多変量データの分析

機械学習の観点で言うと

教師あり学習: 学習データに正解を与えた状態で学習

教師なし学習: 学習データに正解を与えない状態で学習

教師なしデータ
- 次元圧縮: **主成分分析**, 特異値分解
- クラスタリング: **K-means**, 階層クラスタリング

教師ありデータ
- 回帰: 線形回帰, 決定木, ニューラルネットワーク
- 分類: SVM, 判別分析, 単純ベイズ

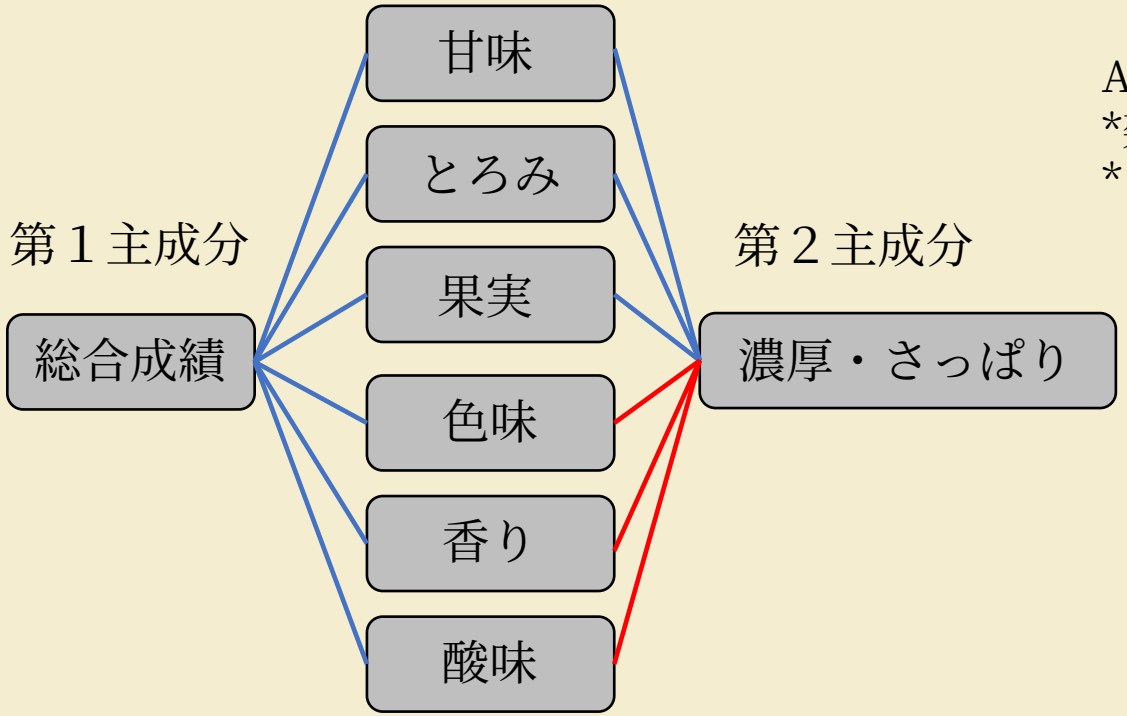
では、主成分分析とK-means法の原理と応用例を紹介していきます。

主成分分析 (PCA)

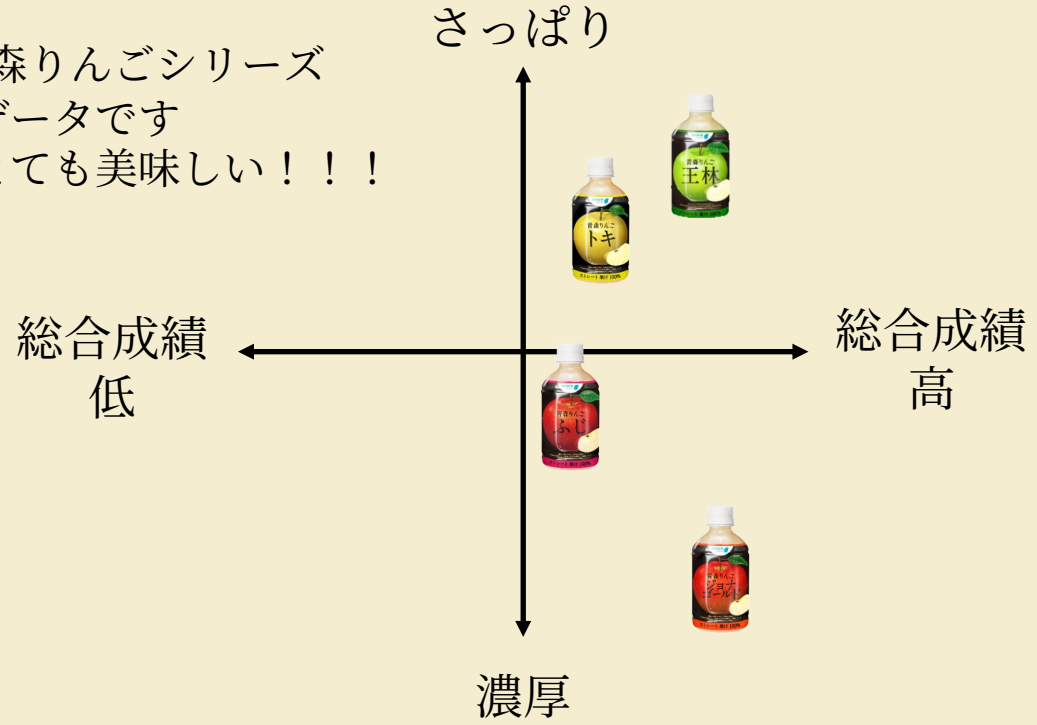
✓ 次元の圧縮に関する手法

多次元データから合成変数を用いて低次元空間に情報を縮約・圧縮

２・３次元へ圧縮 → データの視覚化 → 解釈のしやすさ

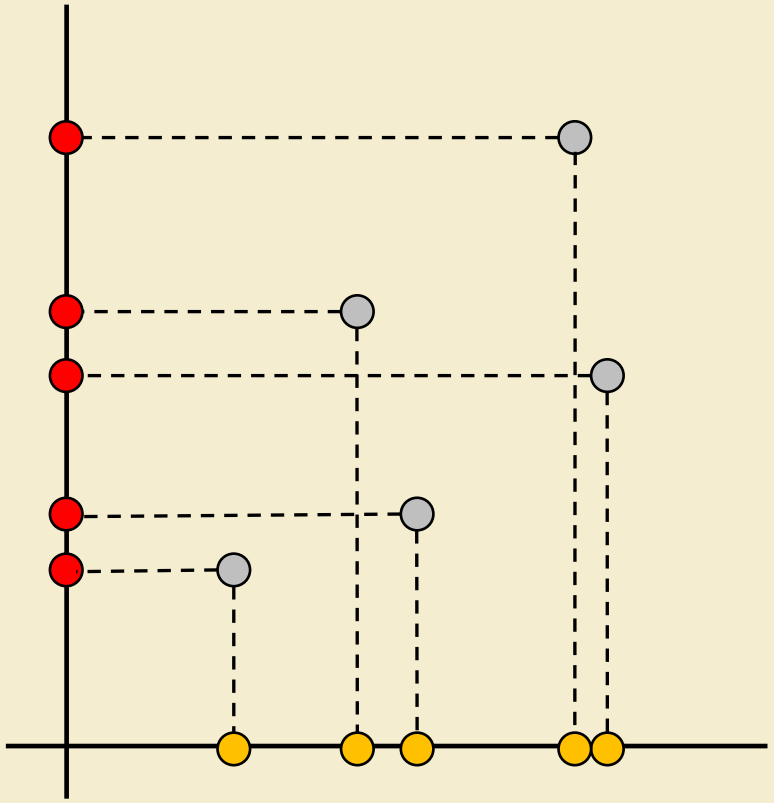


Acure青森りんごシリーズ
*架空のデータです
*どれもとても美味しい！！！

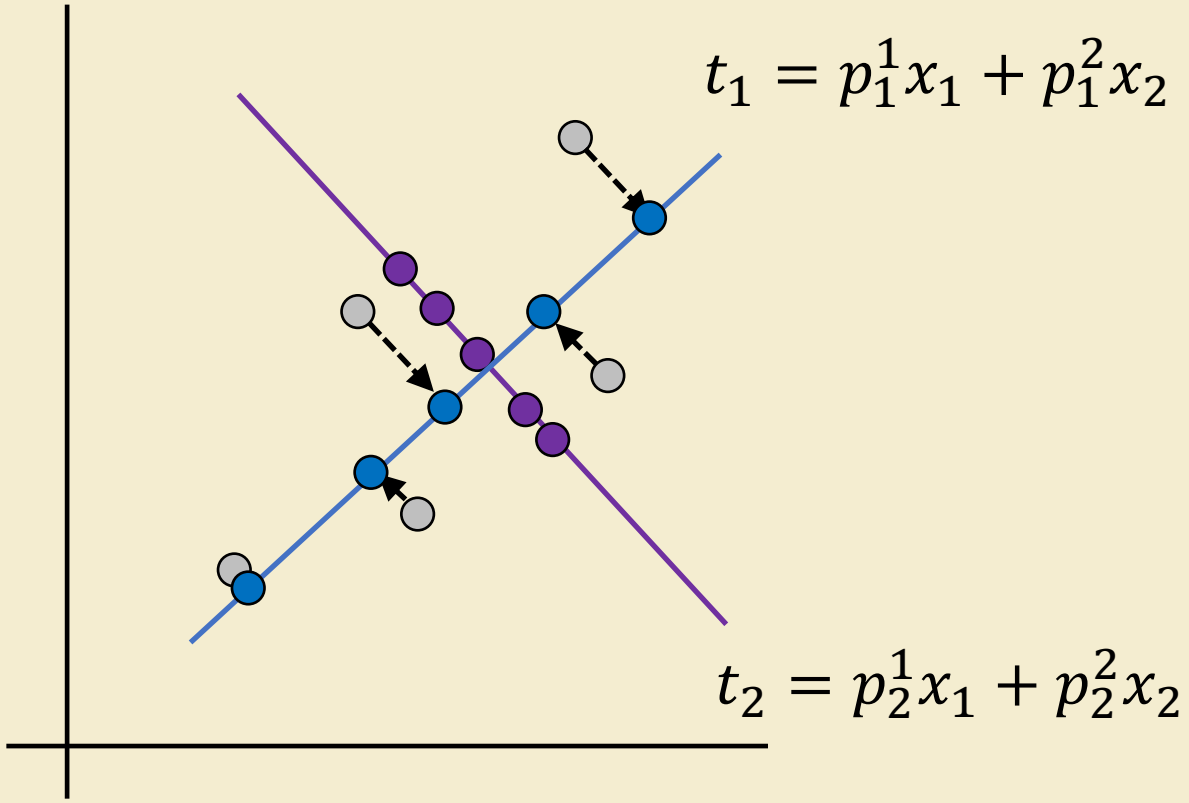


主成分分析の図解

１次元への圧縮



射影したデータの分散の最大化



もちろん次元圧縮は情報量の減少を意味する
データのばらつきが大きいほど情報量が多い

t_i　　第 i 主成分

p_i^j　　第 i 主成分に対応する j 番目の重み

じゃあ、どう解くの？

N個サンプルがある場合

x_1	x_2
x_1^1	x_2^1
x_1^2	x_2^2
・・・	・・・
x_1^n	x_2^n

1個目　$t_1 = p_1^1 x_1 + p_1^2 x_2^1 \quad t_2 = p_2^1 x_1 + p_2^2 x_2$

2個目　$t_1 = p_1^1 x_1 + p_1^2 x_2^1 \quad t_2 = p_2^1 x_1 + p_2^2 x_2$

⋮

n個目　$t_1 = p_1^1 x_1 + p_1^2 x_2^1 \quad t_2 = p_2^1 x_1 + p_2^2 x_2$

行列で書くと

$$\begin{bmatrix} t_1^1 & t_2^1 \\ \vdots & \vdots \\ t_1^n & t_2^n \end{bmatrix} = \begin{bmatrix} x_1^1 & x_2^1 \\ \vdots & \vdots \\ x_1^n & x_2^n \end{bmatrix} \begin{bmatrix} p_1^1 & p_2^1 \\ p_1^2 & p_2^2 \end{bmatrix}$$

t_i^k　第 k個目のサンプルの i 主成分の値

x_i^k　第 k個目のサンプルの i 番目の変数の値

p_i^j　第 i 主成分に対応する j 番目の重み

主成分の値の二乗和を最大化

$$\begin{bmatrix} t_1^1 & t_2^1 \\ \vdots & \vdots \\ t_1^n & t_2^n \end{bmatrix} = \begin{bmatrix} x_1^1 & x_2^1 \\ \vdots & \vdots \\ x_1^n & x_2^n \end{bmatrix} \begin{bmatrix} p_1^1 & p_2^1 \\ p_1^2 & p_2^2 \end{bmatrix}$$

既知　　変数

t_i^k　第 k個目のサンプルの i 主成分の値

x_i^k　第 k個目のサンプルの i 番目の変数の値

p_i^j　第 i 主成分に対応する j 番目の重み

主成分の分散の最大化

主成分の値の二乗和を最大化させることに対応　(変数のセンタリング後)

$$S = \sum_{i=1}^{n} \left(t_1^i\right)^2 = (p_1^1)^2 \sum_{i=1}^{n} \left(x_1^i\right)^2 + 2p_1^1 p_1^2 \sum_{i=1}^{n} x_1^i x_2^i + (p_1^2)^2 \sum_{i=1}^{n} (x_2^i)^2$$

この時、規格化条件　$(p_1^1)^2 + (p_1^2)^2 = 1$

Lagrangeの未定乗数法でSの最大化を行う

Lagrangeの未定乗数法

λを未知の定数としてGが最大になるλ, p_1^1, p_1^2を決める

$$S = (p_1^1)^2 \sum_{i=1}^{n} \left(x_1^i\right)^2 + 2p_1^1 p_1^2 \sum_{i=1}^{n} x_1^i x_2^i + (p_1^2)^2 \sum_{i=1}^{n} \left(x_2^i\right)^2 \qquad \text{制約条件} \quad (p_1^1)^2 + (p_1^2)^2 = 1$$

$$\mathrm{G} = S - \lambda\{(p_1^1)^2 + (p_1^2)^2 - 1\}$$

$$= (p_1^1)^2 \sum_{i=1}^{n} \left(x_1^i\right)^2 + 2p_1^1 p_1^2 \sum_{i=1}^{n} x_1^i x_2^i + (p_1^2)^2 \sum_{i=1}^{n} \left(x_2^i\right)^2 - \lambda\{(p_1^1)^2 + (p_1^2)^2 - 1\}$$

Gの最大値 → Gの極大値 → λ, p_1^1, p_1^2の偏微分が0となるとき

偏微分

偏微分すると…

$$\left(\sum_{i=1}^{n}\left(x_1^i\right)^2-\lambda\right)p_1^1+\left(\sum_{i=1}^{n}x_1^ix_2^i\right)p_1^2=0$$

$$\left(\sum_{i=1}^{n}x_1^ix_2^i\right)p_1^1+\left(\sum_{i=1}^{n}\left(x_2^i\right)^2-\lambda\right)p_1^2=0$$

⇒

$$\begin{bmatrix}\sum_{i=1}^{n}\left(x_1^i\right)^2-\lambda & \sum_{i=1}^{n}x_1^ix_2^i\\ \sum_{i=1}^{n}x_1^ix_2^i & \sum_{i=1}^{n}\left(x_2^i\right)^2-\lambda\end{bmatrix}\begin{bmatrix}p_1^1\\ p_1^2\end{bmatrix}=0$$

書き換えると…

$$\left(\boldsymbol{X}^{\mathbf{T}}\boldsymbol{X}-\boldsymbol{\lambda E}\right)\boldsymbol{p_1}=0$$

$$\boldsymbol{X}=\begin{bmatrix}x_1^1 & x_2^1\\ \vdots & \vdots\\ x_1^n & x_2^n\end{bmatrix}\quad \boldsymbol{E}=\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\quad \boldsymbol{p_1}=\begin{bmatrix}p_1^1\\ p_1^2\end{bmatrix}$$

分散の最大化には, p_1^1= p_1^2=0以外の解が必要!!!　という問題になる

固有値問題へ

$p_1^1 = p_1^2 = 0$以外の解が必要

逆行列を持たない → $(X^{\mathrm{T}}X - \lambda E)$の行列式が0である必要

λを固有値, p_1 , p_2を固有ベクトルとする固有値問題

p_1, p_2を求め, 対応する主成分が得られる！

<u>今までやっていたのは分散共分散行列の固有値問題を解いていることに</u>

$$\sum = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_1^i - m)^2 & \frac{1}{n}\sum_{i=1}^{n}(x_1^i - m)(x_2^i - m) \\ \frac{1}{n}\sum_{i=1}^{n}(x_1^i - m)(x_2^i - m) & \frac{1}{n}\sum_{i=1}^{n}(x_2^i - m)^2 \end{bmatrix}$$

主成分分析の実装

使用するデータ: 1973年の車に関するデータセット (32 × 11)

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4

mpg	Miles/(US) gallon	qsec	1/4 mile time
cyl	Number of cylinders	vs	Engine (0 = V-shaped, 1 = straight)
disp	Displacement (cu.in.)	am	Transmission (0 = automatic, 1 = manual)
hp	Gross horsepower	gear	Number of forward gears
drat	Rear axle ratio	carb	Number of carburetors
wt	Weight (1000 lbs)		

Princomp vs. Prcomp

	Princomp	Prcomp
分析対象となるデータ	R mode　(行数 >> 列数)	R mode (行数 >> 列数) Q mode (行数 << 列数)
目的	サンプルに対する変数の解析 (特徴量の抽出)	変数に対してサンプルの解析 (データ行列Xの転置行列に)
計算アルゴリズム	共分散行列の固有値	特異値分解
固有値	主成分得点の不偏分散	主成分得点の標本分散

- データボックス: イギリスの心理学者レイモンド・キャッテルが導入
  データを3つの次元（サンプル、変数、時間）から考えるもの
- **迷ったら, prcomp()**

なぜ共分散行列に基づくアルゴリズムでは, 行数 >> 列数が問題となるのか…

- データ行列の次元: $n \times v$ （変数: v 個, 観測値: n 個)
- 共分散行列を用いてRモードで解析: 分散共分散行列 C の次元は $v \times v$, ランクは n と v の小さい方
- 共分散行列を用いてQモードで解析: 分散共分散行列 C の次元は $n \times n$, ランクは n と v の小さい方
  $n >$ v の時→ 固有値および固有ベクトルが v 個しか求まらない → 計算ストップ

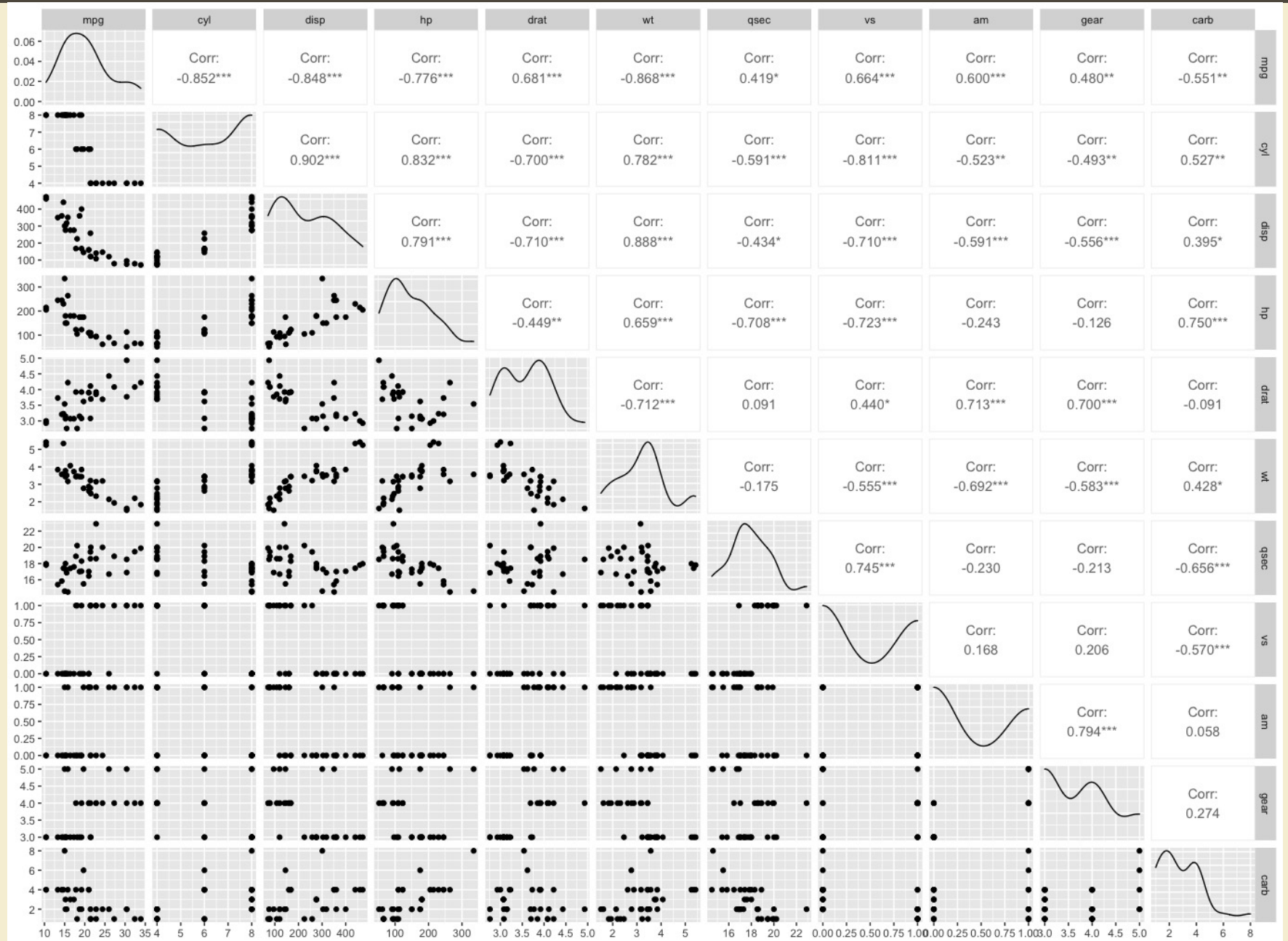
主成分分析の実装

なにはともあれ
データも外観から

```
 9 ggpairs(df)
10 df_filtered <- df %>% select(., -c(am, vs))
```

名義尺度のvs, amを
除いて主成分分析を
してみる

```
12 pca_result <- prcomp(df_filtered, scale=TRUE)
```



データの前処理

centering

各変数の平均を0にする

$$x_i - \bar{\mu}$$

scaling

各変数の分散を１にする

$$x_i / \sigma$$

standardization

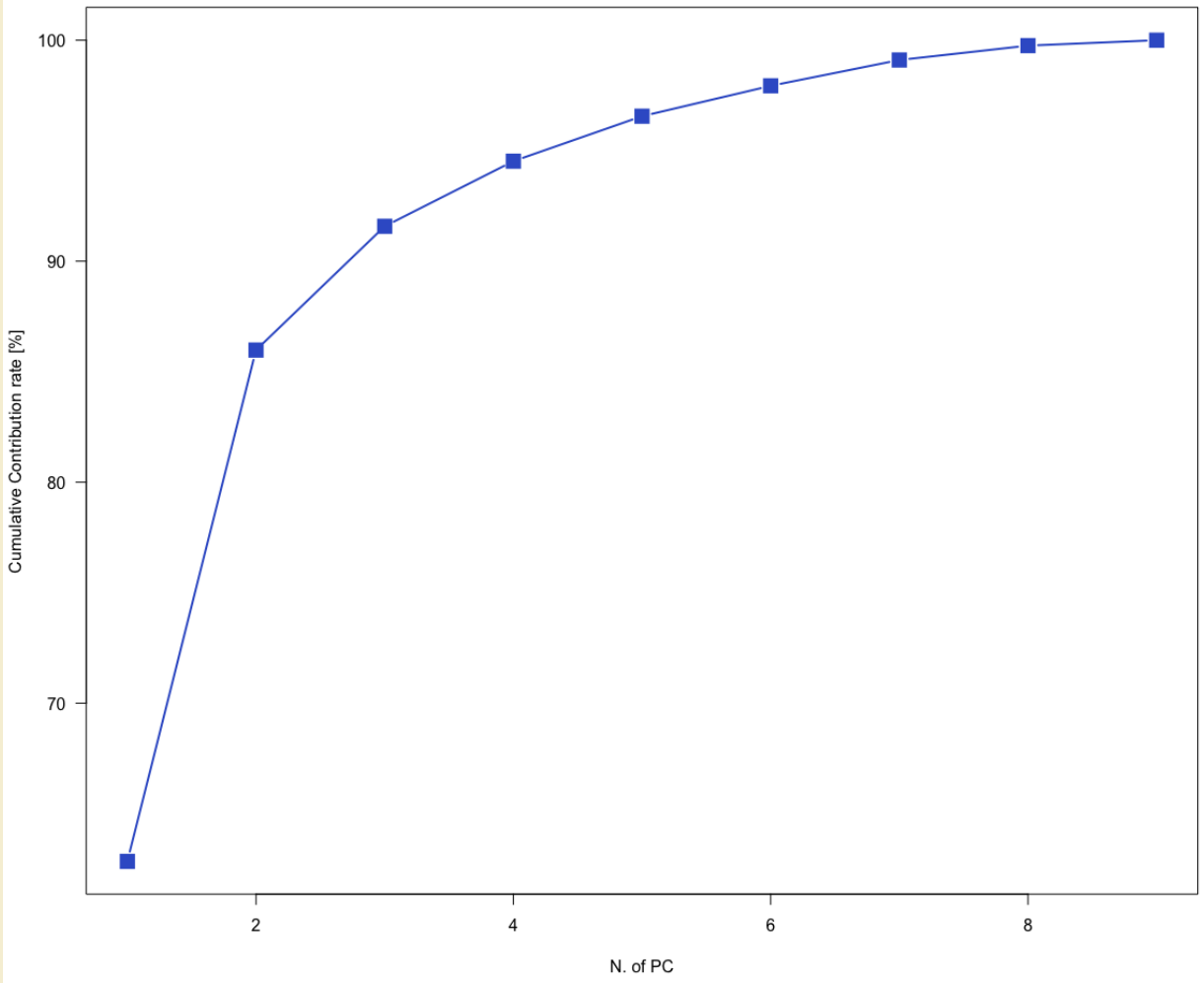
各変数の平均を0、分散を1にする

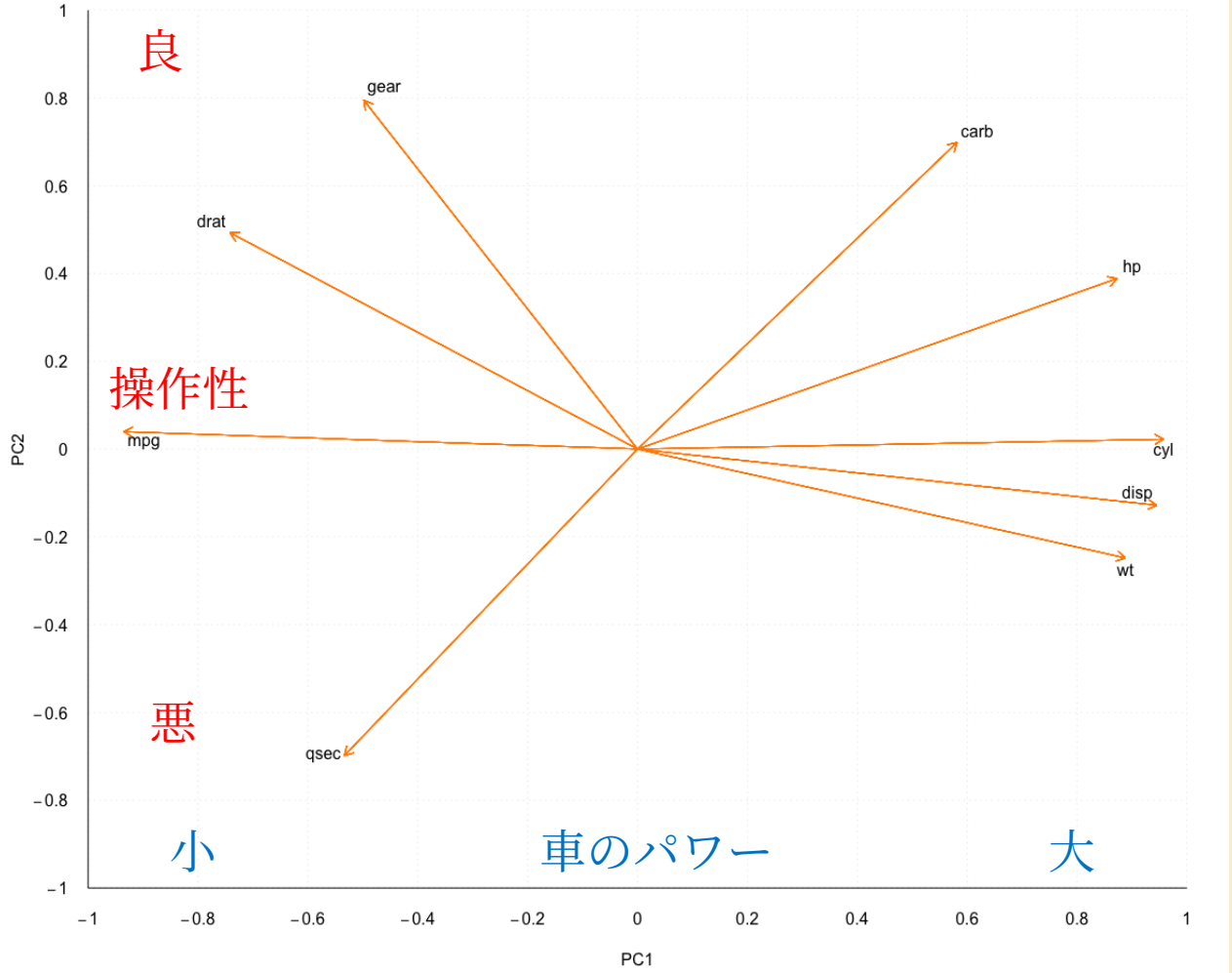
$$\frac{x_i - \bar{\mu}}{\sigma}$$

単位系が異なる場合なども，
各変数が同等の重みを持つようになる

主成分分析の実装

結果の視覚化





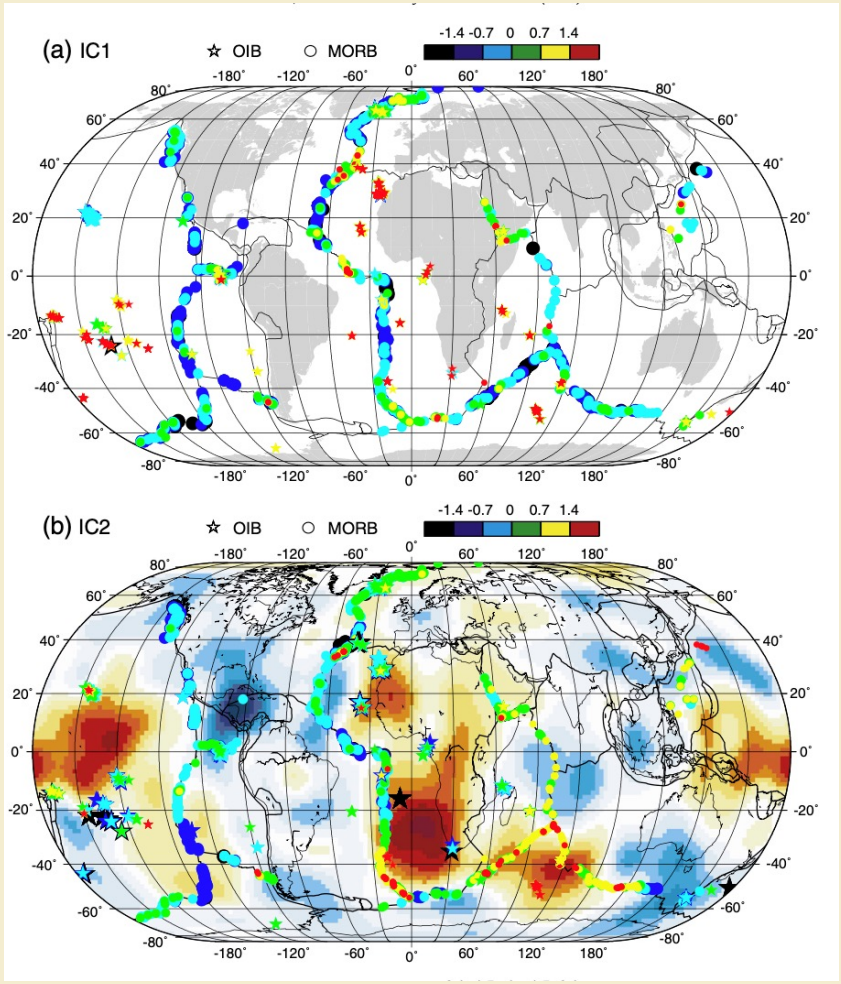
主成分分析のまとめ

- 次元を圧縮, 互いに無相関な成分を計算
- 分散共分散行列の固有値固有ベクトル分解に相当
- N次元のデータからはN個の主成分
  (少ない成分で説明できるほど嬉しい)
- データ構造の本質的な要因を抽出できる可能性

地球化学への応用: 主成分分析

Iwamori + (2010) *EPSL*

独立成分分析(PCAを拡張したもの)を海嶺・海洋島玄武岩の同位体組成解析（Pb, Nd, Sr, Hf）に応用



IC1: MORBと海洋島玄武岩OIB を分ける

IC2: 地理的分布, 特にDUPAL異常を識別

<u>そこから得られた知見</u>

DMM, EMのような地球化学的マントル端成分の相互作用ではなく, 2つの独立な物質分化プロセスの重複を示唆

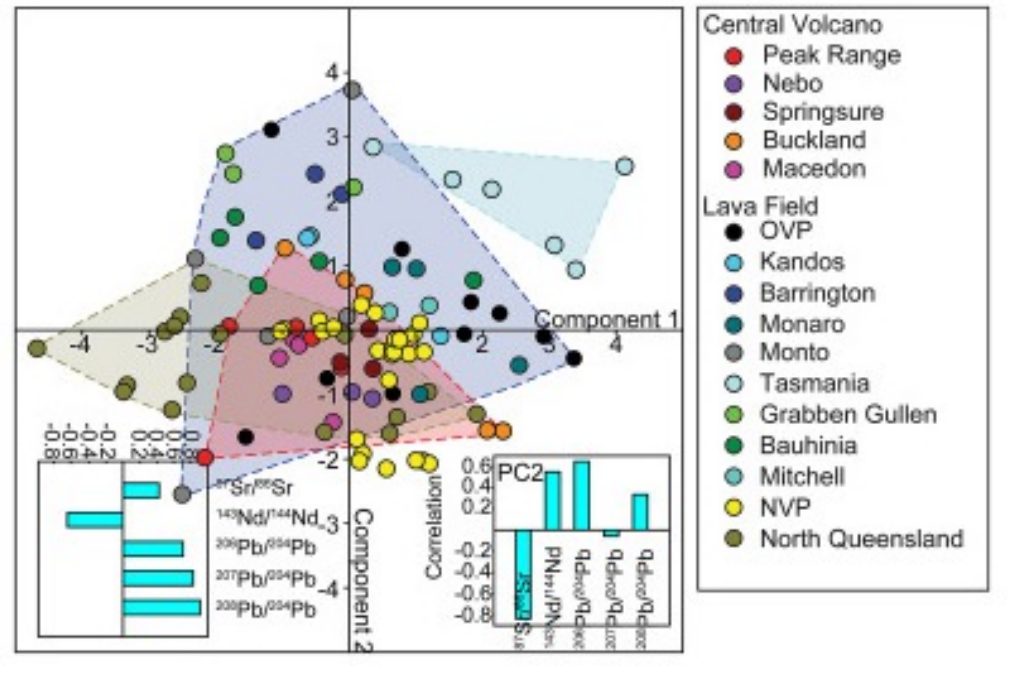
IC1: マントル溶融プロセスで元素分別, 多様性形成

IC2: 沈み込み帯における脱水・加水でIC2方向の多様性

これらが再循環し, 海嶺で再溶融するという, 2つの独立な分別作用の相互重複プロセスがマントルの不均質の原因と推定

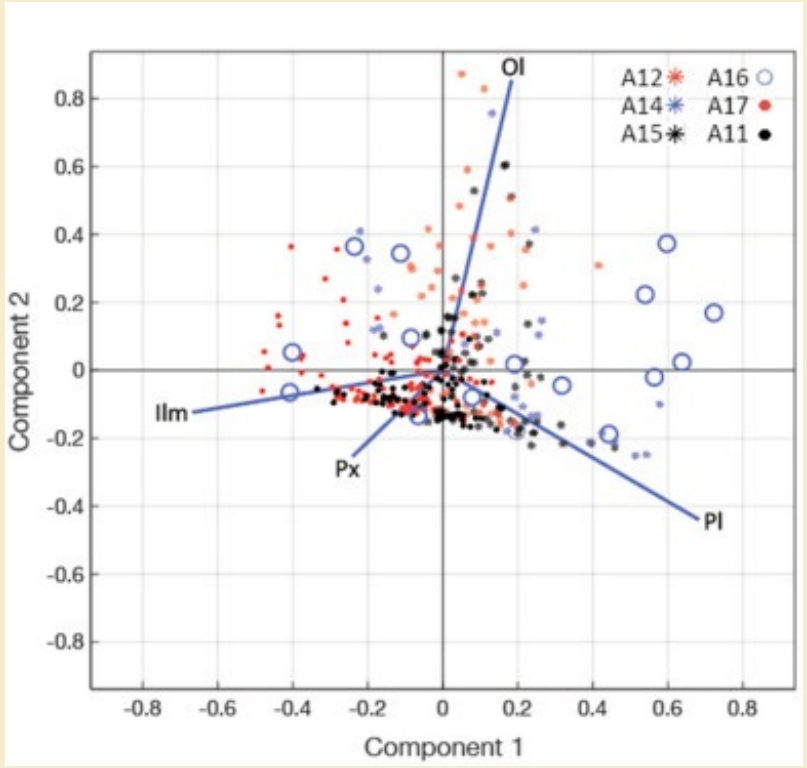
地球化学への応用: PCA

Jones + (2020) *Lithos*



東オーストラリアの新生代火山の
同位体, 微量元素などのデータセットに主成分分析

Cone + (2020) *Icarus*

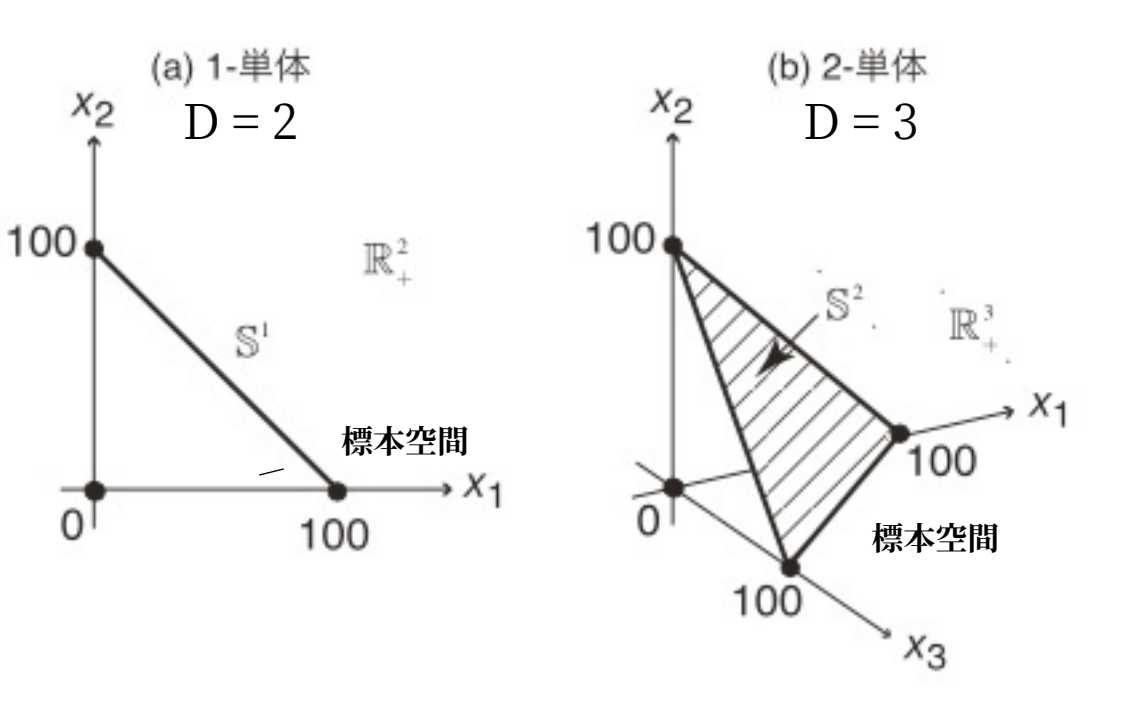


Apollo sampleの玄武岩データベースの
化学組成や鉱物モードに適用

組成データの注意点

Aitchison (1982, 1986)

- 組成データ: 変数が全て正, 変数の総和が定数
- 定数和制約: 変数の独立性は保たれない

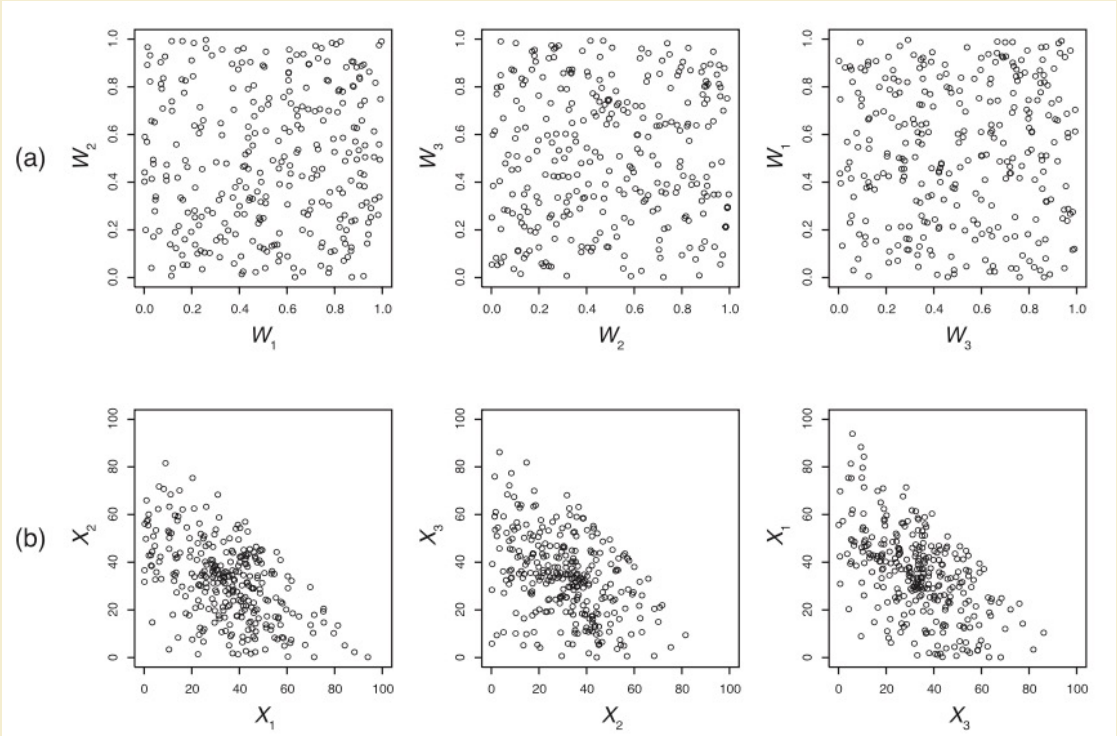


位相幾何学的には D–1 次元空間を単体空間($\mathcal{S}^{D-1}$)

$$\mathbb{S}^{D-1}=\left\{ (x_1,\cdots,x_D) \mid x_i>0 \ (i=1,\cdots,D), \sum_{i=1}^{D} x_i=100 \right\}$$

Butler(1978, 1979a)

統計学的な性質に注意が必要



3次元の組成データは3次元空間に属する 2-単体空間に配列する

→ データはもともとの3次元的な広がりを持たず, ある平面の領域に集約

→ 無理に直行座標系にプロットすると, 実在しない配列や散布度が出現

対数比変換・対数比解析

組成データをS^{D-1}からR^Dに写像する

写像後の対数比データに対して,
実数と同等に演算と統計解析が実行可能

= 定数和制約からデータを開放

相加対数比変換
(alr: additive log-ratio transformation)

$$alr(x) = \left(\ln\frac{x_1}{x_D}, \ln\frac{x_2}{x_D}, \cdots, \ln\frac{x_{D-1}}{x_D}\right)^t = y(y_1, y_2, \cdots, y_{D-1})^t$$

* 変換結果として使用できる変数は
元の変数から一つ減ってしまう

有心対数比変換
(clr: centered log-ratio transformation)

$$clr(x) = \left(\ln\frac{x_1}{g(x)}, \ln\frac{x_2}{g(x)}, \cdots, \ln\frac{x_D}{g(x)}\right)^t = z(z_1, z_2, \cdots, z_D)^t$$

$$g(x) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}$$