主成分分析による次元縮約

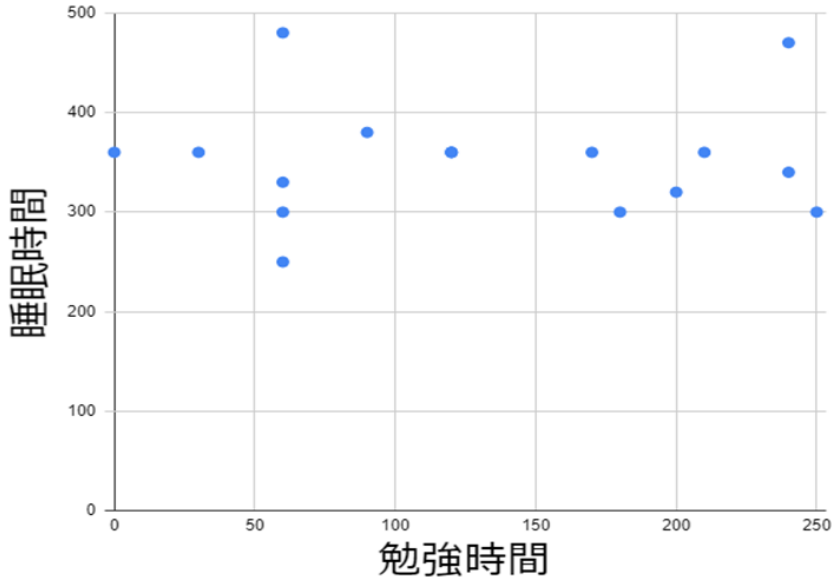
データを圧縮して,関係を見よう!

情報Ⅰ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
⋮	⋮

散布図



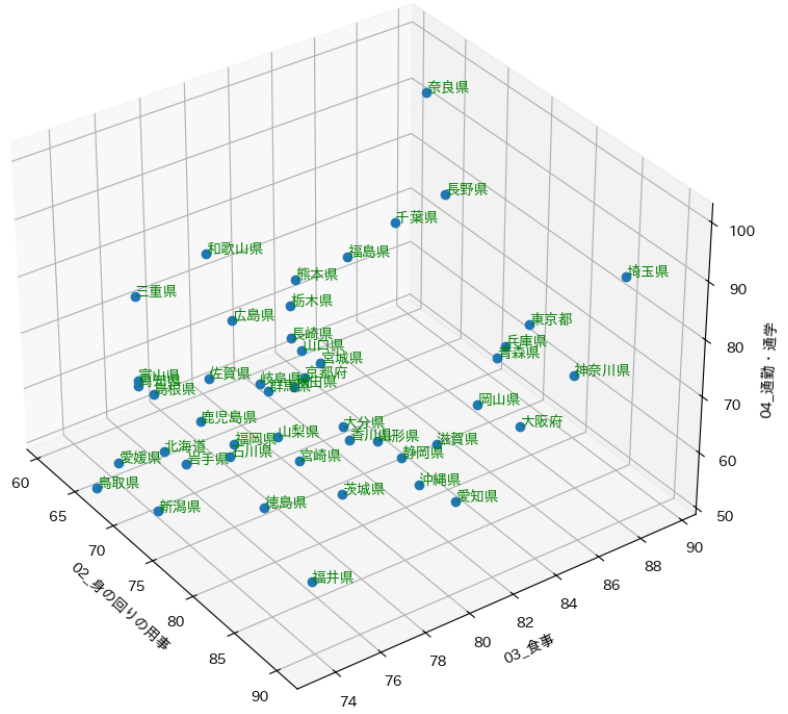
2変数の関係を探る！

データがたくさんあったら・・・？

データ(3変数)

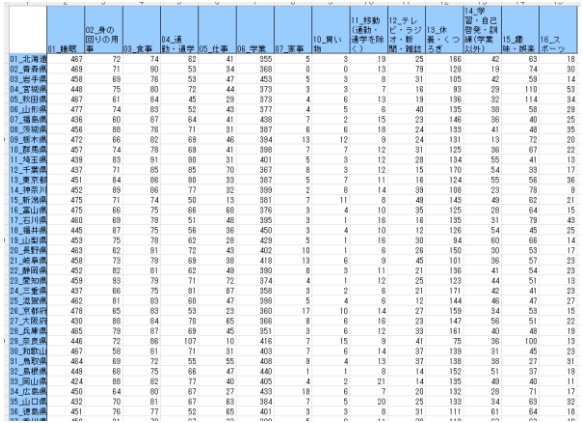
	02_身の回りの用事	03_食事	04_通勤・通学
01_北海道	72	74	62
02_青森県	71	90	53
03_岩手県	69	76	53
04_宮城県	75	80	72
05_秋田県	61	84	45
06_山形県	74	83	52
07_福島県	60	87	64
08_茨城県	88	76	71
09_栃木県	66	82	69
10_群馬県	74	78	69
11_埼玉県	83	91	80
12_千葉県	71	85	85
13_東京都	84	86	80
14_神奈川県	89	86	77
15_新潟県	71	74	50
16_富山県	66	75	66
17_石川県	69	78	51
18_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

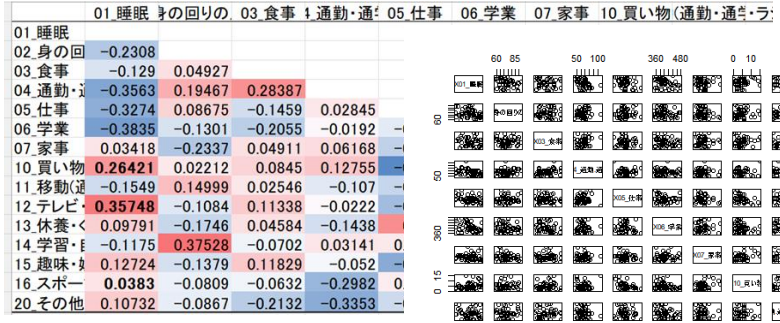


データがもっとたくさんあったら…？

データ



相関行列や散布図行列



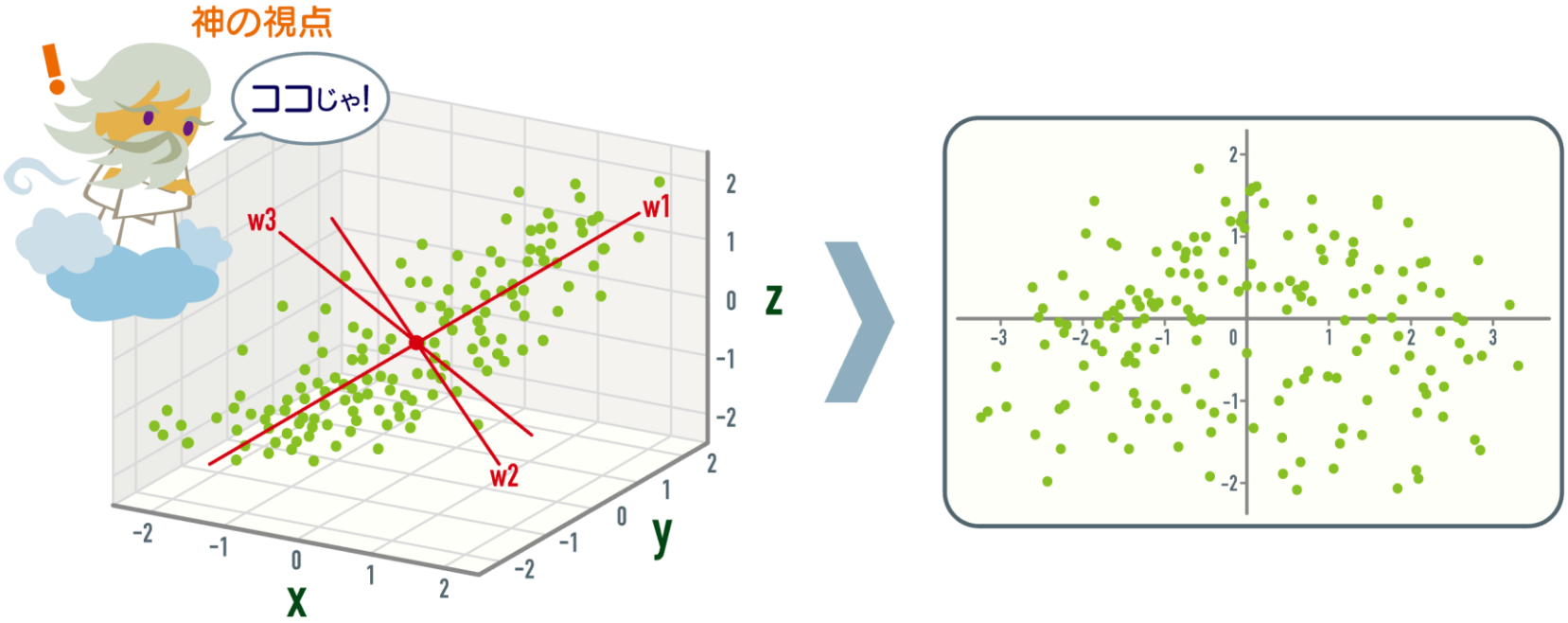
	01_睡眠	02_身の回りの	03_食事	04_通勤・通	05_仕事	06_学業	07_家事	10_買い物(通勤・通	11_移動(通	12_テレビ	13_休養・く	14_学習・自	15_趣味・娯	16_スポー	20_その他
01_睡眠															
02_身の回	−0.2308														
03_食事	−0.129	0.04927													
04_通勤・	−0.3563	0.19467	0.28387												
05_仕事	−0.3274	0.08675	−0.1459	0.02845											
06_学業	−0.3835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

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**

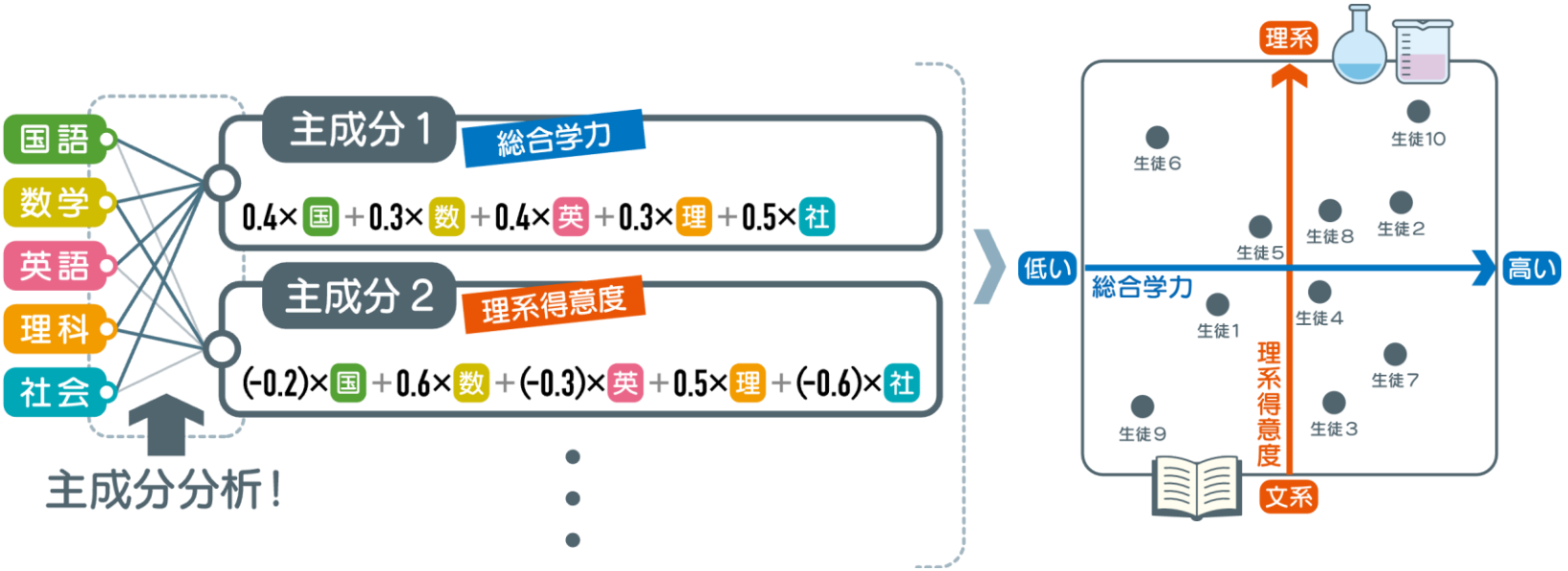
圧縮（要約）のイメージ

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

独立行政法人
統計センター

SSDSE 教育用標準データセット

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

	01 穀類	02 魚介類	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)

syouhi <- read.csv(“syouhi.csv",h=T,row.names=1)

変数「syouhi」に「syouhi.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(syouhi,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)

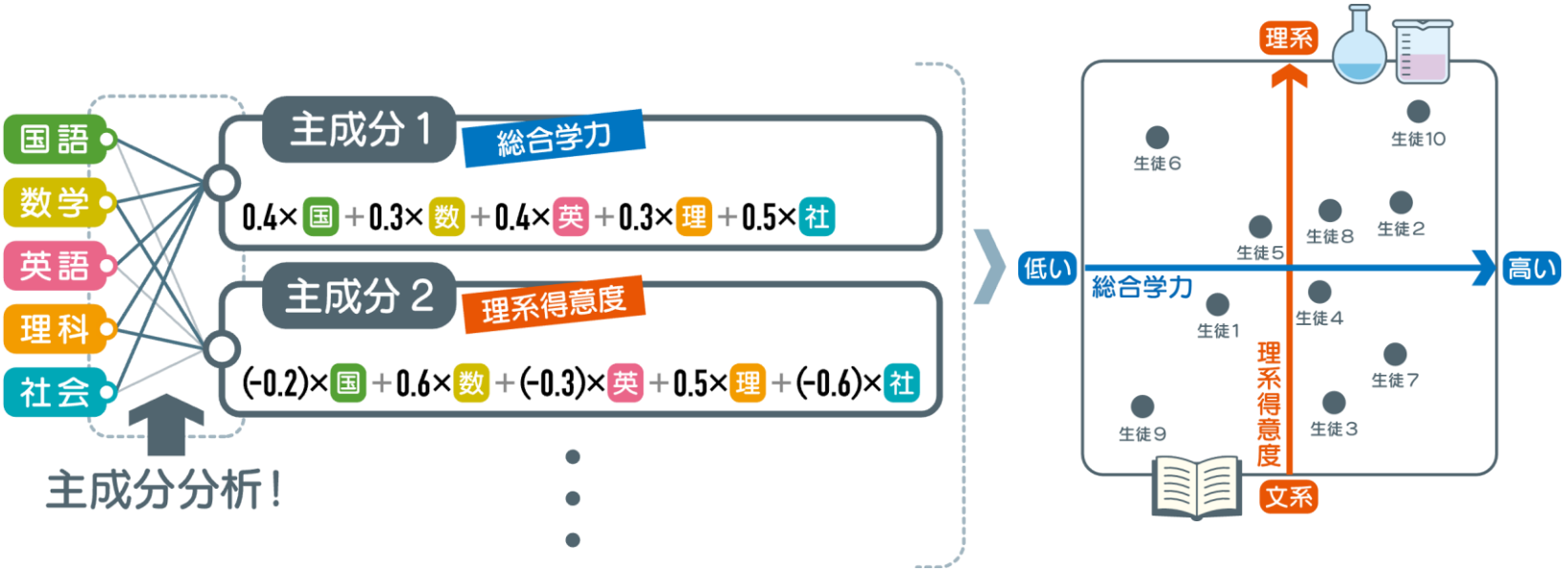
result$xの値を使ってグラフを表示する。横軸が主成分1,縦軸が主成分2となる。

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

ラベルを追加する。
rownames(syouhi) → 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。



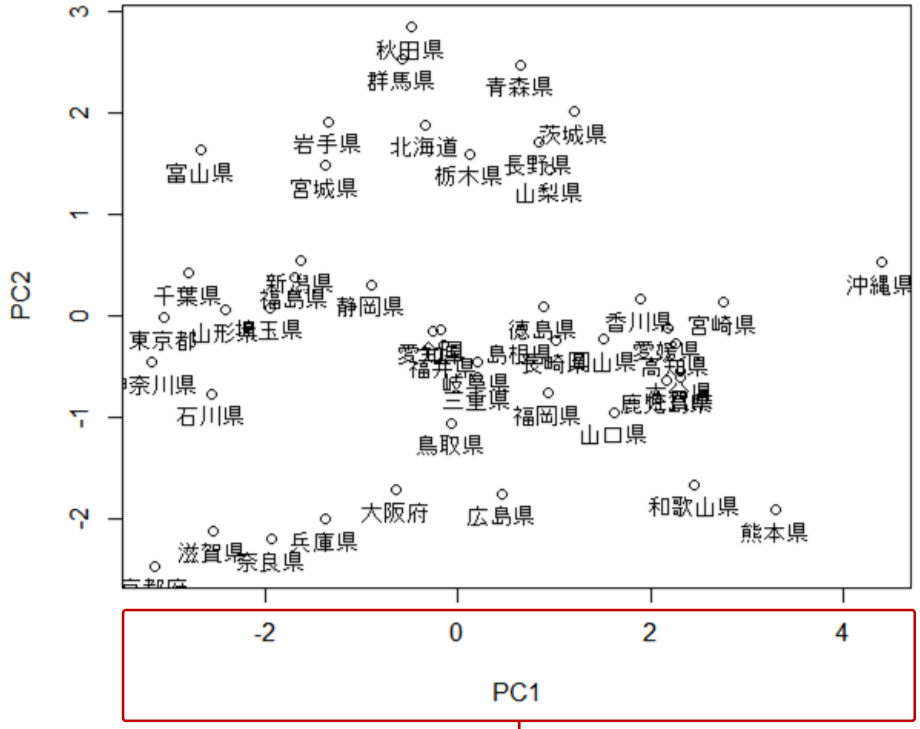
```
Rotation (n x k) = (8 x 8):
                          PC1         PC2         PC3         PC4          PC5         PC6
X01.穀類          -0.38439520 -0.17464131  0.01514826 -0.42250144  0.697522761 -0.25335258
X02.魚介類        -0.40105134  0.17473525  0.23626920 -0.45283487 -0.496366787 -0.01364726
X03.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.(
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.(
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



総合消費度

PC１＝

(-0.38)×穀類 ＋ (-0.40)× 魚介類 ＋…… ＋ (-0.46)×野菜・海藻 ＋……

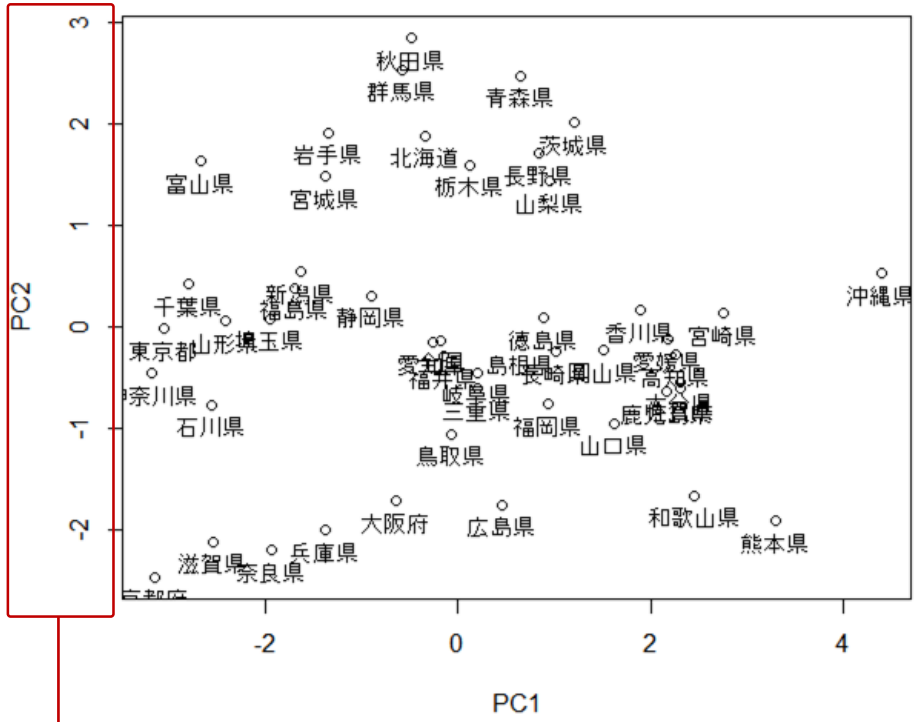


結果の読み取り方

PC Principal Component（主成分）

result

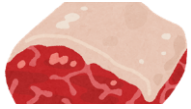
	PC1	PC2	
X01.穀類	-0.38439520	-0.17464131	0.(
X02.魚介類	-0.40105134	0.17473525	0.2
X03.肉類	-0.10261174	-0.64783252	-0.1
牛乳	-0.34313981	-0.28594508	-0.6
卵	-0.06460404	-0.55773922	0.(
X05.野菜.海藻	-0.46319275	0.14350833	0.2
X06.果物	-0.39785698	0.32531796	0.1
X08.菓子類	-0.43140778	0.00412141	-0.1

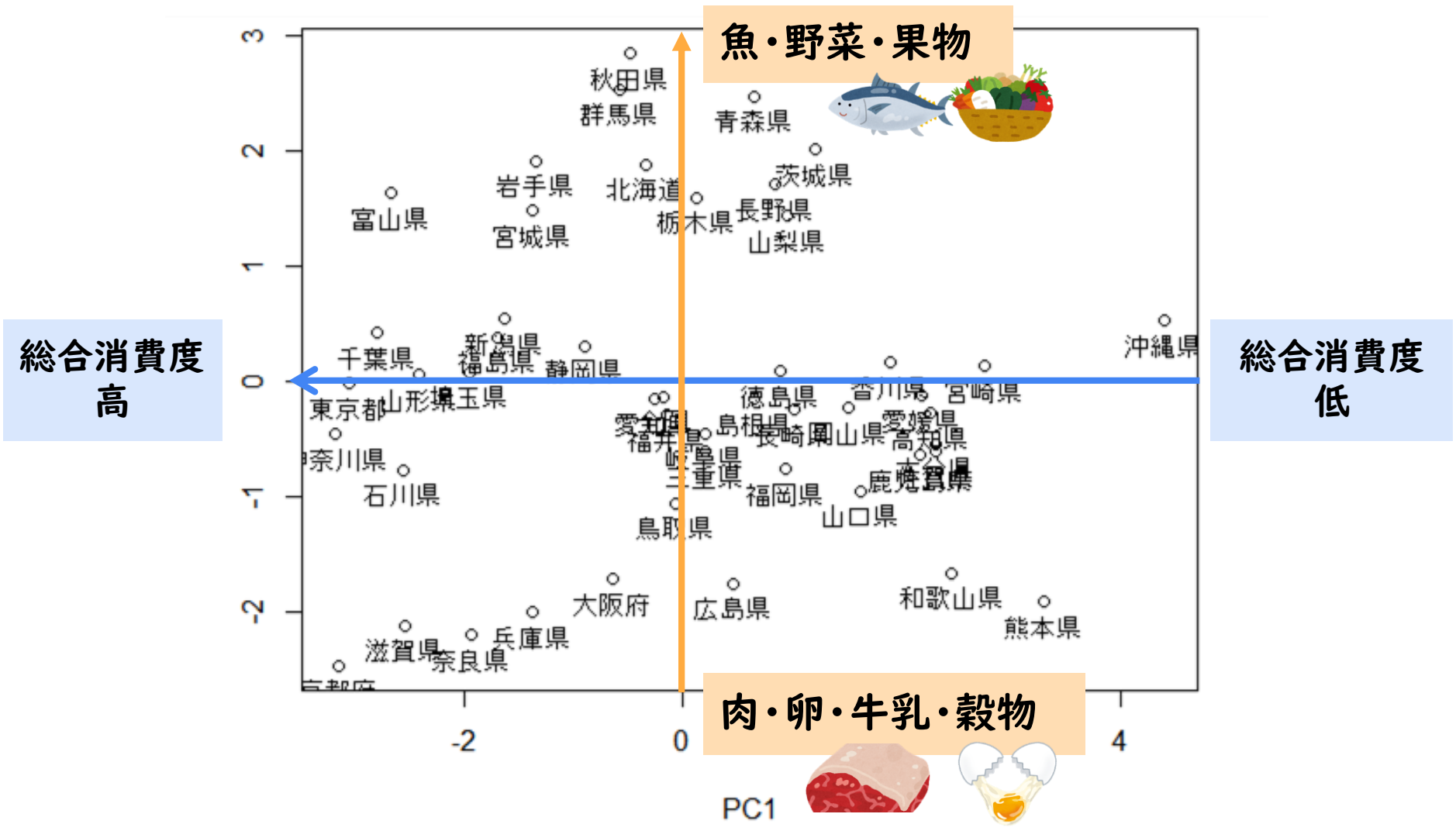


食傾向

PC１＝

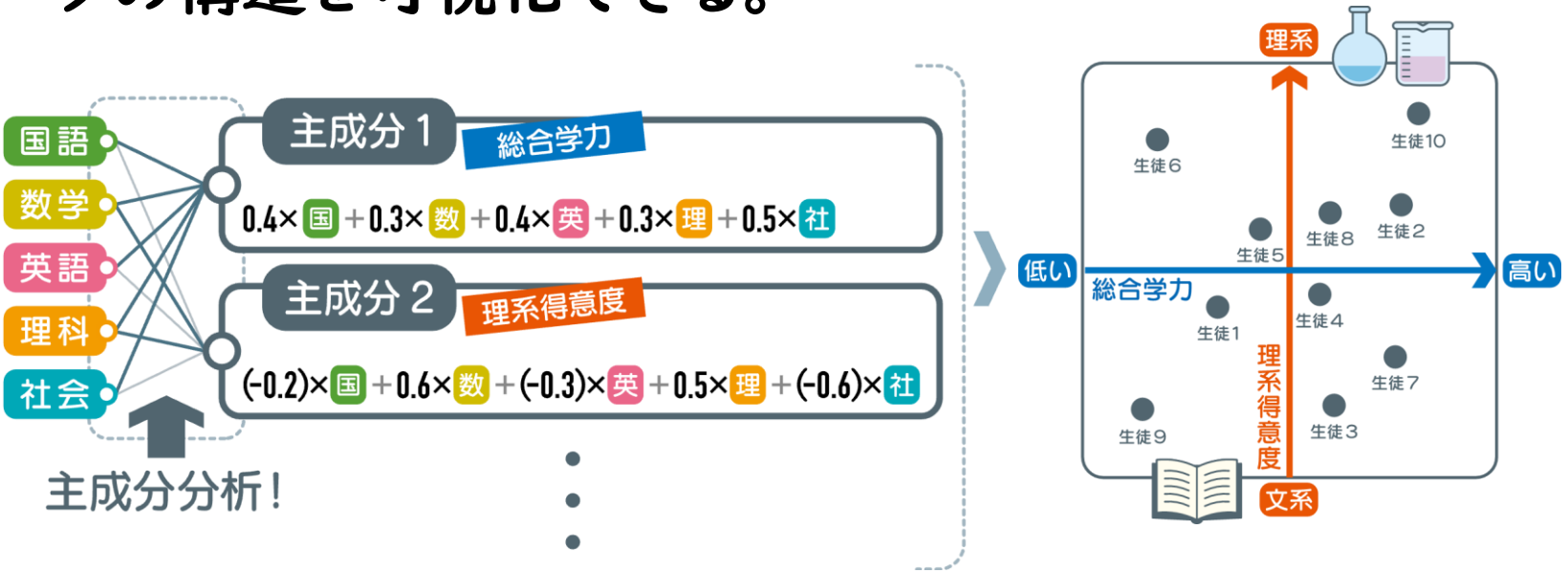
(-0.38)×穀類 ＋ (-0.40)× 魚介類 ＋…… ＋ (-0.46)×野菜・海藻 ＋……


魚・野菜・果物
総合消費度高
総合消費度低
肉・卵・牛乳・穀物
PC1
秋田県
群馬県
青森県
茨城県
岩手県
北海道
富山県
栃木県
長野県
宮城県
山梨県
千葉県
新潟県
福島県
静岡県
沖縄県
東京都
山形県
埼玉県
徳島県
香川県
宮崎県
愛媛県
島根県
長崎県
岡山県
高知県
神奈川県
石川県
三重県
福岡県
鹿児島県
山口県
鳥取県
大阪府
広島県
和歌山県
熊本県
滋賀県
奈良県
兵庫県
京都府
3
2
1
0
-1
-2
-2
0
4

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
```

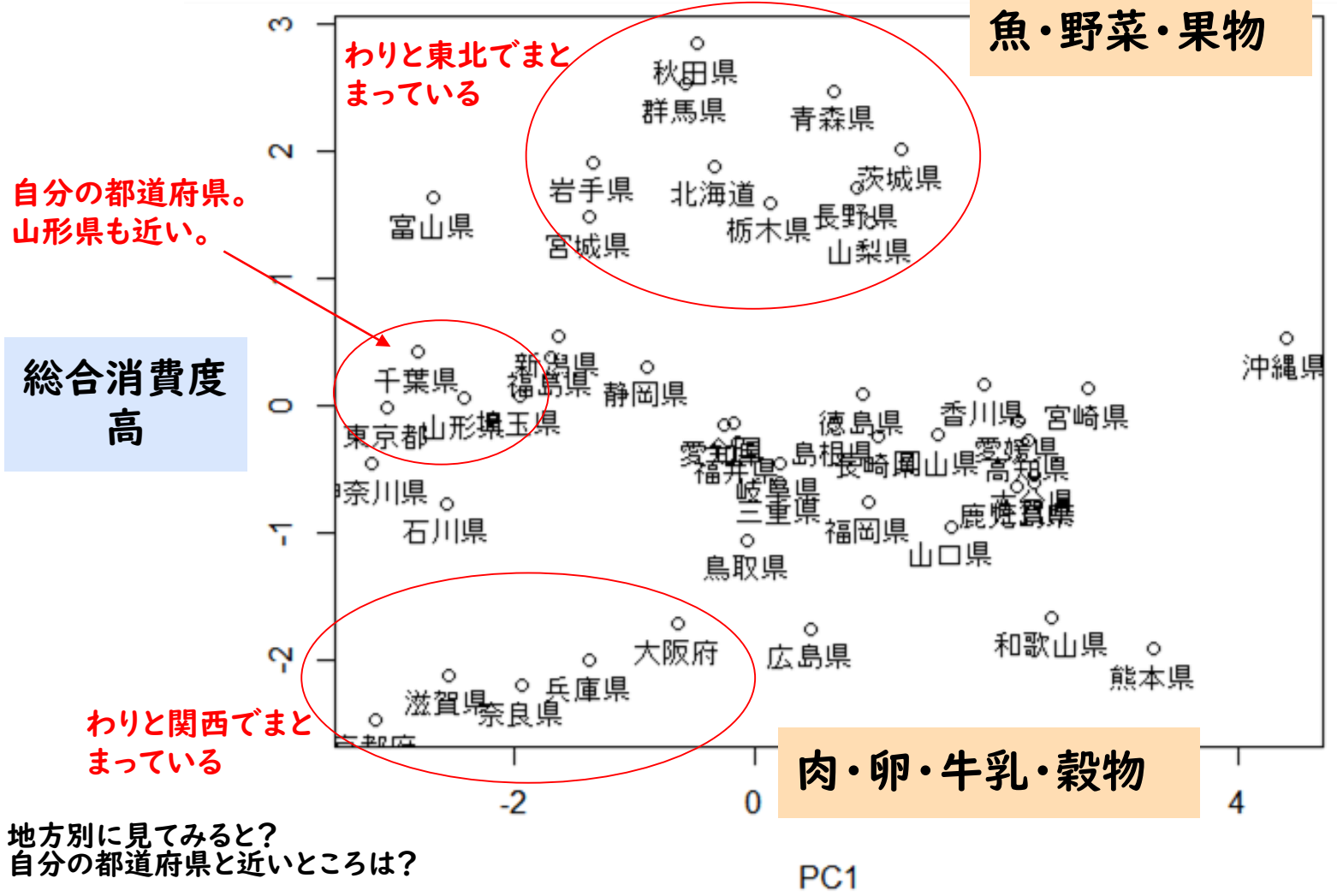
説明

```
> 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）



地方別に見てみると?
自分の都道府県と近いところは?

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