実験系生物学者のための
数理・統計・計算生物学入門コース

バイオマーカー・統計・多変量解析
網羅的解析・遺伝子発現解析
（1／2）

大羽成征（おおばしげゆき）
情報学研究科 講師

講義資料 https://github.com/shigeyukioba/biostat

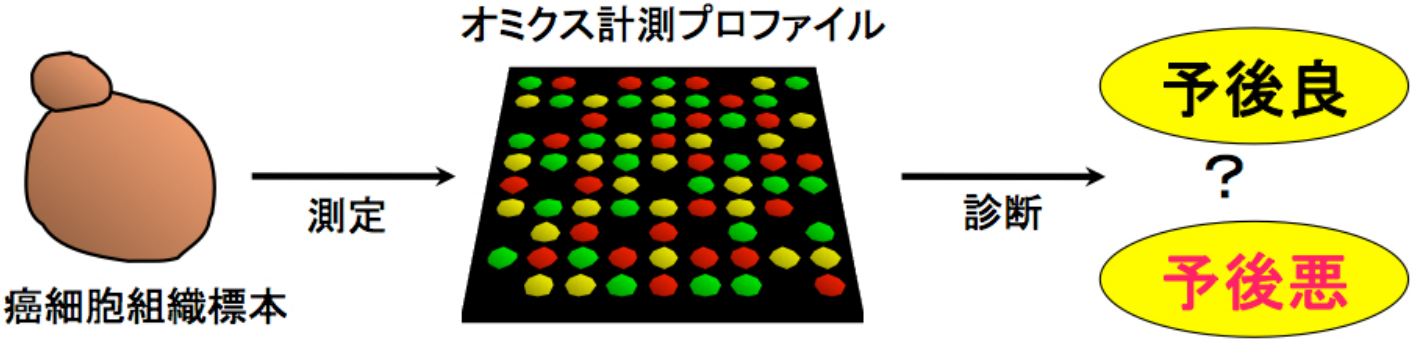
連絡先 Email: oba@i.kyoto-u.ac.jp, Twitter: @shigepong

京都大学
KYOTO UNIVERSITY

バイオマーカーとは

- 生体の状態を知るための計測情報
- 医師による診断の手がかり
- 例:
  - 血糖値・コレステロール値
  - 遺伝子異常
  - 遺伝子発現量

インフォマティクスによる診断マーカー探索



診断用バイオマーカーの性能

- コストが低いこと
  - 金額・時間・労力・侵襲性
- 正解率が高いこと
  - 2種類の正解と2種類の誤り

２種類の誤りと２種類の正解率

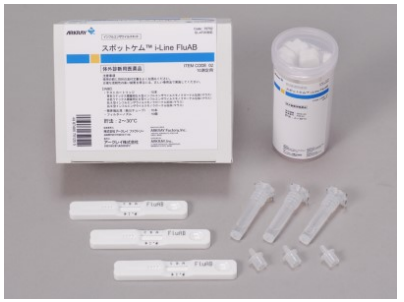
		真実：正例（異常あり）	真実：負例（異常なし）
判定	陽性	TP 正解	FP 偽陽性
	陰性	FN 偽陰性	TN 正解

		真実：正例	真実：負例
判定	陽性	TP/(TP+FN) sensitivity	
	陰性		TN/(FP+TN) specificity

参考用

新型インフルエンザ (A/H1N1) の診断

- 迅速診断キットによる診断
  - 鼻やのどの粘液に試薬を加えて約10分で判定
  - A型（＋／－）B型（＋／－）が分かる
    - 特異度(specificity) 90%程度
    - 感度(sensitivity) 90%程度
- ウィルス学的診断
  - RT-PCR法
    - （温める→待つ→冷やす→待つ）×40回
  - A/H1N1 であるか否かが確実に分かる



アークレイ社HPより

出典：　国立感染症研究所　感染症情報センター
国内医療機関における新型インフルエンザ(A/H1N1)診断の流れ ver.1.6 2009/5/6
GSK社リレンザのページ　http://relenza.jp/clinic/kit.html

余談クイズ

- （仮定1）現在インフルエンザに罹っている患者は、1000人中で1人だとする
- （仮定2）迅速診断キットによるインフルエンザ診断の精度は
  感度 sensitivity（正例中の陽性率）90%
  特異度 specificity（負例中の陰性率）90%　だとする
- Q：あなたを診断した結果が陽性だったとして、あなたがインフルエンザに罹っている確率は？

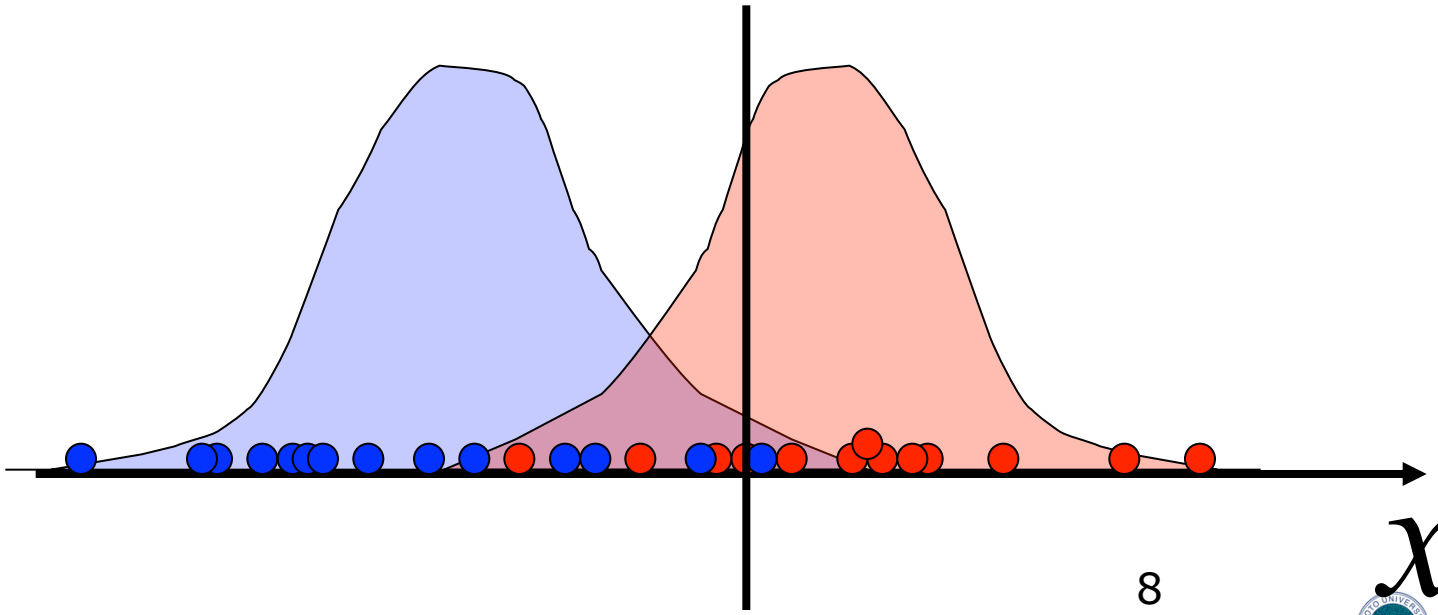
  A:　90%　　B：約10%　　C:約1%　　D:それ以下

- Q:あなたを診断した結果が陰性だった場合は？

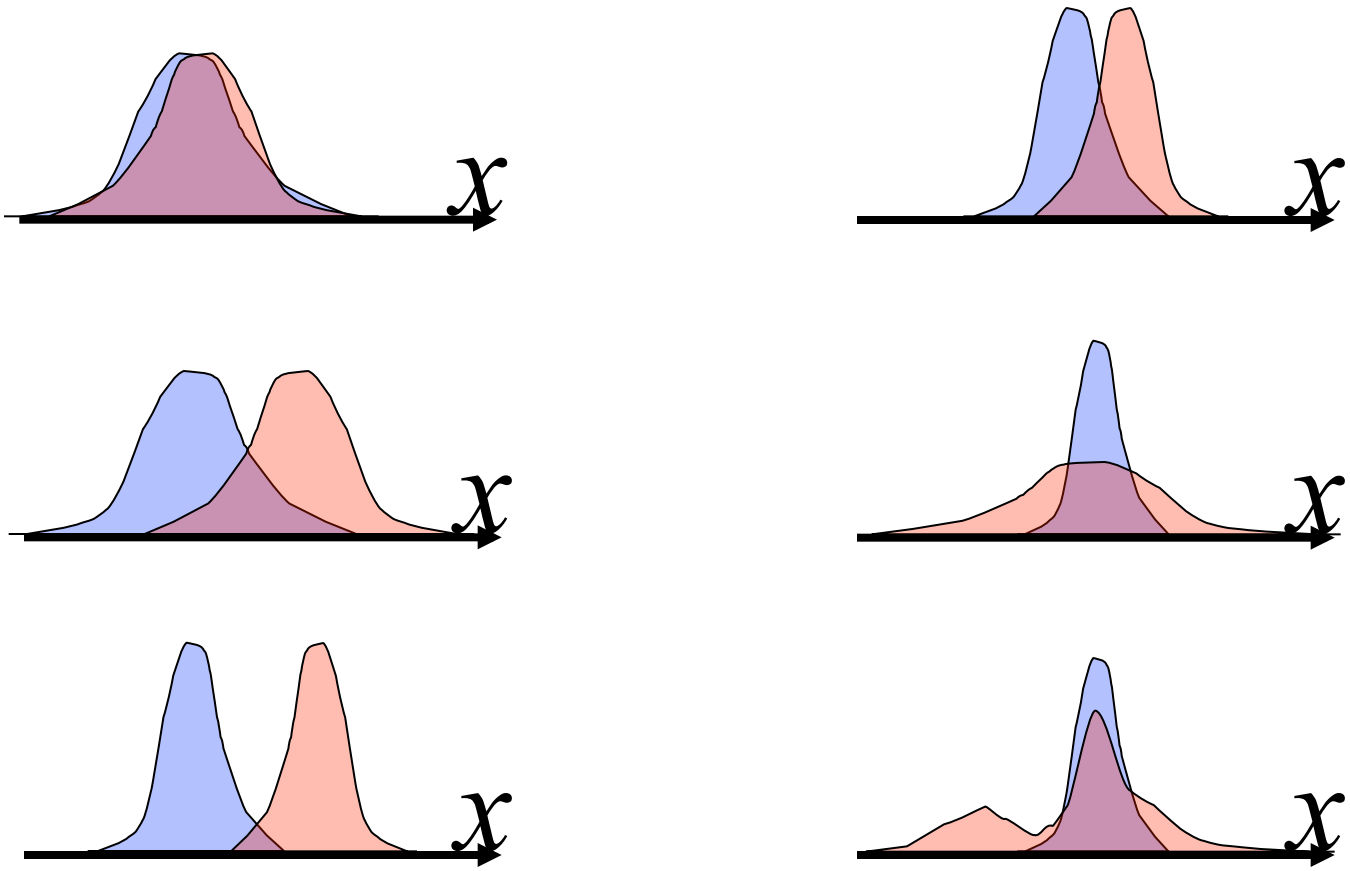
  A:　90%　　B：約10%　　C:約1%　　D:それ以下

連続値をとるマーカー

- 例：コレステロール値、γGDP値
- 大きい値（小さい値）であるほど危険
- しきい値で二値化して診断



連続値マーカーの性能を診断する



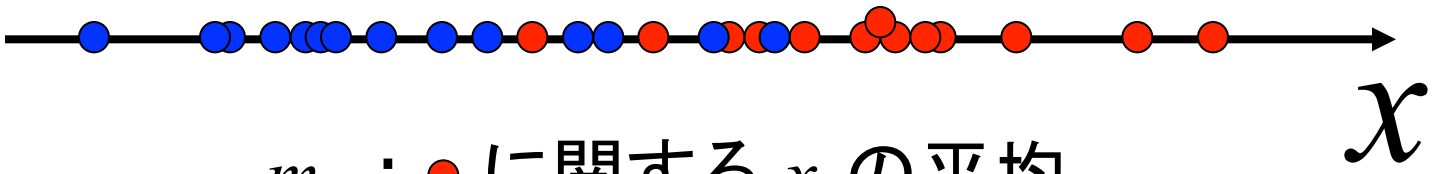
どのマーカーが最も良いマーカか？

t 統計量

有限のデータからマーカーの良さを見積もる方法

- データに基づく t 統計量を求める

$$t = \frac{m_1 - m_2}{s}$$



m_1 ：● に関する x の平均

m_2 ：● に関する x の平均

s ： 群内標準偏差（の定数倍）

- t 統計量の絶対値が大きいほど、良いマーカー

「統計量 statistic(s)」とは？

- 定義：
  - 標本データ数値から計算によって得られる値

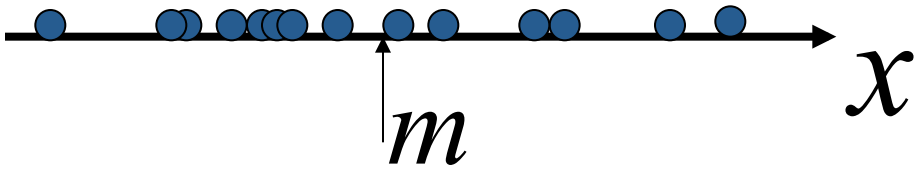
何でも統計量？
平均とか、分散とか、標準偏差とかも？

統計量とは

- 定義
  - a statistic (単数) は標本データ数値から計算によって得られる値
  - 参考:
    statistics (非加算名詞) = 統計学

- 覚えておきたい周辺概念
  - 記述統計学・記述統計量・要約統計量
    descriptive statistics
  - 検定統計量 test statistics
  - 順序統計量 order statistics

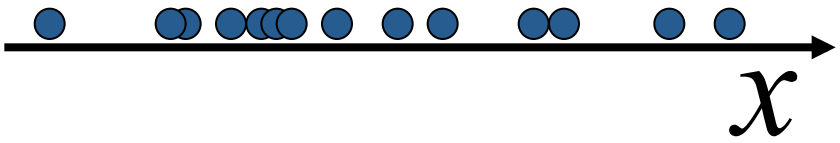
分散 variance



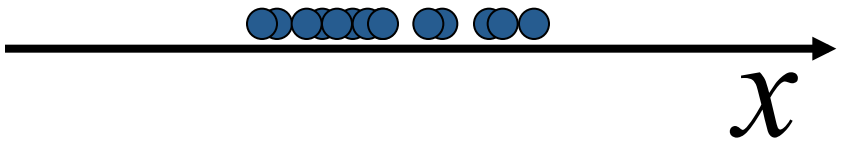
$$Var[x] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2$$

- データ値のばらつきの程度を表す

分散大きい



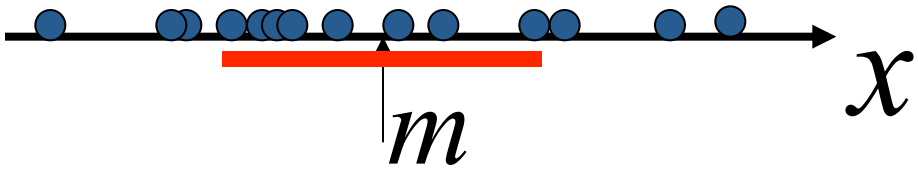
分散小さい



京都大学 KYOTO UNIVERSITY

標準偏差 standard deviation

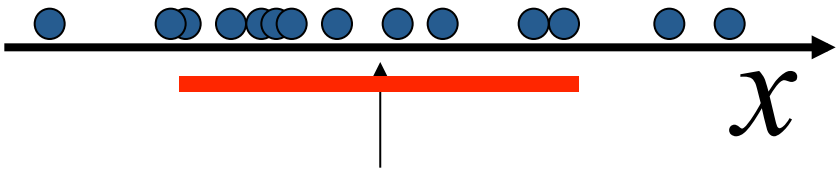
SD.　　std



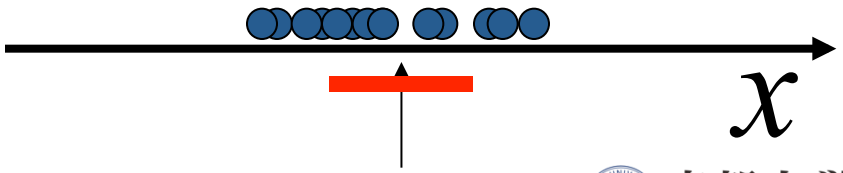
$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2}$$

- データ値のばらつきの程度を表す

標準偏差大きい



標準偏差小さい

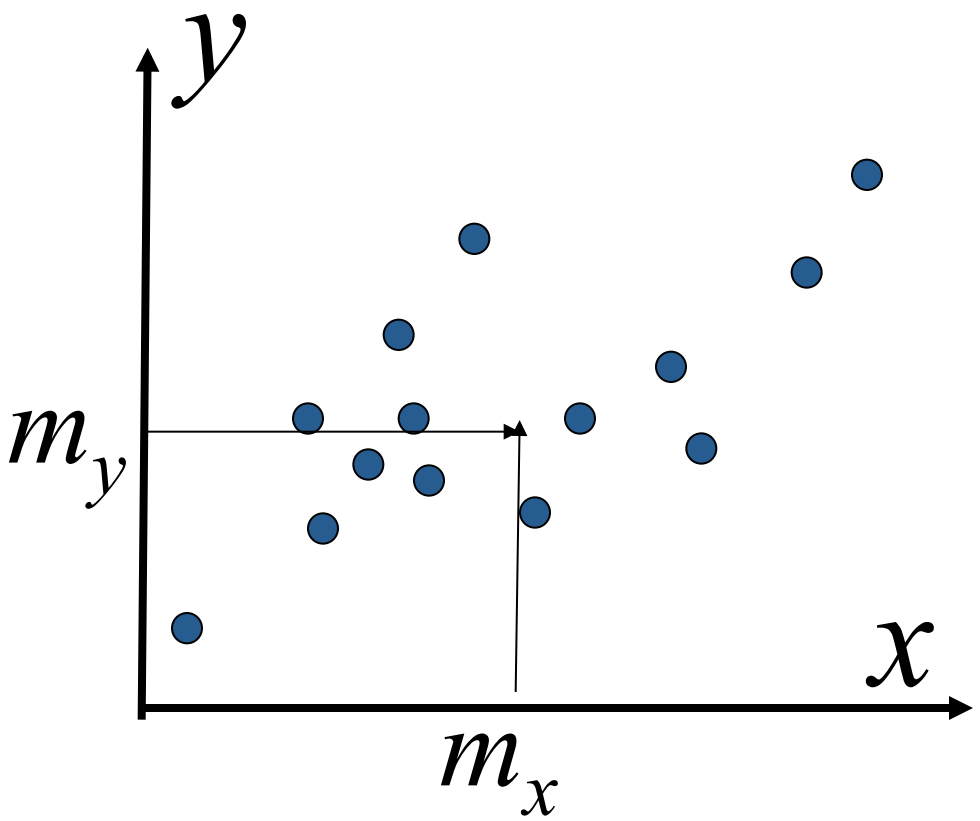


京都大学 KYOTO UNIVERSITY

共分散 covariance

$$\mathrm{cov}[x,y] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)$$

$$= \mathrm{cov}[y,x]$$

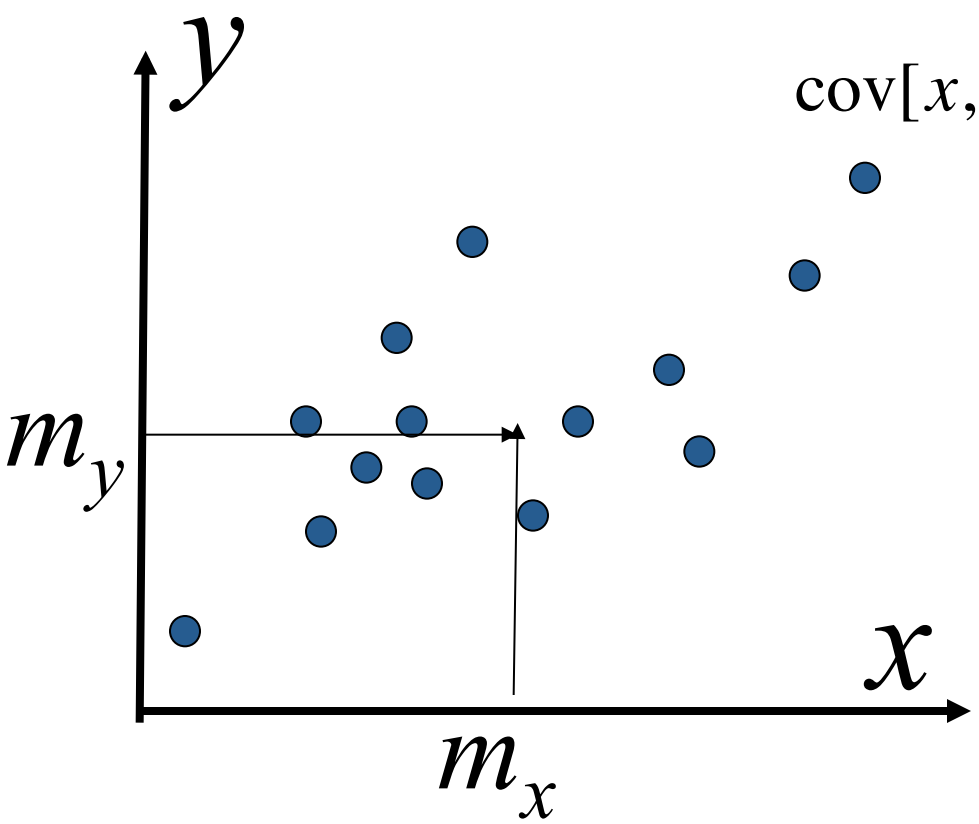


$$Var[x] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_x)^2$$

$$Var[y] = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_y)^2$$

相関係数 correlation coefficient

$$r[x,y] = \frac{\mathrm{cov}[x,y]}{\sqrt{Var[x]Var[y]}}$$



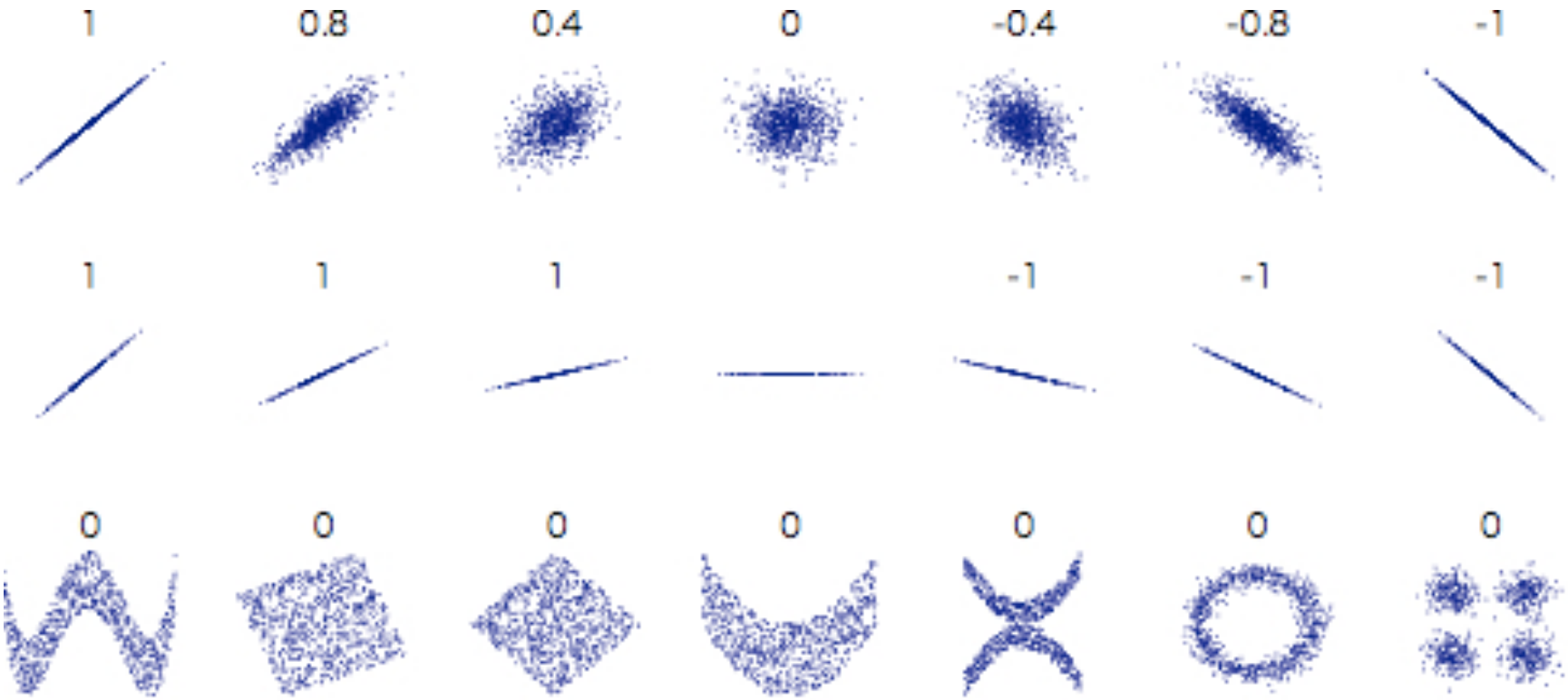
$$\mathrm{cov}[x,y] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_x)(y_i - m_y)$$

$$= \mathrm{cov}[y,x]$$

$$Var[x] = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - m_x)^2$$

$$Var[y] = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - m_y)^2$$

相関係数



Wikipedia 相関係数のページより

京都大学
KYOTO UNIVERSITY

参考用

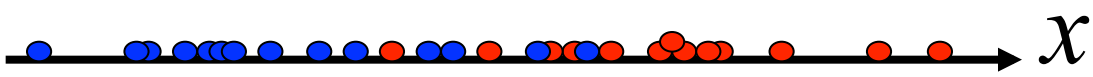
もうすこし正確な定義
等分散仮定のもとでの

t 統計量

$x_{\bullet 1}, \cdots, x_{\bullet N1}$

$x_{\bullet 1}, \cdots, x_{\bullet N2}$

$$t = \frac{m_1 - m_2}{s}$$



m_1 ：●に関する x の平均

m_2 ：●に関する x の平均

s ：群内標準偏差（の定数倍）

$$s = SD_* \sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$SD_1 = \sqrt{\frac{1}{N_1 - 1}\sum_{i=1}^{N}(x_{\bullet i} - m)^2}$$

$$SD_* = \sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2}{N_1 + N_2 - 2}}$$

群内標準偏差

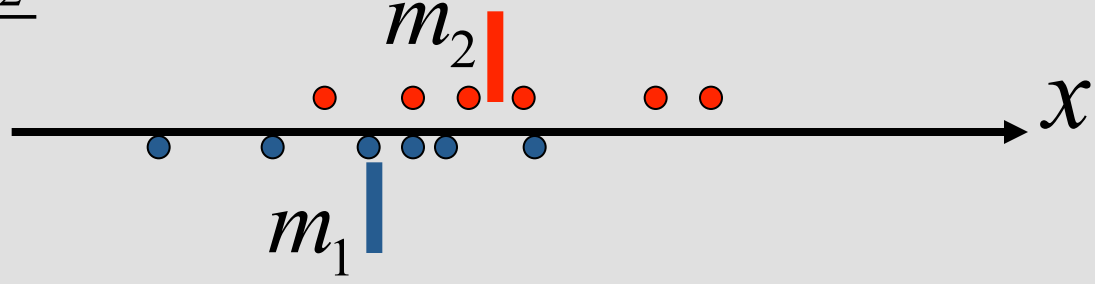
$$SD_2 = \sqrt{\frac{1}{N_2 - 1}\sum_{i=1}^{N}(x_{\bullet i} - m)^2}$$

次回！

統計的仮説検定

サンプル数 N=12 でドヤ顔していいの？

$$t = \frac{m_1 - m_2}{s}$$



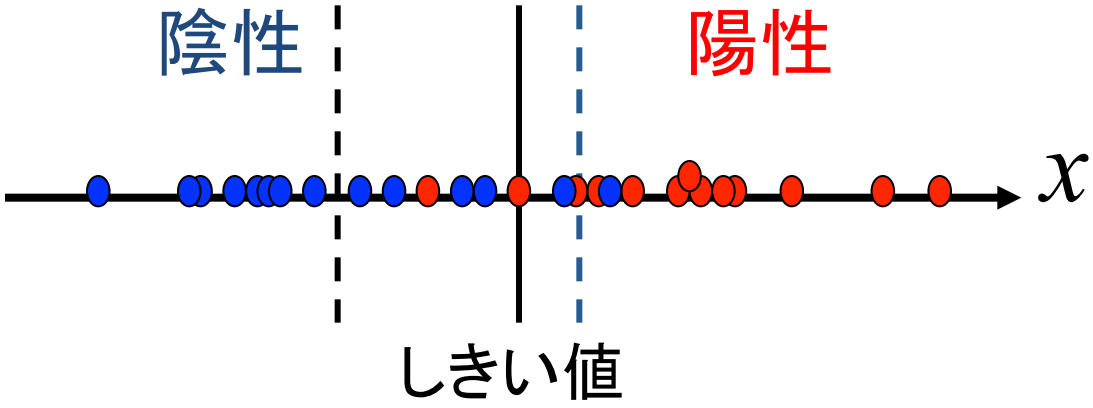
t 統計量の絶対値が大きな値 T=2.0 であったとする。

その良さそうなマーカーは本当に良いマーカーか？

ゼロでない標準偏差　と　有限の標本数　のせいで偶然に良さそうなマーカーに見えただけではないか？

⇒ 偶然ではないことを t 検定で確かめる

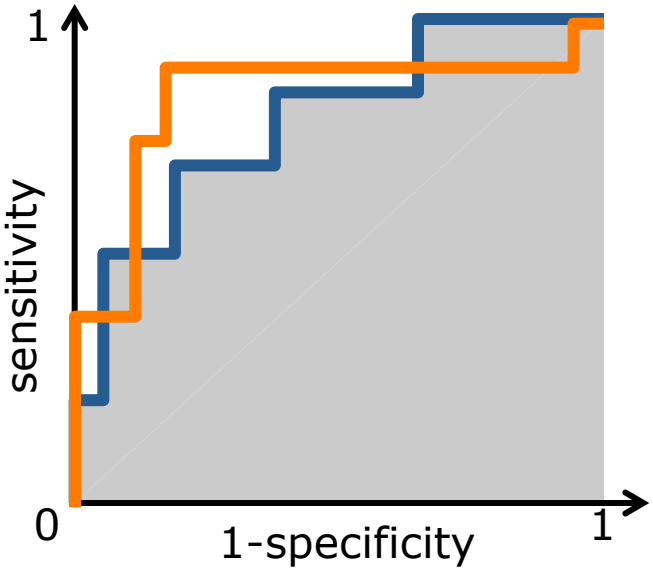
連続値マーカーの二値化とマーカー性能の評価

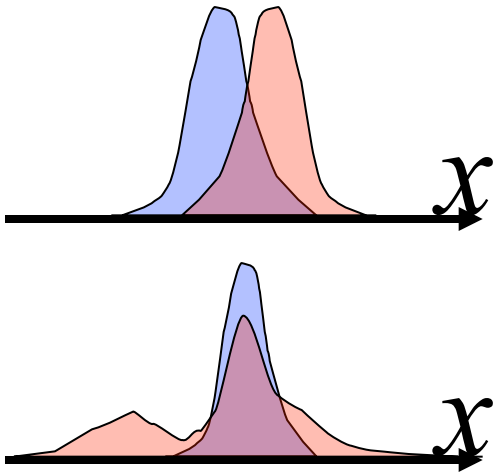


- sensitivity（感度・敏感性：正例中の正解率）
  - tp / (tp+fn)
- specificity（特異性：負例中の正解率）
  - tn / (fp+tn)

ROC 曲線と AUC 基準

- 受信者動作特性曲線
  Receiver Operating Characteristic
- 曲線下面積
  Area Under the Curve
- とくに正例数と負例数がアンバランスである場合に使われる





多変量マーカーを作る！

例）脳腫瘍診断データの解析

Cancer Science

Using gene expression profiling to identify a prognostic molecular spectrum in gliomas

Mitsuaki Shirahata[1,2], Shigeyuki Oba[3], Kyoko Iwao-Koizumi[2], Sakae Saito[2], Noriko Ueno[2], Masashi Oda[1], Nobuo Hashimoto[1], Shin Ishii[3], June A. Takahashi[1], Kikuya Kato[2,*]

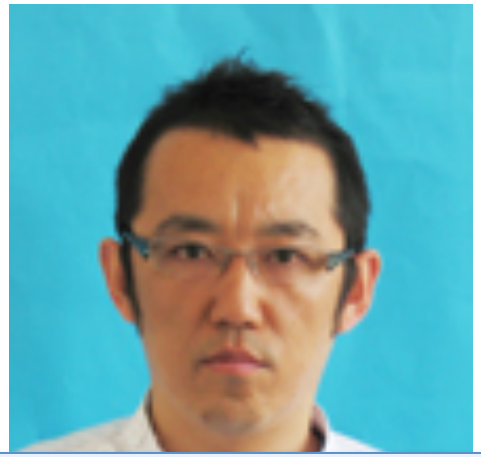
Article first published online: 25 NOV 2008

DOI: 10.1111/j.1349-7006.2008.01002.x

Issue

Cancer Science

Volume 100, Issue 1, pages 165–172, January 2009



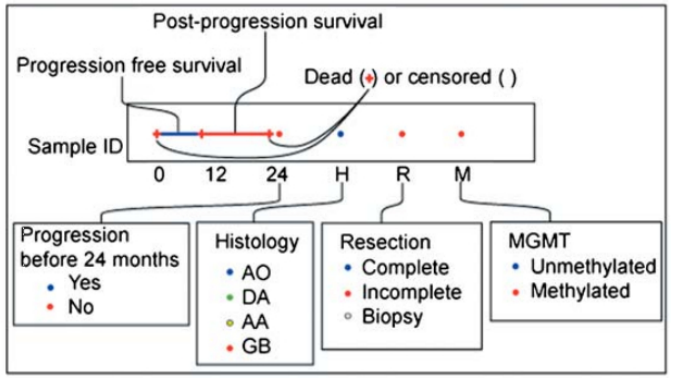
共同研究者 白畑充章先生
静岡大学病院（当時京大医）

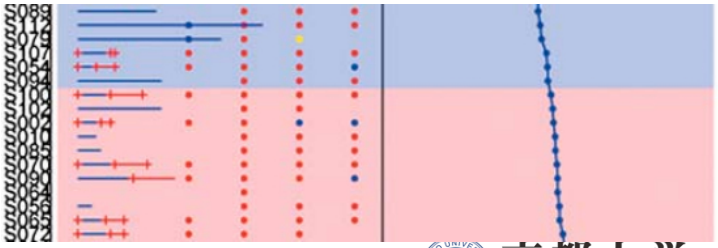
- 遺伝子発現の高次元ベクトルの主成分分析の結果から
  脳腫瘍の既知4分類の違いや、
  予後（死亡率）の違いが読み取れる！

共同研究者 加藤菊也
大阪成人病センター研究所長

脳腫瘍診断データの解析

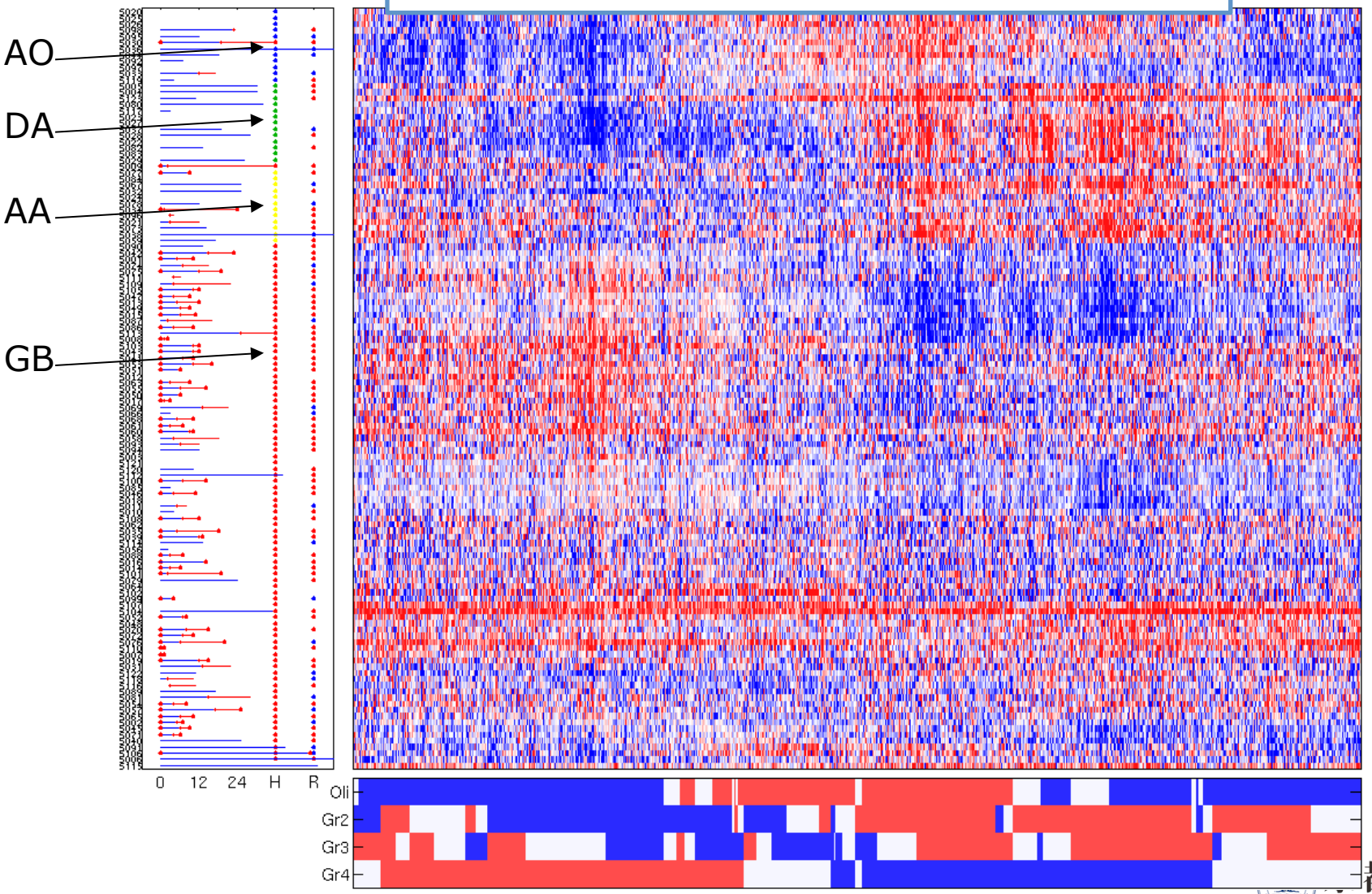
- 脳腫瘍患者の臨床データ
  - 予後
    - 無再発生存期間・再発後生存期間
  - 手術の種類
    - (完全切除、部分切除、細胞診断のみ)
- 脳腫瘍の病理分類 (Histology)
  - GB glioblastoma
  - AA anaplastic astrocytoma
  - AO anaplastic oligodendroglioma
  - DA diffuse astrocytoma



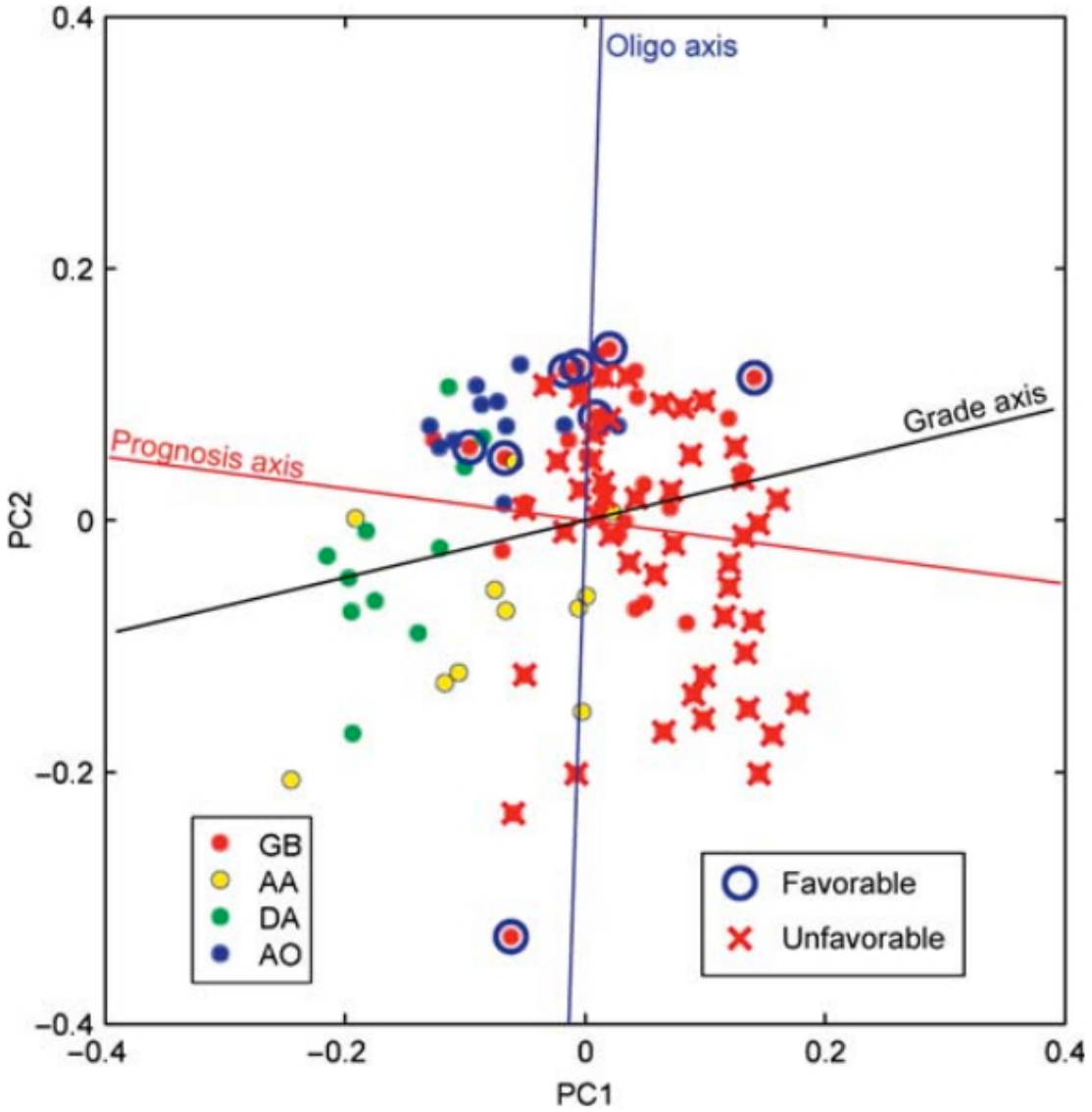


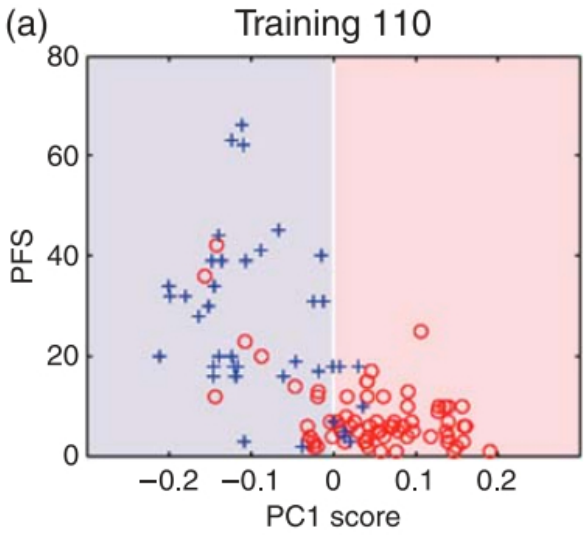
脳腫瘍：遺伝子発現プロファイル

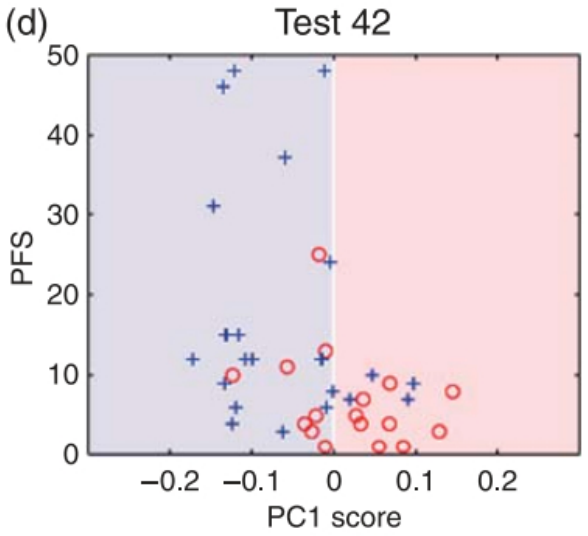
ATAC法（マイクロアレイ代替手法）による
脳腫瘍200症例 × 3000遺伝子の発現プロファイル



脳腫瘍診断データ解析結果



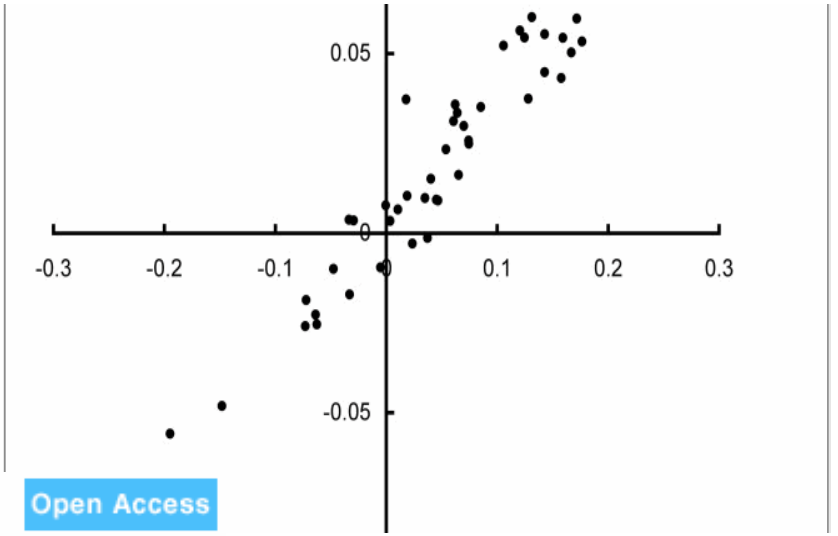




京都大学
KYOTO UNIVERSITY

58遺伝子の発現量に基づく脳腫瘍マーカの再現性がとても良かった

オリジナル：ATAC-PCR法 と、
移行先: RT-PCR法との間の再現性は
相関係数 r = 0.7 程度だったが、
指標同士の再現性は r = 0.94



BMC Medical Genomics

Technical advance

Open Access

Conversion of a molecular classifier obtained by gene expression profiling into a classifier based on real-time PCR: a prognosis predictor for gliomas

Satoru Kawarazaki[1,2], Kazuya Taniguchi[1], Mitsuaki Shirahata[2], Yoji Kukita[1], Manabu Kanemoto[1,2], Nobuhiro Mikuni[2], Nobuo Hashimoto[3], Susumu Miyamoto[2], Jun A Takahashi[4] and Kikuya Kato[1]*

* Corresponding author: Kikuya Kato katou-ki@mc.pref.osaka.jp

▸ Author Affiliations

For all author emails, please log on.

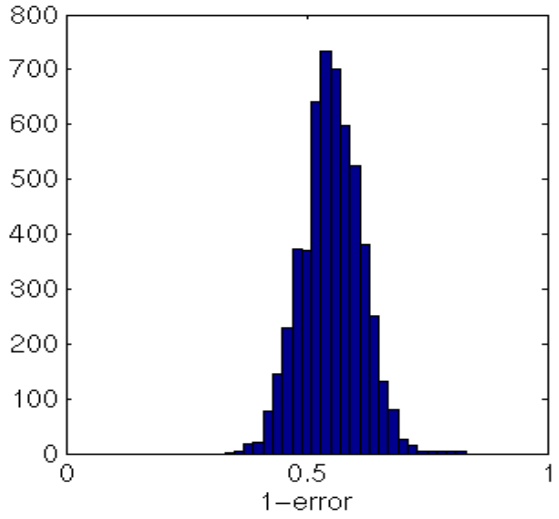
BMC Medical Genomics 2010, **3**:52 doi:10.1186/1755-8794-3-52

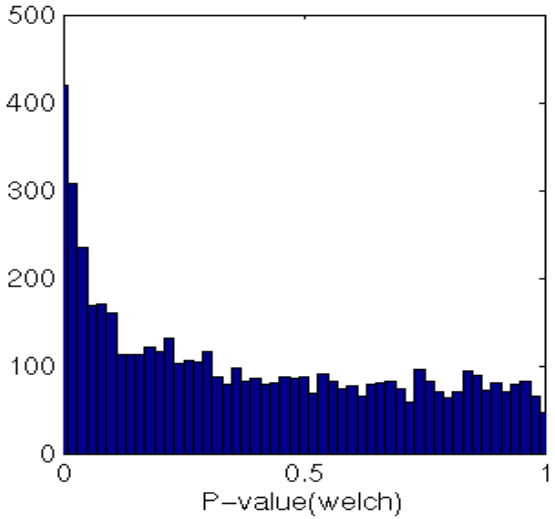
The electronic version of this article is the complete one and can be found online at:
http://www.biomedcentral.com/1755-8794/3/52

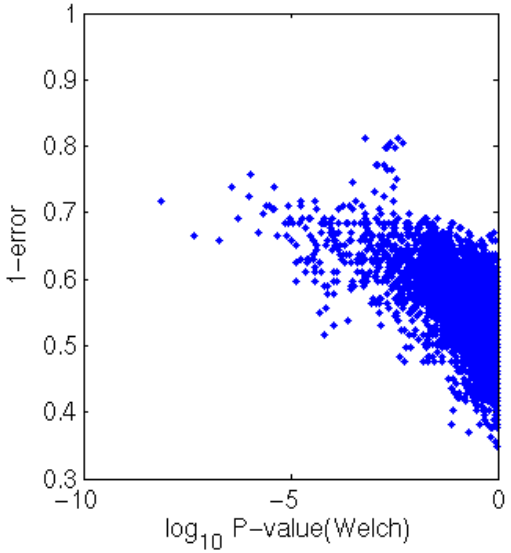
高次元多変量パターンデータの可視化法

- そのいち：ヒストグラムと散布図
- >>>
- そのに：色つき行列
- >>> (越えられない壁) >>>
- うまい可視化法
  - MDS, PCA, ほか…

心得のそのいち：なにはともあれ
ヒストグラムと散布図

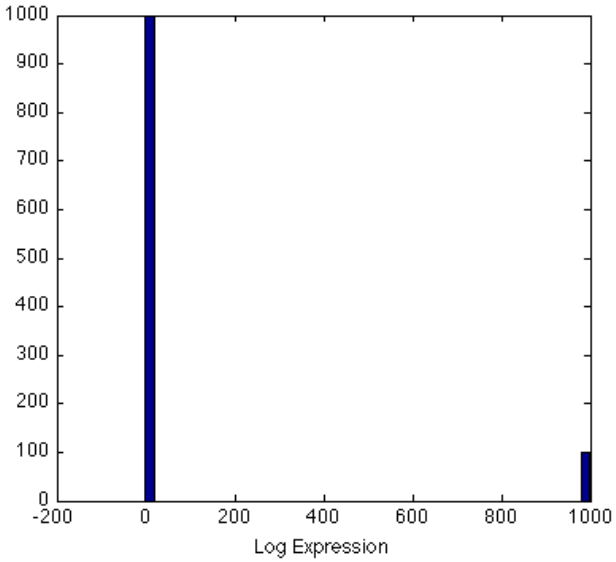


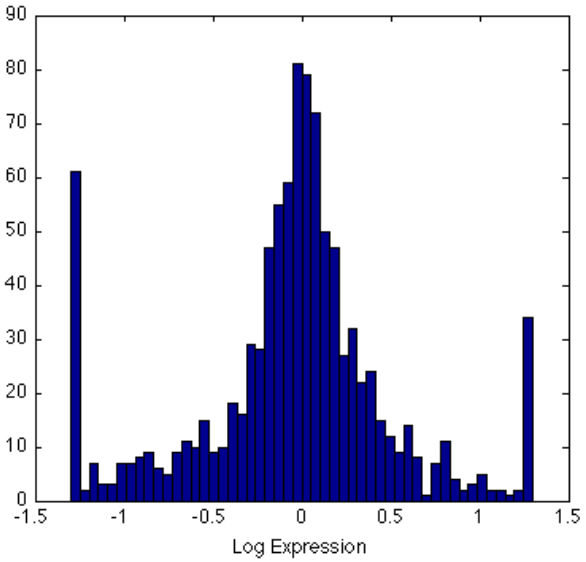




京都大学 KYOTO UNIVERSITY

よくあるタイプの異常値
～ヒストグラム見れば分かる～





外れ値(outlier)： 外れ値を除いた場合の標準偏差の3~5倍以上外側

欠測値(missing value)： (例)欠測扱いとした値に 999 が入っている

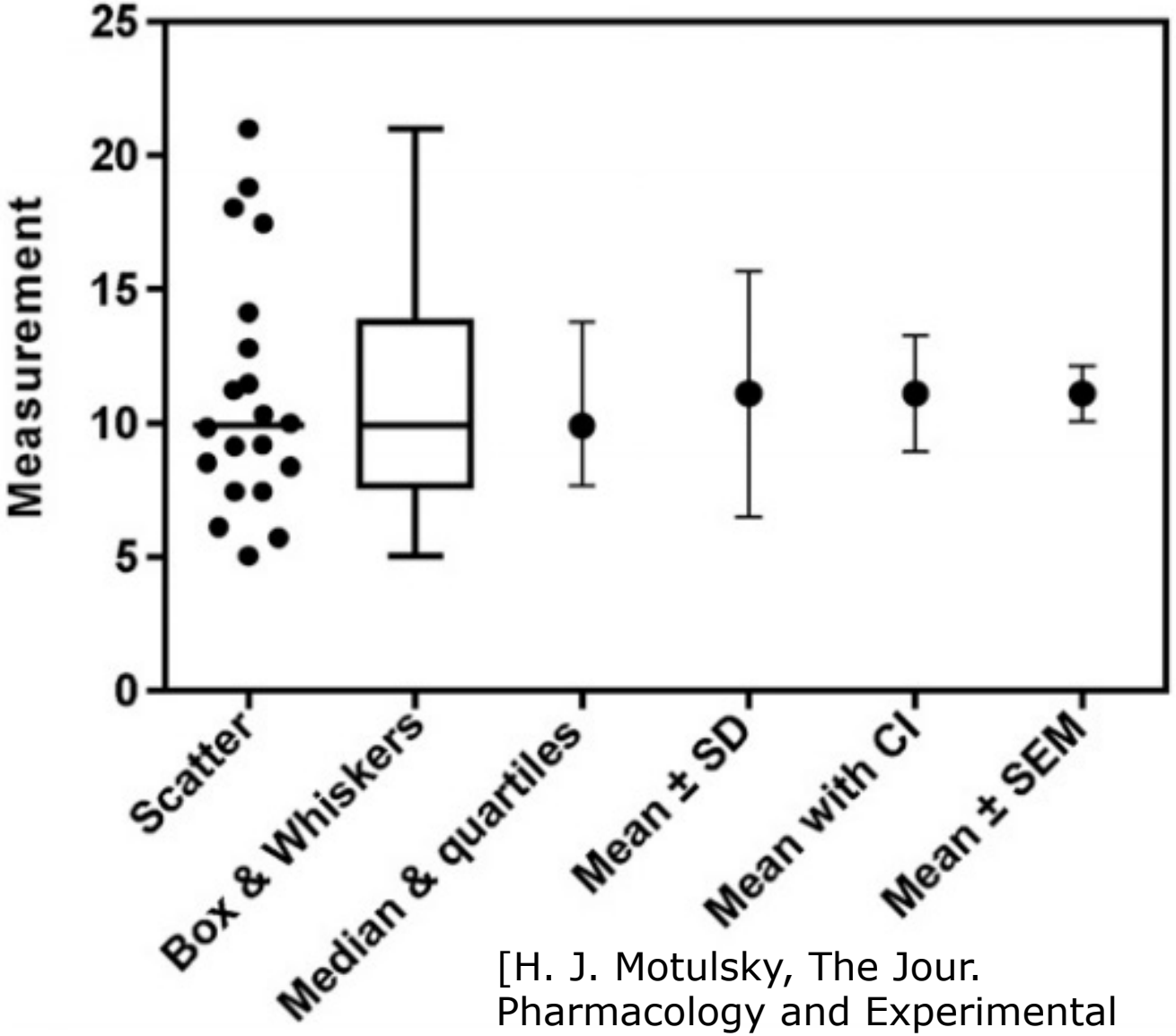
上下打ち切り値(censored value)： (例)-1.3 未満の値はすべて -1.3
1.3 以上の値はすべて 1.3 として扱う

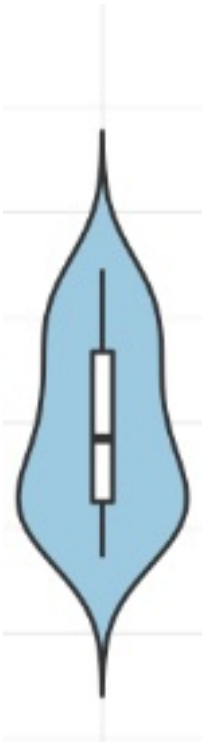
飽和値 (saturated value) :(例) 計測装置の限界により値に上限がある

デジタル化・量子化 (quantization) : (例) A/D変換の限界により値の小さいところで分布が間欠的になる

ヒストグラム・散布図で初めて気づくこと多大！
データを得たら、まずヒストグラム書くこと

ばらつきの可視化方法いろいろ



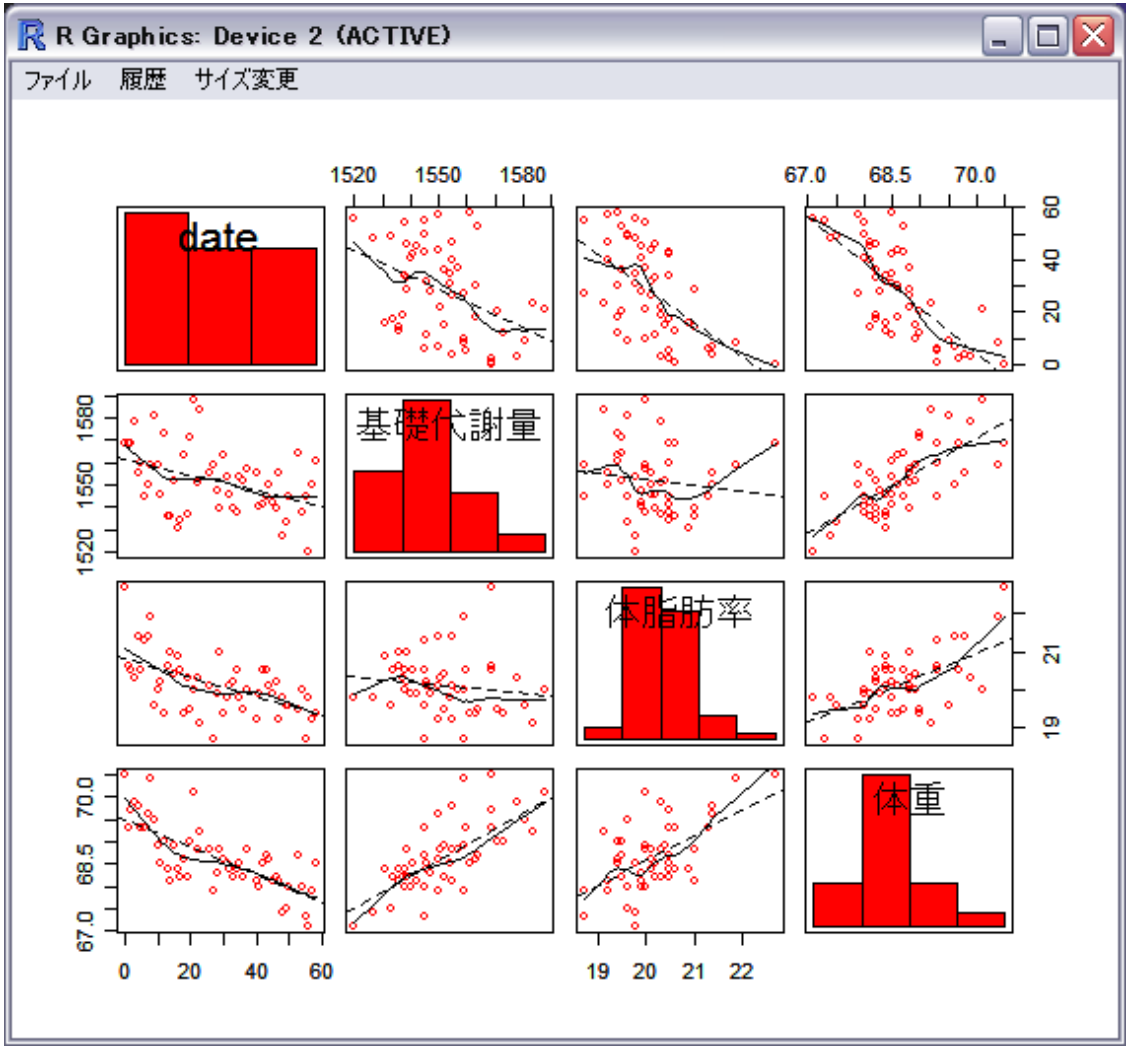


Violin plot

[H. J. Motulsky, The Jour. Pharmacology and Experimental Therapeurics, 2014]

京都大学 KYOTO UNIVERSITY

ヒストグラムと散布図を並べる

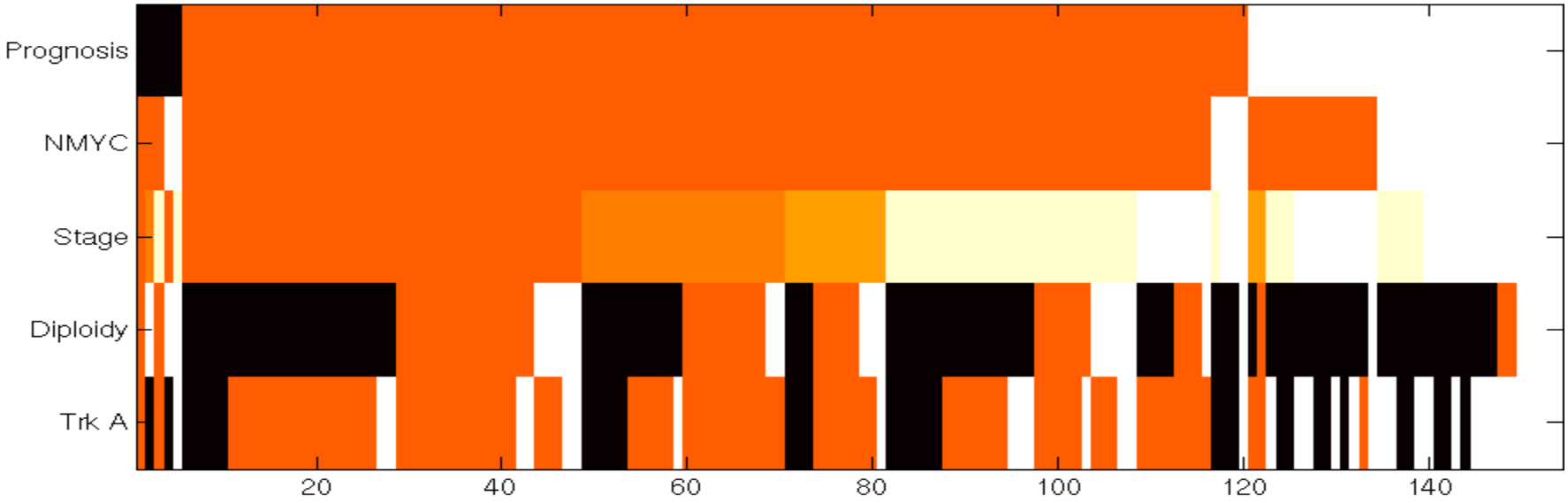


http://plaza.umin.ac.jp/~takeshou/R/Rcmdr02.html

京都大学 KYOTO UNIVERSITY

心得そのに：なにはともあれ色つき行列にしてただ眺める

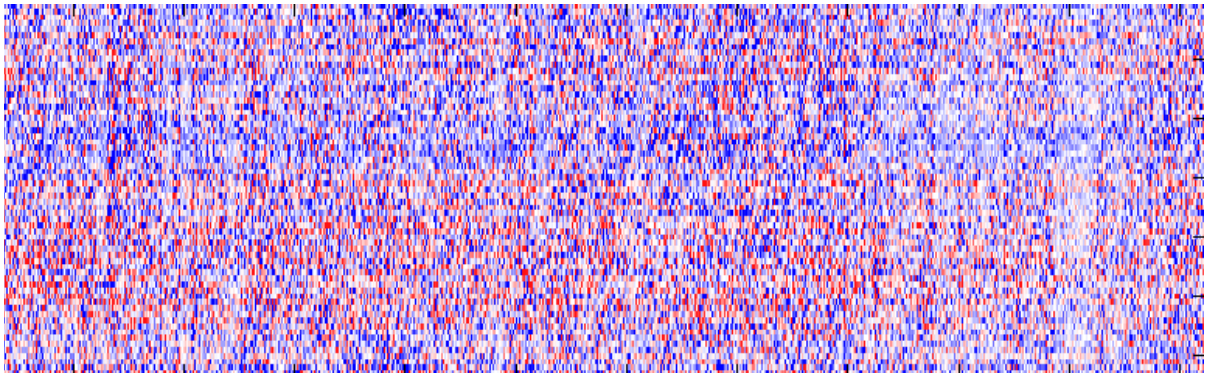
- 例： 小児がん150症例のラベルデータ
  - 縦軸はラベル （ラベルの重要度順にソート）
  - 横軸は症例 （ラベルの値を使ってソート）
  - 白: 予後悪関連、橙: 予後良関連、黒：欠測

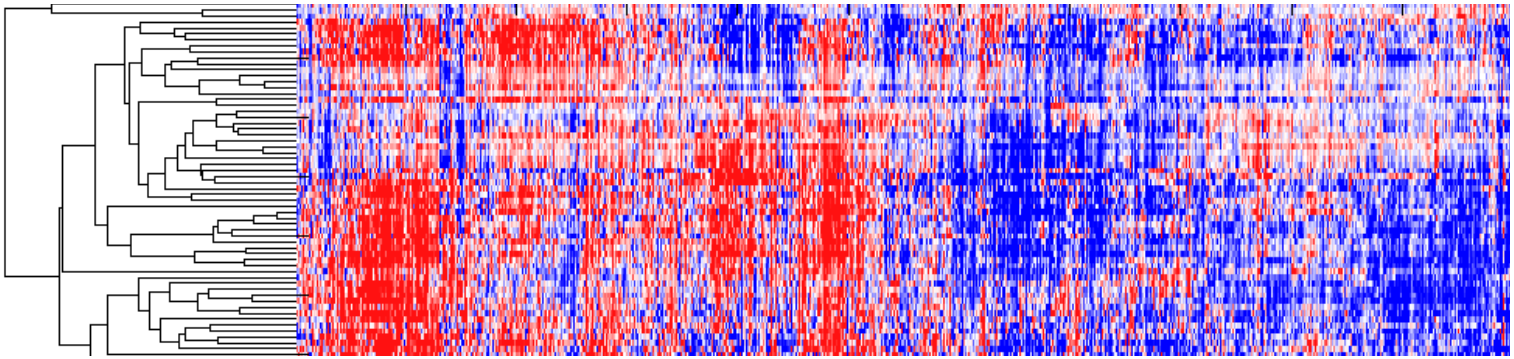


コツ：ラベルデータには、目的に沿った意味がある。
独断・偏見・議論に基づき、重要度をつけると
分かりやすく可視化できる。

色つき行列形データを眺める

1. 数値に色付けしてただ並べる
2. 上手に並べる(階層クラスタリング)



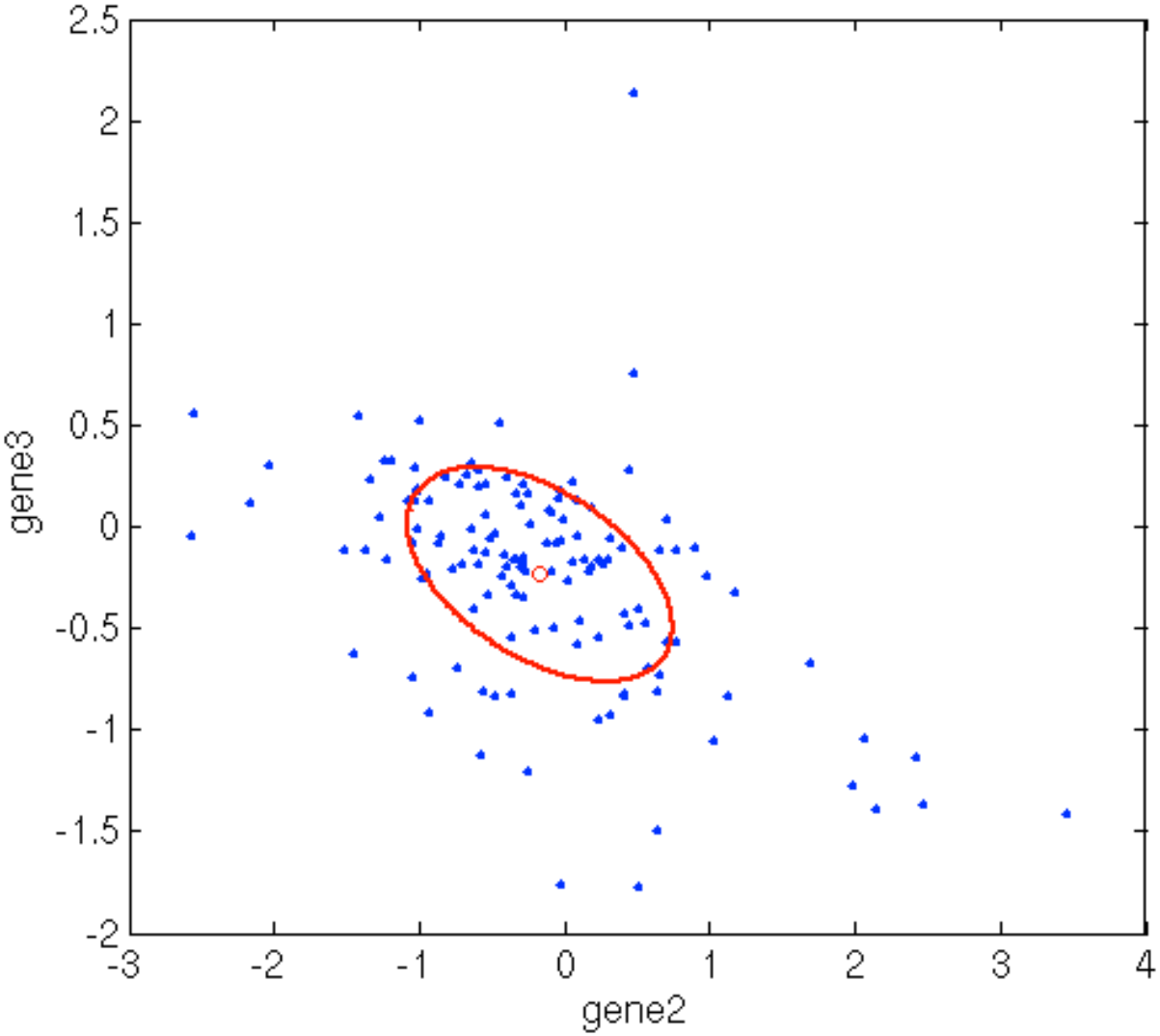


演習：ヒストグラムを描いてみよう
平均と標準偏差を示してみよう

演習資料
https://github.com/shigeyukioba/biostat/

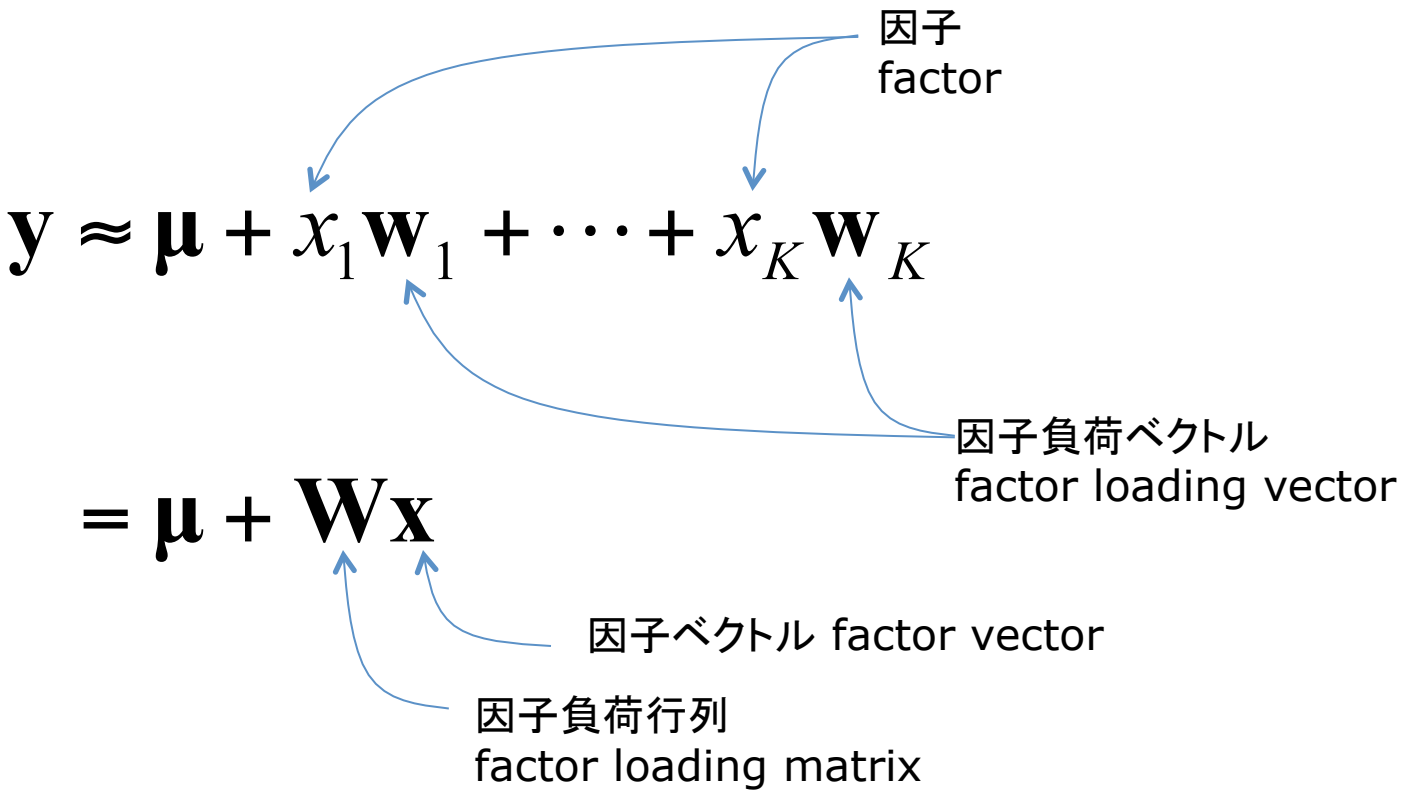
サンプル用遺伝子発現データ行列
NBLexpression.dat

演習：散布図を描いてみよう
共分散楕円を加えてみよう



因子化 factorization

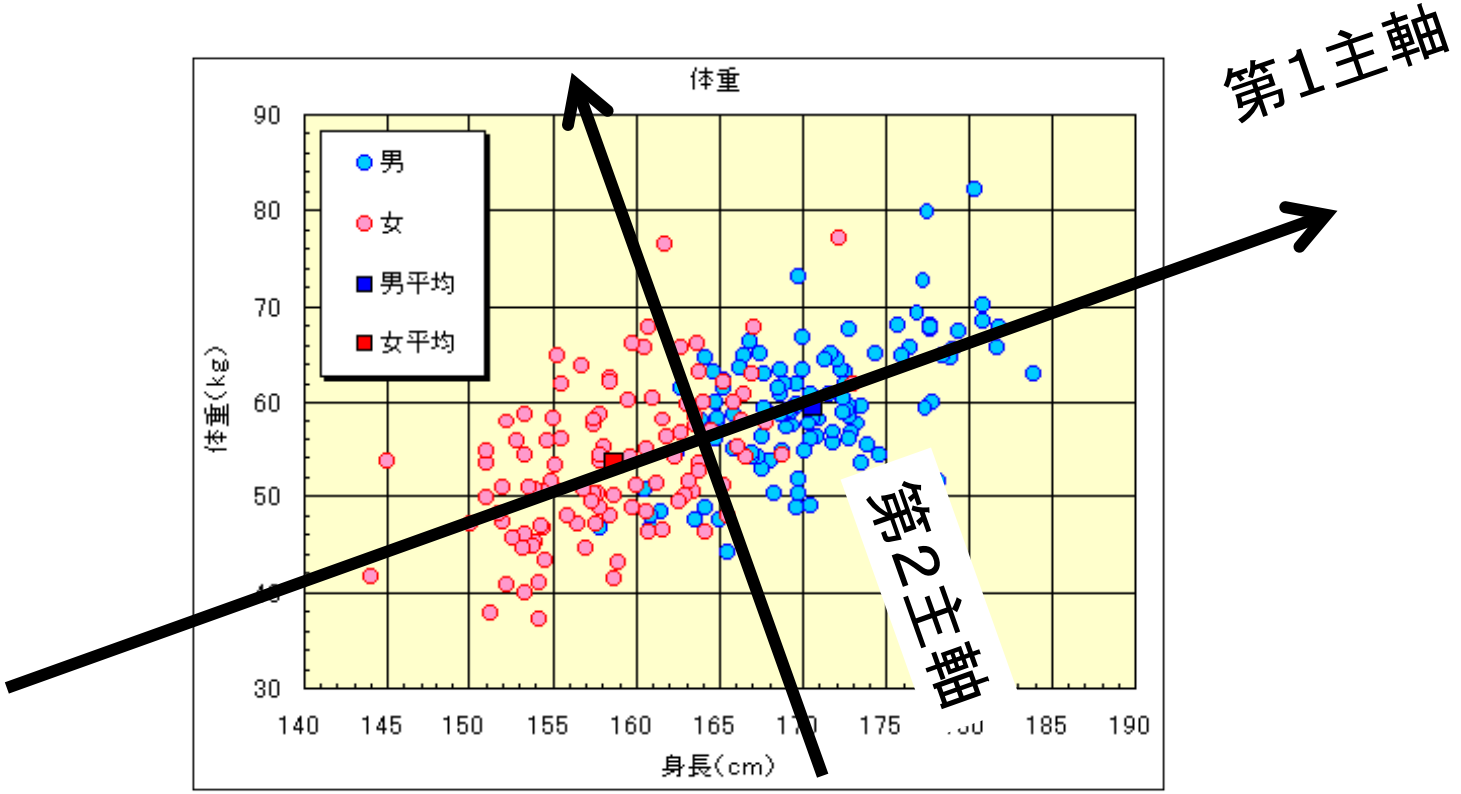
- 高次元ベクトル　$\mathbf{y} \in \Re^M$　の変動を、
  少数K $(K < M)$の成分（因子）の変動で表したい



主成分分析

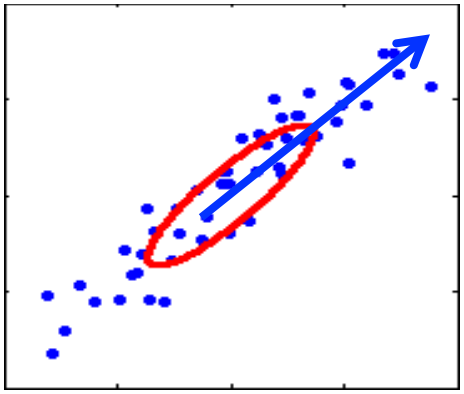
principal component analysis

- 多変量の変動を、少数個の「成分」の変動に分解



資料：http://homepage3.nifty.com/orangejuice/

主成分分析の方法



- 共分散行列 C を固有値分解

$$\mathbf{C} = \mathbf{P}\boldsymbol{\Gamma}\mathbf{P}^T,\ \text{s.t.}\ \mathbf{P}\mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}_2, \boldsymbol{\Gamma} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{pmatrix}$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_M$$

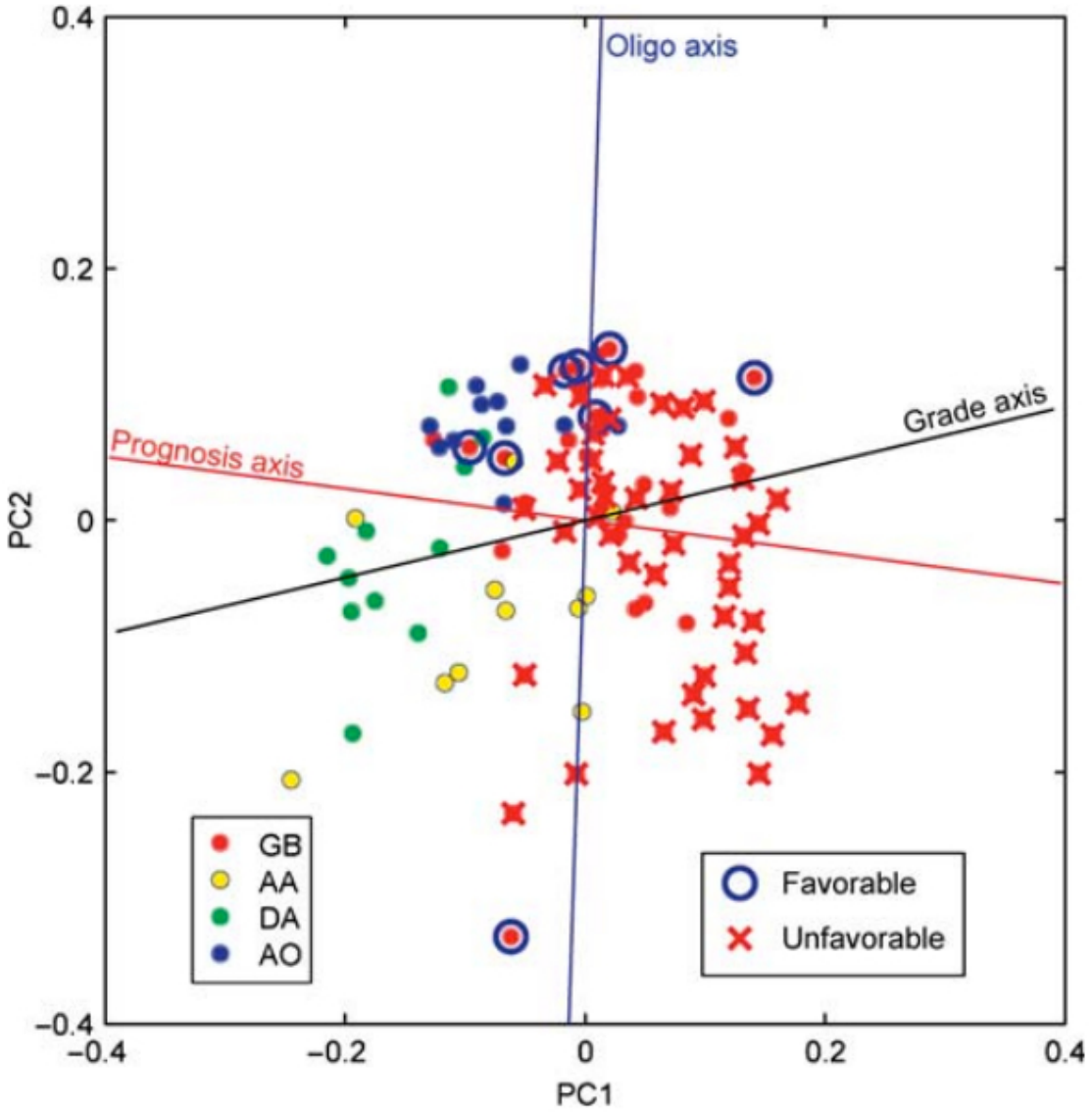
- 固有値の大きい順に K 個の固有ベクトルをとり、基底とする

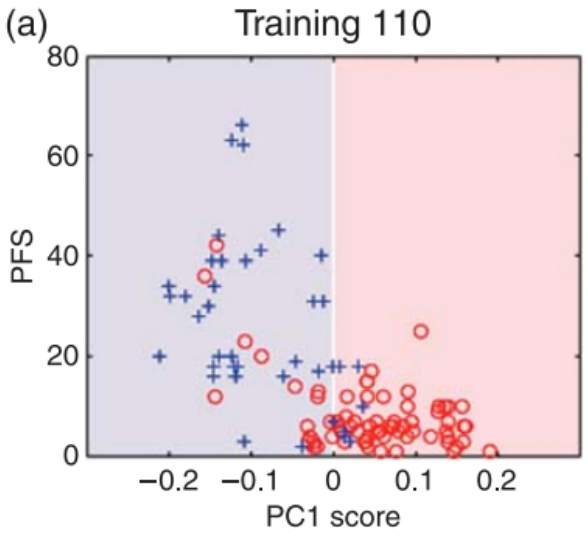
$$\mathbf{W} = \begin{pmatrix} \mathbf{w}_1 & \mathbf{w}_2 \ldots & \mathbf{w}_K \end{pmatrix}$$

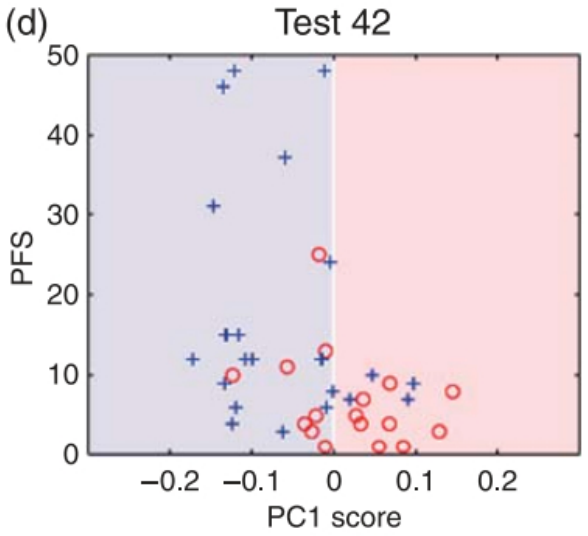
- データ点 y 毎に、主成分得点を求める

$$x_k = \lambda_k^{-1/2}\mathbf{w}_k\mathbf{y}$$

脳腫瘍発現データのPCA結果







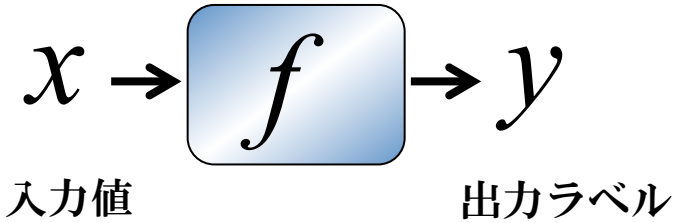
京都大学
KYOTO UNIVERSITY

演習：主成分分析を行ってみよう

判別と判別関数

- 判別関数

$$h(\mathbf{x}) \in \Re$$



- 判別関数による判別

$$h(\mathbf{x}) < 0 \quad \text{ならば} \quad f(\mathbf{x}) = -1$$

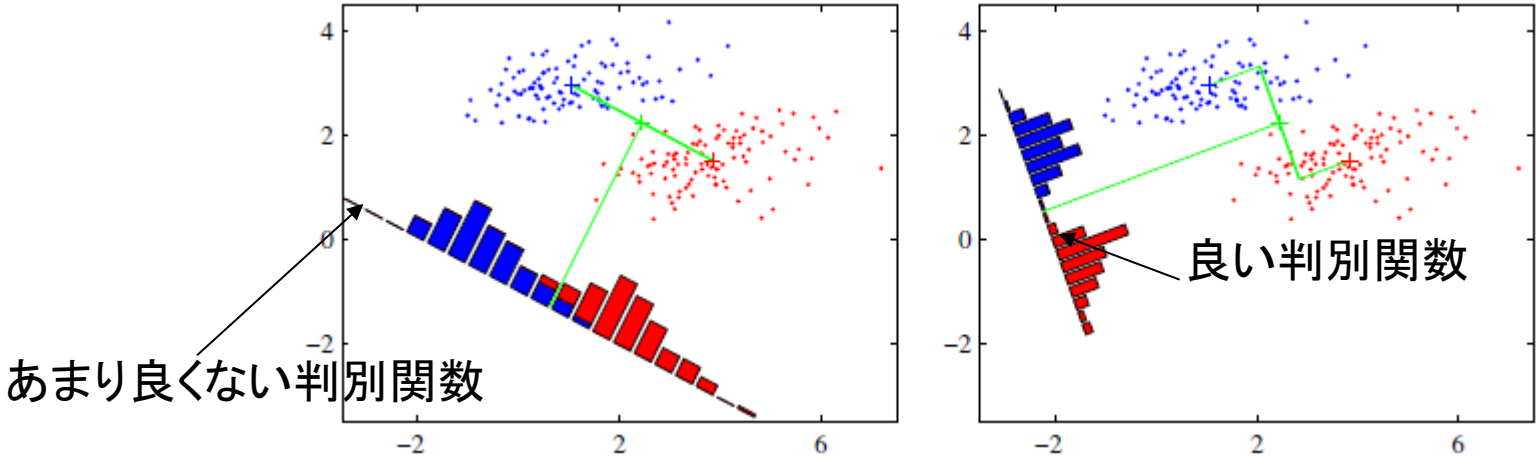
$$h(\mathbf{x}) \geq 0 \quad \text{ならば} \quad f(\mathbf{x}) = +1$$

線形判別器

- 線形判別関数　＝　ベクトル x の線形射影
- 射影方向次第で、判別性能が異なる

$$h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b = b + \sum_{i=1}^{K} w_i x_i$$

  - 最も適切な射影 w を
    ラベル付きデータから学習によって決める。



まとめ

- バイオマーカーの正解率は2種類
  - 感度 Sensitivity, 特異度 Specificity
- 統計量はいろいろある
  - 連続値マーカーの性能を示す「統計量」の例
    t-statistic, AUC score
- 多変量データに出会ったら
  1にヒストグラム、2に散布図、3,4が無くて、
  5に色行列
- 共分散楕円が分かれば主成分分析が分かる

レポート課題(1/2)

- データ可視化の練習
  - 遺伝子の任意の2つ組について、発現量の散布図を描きます。このさい予後良症例を青色で、予後悪症例を赤色で描いてください。
  - 書き方は、
    http://ishiilab.jp/member/oba/seimeidoutai2013
    の Exercise 4. を参考にしてください。
  - これに、赤と青の共分散楕円を重ねて描いてください。
  - 遺伝子の2つ組を、6組ほど任意に試してみて、同様の図を描いてみましょう。

レポート課題(2/2)

- 統計量
  - 遺伝子2つを用いて予後予測を行います。
    可視化した6組のなかで比較したとき、どの組を用いるのがベストでしょうか？まずは目で見て直感的に評価してみてください。
  - 予後予測の良さを評価した基準を定式化するために、どのような統計量を用いるのが良さそうでしょうか？考えてみてください。それを、まずは言葉もしくは図や絵で、可能ならば数式で表現してみてください。