

主成分分析による次元縮約

データを圧縮して,関係を見よう!

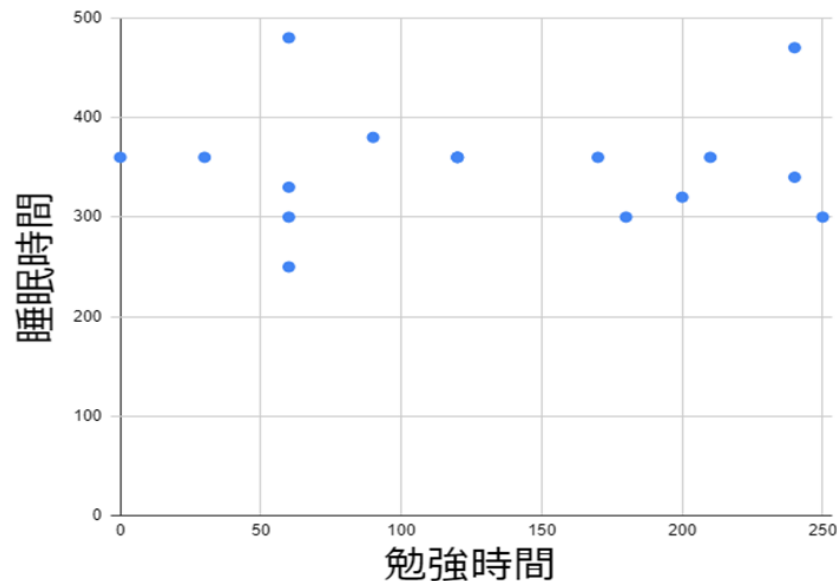
情報1で学んだこと

データ

01_睡眠	06_学業
467	355
469	368
458	453
448	373
467	373
477	377
⋮	⋮
⋮	⋮
⋮	⋮



散布図



2変数の関係を探る!

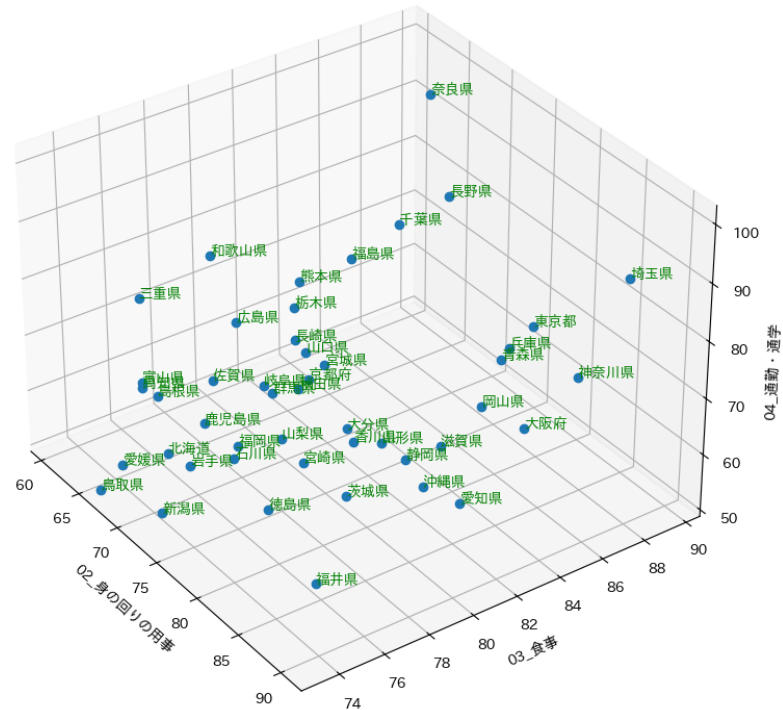
データがたくさんあったら・・・？

データ(3変数)

	02_身の回りの用事	03_食事	04_通勤・通学
01_北海道	72	74	62
02_青森県	71	90	53
03_岩手県	69	76	53
04_宮城県	75	80	72
05_秋田県	61	84	45
06_山形県	74	83	52
07_福島県	60	87	64
08_茨城県	88	76	71
09_栃木県	66	82	69
10_群馬県	74	78	69
11_埼玉県	83	91	80
12_千葉県	71	85	85
13_東京都	84	86	80
14_神奈川県	89	86	77
15_新潟県	71	74	50
16_富山県	66	75	66
17_石川県	69	78	51
18_福井県	87	75	56
19_山梨県	75	78	62
20_長野県	62	91	72
21_岐阜県	73	78	69
22_静岡県	82	81	62
23_愛知県	93	79	71
24_三重県	66	75	81



3次元散布図



データがもったくさんあったら…?

データ

都道府県	11.睡眠	02.身の回りの回	03.食事	04.通勤・通学	05.仕事	06.学業	07.家事	10.買い物(通勤・通学・ラ)	11.移動(通学・通勤)	12.テレビ・ラジオ・新聞・雑誌	13.休養・趣味・娯楽	14.学習・読書・勉強	15.運動	16.スポーツ	20.その他
01.北海道	487	72	74	62	41	355	5	3	19	25	166	42	63	18	
02.青森県	469	71	90	53	34	368	9	3	13	79	129	19	74	30	
03.岩手県	458	69	79	53	47	453	5	3	9	31	105	42	58	14	
04.宮城県	448	75	80	72	44	373	3	3	7	16	83	29	119	53	
05.秋田県	467	61	64	45	29	373	4	6	13	19	136	32	114	34	
06.山形県	477	74	63	52	43	377	4	5	6	40	135	30	58	29	
07.福島県	496	80	87	64	41	436	7	2	15	23	146	36	46	25	
08.茨城県	456	66	76	71	31	397	8	6	18	24	133	41	46	35	
09.栃木県	472	66	62	69	46	384	13	12	9	24	131	13	72	20	
10.群馬県	457	74	79	69	41	398	7	7	12	31	125	36	67	22	
11.埼玉県	439	63	91	69	31	403	5	3	12	29	134	55	41	13	
12.千葉県	437	71	85	85	70	367	8	3	12	15	170	54	33	17	
13.東京都	451	84	96	60	33	397	5	7	11	16	124	55	56	36	
14.神奈川県	452	89	88	77	32	399	2	6	14	39	109	23	78	9	
15.新潟県	475	71	74	50	13	381	7	11	8	49	143	46	62	21	
16.富山県	476	66	75	66	60	376	3	4	10	35	125	29	64	15	
17.石川県	460	69	79	51	48	395	3	1	16	16	135	31	79	43	
18.福井県	445	67	75	56	36	450	3	4	10	12	126	54	45	25	
19.山梨県	453	76	79	62	39	429	5	1	16	30	84	60	66	14	
20.長野県	463	62	91	72	43	402	10	1	6	28	150	30	53	17	
21.岐阜県	458	79	79	69	39	412	10	6	9	45	101	26	57	23	
22.静岡県	452	62	61	62	48	390	8	3	11	21	136	41	54	23	
23.愛知県	459	63	79	71	72	374	4	1	12	25	123	44	51	13	
24.三重県	437	66	75	61	67	350	3	2	6	21	171	42	41	23	
25.滋賀県	462	61	63	60	47	398	5	4	6	12	144	46	47	27	
26.京都府	478	65	83	53	23	360	17	10	14	27	159	34	53	15	
27.大阪府	430	66	64	78	65	366	6	16	23	147	66	61	22	22	
28.兵庫県	465	79	67	69	45	351	3	6	12	33	161	40	48	19	
29.奈良県	446	72	66	107	10	416	7	15	9	41	79	36	100	13	
30.和歌山県	467	59	61	71	40	403	7	6	14	27	139	31	45	23	
31.鳥取県	464	69	72	55	35	405	9	4	13	37	136	38	27	31	
32.徳島県	449	69	75	69	44	405	9	14	12	152	51	27	19	19	
33.香川県	424	68	62	77	40	405	4	2	21	14	135	49	40	11	
34.広島県	450	64	80	67	27	433	18	6	7	20	132	26	71	17	
35.岡山県	432	70	61	67	39	362	7	6	23	63	84	63	32	32	
36.広島県	451	76	77	52	65	401	3	3	9	31	111	61	64	18	



相関行列や散布図行列

	01.睡眠	02.身の回りの回	03.食事	04.通勤・通学	05.仕事	06.学業	07.家事	10.買い物(通勤・通学・ラ)	11.移動(通学・通勤)	12.テレビ・ラジオ・新聞・雑誌	13.休養・趣味・娯楽	14.学習・読書・勉強	15.運動	16.スポーツ	20.その他
01.睡眠															
02.身の回りの回	-0.2308														
03.食事	-0.129	0.04927													
04.通勤・通学	-0.3563	0.19467	0.28387												
05.仕事	-0.3274	0.08675	-0.1459	0.02845											
06.学業	-0.3835	-0.1301	-0.2055	-0.0192	0.01618										
07.家事	0.03418	-0.2337	0.04911	0.06168											
10.買い物(通勤・通学・ラ)	0.26421	0.02212	0.0845	0.12755	-0.107										
11.移動(通学・通勤)	-0.1549	0.14999	0.02546	-0.0222											
12.テレビ・ラジオ・新聞・雑誌	-0.35748	-0.1084	0.11338	-0.0222											
13.休養・趣味・娯楽	0.09791	-0.1746	0.04584	-0.1438											
14.学習・読書・勉強	-0.1175	0.37528	-0.0702	0.03141	0										
15.運動	0.12724	-0.1379	0.11829	-0.052											
16.スポーツ	0.0383	-0.0809	-0.0632	-0.2982	0										
20.その他	0.10732	-0.0867	-0.2132	-0.3353	0										

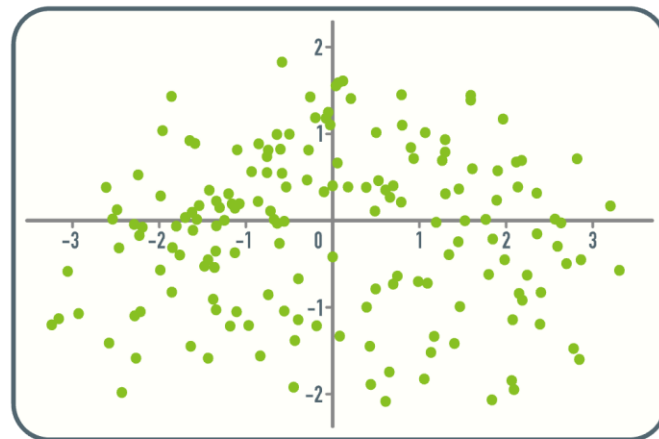
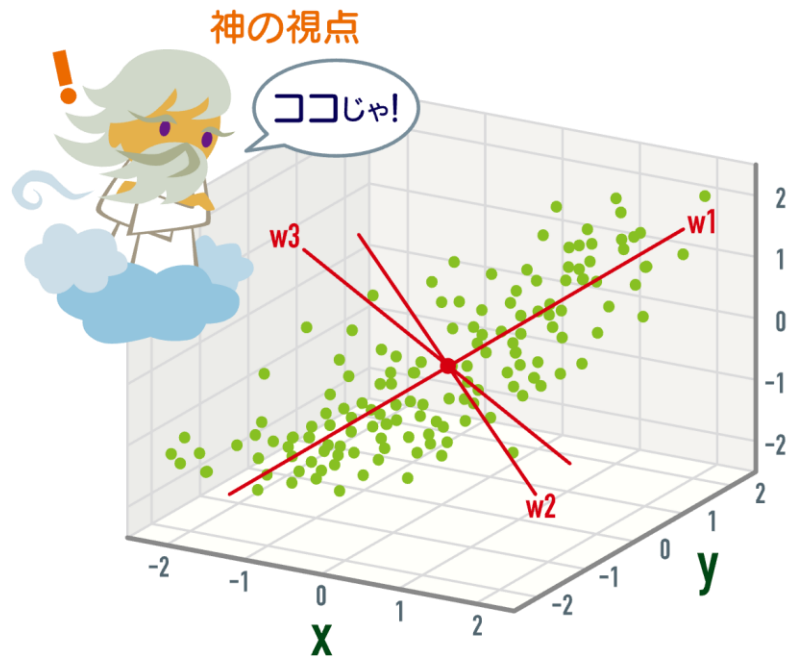


結局全体的には何を言えばいいの？

→こんな時に役立つ1つの方法が、**主成分分析**

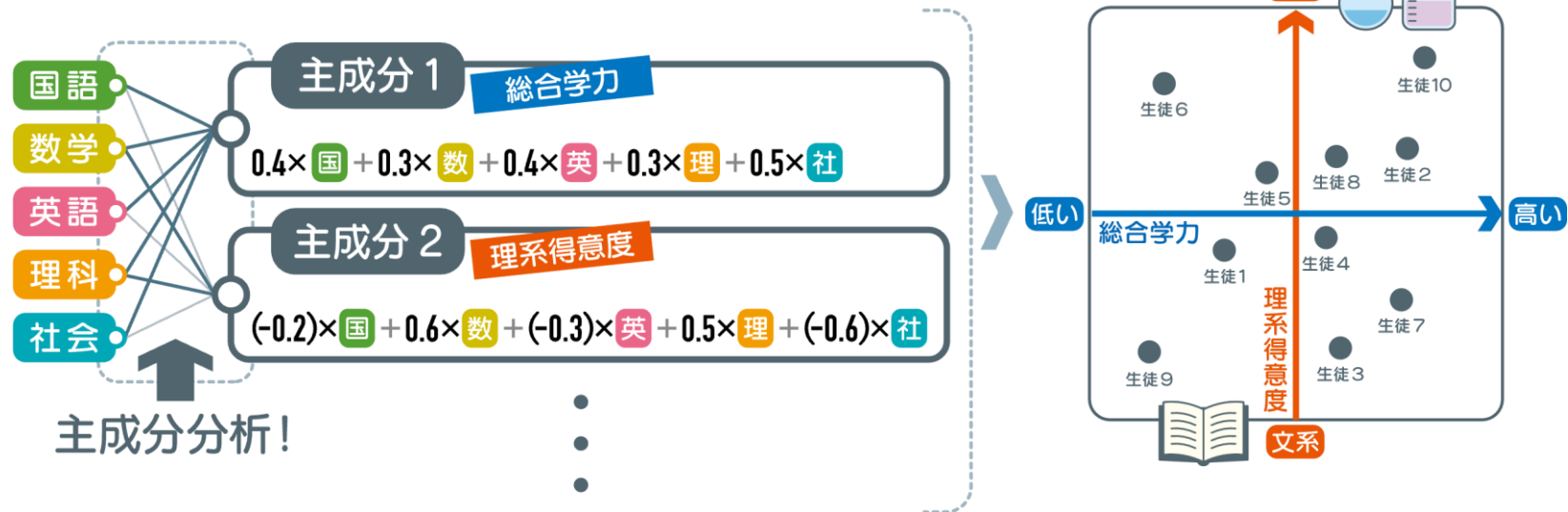
圧縮 (要約) のイメージ

散らばりを最大化するように、空間上で軸を決める。
(一番散らばって見えるところで、見下ろすイメージ)



主成分分析

多次元のデータを低次元の主成分に圧縮（要約）することで、データの構造を可視化できる。



今回使用する題材



<https://www.nstac.go.jp/use/literacy/ssdse/>

	01 穀類	02 魚介*	03 肉類	牛乳	卵	05 野菜	06 果物	08 菓子類
全国	79597	75171	97501	15285	10250	108315	40424	89367
北海道	80069	83434	91587	12873	9054	108284	43848	89616
青森県	74692	87151	89759	13751	7876	112569	38445	82689
岩手県	79493	77147	81637	16556	8921	116270	44146	94418
宮城県	76748	80756	86259	14886	10044	119365	44600	96002
秋田県	70765	83200	84768	13326	9326	120467	47823	88991
山形県	81774	75220	104651	16811	10001	116856	49520	95511
福島県	76382	76925	85656	15018	12258	119124	49191	94925
茨城県	71011	67538	79669	13656	9485	103200	42808	92725
栃木県	76722	70647	85784	14338	9254	110330	42586	93030
群馬県	82055	72087	77366	14965	8706	109748	46874	90262
埼玉県	82383	73742	98466	16105	9845	118512	42548	101101
千葉県	82395	80432	93747	17247	10308	126613	46139	95294
東京都	81898	82370	106706	16524	10137	129296	46646	96423
神奈川県	86334	81864	105787	16034	10637	131479	45025	95843
新潟県	86628	75021	86221	15470	10445	126203	42360	90363
富山県	86991	88471	91485	15420	8888	119643	44898	95398
石川県	87677	76898	103972	16577	9948	111969	40422	105512

SSDSE-家計消費 (SSDSE-C)

- 都道府県庁所在市別の家計消費データを集めたデータセットのうち、一部。
- 都道府県庁所在市別 (→都道府県に読み替え)
- 二人以上の世帯の1世帯当たり (≒人口は関係なし)
- 品目別 (食料の全品目) 年間支出金額を収録。

課題「都道府県ごとに、各食品の消費にはどんな傾向があるだろう？」



今回使用する環境

オープンソースの統計解析向けプログラミング言語



コマンドの解説(一時停止して見て下さい)

```
youhi <- read.csv("youhi.csv",h=T,row.names=1)
```

変数「youhi」に「youhi.csv」ファイルを読み込む。

h=T ヘッダが存在する。 / row.names=1 1列目のデータを行の名前にする。

```
result <- prcomp(youhi,scale=T)
```

prcomp関数で**主成分分析**を実行する(主成分分析= Principal component analysis)。

```
plot(result$x)
```

result\$xの値を使ってグラフを表示する。横軸が主成分1,縦軸が主成分2となる。

```
text(result$x,rownames(youhi),pos=1,cex=1.4)
```

ラベルを追加する。

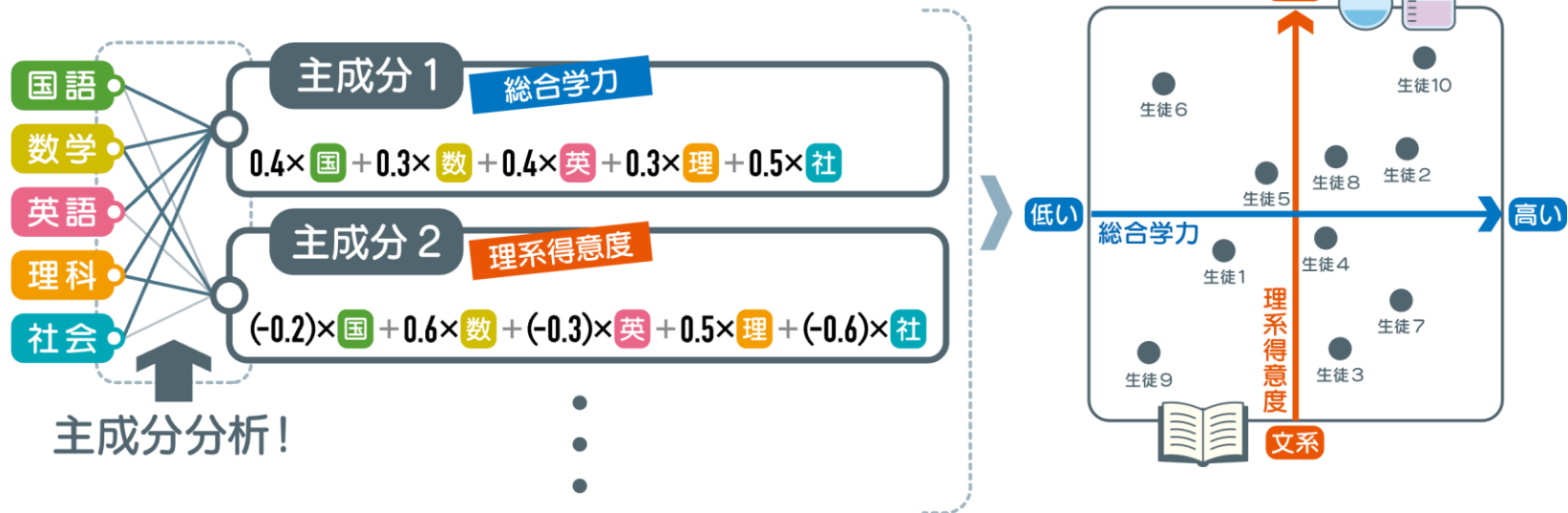
rownames(youhi) → 各データの名前(青森県,……)

pos=1 → ラベルの位置(下)。他の数字指定で,表示位置を変更できる。

cex=1.4 → テキストの表示倍率。文字サイズ(Character Expansion)の頭文字。

復習 主成分分析の目的を思い出そう

多次元のデータを低次元の主成分に圧縮（要約）することで、データの構造を可視化できる。



Rotation (n x k) = (8 x 8):

	PC1	PC2	PC3	PC4	PC5	PC6
X01.穀類	-0.38439520	-0.17464131	0.01514826	-0.42250144	0.697522761	-0.25335258
X02.魚介類	-0.40105134	0.17473525	0.23626920	-0.45283487	-0.496366787	-0.01364726
X03.肉類	-0.10261174	-0.64783252	-0.12590410	-0.30586513	-0.418862100	0.16880933
牛乳	-0.34313981	-0.28594508	-0.62325376	0.33165299	-0.112218298	-0.40510779
卵	-0.06460404	-0.55773922	0.66554837	0.41620098	0.059268557	-0.05112113
X05.野菜.海藻	-0.46319275	0.14350833	0.22969711	-0.06415543	0.006475559	-0.11143289
X06.果物	-0.39785698	0.32531796	0.12453565	0.43804031	-0.195380346	-0.19759489
X08.菓子類	-0.43140778	0.00412141	-0.16859943	0.20896757	0.193172024	0.83002585
	PC7	PC8				
X01.穀類	0.16556111	0.25469156				
X02.魚介類	0.49869834	-0.22907769				
X03.肉類	-0.29774318	0.40961627				
牛乳	0.15073740	-0.32035056				
卵	0.18808561	-0.16453892				
X05.野菜.海藻	-0.75204955	-0.36047664				
X06.果物	0.01345438	0.67163335				
X08.菓子類	0.10668165	-0.06432586				

結果の読み取り方

PC Principal Component (主成分)

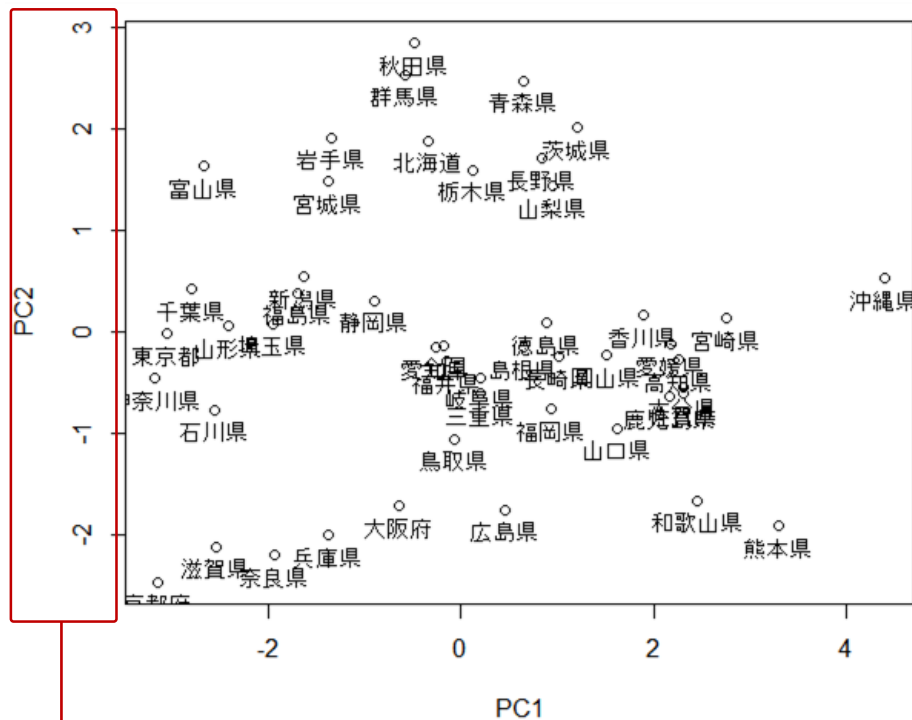
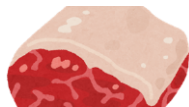
result

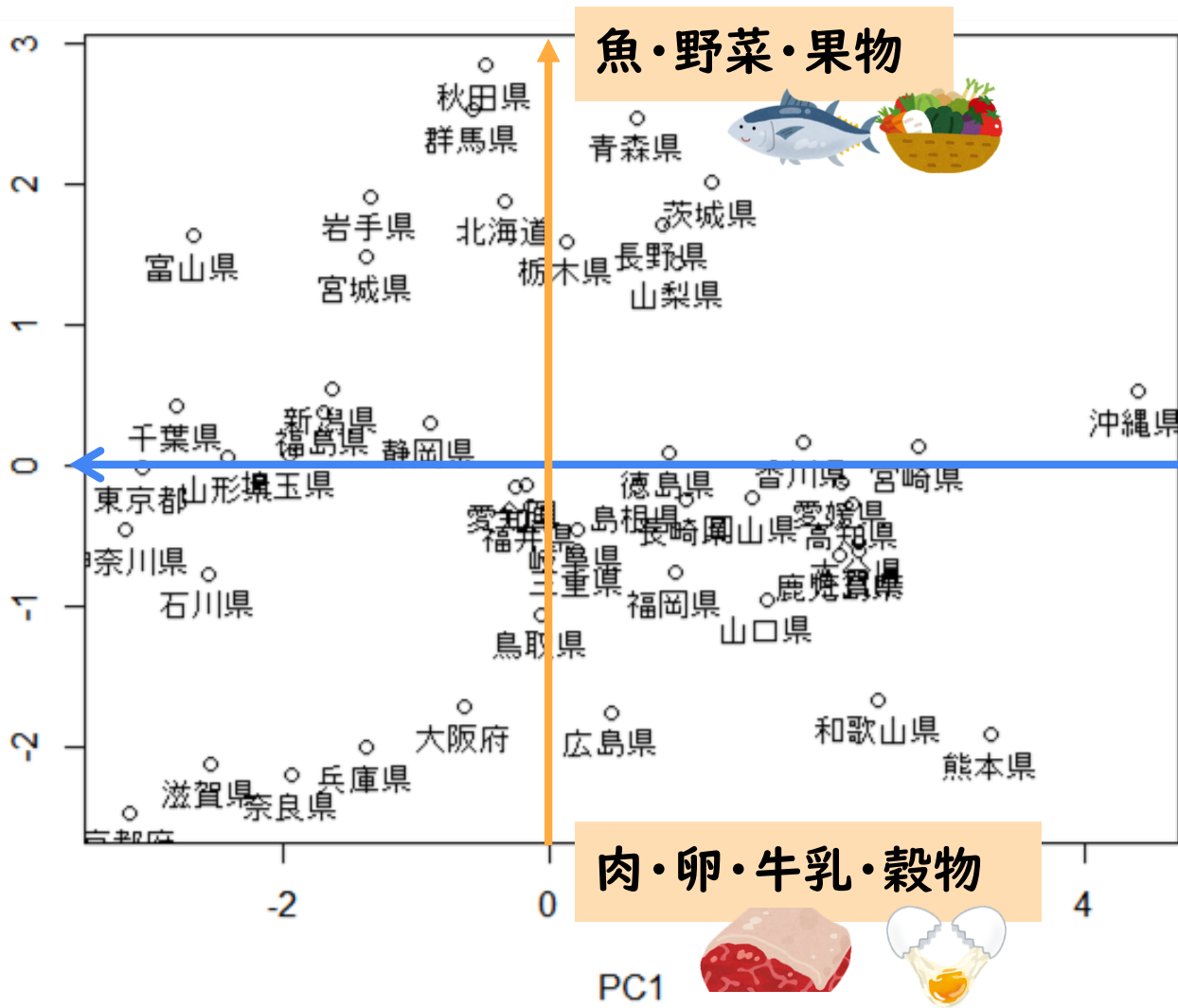
	PC1	PC2	
X01.穀類	-0.38439520	-0.17464131	0.0
X02.魚介類	-0.40105134	0.17473525	0.2
X03.肉類	-0.10261174	-0.64783252	-0.1
牛乳	-0.34313981	-0.28594508	-0.6
卵	-0.06460404	-0.55773922	0.4
X05.野菜・海藻	-0.46319275	0.14350833	0.2
X06.果物	-0.39785698	0.32531796	0.1
X08.菓子類	-0.43140778	0.00412141	-0.1

食傾向

PC1 =

$(-0.38) \times \text{穀類} + (-0.40) \times \text{魚介類} + \dots + (-0.46) \times \text{野菜・海藻} + \dots$





総合消費度
高

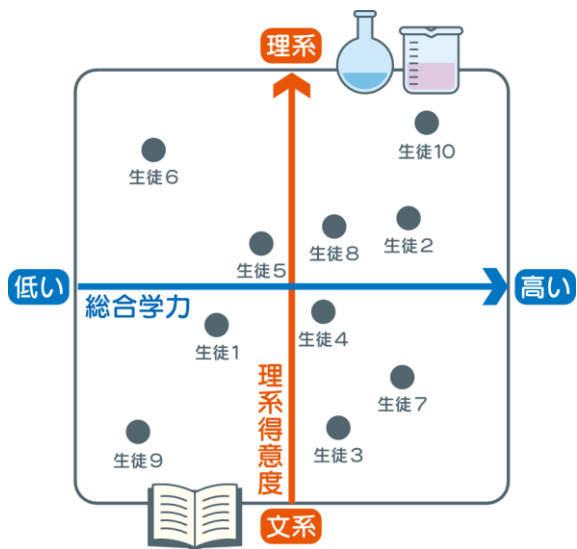
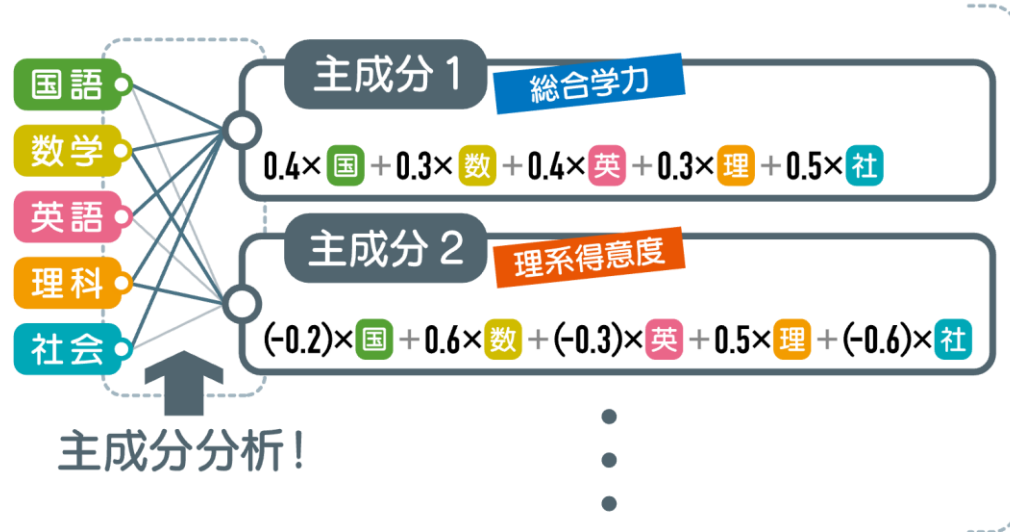
総合消費度
低

肉・卵・牛乳・穀物

PC1

復習 主成分分析の目的を思い出そう

多次元のデータを低次元の主成分に圧縮（要約）することで、データの構造を可視化できる。



2軸でどれくらい説明できているの？

Rotation (n x k) = (8 x 8):

	PC1	PC2	PC3	PC4	PC5	PC6
X01.穀類	-0.38439520	-0.17464131	0.01514826	-0.42250144	0.697522761	-0.25335258
X02.魚介類	-0.40105134	0.17473525	0.23626920	-0.45283487	-0.496366787	-0.01364726
X03.肉類	-0.10261174	-0.64783252	-0.12590410	-0.30586513	-0.418862100	0.16880933
牛乳	-0.34313981	-0.28594508	-0.62325376	0.33165299	-0.112218298	-0.40510779
卵	-0.06460404	-0.55773922	0.66554837	0.41620098	0.059268557	-0.05112113
X05.野菜.海藻	-0.46319275	0.14350833	0.22969711	-0.06415543	0.006475559	-0.11143289
X06.果物	-0.39785698	0.32531796	0.12453565	0.43804031	-0.195380346	-0.19759489
X08.菓子類	-0.43140778	0.00412141	-0.16859943	0.20896757	0.193172024	0.83002585
	PC7	PC8				
X01.穀類	0.16556111	0.25469156				
X02.魚介類	0.49869834	-0.22907769				
X03.肉類	-0.29774318	0.40961627				
牛乳	0.15073740	-0.32035056				
卵	0.18808561	-0.16453892				
X05.野菜.海藻	-0.75204955	-0.36047664				
X06.果物	0.01345438	0.67163335				
X08.菓子類	0.10668165	-0.06432586				

説明

```
> summary(result)
```

```
Importance of components:
```

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.8937	1.3262	0.83490	0.8104	0.72794	0.61549	0.47713	0.40584
Proportion of Variance	0.4483	0.2199	0.08713	0.0821	0.06624	0.04735	0.02846	0.02059
Cumulative Proportion	0.4483	0.6681	0.75526	0.8374	0.90360	0.95096	0.97941	1.00000

- Proportion of variance (寄与率)

データ全体のばらつきをどれくらい説明できているか？

- Cumulative Proportion (累積寄与率)

その成分までで、データ全体のばらつきをどれくらい説明できているか？

2軸でデータ全体の67%を説明できている！

(今回の場合、元のデータは8次元)

魚・野菜・果物

わりと東北でま
まっている

自分の都道府県。
山形県も近い。

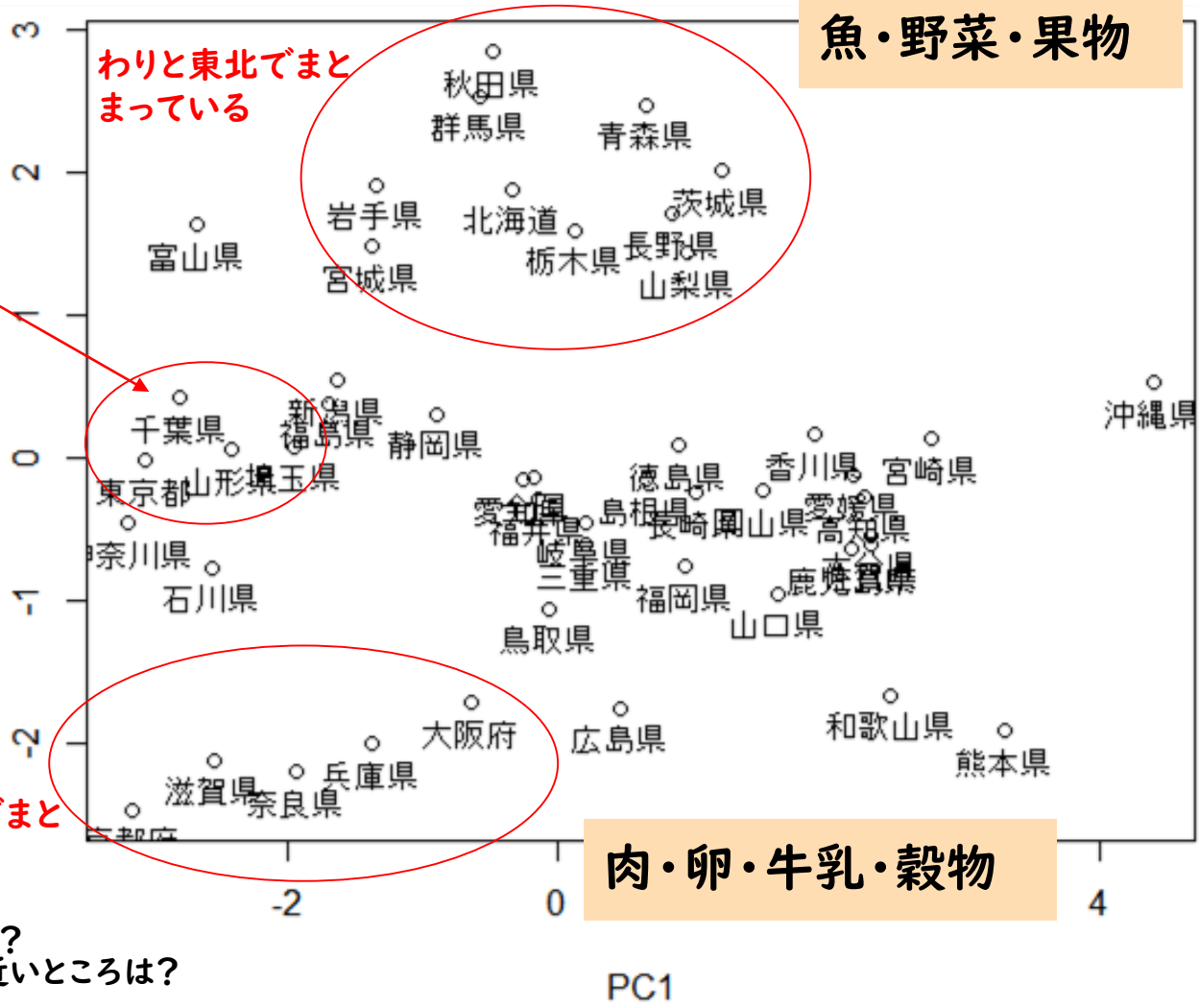
総合消費度
高

総合消費度
低

わりと関西でま
まっている

肉・卵・牛乳・穀物

地方別に見てみると？
自分の都道府県と近いところは？



PC1

おわりに

他にも様々な例でやってみよう!

- 好きなゲームのキャラクタパラメータ
- 野球選手の成績

