多変量解析
（ 特徴空間分割・次元圧縮）

北海道大学　農学研究院
佐藤昌直

モチベーション:

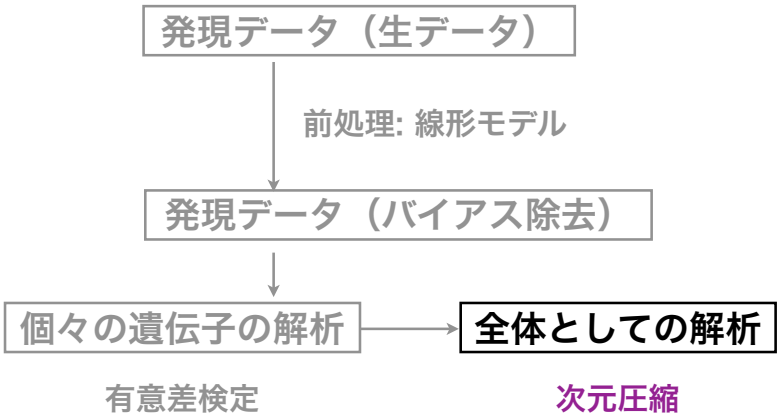
多次元（例: 多パラメーター）を
より少ない指標を使って理解する

↓

N個のサンプルをM個（M < N)の
グループに分類する

→ **人間が新たな解釈を与える**

解析の流れ

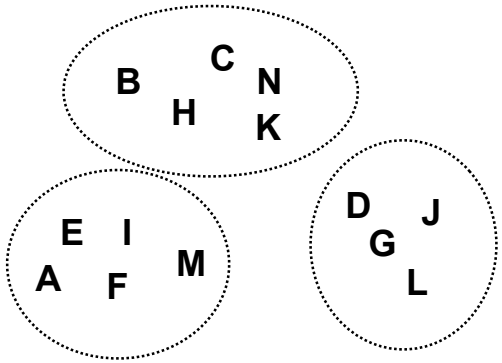


下記のデータセットに含まれる数値を俯瞰してみましょう。 データの特徴を読み取れるでしょうか？

```
inputMatrix<- read.delim("~/data/MS/Sato_A_thaliana-P_syringae_arvRpt2_6h_expRatio_small.txt", header=TRUE, row.name=1)
head(inputMatrix)  #読み込みデータの一部を表示
image(t(inputMatrix))              #カラーコードによって可視化
heatmap(as.matrix(inputMatrix)) #階層クラスタリングで解析し、簡易表示
```

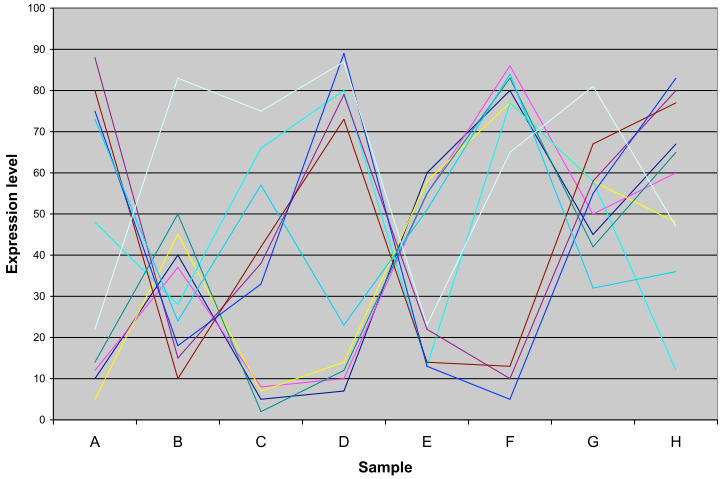
高次元（多パラメーター）データの
認識における問題をどう扱うか？

クラスタリングによる分類



14 プロファイル
→ 3 クラスター

トランスクリプトームデータの
ある一部について可視化してみる



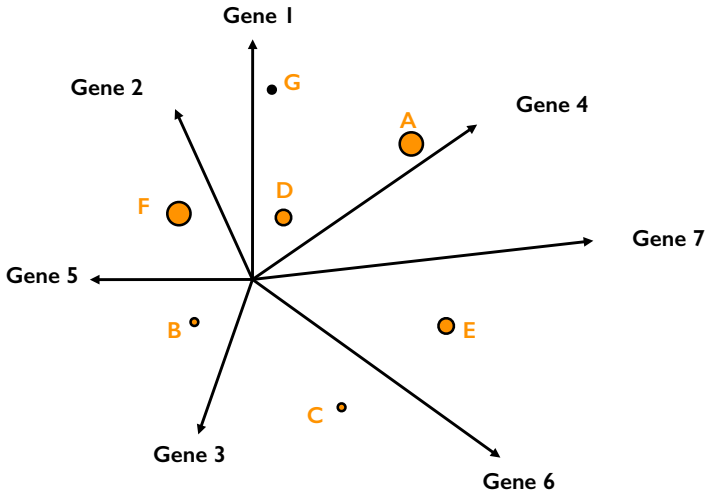
ポイント

多変量解析のポイント

教師有りか無しか
(supervised or unsupervised)？

どのような距離行列を使うか？

7次元の遺伝子発現データセット



コンピューターにどうデータを渡せば
この問題をどう扱えるか？

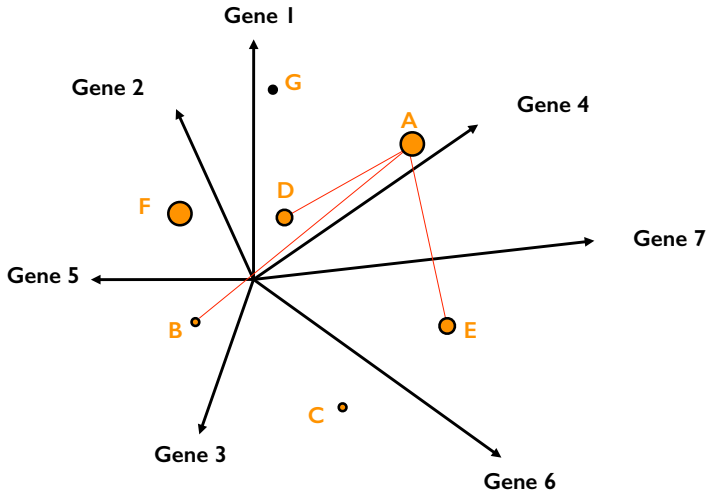
人間　**遺伝子発現プロファイル間のパターンの比較**

問題定義の変換
(生物学の問題を数学の問題に置き換える)

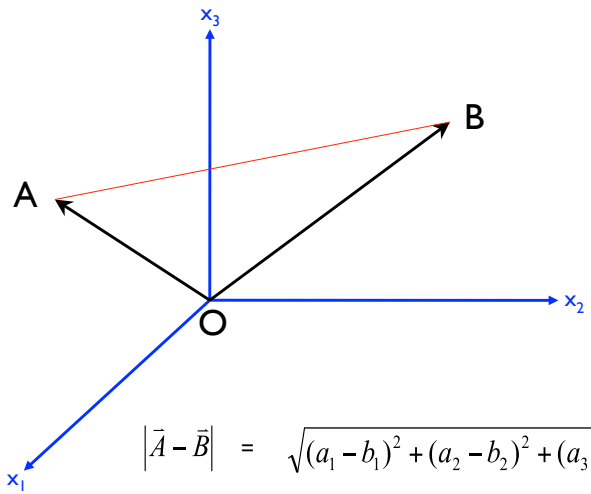
コンピューター　**データの大きさに定義される次元の空間でのデータポイントの分布の比較**

距離の基準を何にするか
距離尺度

7遺伝子の発現プロファイル間の類似性は
7次元空間での距離によって決まる

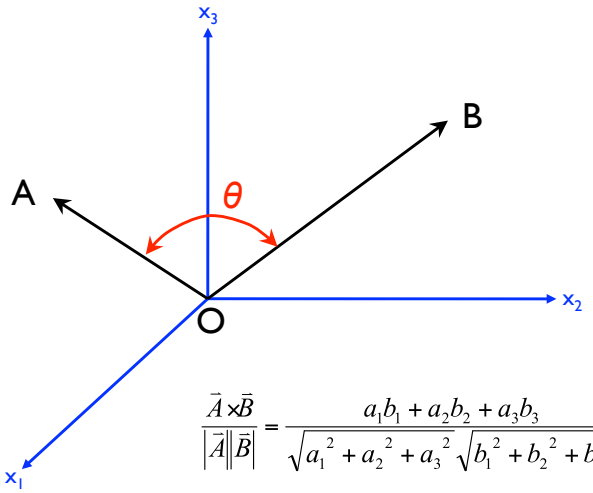


ユークリッド距離



$$\left|\vec{A}-\vec{B}\right| = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+(a_3-b_3)^2}$$

Uncentered Pearson correlation coefficient = cosθ



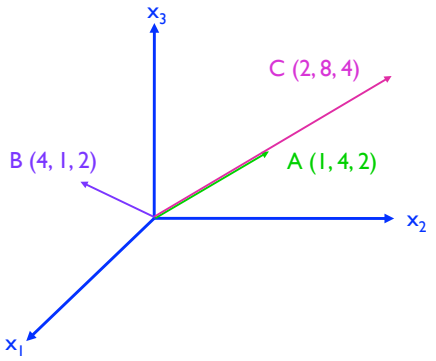
$$\frac{\vec{A}\times\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|}=\frac{a_1b_1+a_2b_2+a_3b_3}{\sqrt{a_1^{\,2}+a_2^{\,2}+a_3^{\,2}}\sqrt{b_1^{\,2}+b_2^{\,2}+b_3^{\,2}}}=\cos\theta$$

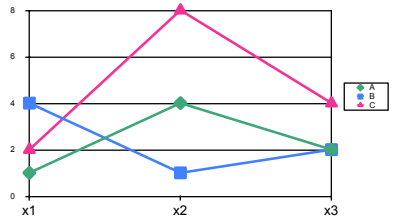
距離尺度の違い→解析対象の違い:

遺伝子発現プロファイルの**形**と**大きさ**

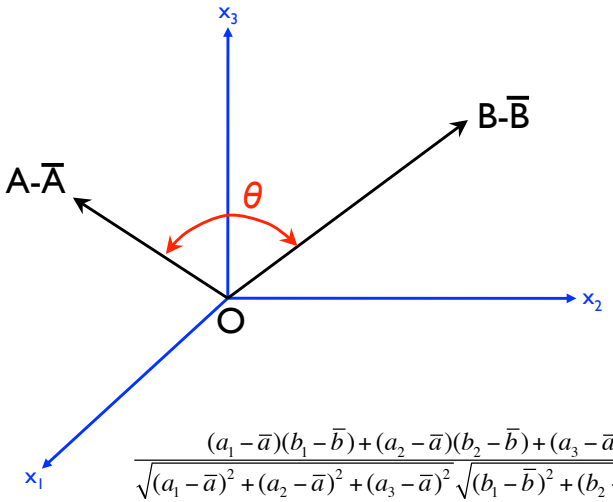
ポイント

- **形**: ベクトルの方向
- **大きさ**: ベクトルのサイズ





相関係数
Pearson correlation coefficient



$$\frac{(a_1-\bar{a})(b_1-\bar{b})+(a_2-\bar{a})(b_2-\bar{b})+(a_3-\bar{a})(b_3-\bar{b})}{\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}\sqrt{(b_1-\bar{b})^2+(b_2-\bar{b})^2+(b_3-\bar{b})^2}}$$

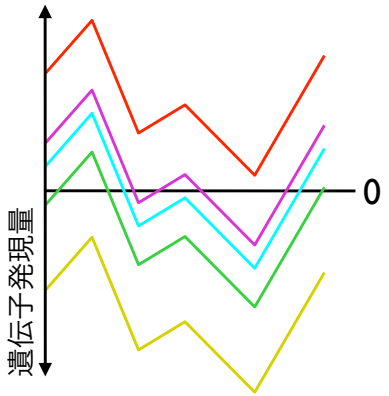
どの距離係数を使うか？

- どんなプロファイルを同じプロファイルと定義するか？
- 距離係数計算の背後にあるものを意識して選択する。

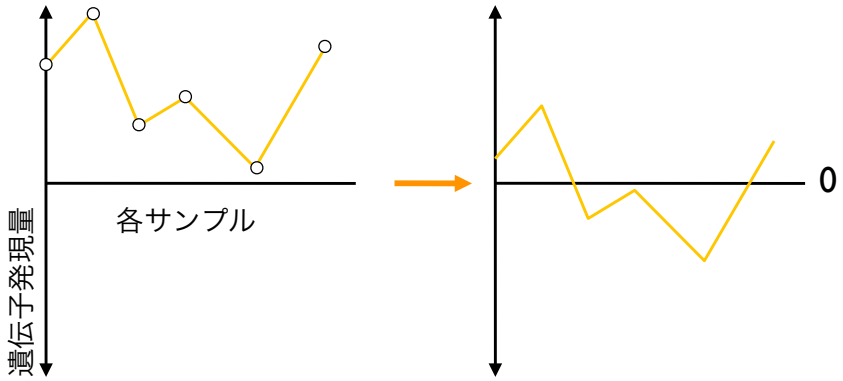
距離係数計算の過程には

- **Centering**: 平均値をゼロにする
- **Scaling**: ベクトルの大きさを1にする

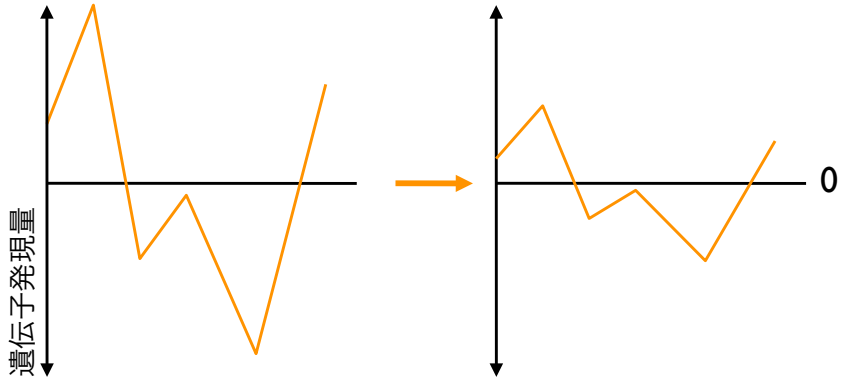
これらはcentering後は
全く同じプロファイルになる



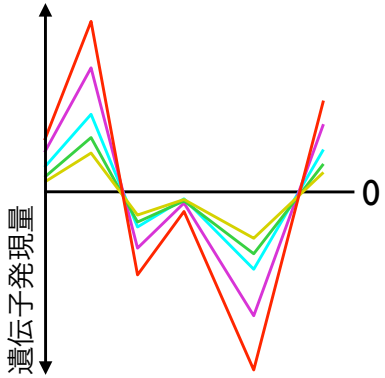
Centering



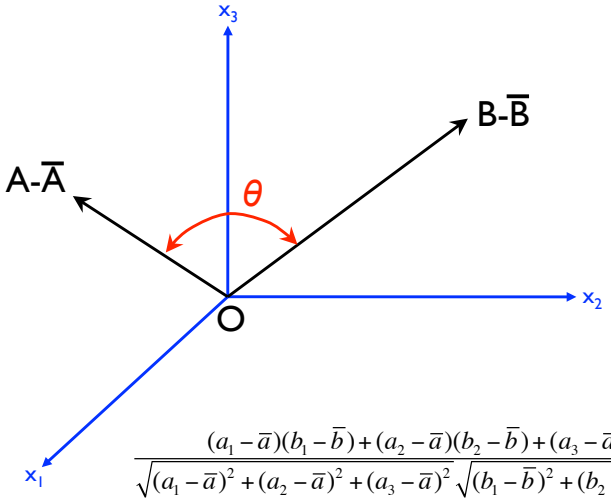
Scaling



これらはscaling後は
全く同じプロファイルになる

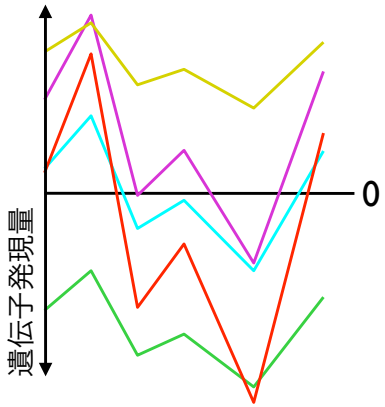


アルゴリズムに注目: 相関係数の場合



$$\frac{(a_1-\bar{a})(b_1-\bar{b})+(a_2-\bar{a})(b_2-\bar{b})+(a_3-\bar{a})(b_3-\bar{b})}{\sqrt{(a_1-\bar{a})^2+(a_2-\bar{a})^2+(a_3-\bar{a})^2}\sqrt{(b_1-\bar{b})^2+(b_2-\bar{b})^2+(b_3-\bar{b})^2}}$$

これらはcentering, scaling後は
全く同じプロファイルになる



ポイント

多変量解析における注意点

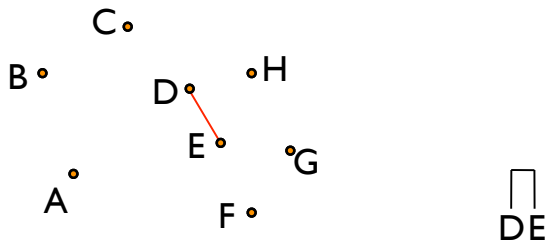
方法依存的に抽出される特徴:

どのような特徴を認識したいのか/
しているのか意識すること

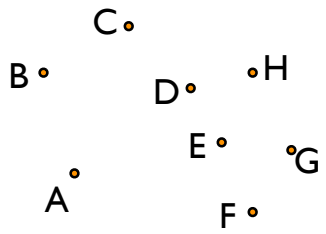
多変量解析の実際

階層クラスタリング

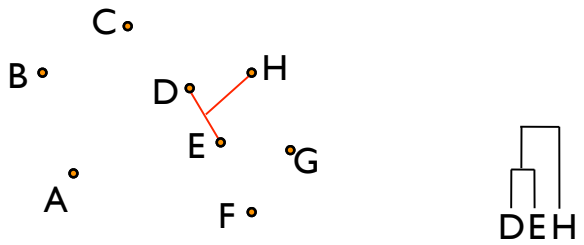
Agglomerative hierarchical clustering



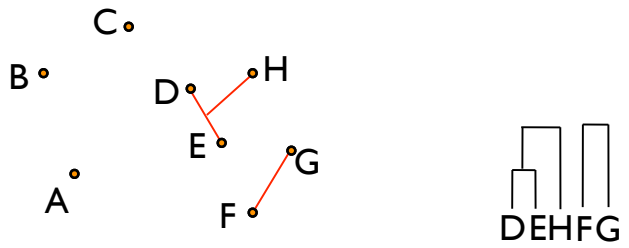
Agglomerative hierarchical clustering



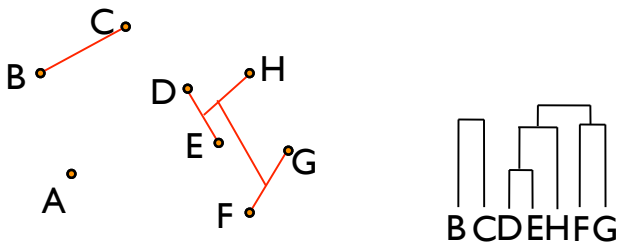
Agglomerative hierarchical clustering



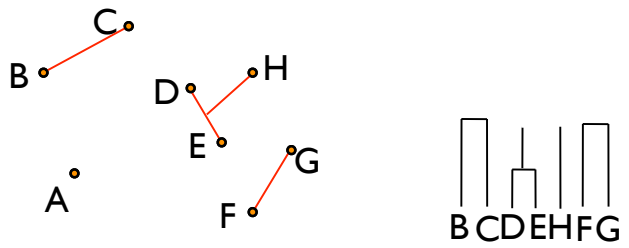
Agglomerative hierarchical clustering



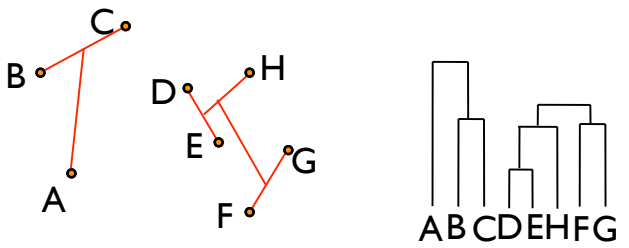
Agglomerative hierarchical clustering



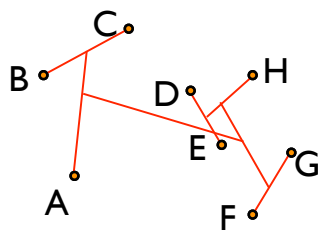
Agglomerative hierarchical clustering

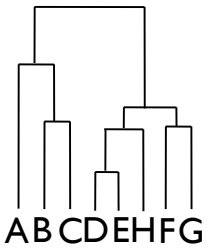


Agglomerative hierarchical clustering

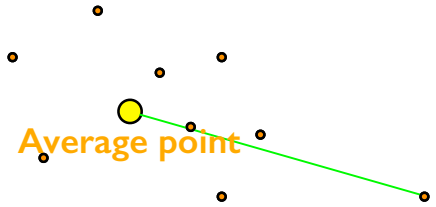


Agglomerative hierarchical clustering

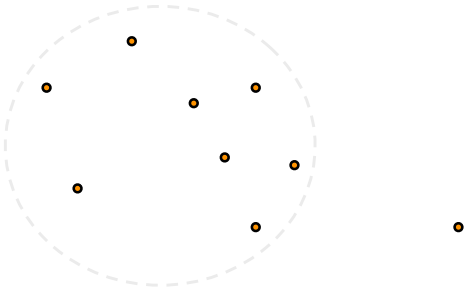




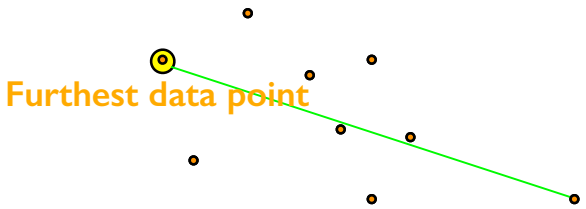
Average linkage



クラスター定義手法



Complete linkage

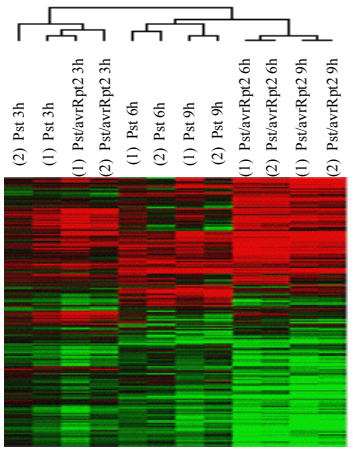


Single linkage

階層クラスタリングの欠点

- Bottom-up: 非常に「手順」依存性
- 一つの距離のみを指標とした
  クラスタリング

階層クラスタリングの利点



- クラスター化してより
  少数のカテゴリーを示
  す
- 人間が認識可能なパター
  ンを示す

「手順依存的」な方法の欠点を補うには？

- 偶然、観察されているクラスターを
  推定する
  → **同じ手順を繰り返す**
  → **クロスバリデーション**

クロスバリデーション

- あるクラスターは必然か偶然か？
  - leave-one out validation:サンプルを一つ抜いてクラスタリングしてみる
- 少数の特定遺伝子がクラスタリングに影響していないか？
  - Bootstrap: 遺伝子サブセットでクラスリングを繰り返してみる

主成分分析

多変量解析(I)のまとめ

教師有りか無しか
(supervised or unsupervised)？

- 事前情報、前提はあるか？
- ある場合はk-means法などの利用を検討

どのような距離行列を使うか？

- プロファイルの大きさ
- プロファイルの角度　など

主成分分析とは？

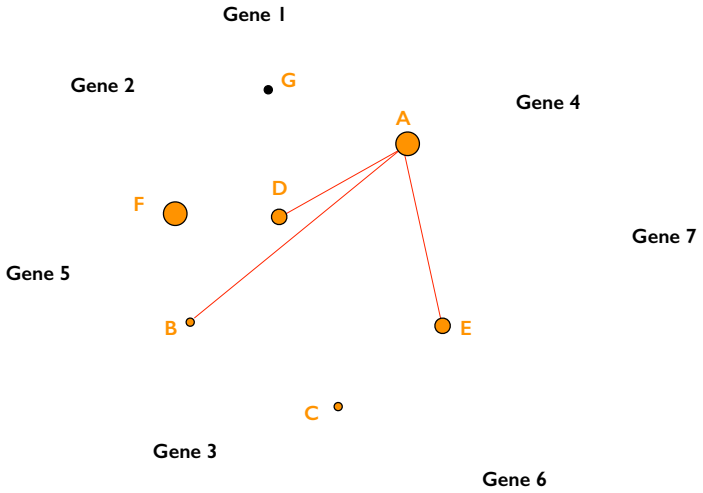
モチベーション:

多数の遺伝子で構成される多次元データ（サンプル）の中で相関のある遺伝子群を使って**新たな軸**を作り、データを見直す

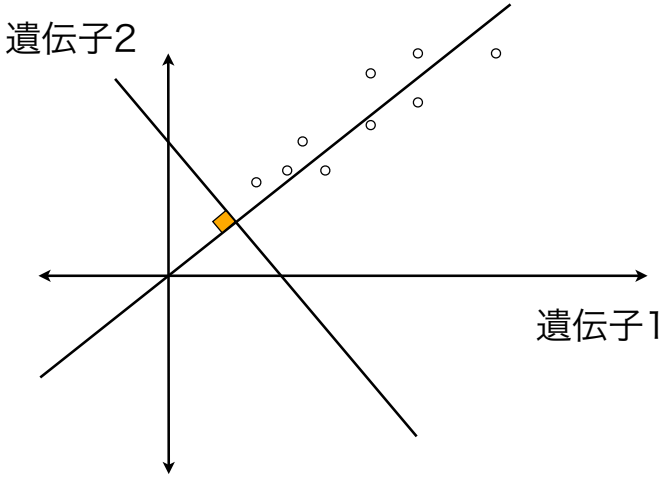
→ **人間が新たな解釈を与える**

階層クラスタリング、k-means法:

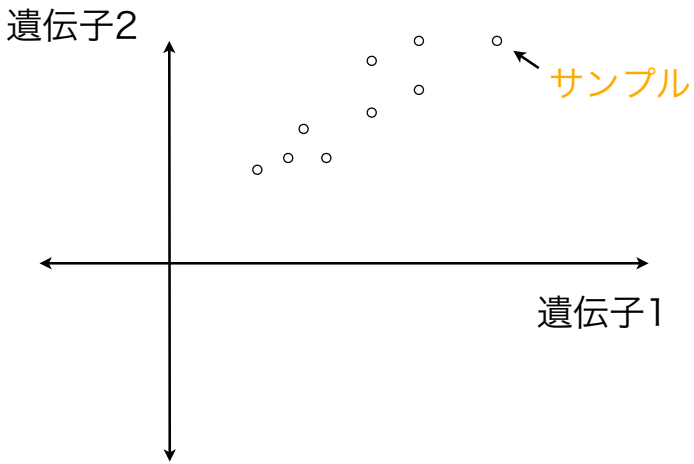
プロファイル間の類似性は空間での**1つの距離**によって決まる



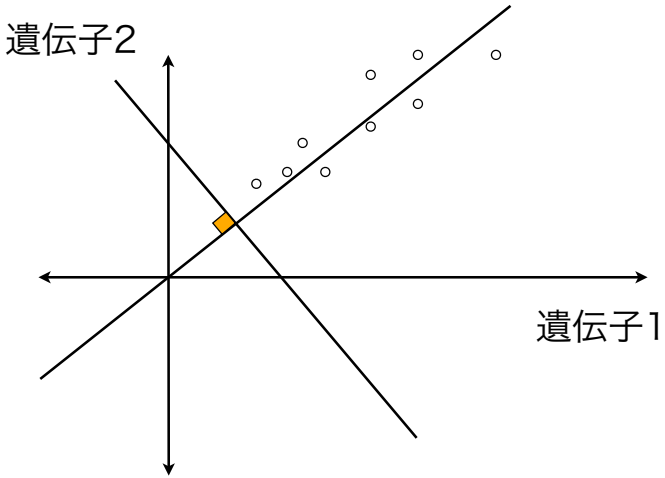
PCAは何をするのか？



PCAは何をするのか？



PCAは何をするのか？



PCAの概略(2次元)

1. 各サンプル（1..n）の観察値(x_n, y_n)を

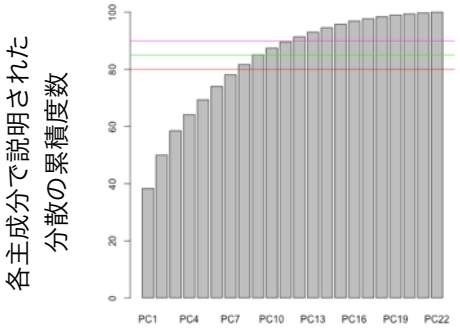
$$u_n = a_1 x_n + b_1 y_n$$
$$v_n = a_2 x_n + b_2 y_n$$

とおく

2. $a^2 + b^2 = 1$, uとvの相関係数0という制約の下でこれを解いて a_n, b_n を求める。

寄与率

- 各主成分が説明する分散の割合

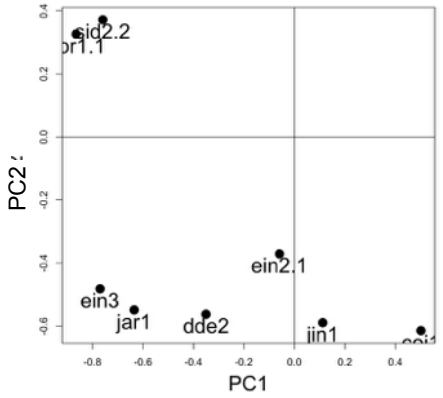


PCAで得られる重要な統計量

- 寄与率
- 因子負荷量
- 主成分得点

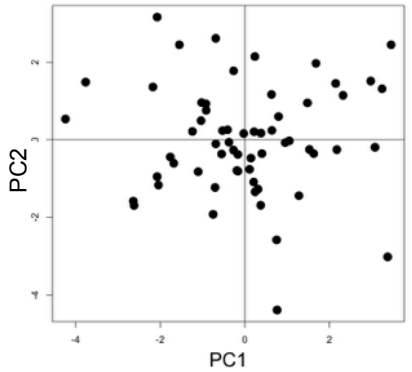
負荷量 loadings

- 得られた主成分と元データのパラメーターの相関
- 各パラメーターがもとのデータの情報をどれだけ有するか



主成分得点 scores

- 各パラメーターの値を各主成分について標準化したもの



標準化: 平均0, SD=1

注意点

1.デフォルトのprincompでは返り値loadingsは因子負荷量ではない。

2. 相関を使うか、分散共分散行列を使うか

主成分分析(まとめ)

- 主成分分析は**データの分散を説明する新たな軸を計算する**方法
  - **寄与率**
  - **因子負荷量**
  - **主成分得点**

多次元尺度構成法

Multi-dimensional scaling(MDS),
Principle coordinate analysis

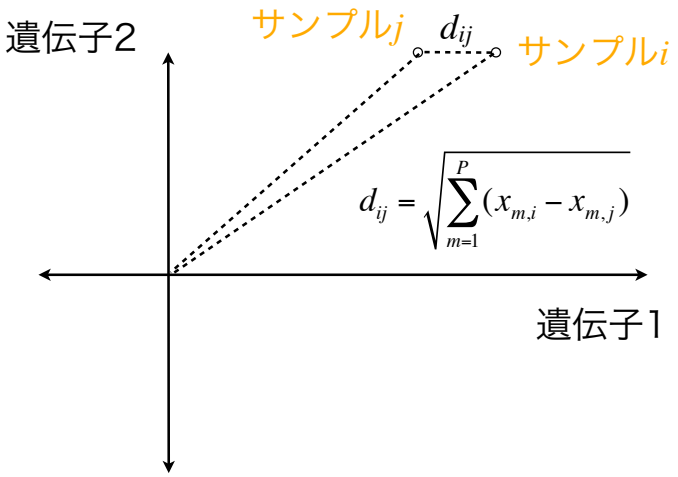
多次元尺度構成法とは？

モチベーション:

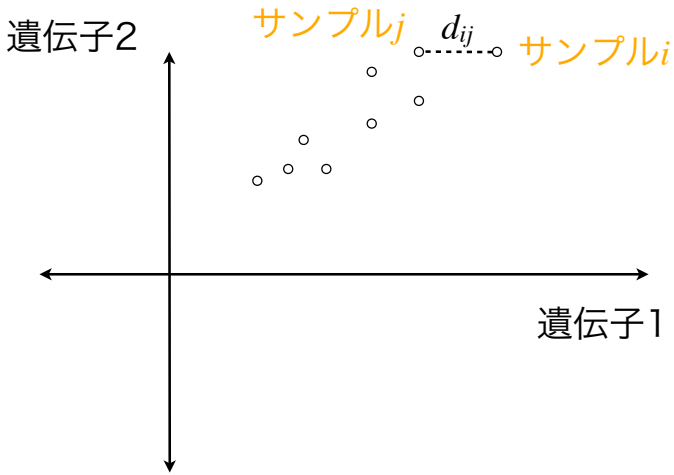
多数の遺伝子で構成される多次元の中で
各サンプル間の違いを低次元で表現する

距離係数を元に次元圧縮するため、非線形
の関係にも対応 (PCA: 分散を使う [線形]。
計算手法によってはPCAと同義になる）

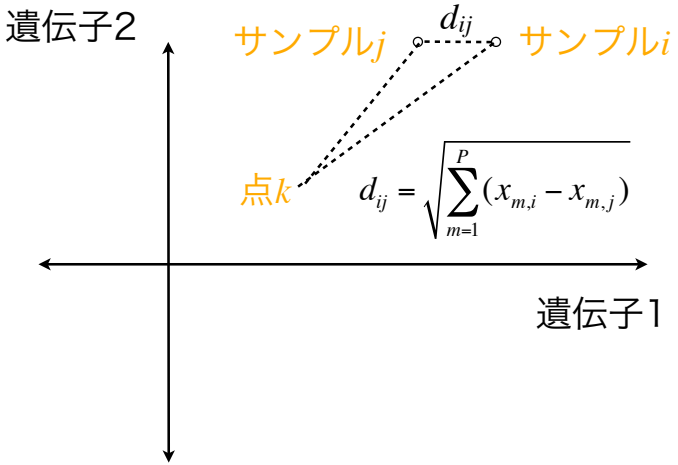
サンプル間の距離をまず計算する



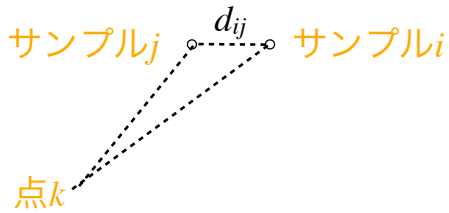
MDSは何をするのか？



この定理はサンプルi,jに対し、どこを原点（点k）としても成り立つ

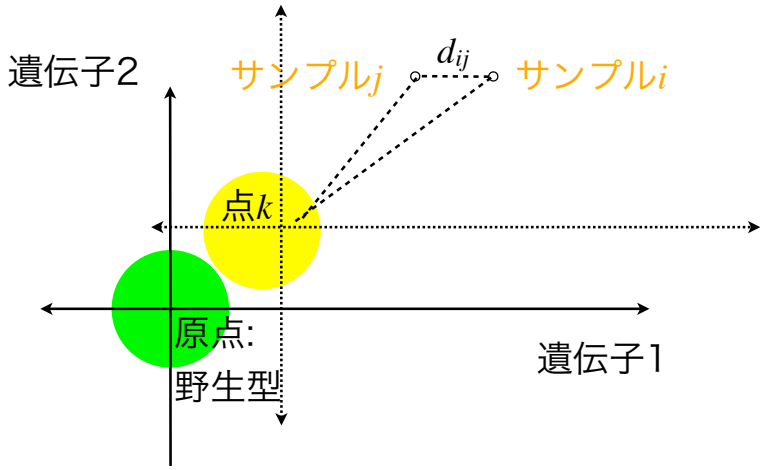


この定理はサンプルi,jに対し、どこを原点（点k）としても成り立つ

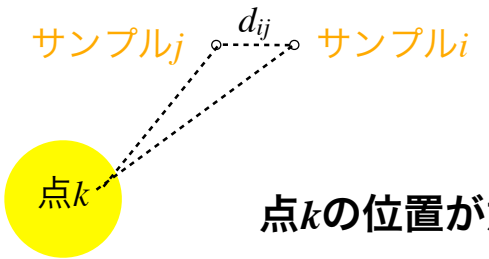


$$d_{ij}^2 = d_{ik}^2 + d_{jk}^2 - 2d_{ik}d_{jk}\cos\theta$$

例:入力データが野生型・変異体プロファイルの比であったら？



この定理はサンプルi,jに対し、どこを原点（点k）としても成り立つ



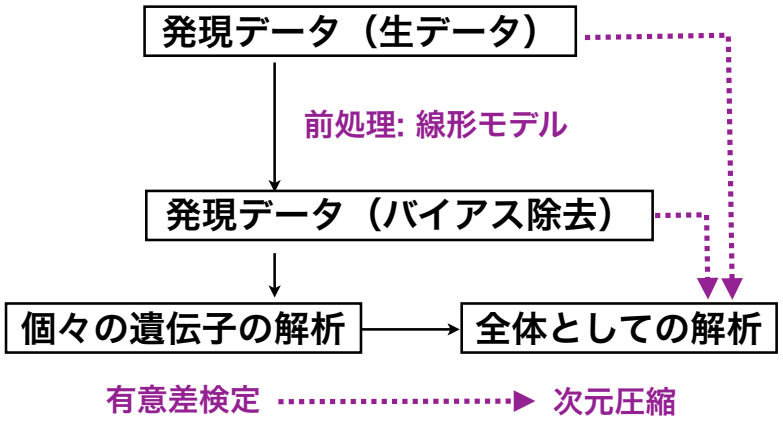
点kの位置が意味を持つことはないのか？

多変量解析(2)のまとめ

PCA/MDS

- データがもつ類似性を低次元で表現し、評価・可視化する
- **重心の置き方に違い**: 入力データをどのように前処理するか

多変量解析をもう一歩進めて:
入力データは何を使うか？



多変量解析をもう一歩進めて:

人間の解釈をアシストするデータ取得を心がける

多変量解析の枠組み

モチベーション:

多次元（例: 多パラメーター）を
より少ない指標を使って理解する

↓

N個のサンプルをM個（M < N)の
グループに分類する

→ **人間が新たな解釈を与える**

コントロール、
指標サンプルは
含められるか？

多変量解析をもう一歩進めて:
研究の目的、実験デザイン、多変量解析

目的	実験デザイン	多変量解析
•何を知りたいか明確に	•**線形モデル**	•入力データ前処理
•実施の制約	•比較、因子	•距離尺度
•予算	•検出力	•アルゴリズム
•時間、労力		

今回のトレーニングコースで
扱わなかった重要項目

- 確率分布
- 回帰、相関
- 線形モデルにおける交互作用
- 非線形クラスタリング・次元圧縮
  - self-organization mapなど