

マイクロアレイデータおよび生物学データの統計解析

金谷重彦(奈良先端科学技術大学院大学・情報科学研究科・情報生命学専攻・比較ゲノム学講座)

1. Bioinformatics

バイオインフォマティクスとは、遺伝情報のコンピュータハンドリングおよび情報解析に基礎をおいた分子生物学と定義でき、究極的には細胞を情報技術で再現することにある。その一つとして細胞をシステムとして記述することが挙げられる。生命を分子生物学的な立場でシステムとして記述する場合、RNA、DNA、代謝物質などの生体物質を要素として規定し、基質-生成物、発現制御関係などの要素間の関係を情報学的に記述することが必要となる。膨大なゲノム情報データと実験技術の進歩に伴うポストゲノム解析による種々の分子生物学的事象の情報（インタラクトーム、トランスクリプトーム、メタボローム）を統合し信頼性の高い要素間の関係を抽出することが、生命システムとしての普遍性および多様性を理解するための重要な課題となっている。トランスクリプトーム解析では、ストレス特異的に発現する遺伝子を探索するなどの目的で複数の実験条件における転写量パターン類似性から遺伝子を分類するために、[1]有意に発現量が変化する実験条件を含む遺伝子のみを統計解析により選択し、[2]複数の実験条件間で相関して発現が変化する遺伝子を分類する。さらに、[3]遺伝子発現プロファイルをもとに診断のための数理モデルを作成するなどの用途が考えられる（図1）。また、メタボローム解析においては、種々の条件で培養された生物サンプルの機器分析チャートをもとに[1]対応するピークごとに行列を作成し、[2]条件の違いにより強度が変化するピークを選択し、[3]総合的に種々の条件における生物サンプルの違いを評価する、あるいは条件特異的に量が増えるあるいは減るピークを探索し、その代謝物がなんであるかについて化学分析を行い、代謝のレベルで生物サンプル間进行分类する。化学の分野のみならず生物学、特にポストゲノム解析において多変量解析は生物などを対象とした測定サンプルを体系的に理解する一助となる。ケモメトリクスでは化学の分野を中心にこのような体系化をめざした。一方、ポストゲノム解析では、生物サンプルについてこのような体系化を目指していることとなる。さらに、このような体系化を含む生物をシステムとして理解するための情報解析がバイオインフォマティクスの一つの目標となっている。

トランスクリプトームデータおよびメタボロームデータを体系的に理解するために多変量解析法が有効な方法でありことから、本講義では、トランスクリプトームおよびメタボロームにおける分析データにおける統計処理法、さらには、これらのデータの体系的理解を目標とした多変量解析法を解説する。

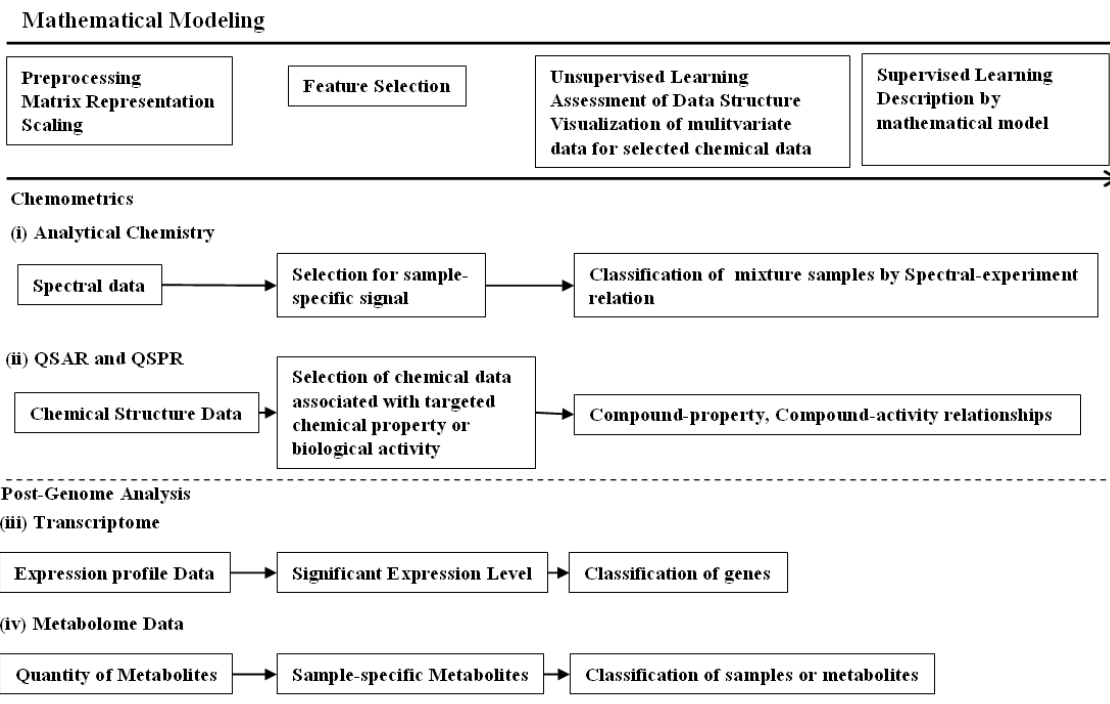


図1 数理モデリング

2.統計学的準備

「統計学は、観測に基づくデータを対象とする数学であり、(i)集団の研究、(ii)変動の研究、(iii)データの簡約方法に関する研究である」と統計学者フィッシャーは言った。生物学者は何かと p-value というもので試験区間の違い、配列間の類似度を表現することに慣れている。この p-value の意味はなにかを理解し、トランスクリプトーム解析やメタボローム解析に役立てる方法を考えよう。トランスクリプトーム解析やメタボローム解析において遺伝子あるいはメタボライトの量を繰り返し測定したときの値の再現性、区間推定がまずはじめに必要となる。この章では、区間推定を中心とした統計学の基礎を整理する。

2.0 基本統計量と用語

Σについて

$$x_1 + x_2 + \dots + x_n$$

をまとめて、

$$\sum_{i=1}^n x_i$$

と表す。

すなわち、

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

である。

問題 1

以下の式を Σ を用いて表せ。

$$(1) x_1 y_1 + x_2 y_2 + \dots + x_n y_n =$$

$$(2) x_1^2 + x_2^2 + \dots + x_n^2 =$$

$$(3) (x_1 x_1 + x_1 x_2 + \dots + x_1 x_n) + (x_2 x_1 + x_2 x_2 + \dots + x_2 x_n) + \dots + (x_m x_1 + x_m x_2 + \dots + x_m x_n) =$$

$$(4) (x_1 x_1 + x_1 x_2 + \dots + x_1 x_n) + (x_2 x_1 + x_2 x_2 + \dots + x_2 x_n) + \dots + (x_n x_1 + x_n x_2 + \dots + x_n x_n) =$$

問題 2

$x_1 = 1, x_2 = 2, x_3 = 3, \dots, x_{10} = 10$ のとき以下の値をもとめよ。

$$(1) x_1 + x_2 + \dots + x_{10} =$$

$$(2) x_1^2 + x_2^2 + \dots + x_{10}^2 =$$

$$(3) (x_1 x_1 + x_1 x_2 + \dots + x_1 x_{10}) + (x_2 x_1 + x_2 x_2 + \dots + x_2 x_{10}) + \dots + (x_{10} x_1 + x_{10} x_2 + \dots + x_{10} x_{10}) =$$

確率変数

おのおのの値に対して出現する確率がそれぞれ決まっている変数を確率変数という。

期待値 (平均値)

確率変数 x 値 $x_1, x_2, \dots, x_i, \dots, x_n$ のそれぞれ確率を $P_1, P_2, \dots, P_i, \dots, P_n$ とする。このとき x の期待値 $E(x)$ は、

$$E(x) = x_1 P_1 + x_2 P_2 + \dots + x_n P_n = \sum_{i=1}^n x_i P_i \quad (2.1)$$

である。いま、 $P_1 = P_2 = \dots = P_i = \dots = P_n = \frac{1}{n}$ のとき式(2.1)は

$$E(x) = \frac{\sum_{i=1}^n x_i}{n} \quad (2.2)$$

となる。

問題 3

(1) 1,2,3,4,5,6 の目の出現確率がそれぞれ 1/6 のさいころがある。このさいころを一回ふったときに出る目の期待値を求めなさい。

(2) 1,2,3,4,5,6 の目の出現確率がそれぞれ 1/12, 1/12, 4/12, 4/12, 1/12, 1/12 のさいころがある。このさいころを一回ふったときに出る目の期待値を求めなさい。

偏差、分散、標準偏差

確率変数 x とその期待値 $E(x)$ との差を **偏差(deviation)** とよぶ。偏差の 2 乗の期待値を **分散(variance)** $V(x)$ という。

$$V(x) = E\left[\{x - E(x)\}^2\right] \tag{2.3}$$

$P_1 = P_2 = \dots = P_i = \dots = P_n = \frac{1}{n}$ のとき

$$\begin{aligned} V(x) &= E\left[\{x - E(x)\}^2\right] \\ &= E\left[\{x_1 - E(x)\}^2 + \{x_2 - E(x)\}^2 + \dots + \{x_n - E(x)\}^2\right] \\ &= \{x_1 - E(x)\}^2 P_1 + \{x_2 - E(x)\}^2 P_2 + \dots + \{x_n - E(x)\}^2 P_n \\ &= \{x_1 - E(x)\}^2 \frac{1}{n} + \{x_2 - E(x)\}^2 \frac{1}{n} + \dots + \{x_n - E(x)\}^2 \frac{1}{n} \\ &= \frac{\sum_{i=1}^n \{x_i - E(x)\}^2}{n} \end{aligned} \tag{2.4}$$

問題 4

(1) 1,2,3,4,5,6 の目の出現確率がそれぞれ 1/6 のさいころがある。このさいころを一回ふったとき出る目の分散を求めなさい。

(2) 1,2,3,4,5,6 の目の出現確率がそれぞれ 1/12, 1/12, 4/12, 4/12, 1/12, 1/12 のさいころがある。このさいころを一回ふったとき出る目の分散を求めなさい。

標準偏差(standard deviation)

分散の平方根を標準偏差という。

$$D(x) = \sqrt{V(x)} \tag{2.5}$$

変動係数(coefficient of variance)

標準偏差 $D(x)$ を期待値 $E(x)$ で割ることにより得られる値を変動係数 CV という。

$$CV = \frac{D(x)}{E(x)} = \frac{\sqrt{V(x)}}{E(x)} \tag{2.6}$$

平均値、分散の和

k 個の確率変数 $x_1, x_2, \dots, x_i, \dots, x_k$ が互いに独立であるとする。定数 $a_1, a_2, \dots, a_i, \dots, a_k$ を係数とする任意の一次式 $z = a_1x_1 + a_2x_2 + \dots + a_kx_k$ の期待値および分散は

$$E(z) = a_1E(x_1) + a_2E(x_2) + \dots + a_kE(x_k) \tag{2.7}$$

$$V(z) = a_1^2V(x_1) + a_2^2V(x_2) + \dots + a_k^2V(x_k) \tag{2.8}$$

である。

問題 5 互いに独立な二つの確率変数 x, y について、 $z = x - y$ の期待値 $E(x - y)$ と分散 $V(x - y)$ を $E(x)$, $E(y)$, $V(x)$, $V(y)$ を用いて表してみよう。

式(2.7)から $z = a_1x + a_2y$ について

$$E(z) = a_1 E(x) + a_2 E(y)$$

となる。 $a_1 = 1$, $a_2 = -1$ を代入すると、

$$E(z) = E(x) - E(y)$$

式(2.8)から $z = a_1 x + a_2 y$ について

$$V(z) = a_1^2 V(x) + a_2^2 V(y)$$

となる。 $a_1 = 1$, $a_2 = -1$ を代入すると、

$$V(z) = V(x) + V(y)$$

ここで重要なことは、互いに独立な二つの確率変数 x, y の差の平均値は二つの平均値の差であるが、これら二つの確率変数 x, y の差の分散はそれぞれの変数の分散の和となるということである。

問題 6 互いに独立な二つの確率変数 x, y のいずれも 0 と 1 の値をとる。また、それぞれの値が出現する確率が、

$$P(x=0) = \frac{1}{2}, P(x=1) = \frac{1}{2}, P(y=0) = \frac{1}{2}, P(y=1) = \frac{1}{2}$$

で与えられるとき、 $z = x - y$ の期待値 $E(x - y)$ と分散 $V(x - y)$ を求めてみよう。

偏りと変動

測定値の誤差は、通常、**偏り (カタヨリ)** と **変動 (バラツキ, ランダム誤差)** の二つの概念により評価される。偏りとは、測定値と測定系から得られる理論値 (真の値) の差のことをいう。一方、変動とは、ある測定系において繰り返し行われた測定値について代表値との誤差をいい、一般には標準偏差を用いることが多い。

例 真の値(100)が既知の対象 X について測定装置 A により繰り返し測定を行ったところ、101, 99, 98, 100 の測定値を得た。一方、測定装置 B により繰り返し測定を行ったところ、110, 111, 109, 108 の測定値を得た。測定装置 A についての測定値の平均値は 99.5、標準偏差は 1.29 である。一方、測定装置 B に関しては、平均値は 109.5、標準偏差は 1.29 である。測定装置 A による測定の平均値と真の値の差は 0.5、一方、測定装置 B については 9.5 となることから、測定装置 B については偏り誤差が大きく含まれていることがわかる。標準偏差については共に 1.29 であるので測定値のバラツキの程度は同等ととらえることができる。

2.1 正規分布

2.1.1 正規分布 $N(\mu, \sigma^2)$

物理量、化学分析における測定値は、測定装置などから生じる偏りによる誤差を除いたとしても、たくさんの偶然的な誤差 (ランダム誤差) が生ずる。このような偶然誤差の分布は正規分布に従う。**正規分布 (normal distribution)** は平均 μ と標準偏差 σ (または分散 σ^2) により規定される分布であり、 $N(\mu, \sigma^2)$ と略記する。その確率密度関数と累積分布関数は式(2.9)および(2.10)により表される。

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \quad (2.9)$$

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} dz \quad (2.10)$$

標準正規分布 $N(0,1^2)$

式(2.9)について平均 $\mu = 0$ 、標準偏差 $\sigma = 1$ とおいた正規分布を標準正規分布(standard normal distribution; $N(0,1^2)$)と呼ぶ。標準正規分布の確率密度関数と累積分布関数は、それぞれ、式(2.11)および(2.12)により表される。

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} \quad (2.11)$$

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \quad (2.12)$$

標準正規分布の性質

$N(0,1^2)$ の確率密度関数と累積分布関数を図 2 に示す。ここで

$$F(1) = \int_{-\infty}^1 f(z) dz = 1 - 0.1587 = 0.8413 \quad (2.13)$$

$$F(2) = \int_{-\infty}^2 f(z) dz = 1 - 0.0228 = 0.9772 \quad (2.14)$$

$$F(3) = \int_{-\infty}^3 f(z) dz = 1 - 0.0013 = 0.9987 \quad (2.15)$$

もまた図中に記述されている。

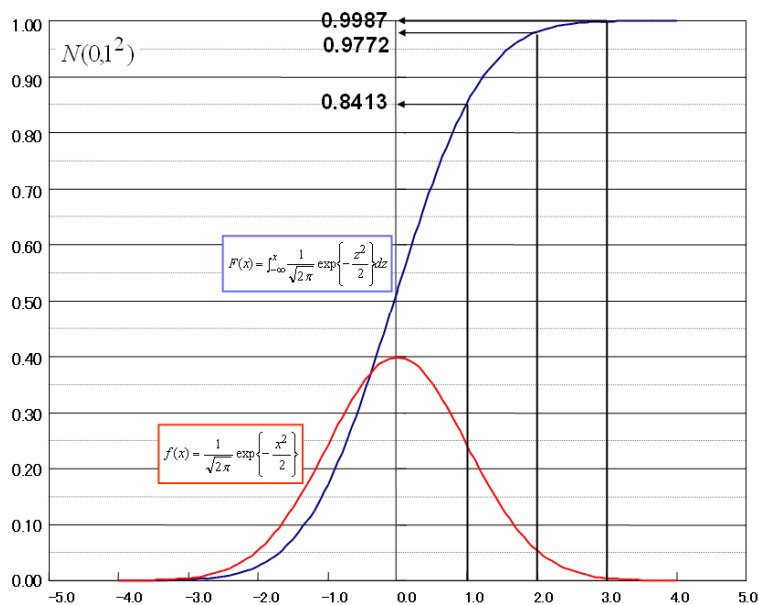


図2 $N(0,1^2)$ の確率密度関数と累積分布関数

(質問) 累積分布関数が確率密度関数を超えているところがあるのはなぜか。

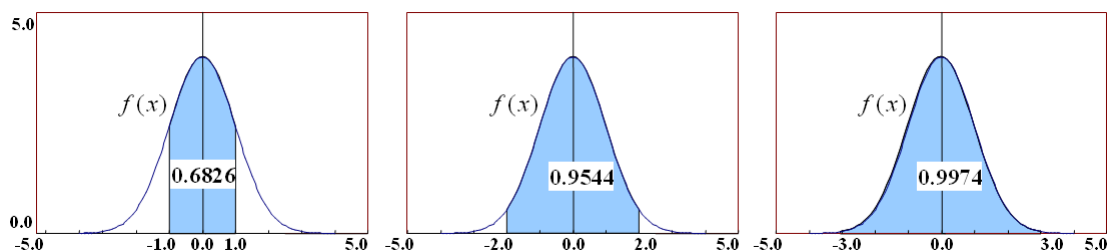


図3 $N(0,1^2)$ の確率密度関数

測定値が正規分布によりランダムに分布するとき、平均値と標準偏差が既知の場合、測定値が $\mu - \sigma \leq x \leq \mu + \sigma$ の区間に存在する確率は

$$P(\mu - \sigma \leq x \leq \mu + \sigma) = 0.6826 \tag{2.16}$$

より、0.6826 となる(図3)。また、 $\mu - 2\sigma \leq x \leq \mu + 2\sigma$ および $\mu - 3\sigma \leq x \leq \mu + 3\sigma$ の区間に存在する確率は

$$P(\mu - 2\sigma \leq x \leq \mu + 2\sigma) = 0.9544 \tag{2.17}$$

$$P(\mu - 3\sigma \leq x \leq \mu + 3\sigma) = 0.9974 \tag{2.18}$$

となる。このことは、測定値が偶然 $\mu - \sigma \leq x \leq \mu + \sigma$ を外れる確率は $1 - 0.6826 = 0.3174$ 、すなわち、おおよそ3回に1回はこの区間をはずれることを意味する。標準偏差の2倍の区間 ($\mu - 2\sigma \leq x \leq \mu + 2\sigma$) から外れる確率は $0.0456 (= 1 - 0.9544)$ となるので、偶然100回の測定のうち約5回は外れる可能性があることを意味する。すなわち、 $\mu - 2\sigma \leq x \leq \mu + 2\sigma$ を外れたデータが同一の母集団からのサンプルである可能性は非常に小さいということになり、なにか異なった原因によるはずれ値の可能性を検討すべきである。このように測定値が正規分布に従うことがわかると得られた値が偶然おこる可能性を分布の平均値と分散の値をもと

に確率により予測することが可能となる。

2.2 平均値の分布

ある母集団から抜き取られた大きな n のサンプル $x_1, x_2, \dots, x_i, \dots, x_n$ から母平均 μ を推定するには、サンプルの平均

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n} \quad (2.19)$$

が用いられる。しかし、推定値 \bar{x} の値はサンプルにより異なり、母平均 μ と一致するとは限らない。繰り返し実験により得られた平均値が統計的にどのように分布するかがわかれば、同一の条件で得られた平均値の再現性を検討することができる。この節では推定値 \bar{x} の分布の性質について述べる。いま、同一の母集団から独立に n 個のサンプルをとり平均値を求めるわけであるから、同一の母集団における確率変数を $x_1, x_2, \dots, x_i, \dots, x_n$ を設定する。このとき、 \bar{x} もまた確率変数である。サンプルから計算された平均値はその一つの実現値であり、当然バラツキをもつ。母平均 μ の推定に用いる標本における平均値 \bar{x} を確率変数とみるとき、 \bar{x} を μ の推定量という。 \bar{x} の期待値 $E(\bar{x})$ 、分散 $V(\bar{x})$ 、および標準偏差 $D(\bar{x})$ はそれぞれ式 (2.20-2.22) により導かれる。

$$\begin{aligned} E(\bar{x}) &= E\left[\frac{1}{n}(x_1 + \dots + x_n)\right] \\ &= \frac{1}{n}E[x_1 + \dots + x_n] \\ &= \frac{1}{n}\{E(x_1) + E(x_2) + \dots + E(x_n)\} \\ &= \frac{1}{n}n\mu \\ &= \mu \end{aligned} \quad (2.20)$$

$$\begin{aligned} V(\bar{x}) &= V\left[\frac{1}{n}(x_1 + \dots + x_n)\right] \\ &= \frac{1}{n^2}V[x_1 + \dots + x_n] \\ &= \frac{1}{n^2}\{V(x_1) + V(x_2) + \dots + V(x_n)\} \\ &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned} \quad (2.21)$$

$$D(\bar{x}) = \frac{\sigma}{\sqrt{n}} \tag{2.22}$$

母平均が μ かつ標準偏差 σ の正規母集団から独立に抜き取られた大きさ n のサンプルの平均 \bar{x} は平均 μ 、標準偏差 σ/\sqrt{n} の正規分布に従う。このことは、測定回数 n によりバラツキの程度は $1/\sqrt{n}$ 小さくなることを意味しており、バラツキの程度を考慮して繰り返し実験の回数を決めるための判断材料となる。ある推定量の期待値が母数(population parameter)と一致するとき、その推定値を普遍推定量とよぶ。不偏推定量の標準偏差が小さいほど精度がよいことを意味する。平均値の場合、サンプル数 n を増やすに従いその精度が増し、その効果は $1/\sqrt{n}$ 程度である。図4に $N(0, 1)$ の正規分布からサンプル数 $n=2, 3, \dots, 10$ と換えたときの平均値の分布を示す。この図から繰り返し測定回数を1回増やすことによる平均値のバラツキの範囲は、2から3回と一回増やした場合と繰り返し測定を9から10回と一回増やした場合で大きく異なることを示しており、繰り返し測定回数を9回から10回に増やすことによる平均値のバラツキの範囲の減少は、繰り返し測定回数を2回から3回に増やすことによる平均値のバラツキの範囲の減少ほど望めないことがわかる。

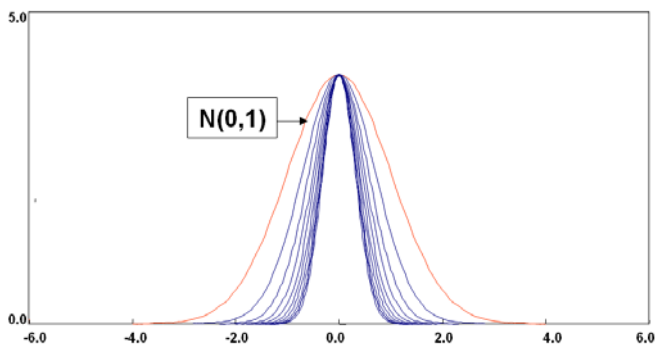


図4 $N(0, 1)$ の正規分布からサンプル数 $n=2, 3, \dots, 10$ と換えたときの平均値の分布

2.2.1 母平均 μ の検定・区間推定

母標準偏差 σ が既知の場合の母平均 μ の検定についてははじめに説明し、次に、母標準偏差 σ が未知の場合の母平均 μ の検定について述べる。実際の測定データの統計解析において母標準偏差 σ が既知であることはまれであるので実務的には **2.2.1.2 母標準偏差 σ が未知の場合の解析手順を理解し、教養として母標準偏差 σ が既知の場合と 2.2.1.1 母標準偏差 σ が未知の場合の違いを理解すればよい。**

2.2.1.1 母標準偏差 σ が既知の場合

(a) 母平均の検定

問題 7 マススペクトルのある標準サンプルに対するピークの m/z 値の平均値 104.840、標準偏差 $\sigma=0.028$ ($N=100$) であった。元旦後、新たに同一の標準サンプルのピークの m/z 値を 10 回測定したところ、104.902, 104.840, 104.839, 104.835, 104.872, 104.870, 104.877, 104.832, 104.900, 104.905。個の値を得た。これらの 10 回の測定結果は以前の m/z 値と同一とみなすことができるであろうか。

まずはじめに、元旦前の平均値が母平均と等しいと仮定する。すなわち、 $\mu=104.840$ とおく。また、母

標準偏差 $\sigma=0.0280$ がわかっているので正規分布が使える。サンプルの平均値 \bar{x} を式(2.23)で変換して得られた u は標準正規分布に従う。

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad (2.23)$$

ここで n は繰り返し測定回数である。いま、得られたサンプルの平均値は104.8672であるので元旦を過ぎてからの測定における $u = u_0$ の値を求めると

$$u_0 = \frac{104.8672 - 104.840}{\frac{0.028}{\sqrt{10}}} = 3.07193$$

$u_0 \geq 3$ となる確率は

$$F(3) = \int_{-\infty}^3 f(z) dz = 1 - 0.0013 = 0.9987$$

から $1 - 0.9987 = 0.0013$ である。すなわち、元旦後に得られた平均値104.8672はめったにおこらない($p=0.0013$)ということになり、「休み中にかまわずに“装置がすねた！“可能性がある」と考えられる。このように仮説 $H_0: u = u_0$ は捨てられたことになる。

平均値の仮説検定を定式化すると以下のようなになる。

1. 平均値に関する仮説を設定する。 $H_0: u = u_0$
2. サンプル平均 \bar{x} と標準偏差 $\frac{\sigma}{\sqrt{n}}$ をもとに $u_o = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$ を計算する。
3. u_o の絶対値がある確率値に対する基準値 $crit(p)$ より大きいとき、すなわち $|u_o| \geq crit(p)$ のとき仮説 $H_0: u = u_0$ は捨てられる。

(b) 母平均の区間推定

測定値の平均値と繰り返し測定回数 n と母標準偏差が既知のとき母平均の区間を推定することができる。

まずはじめに、

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

が $N(0, 1)$ の正規分布に従うとすると $p \geq 0.05$ における母平均の範囲は

$$\left| \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \right| \leq \text{crit}(p) \quad (2.24)$$

を解くことにより得られる。すなわち、

$$\bar{x} - \text{crit}(p) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \text{crit}(p) \frac{\sigma}{\sqrt{n}} \quad (2.25)$$

となる。

上述の例で、元旦前のマススペクトルのある標準サンプルに対するピークの m/z 値の平均値の 95%信頼限界 ($p=0.05$) における測定値の区間を推定する。いま、平均値 $\bar{x} = 104.840$ 、標準偏差 $\sigma = 0.028$ である。95%信頼限界における区間は

$$104.840 - \text{crit}(p = 0.05 / 2) \frac{0.028}{\sqrt{100}} \leq \mu \leq 104.840 + \text{crit}(p = 0.05 / 2) \frac{0.028}{\sqrt{100}}$$

ここで、

$$F(1.96) = \int_{-\infty}^{1.96} f(z) dz = 0.9750$$

より $\text{crit}(p = 0.05 / 2) = 1.96$ となり、

$$104.835 \leq \mu \leq 104.845.$$

が得られる。

(c) 母平均の差の検定と区間推定

平均 μ_1 、分散 σ_1 の正規母集団から独立に n_1 個のサンプルにおける平均値を \bar{x}_1 、平均 μ_2 、分散 σ_2 の正規母集団から独立に n_2 個のサンプルにおける平均値を \bar{x}_2 とする。 $E(\bar{x}_1 - \bar{x}_2)$ および $V(\bar{x}_1 - \bar{x}_2)$ を求めると

$$\begin{aligned} E(\bar{x}_1 - \bar{x}_2) &= E(\bar{x}_1) - E(\bar{x}_2) \\ &= \mu_1 - \mu_2 \end{aligned} \quad (2.26)$$

$$\begin{aligned}
 V(\bar{x}_1 - \bar{x}_2) &= V(\bar{x}_1) + V(\bar{x}_2) \\
 &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}
 \end{aligned}
 \tag{2.27}$$

となるので、

$$D(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
 \tag{2.28}$$

となる。平均 $E(\bar{x}_1 - \bar{x}_2)$ と標準偏差 $D(\bar{x}_1 - \bar{x}_2)$ を用いて仮説 $H_0: \mu_1 = \mu_2$ は以下の手順により検定できる。

1. 平均値に関する仮説を設定する。 $H_0: \mu_1 = \mu_2$
2. サンプル平均 \bar{x}_1 と \bar{x}_2 と標準偏差 $D(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ をもとに $u_o = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ を計算する。
3. u_o の絶対値がある確率値に対する基準値 $crit(p)$ より大きいとき、すなわち $|u_o| \geq crit(p)$ のとき仮説 $H_0: \mu_1 = \mu_2$ は捨てられる。

(d) 母平均の差の区間推定

母平均の差の区間推定は、

$$\left| \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| \leq crit(p)
 \tag{2.29}$$

を解くことにより得られる。すなわち、

$$(\bar{x}_1 - \bar{x}_2) - crit(p) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + crit(p) \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
 \tag{2.30}$$

となる。

2.2.1.2 母標準偏差σが未知の場合

(a) 母平均の検定

母平均 μ と母標準偏差 σ が既知の場合のサンプルの平均値を \bar{x} とすると

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}
 \tag{2.31}$$

は平均0 標準偏差1 の正規分布 $N(0,1)$ に従う。

母標準偏差 σ が未知の場合、計算により得られた不偏分散の平方根 \sqrt{V} を σ の推定値 $\hat{\sigma}$ として、式(2.31)を置き換えることにより以下の統計量 t が得られる。

$$t = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\sqrt{\frac{V}{n}}} \tag{2.32}$$

この t は σ がサンプルから得られた測定値 \sqrt{V} により置き換わっているので $N(0,1)$ の正規分布には従わず、自由度 $\phi = n - 1$ の t 分布に従う。平均値の仮説検定を定式化すると以下ようになる。

1. 平均値に関する仮説を設定する。 $H_0: \mu = \mu_0$
2. サンプル平均 \bar{x} と不偏分散 $V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ をもとに $t_o = \frac{\bar{x} - \mu_0}{\sqrt{\frac{V}{n}}}$ を計算する。
3. t_o の絶対値がある確率値に対する基準値 $t(n-1, p)$ より大きいとき、すなわち $|t_o| \geq t(n-1, p)$ のとき仮説 $H_0: \mu = \mu_0$ は捨てられる。

(b) 母平均の区間推定

母平均の区間推定は式(2.33)により達成される。

$$\bar{x} - t(n-1, p) \sqrt{\frac{V}{n}} \leq \mu \leq \bar{x} + t(n-1, p) \sqrt{\frac{V}{n}} \tag{2.33}$$

問題 8

[A] マススペクトルのある標準サンプルに対するピークの m/z 値の平均値と不偏分散値は実験回数が 2-10 回において全て $\bar{x} = 205.002$ ，および $V = 0.010$ であった。それぞれの実験回数について平均値の t 検定 5% 信頼限界における区間を推定せよ。

[B] マススペクトルのある標準サンプルに対するピークの m/z 値の母分散は既知で $\sigma^2 = 0.010$ である。また、平均値は実験回数が 2-10 回において全て $\bar{x} = 205.002$ であった。それぞれの実験回数について平均値の正規分布 5% 信頼限界における区間を推定せよ。

自由度	1	2	3	4	5	6	7	8	9
t 値	12.706	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262

正規分布において

$$F(1.645) = \int_{-\infty}^{1.96} f(z) dz = 0.975$$

である。

図 5 に問題 6 の結果をプロットする。●は t 分布 5% による平均値の区間、▲は正規分布 5% による平均値の区間を表す。 t 分布あるいは正規分布を使うことは、それぞれ母分散が既知か否かによる。分散値が母分散であると仮定できれば、平均値の区間は▲印の範囲にあり高い精度で推定できる。一方、不偏分散で母分散を代表させるときには、特に実験回数が少ないときの平均値の区間は母分散が既知の時と比べて広くなるのがわかる。通常、母分散がわかることはほとんどないので t 検定を使うことが多い!

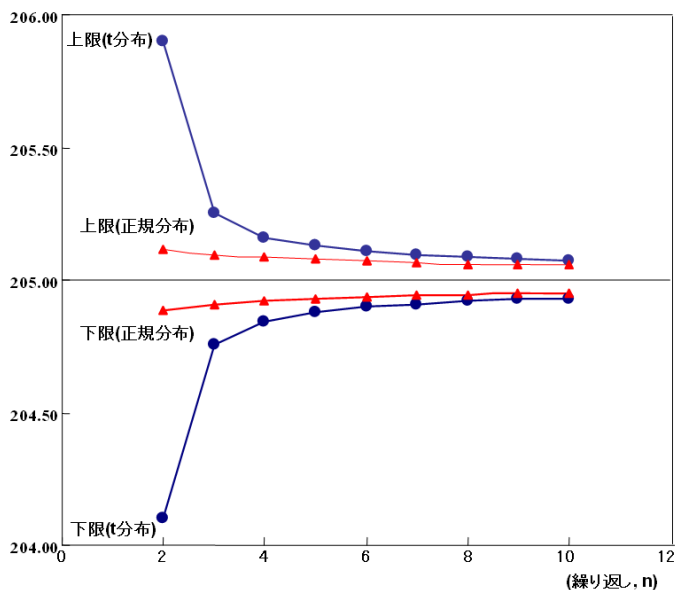


図5 問題6の区間推定 (繰り返し回数と5%信頼限界における区間推定)

(c)母平均の差の検定

(i) $\sigma_x = \sigma_y$ の場合

両サンプルの平方和 $S_x = \sum_{i=1}^{n_x} (x_i - \bar{x})^2$, $S_y = \sum_{i=1}^{n_y} (y_i - \bar{y})^2$ の自由度はそれぞれ $\phi_x = n_x - 1$, $\phi_y = n_y - 1$ であるの

で、和 $S = S_x + S_y$ の自由度は $\phi = \phi_x + \phi_y = n_x + n_y - 2$ となる。よって σ の推定値 $\hat{\sigma}$ は

$$\hat{\sigma} = \sqrt{\frac{S_x + S_y}{\phi_x + \phi_y}} \tag{2.34}$$

となる。このとき式(2.35)における t 値は、自由度 $\phi = \phi_x + \phi_y = n_x + n_y - 2$ の t 分布に従う。

$$t_0 = \frac{\bar{x} - \bar{y}}{\hat{\sigma} \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \tag{2.35}$$

そこで以下の手順で二つの母平均の差を検定できる。

1. 平均値に関する仮説を設定する。 $H_0 : \mu_x = \mu_y$
2. サンプル平均 \bar{x} と \bar{y} と標準偏差 $\hat{\sigma} = \sqrt{\frac{S_x + S_y}{\phi_x + \phi_y}}$ をもとに $t_0 = \frac{\bar{x} - \bar{y}}{\hat{\sigma} \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$ を計算する。
3. t_0 の絶対値がある確率値に対する基準値 $t(n_x + n_y - 2, p)$ より大きいとき、すなわち

$|t_0| \geq t(n_x + n_y - 2, p)$ のとき仮説 $H_0: \mu_x = \mu_y$ は捨てられる。

(ii) $\sigma_x \neq \sigma_y$ の場合

$\sigma_x \neq \sigma_y$ のとき **Welch の検定** を行う。このとき

$$t_0 = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{V_x}{n_x} + \frac{V_y}{n_y}}} \quad (2.36)$$

を使う。ここで $V_x = \frac{\sum_{i=1}^{n_x} (x_i - \bar{x})^2}{n_x - 1}$, $V_y = \frac{\sum_{i=1}^{n_y} (y_i - \bar{y})^2}{n_y - 1}$ である。また使う自由度は

$$\phi = \frac{1}{\frac{c^2}{n_x - 1} + \frac{(1-c)^2}{n_y - 1}} \quad (2.37)$$

ここで

$$c = \frac{V_x}{n_x} \left/ \left(\frac{V_x}{n_x} + \frac{V_y}{n_y} \right) \right. \quad (2.38)$$

である。

(d) 母平均の差の区間推定

(i) $\sigma_x = \sigma_y$ の場合

$$(\bar{x} - \bar{y}) - t(n_x + n_y - 2, p) \hat{\sigma} \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)} \leq \mu_x - \mu_y \leq (\bar{x} - \bar{y}) + t(n_x + n_y - 2, p) \hat{\sigma} \sqrt{\left(\frac{1}{n_x} + \frac{1}{n_y}\right)} \tag{2.39}$$

ここで、 $\hat{\sigma} = \sqrt{\frac{S_x + S_y}{n_x + n_y - 2}}$ である。

2.3 推定と検定 (まとめ)

観測データには必ず誤差が含まれるため、そのままでは自然法則や母集団の特性を正確に反映しているとは限らない。しかし、無限回の実験や調査を繰り返すことによりそのデータの分布を知ることができる。実際に得た有限個の観測データは、母集団から抽出された標本値である。いま、母集団から n 個の標本を抽出し、それらを、

$$x_1, x_2, \dots, x_n$$

とする。これらの確率変数 x_1, x_2, \dots, x_n をもとに一つの統計量、例えば平均値、を計算し y とする。

$$y = y(x_1, x_2, \dots, x_n)$$

平均値の場合、

$$y = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

となる。

母集団から n 個の標本を抽出するたびに得られる統計量 Y も確率変数であるため、 Y における分布 $P(y)$ もまた母集団により決まる。 Y の分布型は当然母集団により決まる。母集団から任意に抽出された一組のデータから算出された Y によって母集団の推定を行うのが推定(estimation)である。 $P(y)$ を確率密度関数であるとすると、

$$\int_{-\infty}^{\infty} P(y) dy = 1$$

である。図6の色つき部分の面積の合計が α であるとき、すなわち、

$$\int_{-\infty}^{a_0} P(y) dy = \frac{\alpha}{2}, \int_{b_0}^{\infty} P(y) dy = \frac{\alpha}{2}$$

のとき、データから算出された統計量 y が $y < a_0, b_0 < y$ である確率を $\Pr(y < a_0, b_0 < y)$ とすると

$$\Pr(y < a_0, b_0 < y) = \alpha \tag{2.40}$$

である。 α が十分小さいとき、このようなことは偶然には起こりにくいと考える、母集団に関する仮説を棄却することになる（平均値の検定の場合、母集団の平均値とは有意に異なるということになる）。これを仮説検定(test statistical hypothesis)という。 α を有意水準(level of significance)とよぶ。

α がたとえば 0.01 であるとすると、100 回に 1 回は $y < a_0, b_0 < y$ のいずれかの条件を満たす状況が起こることになる。すなわち仮説が正しくともこれを棄却することがあり、**第一種の誤り**(error of the first kind)と呼ぶ。一方、仮説が正しくない場合（例えば、計算された平均値は母集団の平均値とは同一とみなせない場合）に、 y が $a_0 < y < b_0$ となる場合もあり、これを**第二種の誤り**(error of the second kind; Type II error)と呼ぶ。仮説が正しくないときにこれを棄却する確率を検出力(power of test)と呼ぶ。この検出力により統計量の**第二種の過誤**の程度を得ることができる。

表 1 検定における 2 種類の誤り

		仮説検定の結果	
		H_0 を採択する。 差なし仮説 (帰無無仮説) を否定せず。 (陰性)	H_0 を棄却する。 差なし仮説を否定する。 (陽性)
事実	H_0 は真である。 (陰性)	True negative (TN) $1 - \alpha$	false positive (FP) 第一種の誤り (type I error) 有意水準 = α
	H_0 は偽である。 (陽性)	false negative (FN) 第二種の誤り (type II error) = β	true positive (TP) 検定力 = $1 - \beta$

問題 9 イヌの妊娠検査薬について統計検定をおこなった。仮説 H_0 : 「イヌが妊娠している」とした場合、

- (a) 検査薬が妊娠していると判定し、イヌが妊娠していない場合、
- (b) 検査薬が妊娠していると判定し、イヌが妊娠している場合
- (c) 検査薬が妊娠していないと判定し、イヌが妊娠していない場合
- (d) 検査薬が妊娠していないと判定し、イヌが妊娠している場合

のそれぞれは、false positive, false negative, true positive, true negative のいずれか

上述の例[式(2.40)]では

$$\Pr(y < a_0) + \Pr(b_0 < y) = \alpha \quad (2.41)$$

となるように a_0, b_0 を選んだ。

一方、

$$\Pr(y < a_1) = \alpha \quad (2.42)$$

あるいは

$$\Pr(b_1 < y) = \alpha \quad (2.43)$$

となるように、 a_1 あるいは b_1 を設定することも原理上はありえる。(2.42)あるいは(2.43)のように、片側のみを考慮する場合の検定を片側検定、また式(2.41)のように両側を考慮する場合の検定を両側検定と呼ぶ。問題7では片側検定を、問題8では区間推定を両側についておこなっている。例えば、問題8において、

$$F(1.645) = \int_{-\infty}^{1.96} f(z)dz = 0.975 \quad (2.44)$$

が示されているのは、

$$\int_0^{1.96} f(z)dz = F(1.645) - 0.5 = 0.975 - 0.5 = 0.475$$

となり、

$$\int_{-1.96}^0 f(z)dz + \int_0^{1.96} f(z)dz = 0.95$$

が得られる。このことは、式(2.41)において

$$\Pr(y < -1.96) + \Pr(1.96 < y) = 1 - 0.95 = 0.05$$

となるための a_0, b_0 が、それぞれ、-1.96, 1.96 であることを示している。そのために、式(2.44)の確率分布が区間推定に必要となる。問題7では、元旦後の測定値が元旦前の測定値に比べて大きいか否かを検定しているため片側検定となっている。両側検定と片側検定のいずれを採用するかは研究者のデータの見方に依存するため解釈したい内容を考慮して適切な選択をする必要がある。

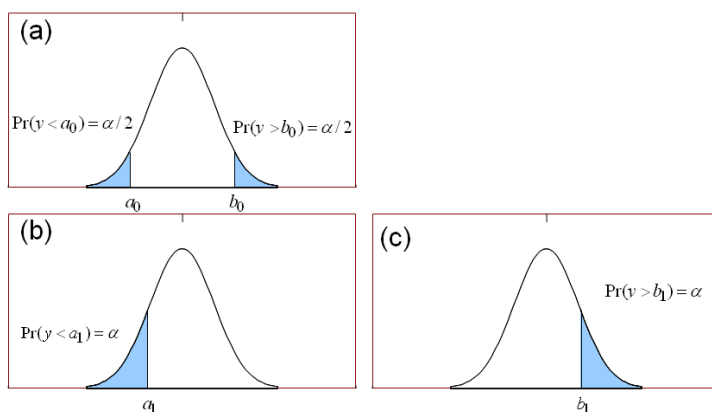


図6 (a) 両側検定における有意水準 α の設定、(b),(c) 片側検定における有意水準 α の設定

ROC 解析

いま N 個のサンプルについて、予め陰性か陽性か既知であり、ある予測システムをもとに陰性か陽性かの予測を行い以下の分割表に分類されたとする。ここで

TN=(事実が陰性で予測結果が陰性であると判定されたデータの数)

TP=(事実が陽性で予測結果が陽性であると判定されたデータの数)

FP=(事実が陰性で予測結果が陽性であると判定されたデータの数)

FN=(事実が陽性で予測結果が陰性であると判定されたデータの数)

とする。

		仮説検定の結果 (予測)	
		(陰性)	(陽性)
事実	(陰性)	TN	FP
	(陽性)	FN	TP

これらの4つの値(TN, TP, FN, FP; $TN + TP + FN + FP = N$)をもとに以下の変量が定義できる。

式による定義	説明
$TPR = \frac{TP}{TP + FN}$	True Positive Rate recall, sensitivity, positive accuracy
$FNR = \frac{FN}{TP + FN}$	False Negative Rate positive error
$TNR = \frac{TN}{TN + FP}$	True Negative Rate specificity, negative accuracy
$FPR = \frac{FP}{TN + FP}$	False Positive Rate negative error
$PPV = \frac{TP}{TP + FP}$	Positive Predicted Value Precision

$NPV = \frac{TN}{TN + FN}$	Negative Predicted Value
$FDR = \frac{FP}{FP + TP}$	False discovery rate
Avg(TPR, TNP)	Macro-average TPR と TNP の算術平均あるいは幾何平均
$BE = \frac{Precision + Recall}{2} = \frac{PPV + TPR}{2}$	Break even
$\frac{Precision \cdot Recall}{BE} = \frac{2(PPV \cdot TPR)}{(PPV + TPR)}$	F-measure

ROC 曲線とは、種々の条件で、サンプルを分割表（表 1）に分類し、横軸に FPR、縦軸に TPR をとり記述した図である。いま、10000 個のサンプルを分類器によって得た分類結果を以下の表とする。

表 2 10000 個のサンプルを分類器によりえられた分割表の例

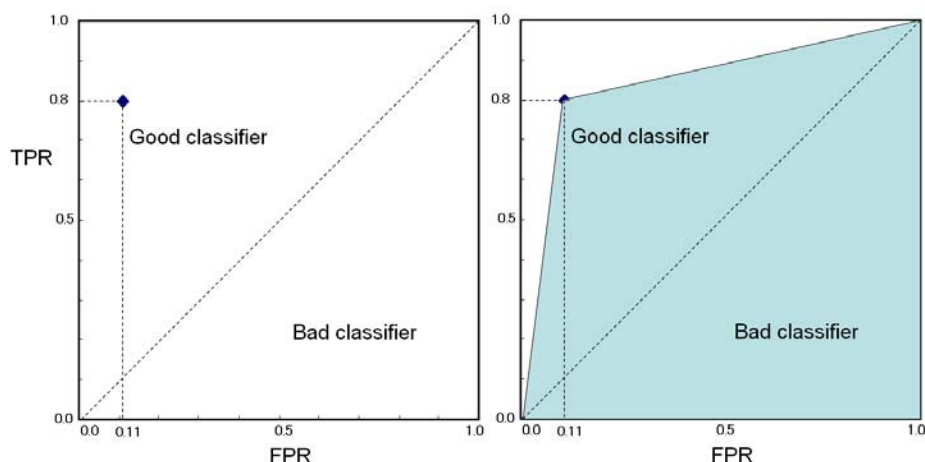
		仮説検定の結果（予測）	
		（陰性）	（陽性）
事実	（陰性）	TN=8000	FP=1000
	（陽性）	FN=200	TP =800

FPR と TPR を求めると、

$$FPR = \frac{FP}{TN + FP} = \frac{1000}{1000 + 8000} = 0.11$$

$$TPR = \frac{TP}{TP + FN} = \frac{800}{200 + 800} = 0.80$$

となる。これを ROC 図にプロットすると以下ようになる。ROC 図において、左上にあるプロットされる分類器は、性能の良いと判断される。こ ROC の面積を(Area under the ROC curve; AUC)を求めることにより分類器の性能を定量的に評価することができる（以下の図左の面積）。AUC は最大が 1 で最も高い性能と対応し、0 が最も低い性能である。



検出力

正常なコインを投げると表の出る確率が 0.5 であるとする。あるコインについて正常なコインか否かを検定することを考える。

コインを投げたときに表の出る確率を P とする。
 帰無仮説 H_0 : 正常なコイン $P = 0.5$
 対立仮説 H_1 : 正常ではないコイン $P \neq 0.5$

以下のルール 1 で適用し、正常でないコイン(イカサマコイン)を正常コインとみなす確率を検討しよう。

ルール 1: 6 回投げて n 回表が出る確率を、 $\Pr(n)$ とするとき、6 回とも表、あるいは 6 回とも裏がでるときこのコインを正常なコインでないとする。 すなわち、 $n=6$ あるいは $n=0$ のとき H_0 を棄却する。

帰無仮説 H_0 と対立仮説 H_1 の関係を概念的に表した図を以下に示す。コインが正常(帰無仮説 H_0 が成り立っているコイン)が棄却される確率(図中 α と対応)を求めよう。このコインが正常な場合、偶然 n 回とも表、あるいは n 回とも裏がでる確率は、

$$\Pr(n) = \binom{6}{n} P^n (1-P)^{6-n}$$

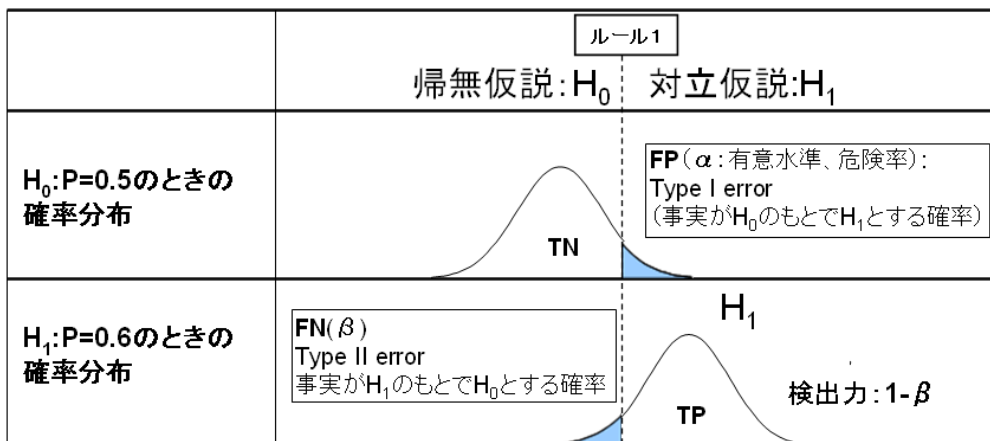
であるので、正常のコインにおいてルール 1 が偶然おこる確率は、

$$\Pr(6) + \Pr(0) = P^6 (1-P)^0 + P^0 (1-P)^6 = 2 \cdot 0.5^6 \approx 0.0312$$

となる。これは、危険率と相当するので、

$$\alpha \approx 0.0312$$

となる。



つづいて、第2種の誤りを起こす確率 β を求めよう。第2種の誤りとは、事実が H_1 のもとで H_0 とする誤りである。そこで、対立仮説 H_1 なるものを仮定しなければならない。そこで、**対立仮説 H_1** : 正常ではないコイン(イカサマコイン) の具体例として、仮に、

対立仮説 H_1 : 正常ではないコイン $P = 0.6$

とする。

$P = 0.6$ のもとで、ルール1の成り立つ確率を求めると、

$$\Pr(n = 6) = \binom{6}{6} 0.6^6 (1 - 0.6)^0 = 0.6^6$$

$$\Pr(n = 0) = \binom{6}{0} 0.6^0 (1 - 0.6)^6 = 0.4^6$$

となる。

ルール1の成り立つ確率は、

$$\Pr(n = 6) + \Pr(n = 0) = 0.6^6 + 0.4^6 = 0.05075$$

となる。

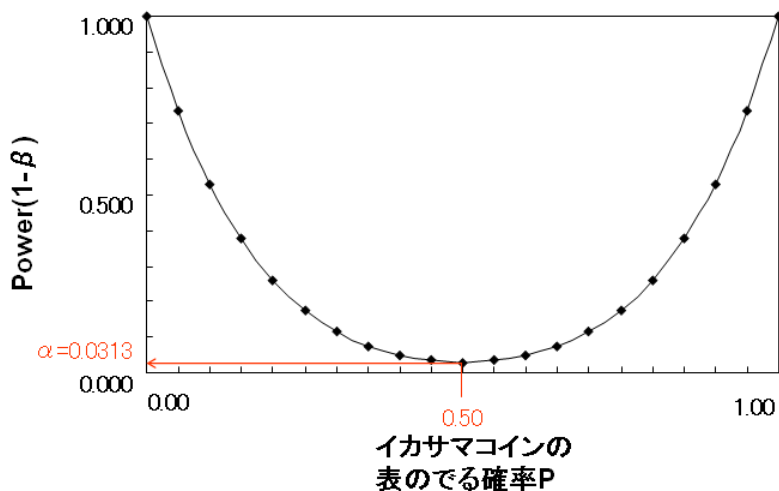
「ルール1の成り立つ」とは「正常でない」と判定されることであるので、ルール1により、イカサマコインが正しく正常でない判定される確率が 0.05075 となる。いま、 H_1 でのもとで H_0 とみなされる確率 β は、

$$\beta = 1 - 0.05075 = 0.94925$$

検出力は

$$1 - \beta = 0.05075$$

となる。すなわち、ルール1で設定した棄却方式では、「イカサマコインを見抜けない確率 β が非常に高く、検出力は非常に小さい」ということになる。ルール1におけるイカサマコインについて設定された表の出る確率 P と検出力の関係を示す。この図から、(当然ながら)ルール1の判定方式ではイカサマコインの表の出る出現確率が 0.5 から離れるに従って、検出力は増加することがわかる。図中、 $P=0.5$ に対する検出力は、正常コインが正常でない判定される確率 α である。



サンプルサイズ

2.4 多重検定

マイクロアレイ解析のように、非常に多くの遺伝子について、t 検定を行い二つの実験間における遺伝子発現量に有意な差がある遺伝子を得る場合に、多重検定の問題が絡んでくる。いま、100 個の遺伝子(N=100) について t 検定により、二つの実験間における発現量において有意な差がある遺伝子を同定することを考えよう。

N 個の遺伝子について

帰無仮説 H_0 : 二つの実験間における遺伝子発現量に有意な差がない。

対立仮説 H_1 : 二つの実験間における遺伝子発現量に有意な差がある。

ここで、有意水準 $\alpha=0.05$ とする。

として t 検定を行う。

1 個の遺伝子のみを検定する場合に**第一種の誤り**を起こす確率は、有意水準と対応して

$$P(1-\alpha, n=1) = 1 - (1-0.05) = 1-0.95 = 0.05$$

である。

2 個の遺伝子を検定する場合に少なくとも一つの遺伝子について**第一種の誤り**がおこる確率は、

$$P(1-\alpha, n=2) = 1 - (1-0.05)^2 = 1-0.9025 = 0.09750$$

同様に

$$P(1-\alpha, n=3) = 1 - (1-0.05)^3 = 1-0.8574 = 0.1426$$

...

となる。検定する遺伝子数と少なくとも一つの遺伝子について**第一種の誤り**がおこる確率の関係を図 7 に示す。

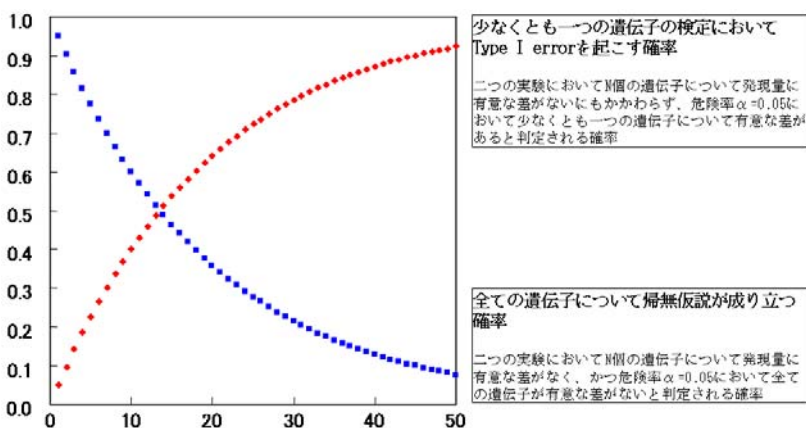


図7 検定する遺伝子数と少なくとも一つの遺伝子について**第一種の誤り**がおこる確率の関係

また、N個の遺伝子について検定を行った場合、**第一種の誤り**がおこる遺伝子数の期待値は、 $N\alpha$ により求めることができる。すなわち、20個の遺伝子について検定を行えば、**第一種の誤り**として判定される遺伝子の期待数は $20 \times 0.05 = 1$ となり、1個の遺伝子については有意な変化がないにもかかわらず偶然有意であると判定されることになる。

ゲノムサイズが小さいバクテリアゲノムにおいても 1000 個程度の遺伝子が含まれており、それぞれの遺伝子について二つの実験間で発現量に有意な差があるかを検定した場合に、

帰無仮説 H_0 : 二つの実験間において発現量の平均値に有意な差がない。

という条件で、有意な遺伝子を探すために、t 検定を行ったとすると、実際に二つの実験間の平均値に有意な差がないにもかかわらず、検定において有意な差があると判定される遺伝子の期待数は

$$1000 \times 0.05 = 50 \text{ 個}$$

となる。この場合、検定において有意な差があると検定された遺伝子数が 50 以下であったとすると、全て**第一種の誤り**の可能性も疑わなければならなくなる。そこで、有意水準 α を補正する方法として、Dunn-Sidak 補正、あるいは Bonferroni 補正などの方法がある。これらの方法では、N個の検定のどこかで**第一種の誤りをおかす確率 α** を検定する遺伝子の数Nにより補正する。この補正值 α' により理論値 (基準 t 値) を求め、検定に用いるという方法である。 α' の算出法を以下に示す。

Dunn-Sidak 補正 $\alpha' = 1 - (1 - \alpha)^{1/N}$

Bonferroni 補正 $\alpha' = \alpha / N$

図8に検定される対象の数と Dunn-Sidak 補正ならびに Bonferroni 補正における補正 α' 値の関係を示す。この図から、10 の検定対象においては、通常有意水準を $\alpha = 0.05$ をもとに計算される補正值を用いるべきであり、Dunn-Sidak 補正值ならびに Bonferroni 補正值はそれぞれ、0.005116、0.005 となる。有意水準による理論統計量を閾値に用いて検定を行うことになる。

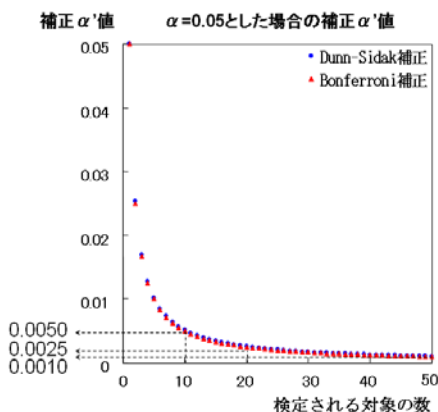


図8 検定される対象の数と Dunn-Sidak 補正ならびに Bonferroni 補正における補正 α' 値

問題 10 $N=100, 1000$ において、有意水準 $\alpha=0.05$ に対する Dunn-Sidak 補正ならびに Bonferroni 補正値を求めよ。

False Discovery Rate (FDR)

$N=20,000$ 遺伝子について、帰無仮説「二つの条件において遺伝子の発現量に有意な差がない」のもとで t 検定を行うことを考える。有意水準を 0.05 とすると、 $N \cdot p=20,000 \cdot 0.05=1000$ 個の遺伝子は、第 I 種の誤りにより対立仮説が成り立つことになる。すなわち、平均値の差の間で有意な差があると評価される。いま、二つの条件で全ての遺伝子の発現量に有意な差がない場合、統計的理論によると $20,000$ 遺伝子それぞれの t 値から得られる p 値ごとの遺伝子の出現頻度は一定になる。すなわち、 p 値に関して、例えば、 0.01 の刻み幅でグラフに表すと遺伝子の出現個数は常に $20,000 \times 0.01=200$ となる (図 9(上))。

一方、 $N=20,000$ 遺伝子における p 値の分布は、データがランダムでなく有意に変化したものが含まれるとすると、 p 値最小 (最大) の領域において、図 9(上)の頻度を上回ることになる (図 9(中)の水色棒グラフ赤線の入った領域)。ここで上回っているということは、

[統計的に有意に発現量に変化した遺伝子数]-[第 I 種の誤りにより帰無仮説が棄却された遺伝子数] >0
 となり真に発現量に変化した遺伝子がこの領域に含まれることを意味する。false discovery rate(FDR)を

$$FDR = FP / (FP + TP)$$

と定義する。ここで、実際には FP, TP は未知である。そこで、適切な基準閾値 $p(\text{crit})$ よりも小さい p 値をもつ遺伝子数 N_{exp} 、第 I 種の誤り帰無仮説が棄却される遺伝子の期待個数 N_{Rand} により、

$$FP = N_{\text{Rand}}$$

$$FP + TP = N_{\text{exp}}$$

となるので、

$$FDR = N_{\text{Rand}} / N_{\text{exp}}$$

と記述できる。

いま、 $FDR=0.1$ であるとする、 $N_{\text{Rand}}=200$ なので $200 / N_{\text{exp}}=0.1$ より $N_{\text{exp}}=2000$ となり、

$$N_{\text{exp}} - N_{\text{Rand}} = 2000 - 200 = 1800$$

すなわち、2000 個の有意に発現量が変化すると判定された遺伝子のうち 1800 個は真に変化したと推定されるということになる (図 9 下)。この FDR 値は、有意な差と判定された遺伝子の中に、どのくらい真に有意な発現量の差と判定される遺伝子が含まれるかを示す指標であるため、検定の信頼性としての指標となる。

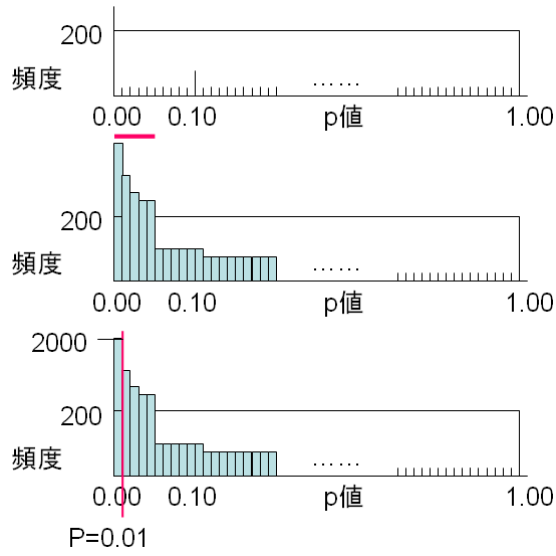


図 9 FDR の説明の図

本章では正規分布と t 分布を中心に平均値の検定ならびに推定について検討した。そのほかの統計解析を要約として表 1 に示す。

表 1 用途と統計分布

用途	確率分布	密度関数
(サンプル数 が大きいと き、例えば m>30)	正規分布 $N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
平均値の差の 検定 平均値の区間 推定	正規分布 $N(0, 1^2)$	$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$
分散の検定・ 推定 適合度の検定 独立性の検定	カイ平方分布 $N(0, 1^2)$ 型分布の n 個の互いに独立な確率変数 $x_j (j=1, \dots, n)$ の平方和 $\chi^2 = x_1^2 + \dots + x_n^2$ は自由度 $\phi = n$ のカイ平方分布を与える。	$f_\phi(\chi^2) = \frac{(\chi^2)^{\frac{\phi}{2}-1} \exp\left\{-\frac{\chi^2}{2}\right\}}{2^{\phi/2} \Gamma\left(\frac{\phi}{2}\right)}$
平均値の差の 検定 平均値の区間 推定	t-分布 $t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{V}{n}}}$ は自由度 $\phi = n - 1$ のテイ分布を与える。	$f_\phi(t) = \frac{1}{\sqrt{\phi} B\left(\frac{\phi}{2}, \frac{1}{2}\right)} \cdot \left(1 + \frac{t^2}{\phi}\right)^{-\frac{\phi+1}{2}}$
幾組かの平均 値の差の検定 (分散分析検 定)	F-分布 二つの確率変数 χ_1^2, χ_2^2 が互いに独立に自由度 ϕ_1, ϕ_2 のカイ平方分布を定めるとき、 $F = \frac{\frac{\chi_1^2}{\phi_1}}{\frac{\chi_2^2}{\phi_2}}$ は 第 1 および第 2 自由度 (ϕ_1, ϕ_2) のエフ分布に従う。	$\varphi(F) = \frac{(\phi_1)^{\frac{\phi_1}{2}} (\phi_2)^{\frac{\phi_2}{2}}}{B\left(\frac{\phi_1}{2}, \frac{\phi_2}{2}\right)} \cdot \frac{F^{\frac{\phi_1}{2}-1}}{(\phi_1 F + \phi_2)^{\frac{(\phi_1+\phi_2)}{2}}}$

3. マイクロアレイデータ解析

3.1. DNA マイクロアレイ実験

DNA アレイの基本原理解は DNA のハイブリダイゼーション(hybridization)であり、サザンブロットやノーザンブロットを大規模かつ並列で行うことで数千から数万個の遺伝子の発現量を測定する方法である。DNA アレイ実験には、GeneChip 法、スポット型アレイ法(cDNA マイクロアレイ法)、および Serial Analysis of Gene Expression (SAGE)法がある。

GeneChip 法は Affymetrix 社が開発した方法であり、各遺伝子の検出に対して最大 40 塩基(通常は 11-20 塩基程度) のオリゴヌクレオチドが使われる。各遺伝子について、他の遺伝子と比べて最も相同性の低い領域をいくつか選択し、それぞれの領域に対して 11-20 塩基からなる完全相補オリゴヌクレオチドとこの配列と 1 塩基置換が中央付近にある不完全オリゴヌクレオチドに対する蛍光強度差から発現量を測定する。

スポット型アレイ法(cDNA マイクロアレイ法)は、基準となる系(対照実験)と目的とする系(目的実験)

それぞれから mRNA を抽出し、それぞれの系において二つの蛍光色素（例えば、Cy3 と Cy5）によりラベルされた cDNA を、マイクロアレイ上にスポットした cDNA と競合ハイブリダイゼーションすることにより、それぞれの実験における 1000-10,000 種類の遺伝子の発現量の相違を検出する方法である(図 10)。Serial Analysis of Gene Expression (SAGE)法では cDNA 断片（タグ）をライゲーション反応により一本につなぎあわせ、その配列を決定する。タグ配列の出現頻度により細胞内の各々の mRNA 量を定量する方法である。

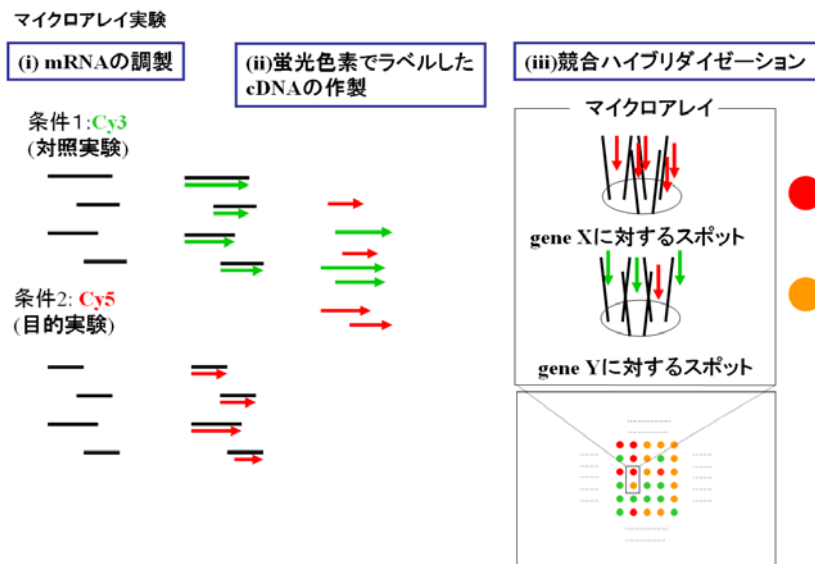


図 10 cDNA マイクロアレイ実験

3.2 シグナル評価

Cy3 と Cy5 の二種類の蛍光色素で cDNA を標識した場合、蛍光色素の蛍光強度が不安定であること、また、標識する過程における cDNA への取り込み効率が異なることから、画像処理の過程で得られたこれらの二色の蛍光強度にはランダムノイズとバイアスの二つの誤差が含まれる。シグナル強度が低いほどこれらの誤差の影響を受けやすい。これらの誤差を軽減化する方法として MA プロット(MA-plot)がある(Dudoit, et al., 2002b; Quackenbush, 2002)。それぞれのスポットに対する蛍光強度の対数比を $M_i = \log(T_i / R_i)$ 、平均対数強度(averaged logarithmic intensity)を $A_i = \{\log(T_i) + \log(R_i)\} / 2$ とする。 A_i を横軸、 M_i を縦軸にプロットすることにより平均対数強度に対する偏り誤差を評価することができる。図 11 上段は MA プロットの一例である。この図から平均対数強度が低い領域において対照実験におけるシグナル強度が低い傾向にあることがわかる。

このような偏り誤差を補正するために $M_i = 0$ の近傍において A_i と M_i の関数

$$M_i = f(A_i) + \varepsilon_i \tag{3.1}$$

を求める。ここで ε_i は i 番目のスポットに対する残差である。式(3.1)により得た関数 f をもとにそれぞれの A_i に対するベースラインを求め、これを

$$\hat{M}_i = f(A_i) \tag{3.2}$$

とする。規格化対数比(normalization log-ratio)は M_i と \hat{M}_i との差 (式(3.3)) により求められる。

$$\tilde{M}_i = M_i - \hat{M}_i \tag{3.3}$$

補正前(図 8 上段)および補正後 (図 8 下段) の MA プロットを図 8 に示す。この図から明らかなように下段において $M_i = 0$ に対して対称的な分布となっていることから、蛍光色素特異的なシグナルによる偏り誤差は軽減されたことになる。この方法では、大多数のスポットに対しては発現強度に大きな差がないことが前提となっている。また、 A_i と M_i の関数 f を求める過程で $M_i = 0$ の近傍の統計的規定法は Schadt et al.(2001)および Tseng et al.(2001)により提案されている。

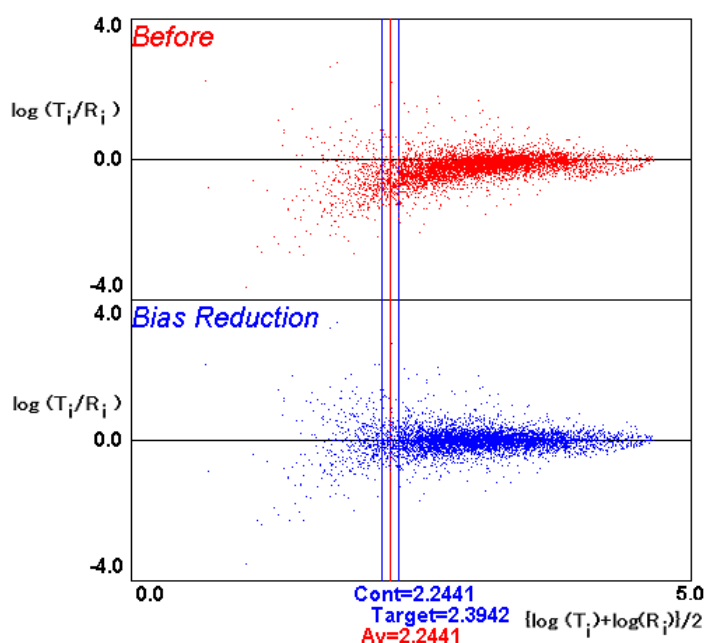


図 11 MA プロット(上段、補正前；下段、補正後)

繰り返し実験がなされている場合のデータの規格化法についての統計処理、スポットされている位置を考慮した規格化、二色の蛍光色素を例えば Cy3 を対照実験および Cy5 を目的実験に対応させた場合とその対応を逆にした場合の二通りの実験を行うことによる規格化についての方法もまた提案されている。

3.3 観測データの分布が正規分布に従うか否かの判定法

観測データの分布が正規分布に従うかどうかを判定できれば、観測データにランダム誤差以外の誤差の性質を理解し、データ解析に必要とされる統計解析法を検討できる。確率プロット法は、観測データの分布が正規分布に従うことを判定する方法であるから観測データにランダム誤差以外の誤差の性質を把握することが可能となる。

3.3.1 正規確率紙

横軸に変数 X を等間隔でとり縦軸に相対累積度数の対数値をプロットする。観測データが正規分布に従っているとすると、標本点は直線上に並ぶ。cDNA マイクロアレイにおいてある補正法 (MA 法) を適用した場合、補正を行わなかった場合の正規確率紙プロットを図 9 に示す。相対累積度数の対数値が $-0.301(\log_{10}(0.50)=-0.301)$ と対応する変数 X の値が平均値 μ と対応し、

$$P(-\infty \leq x \leq \mu + \sigma) = F(1) = 0.8413 \tag{3.4}$$

$$P(-\infty \leq x \leq \mu - \sigma) = 0.1587 \tag{3.5}$$

であることを利用すると、相対累積度数の対数値 $-0.7994(=\log_{10}(0.1587))$ および $-0.0750(=\log_{10}(0.8413))$ は $\mu + \sigma$ 、と対応するため、これらの値により変数 X の標準偏差を求めることができる。図 12 において MA 補正を行わなかった場合の個々の遺伝子発現量の対数比 (破線)、また、MA 補正を行なった場合の個々の遺伝子発現量の対数比 (実線) をプロットした。この図から、MA 補正を行うことにより、広範の対数比において正規性が保たれていることがわかる。一方、MA 補正を行わないことにより、二つの領域で異なる傾きで直線性が観察されることから、異なったランダムな現象が補正を行わないデータに対して存在すると結論付けられる。

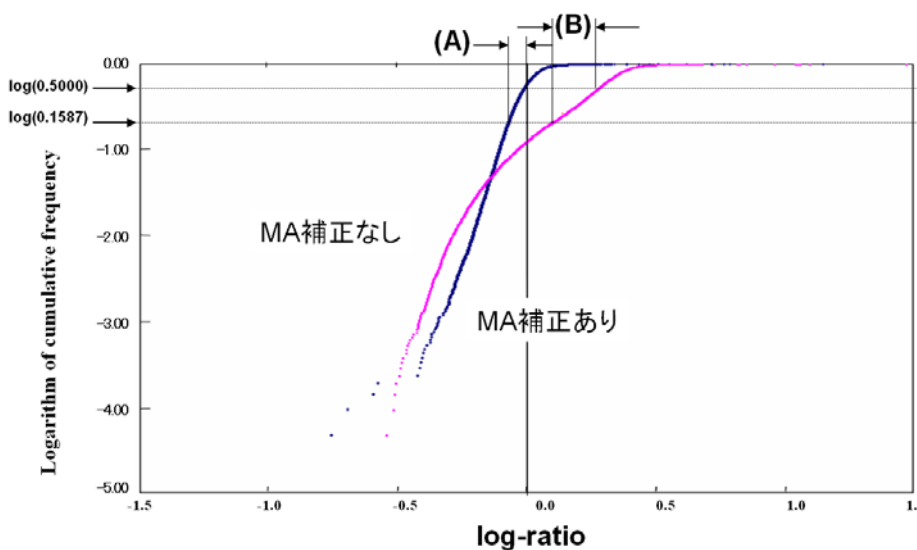


図 12 正規確率紙による MA 補正を行った場合 (実線、標準偏差は(A)と対応) と行わなかった場合 (破線、標準偏差は(B)と対応) のデータの分布 (アラビドプシスにおけるアジレント製マイクロアレイ)

3.3.2 確率プロット

一つの母集団から n 個の標本を無作為に抽出し、その母集団の累積分布関数が $F(x)$ であるという仮説を、標本の累積度数関数 $F_n(x)$ と $F(x)$ の類似性により検討する方法である。仮説が正しいとき n 個の標本と対応する点がほぼ直線に並ぶ。プロットの方法としては P-P プロットと Q-Q プロットがある。母集団から n 個の無作為標本をとり小さい順に並べる。すなわち $x_1 \leq x_2 \leq \dots \leq x_i \leq \dots \leq x_n$ とする。式(3.6)における正規分布を解いて $z_1, z_2, \dots, z_i, \dots, z_n$ を求める。

$$F(z_i) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} dz = \frac{i-\frac{1}{2}}{n} \tag{3.6}$$

P-P プロット

n 個の累積度数の対 $\left(\frac{i-\frac{1}{2}}{n}, F(x_i)\right)$ をプロットする。

Q-Q プロット

n 個の累積度数の対 (z_i, x_i) をプロットする。

Arabidopsis におけるマイクロアレイデータならびにマクロアレイデータについて、遺伝子発現量の対数比について MA 補正を行わなかった場合および行った場合の P-P プロットを例として図 13 に示す。この図から、マイクロアレイデータにおいて補正を行った場合に直線性が成り立つ領域が広いことから一つの正規分布でデータを解釈することがおおよそ可能となったと判断される。

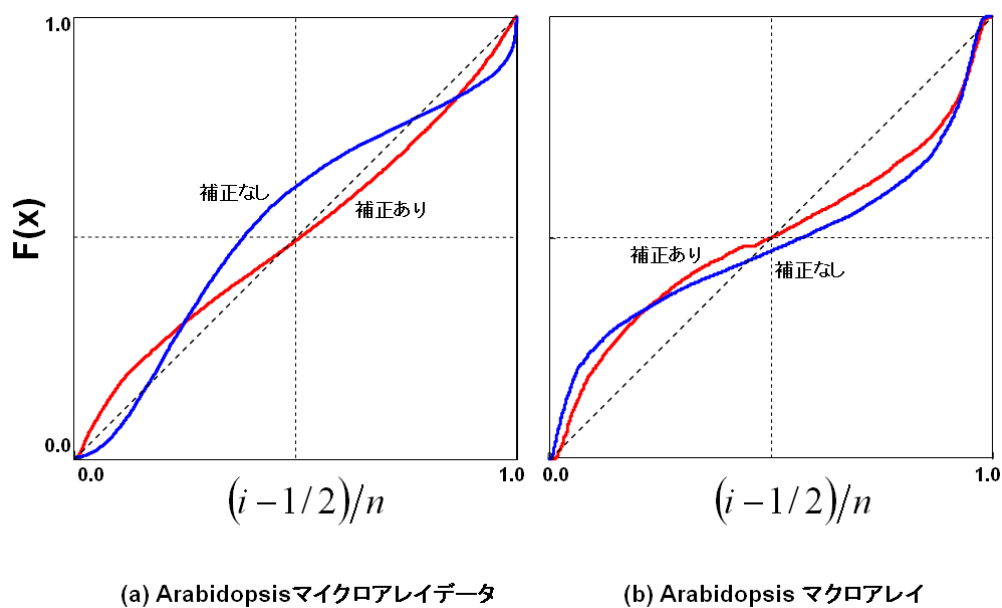


図 13 P-P プロット

3.3.3 標準正規累積分布関数 $F(x)$ の算出法

$$\begin{aligned}
 F(x) &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} dz \\
 &= \frac{1}{2} + \frac{e^{-z^2/2}}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{z^{2k+1}}{1 \cdot 3 \cdot 5 \cdot \dots \cdot (2k+1)}
 \end{aligned} \tag{3.7}$$

と級数展開してこれをプログラミングする。

$$F(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx$$

の計算プログラム例を以下に示す。

```

double lower(double z)
{
    double z2=z*z;
    double t=z*Math.exp(-0.5*z2)/Math.sqrt(2*Math.PI);
    double p=t;
    for(int i=3;i<200; i+=2)
    {
        double prev=p;
        t*=z2/i;
        p+=t;
        if(p==prev) return 0.5+p;
    }
    return (z>0) ? 1:0;
}

```

3.4 2色蛍光型マイクロアレイデータの有効スポットの検出法

2色蛍光型マイクロアレイでは i 番目の遺伝子に対して対数相対値 $\log(T_i/R_i)$ により二つの実験における発現遺伝子量の比を表す。 $\log(T_i/R_i)$ が正のとき対照実験 R に比べて目的とする実験 T で発現量が上昇したことを、また、 $\log(T_i/R_i)$ が負のとき対照実験 R に比べて目的とする実験 T で発現量が減少したことを意味する。 n 回の繰り返し実験を行ったときの i 番目の遺伝子における対数相対値を $x_{i1}, x_{i2}, \dots, x_{in}$ とするとき、これらの値が正あるいは負である有意水準を求めることが必要となる。図 14 はそれぞれ平均値の分布を示しているものとする。この図の(a)および(b)は平均値 \bar{x} が正である。一方、(c)においては \bar{x} は負である。 t 統計量を用いるとこの正および負の値がどの程度、統計的に有意であるかを確率値で表現することができる。

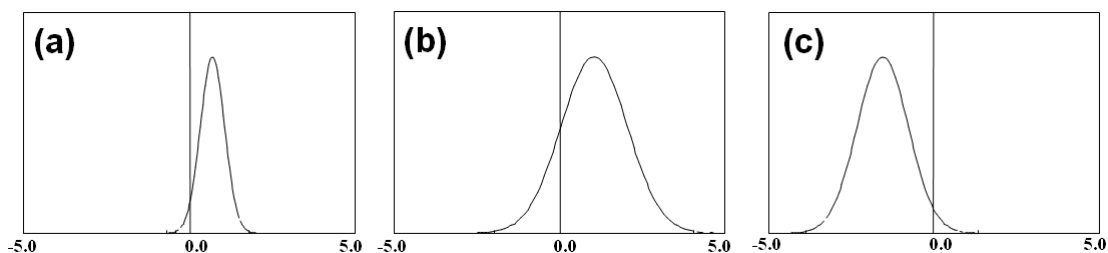


図 14 平均値の分布の例

図 14(a)と(b)は共に平均値が正であるものの(a)は(b)に比べて分散が非常に小さい。平均値と分散に注目して平均値が 0 とみなせるかという仮説を検定できれば、有意水準のもとに対数相対値 $x_{i1}, x_{i2}, \dots, x_{in}$ がどの程度再現性があるかを検討することができる。そのためには、t 分布を用いて平均値が 0 となる有意水準 p を求め、有意水準 p が非常に小さい時、平均値は正または負とみなすことができる。この方法を要約すると以下ようになる。

1. サンプル平均 \bar{x} と不偏分散 $v = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$ をもとに $t_o = \frac{\bar{x} - 0}{\sqrt{\frac{v}{n}}}$ を計算する。
2. 理論分布 $t(n-1, p)$ をもとに $t_o = t(n-1, p)$ となる p 値を求める。この p 値により再現性の程度を評価する。

2 回の cDNA マイクロアレイ測定結果の例を表 2 に示す。[1-8]は 2 回の測定で最も p 値が小さいサンプルである。一方、[9-16]は最も p 値が大きいサンプルである。[1-8]については、有意水準 $p=0.0005$ 以下であるので 1/2000 以下の確率で 0 とみなせることとなる。すなわち、[1-4]は“有意に正”の値である。一方 [5-8]は“有意に負”の値である。一方、[9-16]については $p=0.49$ 以上であるので、0 とみなせる確率はほぼ五分五分の状況にあり、二回に一回は 0 とみなせる。すなわち、正あるいは負とみなすことはできないと判断される。

表 2 2 回の cDNA マイクロアレイ測定結果

	1st	2nd	Av	SD	p
[1] At2g38270	0.32832	0.32831	0.32832	0.00001	0.00001
[2] At2g34620	0.35022	0.35028	0.35025	0.00003	0.00003
[3] At2g18710	0.26919	0.26927	0.26923	0.00004	0.00004
[4] At2g22500	0.42060	0.42043	0.42052	0.00008	0.00006
[5] At4g37870	-0.68499	-0.68533	-0.68516	0.00017	0.00008
[6] At1g69080	-1.29664	-1.29767	-1.29716	0.00052	0.00013
[7] At3g05290	-0.44634	-0.44598	-0.44616	0.00018	0.00013
[8] At2g40940	-0.21412	-0.21395	-0.21404	0.00009	0.00013

```

---
[9] At5g13740    0.03155    -0.03124    0.00015    0.03140    0.49843
[10] At4g32620   -0.09012    0.08925    -0.00043    0.08968    0.49846
[11] At4g09510    0.13364    -0.13245    0.00060    0.13304    0.49857
[12] At3g45620   -0.12305    0.12198    -0.00053    0.12252    0.49861
[13] At4g01560   -0.19832    0.19982    0.00075    0.19907    0.49880
[14] At5g35570   -0.07307    0.07351    0.00022    0.07329    0.49905
[15] At2g47710   -0.05806    0.05785    -0.00011    0.05795    0.49940
[16] At2g27450   -0.06773    0.06796    0.00011    0.06785    0.49947
    
```

2.2.3 t 値から p 値の算出

$\int_{-\infty}^x f_{\phi}(t)dt$ の計算アルゴリズムを以下に示す。

```

double lowerT(int df, double t)
{
    double c2=df/(df + t*t), s=Math.sqrt(1-c2);
    if(t<0)s=-s;
    double p=0;
    for(int i=df%2+2; i<=df; i+=2)
    {
        p+=s;
        s*=(i-1)*c2/i;
    }
    if((df&1)!=0)
        return 0.5+(p*Math.sqrt(c2)+Math.atan(t/Math.sqrt(df)))/Math.PI;
    else
        return (1+p)/2;
}
    
```

4.多変量解析

N 個の遺伝子について M 個の実験を行ったとすると得られるデータは、式 4.1 で示される N x M 行列により表現することができる。

$$\mathbf{X} = \begin{pmatrix}
 x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1M} \\
 x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2M} \\
 \dots & \dots & \dots & \dots & \dots & \dots \\
 x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{iM} \\
 \dots & \dots & \dots & \dots & \dots & \dots \\
 x_{N1} & x_{N2} & \dots & x_{Nj} & \dots & x_{NM}
 \end{pmatrix} \tag{4.1}$$

そこで、i 番目と i 番目の遺伝子について M 個の実験における発現データは式(4.2)および(4.3)に示すベクトルで、また、一方、j 番目と j 番目の実験における N 個の遺伝子の発現データは式(4.4)および(4.5)で表すことができる。

$$\mathbf{x}_i = (x_{i1} \ x_{i2} \ \dots \ x_{ij} \ \dots \ x_{iM}) \tag{4.2}$$

$$\mathbf{x}_j = (x_{j1} \ x_{j2} \ \dots \ x_{ij} \ \dots \ x_{jM}) \tag{4.3}$$

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \dots \\ x_{ij} \\ \dots \\ x_{Nj} \end{pmatrix} \tag{4.4}$$

$$\mathbf{x}_{j'} = \begin{pmatrix} x_{1j'} \\ x_{2j'} \\ \dots \\ x_{ij'} \\ \dots \\ x_{Nj'} \end{pmatrix} \tag{4.5}$$

いま、遺伝子を対象(object)、実験を变量(variable)とする場合に式(4.2)と(4.3)を比較することにより i 番目と i' 番目遺伝子 (対象) を発現データ (多変量ベクトル) の類似性がわかる。また、式(4.4)と(4.5)を比較することにより j 番目と j' 番目の実験 (対象) の類似性がわかる。すべてのデータに対してこのような比較を行い全ての遺伝子あるいは実験間 の関係を明らかにすれば、遺伝子間の発現制御の関係あるいは実験条件の発現遺伝子パターンからの共通性が多少なりとも解明されると期待できる。このように複数のベクトル間を比較し体系的に整理する方法が多変量解析法である。図 15 に示すように解析の流れは(1) データの前処理、(2)特徴抽出、(3)教師なし学習、(4)教師あり学習の4つの過程よりなる。

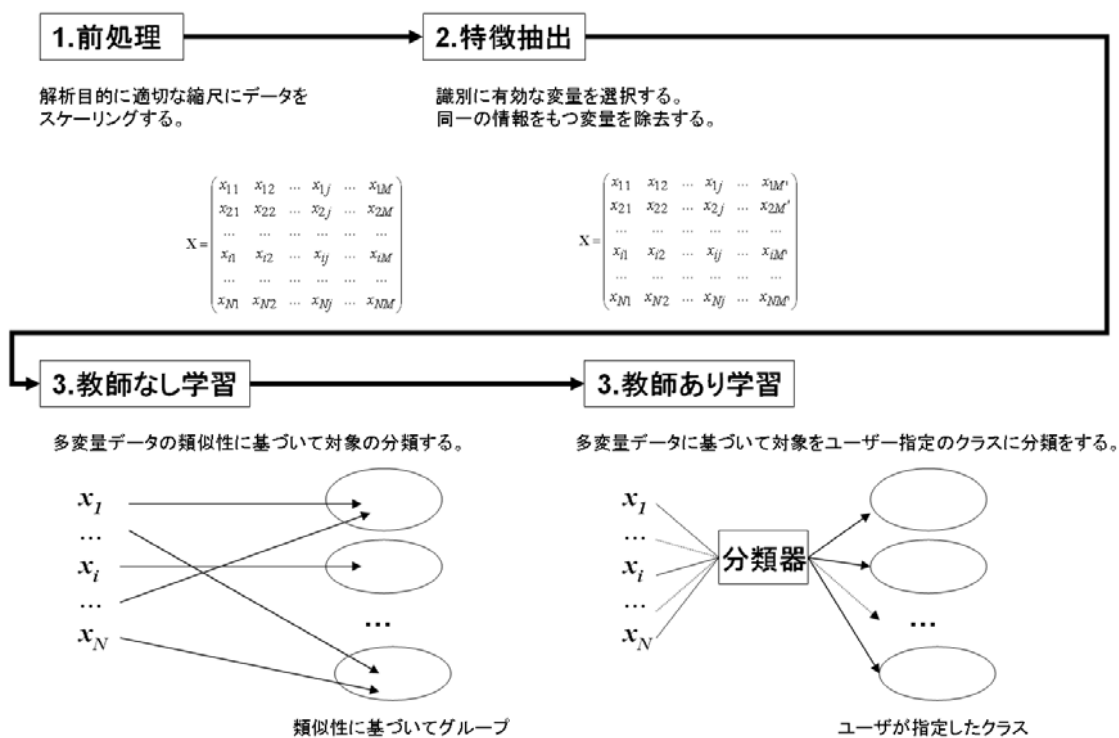


図 15 多変量解析における解析の流れ図

4.1 データの前処理

変量ごとに変動範囲を同等とすることによって、特定の変量のみが多変量解析により得られる結果に影響を及ぼさないように、あらかじめ変量の縮尺を定義するプロセスを前処理という。いま N 個の対象について j 番目の変量の値を x_{ij} とする。全ての変量に対して統計的バラツキを一定にしたいときには、式(4.6)に従って平均 0 分散 1 にスケーリングをする。

$$x'_{ij} = (x_{ij} - \bar{x}_j) / \sqrt{s_j} \tag{4.6}$$

ここで、 $\bar{x}_j = \sum_{i=1}^N x_{ij} / N$ 、 $s_j = \sqrt{\sum_{i=1}^N (x_{ij} - \bar{x}_j)^2 / (N-1)}$ である。

一方、複数の対象について相対値表現において比較すべき場合には式(3.7)により全体の変量の二乗和が 1 となるように規格化する場合が多い。

$$x'_{ij} = x_{ij} / \sqrt{\sum_{i=1}^M x_{ij}^2} \tag{4.7}$$

4.2 特徴抽出

特徴抽出の目的は、データ解析に必要なでない変量を除去することにある。データ解析に不必要な変量が、教師なしおよび教師あり学習において悪い影響を及ぼすことがあるので、このような変量を除去することが必要となる。変量を減らす方法には、(i)同一の情報を有する変量をまとめるなどして変量の数を減らす場合、あるいは、(ii)解析目標にグループ間の識別を考慮するときには、識別に有効な変量を選択することにより変量の数を減らす場合の主に二つの考え方がある。前者(i)について、二つの変量の相関が高いものについては同一の情報であるとみなし合成変量にまとめるなどして変量の数を減らす。一方、(ii)の場合のグループ間の識別に有効な変量のみを解析に用いることにより変量を減らすことができる。この識別に関して有効な変量を決める方法には、前述の t 統計量などが用いられることが多い。

ピアソンの積率相関係数

実験 (変量) 間を比較する場合のピアソンの積率相関係数 (単に相関係数ということが多い) を式(4.8)、また、遺伝子 (対象) 間を比較する場合のピアソンの積率相関係数を式(3.9)に示す。

$$r_{jj'} = \frac{\sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{i'j} - \bar{x}_{j'})}{\sqrt{\sum_{i=1}^N (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^N (x_{i'j} - \bar{x}_{j'})^2}} \tag{4.8}$$

$$r_{ii'} = \frac{\sum_{j=1}^M (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_{j=1}^M (x_{ij} - \bar{x}_i)^2 \sum_{j=1}^M (x_{i'j} - \bar{x}_{i'})^2}} \tag{4.9}$$

ピアソンの相関係数の範囲は $-1 \leq r_{jj'} \leq 1$, $-1 \leq r_{ii'} \leq 1$ であり、負の場合を負の相関、正の場合を正の相関という。

コサイン係数

ピアソン相関係数は平均値を基準とした相関であった。ピアソン相関係数における平均値基準ではなく、二つの実験あるいは遺伝子間の変量について内積を規格化した相関をコサイン係数(式(4.10-11))という。

実験 (変量) 間については

$$\cos_{jj'} = \frac{\sum_{i=1}^N x_{ij}x_{i'j}}{\sqrt{\sum_{i=1}^N x_{ij}^2 \sum_{i=1}^N x_{i'j}^2}} \tag{4.10}$$

遺伝子 (対象) 間については

$$\cos_{ii'} = \frac{\sum_{j=1}^M x_{ij}x_{i'j}}{\sqrt{\sum_{j=1}^M x_{ij}^2 \sum_{j=1}^M x_{i'j}^2}} \tag{4.11}$$

により表現される。

問題 11 異なった培養条件 1-4 について二色蛍光法マイクロアレイ測定をおこない変異株と野生株の間に以下の対数比を得た。遺伝子間のピアソン相関係数およびコサイン係数を求めよ。

遺伝子 ID	実験 1	実験 2	実験 3	実験 4
g1	0.30	0.10	-0.10	-0.30
g2	0.60	0.20	-0.20	-0.60
g3	-0.30	-0.50	-0.70	-0.90

遺伝子間のピアソン相関係数ならびにコサイン係数を以下に示す。二色蛍光法マイクロアレイ測定の場合正のとき、野生株に比べて発現上昇がみられ、負のとき発現減少がみられたことになる。g1 遺伝子と g2 遺伝子は実験 1-4 における発現は上昇、上昇、減少、減少であるから、強度に違いのみが問題となることが判る。一方、g3 遺伝子における発現は、減少、減少、減少、減少となっており、g1 と g2 とは明らかに異なった発現プロファイルとなっている。この場合でもピアソン相関係数は 1.00 となる。一方、コサイン相関係数では、g1 と g3 あるいは g2 と g3 の間で相関が低くなっている。このように、二色蛍光法マイクロアレイ測定のように正、負により意味が異なるときには、ピアソン相関係数よりもコサイン係数を計算する方が適切である場合が多い。

	ピアソン相関係数			コサイン係数		
	g1	g2	g3	g1	g2	g3
g1	1.0	1.00	1.00	1.00	1.00	0.35
g2		1.00	1.00		1.00	0.35
g3			1.00			1.00

4.3 教師なし学習(unsupervised learning)および教師あり学習(supervised learning)

教師なし学習は多変量データの類似性に基づいて対象を分類することにある。このことにより、実験間(変量)の類似性あるいは遺伝子(対象)間の類似性を検討することができる。主な教師なし学習法と視覚化アルゴリズムを表3に示す。一方、いくつかの因子を変量 x_1, \dots, x_M 、説明したい性質、活性あるいは量を変量 y としたときの定量モデルは

$$y = f(x_1, x_2, \dots, x_M)$$

により表すことができる。ここで、変量 x_1, \dots, x_M を記述子(説明変数)、変量 y を目的変数という。いま、 N 個のサンプルについて変量 x_1, \dots, x_M と y が観測されたとすると、これらをマトリックスにより、

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix}$$

により表すことができる。教師あり学習とは、説明変数 x_1, \dots, x_M の関数により目的変数 y を推定するための関数を求める方法であり、種々の方法が提案されている。古典的なものとしては線形識別関数法、ベイズ法などが挙げられる。また、種々の人工ニューラルネットアルゴリズムも提案されている(表4)。

教師あり学習により定量モデルを構築し、外的基準により対象を識別できるか否かを検討する場合においても教師なし学習により多変量空間での対象の分布を把握することにより識別の原理を考察するために必要となることが多い。ここでは、代表的な教師なし学習法として階層的クラスタリング法(Hierarchical clustering)、主成分分析(Principal Component Analysis; PCA)、自己組織化法(Self-organizing mapping; SOM)、k平均法について説明する。

表3 教師なし学習

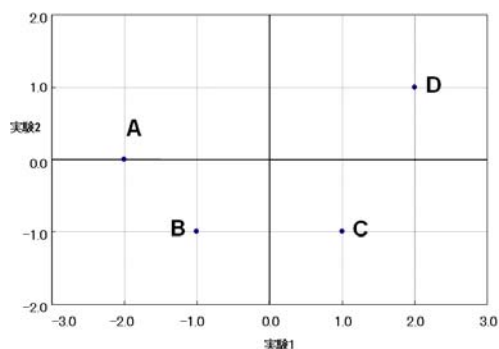
学習法	アルゴリズム
Hierarchical clustering (a) Bottom-up methods (Single linkage, Complete linkage, Mean linkage, Centroid linkage) (b) Top-down methods (Tree structured vector quantization, Macnaughton-Smith algorithm)	対象の多変量空間における分布をデンドログラムに表して理解する方法
K-means method	予めクラス数を設定し、クラスごとの平均ベクトルを対象により最適化した後に、平均ベクトルとの類似性により対象を分類する方法である。
Principal component analysis	多変量からなる多変量空間を少数の線形合成変量に減らすことにより対象の分布を視覚化する。
Self-organizing map	非線形的に対象を分類しその分布を視覚化する。
Multidimensional scaling technique	距離尺度に基づいたマッピングにより対象の分布を視覚化する。

表 4 教師あり学習

k-Nearest neighbor classifiers Discriminant function based on linear discriminant Bayes rule, Maximum likelihood discriminant rule, Fisher linear discriminant analysis, Logistic discrimination, Penalized discriminant analysis Support vector machines Calibration Multiple Linear Regression (MLR) Partial Least Squares(PLS)

4.4 クラスター分析

問題 12



二つの異なる条件でマイクロアレイ実験をおこなったところ4つの遺伝子 A,B,C,D に対して図 10 のような散布図を得た。お互いのユークリッド距離が小さいものから線をつなぎ、4つの遺伝子全てが3本の線でつながったところで終了せよ。M 個の変量における遺伝子 s と t の

ユークリッド距離は $d(\mathbf{x}_s, \mathbf{x}_t) = \sqrt{\sum_{j=1}^M (x_{sj} - x_{tj})^2}$ である。

図 12 二つの実験における二次元プロットの例

問題 12 はクラスター分析のアルゴリズムを基礎である。まずはじめに、距離行列を求めると

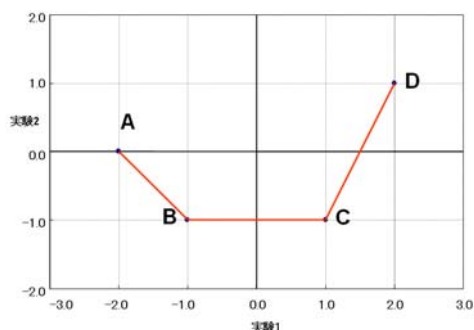
	A	B	C	D
A		1.41	3.16	4.12
B			2.00	3.61
C				2.24

となる。距離の近い順に遺伝子 ABCD をつないでいくと、

A-B $d(A,B)=1.41$

B-C $d(B,C)=2.00$

C-D $d(C,D)=2.24$



となり、以下の図が得られる。

この図はクラスター分析の最小距離法と対応する。

最小距離法

- (i) グループ間の代表距離をそれぞれのグループに属する要素間の最小距離と定義する。
- (ii) はじめに、全ての遺伝子は個別のグループに帰属するものとする。
- (iii) グループ間の距離が小さい順にグループを融合し、一つのグループになったら終了する。

まずはじめに、それぞれのグループを $G_A=\{A\}$ 、 $G_B=\{B\}$ 、 $G_C=\{C\}$ 、 $G_D=\{D\}$ とする。

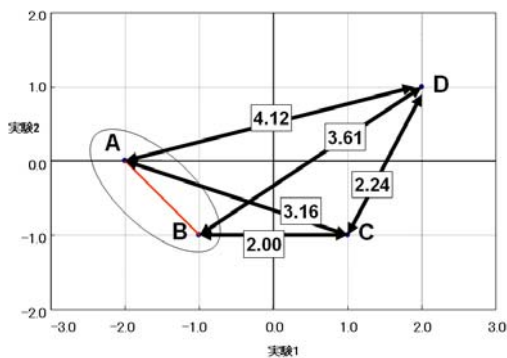
A と B が最小距離にあるので融合されて新たなグループ $G_{AB}=\{A,B\}$ をつくる。

$$d(G_A, G_B) = 1.41$$

このことによりグループは

$$G_{AB}=\{A,B\}, G_C=\{C\}, G_D=\{D\}$$

となる。これらの3つの距離関係を図に表すと以下ようになる。



いま、グループ間の距離はそれぞれの要素間の最小の距離により代表されるので、

$$d(G_{AB}, G_C) = \min\{d(G_A, G_C), d(G_B, G_C)\} = 2.00$$

$$d(G_{AB}, G_D) = \min\{d(G_A, G_D), d(G_B, G_D)\} = 3.61$$

となり、距離行列は以下ようになる。

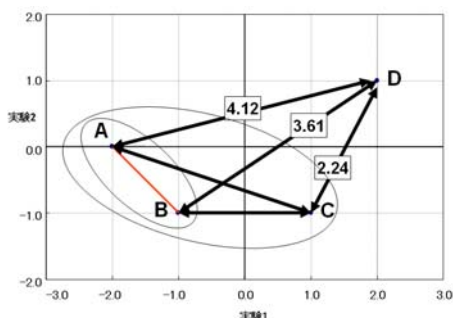
	AB	C	D
AB		2.00	3.61
C			2.24

この中で最小の距離は $G_{AB}-G_C$ であるのでこれらを新たなグループにまとめあげと、

$$G_{\{ABC\}}=\{A,B,C\}, G_D=\{D\}$$

が残る。これらの二つのグループの距離は、下図より、

$$d(G_{\{AB\}C}, G_D) = \min\{d(G_A, G_D), d(G_B, G_D), d(G_C, G_D)\} = 2.24$$



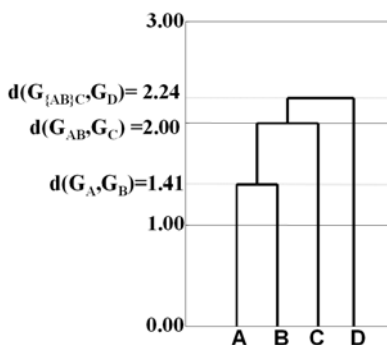
要素がグループ化される様子を整理すると、

$$d(G_A, G_B) = 1.41$$

$$d(G_{AB}, G_C) = 2.00$$

$$d(G_{\{AB\}C}, G_D) = 2.24$$

となりこれをデンドログラムで表すと以下ようになる。



グループ間の代表距離により主に以下の4種の方法が提案されている。

階層的クラスター分析法	グループ間距離
Nearest Neighbor(最近隣法)	それぞれのグループにおける要素間の距離の最小値
Furthest Neighbor(最遠隣法)	それぞれのグループにおける要素間の距離の最大値
Centroid(重心間距離法)	それぞれのグループにおける要素から求めた重心間の距離
Average(平均距離法)	それぞれのグループにおける要素間の距離の平均値

問題 13 問題 12 のデータを用いて重心距離法によりクラスター分析をせよ。グループ間の距離にはユークリッド距離を用いることとする。

根性をだして計算すると、**図 17 右側**のデンドログラムを得るはずである。この図からも明らかに使う代表距離により、デンドログラムの構造が異なるので、目的にあった解析法を選択する必要がある。クラスター分析の説明には類似性の尺度としてユークリッド距離を用いたが、その他にも**表 5**に示すような多数の類似性尺度がある。

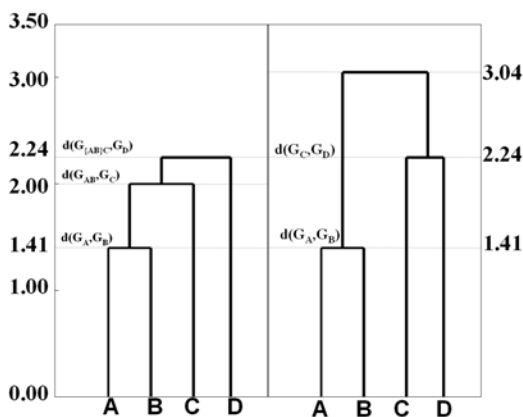
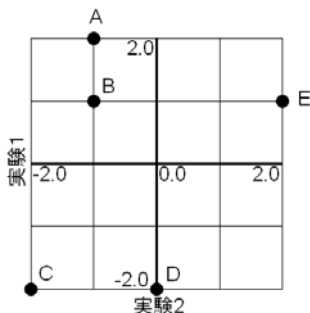


図 17 最近隣法 (左) と重心法 (右) の比較

重心間距離法はグループ間を融合した後の新たなグループの座標を求めることができる。すなわち、重心間距離法は、グループの代表となる座標をデータの解釈に含めることができる利点を有している。

問題 14 二つの異なる条件でマイクロアレイ実験をおこなったところ 5つの遺伝子 A,B,C,D,E に対して左



図のような散布図を得た。最小距離法によりデンドログラムを作成

せよ。なお距離にはユークリッド距離は $d(\mathbf{x}_s, \mathbf{x}_t) = \sqrt{\sum_{j=1}^M (x_{sj} - x_{tj})^2}$

を用いよ。

表 5 類似性の尺度

Minkowski metric

$$d_{Mk}(\mathbf{x}_s, \mathbf{x}_t) = \left\{ \sum_{j=1}^M w_j |x_{sj} - x_{tj}|^\lambda \right\}^{\frac{1}{\lambda}} \quad (\lambda \geq 1)$$

$w_j = 1$ と設定した場合を非標準化距離、 $w_j = 1/(s_{jj}^2)$ と設定したとき標準化距離という。 $\lambda = 1$ のときマンハッタン距離 (Manhattan metric)、 $\lambda = 2$ のときユークリッド距離 (Euclidean metric) という。

Mahalanobis metric

$$d_{Ml}(\mathbf{x}_s, \mathbf{x}_t) = \left\{ (\mathbf{x}_s - \mathbf{x}_t) \mathbf{S}^{-1} (\mathbf{x}_s - \mathbf{x}_t)^T \right\}^{\frac{1}{2}}$$

$$= \sqrt{(x_{s1} - x_{t1}, x_{s2} - x_{t2}, \dots, x_{sM} - x_{tM}) \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1M} \\ s_{21} & s_{22} & \dots & s_{2M} \\ \dots & \dots & \dots & \dots \\ s_{M1} & s_{M2} & \dots & s_{MM} \end{pmatrix}^{-1} \begin{pmatrix} x_{s1} - x_{t1} \\ x_{s2} - x_{t2} \\ \dots \\ x_{sM} - x_{tM} \end{pmatrix}}$$

(s_{ij} は変量 i と j の共分散で表す。)

Canberra metric

$$d_C(\mathbf{x}_s, \mathbf{x}_t) = \sum_{j=1}^M \frac{|x_{sj} - x_{tj}|}{|x_{sj} + x_{tj}|}$$

1-correlation

$$d_{\mathbf{x}_s, \mathbf{x}_t} = 1 - r(\mathbf{x}_s, \mathbf{x}_t) = 1 - \frac{\sum_{j=1}^M (x_{sj} - \bar{x}_s)(x_{tj} - \bar{x}_t)}{\sqrt{\sum_{j=1}^M (x_{sj} - \bar{x}_s)^2} \sqrt{\sum_{j=1}^M (x_{tj} - \bar{x}_t)^2}} \quad \left(\bar{x}_s = \frac{\sum_{i=1}^N x_{sj}}{N} \right)$$

1-cosθ

$$1 - \cos_{\mathbf{x}_s, \mathbf{x}_t} = 1 - \frac{\sum_{j=1}^M x_{sj} x_{tj}}{\sqrt{\sum_{j=1}^M x_{sj}^2} \sqrt{\sum_{j=1}^M x_{tj}^2}}$$

4.5 主成分分析法

4.5.1 主成分分析によるデータ解析で用いるパラメータ

M 個の変量からなる多変量データを式(4.12)で示される合成変量 Z_k ($k=1,2,\dots,M$)により表す。このとき、第 1 主成分を M 次元空間における最大分散を有する合成変量 Z_1 と定義する、つづいて、第 1 主成分と直交し、第 1 主成分の次に大きい分散をもつ軸を第 2 主成分と定義する。このように以下順次、主成分を決めていく方法が主成分分析である。変量間に従属な関係があるとき、M 個より少ない次元でオリジナルデータの分布を説明することが可能になる。また、主成分第 1 軸から順番にいくつかの少ない次元での対象のデータをみることによりオリジナルの多変量空間におけるデータの分布を理解する方法である。

$$\begin{aligned} Z_1 &= a_{11} X_1 + \dots + a_{1j} X_j + \dots + a_{1M} X_M \\ &\dots \\ Z_j &= a_{j1} X_1 + \dots + a_{jj} X_j + \dots + a_{jM} X_M \\ &\dots \\ Z_M &= a_{M1} X_1 + \dots + a_{Mj} X_j + \dots + a_{MM} X_M \end{aligned} \tag{4.12}$$

主成分分析では主成分スコア、寄与率、因子負荷量の三つのパラメータを用いて、多変量空間の対象の分布を解析する。

寄与率

オリジナルの多変量データの全分散に対する k 番目の主成分の分散を **寄与率**(4.13)と呼ぶ。寄与率が高いほどオリジナルデータの分布が反映される。

$$\%Var[Z_k] = \frac{V[Z_k]}{\sum_{u=1}^M V[X_u]} \cdot 100 \tag{4.13}$$

因子負荷量

第 j 番目の変量と第 k 主成分における相関係数を $r(X_j, Z_k)$ で表し、**因子負荷量(Factor Loadings)**とよ

ぶ。

4.5.2 主成分軸の算出法

第1主成分係数 $(a_{11}, \dots, a_{1j}, \dots, a_{1M})$ の算出法を以下に示す。

第1主成分の分散を $V(a_{11}, a_{12}, \dots, a_{1M})$ とする。

$$\begin{aligned} V(a_{11}, a_{12}, \dots, a_{1M}) &= \frac{1}{N-1} \sum_{i=1}^N (z_{i1} - \bar{z}_1)^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left\{ (a_{11}x_{i1} + \dots + a_{1M}x_{iM}) - (a_{11}\bar{x}_1 + \dots + a_{1M}\bar{x}_M) \right\}^2 \\ &= \frac{1}{N-1} \sum_{i=1}^N \left\{ a_{11}(x_{i1} - \bar{x}_1) + \dots + a_{1M}(x_{iM} - \bar{x}_M) \right\}^2 \end{aligned} \quad (4.14)$$

ここで

$$\sum_{j=1}^M a_{1j}^2 = 1 \quad (4.15)$$

と規格化する。

いま、式(4.15)の条件化で式(4.14)を最大化する。関数 $G(a_{11}, a_{12}, \dots, a_{1M}, \lambda)$ を分散と制約条件により式(4.16)により定義し、制約付き最大化問題であるのでラグランジュの未定乗数法により式(4.14)を最大化することができる。

$$G(a_{11}, a_{12}, \dots, a_{1M}, \lambda) = V(a_{11}, \dots, a_{1M}) - \lambda \left(\sum_{j=1}^M a_{1j}^2 - 1 \right) \quad (4.16)$$

式(3.16)をもとに $j=1, 2, \dots, M$ について

$$\frac{\partial G(a_{11}, a_{12}, \dots, a_{1M}, \lambda)}{\partial a_{1j}} = 0 \quad (4.17)$$

を求める。

実際に解くと以下ようになる。

$$\frac{\partial G(a_{11}, a_{12}, \dots, a_{1M}, \lambda)}{\partial a_{1j}} = \frac{1}{N-1} \sum_{i=1}^N \left[2 \left\{ a_{11}(x_{i1} - \bar{x}_1) + a_{12}(x_{i2} - \bar{x}_2) + \dots + a_{1j}(x_{ij} - \bar{x}_j) + \dots + a_{1M}(x_{iM} - \bar{x}_M) \right\} (x_{ij} - \bar{x}_j) \right] - 2\lambda a_{1j} = 0$$

$j=1, 2, \dots, M$

これら M 個の式を整理すると

$$\begin{pmatrix} \text{cov}[X_1, X_1] & \dots & \text{cov}[X_1, X_j] & \dots & \text{cov}[X_1, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_j, X_1] & \dots & \text{cov}[X_j, X_j] & \dots & \text{cov}[X_j, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_M, X_1] & \dots & \text{cov}[X_M, X_j] & \dots & \text{cov}[X_M, X_M] \end{pmatrix} \begin{pmatrix} a_{11} \\ \dots \\ a_{1j} \\ \dots \\ a_{1M} \end{pmatrix} = \lambda \begin{pmatrix} a_{11} \\ \dots \\ a_{1j} \\ \dots \\ a_{1M} \end{pmatrix} \quad (4.18)$$

となる。ここで

$$\text{cov}[X_s, X_t] = \frac{1}{N-1} \sum_{i=1}^N (x_{is} - \bar{x}_s)(x_{it} - \bar{x}_t) \quad (4.19)$$

$$\mathbf{C} = \begin{pmatrix} \text{cov}[X_1, X_1] & \dots & \text{cov}[X_1, X_j] & \dots & \text{cov}[X_1, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_j, X_1] & \dots & \text{cov}[X_j, X_j] & \dots & \text{cov}[X_j, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_M, X_1] & \dots & \text{cov}[X_M, X_j] & \dots & \text{cov}[X_M, X_M] \end{pmatrix} \quad (4.20)$$

とおくと、

$$\mathbf{C} \cdot \mathbf{a}_1 = \lambda \cdot \mathbf{a}_1 \quad (4.21)$$

すなわち

$$\lambda = \mathbf{a}_1^T \cdot \mathbf{C} \cdot \mathbf{a}_1 \quad (4.22)$$

により、固有値 λ が求まり、規格化条件(式(4.15))とあわせて、第1主成分係数 \mathbf{a}_1 が求まる。

第2主成分係数 \mathbf{a}_2 の求め方

第1主成分と第2主成分は直交しているため、

$$\mathbf{a}_1^T \cdot \mathbf{a}_2 = 0 \quad (4.23)$$

また、第1主成分係数と同様に、第2主成分は規格化する。すなわち、

$$\sum_{j=1}^M a_{2j}^2 = 1 \quad (4.24)$$

となる。これら二つの条件のもとで分散を最大にすることは

$$G(a_{21}, a_{22}, \dots, a_{2M}, \lambda, \beta) = \mathbf{a}_2^T \cdot \mathbf{C} \cdot \mathbf{a}_2 - \lambda(\mathbf{a}_2^T \cdot \mathbf{a}_2 - 1) - \beta \mathbf{a}_1^T \cdot \mathbf{a}_2 \quad (4.25)$$

について $j=1, 2, \dots, M$ ごとに

$$\frac{\partial G(a_{21}, a_{22}, \dots, a_{2M}, \lambda, \beta)}{\partial a_{2j}} = 0 \quad (4.26)$$

実際に計算すると

$$2 \cdot \mathbf{C} \cdot \mathbf{a}_2 - 2 \cdot \lambda \cdot \mathbf{a}_2 - \beta \cdot \mathbf{a}_1 = 0 \quad (4.27)$$

が得られる。左から \mathbf{a}_2^T を掛けると

$$2 \cdot \mathbf{a}_2^T \cdot \mathbf{C} \cdot \mathbf{a}_2 - 2 \cdot \lambda \cdot \mathbf{a}_2^T \cdot \mathbf{a}_2 - \beta \cdot \mathbf{a}_2^T \cdot \mathbf{a}_1 = 0 \quad (4.28)$$

ここで、 \mathbf{a}_1 と \mathbf{a}_2 は直交するので $\mathbf{a}_2^T \cdot \mathbf{a}_1 = 0$

また、 \mathbf{a}_2 は大きき1に規格化されているので、 $\mathbf{a}_2^T \cdot \mathbf{a}_2 = 1$ となるため式(4.28)は

$$\mathbf{a}_2^T \cdot \mathbf{C} \cdot \mathbf{a}_2 = \lambda \cdot \mathbf{a}_2^T \cdot \mathbf{a}_2 = \lambda \quad (4.29)$$

となる。

4.5.3 オリジナルの多変量データの分散と主成分スコアの分散との関係

第 k 主成分の分散はその固有値 λ_k と等しい (式(4.30))。

$$\begin{aligned}
 V[Z_k] &= \frac{1}{N-1} \sum_{i=1}^N (z_{i1} - \bar{z})^2 \\
 &= \frac{1}{N-1} \sum_{i=1}^N \{a_{k1}(x_{i1} - \bar{x}_1) + a_{k2}(x_{i2} - \bar{x}_2) + \dots + a_{kM}(x_{i1} - \bar{x}_M)\}^2 \\
 &= \frac{1}{N-1} \sum_{j=1}^M \sum_{s=1}^M \left\{ a_{kj} \cdot a_{ks} \cdot \sum_{i=1}^N (x_{ij} - \bar{x}_j)(x_{is} - \bar{x}_s) \right\} \\
 &= \sum_{j=1}^M \sum_{s=1}^M \{ a_{kj} \cdot \text{cov}[X_j, X_s] \cdot a_{ks} \} \\
 &= \{ a_{k1} \cdot (\text{cov}[X_1, X_1] \cdot a_{k1} + \text{cov}[X_1, X_2] \cdot a_{k2} + \dots + \text{cov}[X_1, X_M] \cdot a_{kM}) \} + \dots + \{ a_{kM} \cdot (\text{cov}[X_M, X_1] \cdot a_{k1} + \text{cov}[X_M, X_2] \cdot a_{k2} + \dots + \text{cov}[X_M, X_M] \cdot a_{kM}) \} \\
 &= \left\{ a_{k1} \cdot \sum_{s=1}^M \text{cov}[X_1, X_s] \cdot a_{ks} \right\} + \dots + \left\{ a_{kM} \cdot \sum_{s=1}^M \text{cov}[X_M, X_s] \cdot a_{ks} \right\} \\
 &= \mathbf{a}_k^T \cdot \mathbf{C} \cdot \mathbf{a}_k = \lambda_k
 \end{aligned}
 \tag{4.30}$$

M個の変数の全分散は、M個の主成分の全分散と等しい。

$$\sum_{k=1}^M V[Z_k] = \sum_{k=1}^M V[X_k]
 \tag{4.31}$$

すなわち、式(4.13)で表される %Var[Z_k]は、もともとの M 個の変量からなるサンプルの分散を基準としたときに第 k 番目の主成分が説明した分散量を表す。

問題 15 4つの遺伝子について3種のマイクロアレイ測定を行ったところ以下のような値を得た。遺伝子を主成分第1軸および第2軸にプロットし、因子負荷量および寄与率を求めてこのデータを解釈せよ。

	実験 1	実験 2	実験 3
遺伝子 1	-1.0	-1.0	2.0
遺伝子 2	1.0	-1.0	0.0
遺伝子 3	-1.0	1.0	0.0
遺伝子 4	1.0	1.0	-2.0
遺伝子 5	0.0	0.0	0.0

このデータのそれぞれの実験の平均値は 0 であるので式(4.20)における分散共分散行列は以下の式になる。

$$\begin{pmatrix}
 \frac{1}{4}((-1)^2 + 1^2 + (-1)^2 + 1^2 + 0^2) & \frac{1}{4}((-1)^2 + 1 \cdot (-1) + (-1) \cdot 1 + 1^2 + 0^2) & \frac{1}{4}((-1) \cdot 2 + 1 \cdot 0 + (-1) \cdot 0 + 1 \cdot (-2) + 0^2) \\
 \frac{1}{4}((-1)^2 + 1 \cdot (-1) + (-1) \cdot 1 + 1^2 + 0^2) & \frac{1}{4}((-1)^2 + (-1)^2 + 1^2 + 1^2 + 0^2) & \frac{1}{4}((-1) \cdot 2 + (-1) \cdot 0 + 1 \cdot 0 + 1 \cdot (-2) + 0^2) \\
 \frac{1}{4}((-1) \cdot 2 + 1 \cdot 0 + (-1) \cdot 0 + 1 \cdot (-2) + 0^2) & \frac{1}{4}((-1) \cdot 2 + (-1) \cdot 0 + 1 \cdot 0 + 1 \cdot (-2) + 0^2) & \frac{1}{4}(2^2 + 0^2 + 0^2 + (-2)^2 + 0^2)
 \end{pmatrix}$$

$$= \begin{pmatrix}
 1 & 0 & -1 \\
 0 & 1 & -1 \\
 -1 & -1 & 2
 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \tag{*}$$

とおき以下の式から λ を求める。

$$\begin{vmatrix} 1-\lambda & 0 & -1 \\ 0 & 1-\lambda & -1 \\ -1 & -1 & 2-\lambda \end{vmatrix} = (1-\lambda)(1-\lambda)(2-\lambda) - (1-\lambda) - (1-\lambda) = 0$$

これを解くと $\lambda=3,1,0$ すなわち、第1主成分の分散は3であり、第2主成分の分散は1でオリジナルデータの分散を全て説明したことになる。すなわち3次元データを2次元で誤差なく見ることが可能となった。第1主成分の寄与率は0.75、第2主成分の寄与率は0.25となる。この値をもとに

$$\sum_{j=1}^3 a_{kj}^2 = 1$$

を考慮して(*)を解く。

まずはじめに、第1主成分ベクトル $\begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix}$ を求める。第1主成分ベクトルは $\lambda=3$ とおくことにより得ら

れる行列、

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix} = 3 \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \end{pmatrix}$$

を解くことによりえられる。この行列を解くと以下の連立方程式が得られる。

$$\begin{cases} a_{11} - a_{13} = 3a_{11} \\ a_{12} - a_{13} = 3a_{12} \\ -a_{11} - a_{12} + 2a_{13} = 3a_{13} \end{cases}$$

これらの3つの式を整理すると $a_{11} = -\frac{1}{2}a_{13}$ 、 $a_{12} = -\frac{1}{2}a_{13}$ 、 $-a_{11} - a_{12} + a_{13} = 0$ となる。はじめの二つの式の a_{11}

および a_{12} を最後に式に代入すると両辺0となるので、この連立方程式から得られる関係式は $a_{11} = -\frac{1}{2}a_{13}$ 、

$a_{12} = -\frac{1}{2}a_{13}$ までである。ここで、 $\sum_{j=1}^3 a_{1j}^2 = 1$ が活躍する。

$a_{11} = -\frac{1}{2}a_{13}$ 、 $a_{12} = -\frac{1}{2}a_{13}$ を $\sum_{j=1}^3 a_{1j}^2 = 1$ に代入すると、

$$\left(-\frac{1}{2}a_{13}\right)^2 + \left(-\frac{1}{2}a_{13}\right)^2 + a_{13}^2 = 1$$

となり、 $\left(\frac{3}{2}a_{13}\right)^2=1$ となるため、 $a_{13}=\pm\sqrt{\frac{2}{3}}$ が得られる。よって

$$(a_{11} \ a_{12} \ a_{13}) = \left(-\frac{1}{2}\sqrt{\frac{2}{3}} \quad -\frac{1}{2}\sqrt{\frac{2}{3}} \quad \sqrt{\frac{2}{3}}\right), \left(\frac{1}{2}\sqrt{\frac{2}{3}} \quad \frac{1}{2}\sqrt{\frac{2}{3}} \quad -\sqrt{\frac{2}{3}}\right)$$

の二つの第1主成分ベクトルが得られる。これらの二つのベクトルは完全に逆向きであるので、データの最大の分散の向きを表す情報としては同じであるので一方のみを取ればよい。はじめのベクトルを第1主成分ベクトルとして採用すると、

$$Z_1 = -\frac{1}{2}\sqrt{\frac{2}{3}}X_1 - \frac{1}{2}\sqrt{\frac{2}{3}}X_2 + \sqrt{\frac{2}{3}}X_3$$

となる。この式を使うと遺伝子1における第1主成分スコアは、

$$\begin{aligned} z_{11} &= -\frac{1}{2}\sqrt{\frac{2}{3}} \cdot (-1) - \frac{1}{2}\sqrt{\frac{2}{3}} \cdot (-1) + \sqrt{\frac{2}{3}} \cdot 2 \\ &= 3\sqrt{\frac{2}{3}} \end{aligned}$$

と求まる。以下遺伝子2・5における主成分スコア $z_{12}=0, z_{13}=0, z_{14}=3\sqrt{\frac{2}{3}}, z_{15}=0$ と求まる。

次に、 $\lambda=1$ のとき

$$\begin{cases} a_{21} - a_{23} = a_{21} \\ a_{22} - a_{23} = a_{22} \\ -a_{21} - a_{22} + 2a_{23} = a_{23} \end{cases}$$

となり、 $a_{23}=0, a_{21}=-a_{22}$ が得られるので $\sum_{j=1}^3 a_{2j}^2=1$ をうまく使って

$$2 \cdot a_{22}^2 = 1$$

となるので

$$(a_{21} \ a_{22} \ a_{23}) = \left(-\sqrt{\frac{1}{2}} \quad \sqrt{\frac{1}{2}} \quad 0\right), \left(\sqrt{\frac{1}{2}} \quad -\sqrt{\frac{1}{2}} \quad 0\right)$$

が得られる。はじめのベクトルを第1主成分ベクトルとして採用すると、第2主成分スコアは

$$Z_2 = -\sqrt{\frac{1}{2}}X_1 + \sqrt{\frac{1}{2}}X_2$$

となる。この式を使い遺伝子1・5における第2主成分スコアを求めると

$z_{21}=0, z_{23}=0, z_{24}=3\sqrt{\frac{2}{3}}, z_{25}=0, z_{22}=0, z_{23}=0, z_{24}=3\sqrt{\frac{2}{3}}, z_{25}=0$ と求まる。このようにして5つの遺伝子に対す

第1および第2主成分スコアは以下の表ようになる。

遺伝子	実験1	実験2	実験3	Z ₁	Z ₂
遺伝子1	-1.0	-1.0	2.0	$3\sqrt{2/3}$	0
遺伝子2	1.0	-1.0	0.0	0	$-2\sqrt{1/2}$
遺伝子3	-1.0	1.0	0.0	0	$2\sqrt{1/2}$
遺伝子4	1.0	1.0	-2.0	$-3\sqrt{2/3}$	0
遺伝子5	0.0	0.0	0.0	0	0
分散	1	1	2	3	1

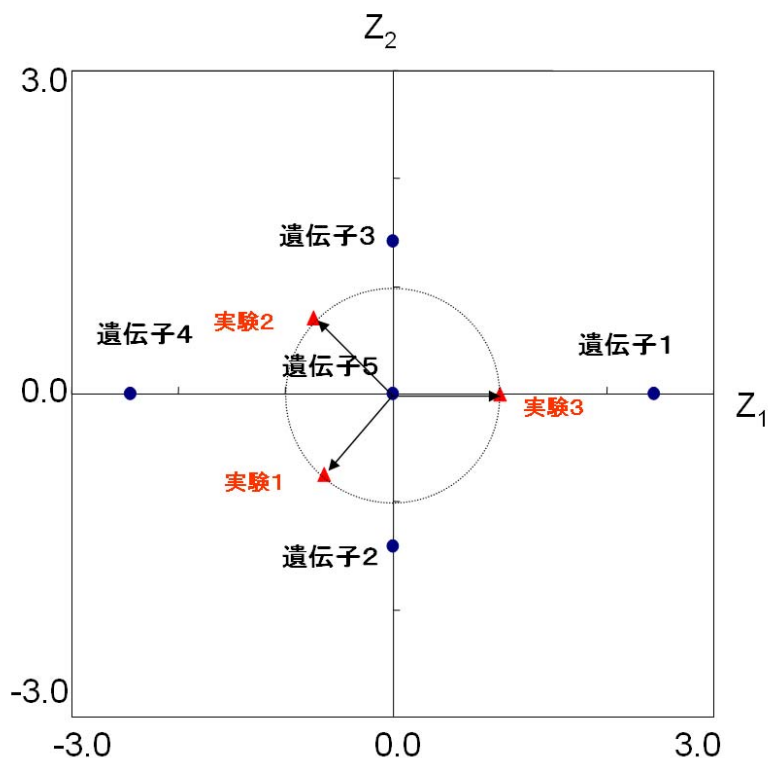
これらの値をもとに実験と主成分スコアとの相関係数（因子負荷量）を求めよう。例えば、実験1と第1主成分スコアの相関を求めると、

$$\begin{aligned}
 r(X_1, Z_1) &= \frac{\sum_{j=1}^5 (x_{j1} - \bar{X}_1)(z_{j1} - \bar{Z}_1)}{\sqrt{\sum_{j=1}^5 (x_{j1} - \bar{X}_1)^2 \sum_{j=1}^5 (z_{j1} - \bar{Z}_1)^2}} \\
 &= \frac{(-1.0-0)\left(3\sqrt{\frac{2}{3}}-0\right) + (1.0-0)(0-0) + (-1.0-0)(0-0) + (1.0-0)\left(-3\sqrt{\frac{2}{3}}-0\right) + (0-0)(0-0)}{\sqrt{\left[(-1.0-0)^2 + (1.0-0)^2 + (-1.0-0)^2 + (1.0-0)^2 + (0-0)^2\right] \left[\left(3\sqrt{\frac{2}{3}}-0\right)^2 + (0-0)^2 + (0-0)^2 + \left(-3\sqrt{\frac{2}{3}}-0\right)^2 + (0-0)^2\right]}} \\
 &= \frac{-6\sqrt{\frac{2}{3}}}{\sqrt{4 \cdot 12}} \\
 &= -\frac{1}{2}\sqrt{2}
 \end{aligned}$$

同様に、がんばって計算すると以下の表を得る。

	Z ₁	Z ₂
実験1	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}\sqrt{2}$
実験2	$-\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{2}$
実験3	1.00	0

主成分スコアと因子負荷量を同一の座標にプロットすると以下のプロットが得られる。



式(3.13)における第1主成分の寄与率は75%、第2主成分の寄与率は25%であり、3つの実験における遺伝子の分布は二つの主成分軸により完全に説明することができた。この図から実験1と2は第1主成分について負に寄与する、一方、実験3は第1主成分に正に寄与する。実験1と実験2は、それぞれ、第2主成分軸に負および正に寄与する。さらに実験3は第2主成分軸に全く寄与しない。実験間相関を求めると

$$r(\text{実験1}, \text{実験2}) = 0$$

$$r(\text{実験1}, \text{実験3}) = -\frac{1}{2}\sqrt{2}$$

$$r(\text{実験2}, \text{実験3}) = -\frac{1}{2}\sqrt{2}$$

となる。図から実験1と実験2のなす角は90度であり、相関がないことを示している。一方、実験1と実験3、図より、実験2と実験3は明らかに負の相関があることがわかる。このように、相関係数ではそれぞれの実験対の関係を見ることのできるのに対して、主成分分析により、3以上の実験間の関係を独立かつ次元を減らした軸上で比較することが可能となる。

次に、遺伝子間関係を見てみよう。遺伝子1は実験1および実験2で負の値であり実験3は正の値をもつ。主成分プロットを見ると遺伝子1はZ1軸上で正の値をもっている。すなわち、実験3の寄与が正であるか、あるいは、実験1あるいは2の寄与が負であるためにこの値を持つことがこの図から読み取れる。

また、遺伝子間の関係が図から一見して理解できる。このように次元を減らすことにより多次元空間のデータの分布を把握することができる。もう一つ問題を解いてみよう。

問題 16 4つの遺伝子について3種のマイクロアレイ測定を行ったところ以下のような値を得た。遺伝子を主成分第1軸および第2軸にプロットせよ。また、因子負荷量および寄与率を求めてこのデータを解釈せよ。(問題 15 の実験 3 の正負が逆になっていることに注意！)

	実験 1	実験 2	実験 3
遺伝子 1	-1.0	-1.0	-2.0
遺伝子 2	1.0	-1.0	0.0
遺伝子 3	-1.0	1.0	0.0
遺伝子 4	1.0	1.0	2.0
遺伝子 5	0.0	0.0	0.0

4.6 k-平均法

N 個の遺伝子にそれぞれに M 種類の実験条件による相対遺伝子発現量が測定されたとすると、N x M の行列によりこれらの発現量を表現することができる。

GENE	X_1	X_2	...	X_j	...	X_{M-1}	X_M
<i>gene</i> ₁	x_{11}	x_{12}	...	x_{1j}	...	x_{1M-1}	x_{1M}
<i>gene</i> ₂	x_{21}	x_{22}	...	x_{2j}	...	x_{2M-1}	x_{2M}
...
...
<i>gene</i> _i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{iM-1}	x_{iM}
...
...
<i>gene</i> _{N-1}	x_{N-11}	x_{N-12}	...	x_{N-1j}	...	x_{N-1M-1}	x_{N-1M}
<i>gene</i> _N	x_{N1}	x_{N2}	...	x_{Nj}	...	x_{NM-1}	x_{NM}

(1) パラメータ設定

分類するクラスの数 **K** ならびに学習回数 **nc** を設定する。また、**ncycle=0** とする。

(2) 初期分類：遺伝子を K 個のクラスに分類する。クラス 1, 2, ..., K に属する遺伝子の数をそれぞれ n_1, n_2, \dots, n_K とする。

	<i>gene</i> ₁ ⁽¹⁾	$x_{11}^{(1)}$	$x_{12}^{(1)}$...	$x_{1j}^{(1)}$...	$x_{1m-1}^{(1)}$	$x_{1m}^{(1)}$
クラス 1	<i>gene</i> ₂ ⁽¹⁾	$x_{21}^{(1)}$	$x_{22}^{(1)}$...	$x_{2j}^{(1)}$...	$x_{2m-1}^{(1)}$	$x_{2m}^{(1)}$

	<i>gene</i> _{n_1} ⁽¹⁾	$x_{n_1 1}^{(1)}$	$x_{n_1 2}^{(1)}$...	$x_{n_1 j}^{(1)}$...	$x_{n_1 m-1}^{(1)}$	$x_{n_1 m}^{(1)}$

4.7 自己組織化法 (SOM)

BL-SOM のアルゴリズムの概念図を図 18 に示す。BL-SOM は、(1)初期値設定、(2)ウェイトベクトルの開発、(3)開発されたウェイトベクトルによる遺伝子の分類の三つのステップからなる。

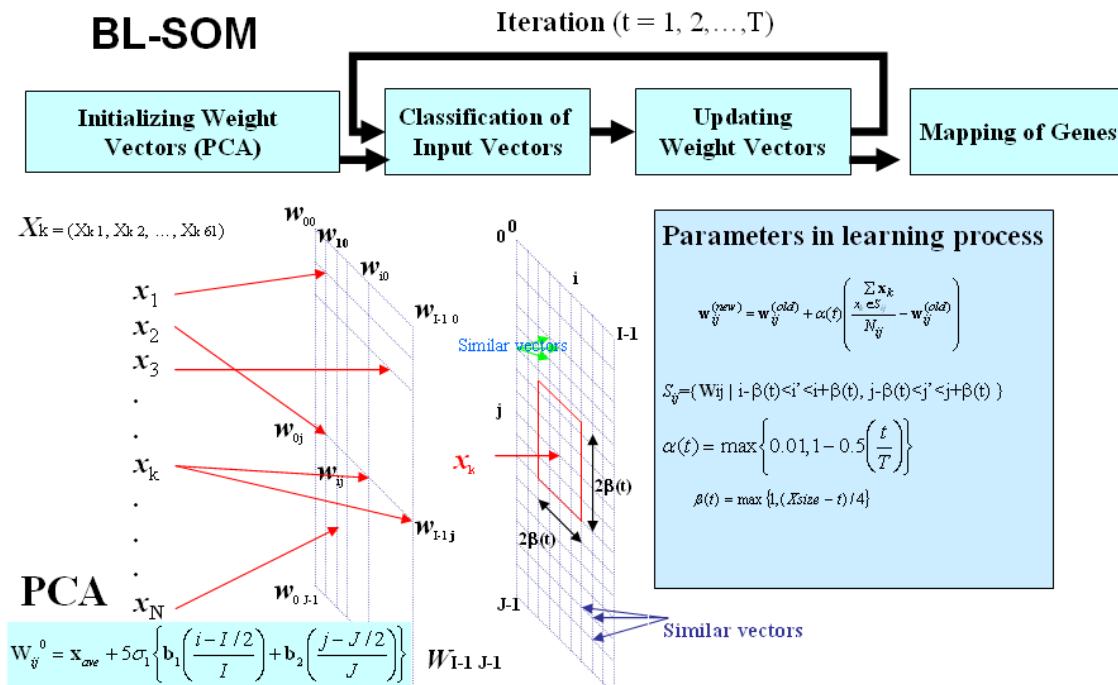


図 18 BL-SOM のアルゴリズムの概念図

(1) 初期値の設定

第 s 番目の遺伝子の M 個のマイクロアレイに対する発現プロファイルを $\mathbf{x}_s = (x_{s1}, x_{s2}, \dots, x_{sM})$ とする。二次元の格子点を i, j ($i = 1, 2, \dots, I, j = 1, 2, \dots, J$) とする。二次元の格子点 (ij) 上に M 次元のリファレンスベクトル $\mathbf{w}_{ij} = (w_{ij1}, w_{ij2}, \dots, w_{ijM})$ を置く。初期リファレンスベクトルは、入力ベクトル \mathbf{x}_i の平均 \mathbf{x}_{ave} を基準とした 2 つの主成分軸により設定する。入力ベクトル \mathbf{x}_i に対して得られた主成分第 1 および第 2 ベクトルを \mathbf{b}_1 および \mathbf{b}_2 とする。これら二つの軸に対する入力ベクトルの標準偏差を σ_1 および σ_2 とする。また、入力ベクトルの平均ベクトルを \mathbf{x}_{ave} としたとき、 \mathbf{w}_{ij} は式 (4.40) により定義した。

$$\mathbf{w}_{ij} = \mathbf{x}_{ave} + 5\sigma_1 \mathbf{b}_1 \left(\frac{i-I/2}{I} \right) + 5\sigma_2 \mathbf{b}_2 \left(\frac{j-J/2}{J} \right) \tag{4.40}$$

ここで、 $J = I \sigma_1 / \sigma_2$ とする。

(2) ウェイトベクトルの開発

コホネンの SOM アルゴリズム (OR-SOM) においては、リファレンスベクトル \mathbf{w}_{ij} に作用する入力ベクトル \mathbf{x}_i により逐次 \mathbf{w}_{ij} を更新する。

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij}^{(old)} + \alpha(t) (\mathbf{x}_k - \mathbf{w}_{ij}^{(old)}) \tag{4.41}$$

ここで $\alpha(t)$ は、学習係数 ($0 < \alpha(t) < 1$) である。また、 ij の近傍に関してもリファレンスベクトルの更

新を行う。この方法ではデータの入力順が \mathbf{w}_{ij} に反映される。すなわち、入力順が後である入力ベクトルほどリファレンスベクトル \mathbf{w}_{ij} に影響を及ぼす。本研究では、入力順によるリファレンスベクトルへの影響を回避するためのアルゴリズムを検討した。はじめに、全ての入力ベクトル \mathbf{x}_k を最小のユークリッド距離を有するリファレンスベクトル $\mathbf{w}_{i', j'}$ に分類する。つぎに次式によって、リファレンスベクトル \mathbf{w}_{ij} を更新する。

$$\mathbf{w}_{ij}^{(new)} = \mathbf{w}_{ij}^{(old)} + \alpha(t) \left(\frac{\sum_{\mathbf{x}_k \in S_{ij}} \mathbf{x}_k}{N_{ij}} - \mathbf{w}_{ij}^{(old)} \right) \quad (4.42)$$

ここで、近傍 S_{ij} は $i - \beta(t) \leq i' \leq i + \beta(t)$ か $j - \beta(t) \leq j' \leq j + \beta(t)$ の条件を満たす格子点 i', j' に分類された入力ベクトル \mathbf{x}_k の集合、 N_{ij} は S_{ij} の要素数である。また、 $\alpha(t)$ は学習係数、 $\beta(t)$ は近傍を決定する数である。学習は次式によって定義する二乗誤差で評価する。

$$e(t) = \sum_{k=1}^N \{\mathbf{x}_k - \mathbf{w}_{i', j'}\}^2 \quad (4.43)$$

ここで i', j' は \mathbf{x}_k が分類された格子点である。 ($k=1, 2, \dots, N$)

(3) 遺伝子の分類

各々の遺伝子について発現プロファイルと全てのウエイトベクトルとの距離を求め、最小の距離を有するウエイトベクトルと対応する格子点に遺伝子を分類する。

5 教師あり学習

定量化学モデリングとはいくつかの因子により一つの化学的性質あるいは量を説明するための数理モデルをつくりことである。いくつかの因子を変量 x_1, \dots, x_M 、説明したい化学的性質、活性あるいは量を変量 y としたときの定量化学モデルは

$$y = f(x_1, x_2, \dots, x_M)$$

により表すことができる。ここで、変量 x_1, \dots, x_M を記述子（説明変数）、変量 y を目的変数という。いま、 N 個のサンプルについて変量 x_1, \dots, x_M と y が観測されたとすると、これらをマトリックスにより、

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{pmatrix}$$

により表すことができる。これらの観測値により $y = f(x_1, x_2, \dots, x_M)$ の関係を推定するための比較的簡単な方法として、説明変数の線形和で目的変数を説明することは化学を含むさまざまな分野で広く行われている。

$$y = a_0 + a_1 x_1 + \dots + a_M x_M$$

このような線形式を推定する方法に線形重回帰分析法(MLR, Multiple Linear Regression Analysis)、ならびに部分最小二乗法(PLS, Partial Least Squares)がある。前者は説明変数間に相関があるときに係数 a_j ($j=1,2,\dots,M$) を安定に求めることが難しくなる。また、説明変数の数がサンプル数を上回る場合、MLR によりモデル式(2)を求めることはできない。これらの二点を克服した方法が Wold により開発され Partial Least Squares(PLS)と名づけられた。目的変数 y がクラスを表す場合、すなわち、サンプルを説明変数によりベクトルで表現し、このベクトルをもとにクラスを予測したいときにおいても、PLS 法を用いることができる。このような目的の方法として種々の教師つき学習が提案されている。古典的なものとしては線形識別関数法、ベイズ法などが挙げられる。また、種々の人工ニューラルネットアルゴリズムも提案されている。以下では特に Calibration 法のアルゴリズムを説明する。

教師あり学習

k-Nearest neighbor classifiers

Discriminant function based on linear discriminant

Bayes rule, Maximum likelihood discriminant rule,

Fisher linear discriminant analysis, Logistic discrimination,

Penalized discriminant analysis

Support vector machines

Calibration

Multiple Linear Regression (MLR)

Partial Least Squares(PLS)

5.1 データの記述

データを以下のように記述する。

サンプルID	目的変数	説明変数
	Y	$X_1 \quad \dots \quad X_j \quad \dots \quad X_M$
1	$\begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_N \end{pmatrix}$	$\mathbf{X}_M = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ x_{N1} & \dots & x_{Nj} & \dots & x_{NM} \end{pmatrix}$
...		
i		
...		
M		

- N 全サンプル数
- M 変数の数
- Y 目的変数
- X_j j 番目の説明変数
- y_i i 番目のサンプルの目的変数の値
- x_{ij} i 番目のサンプルにおける j 番目の説明変数の値

5.2 重回帰分析(Multi-Regression Analysis)

一つの説明目的変数複数 Y を複数の説明変数 ($X_j, j=1,2, \dots, M$) による線形モデル式により表現する(式(5.1))。

$$Y = b_0 + b_1X_1 + \dots + b_jX_j + \dots + b_MX_M \tag{5.1}$$

式(5.1)にデータを代入すると、

$$\begin{aligned} y_1 &= b_0 + b_1x_{11} + \dots + b_jx_{1j} + \dots + b_Mx_{1M} + e_1 \\ y_2 &= b_0 + b_1x_{21} + \dots + b_jx_{2j} + \dots + b_Mx_{2M} + e_2 \\ &\dots \\ y_i &= b_0 + b_1x_{i1} + \dots + b_jx_{ij} + \dots + b_Mx_{iM} + e_i \\ &\dots \\ y_N &= b_0 + b_1x_{N1} + \dots + b_jx_{Nj} + \dots + b_Mx_{NM} + e_N \end{aligned}$$

とかける。ここで、 e_i はサンプル i のデータをモデル式に代入したときの残差である。これら N 個のサンプルの残差の二乗和を G とする。

$$G(b_0, b_1, \dots, b_j, \dots, b_M) = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N (y_i - b_0 - b_1x_{i1} - \dots - b_jx_{ij} - \dots - b_Mx_{iM})^2 \tag{5.2}$$

式(5.2)の残差を最小とするように係数 b_j ($j=0,1,\dots,M$) を求める。このことは

$$\frac{\partial G}{\partial b_j} = 0 \quad (j=0,1,\dots,M) \tag{5.3}$$

を求めることにより達成される。j=0 において

$$\frac{\partial G}{\partial b_0} = 2 \sum_{i=1}^N (y_i - b_0 - b_1 x_{i1} - \dots - b_j x_{ij} - \dots - b_M x_{iM}) (-1) = 0$$

となる。これを整理すると式(5.4)をえる。

$$N b_0 + \sum_{i=1}^N b_1 x_{i1} + \dots + \sum_{i=1}^N b_j x_{ij} + \dots + \sum_{i=1}^N b_M x_{iM} = \sum_{i=1}^N y_i \tag{5.4}$$

$$\sum_{i=1}^N x_{i1} + \sum_{i=1}^N b_1 x_{i1}^2 + \dots + \sum_{i=1}^N b_j x_{i1} x_{ij} + \dots + \sum_{i=1}^N b_M x_{i1} x_{iM} = \sum_{i=1}^N x_{i1} y_i$$

.....

$$\sum_{i=1}^N x_{ij} + \sum_{i=1}^N b_1 x_{ij} x_{i1} + \dots + \sum_{i=1}^N b_j x_{ij}^2 + \dots + \sum_{i=1}^N b_M x_{ij} x_{iM} = \sum_{i=1}^N x_{ij} y_i \tag{5.5}$$

.....

$$\sum_{i=1}^N x_{iM} + \sum_{i=1}^N b_1 x_{iM} x_{i1} + \dots + \sum_{i=1}^N b_j x_{iM} x_{ij} + \dots + \sum_{i=1}^N x_{iM}^2 = \sum_{i=1}^N x_{ij} y_i$$

いま、

$$\mathbf{X}_{M+1} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1M} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{N1} & \dots & x_{Nj} & \dots & x_{NM} \end{pmatrix}, \mathbf{y} = \begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_N \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_0 \\ \dots \\ b_j \\ \dots \\ b_M \end{pmatrix}$$

とすると、式(5.5)は

$$\mathbf{X}_{M+1}^T \mathbf{X}_{M+1} \mathbf{b}_{M+1} = \mathbf{X}_{M+1}^T \mathbf{y} \tag{5.6}$$

となり、係数 \mathbf{b} は

$$\mathbf{b}_{M+1} = (\mathbf{X}_{M+1}^T \mathbf{X}_{M+1})^{-1} \mathbf{X}_{M+1}^T \mathbf{y} \tag{5.7}$$

により求まる。このように、残差を最小とする線形モデルを重回帰分析(Multi Regression Analysis; MRA)とよぶ。ここで、 \mathbf{X}_{M+1} はデータの記述における \mathbf{X}_M に数値 1 の行が追加されているため、このように記述した。

係数の信頼区間

$$b_0 \pm t(N-M-1; \alpha) \sqrt{\left(\frac{1}{N} + \sum_{i=k=1}^N \sum_{i=1}^N \bar{x}_i \bar{x}_i S^{ik} \right) \frac{\sum_{i=1}^N e_i^2}{N-M-1}}$$

$$b_j \pm t(N-M-1; \alpha) \sqrt{\frac{S^{jj} \sum_{i=1}^N e_i^2}{N-M-1}}$$

ここで、

$$\mathbf{S} = \begin{pmatrix} S_{11} & \dots & S_{1j} & \dots & S_{1M} \\ \dots & \dots & \dots & \dots & \dots \\ S_{j1} & \dots & S_{jj} & \dots & S_{jM} \\ \dots & \dots & \dots & \dots & \dots \\ S_{M1} & \dots & S_{Mj} & \dots & S_{MM} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N x_{i1}^2 & \dots & \sum_{i=1}^N x_{i1}x_{ij} & \dots & \sum_{i=1}^N x_{i1}x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^N x_{ij}x_{i1} & \dots & \sum_{i=1}^N x_{ij}^2 & \dots & \sum_{i=1}^N x_{ij}x_{iM} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^N x_{iM}x_{i1} & \dots & \sum_{i=1}^N x_{iM}x_{ij} & \dots & \sum_{i=1}^N x_{iM}^2 \end{pmatrix} \quad \mathbf{S}^{-1} = \begin{pmatrix} S^{11} & \dots & S^{1j} & \dots & S^{1M} \\ \dots & \dots & \dots & \dots & \dots \\ S^{j1} & \dots & S^{jj} & \dots & S^{jM} \\ \dots & \dots & \dots & \dots & \dots \\ S^{M1} & \dots & S^{Mj} & \dots & S^{MM} \end{pmatrix}$$

である。

F 検定

(F-分布の説明を入れる)

説明変数が M 個のときの分散分析表を以下に示す。この表において、 y_i は実測値、 \hat{y}_i は回帰モデル式により算出された推定値、である。分散比 $F = V_R/V_e$ が、F 表の $F(M, N - M - 1; \alpha)$ より大きいとき、回帰モデル式は有意であると判断される。

説明変数がM個の分散分析表

変動因	自由度	平方和	分散	分散比
全体	$N-1$	$S_{yy} = \sum_{i=1}^N (y_i - \bar{y})^2$		
回帰による	M	$S_R = \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2$	$V_R = S_R / M$	V_R / V_e
回帰からの残差	$N-M-1$	$S_e = \sum_{i=1}^N (y_i - \hat{y}_i)^2$ $= S_{yy} - S_R$	$V_e = S_e / (N - M - 1)$	

重相関係数

実測値 y_i と回帰モデル式により算出された推定値 \hat{y}_i との相関係数を重相関係数(multiple correlation)という。

$$R = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}}$$

R の 2 乗を決定係数(coefficient of determination)という。

5.3 主成分回帰法

説明変数行列 \mathbf{X} が共線性を持つとき (説明変数間に強い相関あるとき)、MLR の偏回帰係数 \mathbf{b} を求める操作において逆行列が含まれるために一意には決まらなくなる。式(7)において、説明変数間の相関をなくし

回帰分析を行う方法として主成分回帰法がある。回帰モデル式は

$$Y = b_0 + b_1 Z_1 + \dots + b_j Z_j + \dots + b_L Z_L \tag{5.8}$$

により表現される。ここで、 $Z_j (j=1,2,\dots,L)$ は第j主成分である。

5.4 PLS (Partial Least Square)

線形重回帰分析ではサンプル数 N が説明変数の数 M に比べて大きくない（理想的には $N > 3M$ ）と有意な回帰式は得られない。また、共線性の問題で回帰式が得られない場合もある。PLSでは、これらの2点を克服した線形モデルであり、サンプル数 N に比べて説明変数 M が非常に大きい場合においても適用可能である。

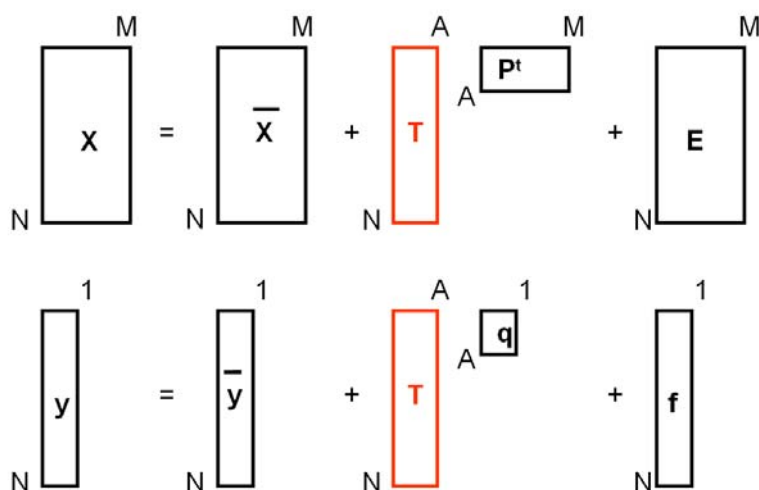
目的変数 \mathbf{y} が一つの性質により定量される場合(PLS1)のアルゴリズムを説明する。PLS1のモデル式を式(5.9)、(5.10)に示す。

$$\mathbf{y} = \bar{\mathbf{y}} + \sum_{k=1}^A \mathbf{t}_k q_k + \mathbf{e} = \bar{\mathbf{y}} + \mathbf{T} \cdot \mathbf{q} + \mathbf{e} \tag{5.9}$$

$$\mathbf{X} = \bar{\mathbf{X}} + \sum_{k=1}^A \mathbf{t}_k \mathbf{p}_k^T + \mathbf{E} = \bar{\mathbf{X}} + \mathbf{T} \cdot \mathbf{P}^T + \mathbf{E} \tag{5.10}$$

式(5.9)および(5.10)において、潜在変数 \mathbf{t}_k により、説明変数 \mathbf{X} と目的変数 \mathbf{y} が関係付けられている。また式(5.9)においては q_k が、式(5.10)においては \mathbf{p}_k によりモデル式が定義されている。これらのパラメータの推定法について以下に示す。

PLS1の概念図



1成分の PLS1

\mathbf{y} , \mathbf{X} はそれぞれの変量について平均値が0となるように規格化されているものとする。A=1 の場合、(5.9)と(5.10)は、

$$\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{e} \quad (5.11)$$

$$\mathbf{X} = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{E} \quad (5.12)$$

いま、 \mathbf{X} の M 個の変数と対応した負荷因子 \mathbf{w} を定義する。ここで

$$\mathbf{w} = (w_1, w_2, \dots, w_M) \quad (5.13)$$

について、

$$\sum_{j=1}^M w_j^2 = 1 \quad (5.14)$$

とする。

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w} \quad (5.15)$$

とおくことにより得られる式(5.16)が、PLS の鍵となる式ある。

$$\begin{aligned} \mathbf{y}\mathbf{t}_1 &= \mathbf{y}\mathbf{X}\mathbf{w} \\ &= \sum_{k=1}^M \left\{ w_k \sum_{i=1}^N y_i x_{ik} \right\} \end{aligned} \quad (5.16)$$

式(5.13)の $\sum_{i=1}^N y_i x_{ik}$ は、 y_i と k 番目の変数の要素 x_{ik} の積であり、仮に

$$y_i = \frac{y'_i - \bar{y}}{\sqrt{\sum_i^N (y'_i - \bar{y})^2}} \quad (5.17)$$

$$x_{ik} = \frac{x'_{ik} - \bar{x}_k}{\sqrt{\sum_i^N (x_{ik} - \bar{x}_k)^2}} \quad (5.18)$$

と規格化しておけば、

$$\sum_{i=1}^N y_i x_{ik} = \frac{\sum_{i=1}^N (x'_{ik} - \bar{x}_k)(y'_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_{ik} - \bar{x}_k)^2 \sum_{i=1}^N (y'_i - \bar{y})^2}} \quad (5.19)$$

となり、 k 番目の変数 x_k と目的変数 y のピアソン相関となる。すなわち、式(5.16)は、それぞれの変数 $k=1,2,\dots,M$ と目的変数 y との積（相関）に重み w_k を掛け合わせた合計値である。(5.14)により規格化されている w_k を適に決めることにより、 $\mathbf{y}\mathbf{t}_1$ を最大化することができれば、変数 x_k ($k=1,2,\dots,M$) と目的変数 y の相関の合計を最大化することができ、式(5.12)を式(5.13)をもとに変形すると式(5.20)が得られる。

$$\begin{aligned} \mathbf{y} &= \mathbf{t}_1 q_1 + \mathbf{e} \\ &= \mathbf{X} \mathbf{w} q_1 + \mathbf{e} \end{aligned} \tag{5.20}$$

式(5.20)より、M 個の変量 x_k ($k=1,2,\dots,M$) と y の相関が最大となるように重み w_k が決められたことになる。

それでは、重み w_k をどのように決めればよいのであろうか。

重み w_k を決めるには、ラグランジュの未定乗数法を使う。いま、制約条件として式(5.14)のもとで式(5.16)を最大化すればよいのであるから、

$$G(\mathbf{w}, \mu) = \sum_{k=1}^M \left\{ w_k \sum_{i=1}^N y_i x_{ik} \right\} - \mu \left\{ \sum_{j=1}^M w_j^2 - 1 \right\} \tag{5.21}$$

とおき、

$$\frac{\partial G(\mathbf{w}, \mu)}{\partial w_j} = 0 \tag{5.22}$$

を解くことにより、重み w_k を求めることができる。式(5.22)を実際にとくと

$$\frac{\partial G(\mathbf{w}, \mu)}{\partial w_j} = \sum_{i=1}^N y_i x_{ij} - 2\mu w_j = 0$$

となり、

$$\sum_{i=1}^N y_i x_{i1} = 2\mu w_1 \tag{5.23-1}$$

$$\sum_{i=1}^N y_i x_{i2} = 2\mu w_2 \tag{5.23-2}$$

...

$$\sum_{i=1}^N y_i x_{ik} = 2\mu w_k \tag{5.23-k}$$

...

$$\sum_{i=1}^N y_i x_{iM} = 2\mu w_M \tag{5.23-M}$$

が得られる。式(5.23-1)から(5.23-M)の二乗和を求めると、

$$\sum_{k=1}^M \left(\sum_{i=1}^N y_i x_{ik} \right)^2 = \sum_{k=1}^M (2\mu w_k)^2 = (2\mu)^2 \sum_{k=1}^M (w_k)^2 = (2\mu)^2$$

となり、

$$2\mu = \sqrt{\sum_{k=1}^M \left(\sum_{i=1}^N y_i x_{ik} \right)^2} \tag{5.24}$$

を得る。すなわち、

$$w_k = \frac{\sum_{i=1}^N y_i x_{ik}}{\sqrt{\sum_{k=1}^M \left(\sum_{i=1}^N y_i x_{ik} \right)^2}} \tag{5.25}$$

$$\mathbf{w} = \frac{\mathbf{X}^T \mathbf{y}}{\|\mathbf{X}^T \mathbf{y}\|} \tag{5.26}$$

を得る。このようにして、 w_k は y_i と x_{ik} の積の $i=1,2,\dots,N$ に関する和により求めることができる。続いて

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w} \tag{5.27}$$

より、

$\sum_{j=1}^M w_j^2 = 1$ を利用して、

ピアソン相関と

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}$$

にする最適化することにより、

$$G(\mathbf{w}) = \mathbf{y}^T \mathbf{t} -$$

$$=$$

PLS1 アルゴリズム

[1] $\mathbf{X}_0 = \mathbf{X} - \bar{\mathbf{X}}$, $\mathbf{y}_0 = \mathbf{y} - \bar{\mathbf{y}}$, $k=1$ と設定する

$$[2] \mathbf{w}_k = \frac{\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}}{\|\mathbf{X}_{k-1}^T \mathbf{y}_{k-1}\|}$$

$$[3] \mathbf{t}_k = \mathbf{X}_{k-1} \mathbf{w}_k$$

$$[4] q_k = \frac{\mathbf{y}_{k-1}^T \cdot \mathbf{t}_k}{\mathbf{t}_k^T \cdot \mathbf{t}_k}, \quad \mathbf{p}_k = \frac{\mathbf{X}_{k-1}^T \cdot \mathbf{t}_k}{\mathbf{t}_k^T \cdot \mathbf{t}_k}$$

$$[5] \mathbf{y}_k = \mathbf{y}_{k-1} + \mathbf{t}_k q_k, \quad \mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \cdot \mathbf{p}_k^T$$

[6] $k \leftarrow k+1$ (k をインクリメント)

[7] 次の PLS 成分が必要なとき [2] へ、不要なとき [8] へ

[8] 終了

PLS モデル式において説明変数 \mathbf{X} により直接 \mathbf{y} を表現すると式(5.11)となる。

$$\mathbf{Y} = [\bar{\mathbf{y}} + (\bar{\mathbf{X}}^T \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q})] + [\mathbf{X}^T \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q}] \tag{5.11}$$

予測誤差が最小の成分を最適成分として用いることが多い。なお、式(3.44)における PRESS(Prediction Residual Sum of Squares)が用いられることが多い。

$$PRESS = \sum_{forTestSet} (y_{obs} - y_{pred})^2 \tag{5.12}$$

また、モデルの精度は

$$R_{pred}^2 = 1 - \frac{\sum (y_{obs} - y_{pred})^2}{\sum (y_{obs} - \bar{y})^2} \tag{5.13}$$

により評価することが多い。

R_{pred}^2 の最大値は 1 であり、 $R_{pred}^2 < 0$ のとき、そのモデルは平均値を予測値とするモデルより悪い結果を与えることとなる。

5.5 PLS2 目的変数 \mathbf{Y} が複数の性質により定量される場合(PLS2)のアルゴリズムを説明する。

$$\mathbf{Y} = \begin{pmatrix} y_{11} \dots y_{2S} \\ \dots \dots \dots \\ y_{21} \dots y_{2S} \\ \dots \dots \dots \\ y_{N1} \dots y_{NS} \end{pmatrix}$$

PLS2 のモデル式を式(5.14)、(5.15)に示す。

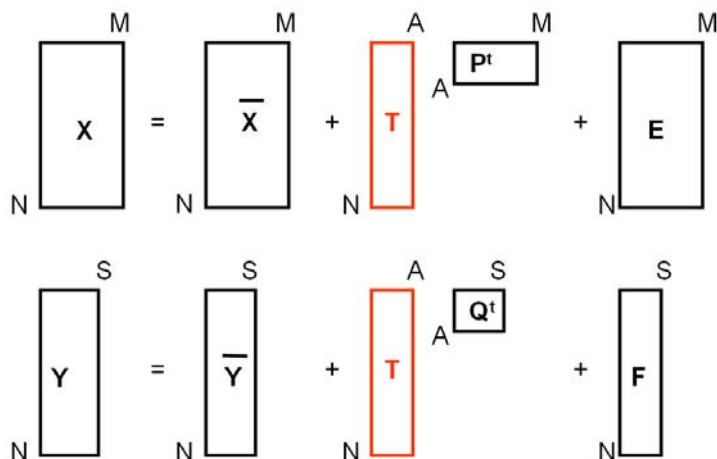
$$\mathbf{Y} = \bar{\mathbf{Y}} + \mathbf{T} \cdot \mathbf{Q} + \mathbf{F} \tag{5.14}$$

$$\mathbf{X} = \bar{\mathbf{X}} + \mathbf{T} \cdot \mathbf{P}^T + \mathbf{E} \tag{5.15}$$

式(5.14)および(5.15)において、潜在変数 \mathbf{t}_k により、説明変数 \mathbf{X} と目的変数 \mathbf{Y} が関係付けられている。こ

ここで、E と F は残差である。また、T、Q、P は説明変数 **X** と目的変数 **Y** を関連付けるパラメータである。これらのパラメータの推定アルゴリズムを以下に示す。

PLS2の概念図



PLS2 アルゴリズム

[1] $\mathbf{X}_0 = \mathbf{X} - \bar{\mathbf{X}}$, $\mathbf{y}_0 = \mathbf{y} - \bar{\mathbf{y}}$, $k=1$ と設定する。また、 \mathbf{u}_0 は $\mathbf{y} - \bar{\mathbf{y}}$ における任意の一変量から選ばれたベクトルとする。

[2]
$$\mathbf{w}_k = \frac{\mathbf{X}_{k-1}^T \mathbf{u}_{k-1}}{\|\mathbf{X}_{k-1}^T \mathbf{u}_{k-1}\|}$$

[3]
$$\mathbf{t}_k = \mathbf{X}_{k-1} \mathbf{w}_k$$

[4]
$$\mathbf{q}_k = \frac{\mathbf{y}_{k-1}^T \cdot \mathbf{t}_k}{\mathbf{t}_k^T \cdot \mathbf{t}_k}, \quad \mathbf{p}_k = \frac{\mathbf{X}_{k-1}^T \cdot \mathbf{t}_k}{\mathbf{t}_k^T \cdot \mathbf{t}_k}, \quad \mathbf{u}_k = \frac{\mathbf{y}_{k-1} \cdot \mathbf{q}_k}{\mathbf{q}_k^T \cdot \mathbf{q}_k}$$

[5]
$$\mathbf{Y}_k = \mathbf{Y}_{k-1} + \mathbf{t}_k \mathbf{q}_k, \quad \mathbf{X}_k = \mathbf{X}_{k-1} - \mathbf{t}_k \cdot \mathbf{p}_k^T$$

[6] $k \leftarrow k+1$ (k をインクリメント)

[7] 次の PLS 成分が必要なとき [2] へ、不要なとき [8] へ

[8] 終了

6.発現プロファイルによる遺伝子分類のさきにあるもの

生物のゲノムにおける全塩基配列が決定されることにより、生物が持つ遺伝子セットが明らかとなる。遺伝子セットの中から有限個の遺伝子を対象に解析を進めることを可能としたことがゲノムプロジェクトの最も重要な点である。研究者は、ゲノム配列が決定されていないときには無限の可能性を考慮する必要があったが、ゲノム配列の決定により予測された遺伝子セットをもとに有限の候補の中で種々の生命現象を議論できるようになった。遺伝子の有無で生物を比較することができるようになった。このようなゲノ

ム解析の進展に伴って、網羅性を重視したハイスループット解析が進められており、多くの”omics”がこれにあたる。予測された遺伝子セットを基としたトランスクリプトームおよびプロテオームレベルでの発現解析が可能となる (図 19)。多変量解析は大量のデータを合理的に分類する方法であるから発現プロファイルにより体系的に遺伝子を分類するためには非常に有効な方法である。遺伝子発現における制御関係をトランスクリプトーム解析のみから求めることは実際には難しく、転写因子の DNA 結合部位(cis-element)の探索解析と発現プロファイル解析を統合して解釈する必要がある。このように生命現象の異なった事象を体系的に整理すること、すなわち、ゲノムからフェノーム至る関係を体系化することがバイオインフォマティクスの使命であり、このように異質のデータを統合的扱い分子生物学の事象でそれぞれの要素の役割が体系化できれば真のシステムズバイオロジーとして生物を理解することができると期待される。

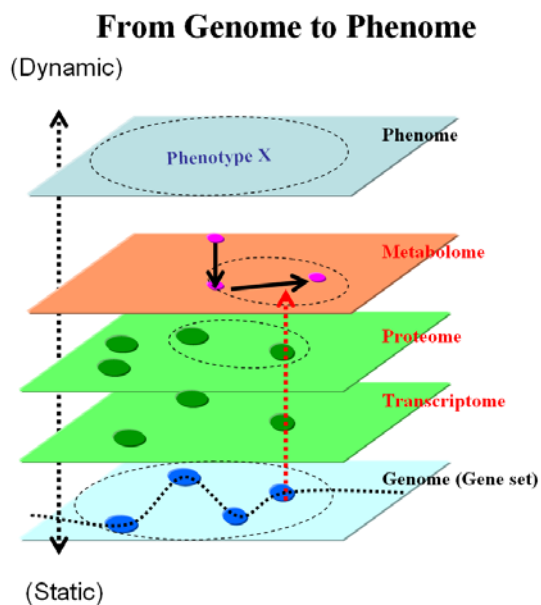


図 19 omics データの階層関係の概念図