主成分分析演習問題

[1] 以下のア ～ シ に適当な数値, 言葉を入れること.

	中間試験 (x)	期末試験 (y)
A 君	10	7
B 君	6	4
C 君	8	3
D 君	7	6
E 君	9	5

データ x の平均値は 8, y の平均値は［ア］となる. 平均値はデータの［イ］を表す. 一方で x の標本分散 s_x^2 は 2, y の標本分散 s_y^2 は［ウ］で与えられる. 一般的に標本分散はデータの［エ］を表す. 上記の対のデータの共分散 s_{xy} は［オ］で与えられることより、その相関係数 r は［カ］となる. 一般的に相関係数は正の相関が強いほど［キ］に近い値をとり、負の相関が強いほど［ク］に近い値をとる.

また相関行列は $R = \begin{pmatrix} [ケ] & [コ] \\ [コ] & [サ] \end{pmatrix}$ となり, $tr(R) =$［シ］となる.

[2] $A = \begin{pmatrix} 6 & -\sqrt{6} \\ -\sqrt{6} & 1 \end{pmatrix}$ とする.

① A の固有値を求めよ.

② A の固有ベクトルを求めよ.

③ A の大きさ 1 の固有ベクトルを求めよ

[3] 以下の記述は主成分分析について述べたものである. 正しいものには T を, 誤っているものには F を記入せよ.

① 主成分分析はデータを標準化するか否かによって, 主成分の値は変わらない (　　)

② 主成分ベクトルは相関行列の大きさ 1 の固有ベクトルのことである (　　)

③ 主成分分析は多次元の情報を低次元の情報に圧縮する統計処理である (　　)

④ biplot における矢印は逆向きにでてくることがある (　　)

[4] 今, 50 人の体育, 数学, 理科の標準化されたデータ (x, y, z) から相関行列を計算したところ $R = \begin{pmatrix} 1 & 0 & -3/5 \\ 0 & 1 & 4/5 \\ -3/5 & 4/5 & 1 \end{pmatrix}$ となった.

① このとき R の固有値を求めよ

② 主成分ベクトルを求めよ.

③ A 君の標準化された体育, 数学, 理科の成績はそれぞれ $\frac{6}{5}, \frac{4}{5}, -\frac{1}{5}$ であった. この時, A 君の第 1 主成分スコアを求めよ.

[5] 3 ページに 2008 年度の各チームの投手全員の防御率, 勝数, 被安打（ヒットを打たれた数）, 被本塁打（ホームランを打たれた数）, 三振（三振を奪った数）, 四球（四球を与えた数）, 死球（デッドボールを与えた数）を調べたデータがある. ただし小さい数の方が良いデータに対しては、マイナスの符号をつけ、データを加工してある. このデータに対して標準化を行った後に主成分分析を行った. (ただし biplot だけは, このページの右側にある)

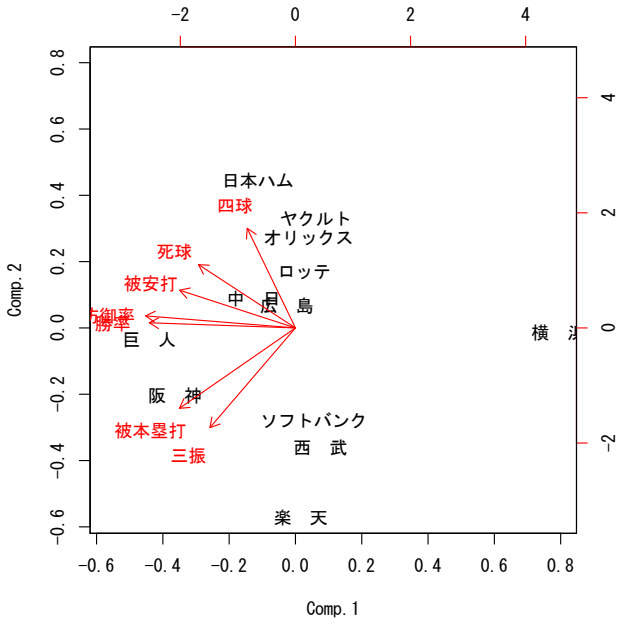
① 第 1 主成分スコアは何を表しているか?

② 第 2 主成分スコアは何を表しているか?

③ 第 3 主成分 (第 3 主成分を含める) までの累積寄与率を求めよ

④ ソフトバンクと特徴が似ているチームを一つ挙げよ

⑤ 三振を多く取れるピッチャーは, どのような特徴があるか? biplot から判断して答えよ.



[6] 中学 1 年生 30 人の身長, 体重, 胸囲, 座高に対して, 相関行列を用いて主成分分析を行った. (生のデータ, 分析結果は 4 ページに載せてある)。

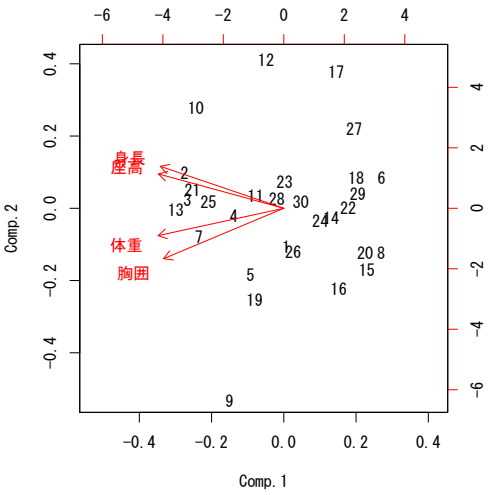
この時, biplot に関して, 以下の a. から d. までの選択肢の中から最もふさわしいものを選ぶこと.

① 第 1 主成分（スコア）は何を表しているか?

a. 総合的な体格の小ささ（右にあるほど体格が小さくなる）
b. 総合的な体格の大きさ（右にあるほど体格が大きくなる）
c. どのくらい痩せているか（右にあるほど痩せている）
d. どのくらい太っているか（右にあるほど太っている）
② 第 2 主成分（スコア）は何を表しているか?
a. 総合的な体格の小ささ（上にあるほど体格が小さくなる）
b. どのくらい身長が低いか（上にあるほど身長は低くなる）
c. どのくらい痩せているか（上にあるほど痩せている）
d. どのくらい太っているか（上にあるほど太っている）
③ 生徒番号 1 の身長における z-score を求めよ.
ただし各項目の平均, 標準偏差, 生徒番号 1 の z-score は以下の値とする

	身長	体重	胸囲	座高
平均	149	38.4	72.2	79.4
標準偏差	7.2	6.7	5.1	4.2
生徒番号 1 の z-score	?	0.388	-0.039	-0.333

④ 生徒番号 1 の 第 2 主成分 の主成分スコアを求めよ.
⑤ 第 3 主成分 (第 3 主成分を含める) までの累積寄与率を求めよ
⑥ biplot **のみ**から判断する時, 生徒番号 17 番の生徒と 18 番の生徒は, どちらが座高が高いと判断できるか? またそう判断した理由を、下記の biplot に補助線等を用いて説明すること



[7] (行列の計算) $\mathbf{a} = \begin{pmatrix} 1 \\ x \end{pmatrix}$, $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ とする.
① $\mathbf{a}^T A \mathbf{a}$ を計算せよ.
② $f(x) = \mathbf{a}^T A \mathbf{a}$ の最小値を求めよ.

解答 [1] ア 5 イ 中心 ウ 2 エ 散らばり オ 1 カ 0.5 キ 1 ク -1 ケ 1 コ 0.5 サ 1 シ 2
[2] ① $\lambda = 7, 0.$ ② $\mathbf{u} = k\begin{pmatrix} -\sqrt{6} \\ 1 \end{pmatrix}$ $(k \neq 0)$, $\mathbf{v} = l\begin{pmatrix} 1/\sqrt{6} \\ 1 \end{pmatrix}$ $(l \neq 0)$. ③ $\mathbf{u} = \pm\frac{1}{\sqrt{7}}\begin{pmatrix} -\sqrt{6} \\ 1 \end{pmatrix}$, $\mathbf{v} = \pm\frac{1}{\sqrt{7}}\begin{pmatrix} 1 \\ \sqrt{6} \end{pmatrix}$. [3] ① F ② T ③ T ④ T(大きさ 1 の固有ベクトルは, 2 つ出てくることを思い出してほしい)
[4] 固有値 $\lambda = 2, 1, 0$, 固有ベクトル $\mathbf{u} = \frac{1}{\sqrt{2}}\begin{pmatrix} -3/5 \\ 4/5 \\ 1 \end{pmatrix}$, $\mathbf{v} = \frac{3}{5}\begin{pmatrix} 4/3 \\ 1 \\ 0 \end{pmatrix}$, $\mathbf{w} = \frac{1}{\sqrt{2}}\begin{pmatrix} 3/5 \\ -4/5 \\ 1 \end{pmatrix}$. ③ $\frac{6}{5}\cdot\left(-\frac{3}{5\sqrt{2}}\right) + \frac{4}{5}\cdot\left(\frac{4}{5\sqrt{2}}\right) - \frac{1}{5}\cdot\frac{1}{\sqrt{2}} = -\frac{7}{25\sqrt{2}}$.
[5] ① 総合的な投手力 ② 三振を奪うことが少なく、ホームランを打たれることも多い。一方で四球やデッドボールが少なく、ヒットを打たれることも少ない。biplot の上にあるほど打たせて取るピッチャーが多いと判断できる。(他に良い意味付けがあるかもしれない。各自考えてみてください) ③ $(3.5278+1.255+0.8712)/7 \fallingdotseq 0.81$. ④ 西武ライオンズ ⑤ 本塁打（ホームラン）を打たれずらい.
[6] ① a. ② c. ③ $\frac{148-149}{7.2} = -0.139$. ④ $-0.139\times0.53+0.388\times(-0.35)+(-0.039)\times(-0.64)+(-0.333)\times0.44 = -0.331$ ⑤ $\frac{3.488+0.355+0.092}{4} = 0.984$. ⑥ 17. (補助線の図は略)
[7] ①

$$\begin{pmatrix} 1 & x \end{pmatrix}\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ x \end{pmatrix} = \begin{pmatrix} 1+2x & 2+x \end{pmatrix}\begin{pmatrix} 1 \\ x \end{pmatrix}$$
$$= 1+2x+(2+x)x = x^2+4x+1$$

② $f(x) = x^2+4x+1$ より $\frac{df(x)}{dx} = 2x+4 = 0$ より $x = -2$ の時, 最小値 $f(-2) = -3$.

[4] の生データ:勝率は高い方が良いが, 一方で防御率は小さい数値の方が良い. 主成分分析する上では, 全ての変数（防御率や勝率など）の値が大きいほど良いということにまとめた方が後々見やすいので、防御率、被安打、被本塁打、四球、死球にはマイナスをつけておく． ここで三振とは、三振を奪った数を意味する. そうすると防御率も大きな数値ほど良い数字であるということになる.

チーム	防御率	勝率	被安打	被本塁打	三振	四球	死球
阪　神	-3.29	0.582	-1261	-85	994	-386	-59
巨　人	-3.37	0.596	-1210	-115	1115	-351	-56
中　日	-3.53	0.511	-1307	-114	989	-336	-57
ヤクルト	-3.75	0.471	-1260	-140	880	-374	-57
広　島	-3.78	0.496	-1286	-121	897	-443	-43
横　浜	-4.74	0.338	-1366	-168	858	-412	-80
日本ハム	-3.54	0.514	-1196	-147	882	-395	-51
西　武	-3.86	0.543	-1324	-114	944	-449	-68
楽　天	-3.89	0.461	-1283	-104	1064	-456	-64
オリックス	-3.93	0.524	-1320	-130	923	-329	-58
ソフトバンク	-4.05	0.454	-1295	-128	1087	-420	-54
ロッテ	-4.14	0.51	-1326	-114	915	-363	-47

主成分と固有値

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
防御率	-0.50	0.07	0.08	-0.13	-0.33	0.42	0.66
勝率	-0.49	0.03	-0.16	0.12	-0.33	-0.78	0.00
被安打	-0.39	0.21	0.50	-0.52	-0.02	0.07	-0.53
被本塁打	-0.39	-0.45	-0.22	0.44	-0.16	0.40	-0.47
三振	-0.29	-0.56	-0.17	-0.40	0.61	-0.13	0.17
四球	-0.16	0.56	-0.74	-0.19	0.18	0.17	-0.15
死球	-0.32	0.35	0.31	0.56	0.59	-0.01	0.10
固有値	3.5278	1.255	0.8712	0.6924	0.4892	0.1173	0.0472

主成分スコア

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
阪　神	-2.36	-0.76	-0.47	0.25	-1	0.38	-0.18
巨　人	-2.87	-0.11	-0.41	-1.45	0.42	-0.45	-0.07
中　日	-0.82	0.37	-1.31	0.01	0.12	0.54	0.4
ヤクルト	0.4	1.32	0.41	-0.37	-0.27	0.35	-0.02
広　島	-0.17	0.29	1.51	1.46	0.21	0.02	0.22
横　浜	5.16	-0.02	-0.34	-0.8	-0.24	0.05	-0.08
日本ハム	-0.74	1.75	1.67	-0.77	-0.36	-0.08	-0.07
西　武	0.49	-1.37	0.06	0.45	-1.26	-0.62	0.21
楽　天	0.11	-2.19	0.61	-0.24	0.2	0.38	-0.23
オリックス	0.26	1.09	-1.42	0.2	0.01	-0.28	0.04
ソフトバンク	0.36	-1.05	0.38	-0.25	1.52	-0.11	0.19
ロッテ	0.17	0.69	-0.68	1.51	0.65	-0.18	-0.41

引用先:日本野球機構

http://bis.npb.or.jp/2008/stats/tmp_c.html

[5] の生データ

生徒番号	身長	体重	胸囲	座高	生徒番号	身長	体重	胸囲	座高
1	148	41	72	78	16	139	34	71	76
2	160	49	77	86	17	149	26	67	79
3	159	45	80	86	18	142	31	66	76
4	153	43	76	83	19	150	43	77	79
5	151	42	77	80	20	139	31	68	74
6	140	29	64	74	21	161	47	78	84
7	158	49	78	83	22	140	33	67	77
8	137	31	66	73	23	152	35	73	79
9	149	47	82	79	24	145	35	70	77
10	160	47	74	87	25	156	44	78	85
11	151	42	73	82	26	147	38	73	78
12	157	39	68	82	27	147	30	65	75
13	157	48	80	88	28	151	36	74	80
14	144	36	68	76	29	141	30	67	76
15	139	32	68	73	30	148	38	70	78

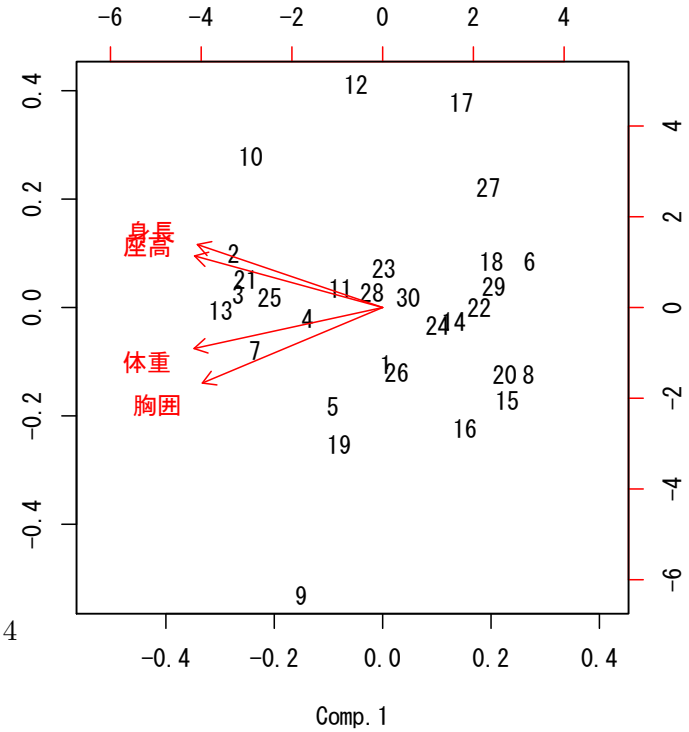
主成分スコア

	Comp.1	Comp.2	Comp.3	Comp.4		Comp.1	Comp.2	Comp.3	Comp.4
1	0.07	?	-0.41	0.12	16	1.55	-0.71	0.17	-0.35
2	-2.81	0.34	-0.34	-0.07	17	1.49	1.25	0.82	0.08
3	-2.73	0.1	0.47	-0.05	18	2.05	0.29	-0.05	-0.15
4	-1.42	-0.05	0.08	-0.19	19	-0.82	-0.81	-0.01	0.23
5	-0.94	-0.58	0.16	0.17	20	2.31	-0.39	0.08	-0.05
6	2.77	0.29	-0.15	-0.03	21	-2.58	0.19	-0.08	0.41
7	-2.41	-0.24	-0.37	0.28	22	1.82	0.02	-0.13	-0.51
8	2.76	-0.39	-0.21	-0.12	23	0.02	0.25	0.47	0.41
9	-1.54	-1.72	0.1	0.21	24	1.04	-0.09	0	0.02
10	-2.49	0.93	-0.4	-0.29	25	-2.13	0.07	0.3	-0.2
11	-0.79	0.13	-0.21	-0.26	26	0.26	-0.37	0.06	0.07
12	-0.5	1.36	-0.41	0.22	27	2	0.74	-0.08	0.48
13	-3.05	0	0.21	-0.61	28	-0.2	0.11	0.51	0.16
14	1.34	-0.07	-0.39	0.03	29	2.1	0.14	0.18	-0.21
15	2.35	-0.54	-0.09	0.11	30	0.48	0.08	-0.29	0.1

主成分（ベクトル）と固有値

	Comp.1	Comp.2	Comp.3	Comp.4
身長	-0.5	0.53	0.03	0.68
体重	-0.51	-0.35	-0.79	-0.07
胸囲	-0.49	-0.64	0.59	0.12
座高	-0.51	0.44	0.2	-0.72
固有値	3.488	0.355	0.092	0.066

- 右の図は biplot:グラフの中の数字は、生データの生徒番号に相当する.
- データ引用先:http://case.f7.ems.okayama-u.ac.jp/statedu/hbw2-book/node78.html
- 固有値を全て足したものが 4 にならないが、これはコンピューターの計算上の誤差なので、気にしないこと.



4