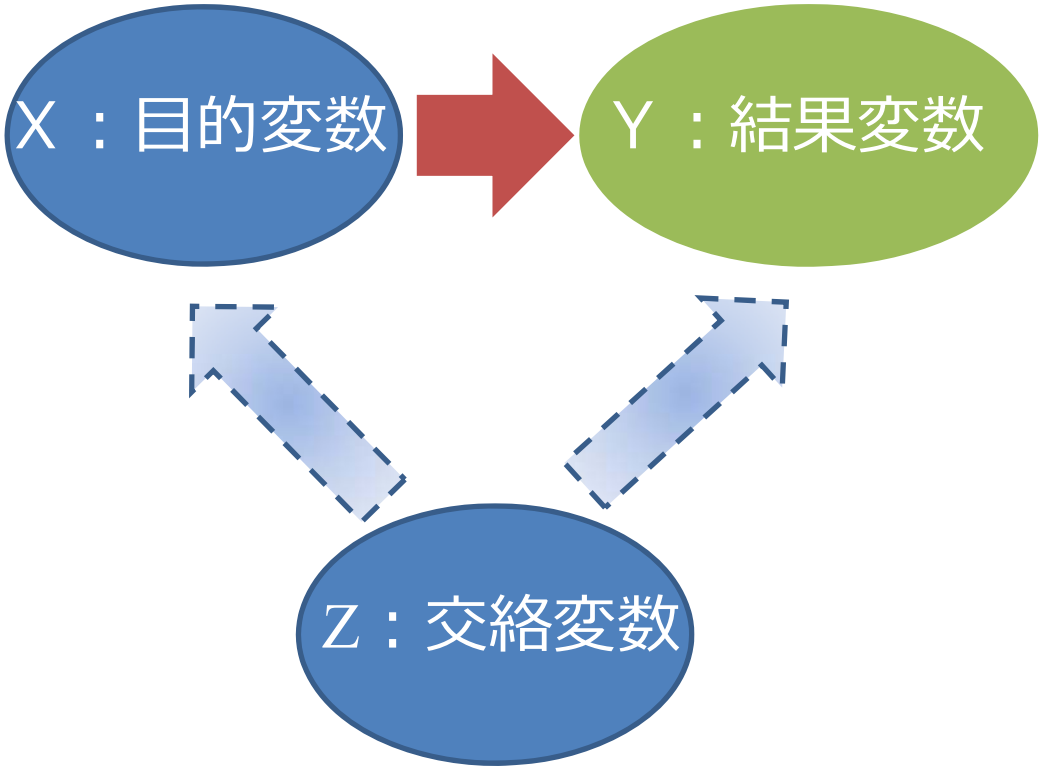
相関構造の分析
主成分分析とクラスター分析
ー分子マーカーデータへの応用を例に

国立国際医療研究センター
医学統計研究室
田中　紀子

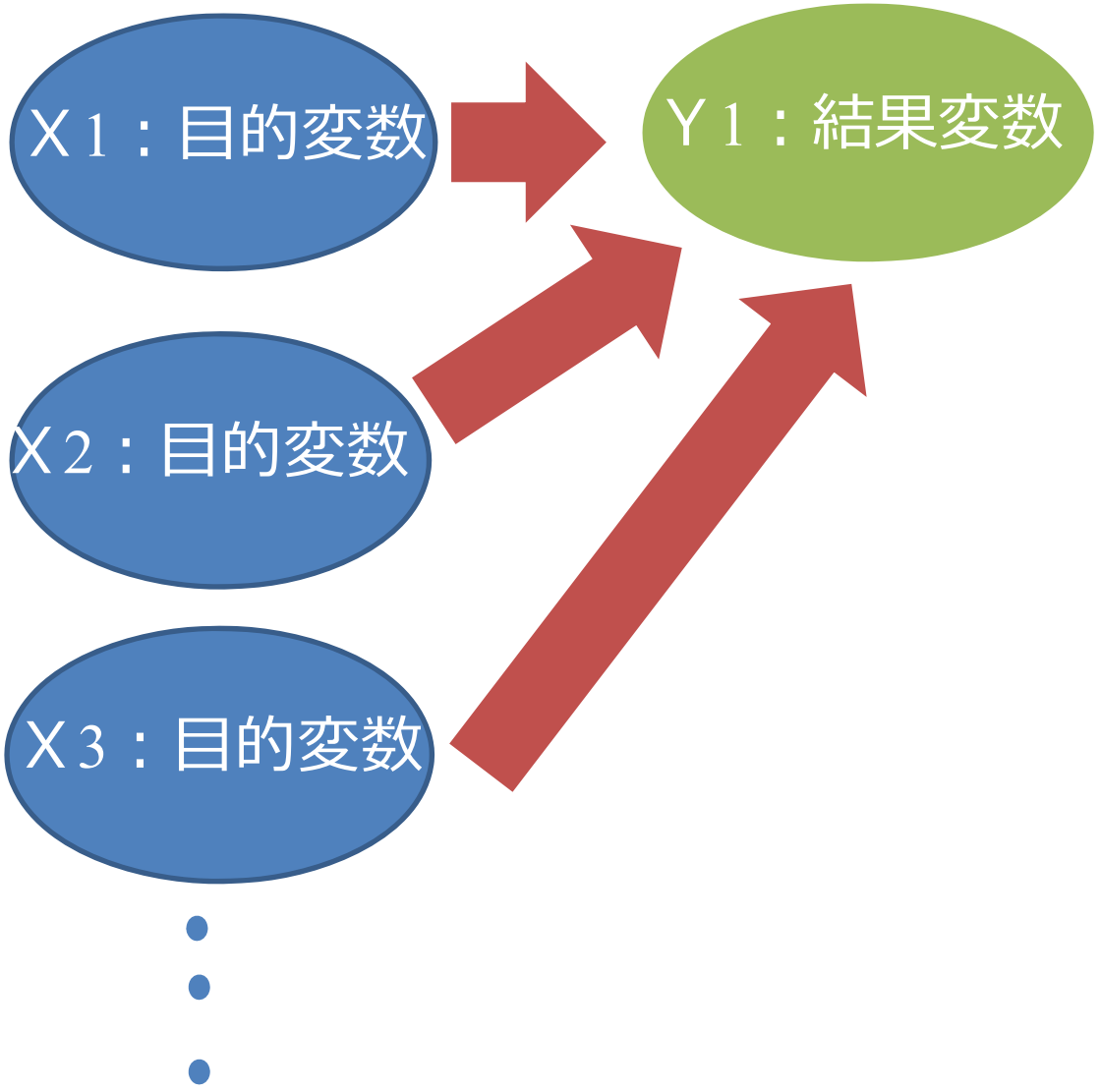
営利目的での利用はご遠慮ください

データを分析する



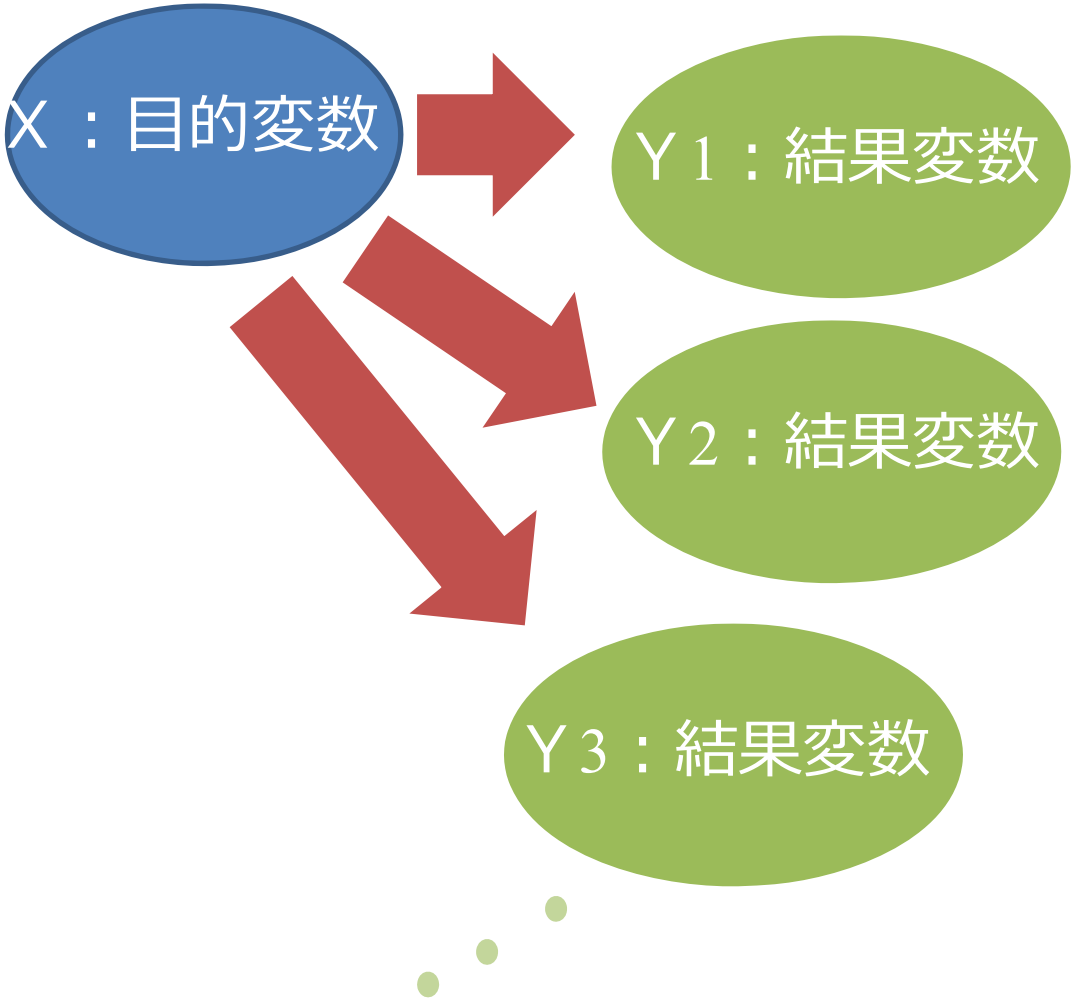
営利目的での利用はご遠慮ください

データを分析する



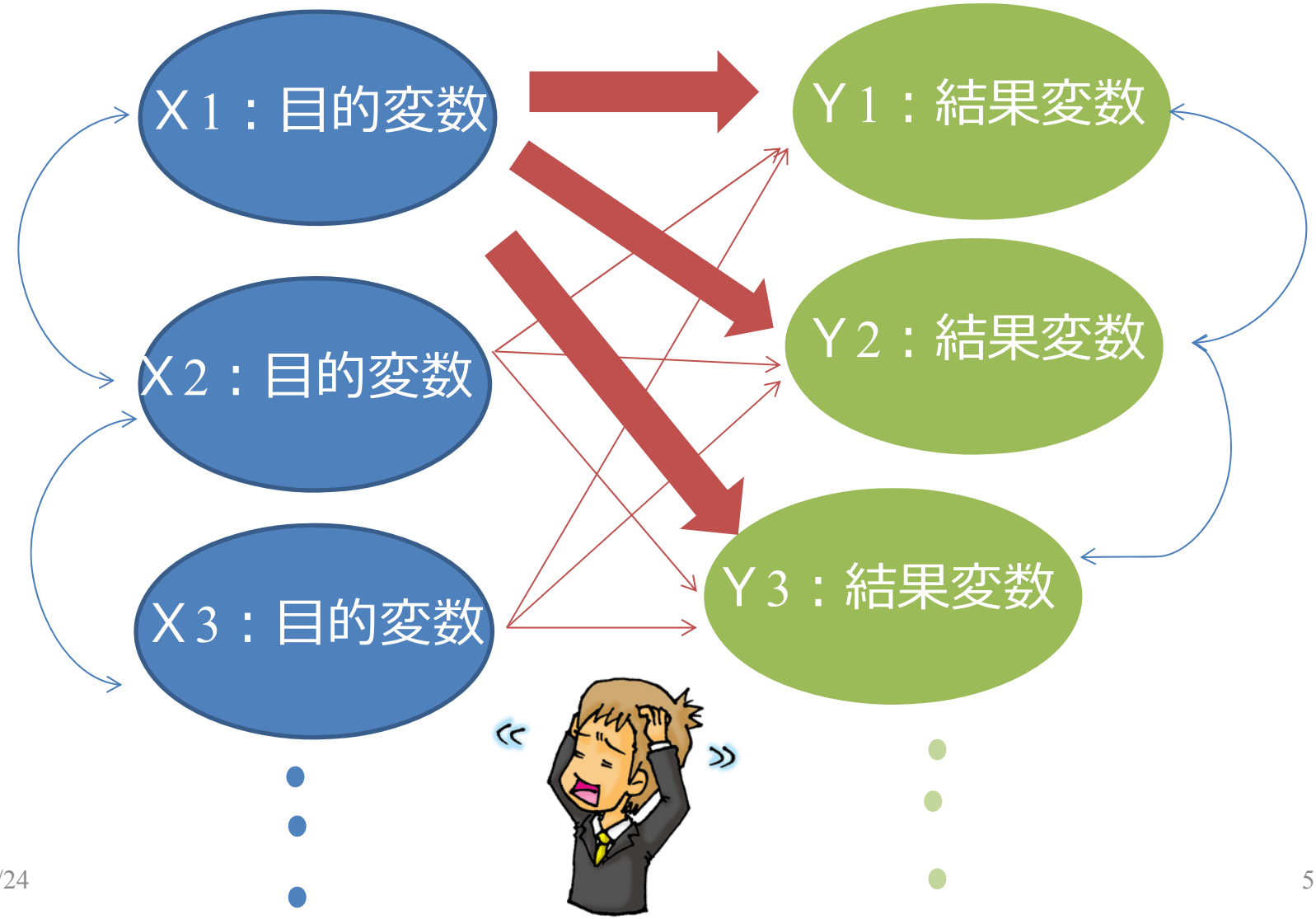
営利目的での利用はご遠慮ください

データを分析する



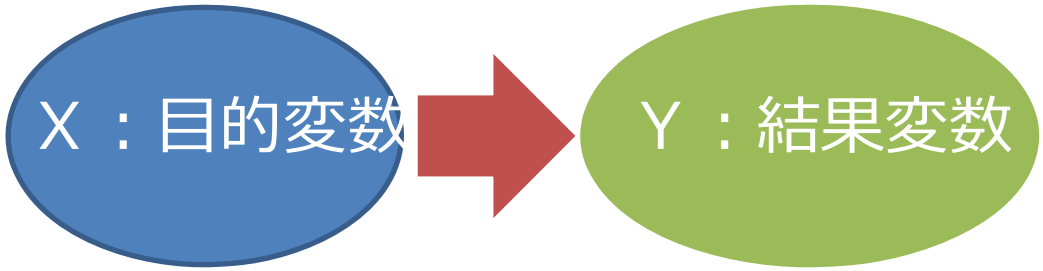
営利目的での利用はご遠慮ください

データを分析する



営利目的での利用はご遠慮ください

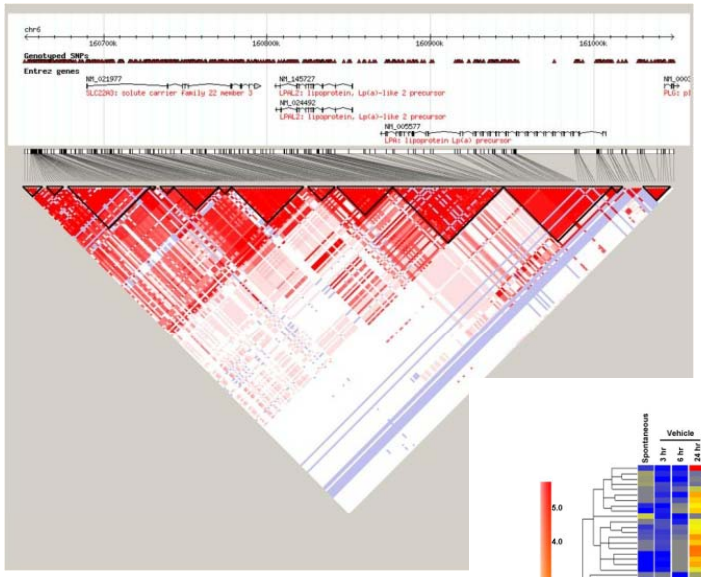
Simpler is the better!

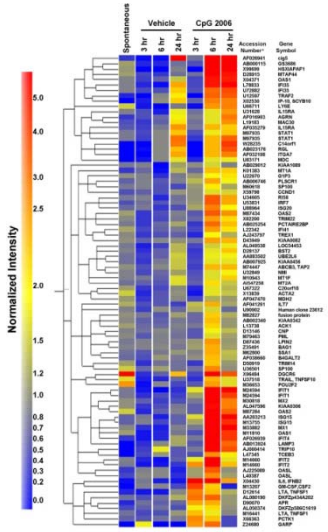


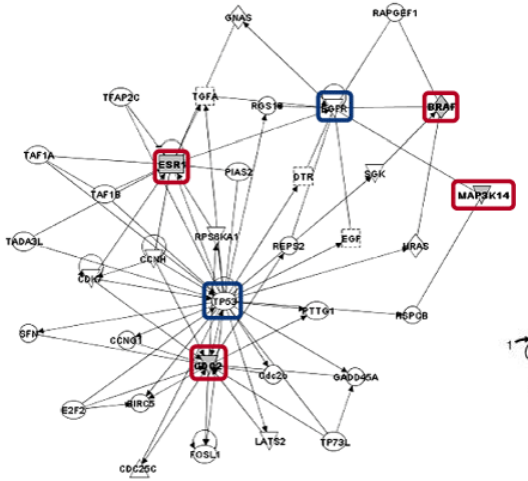
営利目的での利用はご遠慮ください

データの相関構造を探る

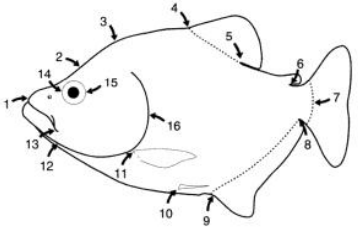
- Genetics, Morphometrics, などでは一般的.

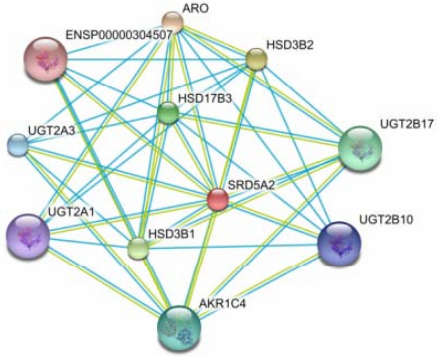


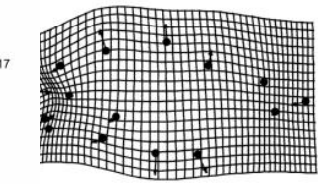




Panel B. WI-38 Cell Hits (BRAF, CDC2, ESR1, and MAPK3K14)







営利目的での利用はご遠慮ください

主成分分析
-Principal Component Analysis (PCA)-

- 互いに相関のある多くの変数から互いに無相関な少数個の変数（主成分）にデータを要約（＝データの次元を縮小：dimensionality reduction）する手法。

- Introduced by Pearson (1901) and Hotelling (1933)

営利目的での利用はご遠慮ください

例：フィッシャーのあやめのデータ

id	Species (Y_1)	Sepal length(x_1)	Sepal Width(x_2)	Petal length(x_3)	Petal Width(x_4)
1	Setosa	50	33	14	2
2	Virginica	64	28	56	22
…					
150	Versicolor	67	30	50	17

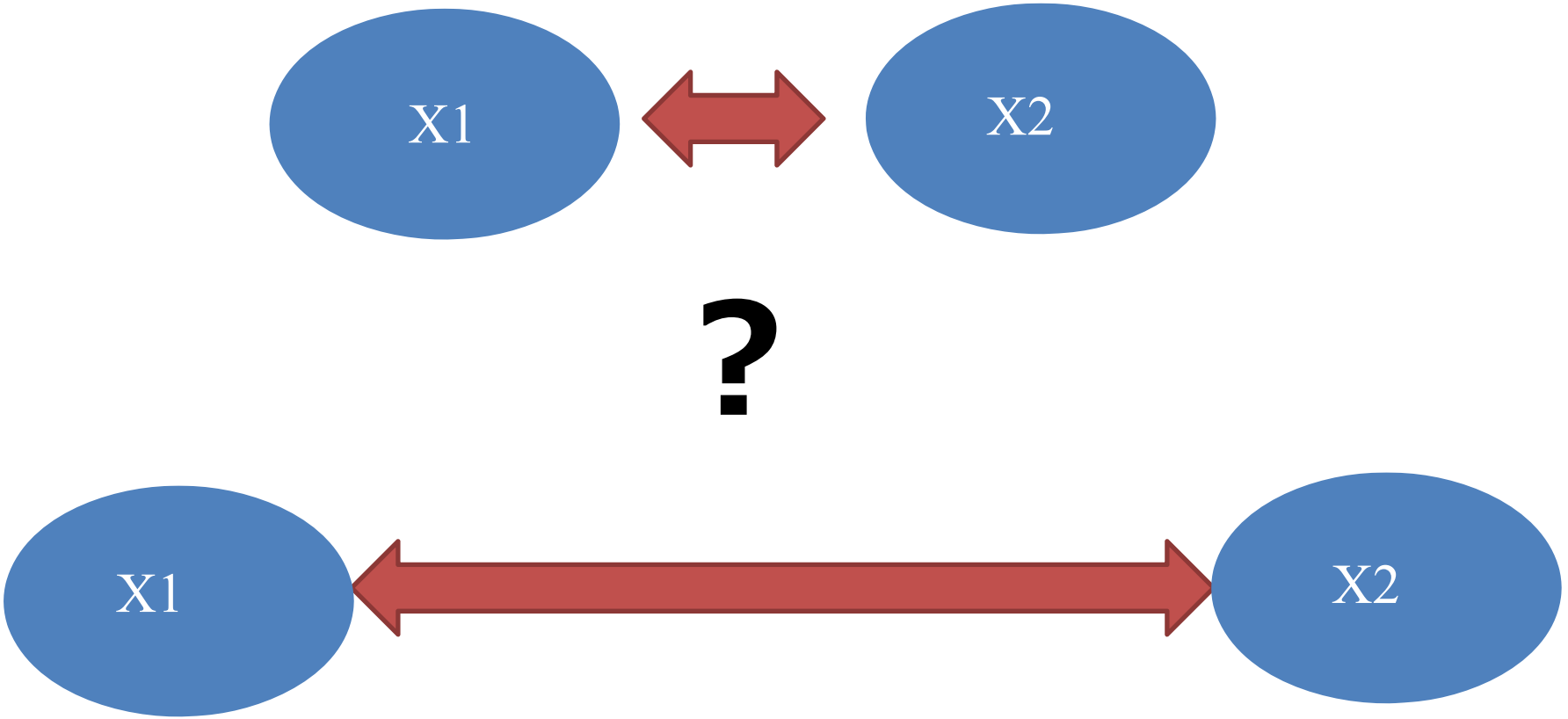
Setosa(n=50)

Virginica(n=50)

Versicolor(n=50)

営利目的での利用はご遠慮ください

2変量の関係

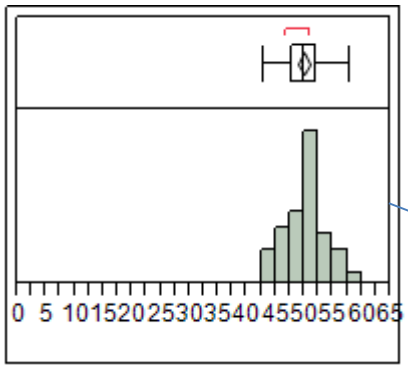


営利目的での利用はご遠慮ください

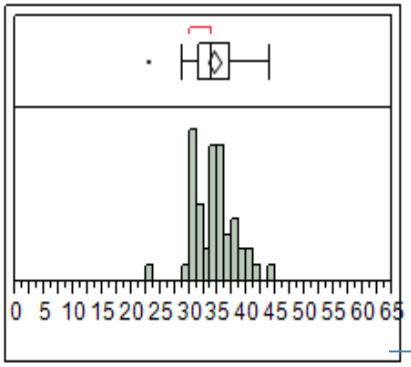
ばらつき（分散）と共分散

-Variance and Covariance-

分散



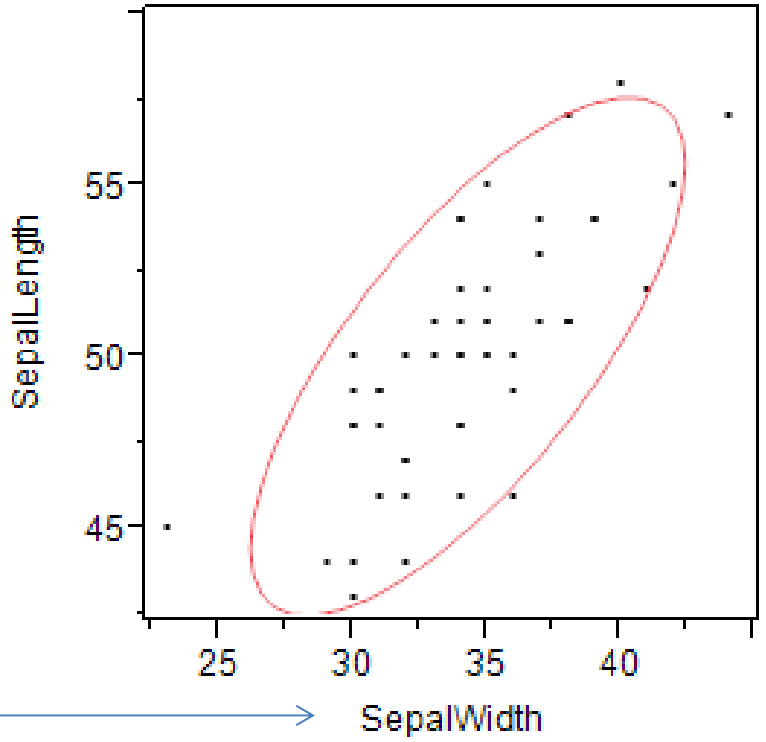
$\mathrm{Var}(X_1)=v(X_1)$
$=\Sigma(X_{i1}-\mu 1)^2/N$
$\fallingdotseq 12.4$



$\mathrm{Var}(X_2)= v(X_2)$
$=\Sigma(X_{i2}-\mu_2)^2/N$
$\fallingdotseq 14.4$

共分散

$\mathrm{Cov}\,(X_1,X_2)= c(x_1,x_2)=\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\overline{x_1})(x_{i2}-\overline{x_2})$ =9.92



共分散＝二つの変数の線形的な関連の強さを測る指標

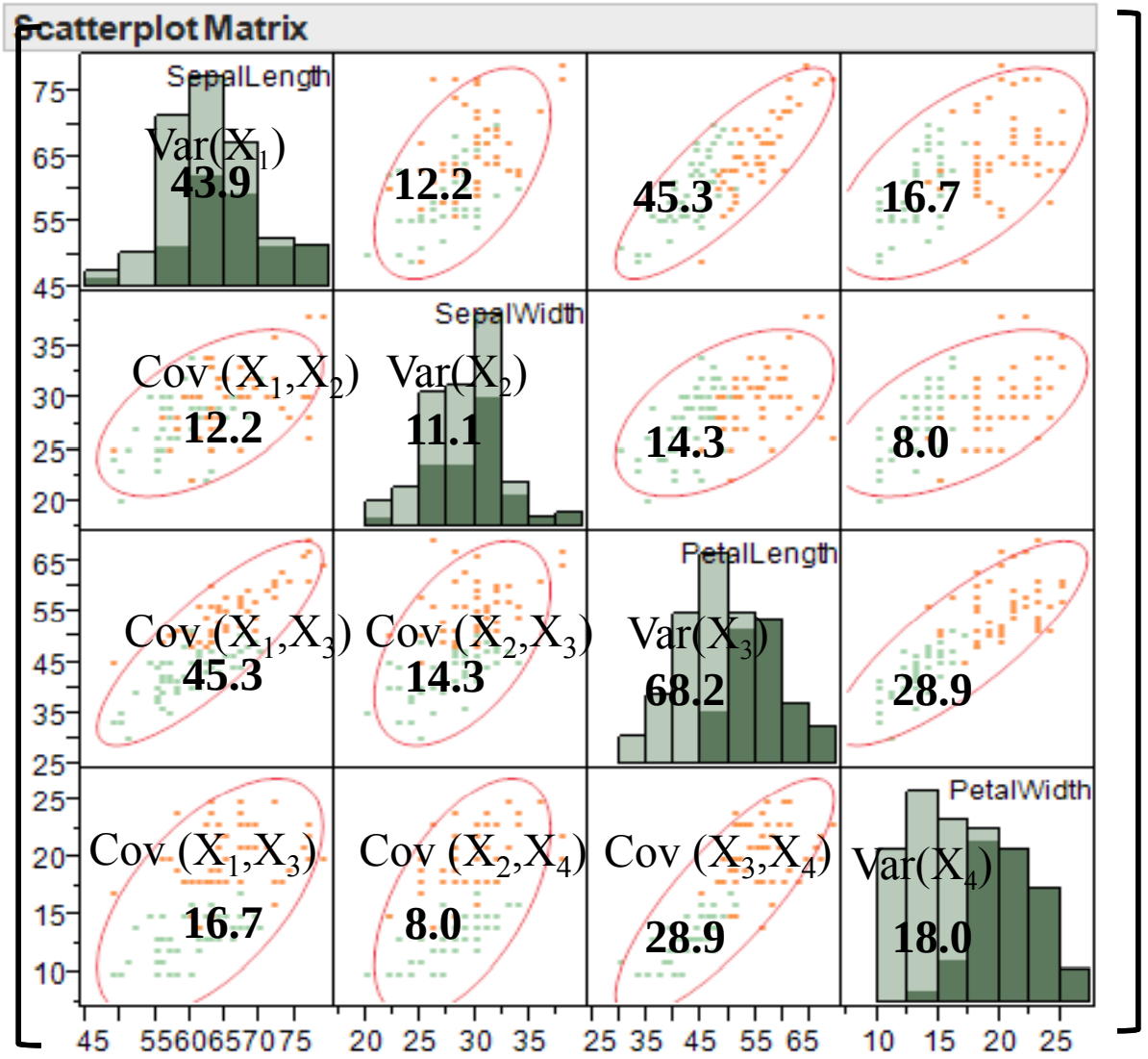
営利目的での利用はご遠慮ください

共分散行列(Covariance Matrix)

Virginica

Versicolor

C=



営利目的での利用はご遠慮ください

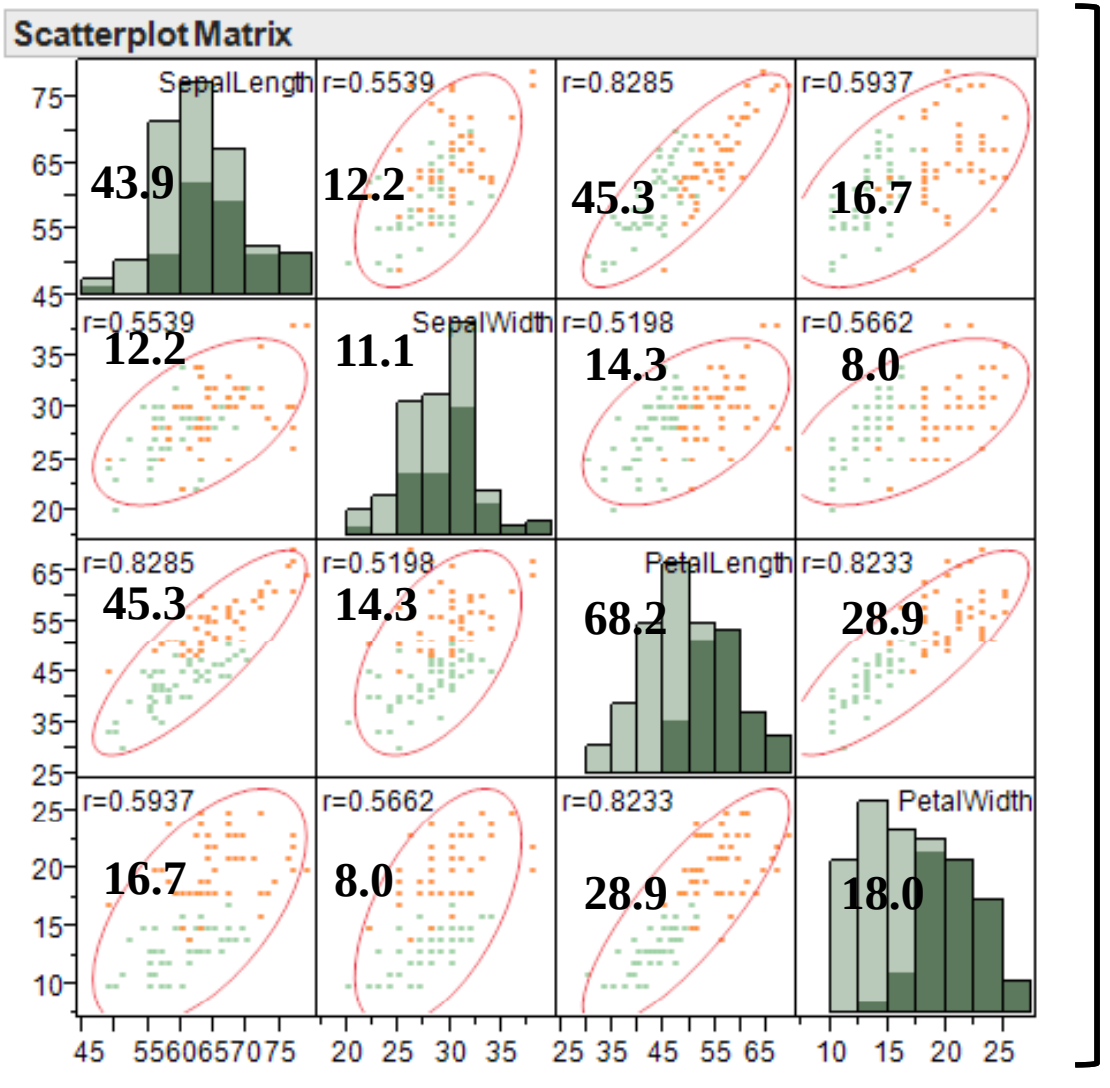
共分散と相関

共分散

$$c(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} - \overline{x_1})(x_{i2} - \overline{x_2})$$

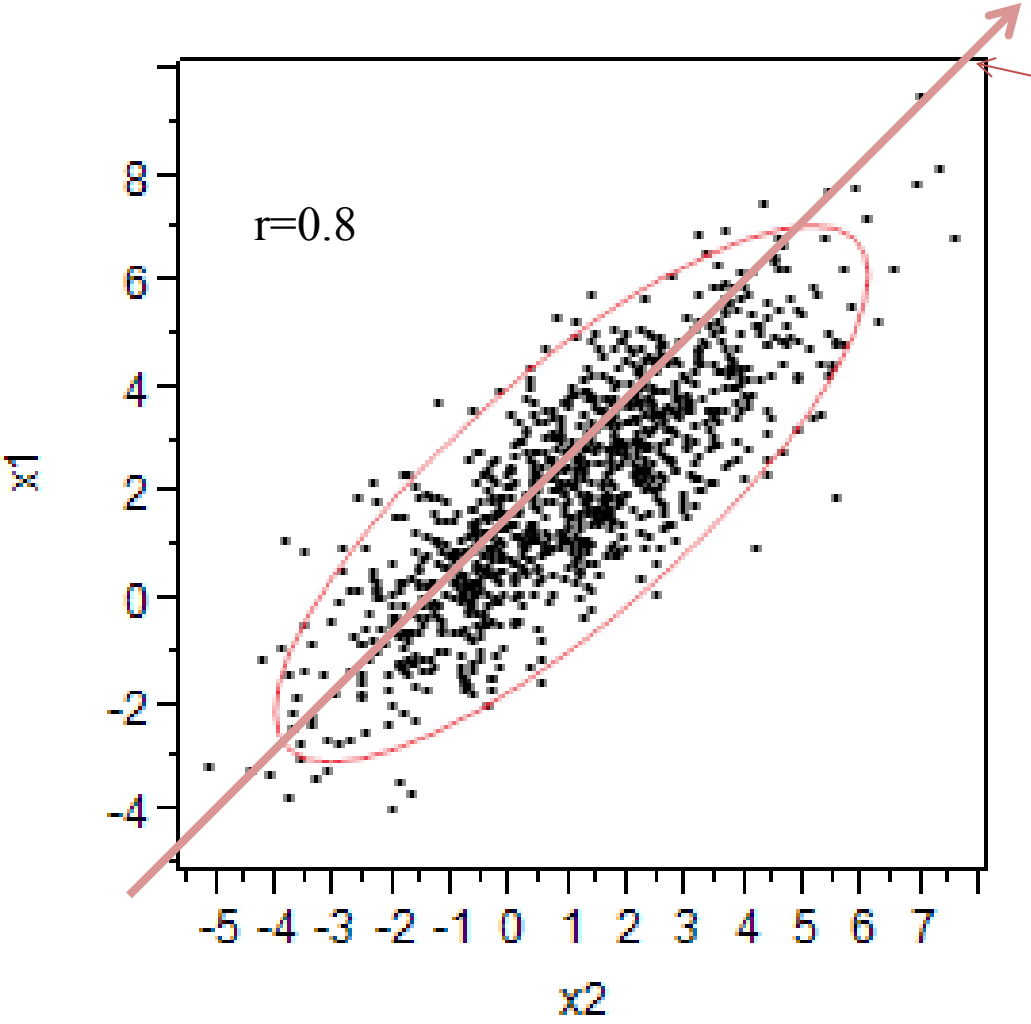
相関係数

$$r(x_1, x_2) = \frac{\sum_{i=1}^{n}(x_{i1} - \overline{x_1})(x_{i2} - \overline{x_2})}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \overline{x_1})^2}\sqrt{\sum_{i=1}^{n}(x_{i2} - \overline{x_2})^2}}$$



営利目的での利用はご遠慮ください

主成分分析の目標



この線の上に
データがすべて
乗ってくれてれ
ばなあ・・・

営利目的での利用はご遠慮ください

主成分の計算

k 個の変数: $x_1, x_2, \ldots, x_k$ から

k 個の新しい変数: $Prin_1, Prin_2, \ldots, Prin_k$: を共分散行列を使って計算

$$Prin_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$$

$$Prin_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k$$

$$\ldots$$

$$Prin_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k$$

ここで

$Prin_k$ はすべて独立

$Prin_1$:

できる限りもとデータのバラツキを説明するように合成された新しい変数

$Prin_2$:

残りのバラツキを説明するように合成された新しい変数

営利目的での利用はご遠慮ください

主成分の計算のために・・
固有値と固有ベクトルの計算

- 固有値λ_i は以下の数式を解くと得られます

$$det(C-\lambda I)=0$$

- 固有ベクトルは次のような条件を満たす行列の列になります

$$C=A\ D\ A^T$$

- ここで

$$D=\begin{pmatrix} \lambda 1\ 0\ .....0 \\ 0 \quad \lambda_2 .....0 \\ 0 \\ 0\ .....\lambda_p \end{pmatrix}$$

営利目的での利用はご遠慮ください

例：2変数の場合・・固有値の計算

- 二つの変数の共分散 $c>0$
- $\mathbf{C}=\begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$　　$\mathbf{C}-\lambda\mathbf{I}=\begin{pmatrix} 1-\lambda & c \\ c & 1-\lambda \end{pmatrix}$

$$det(\mathbf{C}-\lambda\mathbf{I})=(1-\lambda)^2-c^2$$

- 解くと

$$\lambda_1 = 1+c$$

$$\lambda_2 = 1-c < \lambda_1$$

営利目的での利用はご遠慮ください

例：2変数の場合・・固有ベクトルの計算

- すべての固有ベクトルA は以下の条件を満たすので

$$\mathbf{C}A=\lambda A$$

$$A=\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad CA=\begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}=\begin{pmatrix} a_1+ca_2 \\ ca_1+a_2 \end{pmatrix}=\begin{pmatrix} \lambda a_1 \\ \lambda a_2 \end{pmatrix}$$

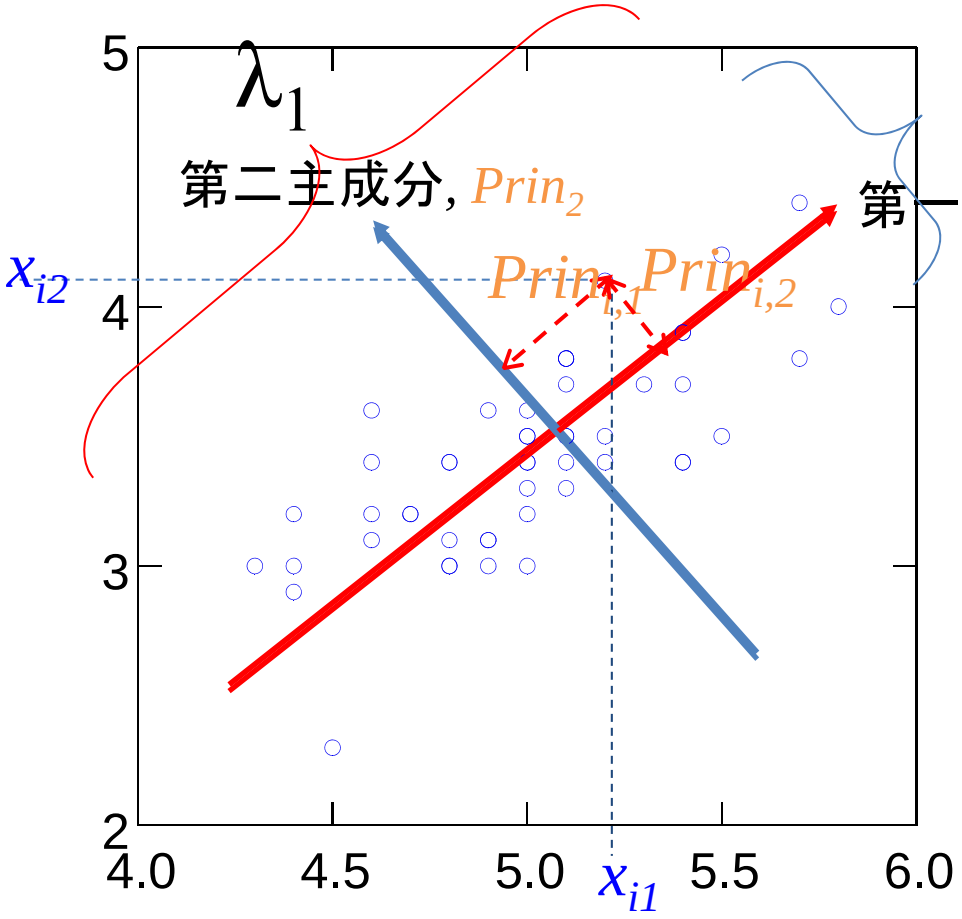
解は

$$A_1=\begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \cdot \quad A_2=\begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

営利目的での利用はご遠慮ください

主成分と固有値と固有ベクトル

固有値： λ_i



固有ベクトル

営利目的での利用はご遠慮ください

主成分と固有値の解釈

- 新しく作った合成変数（主成分）の分散は固有値に一致する。

$$Var(Prin_i) = \lambda_i \text{ for all } i=1\ldots p$$

- 小さい固有値λ_i ⇔ 分散小⇔ 主成分$Prin_i$の方向へはデータはほとんど変化しない

- $\lambda_i / \Sigma \lambda_i$ ＝それぞれの主成分によってデータが説明される割合

営利目的での利用はご遠慮ください

主成分と固有ベクトルの解釈

固有値

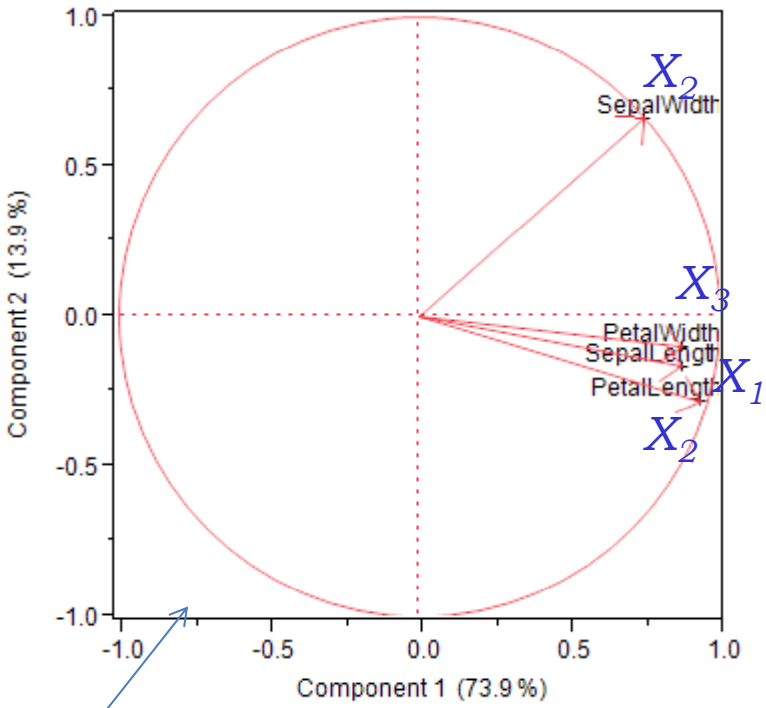
Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent	ChiSquare	DF	Prob>ChiS
1	2.9579	73.948		73.948	285.328	4.057	<.000
2	0.5551	13.878		87.826	80.531	4.364	<.000
3	0.4057	10.143		97.969	56.645	1.754	<.000
4	0.0812	2.031		100.000	0.000	.	

固有ベクトル

Eigenvectors

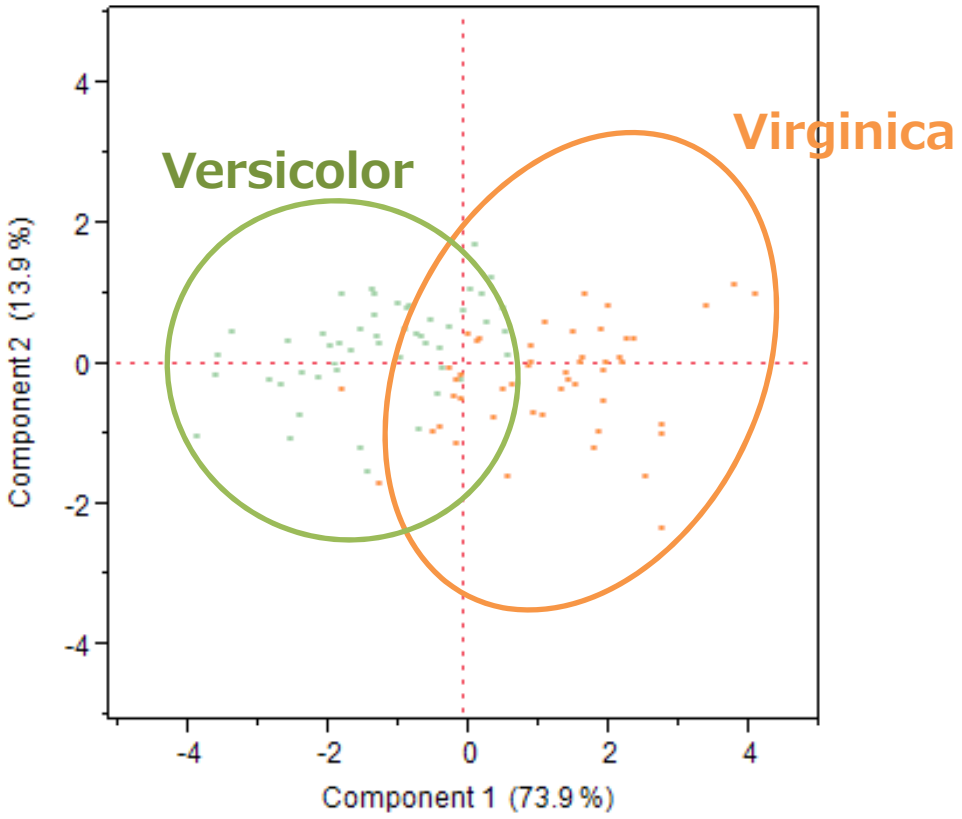
	Prin1	Prin2	Prin3	Prin4
SepalLength	0.50713	-0.22061	0.69307	0.46238
SepalWidth	0.43475	0.88845	0.05559	-0.13625
PetalLength	0.54370	-0.37959	-0.01944	-0.74829
PetalWidth	0.50814	-0.13381	-0.71846	0.45575



- $Prin_1 = 0.51X_1 + 0.43X_2 + 0.54X_3 + 0.51X_4$
- $Prin_2 = -0.22X_1 + 0.88X_2 - 0.37X_3 - 0.37X_4$
- X_2 が第2主成分(Prin2)にとって重要な変数
- 可視化→circle of orrelation

営利目的での利用はご遠慮ください

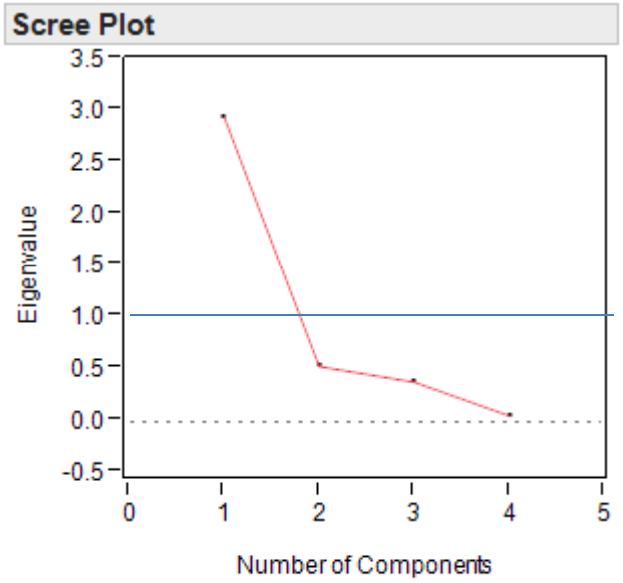
主成分プロット



営利目的での利用はご遠慮ください

主成分の数

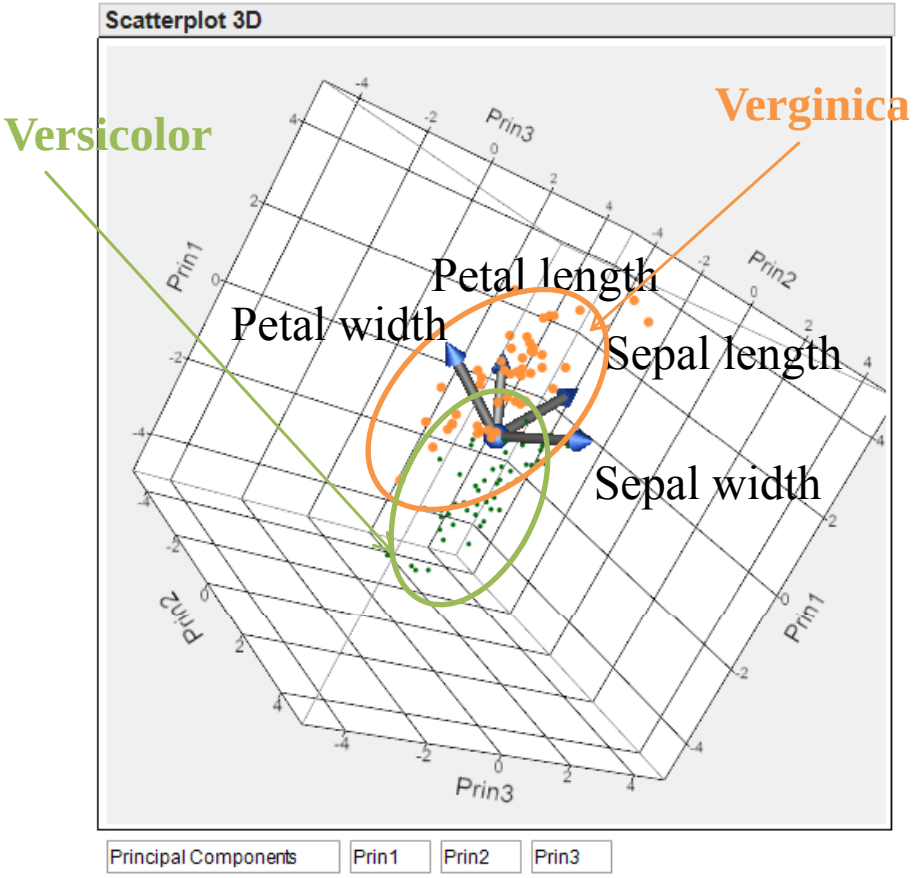
- 50-70%以上データを説明している主成分の数まで、次元を縮小する
- Kaiser criterion: 固有値>1
- Scree plot:

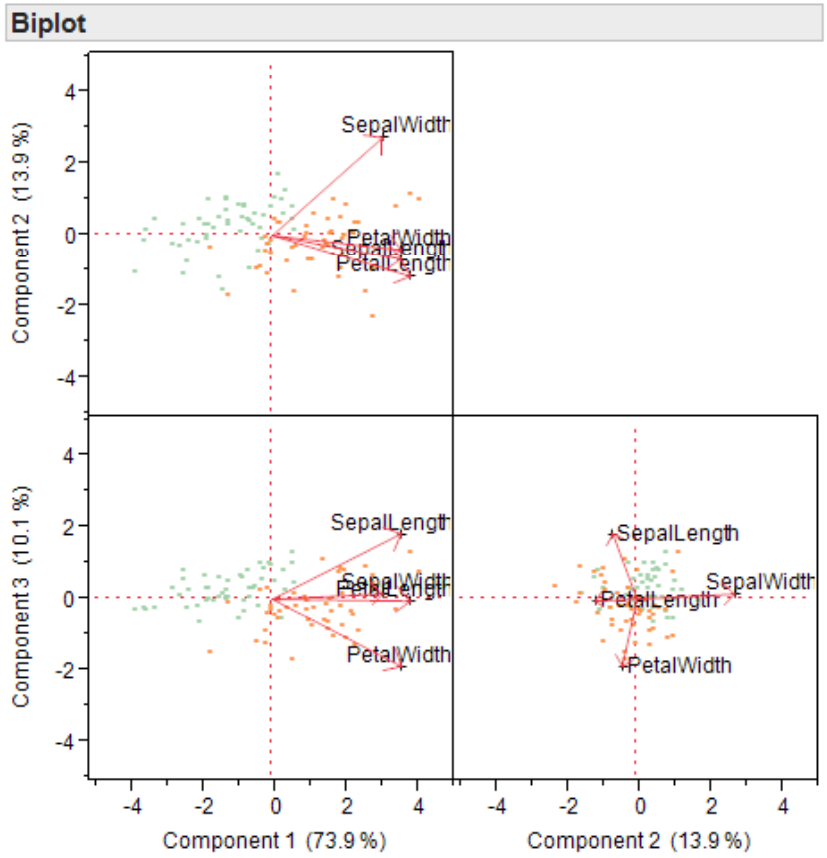


営利目的での利用はご遠慮ください

バイプロット(biplot)

- 変数と人（オブザーベーション）の同時プロット
- 主成分プロットとcorrelation of circleの重ね合わせ図





営利目的での利用はご遠慮ください

分析の注意点

- 主成分分析は、共分散行列を用いているので、変数のスケールをそろえる必要性あり。
  - 変数のスケールを線形変換しただけで、結果は変わる。
    - 標準化
    - 共分散行列の代わりに相関行列を用いる
- 主成分スコアは、解釈が難しい

営利目的での利用はご遠慮ください

分子マーカーデータへの応用

- 変数の多い・・・
- 遺伝子発現データの解析
- 遺伝子多型データの解析
- エピジェネティクスデータの解析
- ・・・

営利目的での利用はご遠慮ください

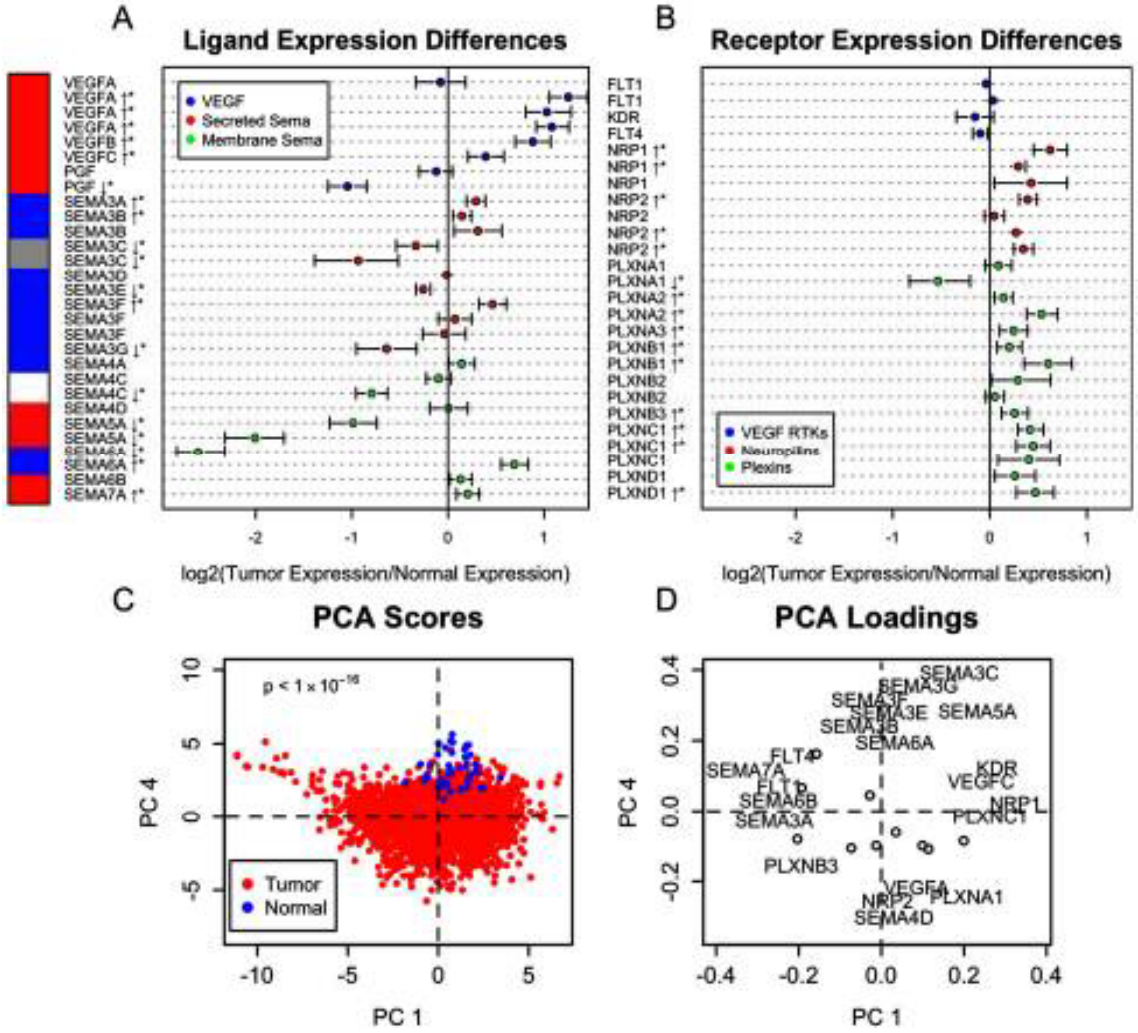
遺伝子発現データの解析

- 例：Triple negative breast cancer(TNBC)患者のVEGF とセマフォリン関連遺伝子群の発現解析　(Bender and Mac Gabhann.Plos One 2013)
  - TNBC 2656サンプル, Normal 42サンプル
  - Normal tissueと発現に差があるのか
  - ＴＮＢＣのサブタイプで発現に差があるのか

営利目的での利用はご遠慮ください

TNBCデータの解析

Bender and Mac Gabhann.Plos One 2013

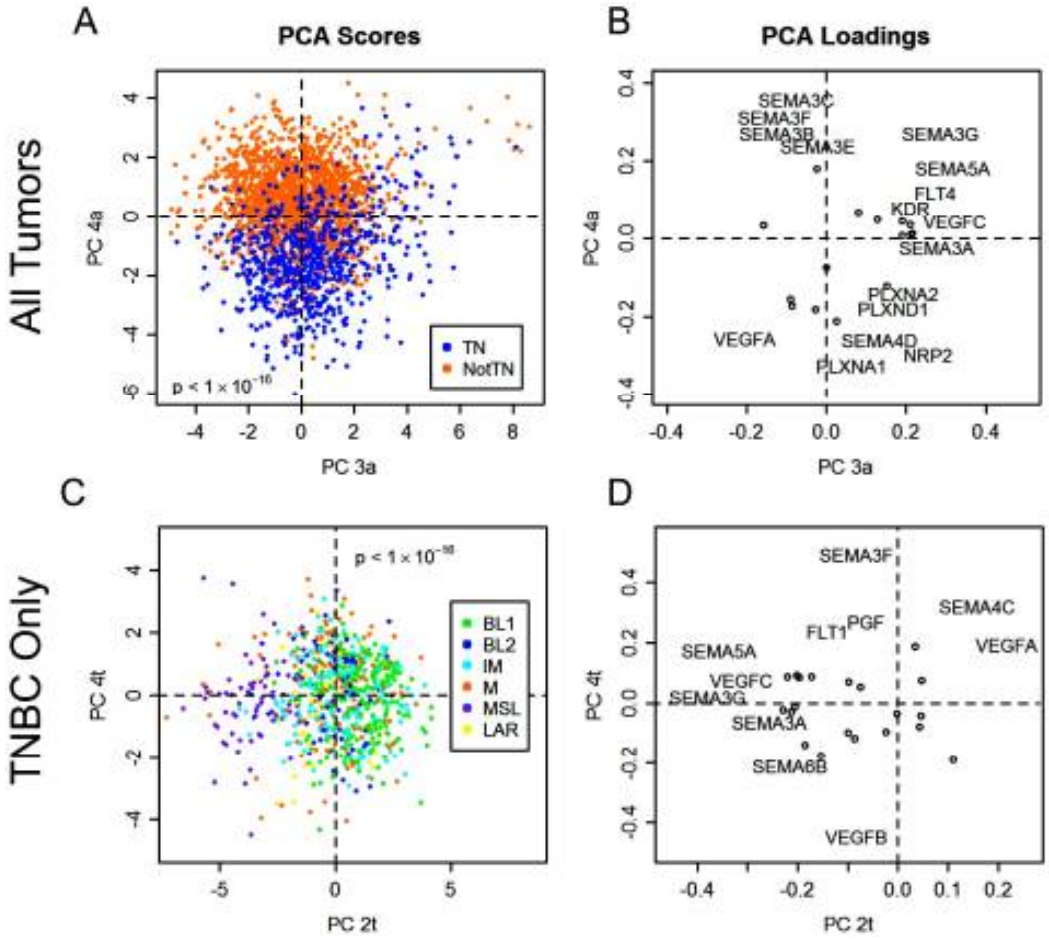


Differences in expression patterns of VEGF- and semaphorin-related genes between normal breast tissue (n=42) and breast tumors (n=2656).

営利目的での利用はご遠慮ください

TNBCデータの解析

Bender and Mac Gabhann.Plos One 2013



Triple negative breast cancers and the mesenchymal stem-like (MSL) subtype of triple negative breast cancers are associated with increased pro-angiogenic gene expression and decreased anti-angiogenic gene expression.

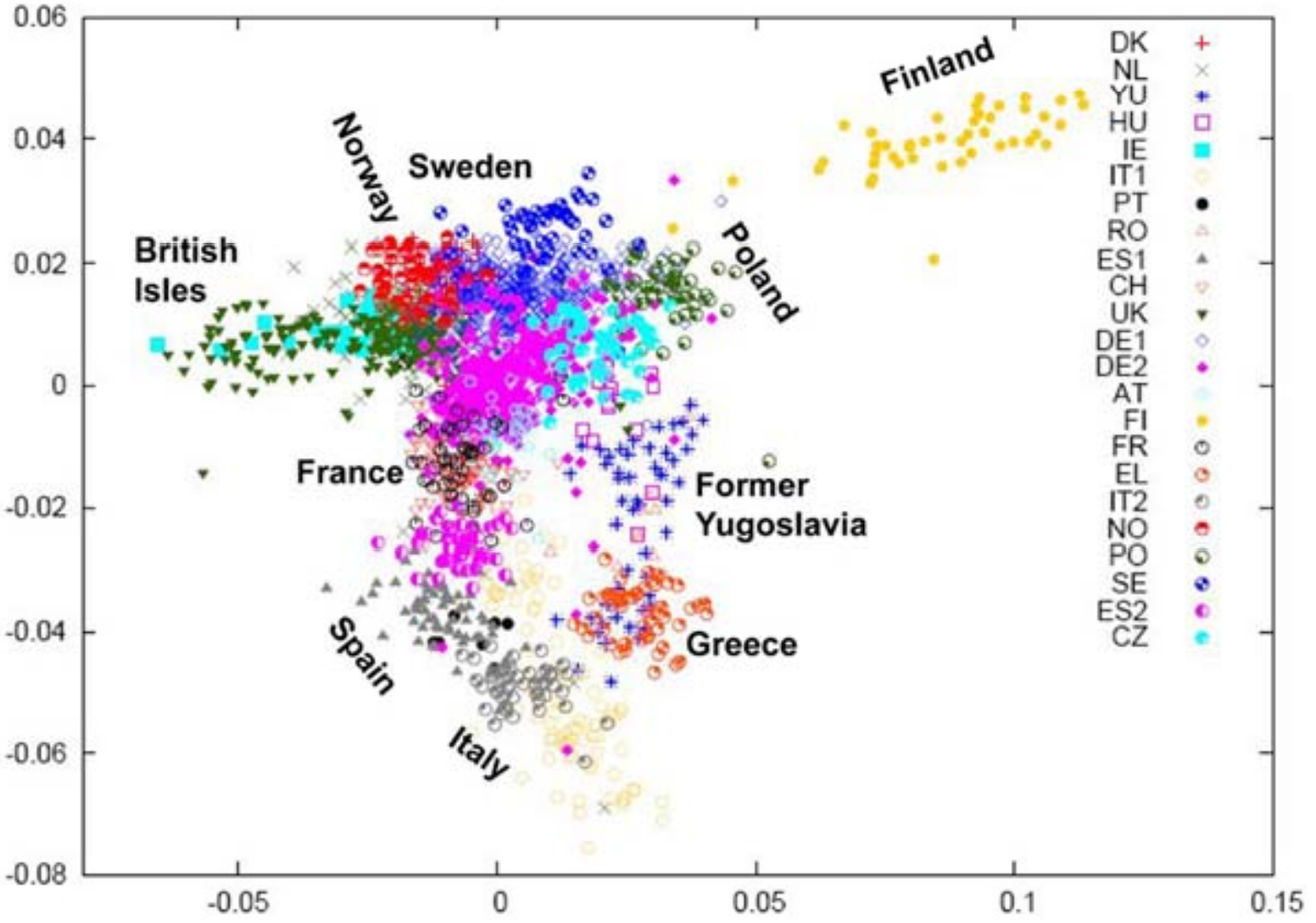
営利目的での利用はご遠慮ください

GWASデータを用いた
集団層化（population stratification) の分析

- 500K <markers で個人の遺伝情報→集団の情報
- 遺伝分散（分散共分散行列）＝ C
- 主成分分析で潜在的集団や測定された集団の遺伝学的背景を分析

営利目的での利用はご遠慮ください

ヨーロッパ集団の遺伝的背景



- Razib, Current Biology 2008

営利目的での利用はご遠慮ください

営利目的での利用はご遠慮ください

データを分類する

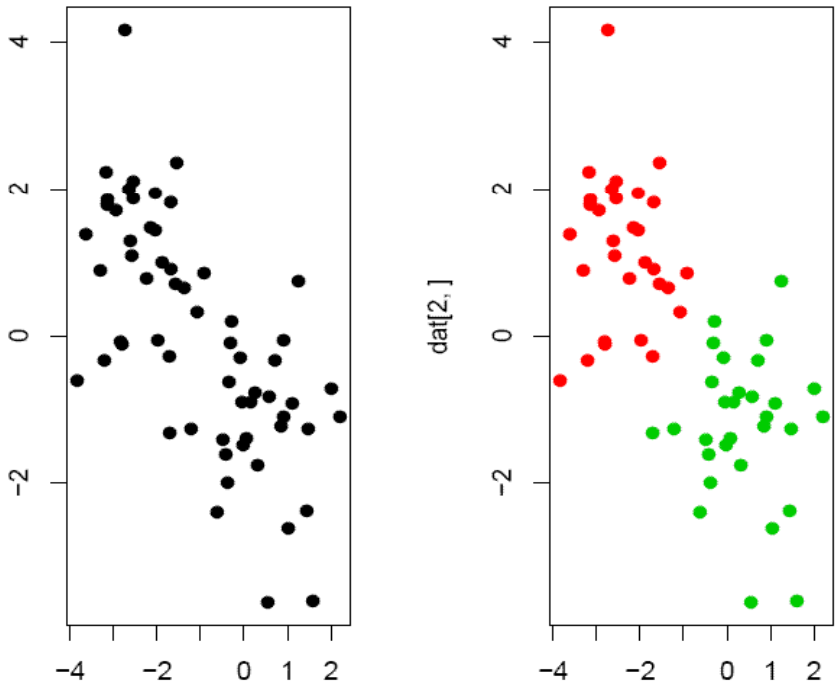
- 教師付き分類（Supervised learning)
  - 判別分析、ＳＶＭ、決定木など・・

- 教師なし分類 (Unsupervised learning)
  - クラスター分析
  - 直接変数（人）を分類する

- 主成分分析
  - 変数（人）をまとめて新しい変数で説明できるかどうかを観察することで、全体からはずれた人を見つけたり、人をグループに分けられるか検討できる。
  - 分類すること自体は実は分析の目的ではない。

営利目的での利用はご遠慮ください

クラスター分析の目的1：
オブザーベーションの分類（例：遺伝マーカー）



- Quality control: 実験がうまくいってないサンプルを見つける
- “既知”の郡分けのバリデーション
- 新しい郡分けの可能性を探る

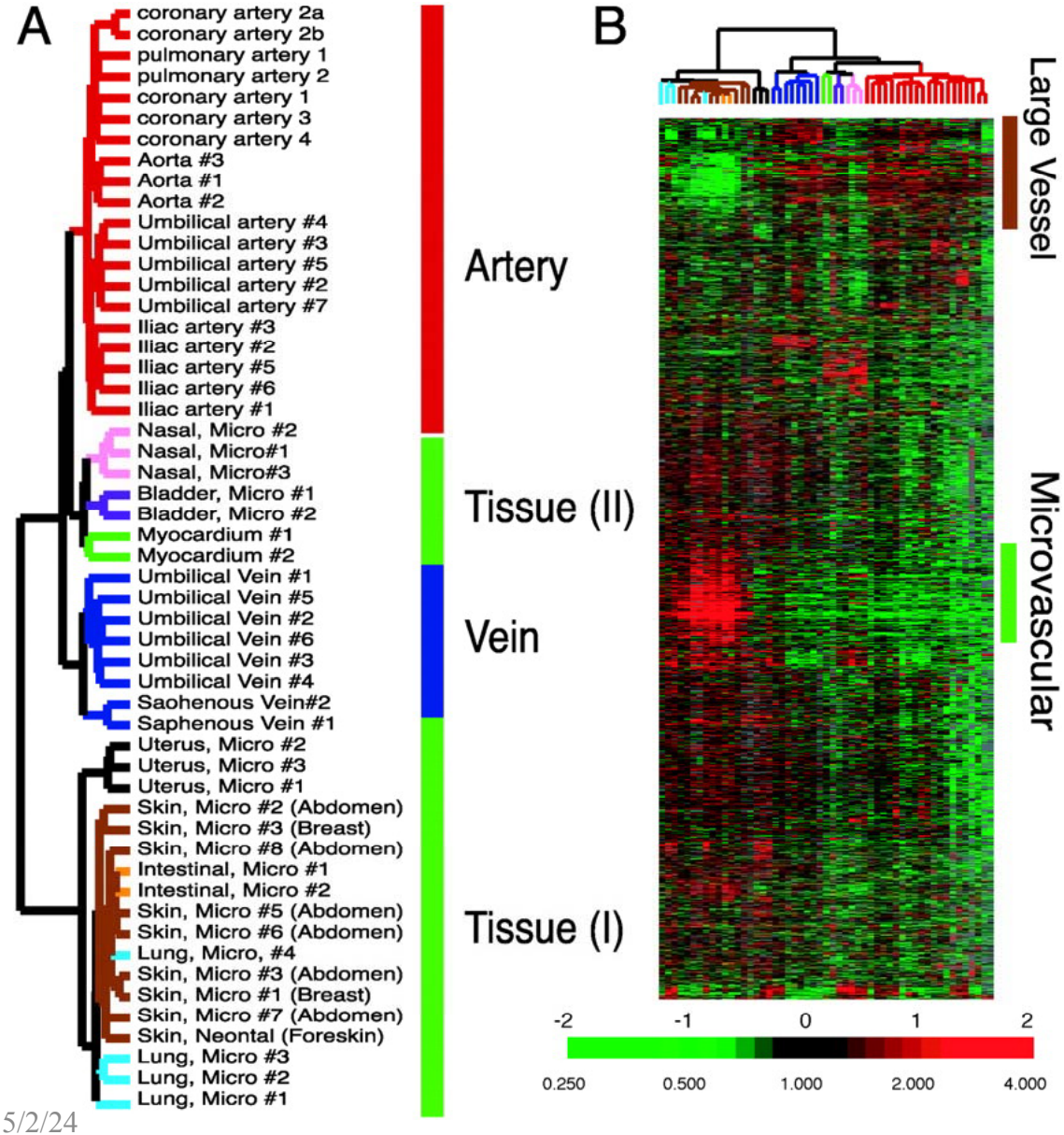
営利目的での利用はご遠慮ください

クラスター分析の目的2:
変数(例:遺伝マーカー)の分類

- 同時制御されている遺伝子を探れる（かも）.
- 時間・空間的に発現パターンが似ている遺伝子群が見つかる（かも）.
- 関係ない／ただのノイズ変数を分別できる（かも）.

営利目的での利用はご遠慮ください

ヒートマップ



Chi et al., PNAS | **September 16, 2003** | vol. 100 | no. 19 | **10623-10628**

“Endothelial cell diversity revealed by global expression profiling”

営利目的での利用はご遠慮ください

クラスター分析

クラスター分析には二つの材料が必要:

- **Distance measure**:
  人（変数）間距離を測る指標.
- **Cluster algorithm:**
  人（変数）をまとめたグループの作り方

営利目的での利用はご遠慮ください

distance measures

二つの変数*x*, *y*の*n*人分のデータ

$\boldsymbol{x} = (x_1, \ldots, x_n), \boldsymbol{y} = (y_1, \ldots, y_n)$

- ユークリッド距離: $d_E(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- マンハッタン距離: $d_M(x,y) = \sum_{i=1}^{n}\left|x_i - y_i\right|.$
- 相関距離: $d_C(x,y) = 1 - \frac{\sum_{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}(x_i - \bar{x})^2 \sum_{i=1}(y_i - \bar{y})^2}}.$

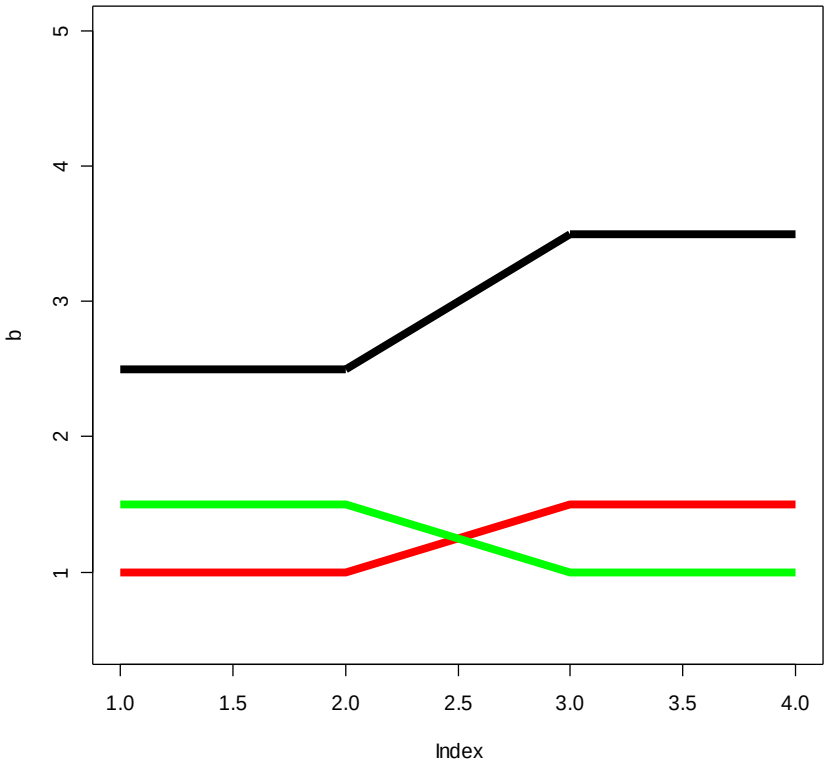
営利目的での利用はご遠慮ください

どのdistance measure を使う?

- 分野（データ）によってきまっている場合もある。

- 相関距離は、相対的な距離:

$d_c(x, y) = d_c(ax+b, y)$ if $a > 0$.



$\boldsymbol{x} = (1, 1, 1.5, 1.5)$
$\boldsymbol{y} = (2.5, 2.5, 3.5, 3.5) = 2\boldsymbol{x} + 0.5$
$\boldsymbol{z} = (1.5, 1.5, 1, 1)$

$d_c(\boldsymbol{x}, \boldsymbol{y}) = 0$, $d_c(\boldsymbol{x}, \boldsymbol{z}) = 2$.
$d_E(\boldsymbol{x}, \boldsymbol{z}) = 1$, $d_E(\boldsymbol{x}, \boldsymbol{y}) \sim 3.54$.

営利目的での利用はご遠慮ください

どの distance measure を使う?

- ユークリッド距離とマンハッタン距離はどちらも絶対距離. マンハッタン距離のほうがよりはずれ値に頑健.

- 標準化（standardization）: $x \mapsto \frac{x - \overline{x}}{\hat{\sigma}_x}$

を行うと、相関距離とユークリッド距離は同じ:

$$d_E(x_1, x_2)^2 = 2n d_C(x_1, x_2).$$

営利目的での利用はご遠慮ください

主な Clustering　アルゴリズム

- 分割法（Partitioning approach）:
  - ある規則に基づいて、データを分割していく方法：例、2乗和を最小にするなど。
  - k-means, k-medoids, CLARANS
- 階層的方法（Hierarchical approach）:
  - ある規則に基づいて、データから階層構造を作っていく方法
  - Diana, Agnes, BIRCH, ROCK, CAMELEON
- 密度に基づく方法（Density-based approach）:
  - データの密度や密度関数に基づいて分類する方法
  - DBSACN, OPTICS, DenClue

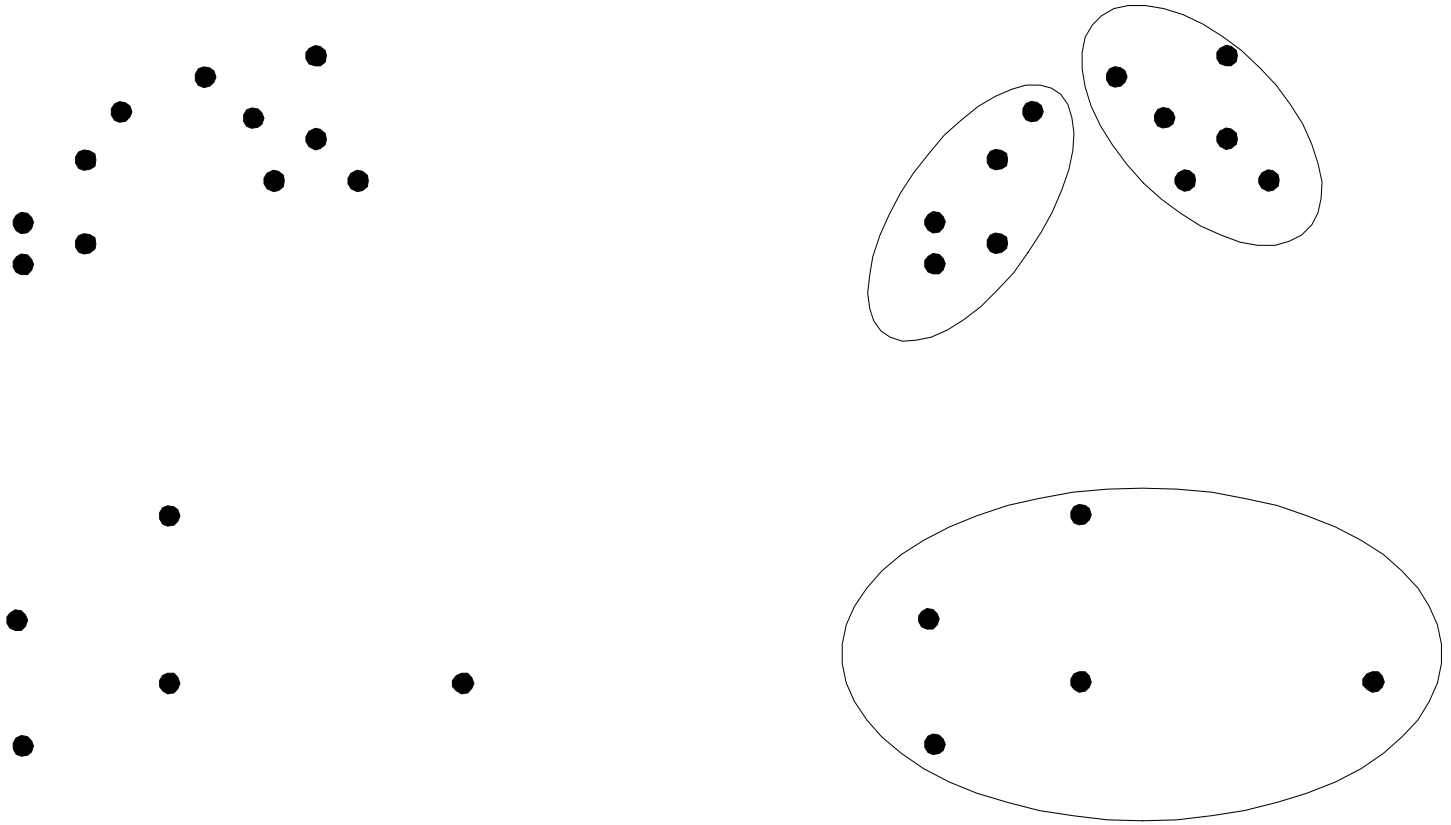
営利目的での利用はご遠慮ください

主なClustering アルゴリズムーおまけ

- 格子を用いる手法（Grid-based approach）:
  - データをいくつか細かく分割して、同じ細分割内のデータを均質に扱う方法
  - STING, WaveCluster, CLIQUE
- モデルに基づく方法（Model-based）:
  - 仮説に基づいたモデルを構築し、そこにフィットするクラスターを形成させる
  - EM, SOM, COBWEB
- データのパターンに基づく方法（Frequent pattern-based）:
  - pCluster

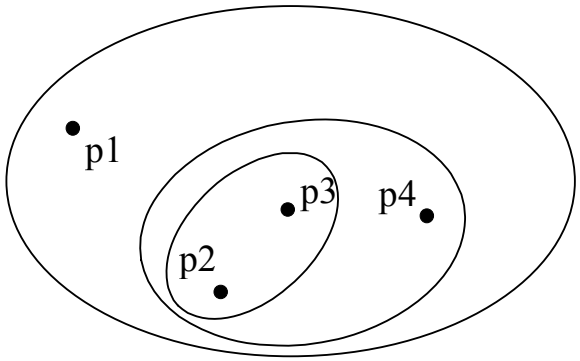
営利目的での利用はご遠慮ください

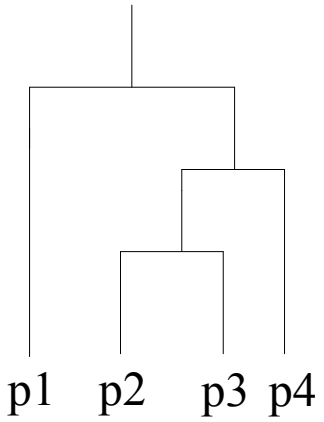
分割法

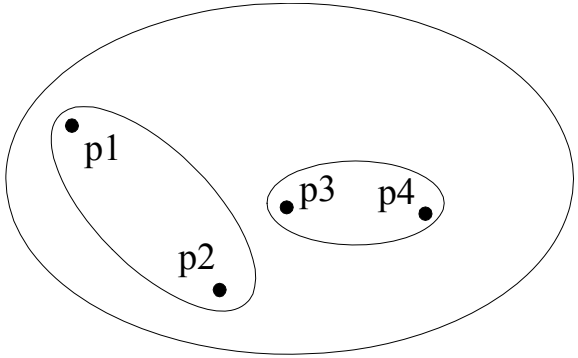


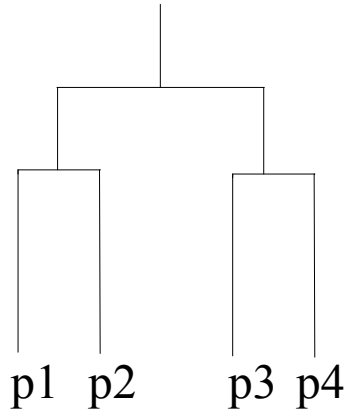
営利目的での利用はご遠慮ください

階層的方法









営利目的での利用はご遠慮ください

K-means法

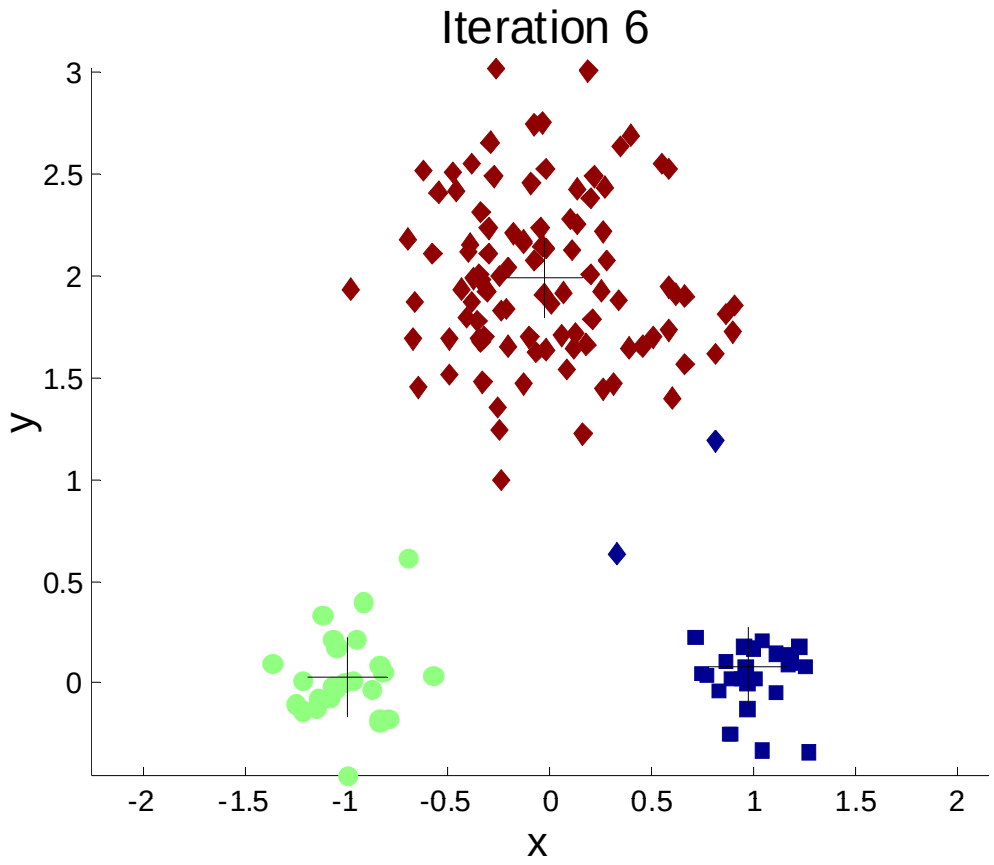
- 分割法
- クラスターの重心点 （セントロイド：centroid） を各クラスターの代表点とする
- それぞれの点（データ）はもっとも近いセントロイドのあるクラスターに分類される
- 分析の前に、クラスターの数、Kは決めなければいけない

以下アルゴリズム

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

営利目的での利用はご遠慮ください

K-meansクラスタリング法適用例



営利目的での利用はご遠慮ください

あやめのデータへの適用

K Means NCluster=3

Columns Scaled Individually

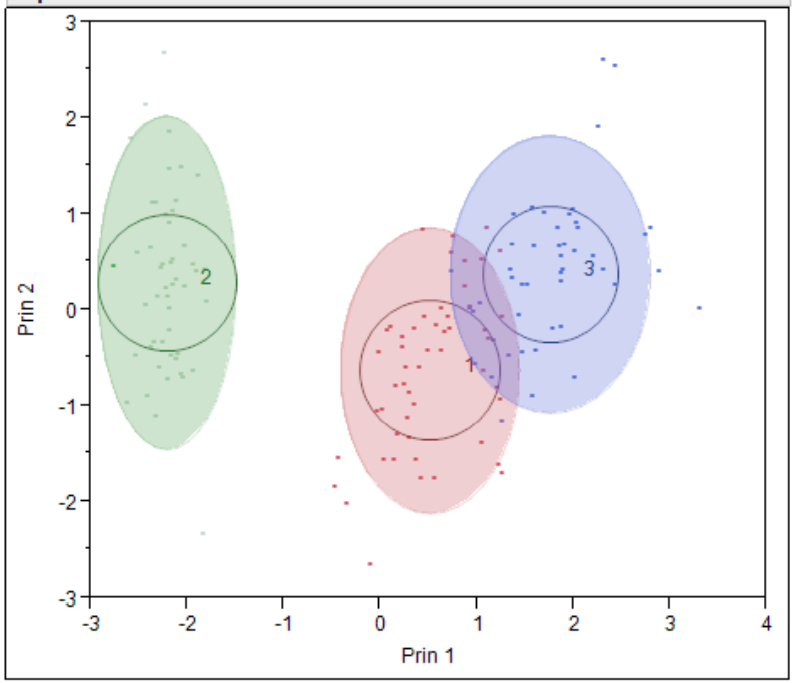
Cluster Summary

Cluster	Count	Step	Criterion
1	52	9	0
2	50		
3	48		

Cluster Means

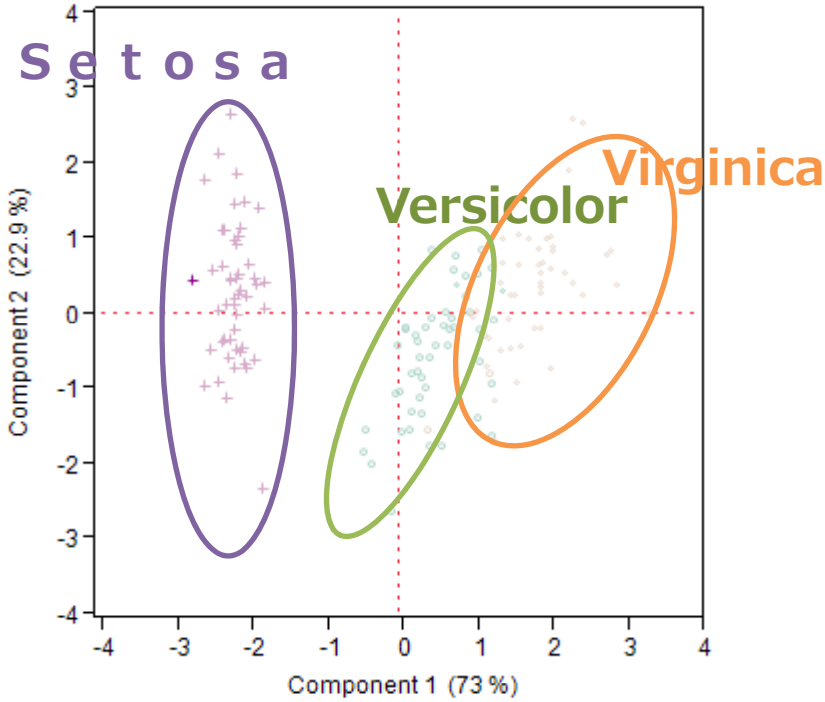
Cluster	SepalLength	SepalWidth	PetalLength	PetalWidth
1	59.1346154	27.3846154	42.9615385	13.25
2	50.06	34.28	14.62	2.46
3	66.3958333	30.1666667	55.6666667	20.5625

Biplot



Eigenvalues

2.9184978 0.9140305 0.1467569 0.0207148



営利目的での利用はご遠慮ください

K-Means 法の利点と欠点

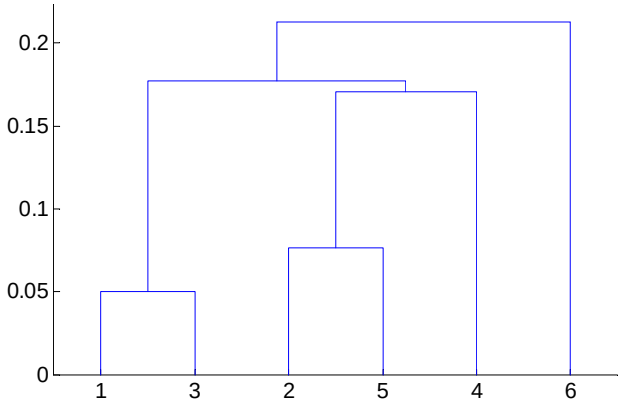
- 利点: 効率がよい　（大きなデータでも速く計算できる）

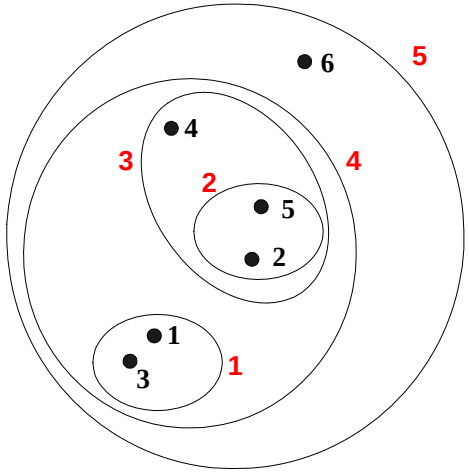
- 欠点
  - 先にクラスター数を決定しないと分析できない
  - ノイズやはずれ値をコントロールできない
  - 平均が重心になっていないようなデータには向いてない

営利目的での利用はご遠慮ください

階層的方法

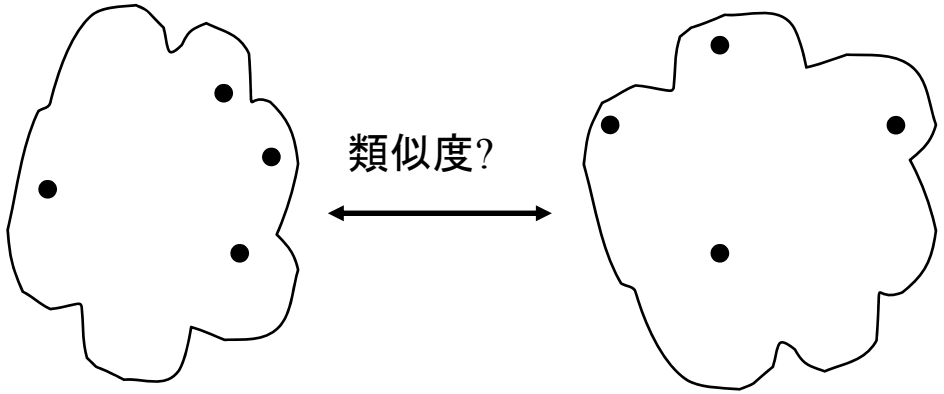
- 階層的なクラスターを作っていく
- 結果はデンドグラムとして視覚的に理解できる





営利目的での利用はご遠慮ください

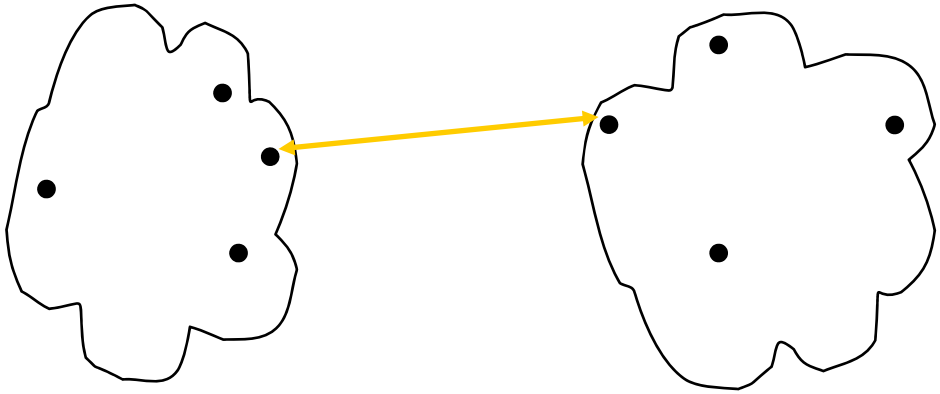
クラスター間距離の決め方



- 最短距離(single linkage)
- 最長距離 (complete linkage)
- 群平均 (average linkage)
- セントロイド間距離
- そのほか、関数を使う方法など
  - ウォード法など

営利目的での利用はご遠慮ください

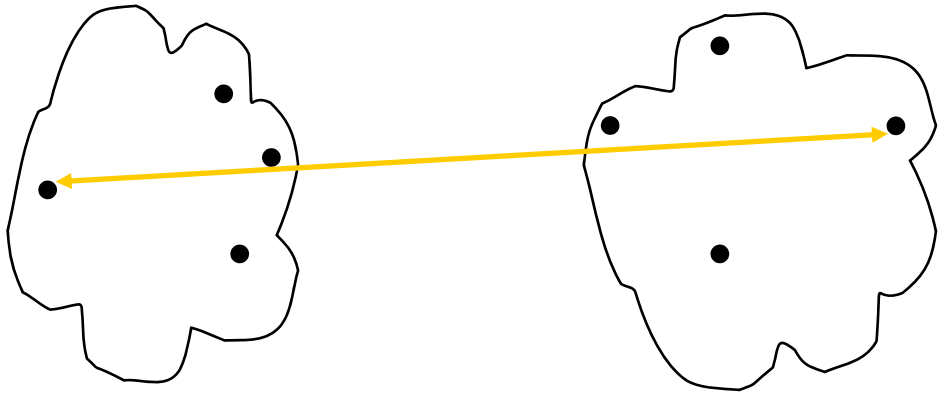
クラスター間距離の決め方



- 最短距離(single linkage)
- 最長距離 (complete linkage)
- 群平均 (average linkage)
- セントロイド間距離
- そのほか、関数を使う方法など
  - ウォード法など

営利目的での利用はご遠慮ください

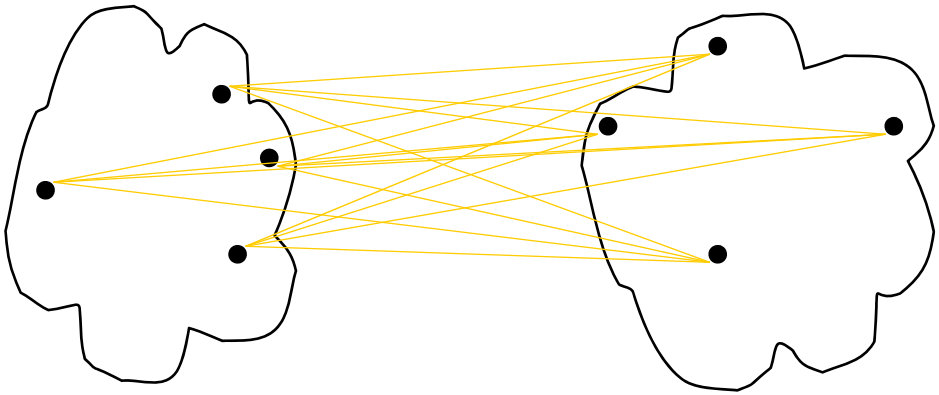
クラスター間距離の決め方



- 最短距離(single linkage)
- 最長距離 (complete linkage)
- 群平均 (average linkage)
- セントロイド間距離
- そのほか、関数を使う方法など
  - ウォード法など

営利目的での利用はご遠慮ください

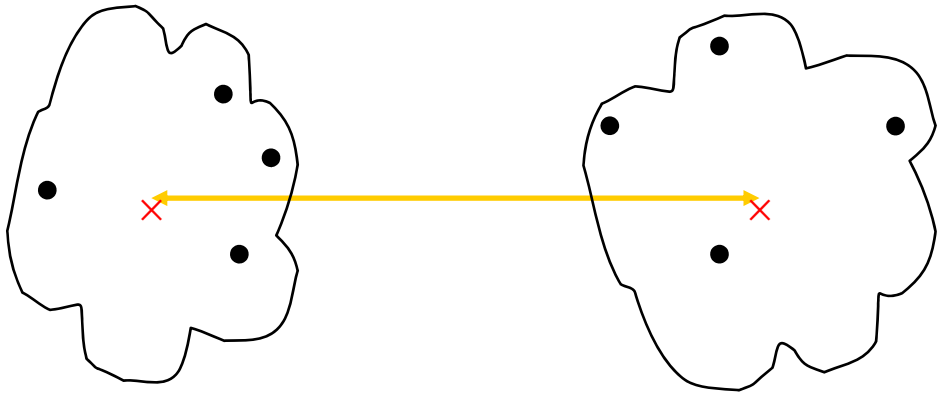
クラスター間距離の決め方



- 最短距離(single linkage)
- 最長距離 (complete linkage)
- 群平均 (average linkage)
- セントロイド間距離
- そのほか、関数を使う方法など
  - ウォード法など

営利目的での利用はご遠慮ください

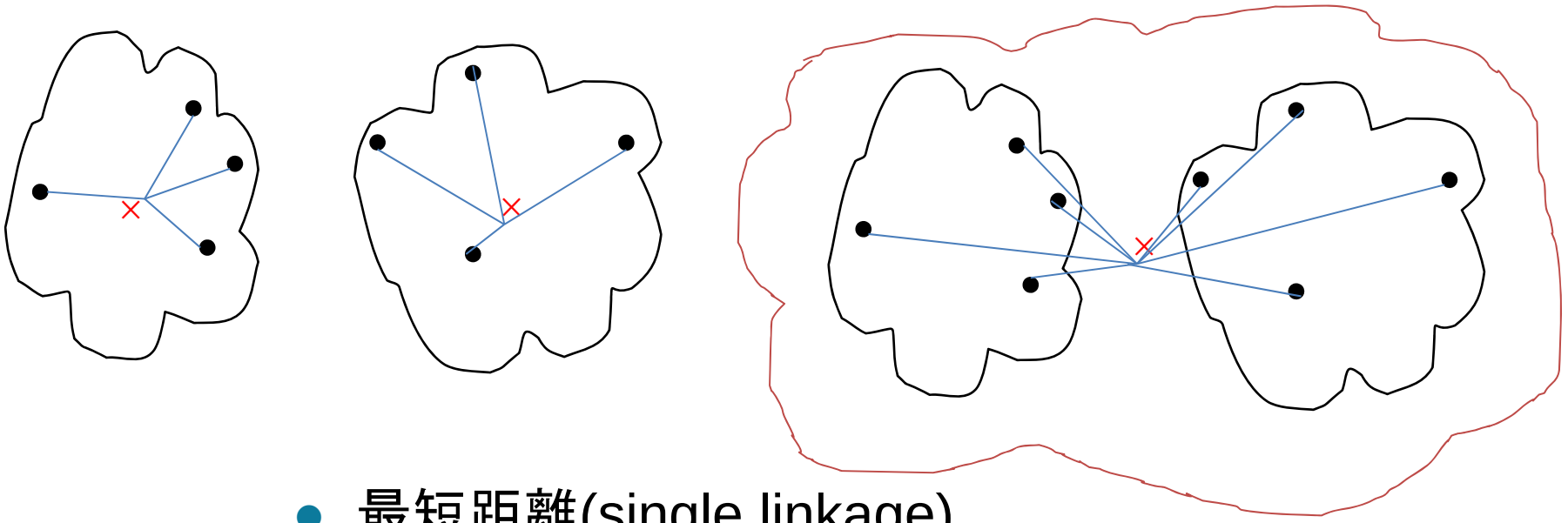
クラスター間距離の決め方



- 最短距離(single linkage)
- 最長距離 (complete linkage)
- 群平均 (average linkage)
- セントロイド間距離
- そのほか、関数を使う方法など
  - ウォード法など

営利目的での利用はご遠慮ください

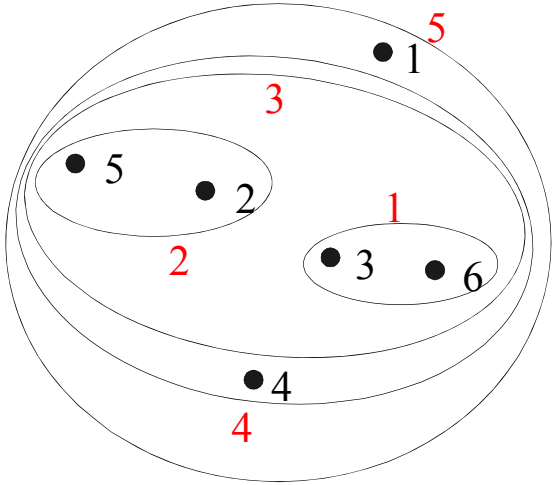
クラスター間距離の決め方



- 最短距離(single linkage)
- 最長距離 (complete linkage)
- 群平均 (average linkage)
- セントロイド間距離
- そのほか、関数を使う方法など
  - ウォード法など

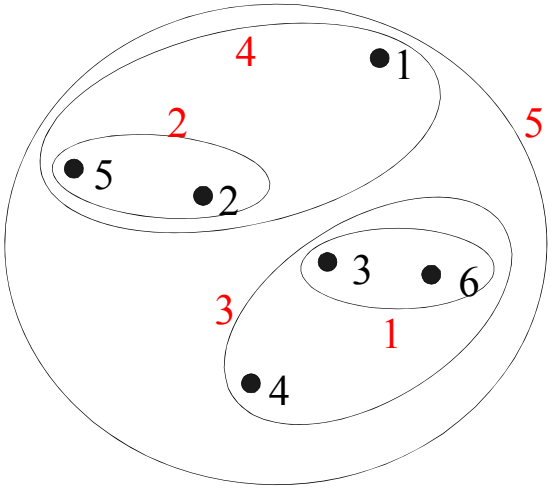
営利目的での利用はご遠慮ください

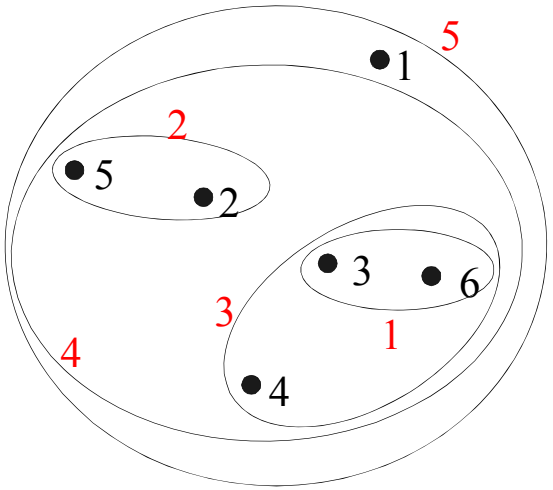
クラスター間距離の設定による比較



最短

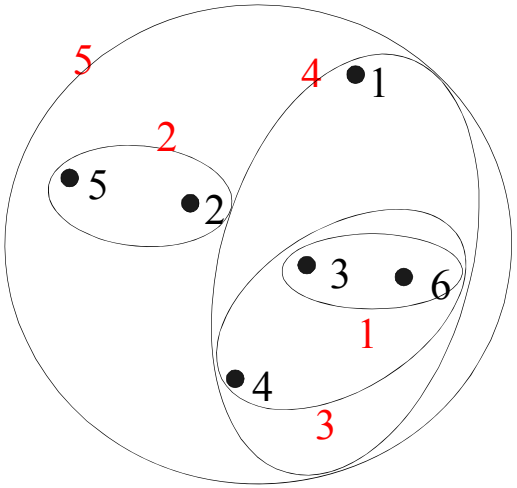
最長





Group Average

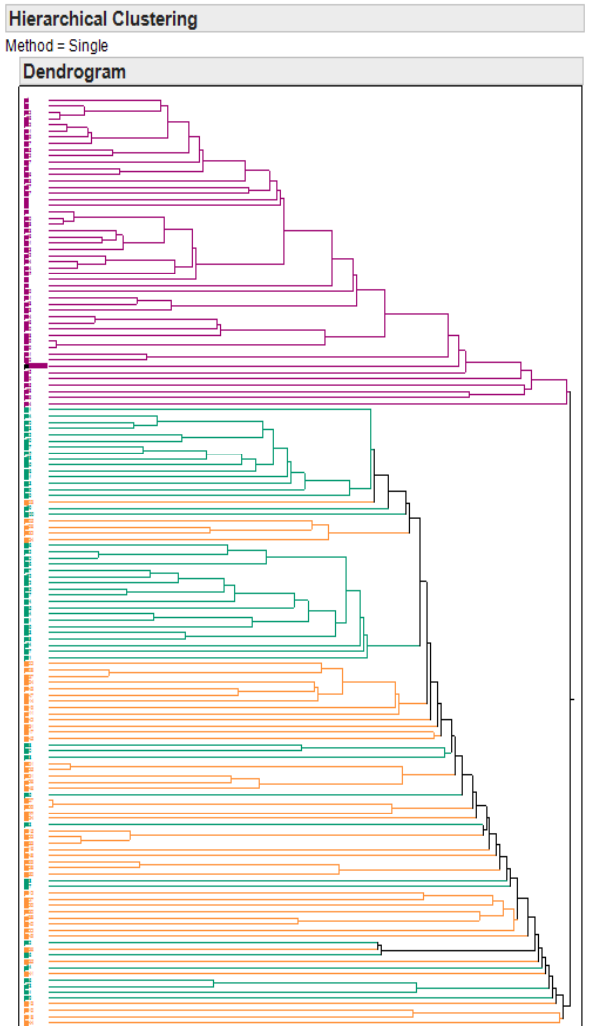
Ward法



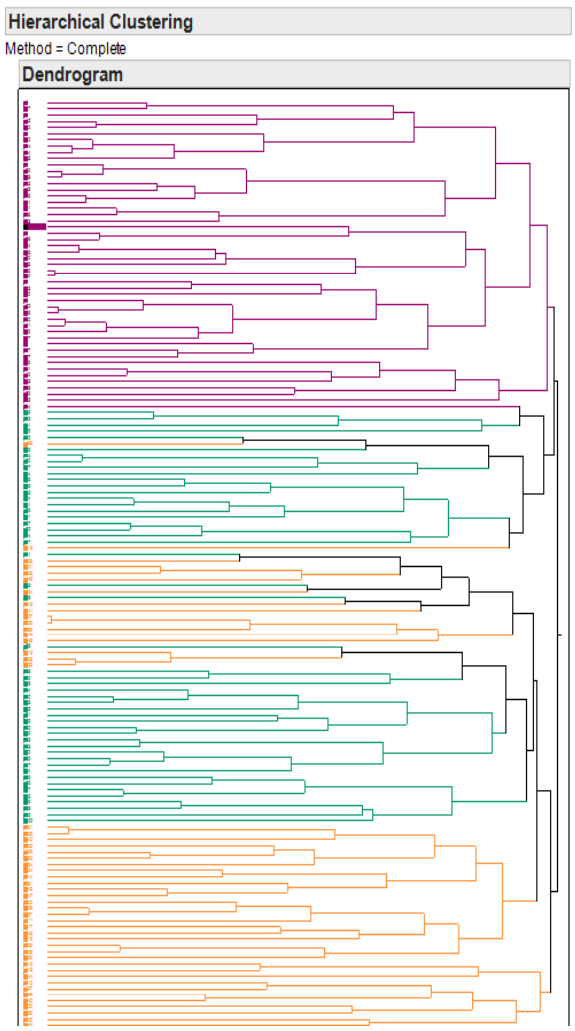
営利目的での利用はご遠慮ください

クラスター間距離の設定による結果の比較
あやめのデータへの適用

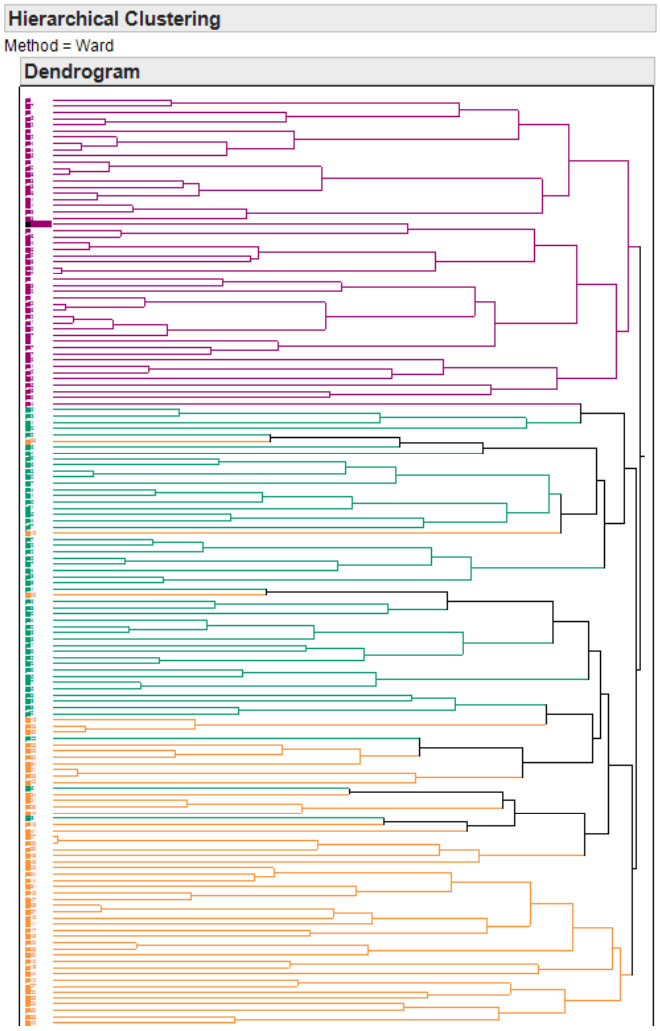
最短距離



最長距離



ウォード法



営利目的での利用はご遠慮ください

階層的クラスタリング手法の利点と欠点

- 利点
  - K-means法のように、クラスター構造が計算の過程で変わることはない。
  - 視覚的にわかりやすい
  - 特に目的関数を設定する必要がない

- 欠点:
  - はずれ値やノイズに影響を受けやすい
  - 大きさや形の異なるクラスターを作ることが難しい
  - 大きなクラスターを見逃しやすい

営利目的での利用はご遠慮ください

クラスタリング結果の妥当性

- モデルや目的関数がない手法は特に結果の妥当性を検討することが難しい
- 探索目的なので、見る人が見ればそこにクラスターがあればよい？
- しかし・・・
  - ノイズの影響は受けてない？
  - アルゴリズムによって結果が変わるんじゃない？
  - クラスター間の比較や意味づけは正しい？

営利目的での利用はご遠慮ください

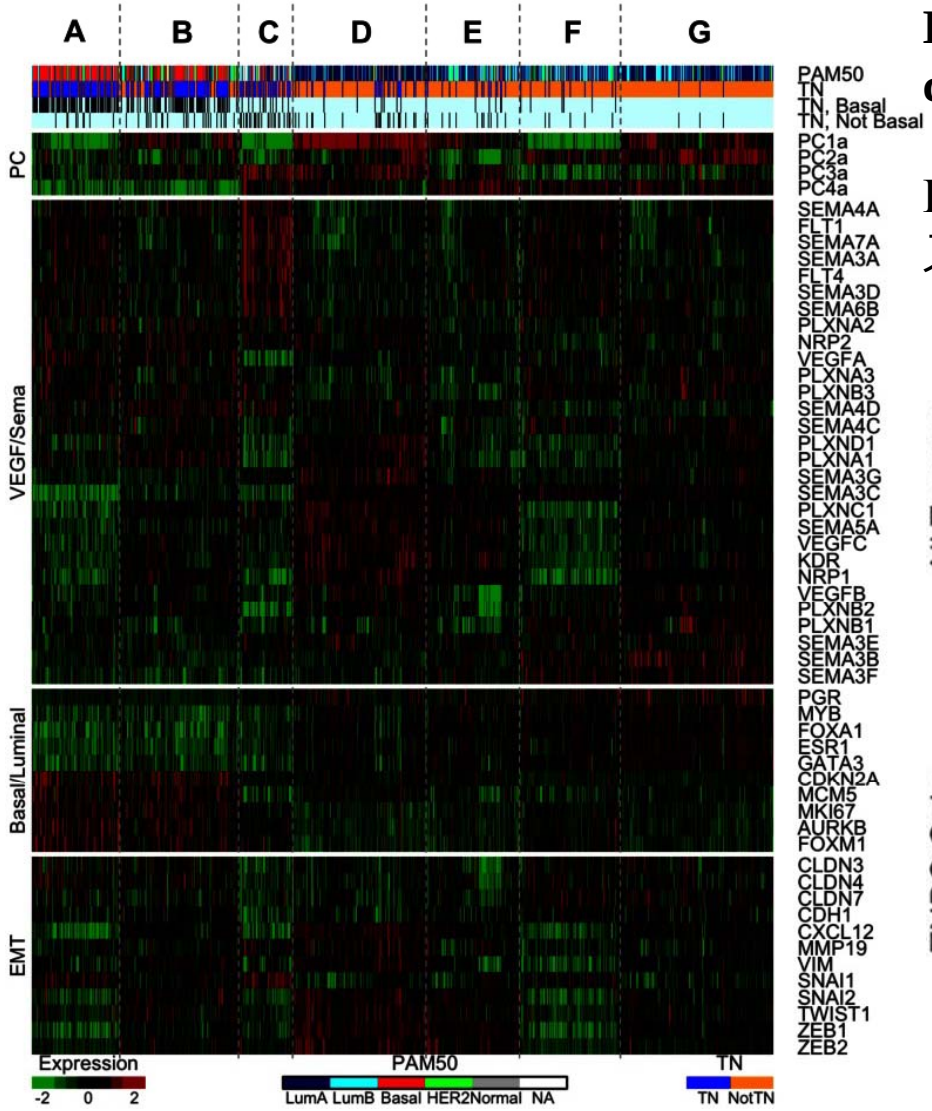
クラスタリング結果の妥当性の検討

1. 先行研究や知見との比較
2. ＰＣＡやほかの探索的手法の結果との比較
3. LOOCVなどで感度分析
4. Meta-classification (Alexe, 2005)
5. Bootstrap法によるクラスターのp値の計算
   （Shimodaira, 2002)

営利目的での利用はご遠慮ください

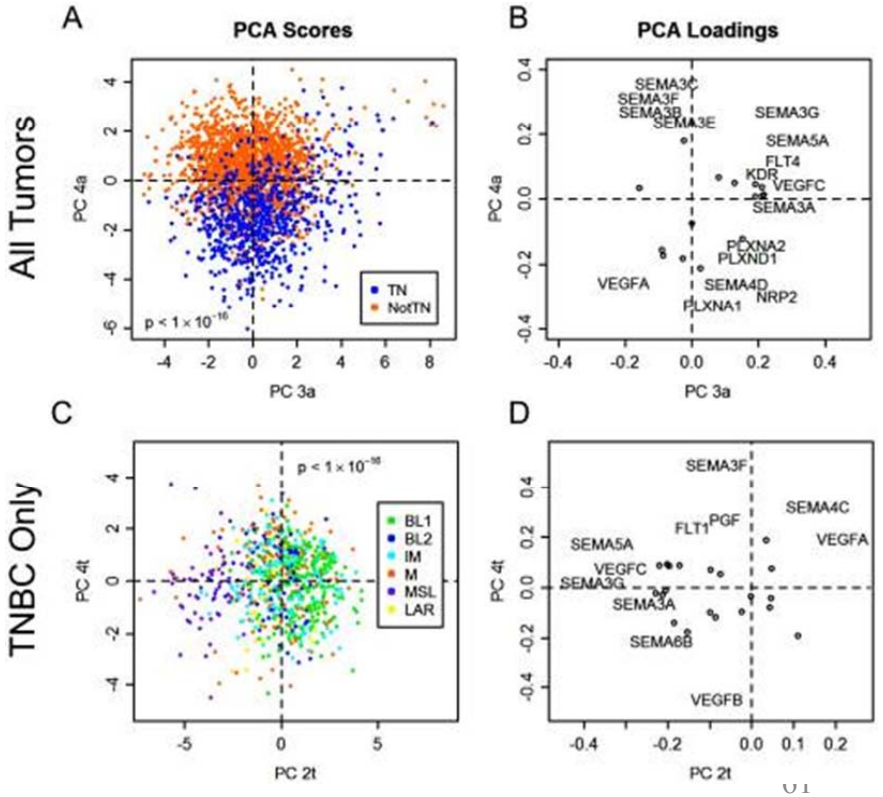
ＴＮＢＣデータへの適用

Bender and Mac Gabhann.Plos One 2013



Heatmap of the 7 VEGF/Sema-based tumor clusters.

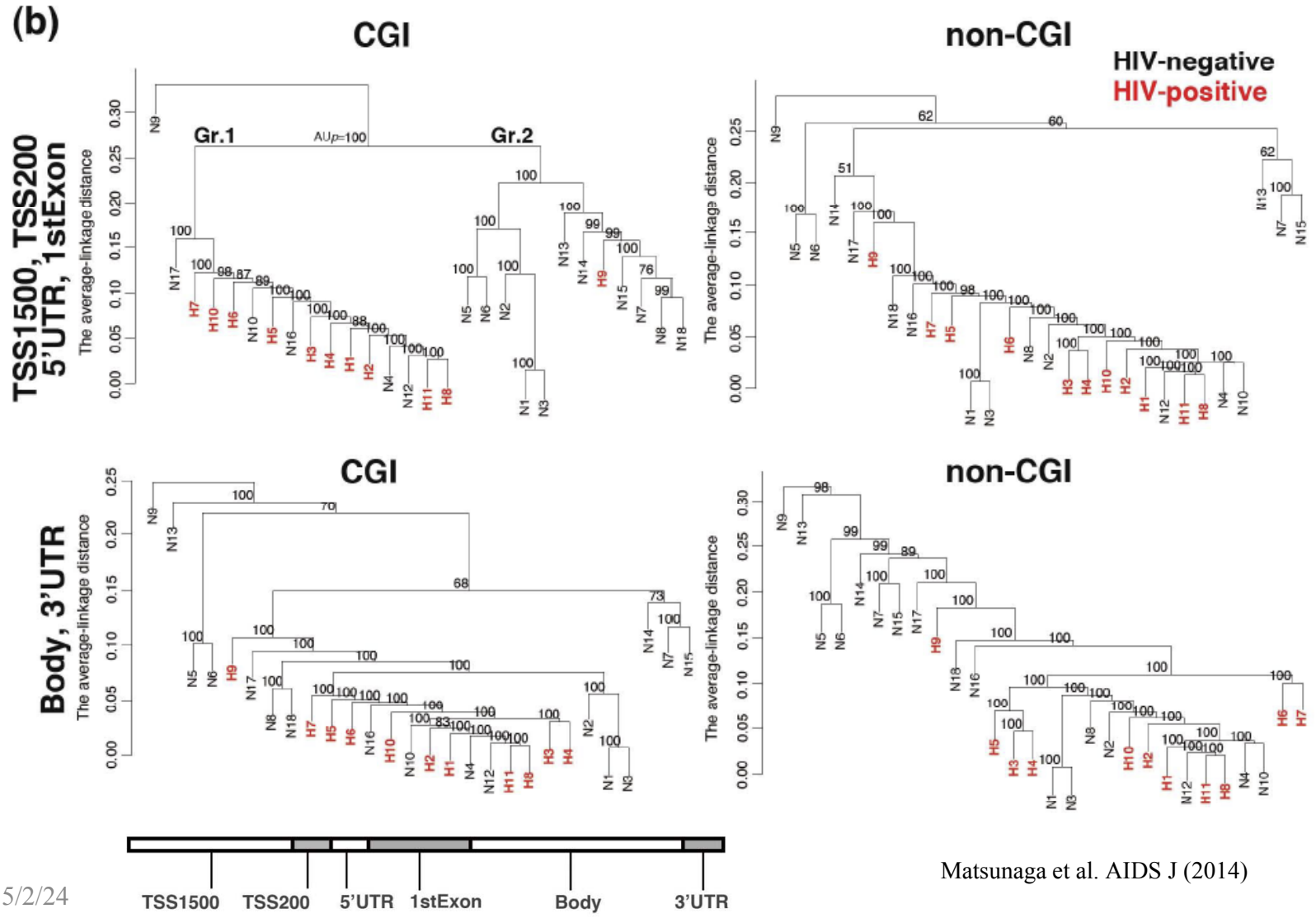
K-means法により、サンプルを７つのクラスターに分類した結果



営利目的での利用はご遠慮ください

エピジェネティクスデータへの適用１：

HIV陽性lymphomaのエピジェネティクス的特徴を探る

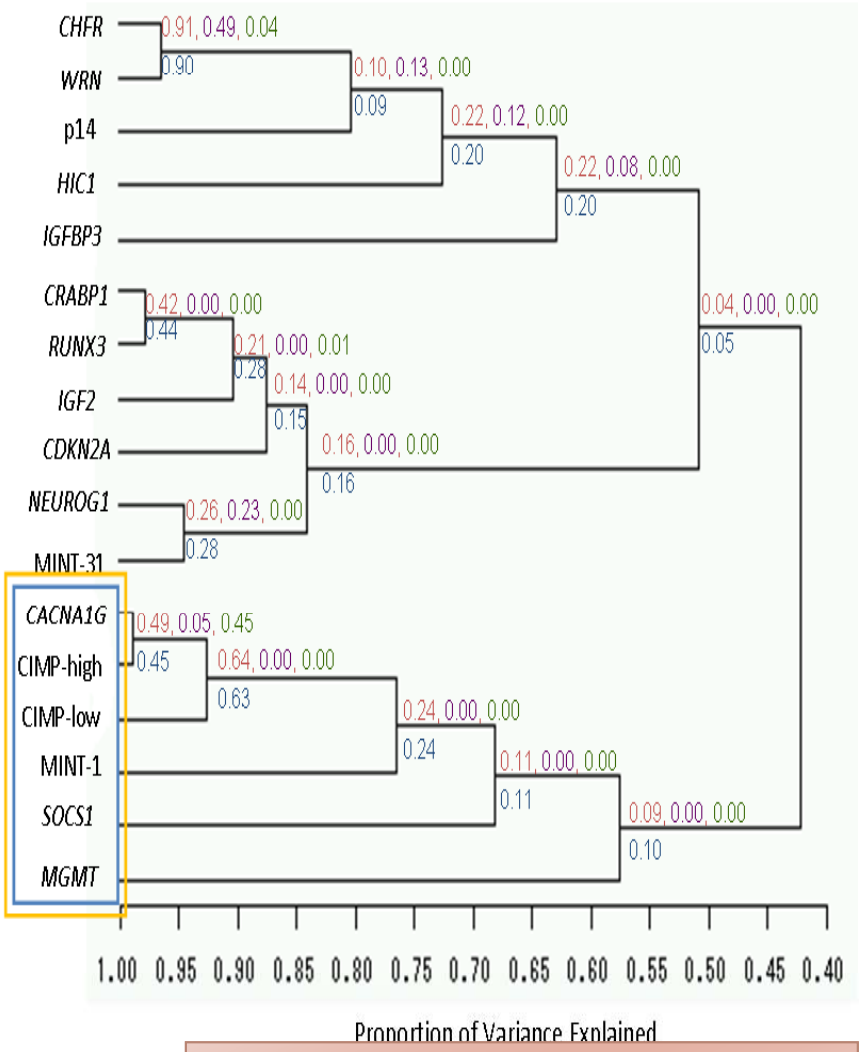


Matsunaga et al. AIDS J (2014)

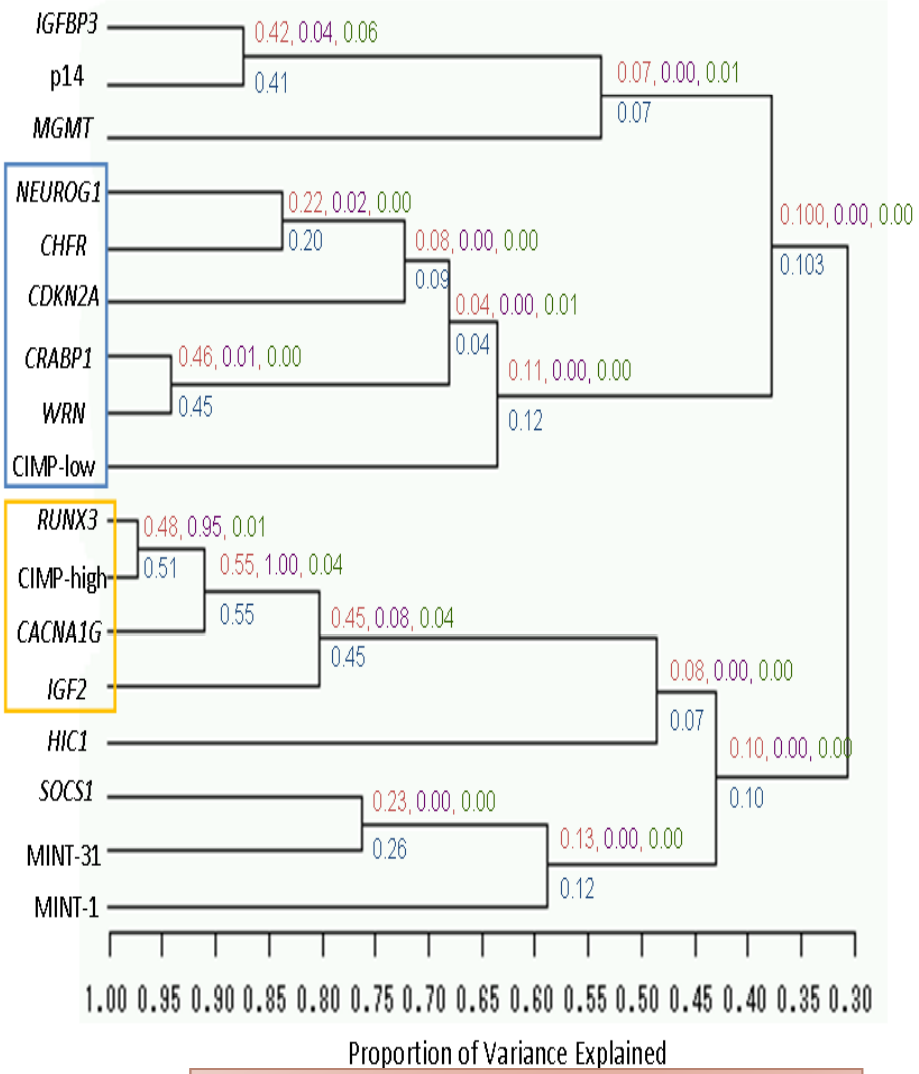
営利目的での利用はご遠慮ください

エピジェネティクスマーカーへの適用例２：
大腸がんのサブタイプをＣＩＭＰ（CpG island methylator phenotype）マーカーで探る

Tanaka et al. AJP 2010.



BRAF mutant group



KRAS mutant group

営利目的での利用はご遠慮ください

営利目的での利用はご遠慮ください

まとめ

- 主成分分析で、

データの次元を縮小≒新しい少ない数の変数でデータの相関構造を分析する

- クラスター分析で

人（オブザーベーション）や変数の分類を行うことで、データの相関構造を分析する

営利目的での利用はご遠慮ください

参考文献

- Fisher, R. A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7, 179–188.
- Bender RJ1, Mac Gabhann F. (2013) Expression of VEGF and semaphorin genes define subgroups of triple negative breast cancer. PLoS One.8(5), e61788.
- Ester M., Kriegel H.-P., Sander J., Xu X. (1996) A density-based algorithm for discovering clusters in large spatial databases. KDD.
- M. Ankerst, M. Breunig, H.-P. Kriegel, and J. Sander. (1999) Optics: Ordering points to identify the clustering structure, SIGMOD.
- A. Hinneburg, D.l A. Keim. (1998) An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD.
- R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan. (1998) Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD.
- Alexe G, Bhanot G, Venkataraghavan B, Ramaswamy R, Lepre J, Levine AJ, Stolovitzky G. Robust diagnosis of non-Hodgkin lymphoma phenotypes validated on gene expression data from different laboratories.(2005) Proc IEEE Comput Syst Bioinform Conf. 322-5.

営利目的での利用はご遠慮ください