いたときと比べて，1 変数だけでもとの 2 変数の 95.7%の情報量を持っていることになる。

⑩ Cumulative： 累積寄与率。寄与率を上から下へ順に累計した値。何番目の主成分までを採択すればよいかを判断するときに用いる。だいたい 70% くらいまでを採択する。この場合，第 1 主成分だけで 70%を超えているのでそれで十分であろう。

⑪ もし，第 1 主成分の寄与率が 60%，第 2 主成分の寄与率が 40%，という場合は，第 1 主成分だけで総合指標にするわけにはいかない。第 1 と第 2 の二つの主成分を用いなければならないだろう。これは変数の数の節約にはならない。ただ，違った視点から解釈することはできる。後掲。

⑫ Eigenvectors： 固有ベクトル（重み係数）。$z = l_1 x_1 + l_2 x_2$ における l_1，l_2 を表す。この場合，第 1 主成分については $l_1 = 0.3119$，$l_2 = 0.9501$ である。よって，第 1 主成分を z_1 とすると，

$$z_1 = 0.3119x_1 + 0.9501x_2 \quad \text{.....(2.21)}$$

である。この式で計算されたzの値を主成分スコアという。

⑬ 主成分スコアは，第1主成分と第2主成分のそれぞれについて計算される。それぞれに変数名として，Prin1，Prin2 が自動的につく。

⑭ また主成分スコアは，個体ごとに計算される。例えば，第 1 主成分について，

北海道の第 1 主成分スコア＝　178.5

青森県の第 1 主成分スコア＝－128.1

・・・・

のようになる。

⑮ ここでは，第 1 主成分z_1は寄与率が 95.7%と大きいので，もとの 2 変数を合成した総合指標だとみなせる。つまり対個人サービスと対事業所サービスの両方の集積度を総合して表していると考えられる。z_1の値が大きいほど集積度は高く，小さいほど集積度は低いことになる。

⑯ なお SAS では，主成分スコアは自動的に「平均＝ゼロ」になるように調整された結果がプリントされる。よってプラスとマイナスの値が出てくる。

課題10　2変数の主成分分析の結果の読み取り
**　表2.3のSASプログラムを実行しなさい。その結果をもとに以下の文の(1)～(5)の個所に適切な数値または語句を解答しなさい。**

**　x2(対事業所サービス)の平均値は(1)である。x1(対個人サービス)の標準偏差は(2)である。第2主成分の寄与率は(3)%である。第1主成分の重み係数について，「対個人サービス」と「対事業所サービス」とのどちらが大きいかといえば，(4)の方が大きい。東京都の第1主成分のスコアは(5)である。**

1．提出先：tak@reitaku-u.ac.jp

2．件名：datB10 注）必ず半角とすること，大文字と小文字の区別をすること

3．提出物の例

```
+-------------------------------------------------------------------------+
|  課題10 学籍番号　氏名　ユーザID                                        |
|                                                                         |
|  (1) xxxx.x（小数2桁目を四捨五入して小数1桁まで）                       |
|  (2) xxxx.x（小数2桁目を四捨五入して小数1桁まで）                       |
|  (3) x.xx（小数3桁目を四捨五入して小数2桁まで）                         |
|  (4)                                                                    |
|  (5) xxxx.x　（小数2桁目を四捨五入して小数1桁まで）                     |
+-------------------------------------------------------------------------+
```

以上

3. 主成分分析の幾何学的な意味

これまでの 2 変数の分析を例に, 主成分分析を幾何学的な観点から理解しておこう。図 3.1 は, x_1 を横軸, x_2 を縦軸にとってそれぞれの地域の散布状況を見たものである。直観的に 2 つの変数の間の相関が強いことがわかる。つまり, x_1 の値が大きければ (小さければ) x_2 の値も大きい (小さい)。こうした場合, 主成分とは散布している地域全体の「中心を貫くような直線」として表される (図 3.2)。言い換えると, 主成分は「散布しているどの地域 (点) からも最も近い距離になるような直線」として求められる。以下その点を詳しく見ていこう。

図 3.1 産業集積でみた地域の散布図

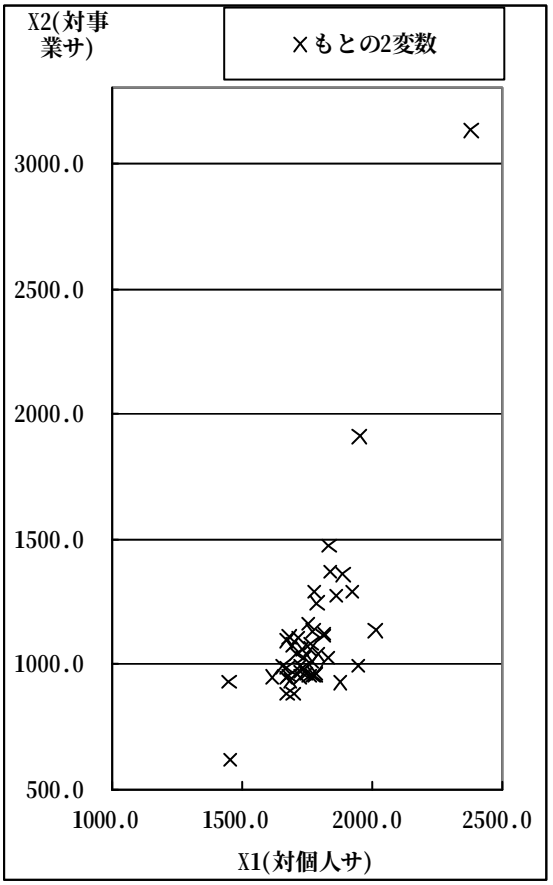
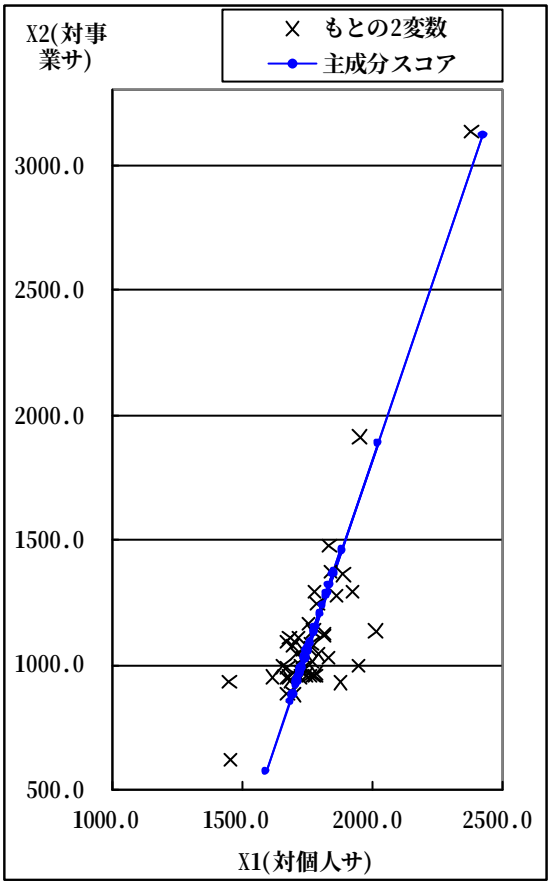
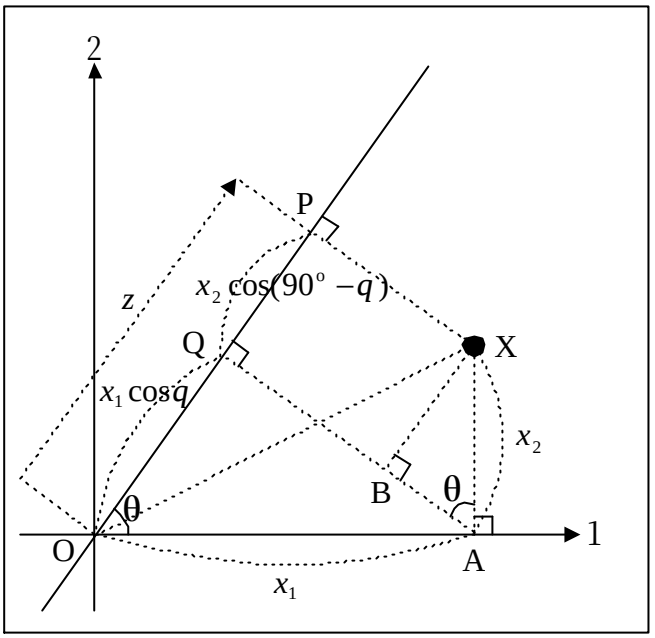


図 3.2 主成分の位置づけ



改めて横軸に x_1, 縦軸に x_2 をとった座標平面上に一つの地域 X がプロットされているとする (図 3.3)。説明のため地域 X の座標 (x_1, x_2) は, 地域全体の平均値からの偏差を表しているものとする。言い換えれば, 図 3.2 の縦軸と横軸とを地域全体の平均値の位置 (重心) に平行移動して描き直したものが図 3.3 である。原点 O は地域全体の重心に当る。

図 3.3　もとの 2 変数と主成分との関係



3.1　主成分が求められたときの解釈

いま変数x_1とx_2とによる主成分が,

$$z = l_1 x_1 + l_2 x_2 \quad \text{.....(2.1)}$$

と求められたとする。このとき主成分の直線（主成分軸）と横軸とがなす角度をqとすると, 重み係数l_1, l_2は次のような値をとる。

$$l_1 = \cos q\,,\ l_2 = \cos(90^\circ - q) = \sin q \quad \text{.....(3.1)}$$

ここから主成分スコアを表す(2.1)式は次のように書き換えられる。

$$\begin{aligned} z &= x_1 \cos q + x_2 \cos(90^\circ - q) \\ &= \overline{\mathrm{OQ}} + \overline{\mathrm{QP}} \\ &= \overline{\mathrm{OP}} \end{aligned} \quad \text{.....(3.2)}$$

すなわち,

「点X から主成分軸に下ろした垂線の足をP とすると, $z = \overline{\mathrm{OP}}$ である。」.....(3.3)

（点X を主成分軸の上に投影した点をP とすれば$z = \overline{\mathrm{OP}}$ である）

3.2　主成分の求め方の幾何学的解釈（垂線の二乗和の最小化）

逆に, 幾何学的に主成分軸を求める場合は次のように考える。

「基準 C'： 散布しているどの点からも最も近い距離になるような直線」.....(3.4)

より厳密には,

「基準 C： 各点からの垂線の長さの二乗和が最小になるような直線」.....(3.5)

を求めるという考え方である。

この基準から次のようにして主成分を求めることができる。いま図 3.3 の主成分軸が未知で，それが横軸となす角度をq とする。点X から主成分軸に下ろした垂線の足をP とする。また点X から横軸に下ろした垂線の足をA とする。ここで⊿OXP と⊿OXA とに注目すると，ともに直角三角形であるからピタゴラスの定理により，

$$\overline{\mathrm{OX}}^2 = \overline{\mathrm{XP}}^2 + \overline{\mathrm{OP}}^2 = \overline{\mathrm{XP}}^2 + z^2 \quad \text{.....(3.6)}$$

$$\overline{\mathrm{OX}}^2 = \overline{\mathrm{XA}}^2 + \overline{\mathrm{OA}}^2 = x_1^2 + x_2^2 \quad \text{.....(3.7)}$$

である。これにより，垂線 XP の長さの二乗は

$$\overline{\mathrm{XP}}^2 = x_1^2 + x_2^2 - z^2 \quad \text{.....(3.8)}$$

と表される。この関係はすべての点（地域）について成立する。全部で n 個の点があるとして，この関係式をすべての点について合計すると，

$$\sum \overline{\mathrm{XP}}^2 = \sum x_1^2 + \sum x_2^2 - \sum z^2 \quad \text{.....(3.9)}$$

である。左辺は「各点からの垂線の長さの二乗和（これを d とおく）」，右辺の第 1 項は変数 x_1 の分散 v_1 の n 倍，第 2 項は変数 x_2 の分散 v_2 の n 倍，第 3 項は主成分 z の分散 v_z の n 倍である。よって上の式は，

$$d = n(v_1 + v_2 - v_z) \quad \text{.....(3.10)}$$

と表される。この式において $v_1 + v_2$（全分散）は一定であるから，結局 d を最小化することは，v_z を最大化することと同値である。すなわち，

「基準 C： 各点からの垂線の長さの二乗和（d）を最小にする」

⇔ 「基準 A： 主成分 z の分散（v_z）を最大化する」 .....(3.11)

が言える。よって，いずれの基準を用いても同じ主成分の解が得られる。逆に，基準 A と基準 C とは主成分の持つ 2 つの特徴を表している。

4. 相関行列を用いた主成分分析

4.1 基準化

体重（kg）と身長（cm）という2 つの指標から，体の大きさを表す総合指標を作成することを考えてみよう。これまでどおりの方法で主成分を求めてもよいだろうか。問題は次の点にある。

① 体重をkg でなくg で表示すると体重の分散は大きくなる（10^6 倍になる）。身長をcm でなくm で表示すれと身長の分散は小さくなる（1/10000 になる）。この 2 つの指標を使って主成分分析を行うと，分散の大きい変数の重み係数が大きくなる。つまり，変数の単位の取り方によって解が変わってくることになる。これは具合が悪い。

② では，先に行った2つの産業の集積度の総合指標を求める分析には問題はないのだろうか。2 つの変数とも単位は「従業者数/人口 10000 人」であった。しかし，かといって同じ単位の下であれば，対個人サービスが100 変化することと対事業所サービスが100 変化することとは同じ変化の大きさだと言えるのだろうか。表 2.4 によれば，対事業所サービスの分散の方が対個人サービスよりも大きい。ということは，同じ100 の変化であっても，対事業所サービスより対個人サービスの方が，より起こりにくい変化が起こったことにならないか。つまり変化を表す 1 単位が異なるのではないか。

このように，異なる単位（Kg, cm, 人など）の変数を用いる場合，あるいは異なる大きさの分散の変数を用いる場合，これまでどおりの方法で主成分を求めてもよいのかという疑問が生ずる。いずれも，

「変化の大きさを表す 1 単位の大きさ（分散または標準偏差）が
ふぞろいのままでは，分散の大きな変数が重み係数の決定に
大きく影響することになる」 .....(4.1)

からである。そこでふつう主成分分析を行う場合，変化の大きさを表す 1 単位の大きさをそろえるため，

「もとの変数から平均値を引いて平均がゼロになるように変換し，
さらに標準偏差で割って標準偏差を1 単位とした数値に変換する」.....(4.2)

という変換を行う。これを「基準化」という。例として 2 つの変数 x_1，x_2 をあげる。それぞれの平均を m_1，m_2，標準偏差を s_1，s_2 とすれば，基準化された変数 $\tilde{x}_1$，$\tilde{x}_2$ は次のように表される。

$$\tilde{x}_1 = \frac{x_1 - m_1}{s_1}, \quad \tilde{x}_2 = \frac{x_2 - m_2}{s_2} \qquad \text{.....(4.3)}$$

変数 $\tilde{x}_1$，$\tilde{x}_2$ はともに平均ゼロ，標準偏差 1（よって分散 1）になる。これを用いて主成分

$$z = l_1\tilde{x}_1 + l_2\tilde{x}_2 \qquad \text{.....(4.4)}$$

を求めることになる。それには，

「基準化した変数 $\tilde{x}_1$，$\tilde{x}_2$ の分散共分散行列を用いて
これまでと同じく固有値と固有ベクトルを求めればよい。」 .....(4.5)

ただここで統計的には,

基準化した変数$\tilde{x}_1$, $\tilde{x}_2$の分散共分散行列」

＝「もとの変数x_1, x_2の相関行列」.....(4.6)

である。よって, 主成分を求めるには,

「x_1, x_2の相関行列から固有値と固有ベクトルを求めればよい。」.....(4.7)

もとの変数x_1, x_2の相関行列を

$$R=\begin{pmatrix}1 & r_{12}\\ r_{21} & 1\end{pmatrix}$$.....(4.8)

(注) $r_{12}=r_{21}=$変数x_1, x_2の相関係数

とすれば,

$$R\boldsymbol{l}=l\boldsymbol{l},\quad l_1^2+l_2^2=1$$ .....(4.9)

を解けばよい。

4.2 相関行列による主成分分析

(1) princomp による実行（cov オプションを付けない）

x_1（対個人サービス）とx_2（対事業所サービス）の 2 変数の相関行列を用いて, SAS で主成分分析を行ってみよう（表 4.1 参照）。

① princomp プロシジャで主成分分析を実行する。
② data=入力データセット名： 分析に用いるSASデータセット名を指定する。ここではindus。
③ 相関行列から固有値と固有ベクトルを求めるため cov オプションをつけない。
④ out=出力データセット名: 分析結果の主成分スコアを出力するSASデータセット名を指定する。ここでは prin。第 1 主成分, 第 2 主成分には prin1, prin2 という変数名がつく。
⑤ var： 主成分分析に用いる変数名を指定する。ここでは, x1 と x2。

表 4.1 SAS プログラム： 2 変数による主成分分析 (相関行列による)

```
/*********************************************************************
#                                                                    #
#  datB_prin02.sas: 2変数による主成分分析 ( 相関行列による )          #
#                                                                    #
*********************************************************************/
options  nocenter nodate  nonumber  linesize=100   pagesize=500;

/*---------------------------------+
| ... データセットindusを準備する   |
+---------------------------------*/
data indus;
  infile 'f:\datB\datB_jigyosho2.txt';
  length  pref $ 8;
  input   pref $  x01-x15 pop;
  x1=(x05 + x08 + x09 + x10 + x15)/pop * 10000;          /* 1.対個人サ  */
  x2=(x04 + x06 + x11 + x12 + x13 + x14)/pop * 10000;    /* 2.対事業所サ*/
  x3=x03/pop * 10000;                                    /* 3.製造      */
  x4=x07/pop * 10000;                                    /* 4.公的機関  */
  x5=x02/pop * 10000;                                    /* 5.インフラ  */
  x6=x01/pop * 10000;                                    /* 6.農林魚鉱  */
  label
  x1='1.対個人サ '
  x2='2.対事業サ '
  x3='3.製造     '
  x4='4.公的機関 '
  x5='5.インフラ '
  x6='6.農林魚鉱 ';
run;

/*---------------------------------+
| ... 2変数の主成分分析を行う      |
+---------------------------------*/
proc princomp  data=indus      out=prin;
  title    ' ';
  title2   '表  2変数による主成分分析 ( 相関行列による )';
  title3   '      (1) 対個人サービス × 対事業所サービス';
  title4   '      (2) 人口10000人当り従業者数による(2001年)';
  var    x1 x2;
run;

/*---------------------------------+
| ... 主成分スコアをプリントする   |
+---------------------------------*/
proc print data=prin;
  title   ' ';
  title2  '表  主成分スコア';
  format prin1-prin2 8.3;
  var pref prin1-prin2;
run;
```

データセットindusを作成する。

1. princompプロシジャ(主成分分析)を実行する。
2. data=indus: データセットindusを使用する。
3. 相関行列から固有値を求めるため，covオプションをつけない。
4. out=prin: 分析結果の主成分スコアをデータセットprinに出力する。第1主成分，第2主成分には自動的にprin1，prin2という変数名がつく。

変数x1とx2の2変数を用いて主成分分析を行うことを指定する

1. printプロシジャを実行する。
2. data=prin: データセットprinを使用する。主成分スコアが記録されている。

出力のフォーマットを8.3に設定

第1主成分スコア(prin1)，第2主成分スコア(prin2)をプリントするように指定。

(2) 分析結果の読み方

相関行列を用いた主成分分析の出力結果は，分散共分散行列を用いたときとよく似ている（表 4.2）。しかし結果の読み取り方がやや異なるところがあるので注意が必要である。

表 4.2 SAS 実行結果：2 変数による主成分分析 (相関行列による)

```
表　2変数による主成分分析 ( 相関行列による )
      (1) 対個人サービス × 対事業所サービス
      (2) 人口10000人当り従業者数による(2001年)

The PRINCOMP Procedure

Observations          47        <--- 観測値の件数
Variables              2        <--- 使用した変数の数

           Simple Statistics                    <--- 基礎統計
                    x1                  x2      <--- 使用した変数
Mean        1768.121330         1115.747001    <--- 平均値（基準化前のものが表示される）
StD          139.543957          358.192036    <--- 標準偏差（基準化前のものが表示される）

                Correlation Matrix              <--- 相関行列（もとの変数を基準化したときの
                                                     分散共分散行列に当る）
                                x1          x2
                                                 x1とx2の相関係数(r12)
x1      1.対個人サ          1.0000      x.xxxx
x2      2.対事業サ          x.xxxx      1.0000
                                                 x2とx2との相関係数(r2)＝1
                                                 これは基準化した変数の分散でもある。
                                                 x1とx1とについても同じ。
 x1とx2の相関係数(r21)

          Eigenvalues of the Correlation Matrix  <--- 相関行列の固有値

        Eigenvalue    Difference    Proportion    Cumulative
         固有値          差           寄与率        累積寄与率
   1    x.xxxxxxxx    x.xxxxxxxx      x.xxxx        x.xxxx
   2    x.xxxxxxxx                    x.xxxx        1.0000
                                                 全分散に占める第1主成分の割合
                                                   ＝λ1 ／(もとの変数の数)
 第1番目に大きい固有値λ1(第1主成分の分散)          全分散に占める第2主成分の割合
 第2番目に大きい固有値λ2(第2主成分の分散)            ＝λ2 ／(もとの変数の数)
 (注1) 固有値の合計 ＝ 全分散 ＝ v1 + v2
       ＝ 1 + 1 ＝ もとの変数の数                  (注) ふつう100%を掛けて%表示
 (注2) もとの変数は基準化してあるので分散                にする。
       は1である。

          Eigenvectors      <--- 固有ベクトル

                          Prin1         Prin2     <--- 主成分1,2の変数名（自動的にPrinが付く）
x1      1.対個人サ      x.xxxxxx      x.xxxxxx
x2      2.対事業サ      x.xxxxxx      x.xxxxxx
                                                 第2番目の固有値に対応する固有ベクトル
                                                 ( 第2主成分を求めるための重み係数 )
 第1番目の固有値に対応する固有ベクトル
 ( 第1主成分を求めるための重み係数 )

表　主成分スコア

OBS      pref          Prin1        Prin2

  1     北海道         xx.xxx       xx.xxx
  2     青森県         xx.xxx       xx.xxx
  3     岩手県         xx.xxx       xx.xxx
  :       :              :            :
  :       :              :            :
  :       :              :            :
 45     宮崎県         xx.xxx       xx.xxx
 46     鹿児島県       xx.xxx       xx.xxx
 47     沖縄県         xx.xxx       xx.xxx
                                                 本文中の(4.4)式によって求めた第1主成分のスコア
                                                 (値)と第2主成分のスコア。
```

① Observations：観測値の件数。予定した通りの件数が入力されたかチェックする。

② Variables： 使用した変数の数。

③ Simple Statistics： 基礎統計。使用した変数の平均値（Mean）と標準偏差（STD）。基準化する前の値が表示される。

④ Correlation Matrix： 相関行列。基準化した変数の分散共分散行列に等しい。

$$R=\begin{pmatrix}1 & r_{12}\\ r_{21} & 1\end{pmatrix}$$

⑤ "Total Variance"（全分散）は出力されない。が，基準化された変数の分散はすべて 1 である。よって，

全分散＝もとの変数の数.....(4.10)

である。この場合，

全分散$=v_1+v_2=1+1=2$.....(4.11)

である。分散は情報量を表すから，全分散とは使用したすべての変数が持っている総情報量を意味する。このうちできるだけ多くの情報量を，数少ない指標にまとめなおそうとする（合成する）のが主成分分析である。この例では，もとの変数の数が 2 つであるから，それより少ない 1 つの指標にまとまるのが望ましい。

⑥ Eigenvalue： 相関行列の固有値。(4.9)式の l に当る。これは主成分の分散v_zに等しい。大きい順にたてにプリントされる。数学的に固有値は，もとの変数の数（行列 R の次数）だけある。ここでは 2 つである。

(a) どの主成分までを総合指標として採択するかは
次の寄与率を参考に決める。.....(4.12a)

(b) または分散が 1 より大きい主成分を採択する。
なぜなら，もとの変数を基準化したものは分散が 1 であり，
総合指標としてはそれ以上の情報量を持っていることが
望ましいからである。.....(4.12b)

⑦ なお，すべての固有値の合計は，もとの変数の全分散（もとの変数の数）に等しい。

$l_1+l_2=v_1+v_2=1+1=2$ .....(4.13)

⑧ Proportion： 寄与率。それぞれの主成分の分散が，全分散の何割を占めるかを表す。

第 1 主成分の寄与率$=l_1/(v_1+v_2)$.....(4.14)

第 2 主成分の寄与率$=l_2/(v_1+v_2)$ .....(4.15)

一般にこれに 100%を掛けて%表示で表す。

⑨ 寄与率が大きいほど，その主成分が多くの情報量を集めて合成されたことを意味する。

⑩ Cumulative： 累積寄与率。寄与率を上から下へ順に累計した値。何番目の主成分までを採択すればよいかを判断するときに用いる。だいたい 70%くらいまでを採択する。

⑪ もし，第 1 主成分の寄与率が 60%，第 2 主成分の寄与率が 40%，という場合は，第 1 主成分だけで総合指標にするわけにはいかない。第 1 と第 2 の二つの主成分を用いなけ

ればならないだろう。これは変数の数の節約にはならないが，違った視点から個体の分類規準とすることができる。

⑫ Eigenvectors：固有ベクトル（重み係数）。主成分の式

$$z = l_1\tilde{x}_1 + l_2\tilde{x}_2 \quad \text{.....(4.4)}$$

におけるl_1，l_2を表す。ここで$\tilde{x}_1$，$\tilde{x}_2$は基準化した変数である（4.3式）。

⑬ いまある主成分の固有ベクトルについて，k番目の重み係数l_kが大きな値（プラスでもマイナスでも）をとっているならば，その主成分の性格にはもとの変数x_kの影響が強いということができる。変数x_kが効いているという言い方をすることもある。

⑭ 実は，2 変数の相関行列を用いた主成分分析では，それぞれの変数の重み係数の値（プラスかマイナスかを問わず）は常に等しい。よって，効き方は等しい。これは 2 変数という特殊性によるものである。なお 3 変数以上になると，重み係数の値は多様である。

⑮ (4.4)式で計算されたzの値を主成分スコアという。主成分スコアは，第 1 主成分と第 2 主成分のそれぞれについて計算される。それぞれに変数名として，Prin1，Prin2 が自動的につく。主成分スコアは，個体ごとに計算される。

課題11　2変数の相関行列による主成分分析の結果の読み取り(a)

x1(対個人サービス)とx2(対事業所サービス)の相関行列を用いて主成分分析を行いなさい(表4.1, 表4.2)。その結果をもとに以下の文の(1)～(6)の個所に適切な数値を解答しなさい。

x1とx2との相関係数は(1)である。第1主成分の分散は(2)で寄与率は(3)である。第1主成分と第2主成分の分散の合計は(4)である。第1主成分における，x1の重み係数は(5)であり，x2の重み係数は(6)である。

1．提出先: tak@reitaku-u.ac.jp
2．件名:　datB11 注）必ず半角とすること，大文字と小文字の区別をすること
3．提出物の例

```
+-------------------------------------------------------------------------+
|  課題11 学籍番号　氏名　ユーザID                                        |
|                                                                         |
|  (1) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
|  (2) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
|  (3) xxxxxx（小数2桁目を四捨五入して小数1桁まで，%表示で）              |
|  (4) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
|  (5) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
|  (6) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
+-------------------------------------------------------------------------+
```

以上

課題12　2変数の相関行列による主成分分析の結果の読み取り(b)

課題11と同じ分析結果をもとに以下の文の(1)～(3)の個所に適切な数値を解答しなさい。

ある地域について，x1=1834.8，x2=1475.4としたとき，それぞれを基準化した値は(1)，(2)である。これを用いて，この地域の第1主成分のスコアを計算すると(3)となる。

1．提出先: tak@reitaku-u.ac.jp
2．件名:　datB12 注）必ず半角とすること，大文字と小文字の区別をすること
3．提出物の例

```
+-------------------------------------------------------------------------+
|  課題12 学籍番号　氏名　ユーザID                                        |
|                                                                         |
|  (1) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
|  (2) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
|  (3) xxxxxx（小数4桁目を四捨五入して小数3桁まで）                       |
+-------------------------------------------------------------------------+
```

以上

5. 多変量の主成分分析

5.1 はじめに

これまでの2変量の分析から, より一般的な多変量の主成分分析について考えてみよう。目的は, 個々の個体について観測したp個の特性値$(x_1, x_2, \cdots, x_p)$の持つ情報を, より少ないm個の総合特性値つまり主成分$(z_1, z_2, \cdots, z_m)$に要約することである（ここで$p \geqq m$）。2変量の場合と比べて考え方が変わるわけではないが, 定式化のための行列表記が多くなる。

それより重要なのは,

① 分析結果として複数の主成分を採択する場合が多くなること,

② その場合, それぞれの主成分の意味づけが重要な作業になること,

という点である。

5.2 表記

全部でn個の個体（サンプル）があるとし, i番目の個体のp個の特性値を次のように縦ベクトル$\boldsymbol{x}_i$で表す。

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad i = 1,2,\cdots,n \qquad \cdots\cdots(5.1)$$

データ行列を次のように定める。なお, $\boldsymbol{x}_i'$は$\boldsymbol{x}_i$の転置を表す。

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad (n \times p) \qquad \cdots\cdots(5.2)$$

↑ ⋮ **サンプル数** ⋮ ↓

← **変数(特性値)** →

p個の特性値の平均値ベクトルを次のように縦ベクトル$\boldsymbol{m}$で表す。

$$\boldsymbol{m} = \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_p \end{pmatrix} \qquad \cdots\cdots(5.3)$$

平均値行列を次のように定める。

$$\boldsymbol{M}=\begin{pmatrix}\boldsymbol{m}'\\ \boldsymbol{m}'\\ \vdots\\ \boldsymbol{m}'\end{pmatrix}=\begin{pmatrix}m_1 & m_2 & \cdots & m_p\\ m_1 & m_2 & \cdots & m_p\\ \vdots & \vdots & \ddots & \vdots\\ m_1 & m_2 & \cdots & m_p\end{pmatrix}\quad(n\times p)\quad\begin{matrix}\uparrow\\ \vdots\\ \text{サンプル数}\\ \vdots\\ \downarrow\end{matrix}\qquad\text{.....(5.4)}$$

$\leftarrow$ 変数(特性値) $\rightarrow$

分散共分散行列は次のように表される。

$$\begin{aligned}\boldsymbol{V}&=\frac{1}{n}(\boldsymbol{X}-\boldsymbol{M})'(\boldsymbol{X}-\boldsymbol{M})\\ &=\frac{1}{n}(\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{X}'\boldsymbol{M}-\boldsymbol{M}'\boldsymbol{X}+\boldsymbol{M}'\boldsymbol{M})\\ &=\frac{1}{n}(\boldsymbol{X}'\boldsymbol{X}-\boldsymbol{M}'\boldsymbol{M})\\ &=\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}-\tilde{\boldsymbol{M}}\end{aligned}\qquad\text{.....(5.5)}$$

ここで,

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{M}=\frac{1}{n}\boldsymbol{M}'\boldsymbol{X}=\frac{1}{n}\boldsymbol{M}'\boldsymbol{M}=\begin{pmatrix}m_1m_1 & m_1m_2 & \cdots & m_1m_p\\ m_2m_1 & m_2m_2 & \cdots & m_2m_p\\ \vdots & \vdots & \ddots & \vdots\\ m_pm_1 & m_pm_2 & \cdots & m_pm_p\end{pmatrix}=\tilde{\boldsymbol{M}}\qquad\text{.....(5.6)}$$

である。

p 個の特性値の標準偏差を対角要素に持つ対角行列を次のように表す。

$$\boldsymbol{S}=\begin{pmatrix}s_1 & 0 & \cdots & 0\\ 0 & s_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & s_p\end{pmatrix}\qquad\text{.....(5.7)}$$

もとの変量を, 平均ゼロ, 分散 1 とした基準化データは次のように表される。

$$\boldsymbol{X}_S=(\boldsymbol{X}-\boldsymbol{M})\boldsymbol{S}^{-1}\quad(n\times p)\qquad\text{.....(5.8)}$$

相関行列は次のように表される。

$$\begin{aligned}\boldsymbol{R} &= \frac{1}{n}\boldsymbol{X}_s{}'\,\boldsymbol{X}_s \\ &= \frac{1}{n}\boldsymbol{S}^{-1}(\boldsymbol{X}-\boldsymbol{M})'(\boldsymbol{X}-\boldsymbol{M})\boldsymbol{S}^{-1} \\ &= \boldsymbol{S}^{-1}\boldsymbol{V}\boldsymbol{S}^{-1}\end{aligned} \quad \text{.....(5.9)}$$

5.3 分散最大化による主成分の導出

(1) 問題の定式化

ある個体について観測した p 個の特性値 $\boldsymbol{x}' = (x_1, x_2, \cdots, x_p)$ があるとき, その総合特性値 z を次のように定義する。なお, 個体を示す添え字を省略した。

$$z = l_1 x_1 + l_2 x_2 + \cdots + l_p x_p \quad \text{.....(5.10)}$$

$\boldsymbol{l}' = (l_1, l_2, \cdots, l_p)$ は未知の重み係数であり, これを求めるのが課題である。もとの p 個の特性値が持つ情報量はそれぞれの分散で表される。それを合成した総合特性値についても保持する情報量を最大にしたい。よって問題は,

z の分散（情報量）が最大になるように係数 $\boldsymbol{l}$ を定めよ.....(5.11)

となる。なお係数 $\boldsymbol{l}$ に制約がないと z の分散が無限に大きくなる。そこで,

$l_1^2 + l_2^2 + \cdots + l_p^2 = 1$　（つまり $\boldsymbol{l}'\boldsymbol{l} = 1$）.....(5.12)

という制約を設ける。改めて問題は次のように定式化できる。

分散最大化による主成分の導出

z の分散が最大になるように係数 $\boldsymbol{l}$ を定めよ。ただし, $\boldsymbol{l}'\boldsymbol{l} = 1$ とする。

$$\begin{aligned} &Var(z) \rightarrow \max \\ &s.t.\ \ \boldsymbol{l}'\boldsymbol{l} = 1 \end{aligned} \quad \text{.....(5.13)}$$

(2) 第１主成分の導出

個体が全部で n 個あるとする。i 番目の個体の総合特性値を z_i とする。

$$z_1 = l_1 x_{11} + l_2 x_{12} + \cdots + l_p x_{1p} \quad \text{.....(5.14)}$$

$$z_2 = l_1 x_{21} + l_2 x_{22} + \cdots + l_p x_{2p} \quad \text{.....(5.14)}$$

$$\cdots$$

$$z_n = l_1 x_{n1} + l_2 x_{n2} + \cdots + l_p x_{np} \quad \text{.....(5.14)}$$

ここで, $\boldsymbol{z}' = (z_1, z_2, \cdots, z_n)$, 係数 $\boldsymbol{l}$, データ行列 $\boldsymbol{X}$ を用いて行列表記すると

$$\boldsymbol{z} = \boldsymbol{X}\boldsymbol{l} \quad \text{.....(5.15)}$$

分散は次のように表される。

$$Var(z) = \frac{1}{n}(\boldsymbol{X}\boldsymbol{l} - \boldsymbol{M}\boldsymbol{l})'(\boldsymbol{X}\boldsymbol{l} - \boldsymbol{M}\boldsymbol{l})$$
$$= \boldsymbol{l}'\frac{1}{n}(\boldsymbol{X} - \boldsymbol{M})'(\boldsymbol{X} - \boldsymbol{M})\boldsymbol{l} \quad \text{.....(5.16)}$$
$$= \boldsymbol{l}'\boldsymbol{V}\boldsymbol{l}$$

$\boldsymbol{V}$はもとの特性値$\boldsymbol{x}$の分散共分散行列である。以上をもとに分散最大化を解く。λをラグランジェ乗数として,

$$W = \boldsymbol{l}'\boldsymbol{V}\boldsymbol{l} - l(\boldsymbol{l}'\boldsymbol{l} - 1) \quad \text{.....(5.17)}$$

とおいて, 一階の条件を求める。

$$\frac{\partial W}{\partial \boldsymbol{l}} = 2\boldsymbol{l}'\boldsymbol{V} - 2l\boldsymbol{l}' = \boldsymbol{0}' \quad \text{.....(5.18)}$$

$$\frac{\partial W}{\partial l} = -(\boldsymbol{l}'\boldsymbol{l} - 1) = 0 \quad \text{.....(5.19)}$$

$$\therefore \quad \boldsymbol{V}\boldsymbol{l} = l\boldsymbol{l} \quad \text{.....(5.20)}$$

ここから, λは$\boldsymbol{V}$の固有値として求められ, $\boldsymbol{l}$はそれに対応する固有ベクトルとして求められる。このとき,

$$Var(z) = \boldsymbol{l}'\boldsymbol{V}\boldsymbol{l} = l\boldsymbol{l}'\boldsymbol{l} = l \quad \text{.....(5.21)}$$

となってλはzの分散を意味している。したがって, 分散最大化の解は次のように整理できる。

① $\boldsymbol{V}$の最大固有値λがzの分散を最大化したときの値である。.....(5.22)
② 最大固有値λに対応する固有ベクトルを係数$\boldsymbol{l}$とすればよい。.....(5.23)

こうして求められたzを第1主成分と呼び, 改めてz_1と表す（ここでの添え字は個体ではなく主成分を表す）。係数ベクトルを$\boldsymbol{l}_1$とする。またそのときの分散の最大値をl_1とする。添え字の表記を明確にしておく。

③ i番目の個体の第1主成分のスコア（値）をz_{i1}と表す。データベクトルを次のように表す。
$$\boldsymbol{z}_1' = (z_{11}, z_{21}, \cdots, z_{n1}) \quad \text{.....(5.24)}$$
④ 第1主成分の係数ベクトルを次のように表す。
$$\boldsymbol{l}_1' = (l_{11}, l_{21}, \cdots, l_{p1}) \quad \text{.....(5.25)}$$
⑤ もとのデータ行列$\boldsymbol{X}$との関係は, 次のように表される。
$$\boldsymbol{z}_1 = \boldsymbol{X}\boldsymbol{l}_1 \quad \text{.....(5.26)}$$
⑥ $\boldsymbol{V}\boldsymbol{l}_1 = l_1\boldsymbol{l}_1$ .....(5.27)
⑦ $Var(z_1) = l_1$ .....(5.28)

(3) 第k主成分の導出

さて次に, この第1主成分と無相関で, 2番目に分散の大きな総合特性値を求めることができる。これを第2主成分z_2という。同様にして, 第1, 2, ･･･, $k-1$主成分と無相関で, k番目に分散の大きな総合特性値を第k主成分z_kという。いま第1, 2, ･･･, $k-1$主成分が得られたと

して, 第 k 主成分 z_k は次のようにして求められる。

分散最大化による第 k 主成分の導出

$Var(z_k) \to \max$ .....(5.29)

s.t.

$\boldsymbol{l}_k' \boldsymbol{l}_k = 1$

$\boldsymbol{l}_j' \boldsymbol{l}_k = 0 \quad (j = 1,2,\cdots,k-1)$ (注) 無相関の条件。後記。

$m_1, m_2, \cdots, m_{k-1}, l_k$ をラグランジェ乗数として,

$$W = \boldsymbol{l}_k' \boldsymbol{V} \boldsymbol{l}_k - \sum_{j=1}^{k-1} m_j \boldsymbol{l}_j' \boldsymbol{l}_k - l_k (\boldsymbol{l}_k' \boldsymbol{l}_k - 1) \quad \text{.....(5.30)}$$

とおいて, 一階の条件を求める。

$$\frac{\partial W}{\partial \boldsymbol{l}_k} = 2\boldsymbol{l}_k' \boldsymbol{V} - \sum_{j=1}^{k-1} m_j \boldsymbol{l}_j' - 2 l_k \boldsymbol{l}_k' = \boldsymbol{0}' \quad \text{.....(5.31a)}$$

$$\boldsymbol{l}_k' \boldsymbol{l}_k = 1 \quad \text{.....(5.31b)}$$

$$\boldsymbol{l}_j' \boldsymbol{l}_k = 0 \quad (j = 1,2,\cdots,k-1) \quad \text{.....(5.31c)}$$

ここで, $\partial W / \partial \boldsymbol{l}_k = \boldsymbol{0}'$ 式の両辺に右から $\boldsymbol{l}_r$ $(r = 1,2,\cdots,k-1)$ を掛ける。

$$2\boldsymbol{l}_k' \boldsymbol{V} \boldsymbol{l}_r - \sum_{j=1}^{k-1} m_j \boldsymbol{l}_j' \boldsymbol{l}_r - 2 l_k \boldsymbol{l}_k' \boldsymbol{l}_r = 0 \quad \text{.....(5.32a)}$$

$$2\boldsymbol{l}_k' (l_r \boldsymbol{l}_r) - m_r - 0 = 0 \quad \text{.....(5.32b)}$$

（$\because$ $\boldsymbol{V}\boldsymbol{l}_r = l_r \boldsymbol{l}_r$ (l_r は $\boldsymbol{V}$ の固有値), $\boldsymbol{l}_j' \boldsymbol{l}_r = 0$ $(j \neq r)$,
$\boldsymbol{l}_j' \boldsymbol{l}_r = 1$ $(j = r)$, $\boldsymbol{l}_k' \boldsymbol{l}_r = 0$）

$$0 - m_r - 0 = 0 \quad \text{.....(5.33c)}$$

（$\because$ $\boldsymbol{l}_k' \boldsymbol{l}_r = 0$）

$\therefore \quad m_j = 0 \quad (j = 1,2,\cdots,k-1)$

これを $\partial W / \partial \boldsymbol{l}_k = \boldsymbol{0}'$ に適用して,

$$\boldsymbol{V} \boldsymbol{l}_k = l_k \boldsymbol{l}_k \quad \text{.....(5.34)}$$

となる。これは先に第 1 主成分を求めたのと同様の解である。ただ, $\boldsymbol{V}$ に関する第 1, 2, ⋯, $k-1$ 番目の固有値と固有ベクトルは, 第 1, 2, ⋯, $k-1$ 番目の主成分として既に使用されている。よって, ここでは第 k 番目の固有値 l_k を第 k 主成分 z_k の分散とし, それに対応する固有ベクトル $\boldsymbol{l}_k$ を係数ベクトルとすればよい。

(4) 相関行列からの主成分の導出

通常はもとの変量の分散共分散行列 $\boldsymbol{V}$ を用いるのではなく, 基準化した変量の分散共分散行列 $\boldsymbol{R}$ (もとの変量の相関行列) を用いて,

$\boldsymbol{Rl} = l\boldsymbol{l}$ .....(5.35)

を解いて, 固有値 l と固有ベクトル $\boldsymbol{l}$ を求める。

主成分は相関行列 $\boldsymbol{R}$ の次数（もとの変数の数）だけ存在する。

また, 主成分スコアは, 基準化した変数 $\tilde{x}_i$ と固有ベクトル（重み係数）$\boldsymbol{l}$ を用いて,

$z = l_1\tilde{x}_1 + l_2\tilde{x}_2 + \cdots + l_p\tilde{x}_p$ .....(5.35a)

と計算される。

5.4 相関行列による多変量主成分分析

(1) princomp プロシジャによる実行（cov オプションを付けない）

x_1（対個人サービス），x_2（対事業所サービス），x_3（製造），x_4（公的機関），x_5（インフラ），x_6（農林魚鉱）の 6 変数の相関行列を用いて主成分分析を行ってみよう（表 5.1 参照）。

① princomp プロシジャで主成分分析を実行する。
② data=入力データセット名： 分析に用いるSASデータセット名を指定する。ここではindus。
③ 相関行列から固有値と固有ベクトルを求めるため cov オプションをつけない。
④ out=出力データセット名: 分析結果の主成分スコアを出力するSAS データセット名を指定する。ここではprin。第 1 主成分, 第 2 主成分...第 6 主成分には prin1, prin2...prin6 という変数名が自動的につく。
⑤ var： 主成分分析に用いる変数名を指定する。ここでは, x1, x2...x6。

表 5.1 SAS プログラム： 6 変数による主成分分析 (相関行列による)

```
/*************************************************************************
#  datB_prin03.sas: 6変数による主成分分析 ( 相関行列による )          #
*************************************************************************/
options  nocenter nodate  nonumber  linesize=100   pagesize=500;
data indus;
  infile 'f:\datB\datB_jigyosho2.txt';
  length  pref $ 8;
  input   pref $  x01-x15 pop;
  x1=(x05 + x08 + x09 + x10 + x15)/pop * 10000;          /* 1.対個人サ   */
  x2=(x04 + x06 + x11 + x12 + x13 + x14)/pop * 10000;    /* 2.対事業所サ*/
  x3=x03/pop * 10000;                                    /* 3.製造       */
  x4=x07/pop * 10000;                                    /* 4.公的機関   */
  x5=x02/pop * 10000;                                    /* 5.インフラ   */
  x6=x01/pop * 10000;                                    /* 6.農林魚鉱   */
  label
  x1='1.対個人サ '
  x2='2.対事業サ '
  x3='3.製造     '
  x4='4.公的機関 '
  x5='5.インフラ '
  x6='6.農林魚鉱 ';
run;
proc princomp  data=indus  out=prin;
  title    ' ';
  title2   '表  6変数による主成分分析 ( 相関行列による )';
  title3   '    ( 人口10000人当り従業者数による: 2001年 )';
  var    x1-x6;
run;
proc print data=prin;
  title    ' ';
  title2  '表  主成分スコア ( 相関行列による )';
  format prin1-prin6 8.3;
  var pref prin1-prin6;
run;
```

データセットindusを作成する。

1. princompプロシジャ(主成分分析)を実行する。
2. data=indus: データセットindusを使用する。
3. 相関行列から固有値を求めるため, covオプションをつけない。
4. out=prin: 分析結果の主成分スコアをデータセットprinに出力する。第1主成分, 第2主成分...第6主成分には自動的にprin1, prin2...prin6という変数名がつく。

変数x1,x2...x6の6変数を用いて主成分分析を行うことを指定する

1. printプロシジャを実行する。
2. data=prin: データセットprinを使用する。主成分スコアが記録されている。

出力のフォーマットを8.3に設定

第1主成分～第6主成分のスコア(prin1～prin6)をプリントするように指定。

(2) 分析結果の読み取り

表 5.2 SAS 実行結果：6変数による主成分分析 (相関行列による)

表　6変数による主成分分析 (相関行列による)
　　(人口10000人当り従業者数による：2001年)

The PRINCOMP Procedure

Observations　47　<--- ***観測値の件数***
Variables　6　<--- ***使用した変数の数***

平均値（基準化前のものが表示される）

Simple Statistics　<--- ***基礎統計***

	x1	x2	x3	x4	x5	x6
Mean	1768.121330	1115.747001	867.7966571	311.5294310	446.2727569	32.00538426
StD	139.543957	358.192036	284.5045206	60.3185534	87.5523580	19.57319514

標準偏差（基準化前のものが表示される）

Correlation Matrix　<--- ***相関行列（もとの変数を基準化したときの分散共分散行列に当る）***

相関係数

		x1	x2	x3	x4	x5	x6
x1	1.対個人サ						
x2	2.対事業サ						
x3	3.製造						
x4	4.公的機関						
x5	5.インフラ						
x6	6.農林魚鉱						

Eigenvalues of the Correlation Matrix　<--- ***相関行列の固有値***

	Eigenvalue ***固有値***	Difference ***差***	Proportion ***寄与率***	Cumulative ***累積寄与率***
1		0.46023348		
2		0.84644164		
3		76797913		
4		324684		
5		1093		
6				

第1番目に大きい固有値λ_1(第1主成分の分散)
第2番目に大きい固有値λ_2(第2主成分の分散)
.....(以下同様)
(注1) ***固有値の合計 ＝ 全分散 ＝ v1+v2+...+v6***
＝ 1+1+...+1 ＝もとの変数の数 ＝6
(注2) ***もとの変数は基準化してあるので分散は１である。***

全分散に占める第1主成分の割合
＝λ_1 ／(もとの変数の数)
全分散に占める第2主成分の割合
＝λ_2 ／(もとの変数の数)
.....(以下同様)

(注) ***ふつう100%を掛けて%表示にする。***

Eigenvectors　<--- ***固有ベクトル***

固有ベクトル

		Prin1 ***第1主成分***	Prin2 ***第2主成分***	Prin3 ***第3主成分***	Prin4 ***第4主成分***	Prin5 ***第5主成分***	Prin6 ***第6主成分***
x1	1.対個人サ			-.122025	0.212952	0.413537	-.553346
x2	2.対事業サ			-.061785	-.437390	-.135485	0.565619
x3	3.製造			0.815547	0.360951	0.217056	0.234845
x4	4.公的機関			-.173061	0.671542	-.176018	0.384879
x5	5.インフラ			0.511571	-.242888	-.529202	-.366290
x6	6.農林魚鉱			0.156591	-.350846	0.672667	0.190868

表　主成分スコア（ 相関行列による ）

OBS	pref	Prin1 *第1主成分*	Prin2 *第2主成分*	Prin3 *第3主成分*	Prin4 *第4主成分*	Prin5 *第5主成分*	Prin6 *第6主成分*
1	北海道				-0.878	0.636	-0.015
2	青森県				-0.033	-0.272	0.183
3	岩手県				-0.846	0.621	0.412
4	宮城県				-0.959	-0.224	-0.409
5	秋田県				-0.322	-0.328	-0.104
6	山形県				0.316	-0.702	0.072
7	福島県				-0.603	-0.430	-0.353
8	茨城県				-0.299	0.227	-0.125
9	栃木県				-0.131	0.661	-0.260
10	群馬県				0.327	0.399	0.051
11	埼玉県				-0.781	-0.472	0.556
12	千葉県				-0.911	-0.076	-0.171
13	東京都				-0.374	-0.062	0.867
14	神奈川県				-0.726	-0.120	-0.218
15	新潟県				-0.767	-0.502	-0.037
16	富山県				-0.076	-0.196	0.016
17	石川県				0.544	-0.204	-0.316
18	福井県				0.528	-0.962	-0.478
19	山梨県				0.626	0.322	-0.709
20	長野県				-0.061	0.271	0.008
21	岐阜県				0.057	0.422	-0.058
22	静岡県				0.394	0.416	0.145
23	愛知県				-0.166	0.159	0.097
24	三重県				0.419	0.519	0.253
25	滋賀県				1.359	0.218	0.549
26	京都府				1.284	0.846	-0.465
27	大阪府				-0.626	-0.079	0.131
28	兵庫県				0.055	0.086	-0.152
29	奈良県				0.452	-0.387	0.649
30	和歌山県				0.473	-0.169	0.221
31	鳥取県				0.354	-0.101	0.456
32	島根県				0.437	-0.531	0.567
33	岡山県				-0.023	-0.109	0.182
34	広島県				-0.014	-0.341	0.146
35	山口県				0.368	-0.621	-0.373
36	徳島県				0.248	-0.298	-0.043
37	香川県				0.046	-0.304	0.085
38	愛媛県				-0.636	0.099	-0.002
39	高知県				0.234	0.461	-0.810
40	福岡県				-0.442	-0.355	-0.624
41	佐賀県				0.629	-0.045	-0.001
42	長崎県				0.315	0.555	0.333
43	熊本県				0.211	0.271	0.042
44	大分県				0.072	0.022	-0.346
45	宮崎県				-0.401	0.690	0.110
46	鹿児島県				-0.303	0.941	0.081
47	沖縄県				0.634	-0.952	-0.139

① Observations：観測値の件数。予定した通りの件数が入力されたかチェックする。

② Variables： 使用した変数の数。

③ Simple Statistics： 基礎統計。使用した変数の平均値（Maen）と標準偏差（STD）。基準化する前の値が表示される。

④ Correlation Matrix： 相関行列 R。基準化した変数の分散共分散行列に等しい。

⑤ "Total Variance"（全分散）は出力されない。が, 基準化した変数の分散はすべて 1 であ

る。よって,

全分散＝もとの変数の数＝p.....(5.36)

である。分散は情報量を表すから, 全分散とは使用したすべての変数が持っている総情報量を意味する。このうちできるだけ多くの情報量を, 数少ない指標にまとめなおそうとする（合成する）のが主成分分析である。

⑥ Eigenvalue： 相関行列の固有値。(5.35)式の l に当る。これは主成分の分散 v_z（情報量）に等しい。大きい順にたてにプリントされる。数学的に固有値は, もとの変数の数（行列 R の次数 p）だけある。この値が大きい主成分ほど総合指標としての意味がある。

(a) どの主成分までを総合指標として採択するかは
寄与率（下記参照）を参考に決める。.. .....(5.37a)

(b) または分散が 1 より大きい主成分を採択する。
なぜなら, もとの変数を基準化したものは分散が 1 であり,
総合指標としてはそれ以上の情報量を持っていることが
望ましいからである。.....(5.37b)

⑦ なお, すべての固有値の合計は, 基準化した変数の全分散（もとの変数の数）に等しい。

$$l_1 + l_2 + \cdots + l_p = 1+1+\cdots+1 = p \quad \text{.....(5.38)}$$

⑧ Proportion： 寄与率。それぞれの主成分の分散が, 全分散の何割を占めるかを表す。

第 1 主成分の寄与率＝ l_1 / p .....(5.39a)

第 2 主成分の寄与率＝ l_2 / p .....(5.39b)

・・・・・（以下同様）

一般にこれに 100%を掛けて%表示で表す。

寄与率が大きいほど, その主成分が多くの情報量を集めて合成されたことを意味する。

⑨ Cumulative： 累積寄与率。寄与率を上から下へ順に累計した値。何番目の主成分までを採択すればよいかを判断するときに用いる。だいたい 70%くらいまでを採択する。

⑩ もし, 第 1 主成分の寄与率が 40%, 第 2 主成分の寄与率が 35%, という場合は, 第 1 主成分だけで総合指標にするわけにはいかない。第 1 と第 2 の二つの主成分を用いなければならないだろう。

⑪ Eigenvectors： 固有ベクトル（重み係数）$\boldsymbol{l}$。主成分の式

$$z = l_1\tilde{x}_1 + l_2\tilde{x}_2 + \cdots + l_p\tilde{x}_p \quad \text{.....(5.35a)}$$

におけるl_1, $l_2 \ldots l_p$ を表す。ここで $\tilde{x}_1$, $\tilde{x}_2 \ldots \tilde{x}_p$ は基準化した変数である（4.3 式）。

⑫ いまある主成分の固有ベクトルについて, k 番目の重み係数l_k が大きな値（プラスでもマイナスでも）をとっているならば, その主成分の性格にはもとの変数 x_k の影響が強いということができる。変数x_k が効いているという言い方をすることもある。

⑬ (5.35a)式で計算されたz の値を主成分スコアという。主成分スコアは, 第 1 主成分～第 p 主成分のそれぞれについて計算される。それぞれに変数名として, Prin1, Prin2...が自動的につく。主成分スコアは, 個体ごとに計算される。

課題13　6変数による産業構造の分析(a)
　x1（対個人サービス），x2（対事業所サービス），x3（製造），x4（公的機関），x5（インフラ），x6（農林魚鉱）の6変数の相関行列を用いて主成分分析を実行しなさい。その結果をもとに以下の文の(1)～(10)の個所に適切な数値または語句を解答しなさい。

　対個人サービスと最も相関が強い産業は(1)である。いま相関係数の絶対値が0.6以上のとき，相互の相関が強いということにする。すると，1つも相関の強い相手を持たない産業は(2)である。これは立地の独立性が高い産業，つまり近くに関連する産業を必要としない産業であると考えられる。
　第1主成分の分散は(3)で，その寄与率は%で表示すると(4)である。基準化した変数で見ると，もとの全分散は(5)である。それを全情報量とすると，そのうちの70%以上の情報量を集めるには，第(6)主成分から第(7)主成分までを採択したほうがよい。
　一方，分散が1以上の主成分を採択することにすると，第(8)主成分から第(9)主成分までを採択したほうがよいことになる。その場合，これらの累積寄与率は%で表示すると(10)である。

1．提出先: tak@reitaku-u.ac.jp
2．件名:　datB13 注）必ず半角とすること，大文字と小文字の区別をすること
3．提出物の例

```
+-------------------------------------------------------------------------+
| 課題13 学籍番号　氏名　ユーザID                                        |
| (1) xxxxxx (語句で)                                                     |
| (2) xxxxxx (語句で)                                                     |
| (3) xxxxxx (小数3桁目を四捨五入して小数2桁まで)                         |
| (4) xxxxxx (小数2桁目を四捨五入して小数1桁まで，%表示で)                |
| (5) xxxxxx (小数2桁目を四捨五入して小数1桁まで)                         |
| (6) xxxxxx (小数なし)                                                   |
| (7) xxxxxx (小数なし)                                                   |
| (8) xxxxxx (小数なし)                                                   |
| (9) xxxxxx (小数なし)                                                   |
| (10) xxxxxx (小数2桁目を四捨五入して小数1桁まで，%表示で)               |
+-------------------------------------------------------------------------+
```

以上

課題14　6変数による産業構造の分析(b)
　課題13と同じ分析結果をもとに，以下の文の(1)～(10)の個所に適切な語句を解答しなさい。

　第1主成分の重み係数を見ると，(1)産業だけがマイナスでそれ以外の産業はプラスである。よって第1主成分は，(1)産業の独立した集積度を表す軸だと考えられる。この場合，主成分スコアは，(1)産業の集積度が大きいほどマイナスの値をとる。(1)産業の集積度が最も大きい地域を大きい方から順に2つあげると(2)，(3)である。逆にプラスの方は(1)以外の産業の集積が大きいところである。その最も大きい地域を大きい方から順に2つあげると(4)，(5)である。
　第2主成分の重み係数を見ると，(6)，(7)という産業の値がプラスで大きい。よって第2主成分は，都市化の度合いを表していると考えられる。都市化の度合いが最も高い地域を高い方から順に2つあげると(8)，(9)である。また都市化の度合いが高い方から24番目の地域は(10)である。
（注）SASによる主成分スコアのソート方法(課題13のSASプログラムに続けて)

```
proc sort data=prin out=sorted;
  by descending prin2;
run;
proc print data=sorted;
  title  '表　主成分スコア(相関行列による，prin2ソート済み)';
  format prin1-prin6 8.3;
  var pref prin1-prin6;
run;
```

1．提出先: tak@reitaku-u.ac.jp
2．件名:　datB14 注）必ず半角とすること，大文字と小文字の区別をすること
3．提出物の例

```
+-------------------------------------------------------------------------+
| 課題14 学籍番号　氏名　ユーザID                                        |
| (1) xxxxxx (産業で)                                                     |
| (2) xxxxxx (地域で)                                                     |
| (3) xxxxxx (地域で)                                                     |
| (4) xxxxxx (地域で)                                                     |
| (5) xxxxxx (地域で)                                                     |
| (6) xxxxxx (産業で)                                                     |
| (7) xxxxxx (産業で)                                                     |
| (8) xxxxxx (地域で)                                                     |
| (9) xxxxxx (地域で)                                                     |
| (10) xxxxxx (地域で)                                                    |
+-------------------------------------------------------------------------+
```

以上