In Silico Data
Miracles by the KY-methods

◇第二部　化学多変量解析／パターン認識（ケモメトリックス（Chemometrics））関連(2)

Part2. Chemical multivariate analysis / pattern recognition (Chemometrics)

株式会社　インシリコデータ
湯田　浩太郎

Contents:

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター　：Compound related parameters

■　トポロジカルデータ
分子構造インデックス：原子数（原子種）、結合数（結合種）、リング数、その他
様々なインデックス値：ＨＯＳＯＹＡインデックス、分子結合インデックスＭＣ値
パス値インデックス、

■　トポグラフィカルデータ
化合物の３次元的形状に関するパラメータ
化合物全体構造　：ボックスパラメータ、対称パラメータ、
立体格子パラメータ、その他
化合物部分構造　：ステリモルパラメータ、

■　物理化学データ
分子に関する様々な物性データ　：　分子屈折率、分子量、ＬＯＧＰ、融点、沸点
分子容積、分子表面積、その他
分子軌道法より得られる様々なパラメータ：電子密度、ＨＯＭＯ、ＬＵＭＯ、他
分子力学計算から得られるパラメータ：　種々歪みエネルギー
種々スペクトルより得られるデータ：　種々スペクトルデータ

■　その他のデータ
部分構造パラメータ：　部分構造の有無、部分構造数、
部分構造単位の様々なパラメータ値計算、
演算パラメータ１　：　記述子間の演算により得られるパラメータ（＋－ｘ÷ Log）
演算パラメータ２　：　他の解析手法より算出されたパラメータ
ダミーパラメータ　：　有るパターン存在の有無（１／０）に関するパラメータ

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:トポロジカルパラメーター

Compound related parameters: Topological parameters

□　トポロジカルデータの特徴

トポロジカルデータは化合物を構成する原子と結合とを、それぞれノードとエッジとに定義する。　トポロジー的な問題として化合物構造式を捕らえ、化合物の原子（ノード）間の相互的な関係（つながり状態）を数値データに変換したものである。

このトポロジカルデータの特徴を簡単にまとめると以下のようになる。

長所：　化合物の複雑な結合情報を数値データに変換できるので、通常の数値データでは説明出来ないような複雑な情報を扱う事が可能となる。　この結果、分類能が飛躍的に向上することが期待される。

短所：　数値データの変換ルールと化合物構造式との関係が不明な時が多い。　アルゴリズムが数値データへの変換の為のルールとなっている事が多く、最終目的である目的変数に対する情報の説明や解釈が困難な事が多い。　即ち、分類の為のデータに陥り易く、分類だけが目的の時は強力なパラメータとなりうるが、そのパラメータの持つ意味（情報）を解釈する事が重要となる解析には不向きである。

このトポロジカルデータは現在様々なものが提唱されている。　特に有名なものとして化合物の物性予測に用いられる事の多いＨＯＳＯＹＡ　ＩＮＤＥＸ　と、構造活性相関分野で利用実績の多い分子結合インデックス（Ｍ．Ｃ．）（Molecular Connectivity Index）等が有名である。

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:トポロジカルパラメーター

Compound related parameters: Topological parameters

□　ＭＣＩ値の算出法

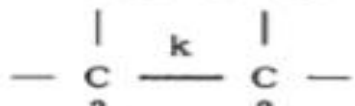
まず化合物を構成している個々の結合についてＣ$_k$値を求める。　続いて、このＣ$_k$値を化合物中の総ての結合について総和した値が分子に対するＭＣＩ値となる。

$$MCI = \sum_{k=1}^{m} C_k = \sum_{k=1}^{m} \frac{1}{[L_i \cdot L_j]^{1/2}}$$

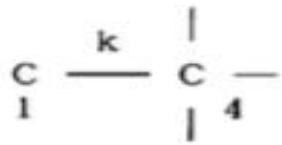
k：　ある一つの結合のＩＤ番号
i：　結合ｋを形成する原子２個のうちの一つの原子に関するＩＤ
j：　結合ｋを形成する原子２個のうちｉ以外の原子に関するＩＤ

上式中、L_iは原子ｉの結合の多重度であり、L_jは原子ｊの多重度を示している。この多重度とは現在注目している原子から飛び出している結合の数を意味し、この時水素原子とつながっている結合の数は無視して計算する。

例）　Ｃ$_k$値の求め方



$$C_k = \frac{1}{[3 \cdot 3]^{1/2}} = \frac{1}{3}$$



$$C_k = \frac{1}{[1 \cdot 4]^{1/2}} = \frac{1}{2}$$

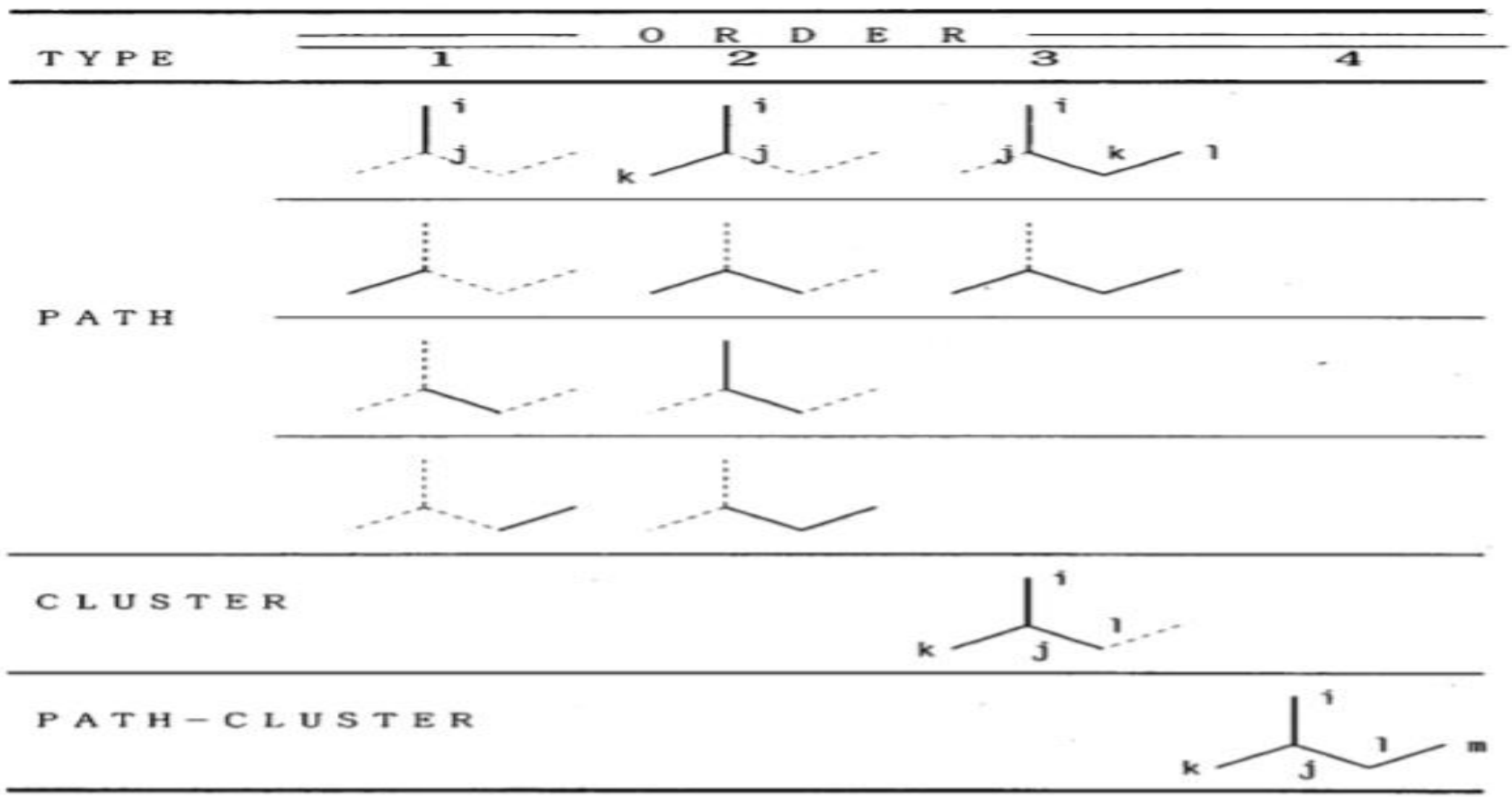
口解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:トポロジカルパラメーター

Compound related parameters: Topological parameters

例）ＭＣＩにおける次数と結合タイプの概念及びＣｘ 計算式



□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:トポロジカルパラメーター

Compound related parameters: Topological parameters

□　ＭＣＩへの結合次数及び結合タイプの導入

Ｃｋ値を求め、この値を基準としてＭＣＩを求める時、化合物構造式の複雑さを情報として取り入れるべく結合次数（ＢＯＮＤ　ＯＲＤＥＲ）という概念と結合タイプ（ＢＯＮＤ　ＴＹＰＥ）という２つの概念を導入する。

・結合次数（ＢＯＮＤ　ＯＲＤＥＲ）はＣｋを求める時の対象となる結合と、その結合を形成する原子の数を拡大してゆくものである。
・結合タイプ（ＢＯＮＤ　ＴＹＰＥ）とは、結合が複数集まって一つのＣｋを形成する時の集合形態に関する情報である。

□　結合次数（ＢＯＮＤ　ＯＲＤＥＲ）について

結合次数は基本となるＣｋ値を求める時に対象とする結合や原子数を規定するものである。　次数が小さければＭＣＩの値は大きく、次数が増大するにつれてＭＣＩの値は小さくなる。

□　結合タイプ（ＢＯＮＤ　ＴＹＰＥ）について

ＴＹＰＥはＣｓとしてまとまった単位（特に次数が大きくなった時）の形を規制するものである。

・ＰＡＴＨは最も単純な形をしており、結合が直線上に繋がっているものを意味する。この時、次数が１のものは直線であり、ＰＡＴＨとみなす。
・ＣＬＵＳＴＥＲは分岐した形状を持つＣｓとなる。　従って、次数が３以上で現れる
・ＰＡＴＨ－ＣＬＵＳＴＥＲはＣｓ内部にＰＡＴＨ部分とＣＬＵＳＴＥＲ部分を持つ。

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:トポロジカルパラメーター

Compound related parameters: Topological parameters

例）　$^{1}\chi_{p}$ ： 結合次数１、ＰＡＴＨタイプのＣ$_{k}$ 値を基本として求めたＭＣＩ値

$^{4}\chi_{pc}$ ： 結合次数４、ＰＡＴＨ－ＣＬＵＳＴＥＲタイプのＣ$_{k}$ 値を基本として求めたＭＣＩ値

$^{1}\chi_{p}$ ： 結合次数１、ＰＡＴＨタイプのＣ$_{k}$ 値を基本として求めたＭＣＩ値にリング補正を加えた値

$^{1}\chi_{pv}$ ： 結合次数１、ＰＡＴＨタイプのＣ$_{k}$ 値の計算にヘテロ原子を考慮して求めたＭＣＩ値

□ 次数（ＯＲＤＥＲ）が異なる時のＣ$_{k}$ の計算式（次数１～４について）

$$\text{次数 1} = \sum_{s=1}^{N_a} (\delta_i \delta_j)_s^{-1/2}$$

$$\text{次数 2} = \sum_{s=1}^{N_b} (\delta_i \delta_j \delta_k)_s^{-1/2}$$

$$\text{次数 3} = \sum_{s=1}^{N_c} (\delta_i \delta_j \delta_k \delta_l)_s^{-1/2}$$

$$\text{次数 4} = \sum_{s=1}^{N_d} (\delta_i \delta_j \delta_k \delta_l \delta_m)_s^{-1/2}$$

口解析に使うパラメーター

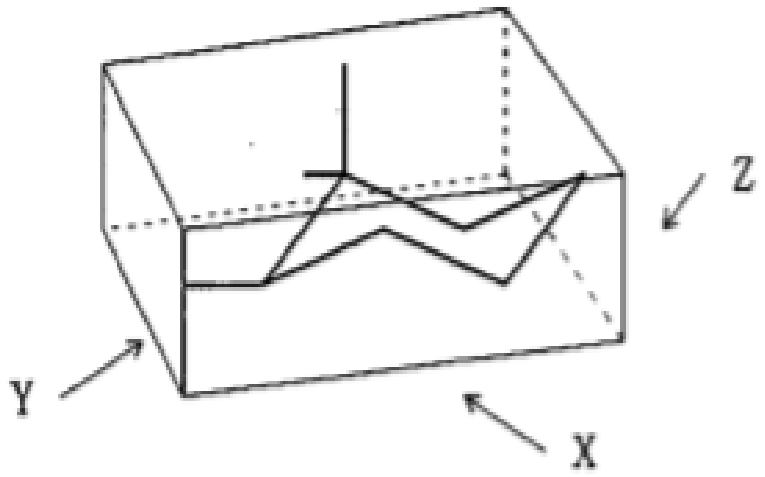
Parameters used for analysis

◆化合物関連パラメーター:トポグラフィカルパラメーター

Compound related parameters: Topographic parameters

② 分子全体の形状に関する幾何学的情報（ボックスパラメータ）

化合物の3次元構造式をそのまま長方形のボックスにいれる。 このボックスの各軸の長さとその比とをパラメータとする。



パラメータ1＝	X
パラメータ2＝	Y
パラメータ3＝	Z
パラメータ4＝	X／Y
パラメータ5＝	X／Z
パラメータ6＝	Y／Z

このパラメータにより、分子全体の立体的な形状についての情報がえられる。 例えば、分子が平面に近い、細長い、立方体に近い等の情報である。

口解析に使うパラメーター

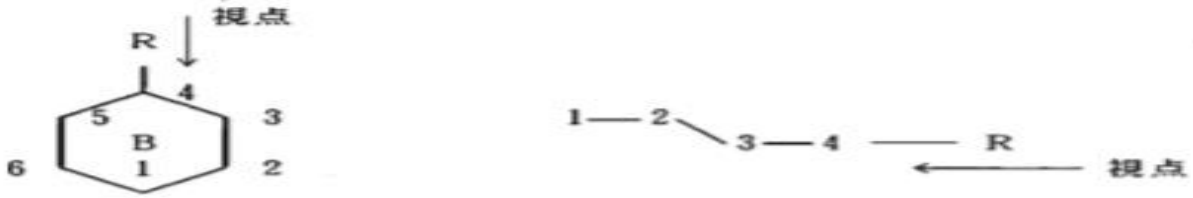
Parameters used for analysis

◆化合物関連パラメーター:トポグラフィカルパラメーター

Compound related parameters: Topographic parameters

①ＳＴＥＲＩＭＯＬ　ＰＡＲＡＭＥＴＥＲ

このパラメータは化合物の置換基Ｒの３次元的立体情報を記述するのに用いられる。特に、重回帰手法によるＨＡＮＳＣＨ／ＦＵＪＩＴＡ法等に用いられて数多くの実績を有する構造活性相関には重要なパラメータである。



パラメータは化合物の基本構造部分（図中１～６で示されるＢ部分）と置換基Ｒ部分とに分けた時、基本構造部分と置換基Ｒ部分とが直結している結合をＲの方からＢに向かって見た時の置換基Ｒの占める空間上の領域をそれぞれの軸方向について分割した時の値を要素データとするものである。

STERIMOL＝（Wl．Wr．Ws．Wn．L）

＝（1.5．2.5．2.0．1.5，4.0）

口解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:物理化学的パラメーター

Compound related parameters: Physicochemical parameters

◇種々の物性:

分子量、 融点、 沸点、分子屈折率、 LogP、 Hammetσ、 その他

◇分子軌道法関連パラメーター:

電子密度、 HOMO、 LUMO、 分極率、 双極子モーメント、その他

◇分子力学関連パラメーター:

結合エネルギー、トーションエネルギー、 水素結合エネルギー、 その他

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:物理化学的パラメーター （LogP）

Compound related parameters: Physicochemical parameters （LogP）

□ LOGPパラメータの定義式

・HANSCH-REOによるフラグメント付加方式によるLOGP値推算。

$$LOGP = LOG \frac{[\ C\ ]\ lipid}{[\ C\ ]\ aqueous}$$

[C] lipid : 平衡状態における油層中の濃度

[C] aqueous : 平衡状態における水層中の濃度

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター:物理化学的パラメーター （LogP）

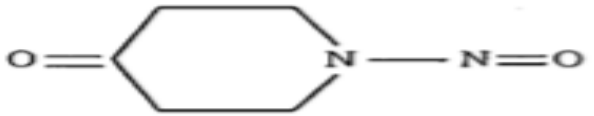
Compound related parameters: Physicochemical parameters （LogP）

① フラグメント付加方式によるＬＯＧＰ推算式

$$LOGP = \sum_{i=1}^{n} a_i f_i + \sum_{j=1}^{m} b_j F_j$$

a_i : i番目のフラグメントの出現回数
f_i : i番目のフラグメントに対するフラグメント定数値
b_j : j番目の修正因子の出現回数
F_j : j番目の修正因子の修正定数値

ＬＯＧＰ値計算例）



4-KETO-N-NITROSO-PIPERIDINE

フラグメント定数

フラグメント	出現回数	フラグメント定数	総　和
— CH_2 —	4	0.66	2.64
ケトン	1	−1.90	−1.90
N−N=O	1	−2.45	−2.45

修　正　定　数

リングボンド	(n−1) (−0.09) = 5 (−0.09)	−0.45
極性基修正	2 X [−(0.20) (f₁ + f₂)]	
	2 X [−(0.20) (−2.45 − 1.90)]	
	2 X [0.87]	1.74
	$LOGP_{CALC}$ =	−0.42
	$LOGP_{obs}$ =	−0.47

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター：その他のパラメーター

Compound-related parameters: Other parameters

◇部分構造パラメーター　Substructure parameter

部分構造を定義し、解析対象とする化合物中に定義した部分構造が含まれているかいないかを利用する Define a partial structure and use whether or not the defined partial structure is included in the compound to be analyzed

部分構造カウントの結果データの表記手法により以下の3種類に大別できる

1. **部分構造があるかないかのバイナリーデータBinary data with or without substructure**
   1または0のバイナリーデータ
   ＊MACCS，PubChem, Daylight等が提供している
2. **内包される部分構造の数をカウント Count the number of substructures included**
   整数値のパラメーター
   ＊フィンガープリントの拡張版
3. **内包された部分構造の隣接原子の情報を加味して数値化**
   Quantify by taking into account the information of neighboring atoms of the included partial structure
   連続変数パラメーター
   ＊ADAPT(Automated Data Analysis by the Pattern recognition)で開発/採用

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター：その他のパラメーター
Compound-related parameters: Other parameters
◇部分構造パラメーター　Substructure parameter

1．部分構造があるかないかのバイナリーデータ
Binary data with or without substructure
1または0のバイナリーデータ
＊MACCS，PubChem, Daylight等が提供
＊この種のパラメーターは一般的に「フィンガープリント"Fingerprint"」と呼ばれる
＊パラメーター数は数百から千種類提供される
＊検索される部分構造は提供元の適用目的等により変化する

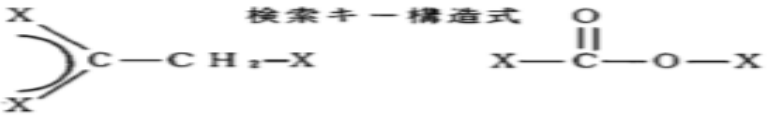
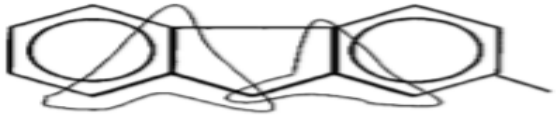
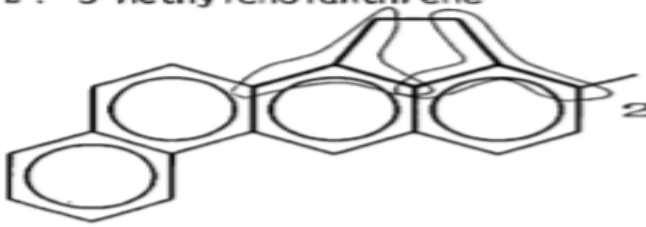
□適用目的
＊多変量解析／パターン認識による解析のパラメーターとして適用
＊化合物の類似度評価等に利用される事が多い

口解析に使うパラメーター

Parameters used for analysis

2. 内包される部分構造の数をカウント Count the number of substructures included
3. 内包された部分構造の隣接原子の情報を加味して数値化

Quantify by taking into account the information of neighboring atoms of the included partial structure

被検索化合物	検索キー構造式 			
	部分構造数	MCI	部分構造数	MCI
1. 2-Aminofluorene 	2	3.166	0	0.0
2. 3-Methylcholanthrene 	2	3.301	0	0.0
3. Safrole 	1	2.690	0	0.0
4. Ethyl acetate 	0	0.0	1	1.904

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター：その他のパラメーター
Compound-related parameters: Other parameters
◇部分構造パラメーターの特徴　Feature of substructure parameter

特徴 characteristics：

①部分構造パラメーターは化合物構造式に直結するパラメーターのため、
他のパラメーターと比較して要因抽出がしやすい
②部分構造パラメーター情報は、薬理活性や毒性等のコントロールに必要な情報を
化合物の構造情報として捉えることが出来る
③部分構造パラメーターの情報は化合物合成等に反映しやすい
④データ解析のみならず、化合物検索等にも適用できる

①Since partial structure parameters are directly linked to the compound structural formula, it is easier to extract the factors than other parameters
(2) Partial structure parameter information can capture information necessary for control of pharmacological activity, toxicity, etc. as structural information of compounds.
(3) Information on partial structure parameters is easily reflected in compound synthesis, etc.
(4) Not only data analysis but also compound search

□解析に使うパラメーター

Parameters used for analysis

◆化合物関連パラメーター：その他のパラメーター
Compound-related parameters: Other parameters
◇部分構造パラメーターの特徴　Feature of substructure parameter

留意点：Points to remember
①設定する部分構造の内容が実施目的の実現性に大きな影響を及ぼす
例：薬理活性では、ファーマコフォア等に留意した部分構造
　化合物毒性では、毒性要因に関連する部分構造
②部分構造の設定にノウハウや経験がある程度実施結果に影響することがある

(1) The content of the partial structure to be set has a great influence on the feasibility of the implementation purpose
Example: In pharmacological activity, partial structure with attention paid to pharmacophore, etc.
In compound toxicity, partial structure related to toxicity factors
(2) Know-how and experience may affect the implementation results to some extent in setting the partial structure

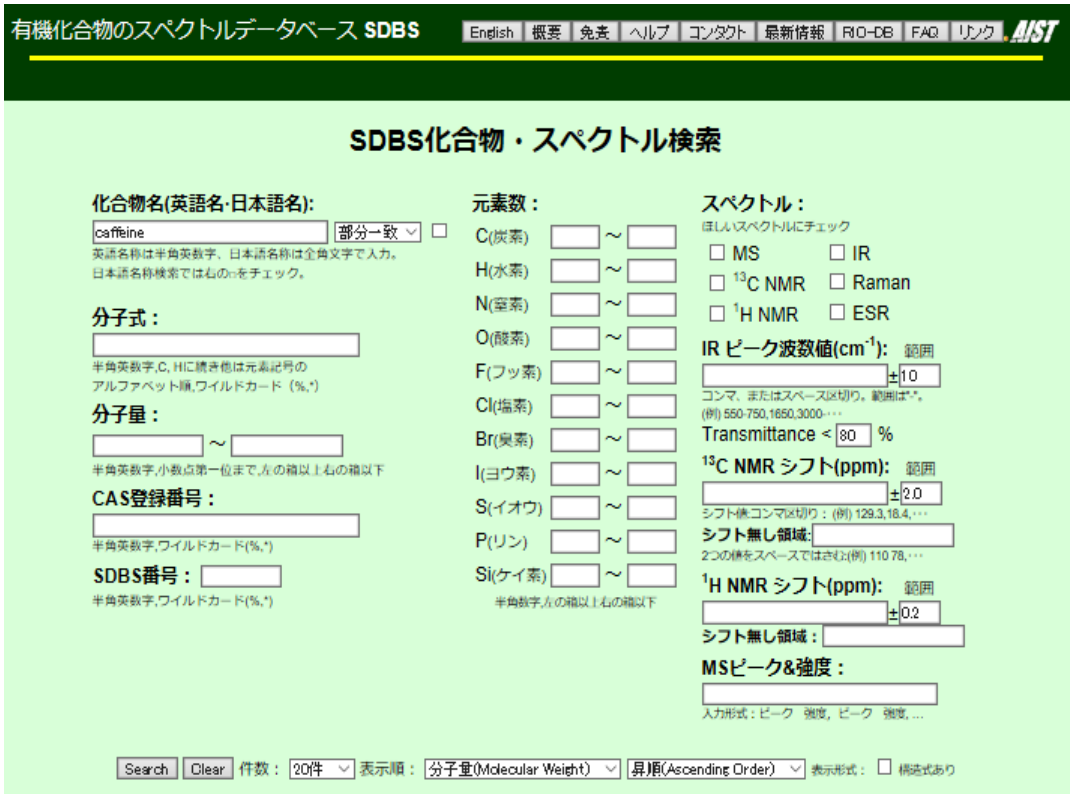
□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

■ Spectral database of organic compounds　**SDBS**

トップ画面



https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

■有機化合物のスペクトルデータベース SDBS

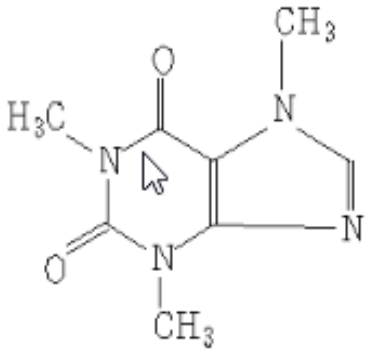
Spectral database of organic compounds **SDBS**

SDBS No: 1898 **CAS Registry No.:** 58-08-2

DOI:

Molecular Formula: $C_8H_{10}N_4O_2$ **Molecular Weight:** 194.2

SDBS-NO= 1898
CAFFEINE



Compound Name:

caffeine
1,3,7-trimethyl-3,7-dihydro-1H-purine-2,6-dione
1,3,7-trimethyl-3,7-dihydro-purin-2,6-dion, kaffein
1,3,7-trimethyl-3,7-dihydro-purine-2,6-dione
1,3,7-trimethylxanthine
1H-purine-2,6-dione, 3,7-dihydro-1,3,7-trimethyl-
3,7-dihydro-1,3,7-trimethyl-1H-purine-2,6-dione
theine

InChI:

InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

InChIKey:

RYYVLZVUVIJVGH-UHFFFAOYSA-N

Publisher:

National Institute of Advanced Industrial Science and Technology (AIST)

https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

□解析に使うパラメーター

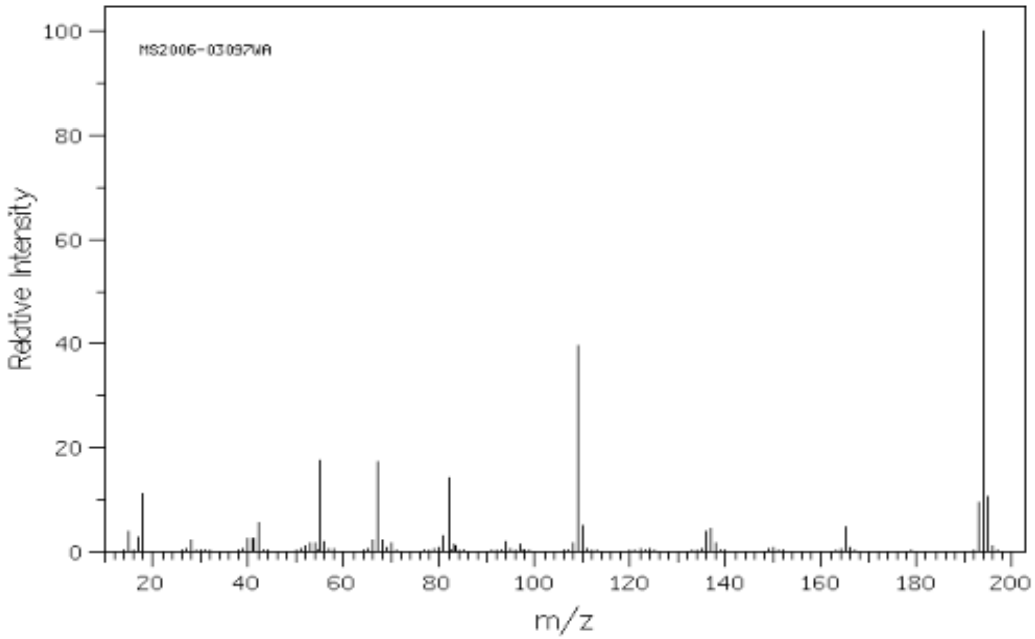
Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

Mass

SDBS-Mass

MS2006-03097WA　　SDBS NO.　1898
caffeine
C8H10N4O2　　(Mass of molecular ion:　194)



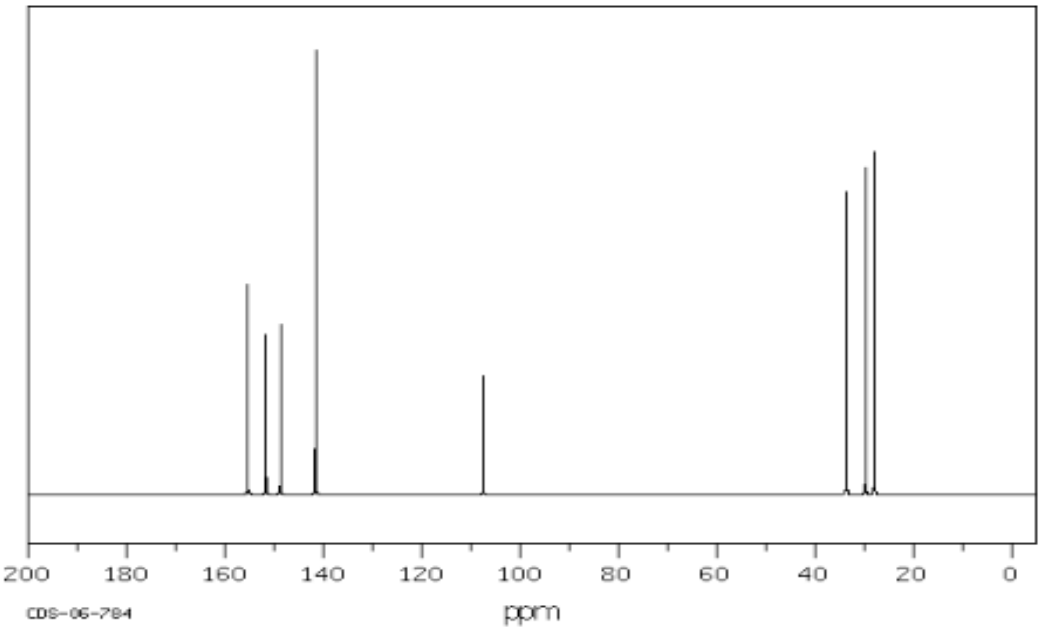
https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

^{13}C NMR : in $CDCl_3$



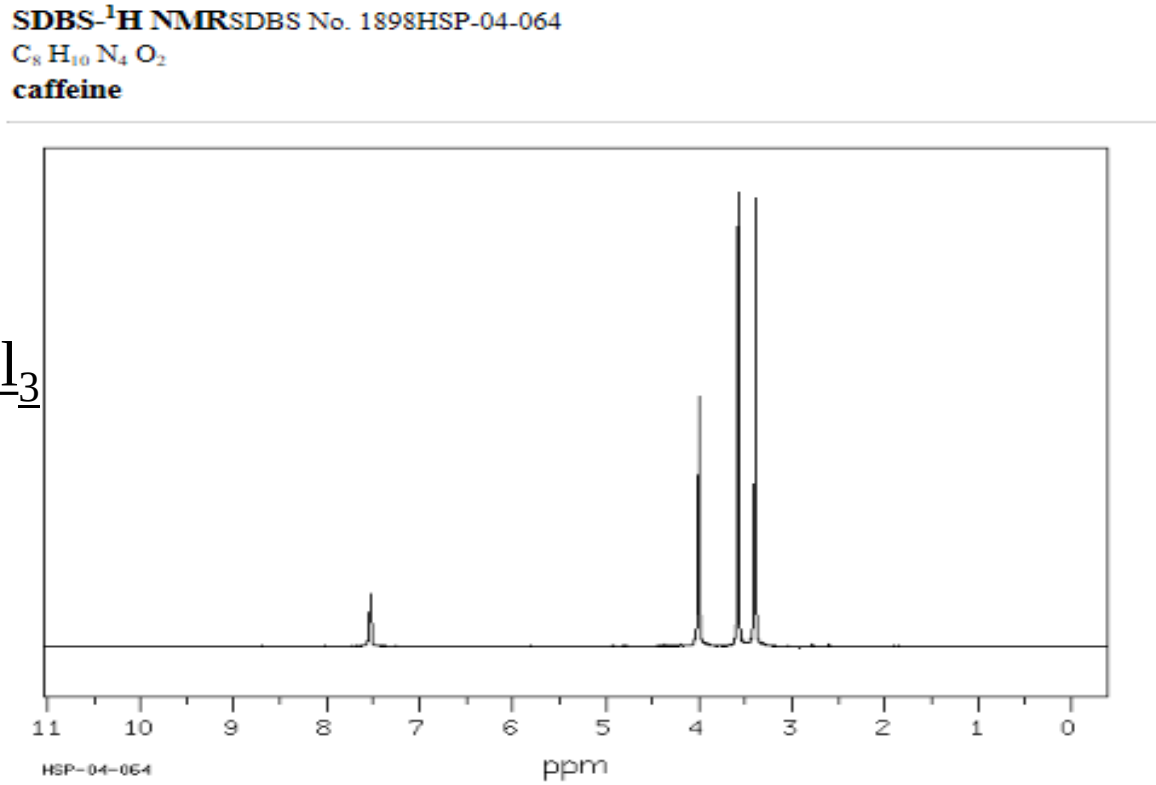
https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

^{1}H NMR : 90 MHz in $CDCl_3$



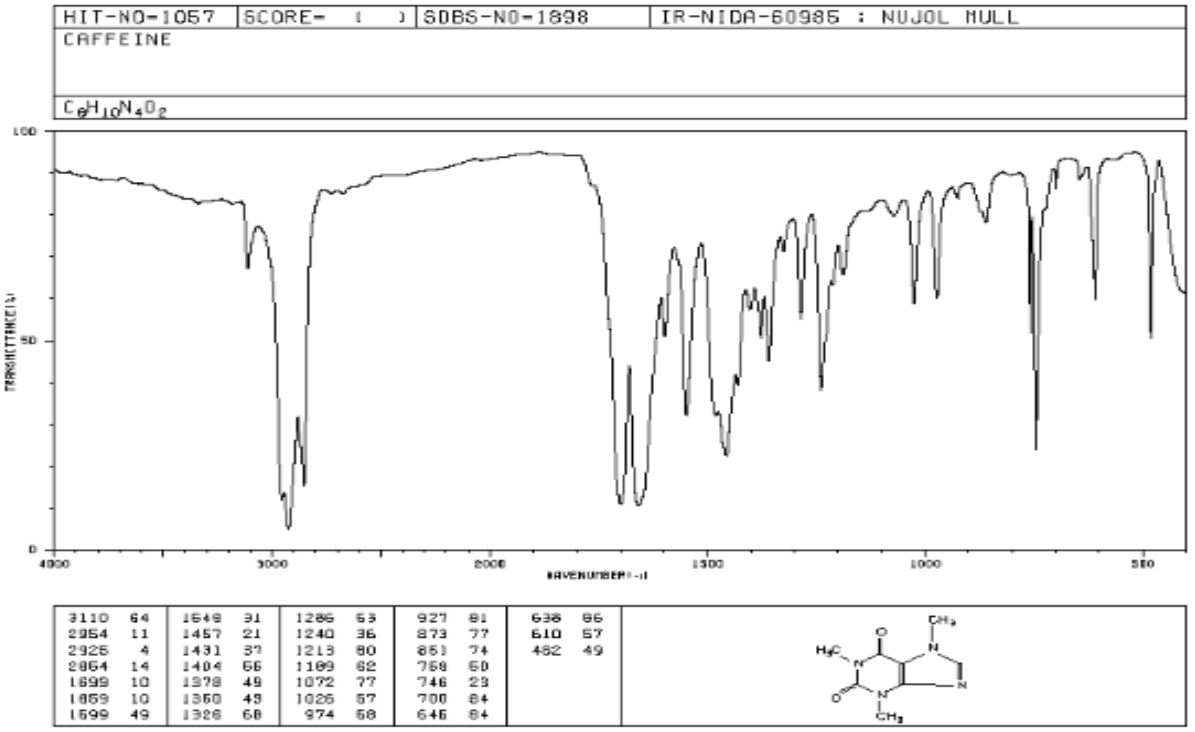
https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

口解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

IR：nujol mull



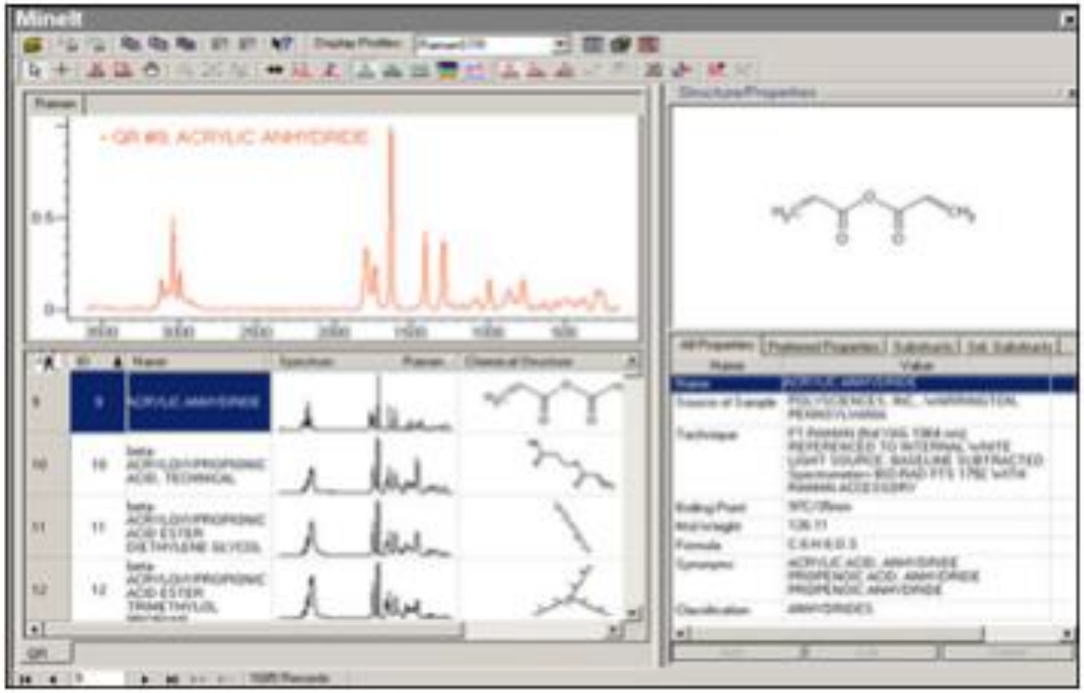
https://sdbs.db.aist.go.jp/sdbs/cgi-bin/direct_frame_top.cgi

口解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

Ramanデータ



BIO-RADのスペクトルデータベースより

http://www.bio-rad.com/

http://www.bio-rad.com/ja-jp/product/raman-spectral-databases?ID=N0ZXPS4VY

□解析に使うパラメーター

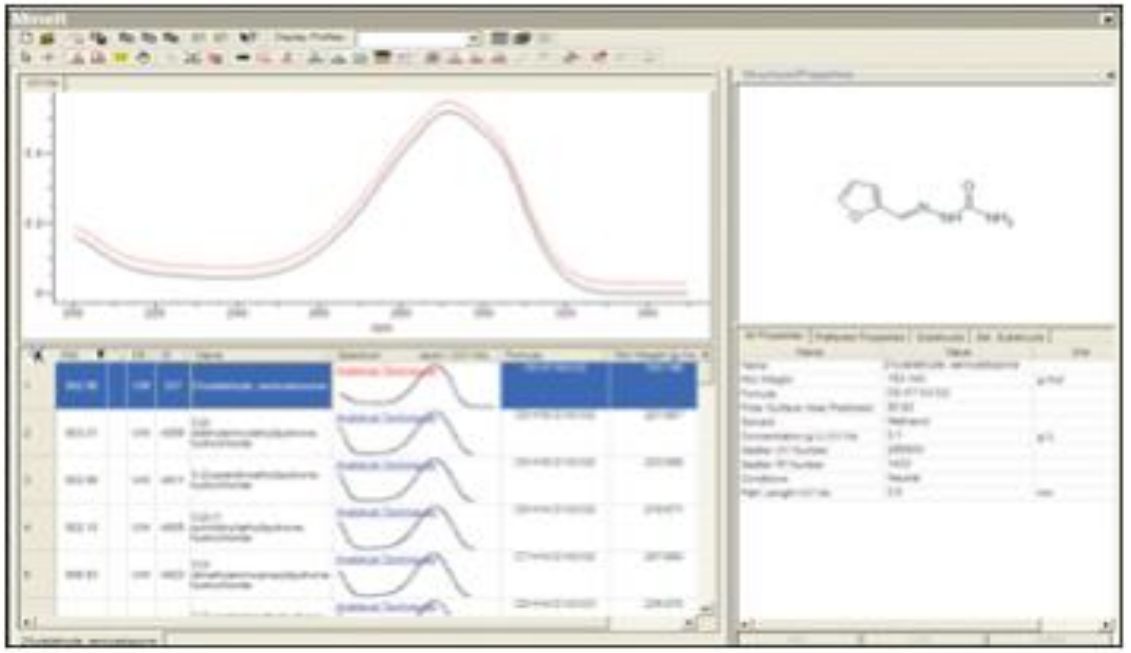
Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

UV-Visスペクトル

BIO-RADのスペクトルデータベースより

http://www.bio-rad.com/



http://www.bio-rad.com/ja-jp/product/uv-vis-spectral-databases?ID=NH262L4VY

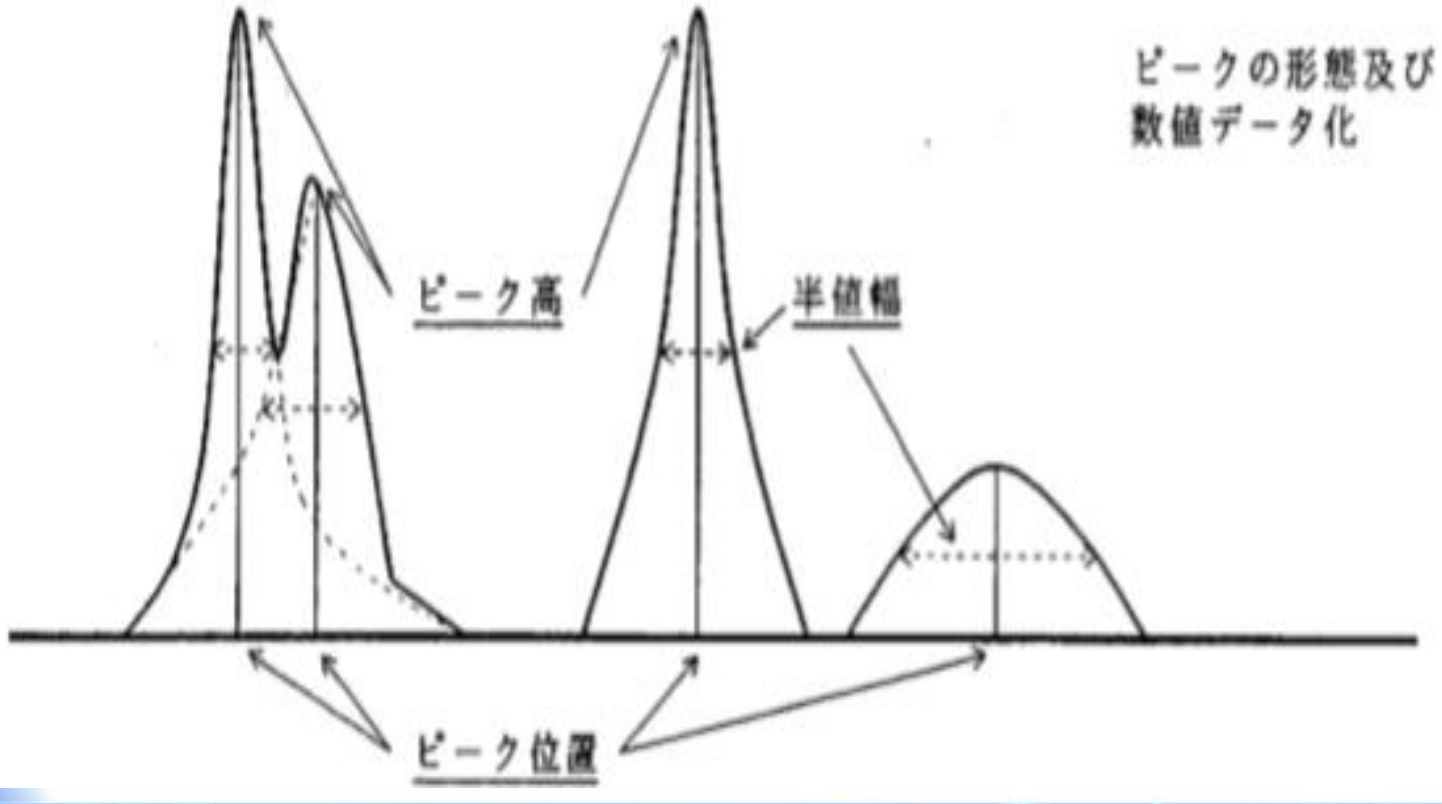
□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

＊機器スペクトルデータは殆どの場合アナログデータなので、デジタルの数値データに変換することが必要である。

Since instrument spectral data is mostly analog data, it must be converted to digital numerical data.



□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

特徴 Characteristic: :

①機器があれば数値データとして簡単に蓄積できる

If there is a device, it can be easily stored as numerical data

②スペクトルチャートは様々な実験過程で種々蓄積される

Spectrum charts are accumulated in various experimental processes.

留意点　Points to remember: :

①一般的にパラメーター数が極めて大きくなりやすい

結果として、データ解析手法が限定、適用不可となる可能性が高くなる

②スペクトルデータは多重共線性が極めて高い

①と②の特徴により、データ解析実施においては次元圧縮・統合等が必要で、解析手法もPLSやPCAと制限されることが多い

③スペクトルチャートの測定条件等統一が必要

出来れば測定機器メーカーや機種も統一

例：60MのH-NMRデータと90MのH-NMRデータを混在してのデータ解析は無意味

・測定条件等が統一されないとデータ解析の精度が保証されにくくなる

□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

＊例え同じスペクトルで、測定環境条件を同じにしたとしても測定機器やソフトウエアの違いにより多種多様のフォーマットがある。
＊スペクトルデータベース間でのデータのやり取りにはファイルフォーマト変換が重要

NMR

Vendor	File Format	Required Parameter Files	Optional Parameter Files
ACD/Labs	*.spectrus, *.esp, *.txt		
Acorn NMR, Inc.	*.fid, *.nmr, *.2d		
Agilent (Varian, Inc.)	data, *.fdf, fid0001.fdf, *.txt, fid, phasefile	acq, proc, procpar	acq_2, text
ASCII†	*.txt; *.prn, *.csv, *.asc		
Bruker Corporation	ser, rr, .fid, *r, 1i, 2rr, *.* (DISNMR)	acqus, procs, acqu2, proc2s, *.fqs, *.fa1, *.fa2, *.fp1, *.fp2	title, intrng, *.tit, *.ti2
GE	*.raw, *.* (Nicolet)		
JCAMP†	*.dx; *.jdx		
JEOL Ltd.	*.als, *.jdf, *.nmfid, *.nmf, *.nmdata, *.nmd, *.gxd, *.bin, *.* (Delta)	*.gxp, *.hdr	exp.param, exp.par
Lybrics	*.*		
MSI Felix	*.*		
Tecmag	*.tnt, *.* (MacNMR)		
Thermo Scientific†	*.spc		

https://www.acdlabs.com/products/fileformats/

□解析に使うパラメーター

Parameters used for analysis

◇ 機器スペクトルパラメーター　Instrument spectral parameters

Chromatography

Vendor	Data Format	Extension	Comments
Agilent Technologies	1100 Series LC/MSD Quad and Ion Trap Systems	*.ms, *.yep	UV, LC-UV and LC-MS
	ChemStation Rev. B.02.01, B.03.01, B.04.01, B.04.02, B.04.03, Rev. C.01.04	*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	Open Lab C v.1.04	*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	Open Lab Rev. C.01.07	*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	EZChrom	*.dat	UV traces only.
AB SCIEX	[illegible]	*.wiff	LC-UV and LC-MS
Bruker Daltonics and Agilent Technologies			LC-UC and LC-MS
Bruker	Compass (accurate mass data)	*.D	LC-MS, LC-UV, UV Entire *.D folder structure should be used.
Shimadzu Corporation	LCMS-IT-TOF	*.lcd	LC-MS and LC-UV. Requires vendor software on same computer.
	LCMSsolution	*.qld	LC-MS, LC-UV and UV traces May require vendor software on same computer
Thermo Scientific	Xcalibur	*.raw	LC-MS, LC-UV and UV traces
	Chromeleon® 6		UV and LC-UV, via Connect to

Mass Spectrometry

Vendor	Data Format	Import	Export	Extension	Comments
ACD/Labs	ACD/Labs	✔	✔	*.spectrus, *.esp	
Agilent Technologies	1100 Series LC/MSD Quad and Ion Trap Systems	✔		*.ms, *.yep	DAD data and single chromatogram curve are imported also. Splitter available
	ChemStation	✔		*.ms	Splitter available
	LC TOF	✔		*.wiff	
	MassHunter (6000 Series)	✔		*.bin	Entire *.D folder structure should be used. Agilent component requires Microsoft .NET version 2. DAD can be imported (V12) and MS/MS split controlled in newer versions
	Open Lab C v.1.04	✔		*.D	UV, LC-UV and LC-MS Entire *.D folder should be used *.ms, *.ch, *.uv
	Open Lab Rev. C.01.07	✔		*.D	UV, LC-UV and LC-MS Entire *.D folder should be used
AB SCIEX	Analyst	✔		*.wiff	Single mass spectra, LC-MS and most LC-MS[n] imported. Splitter available. UV data not currently imported
		✔		See above	LightSight—Spectra are 'pushed' via ACD/Labs (NetCDF)
	Analyst QS	✔		*.wiff	Single mass spectra, LC-MS and most LC-MS[n] imported. Splitter available
	Analyst TF	✔		*.wiff	
Applied Biosystems	Mariner Data Explorer ASCII LC/MS	✔		*.txt	LC-MS data only
Bruker Daltonics and Agilent Technologies	Agilent or Bruker LC/MS Ion Trap	✔		*.yep	LC-MS and DAD data

Optical Spectropy

Vendor	Data Format	File Format	Comments
ACD/Labs	ACD/Labs	*.spectrus, *.esp	
Agilent Technologies	HP 84552A	*.wav	
	ChemStation†	*.uv	
ASCII single, dual and multicolumn		*.txt; *.prn; *.csv; *.asc	
Bruker	OPUS	*.*	
DeltaNu		*.spc	
Dionex	Chromeleon†		"Connect to" ability available
Foss NIRSystems		*.da	
JASCO Corporation	J-700	*.js	
JCAMP, JCAMP multispectra		*.dx; *.jdx	
LabControl		*.uvd; *.irs	
MATLAB DSO†		*.mat†	
Ocean Optics			
PerkinElmer Instruments			
Shimadzu	IR	*.irs	
Thermo Scientific	Galactic	*.spc	
	Mattson	*.*	
	Nicolet OMNIC	*.spa; *.spg	
Varian	Cary UV	*.b*; *.d*	
Waters Corporation†	Empower and Empower 2		"Connect to" ability available
	MassLynx	*.inf	
	Millennium®†		"Connect to" ability available

https://www.acdlabs.com/products/fileformats/

□解析に使うパラメーター

Parameters used for analysis

◆ 医療関連データ Medical data

＊医療分野でのデータ解析はデータのとり方で様々な解析を実施できる

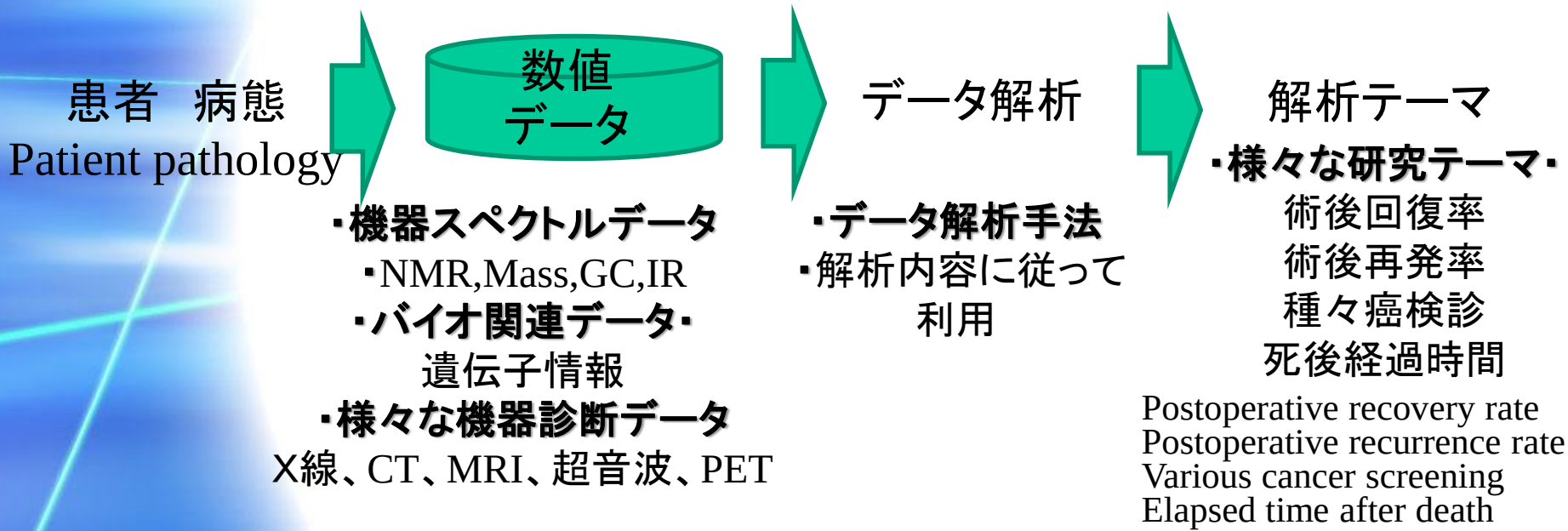
＊現在は医療関連は統計解析を適用し、様々な病気との因果関係や薬効検証等が中心

＊今後は**アイディア次第で様々な解析**を実施できるようになる

- Data analysis in the medical field can be performed in various ways depending on how the data is collected.

* Currently, statistical analysis is applied to medical-related, focusing on causal relationships with various diseases and drug efficacy verification

* In the future, various analyzes will be possible depending on the idea.



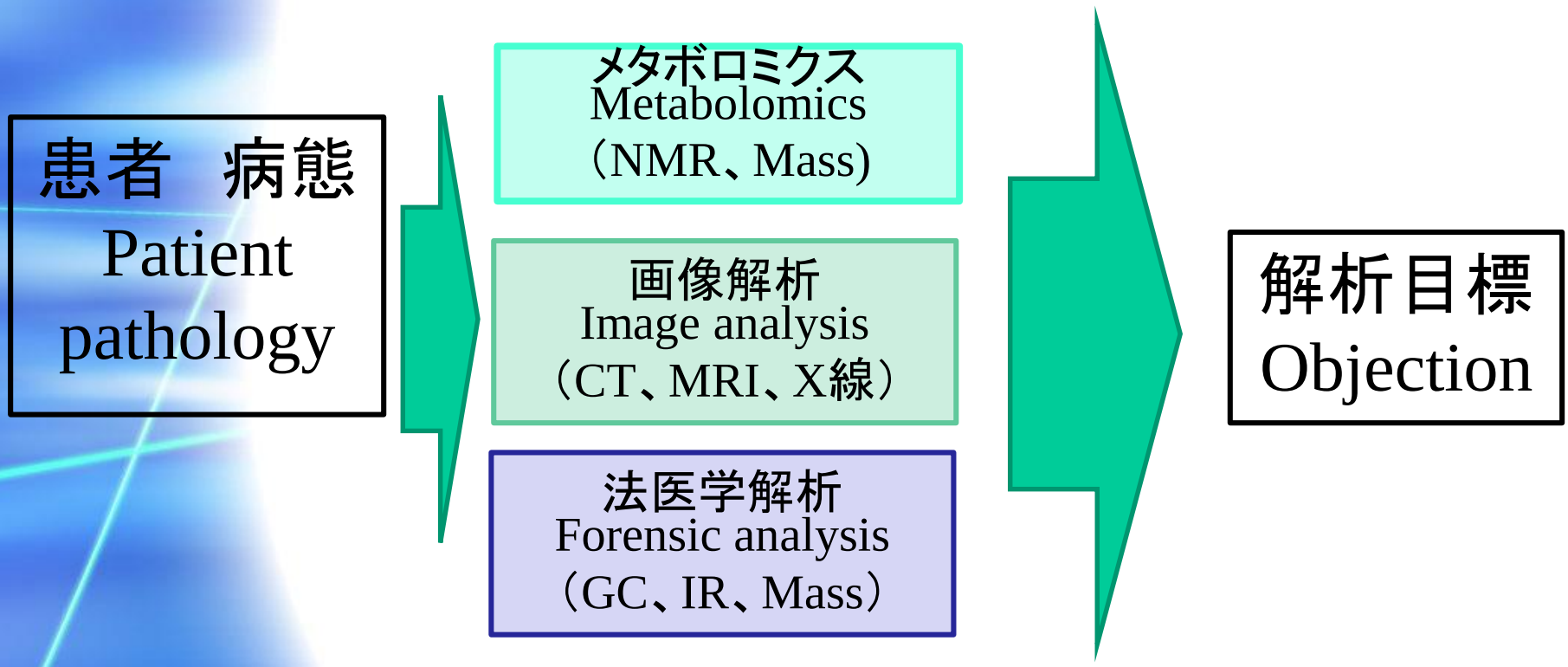
口解析に使うパラメーター

Parameters used for analysis

◆ 医療関連データ Medical data

研究分野により名前や分析機器の種類が変わる

The name and type of analytical instrument varies depending on the research field



□解析に使うパラメーター

Parameters used for analysis

□パラメーターの操作や選択に関連する事項

Matters related to parameter operation and selection

1. パラメーターの単位等に関する留意事項

Notes on parameter units, etc.

2. パラメーター選択について

About parameter selection

3. サンプルの選択について

About sample selection

4. データ解析を保証するための制限事項等

Restrictions to ensure data analysis

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの正規化　Parameter normalization

桁数が大きく異なるパラメーターをデータ解析時に混在すると、要因解析等を困難にする要因となる。
このようなパラメーター群を利用する場合は、通常オートスケーリングを実施する。
オートスケーリングは数値データを平均0で標準偏差1のパラメーターに変換する技術である。

If parameters with significantly different numbers of digits are mixed during data analysis, it becomes a factor that makes factor analysis difficult.
When using such a parameter group, auto scaling is usually performed.
Autoscaling is a technique for converting numerical data into parameters with an average of 0 and a standard deviation of 1.

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの正規化　Parameter normalization

■桁数の異なるパラメーター事例　Parameter examples with different numbers of digits
以下には値の桁数が異なるパラメーターや、正／負の両方を取る値等列挙する
Below are listed parameters with different numbers of values, values that take both positive and negative values, etc.

①桁数の大きくなるパラメーター；分子量（数百から千）、
②値が一桁単位から二けた単位：原子数や環の数等
③値が一桁程度：バイナリーデータ、フラグメントデータ、
④値が小数点以下となる：分子軌道法関連パラメーター
（電子密度、HOMO/LUMO、自由エネルギー、超分極率、ヤング率、その他）
⑤値が生と負の両方を取る：LogPパラメーター、

① Parameters with a large number of digits; molecular weight (hundreds to thousands)
②Values are in single-digit to double-digit units: number of atoms, number of rings, etc.
③ The value is about one digit: binary data, fragment data,
④ Value is below the decimal point: Parameters related to molecular orbital method
(Electron density, HOMO / LUMO, free energy, hyperpolarizability, Young's modulus, etc.)
⑤ Value takes both raw and negative: LogP parameter,

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの正規化：オートスケーリング

Parameter normalization: autoscaling

①解析時に用いるパラメーターは、桁数が揃っていることが望ましい

②同じ種類のデータを用いる時は桁数がそろっている場合が多いので問題ないが、化学分野で扱うデータは内容が多種多様だけでなく、その単位（桁数）も大きく異なる場合が多い

(1) It is desirable that the parameters used for analysis have the same number of digits.

(2) When using the same type of data, there are many cases where the number of digits is the same, so there is no problem, but the data handled in the chemical field is not only diverse but also the units (number of digits) are often very different.

$$Q = \sum_{i=1}^{n} (W_i - \overline{W}) \qquad (1)$$

$$W_i' = \frac{W_i - \overline{W}}{Q} \qquad (2)$$

Qは用いるデータの変量を示す。 W_i'はオートスケーリング後のWの値の内k番目のW_iの値、$\overline{W}$は用いる記述子の平均値である。

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択
（特徴抽出：Feature selection or Parameter selection）

＊データ解析の信頼性を保つためにはサンプル数をパラメーター数で割った値である、信頼性指標を守ることが求められる。
このため、少ないサンプル数の場合はパラメーター数を減少させることが必要。

*現在のケモメトリックス解析では、先に述べた化学関連パラメーターはプログラムにより1サンプル（化合物）あたり数千パラメーター発生することが出来る。

＊化学関連研究ではサンプル数が少ないことが多い。このため、解析精度を保つためにもパラメーター数を減少する特徴抽出が極めて重要である。

＊解析に重要なパラメーターは、Intrinsic Parameter、反対がNon-intrinsic parameter。

* To maintain the reliability of data analysis, it is required to observe the reliability index, which is the value obtained by dividing the number of samples by the number of parameters.
For this reason, it is necessary to reduce the number of parameters when the number of samples is small.

* In the current chemometric analysis, the above-mentioned chemistry-related parameters can generate thousands of parameters per sample (compound) by the program.

* In chemistry-related research, the number of samples is often small. For this reason, feature extraction that reduces the number of parameters is extremely important in order to maintain analysis accuracy.

* Intrinsic parameters are important parameters for analysis, and non-intrinsic parameters are the opposite.

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択
（特徴抽出：Feature selection or Parameter selection）

＊データ解析手法によってはPLSのようにパラメーター数を形式的に減少させることが出来る手法があり、パラメーターを多数発生できる化学分野ではよく利用される。
しかし、PLSは一種の次元変換／圧縮手法で、パラメーターが多いとき緊急避難的に適用される手法。データ解析で重要な要因解析は出来なくなり、分類率や予測率もキレが良くないことになる。

* Some data analysis methods, such as PLS, can formally reduce the number of parameters, and are often used in the chemical field where many parameters can be generated. However, PLS is a kind of dimensional transformation / compression technique that is applied as an emergency evacuation when there are many parameters. It is impossible to analyze important factors in data analysis, and the classification rate and prediction rate are not good

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択

（特徴抽出：Feature selection or Parameter selection）

特徴抽出手法は大きく以下に示される4種類に分類される

①パラメーターを構成する値の特徴と、統計的特性を利用した特徴抽出

Feature extraction using parameters that make up parameters and statistical characteristics

0値チエック、 同値データ出現率、フィッシャー比

＊ニクラス分類および重回帰（フィッテイング）の両方で利用可能

②パラメーター間の相互関係に注目した特徴抽出（相関係数によるアプローチ）

Feature extraction focusing on correlation between parameters (correlation coefficient approach)

単相関、多重相関

＊ニクラス分類および重回帰（フィッテイング）の両方で利用可能

③個々のデータ解析手法の特徴を利用した特徴抽出

Feature extraction using features of individual data analysis methods

個々のデータ解析手法の特徴や機能を用いることで特徴抽出を行う

＊データ解析の種類によりニクラス分類や重回帰に適用される

④最適化等の手法を利用することで特徴抽出を行う

Perform feature extraction by using techniques such as optimization

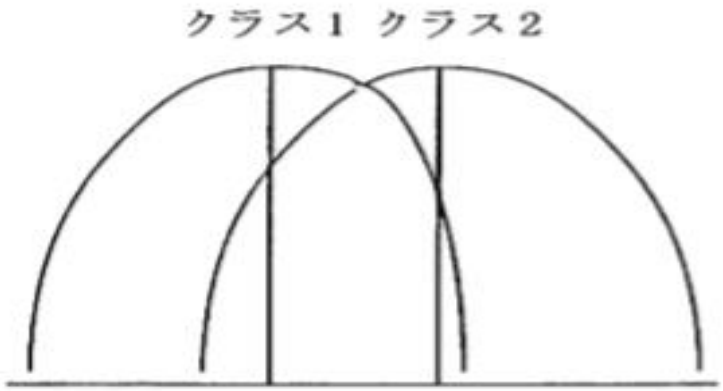
遺伝的アルゴリズム等が利用され、ニクラス分類や重回帰に適用される

□解析に使うパラメーター

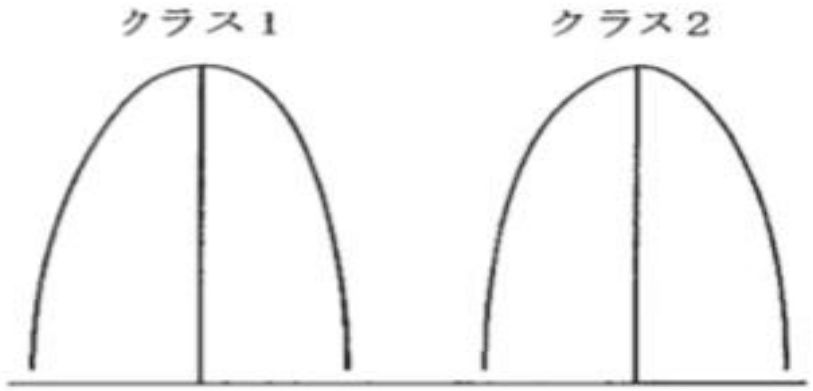
Parameters used for analysis

◇パラメーターの選択：フィッシャー比

Parameter selection: Fisher ratio



フィッシャー比小（分類能小）



フィッシャー比大（分類能大）

図　．フィッシャー比の値と二クラス分類に対するパラメータの関係
右図は一個のパラメータのみで100%分類が可能な例である

$$F = (\bar{X}1 - \bar{X}2)^2 / (\sigma 1^2 - \sigma 2^2)$$

$\bar{X}1$：パラメータのクラス１の平均値
$\bar{X}2$：パラメータのクラス２の平均値
$\sigma 1$：パラメータのクラス１の分散
$\sigma 2$：パラメータのクラス２の分散

口解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：バリアンスウエイト法
Parameter selection: Variance weight method

パーセプトロンの特性を利用した特徴抽出手法
Feature extraction method using characteristics of perceptron

$$VW_j = \frac{V_j}{\overline{W_j}} \quad \cdot$$

$$V_j^2 = \frac{1}{(n_k - 1)} \sum_{k=1}^{n} (W_{jk} - \overline{W_j})^2 \quad \cdot$$

式中jはパラメータの、kはウエイトベクトルのインデックスである。$\overline{W_j}$はj番目のウエイトベクトルの平均値、n_kは用いたウエイトベクトルの数である。

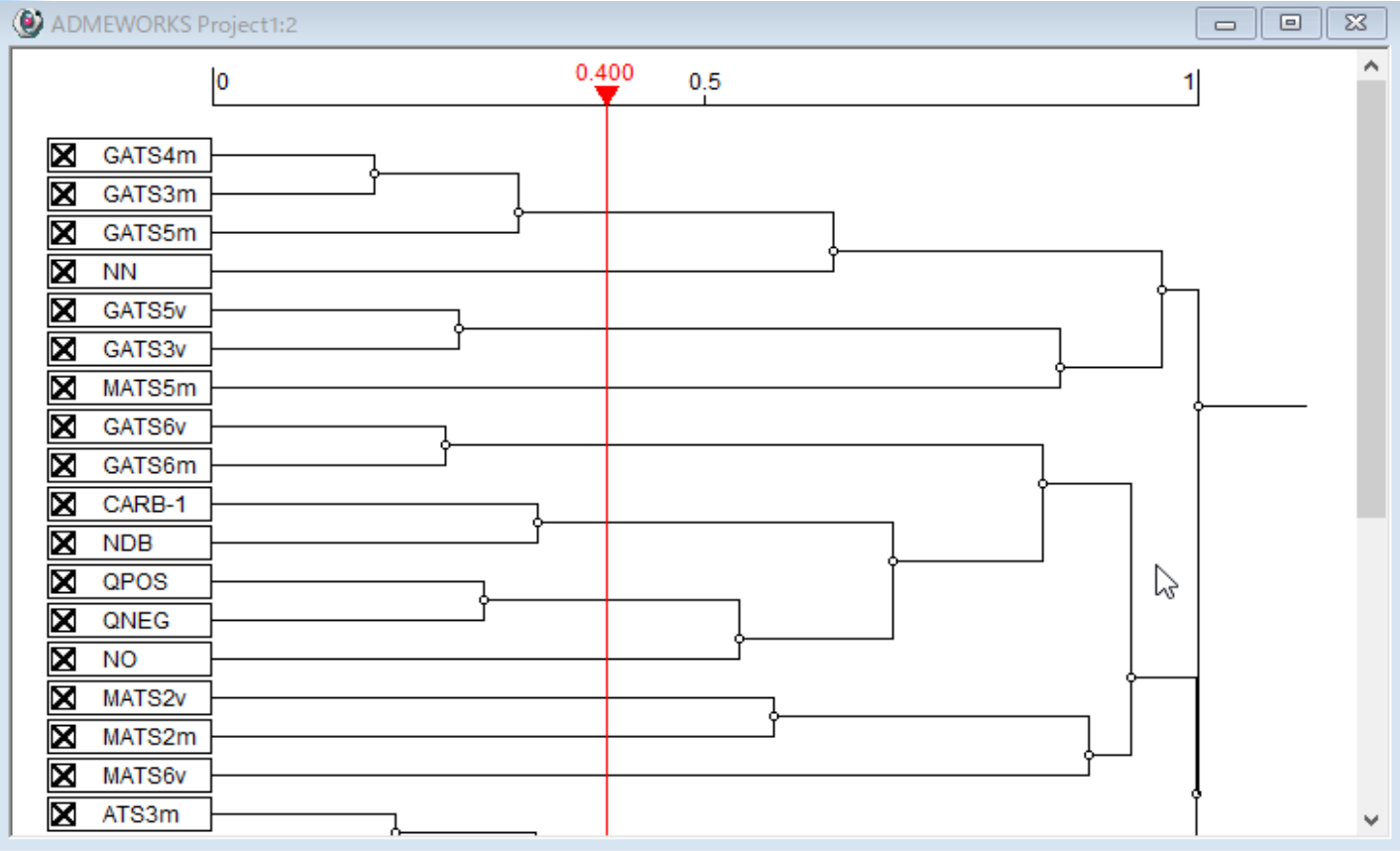
＊1）Jurs P.C. et al., J.C.I.C.S,

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：クラスタリング

Parameter selection: Clustering



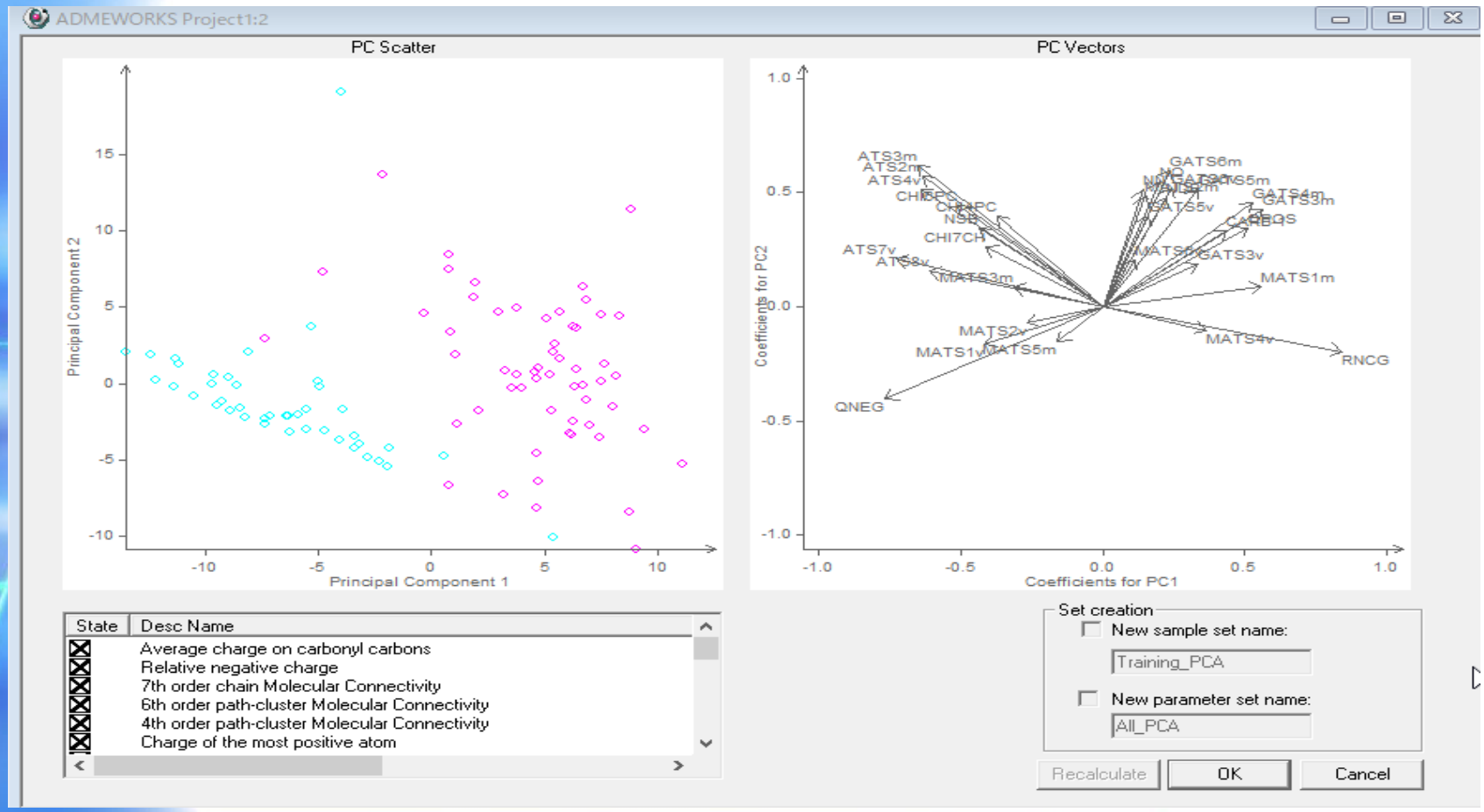
ModelBuilderの画面より

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：主成分分析法（ＰＣＡ）

Parameter selection: Principal component analysis (PCA)



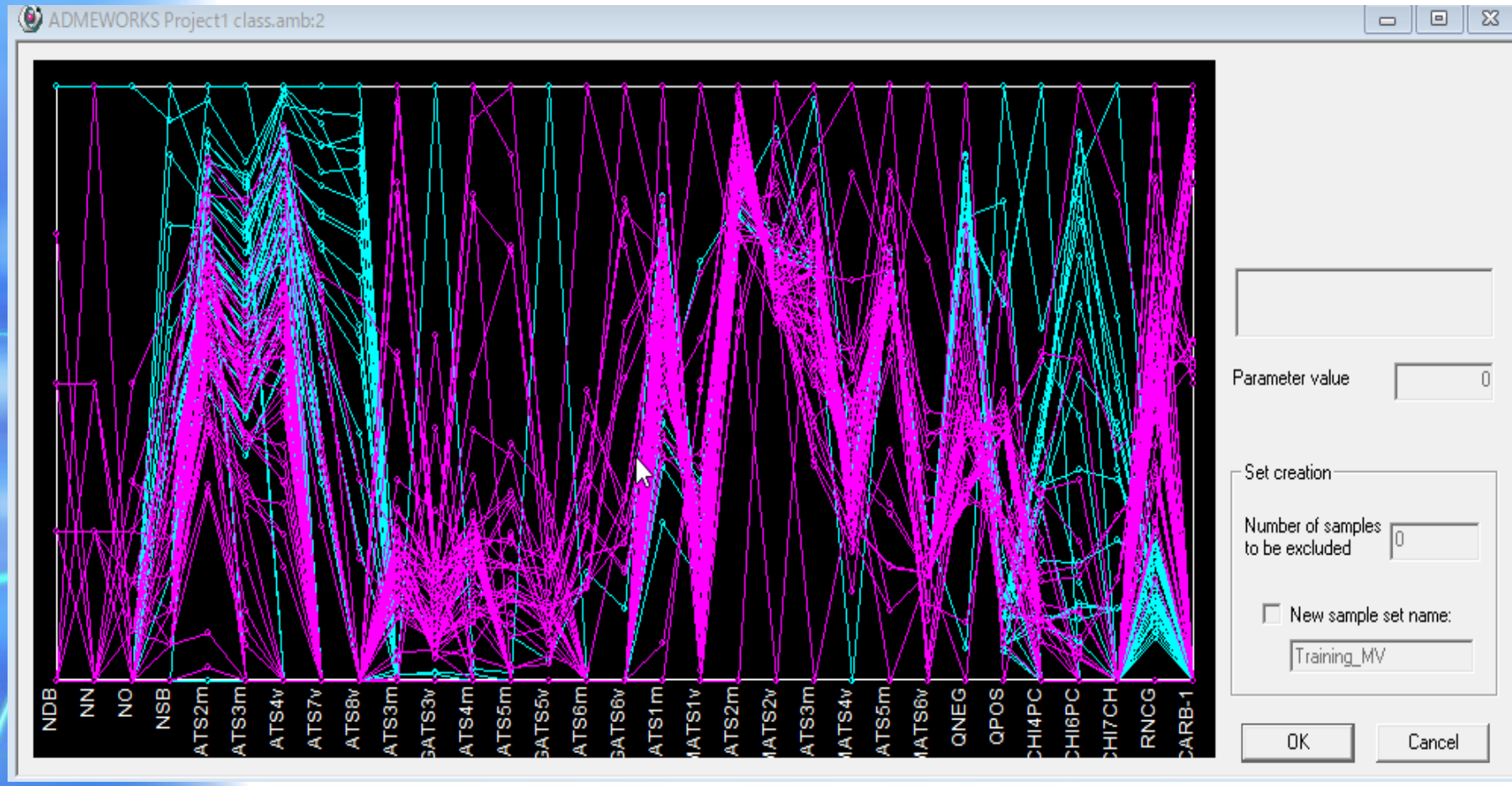
ModelBuilderの画面より

口解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：ラインチャート

Parameter selection: Line chart



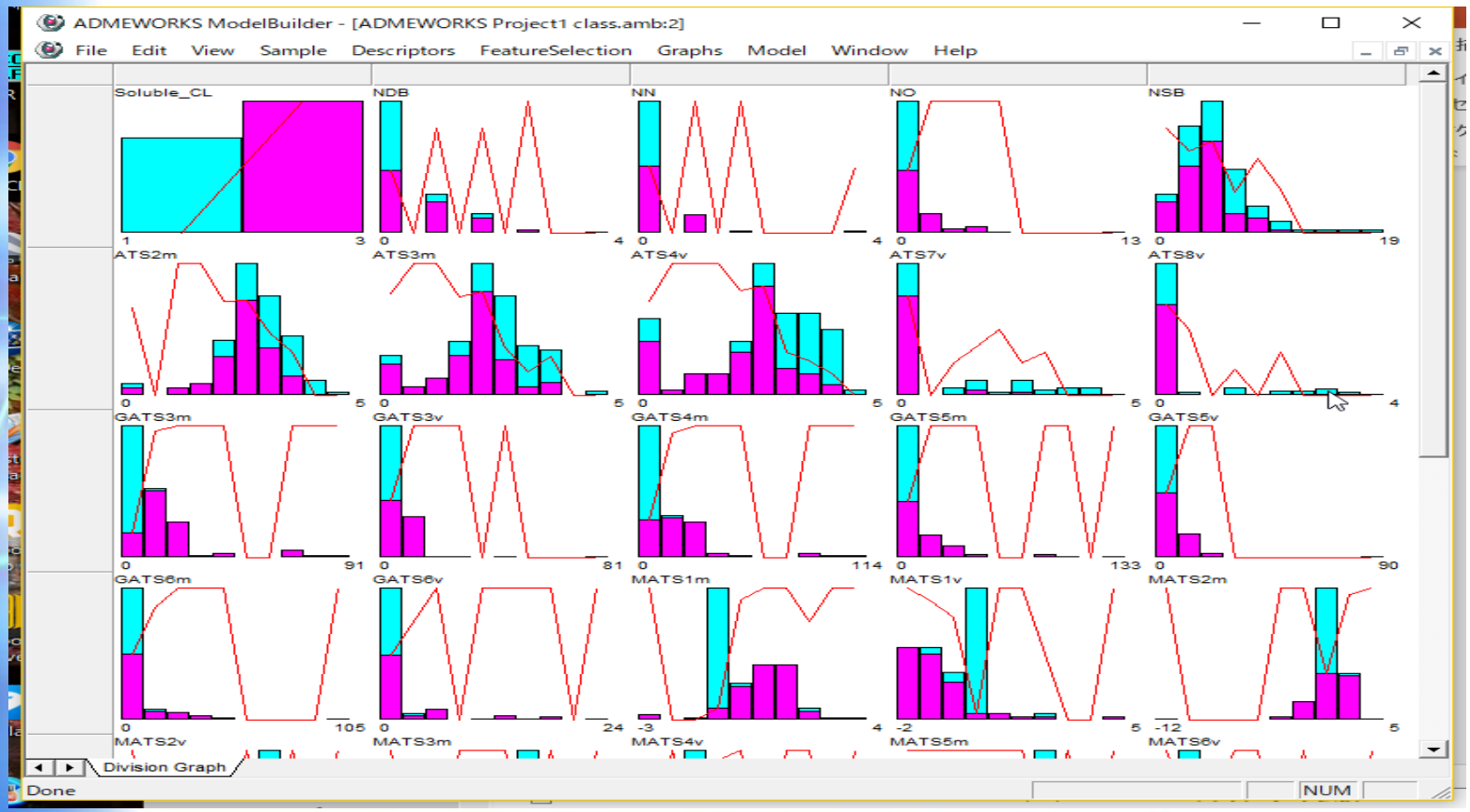
ModelBuilderの画面より

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：クラスディビジョンマップ

Parameter selection: Class division map



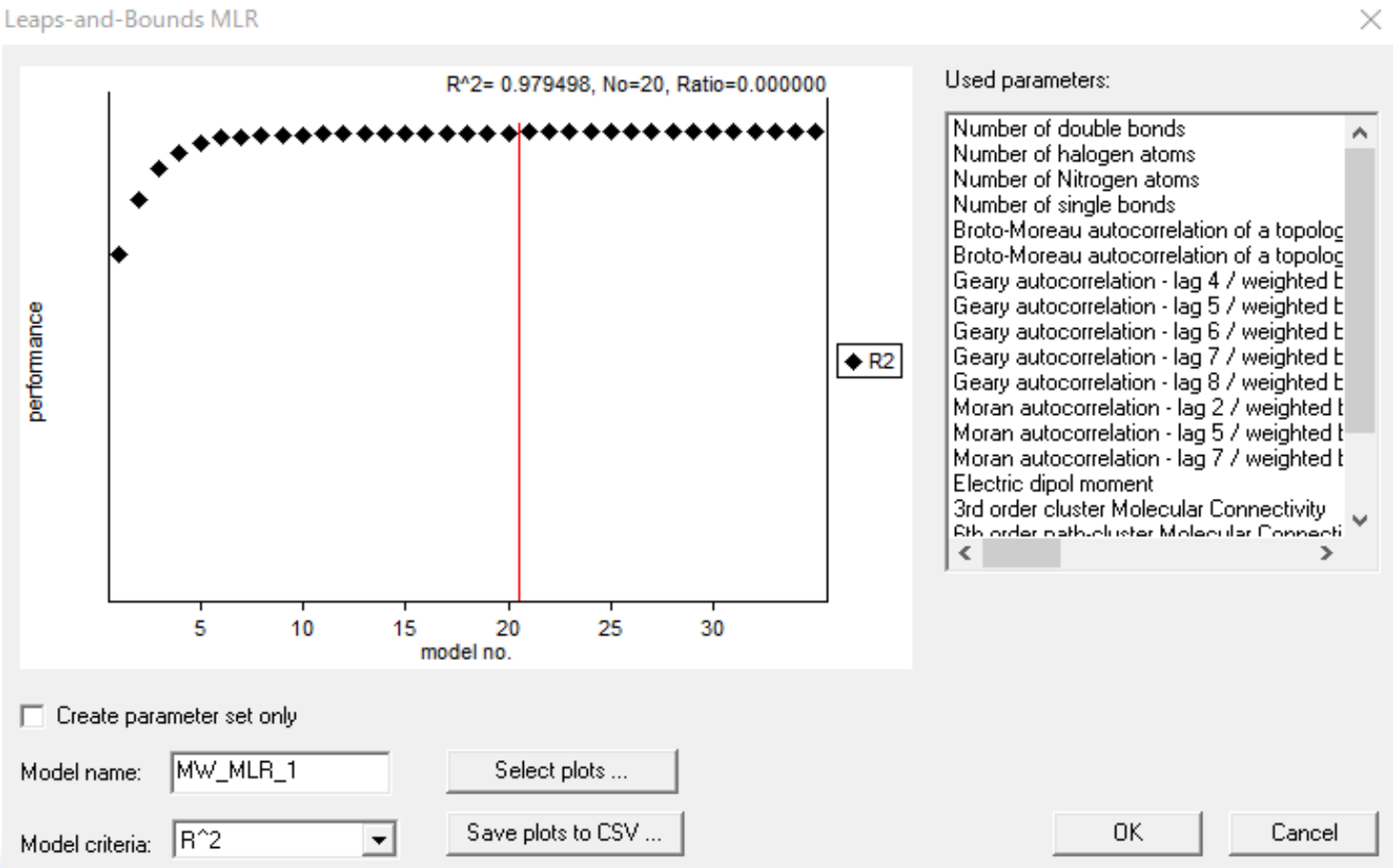
ModelBuilderの画面より

口解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：重回帰

Parameter selection: Multiple regression



ModelBuilderの画面より

□解析に使うパラメーター

Parameters used for analysis

◇パラメーターの選択：データ解析手法の特徴を利用した特徴抽出

Parameter selection: Feature extraction using data analysis method features

クラス分類と重回帰の両方に適用

■二クラス分類　Binary classification

パーセプトロンによる特徴抽出

Feature extraction by perceptron

①ウエイトサイン法 Weight sign method

②バリアンスウエイト法

Variance weight sign method

■多クラス分類 Multi-class classification

SIMCA法の指標を用いた特徴抽出

Feature extraction using SIMCA index

①モデリングパワー

Modeling power

②ディスクリミネイティングパワー

Discriminating power

□ニューラルネットワーク

Neural network

①忘却学習法

Forgetting learning method

②消滅学習法

Erase learning method

□PCA（主成分分析）

PCA (principal component analysis)

①因子負荷量の適用

Application of factor loading

□遺伝的アルゴリズム適用

Genetic algorithm application

種々の解析手法と組み合わせて利用される

Used in combination with various analysis methods

■重回帰（フィッテイング）

Multiple regression (fitting)

①前進選択法

②後進選択法

③T検定適用

④総当たり法

① Forward selection method

② Reverse selection method

③T test application

④ Round-robin method

□解析に使うパラメーター

Parameters used for analysis

◇サンプルの選択 Sample selection

データ解析で判別関数や重回帰式をリファイニングしてゆく過程でサンプルも抽出する必要がある

It is necessary to extract samples in the process of refining discriminant functions and multiple regression equations in data analysis

■ニクラス分類では誤分類の原因となる「インライヤー」の除去

Removal of “inliner” that causes misclassification in 2-class classification

繰り返し最適化の過程で分類できずに残ったサンプル群の除去

Removal of sample groups that could not be classified in the process of repeated optimization

「インライヤー」が特定されたことは、重要な情報源となる

The identification of “inliers” is an important source of information

■重回帰では相関係数を下げる原因となる「アウトライヤー」の除去

Eliminate "outliers" that cause the correlation coefficient to drop in multiple regression

回帰式をプロットし、アウトライヤーを確認して取り除く

Plot regression equation, check outlier and remove

「アウトライヤー」が特定されたことは、重要な情報源となる

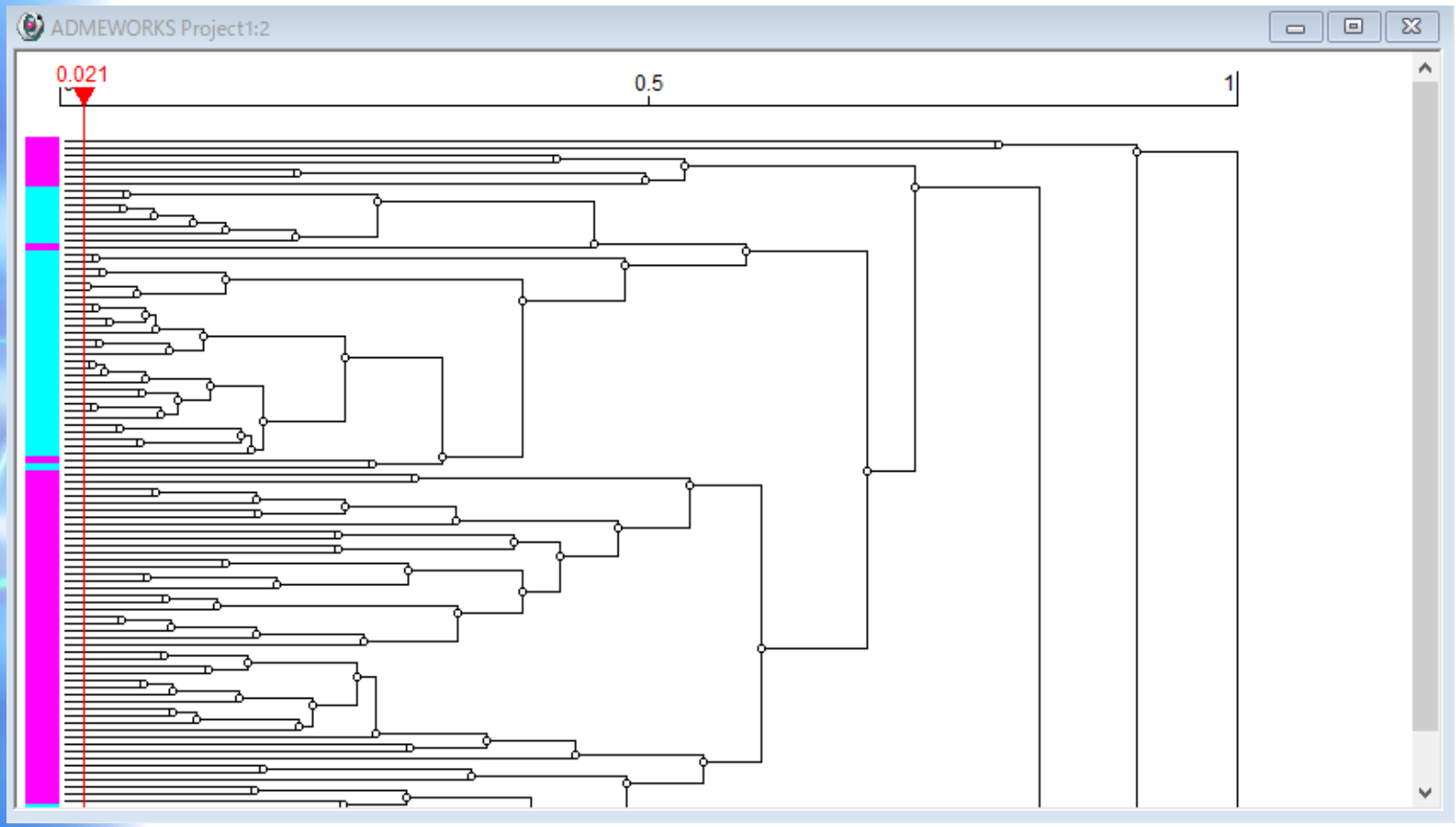
The identification of “outliers” is an important source of information

□解析に使うパラメーター

Parameters used for analysis

◇サンプルの選択：クラスタリング

Sample selection: Clustering



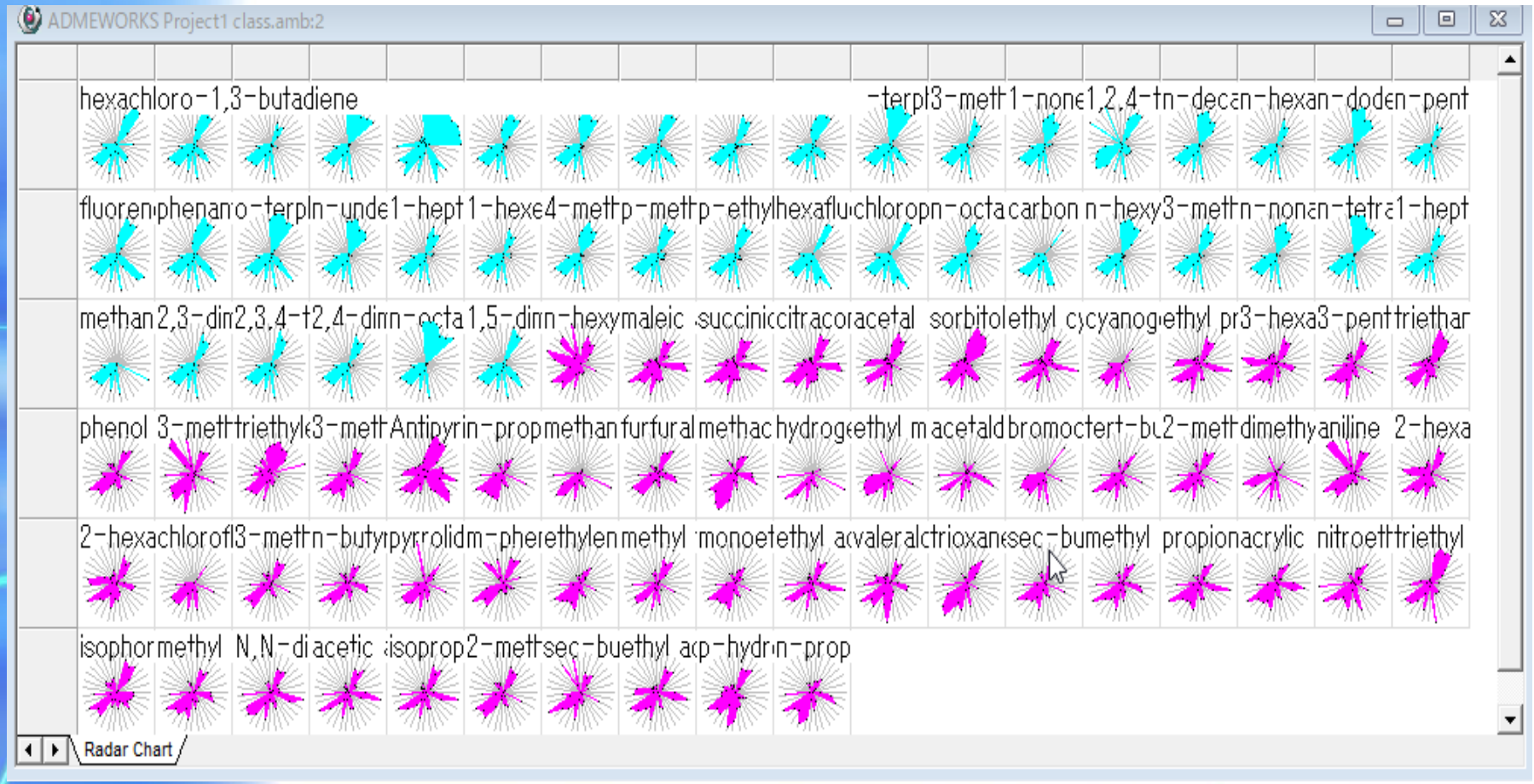
ModelBuilderの画面より

口解析に使うパラメーター

Parameters used for analysis

◇サンプルおよびパラメーター選択：レーダーチャート

Sample and parameter selection: Radar chart



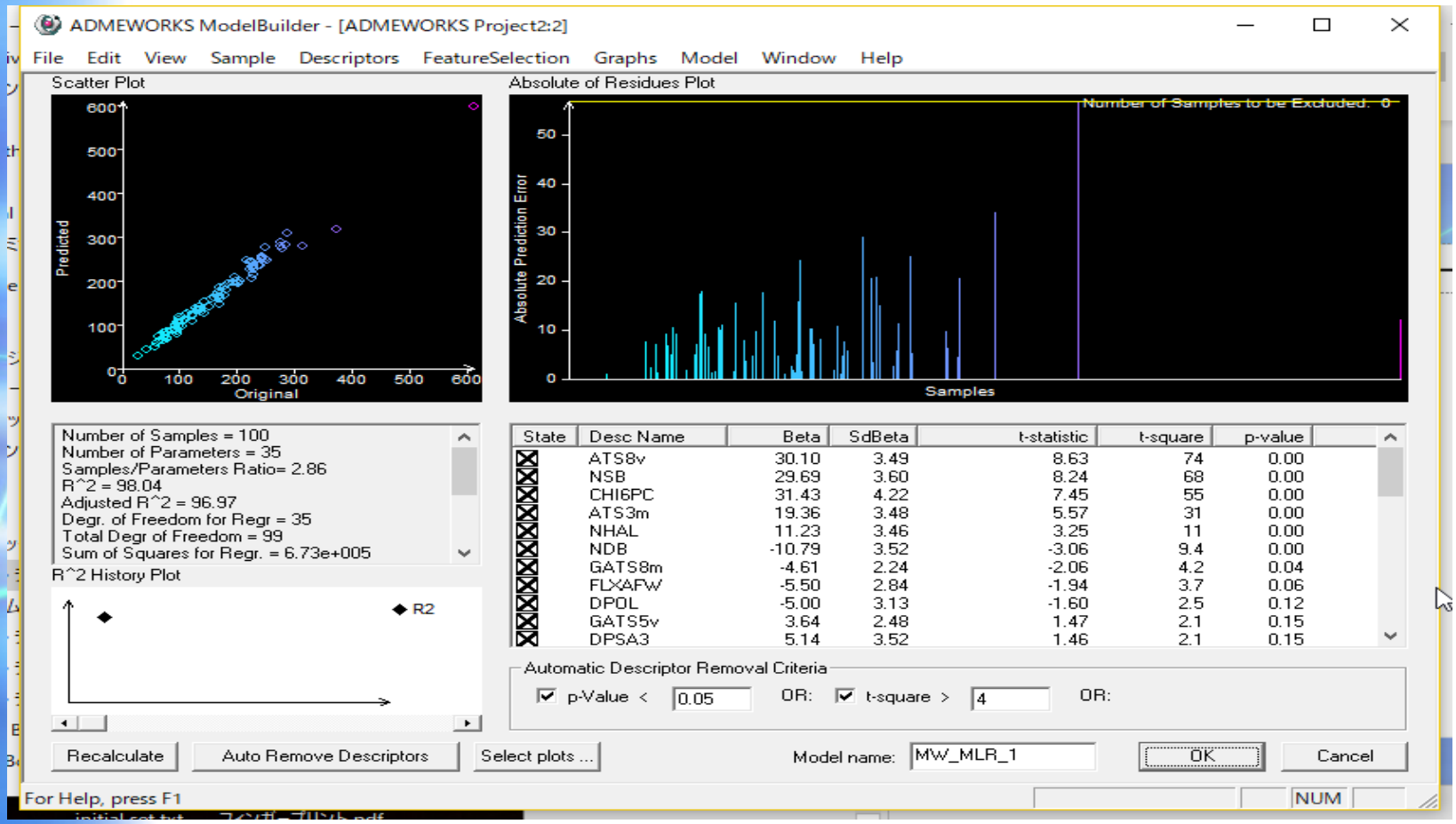
ModelBuilderの画面より

□解析に使うパラメーター

Parameters used for analysis

◇サンプルおよびパラメーター選択：重回帰

Sample and parameter selection: multiple regression



ModelBuilderの画面より

□データ解析関連技術の展開
Development of data analysis related technologies

■解析原理・化合物操作・データ解析での留意事項（1）
Notes on analysis principle, compound manipulation and data analysis

多変量解析／パターン認識によるデータ解析⇒宝箱・発見型アプローチ
化合物関連：一元一項対応、一次元表記、二次元表記、三次元表記
三次元の扱い問題：ローカル・グローバルミニマ対応
確率問題⇒偶然相関
フィッテイング⇒オーバーフィッテイング（過剰適合）
線形/非線形問題；
特徴抽出（パラメーター手法）；種類、特徴、限界
サンプル関連；総サンプル数、最小サンプル数、ポピュレーション（絶対数、クラス比率）
パラメーター関連：種類、単位の違い、オートスケーリング、最少パラメーター数

Data analysis by multivariate analysis / pattern recognition ⇒ treasure chest / discovery approach
Compound-related: one-to-one correspondence, one-dimensional notation, two-dimensional notation, three-dimensional notation
3D handling issues: Local and global minimal
Probability problem ⇒ accidental correlation
Fitting ⇒ Overfitting (overfit)
Linear / nonlinear problems;
Feature extraction (parameter method); type, feature, limit
Sample related: total sample number, minimum sample number, population (absolute number, class ratio)
Parameters: Type, unit difference, auto scaling, minimum number of parameters

■解析原理・化合物操作・データ解析での留意事項（2）

Notes on analysis principle, compound manipulation and data analysis

ニューラルネットワーク⇒中間層のユニット数⇒層の重なり
問題点；チャンスコリレーション、非線形性
パラメーター数が多いとき：パラメーター圧縮　PCA　PLS
パラメーター数が少ない時；成功率低下、解析不能、物性式等
要因解析：パラメーターの読解力、分類力が強いパラメーター
KY法：サンプル数フリー、ポピュレーションフリー
ニクラス分類；完全分類、
重回帰；高相関係数、高絶対係数
分類/予測⇒クロスバリデーション
外挿と内挿の違い

Neural network-> number of units in the middle layer-> layer overlap
Problems: Chance correlation, nonlinearity
When there are many parameters: Parameter compression PCA PLS
When the number of parameters is small; decrease in success rate, inability to analyze, physical properties, etc.
Factor analysis: Parameters with strong reading and classification ability
KY method: Sample number free, population free
2-class classification; complete classification,
Multiple regression; high correlation coefficient, high absolute coefficient
Classification / Prediction ⇒ Cross validation
Difference between extrapolation and interpolation

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

データ解析結果を保証する3段階＋1項目の保証手続き

1. 実施したデータ解析自体が正しい状態にあるか否かのチェック
   データ解析をする前の前提条件
2. データ解析手法自体が有する制限事項や適用限界等
   データ解析手法自体が適用対象や適用限界を有する
3. データ解析の結果が解析的に良好であるか、否か
   一般的には、分類率、予測率、相関係数、絶対係数等の指標を用いる
4. その他
   プログラムの制限事項、化合物を扱うときの特殊事項、その他
   個々のプログラム上での制限事項、バージョンの違い、化合物に
   起因する特殊問題等への注意が必要

3 steps + 1 item guarantee procedure to guarantee data analysis results

1. Check that the data analysis performed is in the correct state
Preconditions before data analysis
2. Restrictions and application limits of the data analysis method itself
The data analysis method itself has application targets and application limits
3. Whether the data analysis results are analytically good or not
Generally, indicators such as classification rate, prediction rate, correlation coefficient, and absolute coefficient are used.
4). Other
· Program restrictions, special matters when handling compounds, etc.
· Restrictions on individual programs, version differences, compound need to pay attention to special problems

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

1. 実施したデータ解析自体が正しい状態にあるか否かのチェック
   Check that the data analysis performed is in the correct state

データ解析をする前の前提条件 Preconditions before data analysis

＊サンプル数と解析に用いるパラメーター数との比が指針となる

The guideline is the ratio between the number of samples and the number of parameters used in the analysis.

＊データ解析で避けなければならない以下の問題を避けることを保証するパラメーターである

This parameter ensures that the following problems that must be avoided in data analysis are avoided:

- **過剰適合**：Over Fitting
- **偶然相関**：Chance Correlation

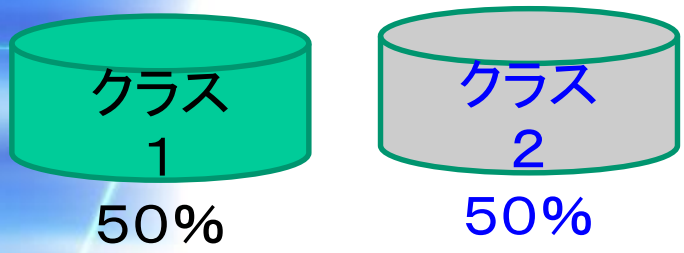
＊上記二問題はクラス分類や重回帰（フィッテイング）手法の総てのデータ解析手法に適用される問題である

The above two problems apply to all data analysis methods such as classification and multiple regression (fitting) methods.

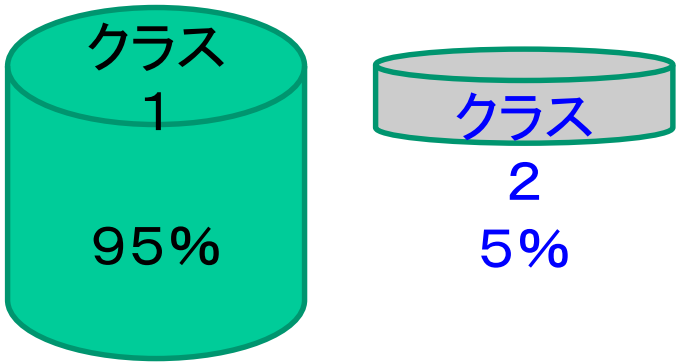
□サンプルポピュレーション
Sample population

□サンプルポピュレーションが問題になるのは判別分析等を適用する場合
＊サンプルポピュレーションがクラス間で大きな差異がない時はデータ解析を問題なく実施できる
＊サンプルポピュレーションがクラス間で大きな違いがある場合、データ解析結果がクラスポピュレーションの影響を強く受ける
＊クラス分類では、ポピュレーションが多いクラスの分類率が高くなる。

Sample population is a problem when discriminant analysis is applied
* Data analysis can be performed without problems when sample populations are not significantly different between classes.
* If the sample population varies greatly between classes, the data analysis results will be strongly influenced by population
* Class classification increases the classification rate of classes with a large population.



分類率に関する寄与は略同等となる
The contribution regarding the classification rate is almost the same.



創出される判別関数は、総てのサンプルをクラス1に判定する。しかし、分類率は95%と高い値になる。
The discriminant function created determines all samples as class 1. However, the classification rate is as high as 95%.

◇総サンプル数／クラスサンプル数での制限事項

Restrictions on the total number of samples / class samples

＊データ解析を行う時はデータ解析結果の信頼性を保証することが必要である
＊一般的にサンプル分布を基本とする統計解析ではサンプル数を多くそろえることが必要なことは
よく知られている
＊多変量解析／パターン認識を行う時の最小サンプル数はどの程度必要なのだろうか
＊他の分野と異なり、薬理活性や毒性等の分野ではサンプルを多数集めることは殆どできない
＊この点で、信頼性を保ちながら多変量解析／パターン認識を行うための最小サンプル数が重要になる

* When performing data analysis, it is necessary to guarantee the reliability of the data analysis results.
* In general, statistical analysis based on sample distribution requires a large number of samples.
well known
* How many minimum samples are required for multivariate analysis / pattern recognition?
* Unlike other fields, it is almost impossible to collect many samples in fields such as pharmacological activity and toxicity.
* In this respect, the minimum number of samples is important for multivariate analysis / pattern recognition while maintaining reliability.

＊信頼性の高いデータ解析を実施するための**最少サンプル数**は、解析に用いた
パラメーターの数との関係で決まる
＊データ解析信頼性は以下のパラメーターを基準として設定される

* The minimum number of samples for performing highly reliable data analysis is determined by the relationship with the number of parameters used in the analysis.
* Data analysis reliability is set based on the following parameters.

◇総サンプル数／クラスサンプル数での制限事項

Restrictions on the total number of samples / class samples

2クラス分類の信頼性
Reliability of 2-class classification

$$\frac{\text{総サンプル数}}{\text{総パラメーター数}} \geqq 4$$

重回帰の信頼性
Reliability of multiple regression

$$\frac{\text{総サンプル数}}{\text{総パラメーター数}} \geqq 5$$

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

□ 「偶然性」問題における次元数とサンプル数との関係（一般化）
次元（記述子）が一つ増える毎に分類可能な場合の数は2倍ずつ増加する。
従って、次元数dにより定まる分割可能な場合の数Rは以下の式で示される。

$$R = 2^{d} \quad \text{(1)}$$

この結果、次元数がNで、サンプル数が2^{N}以下の時には必ず分類出来、この分類結果は偶然により支配されている事は明白である。
一方、サンプル数がnの時、このサンプルを2クラスに分類出来る場合の数Cは単なる組み合わせ問題であり、以下の式で示される。

$$C = \frac{1}{2} \sum_{k=1}^{n} \frac{n!}{k \times (n-k)!}$$

これらの項目を考慮し、与えられた記述子（次元数）dでサンプルnを2分割出来る可能性Pは

$$P = \frac{\text{サンプルnに対する2分割の場合の数}}{\text{記述子dによる2分割の場合の数}} = \frac{C}{R}$$

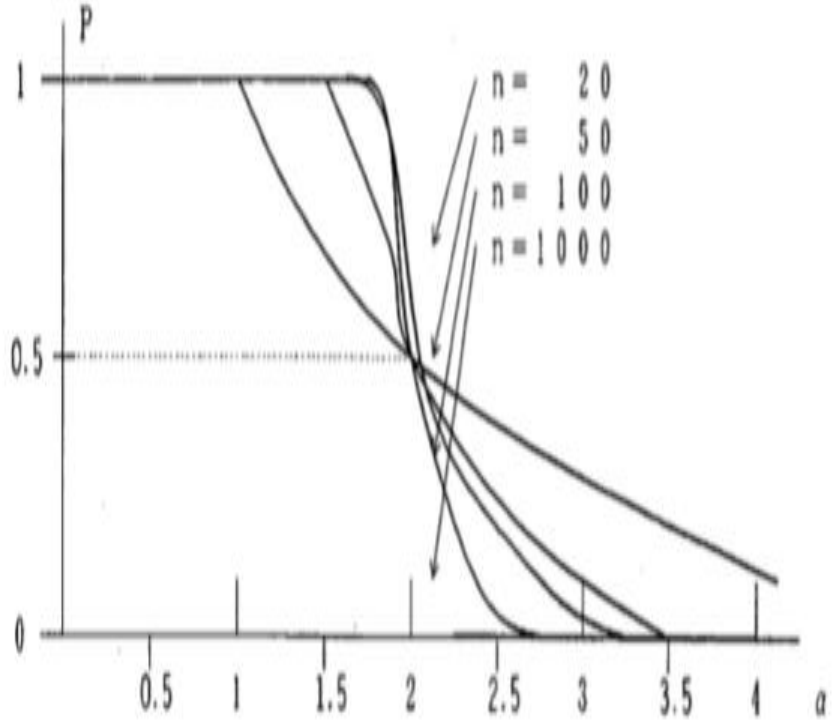


図4. 2分割の可能性に対するα（サンプル数n／次元数d）とPの関係

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

■解析信頼性の簡単な事例　A simple example of analysis reliability

□<u>ニクラス分類 Binary classification</u>

例1：100サンプルを1パラメーターで100%分類　⇒　**信頼性指標＝100**

Example 1: 100 samples are classified as 100% by one parameter ⇒ Reliability index = 100

この解析のパラメーターはクラス分類に極めて**重要な情報を持つ**

This parameter has extremely important information for classification

---

例2：100サンプルを1000パラメーターで100%分類　⇒　**信頼性指標＝0．1**

Example 2: 100 samples are classified 100% with 1000 parameters ⇒ Reliability index = 0.1

1000個のパラメーターはクラス分類に**重要な情報を持たない**

1000 parameters do not have important information for class classification

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

■解析信頼性の簡単な事例　A simple example of analysis reliability

□**重回帰（フィッテイング）Multiple regression (fitting)**

例１：１００サンプルを１パラメーターで１００％分類　⇒　**信頼性指標＝１００**
Example 1: 100 samples are classified as 100% by one parameter ⇒ Reliability index = 100

このパラメーターはクラス分類に極めて**重要な情報を持つ**
This parameter has extremely important information for classification

---

例２：１００サンプルを１０００パラメーターで１００％分類　⇒　**信頼性指標＝０．１**
Example 2: 100 samples are classified 100% with 1000 parameters ⇒ Reliability index = 0.1

１０００個のパラメーターはクラス分類に**重要な情報を持たない**
1000 parameters do not have important information for class classification

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

■解析信頼性の簡単な事例　A simple example of analysis reliability

□<u>ニクラス分類　Binary classification</u>

例1：100サンプルを1パラメーターで100%分類

Example 1: 100 samples are classified as 100% by one parameter

ポジかネガの100サンプルの可能な組み合わせの場合の数は2^{100} となる。
パラメーターが二値パラメーターであれば、表現できる場合の数は2。
従って、1パラメーターで100サンプルを二分割できる確率は、
Ｐ＝２／2^{100} で、殆ど0であり、**チャンスコリレーション（偶然相関）**はない。

The number of possible combinations of 100 positive or negative samples is 2^{100}.
If the parameter is a binary parameter, the number that can be expressed is 2.
Therefore, the probability that 100 samples can be divided into two with one parameter is $P = 2/2^{100}$, almost zero, and there is no chance correlation.

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

■解析信頼性の簡単な事例　A simple example of analysis reliability

□二クラス分類　**Binary classification**

例2：100サンプルを1000パラメーターで100%分類

ポジかネガの100サンプルの可能な組み合わせの場合の数は2^{100} となる。
パラメーターが二値パラメーターであれば、表現できる場合の数は2^{1000} である。
従って、1パラメーターで100サンプルを二分割できる確率は、
$P = 2^{1000} / 2^{100}$ で、Pは極めて大きい値となる。
例2の条件下では、100サンプルの100%分類は確実に実現する。即ち、
チャンスコリレーション（偶然相関）が発生する。

The number of possible combinations of 100 positive or negative samples is 2100.
If the parameter is a binary parameter, the number of cases that can be expressed is 21000.
Therefore, the probability that 100 samples can be divided into two with one parameter is $P = 2^{1000}/2^{100}$, and P is a very large value.
Under the conditions of Example 2, 100% classification of 100 samples is reliably realized.
That is, chance correlation (accidental correlation) occurs.

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

◇重回帰で相関係数や絶対係数を１にする方法
Method to set correlation coefficient and absolute coefficient to 1 by multiple regression

目的変数として薬理活性のED50を用いて、100個のサンプルを用意。なお、これら100個のサンプルにはあらかじめ1から100番までの任意のID番号を付ける。

100 samples were prepared using ED50 of pharmacological activity as the objective variable. These 100 samples are given arbitrary ID numbers from 1 to 100 in advance.

1. サンプルデータとして薬理活性のED50値を持つ100個の化合物を用意。
2. 使用するパラメータとしてサンプル数と同じ100パラメータを用意します。
3. 各パラメータは化合物のID番号の部分を1とし、残りはすべて0とします。
4. 100サンプルのED50を目的変数、100個のパラメータを説明変数として重回帰を実行します。

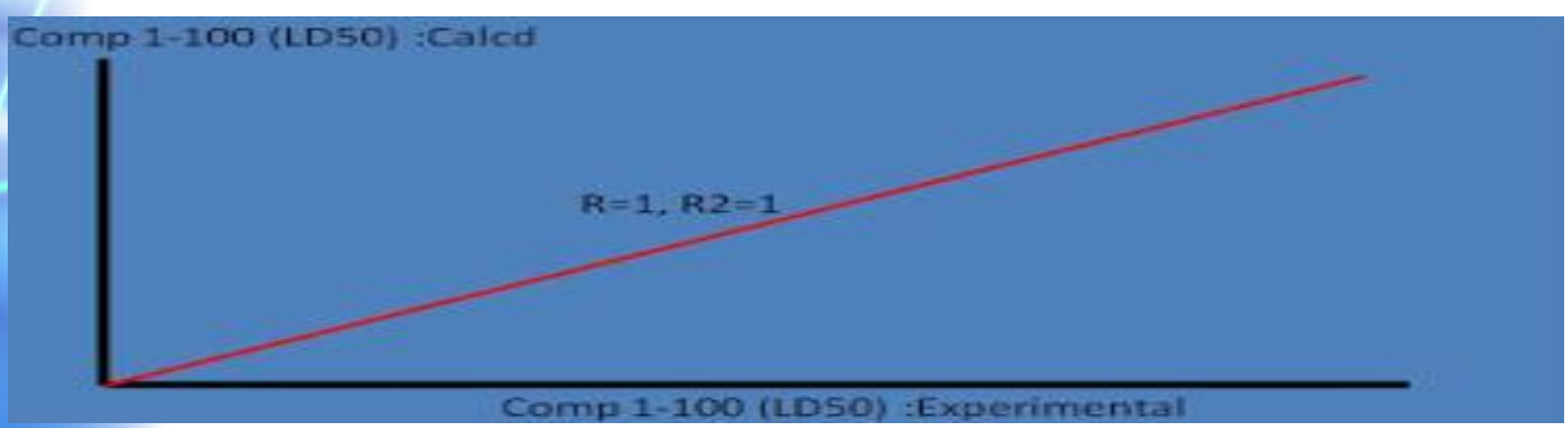
1. Prepare 100 compounds with ED50 values of pharmacological activity as sample data.
2. Prepare 100 parameters that are the same as the number of samples.
3. For each parameter, the ID number of the compound is 1 and the rest are all 0.
4. Perform multiple regression with 100 samples of ED50 as the target variable and 100 parameters as explanatory variables.

実行結果:
相関係数R=1、 絶対係数R2=1
Execution result:
Correlation coefficient R = 1, absolute coefficient R2 = 1

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

◇重回帰で相関係数や絶対係数を１にする方法
Method to set correlation coefficient and absolute coefficient to 1 by multiple regression

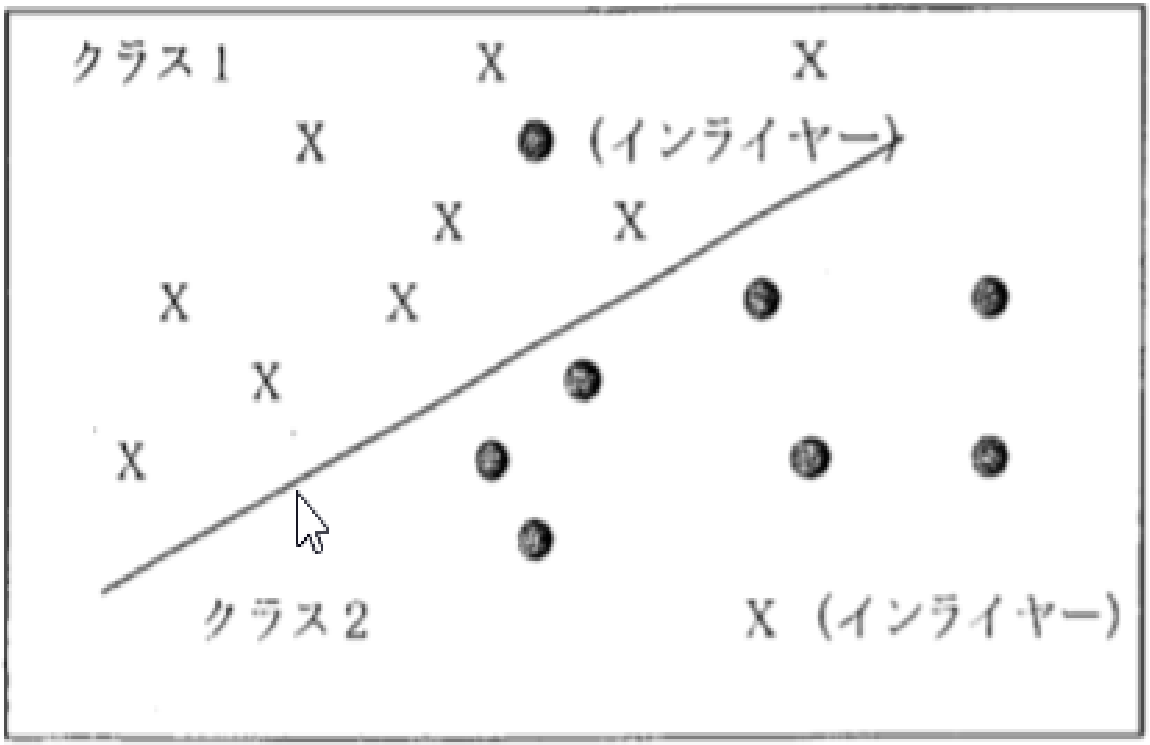
	LD50	param1	param2	***	***	param99	param100
comp1	58	1	0	0	0	0	0
comp2	132	0	1	0	0	0	0
comp3	12	0	0	1	0	0	0
comp99	728	0	0	0	0	1	0
comp100	311	0	0	0	0	0	1



◇取り出すサンプル数の限界
Limit on number of samples

＊**ニクラス分類**では誤る分類するサンプル（インライヤー）を取り除くことで判別関数の予測精度を高める。
このサンプル取り出しは初期サンプル数の約10%を上限にするとされている。
* In the two-class classification, the prediction accuracy of the discriminant function is improved by removing erroneously classified samples (inliers).
This sample removal is supposed to have an upper limit of about 10% of the initial number of samples.

＊**重回帰（フィッテイング）**でもサンプル（アウトライヤー）を取り除くことで重回帰式の相関／絶対係数を高くする。
このサンプル取り出しはニクラス分類と同様に初期サンプル数の約10%を上限にするとされている。
* In the multiple regression (fitting), the correlation / absolute coefficient of the multiple regression equation is increased by removing the sample (outlier).
Similar to the two-class classification, this sample extraction is supposed to have an upper limit of about 10% of the initial number of samples.



◇クラスサンプル数の限界 Class sample limit

クラスポピュレーション ≧ パラメーター数
Class population ≧ Number of parameters

ポピュレーション例：
総サンプル数100、クラス分類率100%、でもどちらの結果を信用しますか
Population example:
Which result do you trust, with a total sample count of 100 and a classification rate of 100%

①クラス1；99サンプル、 クラス2；1サンプル
②クラス1；50サンプル、 クラス2；50サンプル
① Class 1; 99 samples, Class 2; 1 sample
② Class 1; 50 samples, Class 2; 50 samples

◇分類率と予測率 Classification rate and prediction rate

＊二クラス分類では分類率と予測率が作成された判別関数の精度を示す指標として利用される
* In the two-class classification, the classification rate and the prediction rate are used as an index indicating the accuracy of the discriminant function that was created.

＊分類率はデータ解析に用いたサンプルを判定するものなので、サンプルに困ることはない
＊予測率は予測項目の実測値が無いので、予測率を出すことは出来ない
＊予測率算出には、クラス既知のサンプルを使って仮の値を出すことが出来る
＊一般的にはクロスバリデーションと呼ばれており、様々な手法が展開されている
＊“Leave N Out”法が最も良く利用されている
　クラス既知のサンプルのなかからN個のサンプルを取り出す。このNサンプルをクラス未知とし、残る(T－N)個のサンプルを用いて予測モデルを構築し、この予測モデルを用いて取り出されたNサンプルについて予測を行う。この手順を総てのサンプルについて繰り返して、全体の予測値を出す。
　この時、パラメーターセットは同じものを利用する。

* Since the classification rate determines the sample used for data analysis, there is no problem with the sample.
* Since there is no actual measurement value of the prediction item, the prediction rate cannot be obtained.
* For calculating the prediction rate, a temporary value can be obtained using a sample of known class
* Generally called cross-validation, various methods have been developed.
* “Leave N Out” method is the most commonly used.
* N samples are taken out of samples of known class. The N samples are class unknown, a prediction model is constructed using the remaining (TN) samples, and prediction is performed on the N samples extracted using the prediction model. This procedure is repeated for all samples to give an overall predicted value. At this time, the same parameter set is used.

◇分類率と予測率　Classification rate and prediction rate

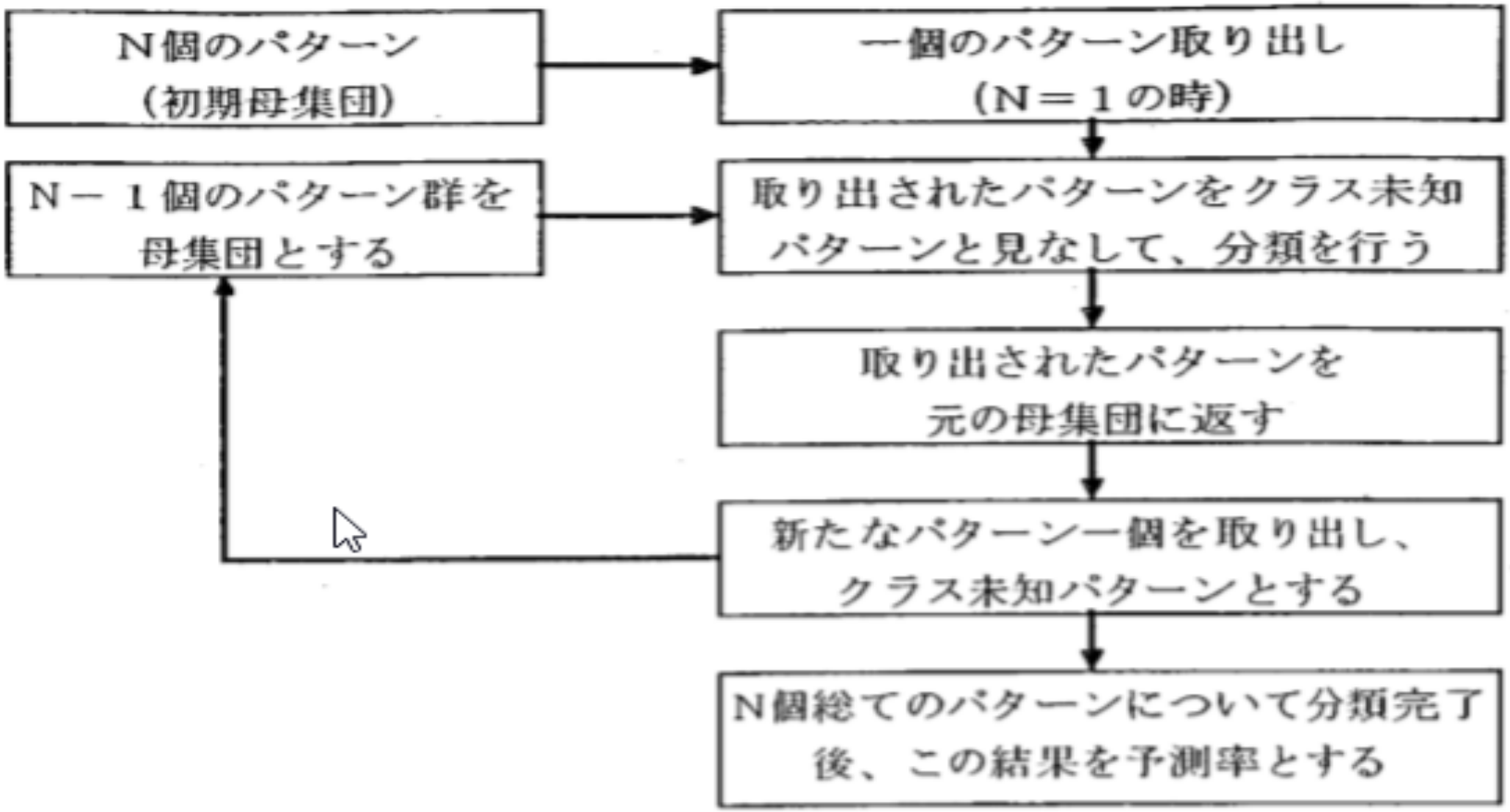


図　．リーブワンアウト法による予測率計算の流れ図

◇総サンプル数／クラスサンプル数
Total number of samples / class samples

＊総サンプル数は使ったパラメーターの数で決まる
＊データ解析時に用いるサンプル数を4（二クラス分類）か5（重回帰）で割った値のパラメーターを用いてデータ解析を行えば、データ解析の信頼性が保証される

* The total number of samples is determined by the number of parameters used.
* Reliability of data analysis is guaranteed if data analysis is performed using parameters obtained by dividing the number of samples used in data analysis by 4 (two-class classification) or 5 (multiple regression)

◇総サンプル数／クラスサンプル数
Total number of samples / class samples

100サンプル用いた場合、解析に用いたパラメーターが
2クラスタ分類　：　25以下　⇒　解析信頼性が保証
25以上　⇒　解析信頼性が低い
重回帰　：　20以下　⇒　解析信頼性が保証
20以上　⇒　解析信頼性が低い

When 100 samples were used, the parameters used for the analysis were
2 cluster classification: 25 or less ⇒ Guaranteed analysis reliability
25 or more ⇒ Low analysis reliability
Multiple regression: 20 or less ⇒ Guaranteed analysis reliability
20 or more ⇒ Low analysis reliability

結論：最小サンプル数の縛りはない
用いたサンプル数でパラメーター数の縛りが発生

Conclusion: There is no restriction on the minimum number of samples
The number of parameters used is limited by the number of samples used

◇総サンプル数／クラスサンプル数
Total number of samples / class samples

＊クラスの最少サンプル数は使えるパラメーターの数と直結する
＊解析信頼性を保ってクラス分類を行う場合、総サンプル数よりも最小クラスのサンプル数がパラメーターの利用制限に強い影響を及ぼす

* The minimum number of samples in a class is directly linked to the number of usable parameters.
* When classifying with analysis reliability maintained, the number of samples in the smallest class has a greater influence on the parameter usage limit than the total number of samples.

１００サンプルでクラス１が９０でクラス２が１０の場合
２クラスタ分類　：　１０**以下**　⇒　解析信頼性が保証
１０**以上**　⇒　解析信頼性が低い

100 samples, class 1 is 90 and class 2 is 10
2 cluster classification:　10 or less ⇒ Guaranteed analysis reliability
10 or more ⇒ Low analysis reliability

結論：クラスポピュレーションの最小クラスのサンプル数
この値を超えた数のパラメーターは使えない

Conclusion: Number of samples in the smallest class of class population
The number of parameters exceeding this value cannot be used

＊パラメーターを多くしたい時は、最小クラスのサンプル数を増やす＊
If you want more parameters, increase the number of samples in the smallest class

◇線形／非線形問題　Linear / nonlinear problems

＊線形および非線形に関する問題はデータ解析を行う時に常に考慮すべき事項である

＊データ解析の**外挿性や内挿性**に関係する

＊**過剰適合**を起こしている場合、非線形解析の方が発生しやすい

＊データ解析の簡易度で考える場合、線形解析よりも非線形解析の方が成功率は高い

＊判別分析や重回帰を行う場合、分類率や相関係数、絶対係数値は線形解析よりも
非線形解析の方が高い／良好な結果を導きだす

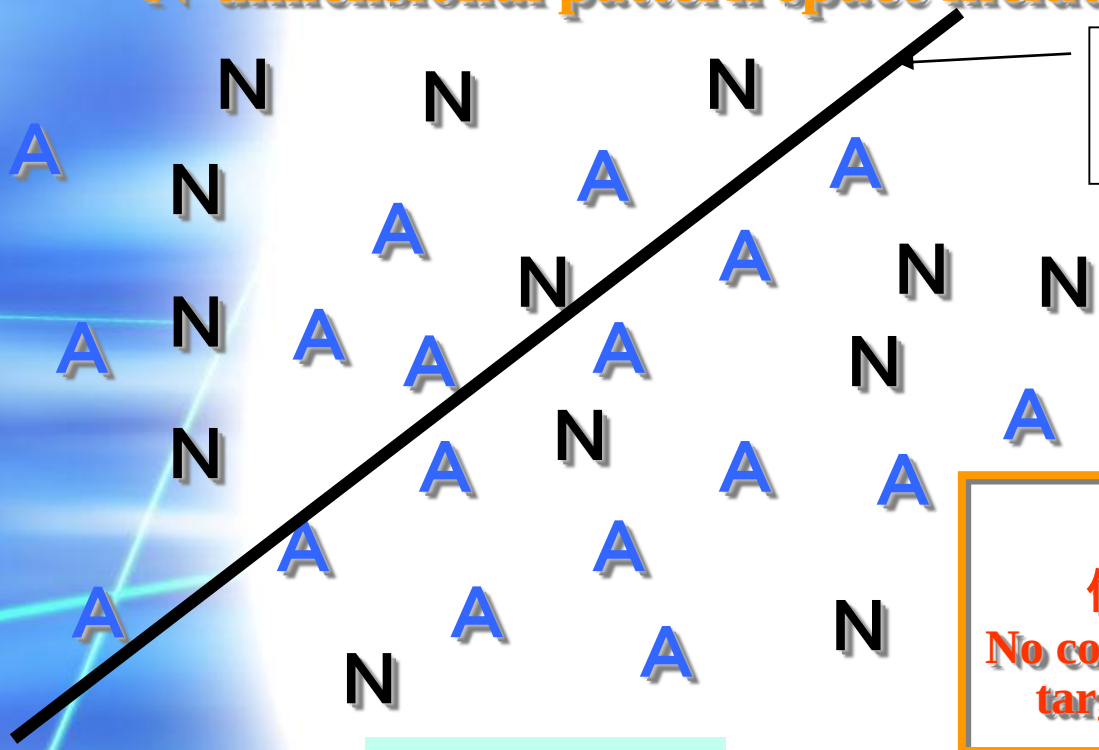
＊N次元サンプル空間での問題を考えた場合、
・線形での解析は**サンプル空間を作り直して**分類や重回帰を実施
・非線形解析の場合、**サンプル空間の形に合わせて**判別関数や重回帰式を算出

* Linear and non-linear problems should always be considered when performing data analysis
* Related to extrapolation and interpolation of data analysis
* Nonlinear analysis is more likely to occur when overfitting occurs.
* When considering the simplicity of data analysis, the success rate of nonlinear analysis is higher than that of linear analysis.
* When performing discriminant analysis and multiple regression, the classification rate, correlation coefficient, and absolute coefficient value are higher in the non-linear analysis than in the linear analysis.
* When considering problems in the N-dimensional sample space,
・For linear analysis, sample space is recreated and classification and multiple regression are performed.
・For non-linear analysis, calculate discriminant function and multiple regression equation according to the shape of the sample space.

◆線形分類によるクラス分類
Class classification by linear classification

＊分類できない場合 classification is not possible

■ノイズデータを含んだN次元パターン空間
N-dimensional pattern space including noise data



線形判別関数
Linear discriminant function

$$Y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + \mathrm{const.}$$

目的活性とパターン空間に
何ら相関関係が見出せない
No correlation can be found between target activity and pattern space

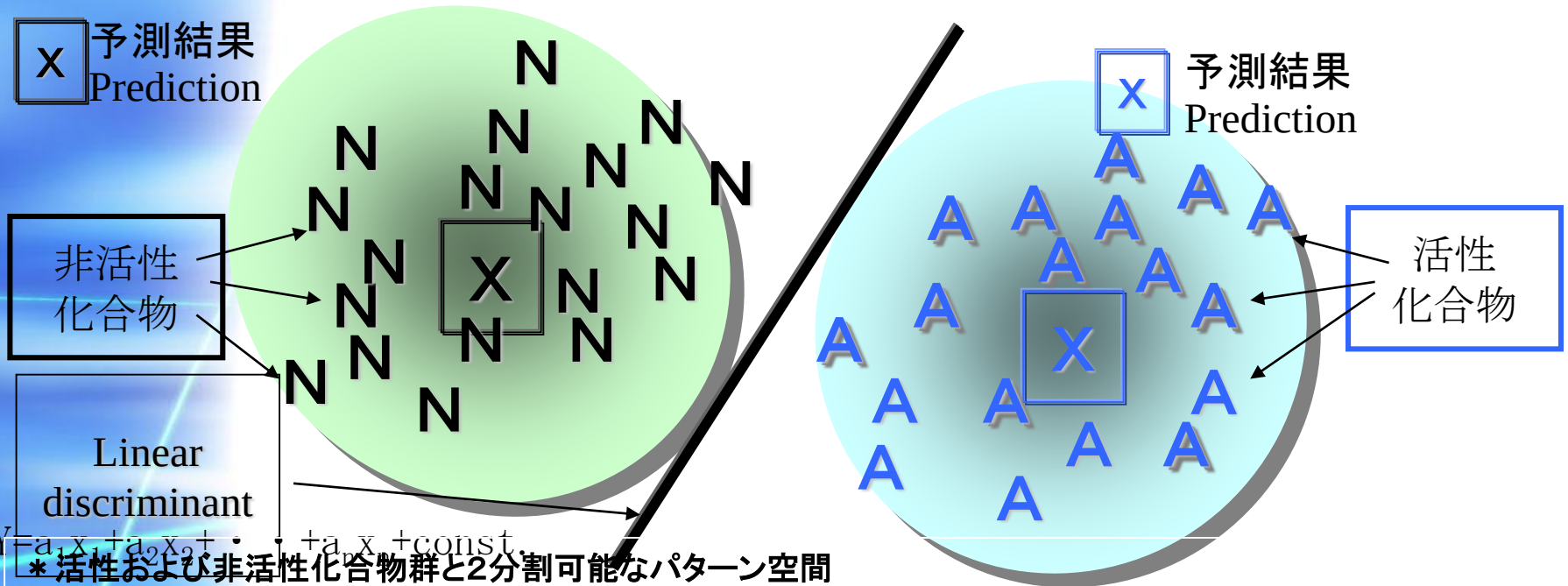
A；活性化合物
N；非活性化合物

◆線形分類によるクラス分類
Class classification by linear classification

＊分類可能な場合 classification is possible

■目的活性と相関の高いパラメータ群により形成されるN次元空間

N-dimensional space formed by parameter groups highly correlated with target activity



＊活性および非活性化合物群と2分割可能なパターン空間

＊本空間を作るパラメータ群には、活性／非活性と分類するに重要な情報 （科学的根拠／相関）を持つ

*** Active and inactive compound groups and pattern space that can be divided into two**

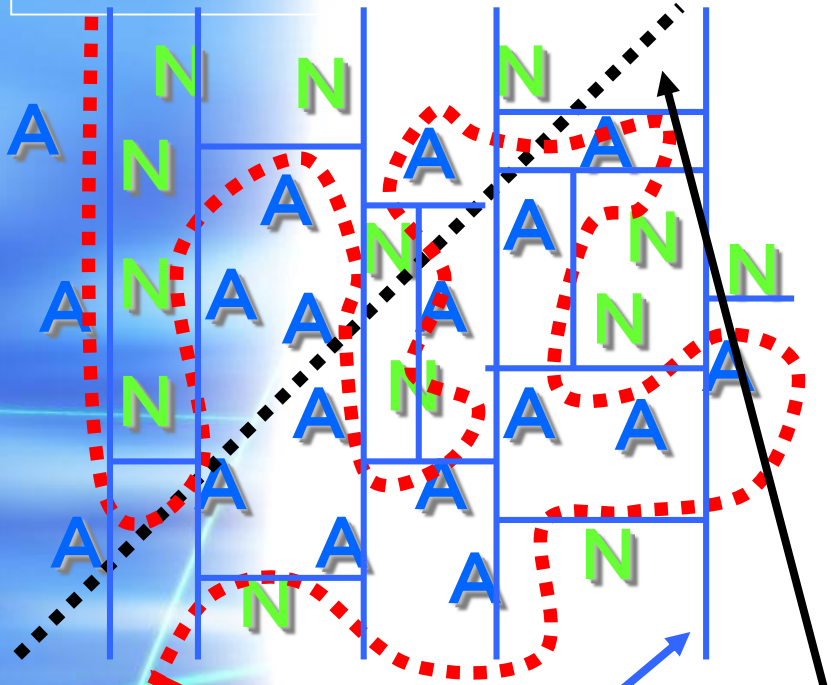
*** The parameters that make up this space have important information (scientific basis / correlation) for classification as active / inactive.**

◆非線形分類によるクラス分類

Class classification by non-linear classification

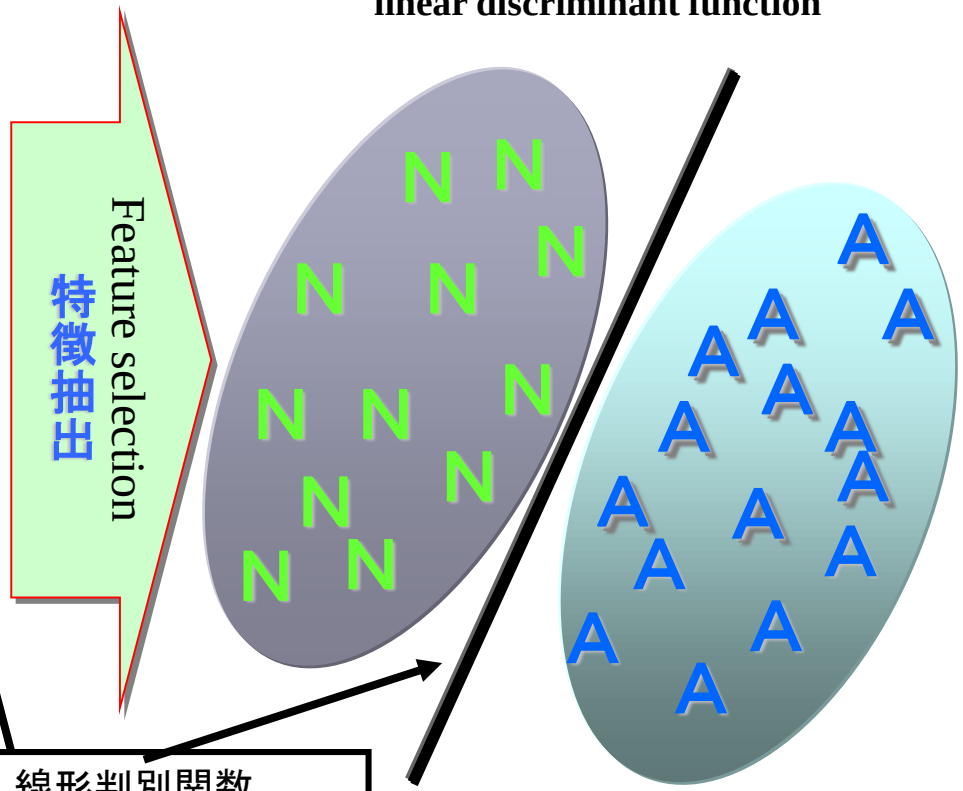
線形判別関数で2分割不可能なパターン空間

Pattern space that cannot be divided into two by linear discriminant function



線形判別関数で2分割可能なパターン空間

Pattern space that can be divided into two by linear discriminant function



ニューラルネットワーク
N.N.

決定木
Decision tree

線形判別関数
Linear discriminant function

科学に基づいたパターン空間
Science-based pattern space

パターン分布に合わせた分類

Classification according to pattern distribution

◆ニューラルネットワークによる非線形分類

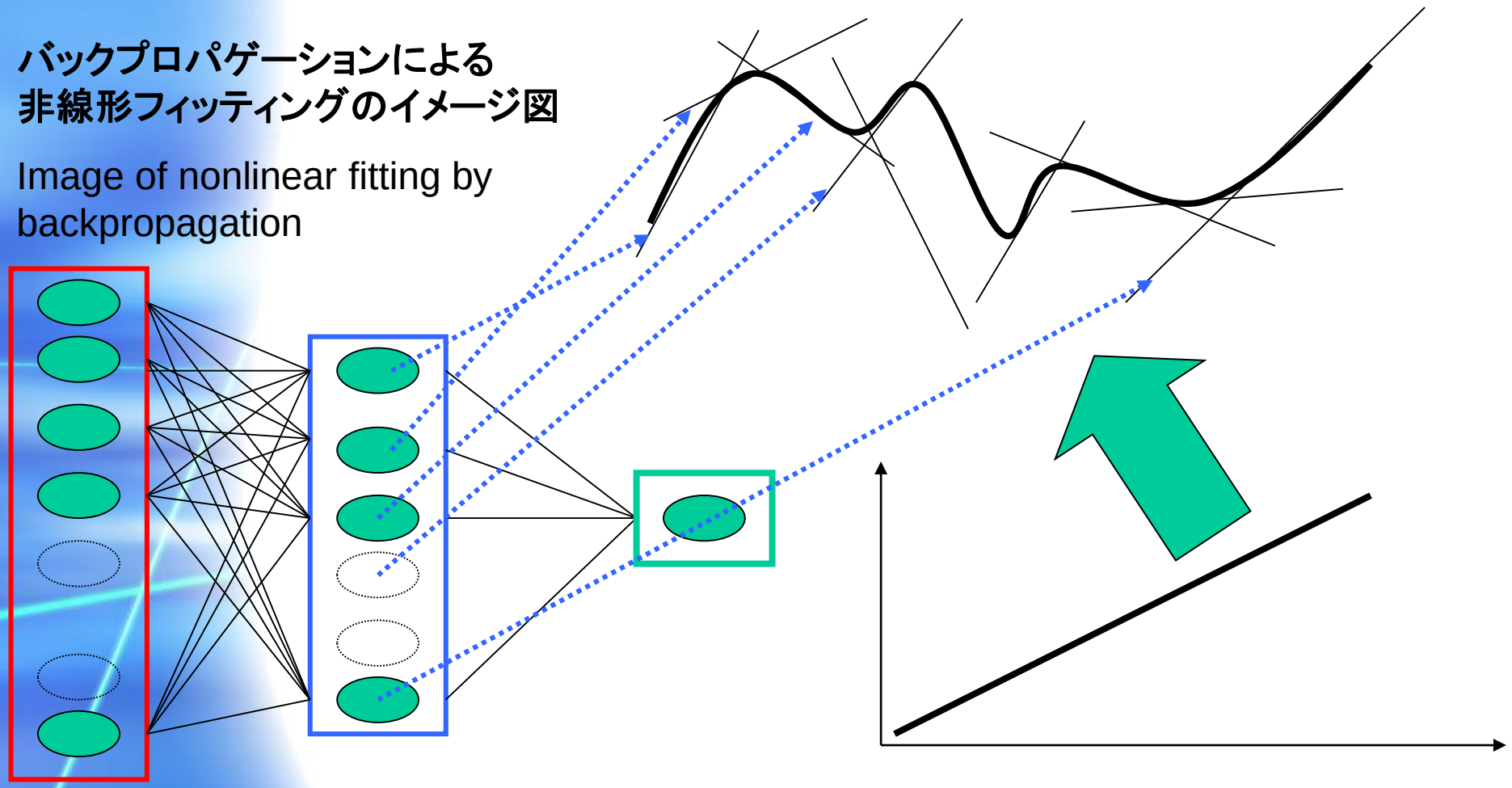
Nonlinear classification by neural network

ニューラルネットワーク：バックプロパゲーション

Neural network: Backpropagation

バックプロパゲーションによる
非線形フィッティングのイメージ図

Image of nonlinear fitting by backpropagation



◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

□ニューラルネットワークによる二クラス分類 Two-class classification by neural network

ニューラルネットワークのネットワーク構造により、パラメーターが表現できる場合の数は単純パーセプトロンと比較して極端に大きい値となる。

Due to the network structure of the neural network, the number of parameters that can be expressed is extremely large compared to a simple perceptron.

例：100サンプルの二クラス分類で、ニューラルネットワークで、入力パラメーター数は10とし、中間層のユニット数も10とした場合の100%分類の可能性は以下となる。

Example: 100-sample two-class classification, with a neural network, the number of input parameters is 10 and the number of units in the middle layer is also 10. The possibility of 100% classification is as follows.

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

□ニューラルネットワークによる二クラス分類　Two-class classification by neural network

ポジかネガの100サンプルの可能な組み合わせの場合の数は2^{100} となる。
一方、パラメーターが二値パラメーターであれば、ニューラルネットワークで表現できる場合の数は入力層で2^{10}。これに中間層で生起される場合の数は、$(2^{10})^{10}$**となる。**
従って、1パラメーターで100サンプルを二分割できる確率は、

$P = (2^{10})^{10} / 2^{100}$ で、極めて大きな値となる。

この結果、チャンスコリレーション（偶然相関）は必然的に発生する。

The number of possible combinations of 100 positive or negative samples is 2^{100}.
On the other hand, if the parameter is a binary parameter, the number that can be expressed by the neural network is 2^{10} in the input layer. The number when this occurs in the intermediate layer is $(2^{10})^{10}$.
Therefore, the probability that 100 samples can be divided into two with one parameter is

$P = (2^{10})^{10}/2^{100}$, which is an extremely large value.

As a result, chance correlation (incidental correlation) inevitably occurs.

◇ケモメトリックス解析を保証するための最低限の制限事項
Minimum restrictions to ensure chemometric analysis

□ニューラルネットワークによる二クラス分類　Two-class classification by neural network

ニューラルネットワークの二クラス分類は**中間層のユニット数**が大きくなると場合の数が拡大し、**中間層の層数**が拡大するとさらに場合の数は急激に拡大する。
最近の**深層学習**を行う**多層のネットワーク構造**では、**チャンスコリレーション**の影響を少なくするべく、学習用サンプル数を極めて大きくすることが必要となる。

In the two-class classification of the neural network, the number of cases increases as the number of units in the intermediate layer increases, and the number of cases increases rapidly as the number of layers in the intermediate layer increases.
In recent multi-layer network structures that perform deep learning, it is necessary to increase the number of learning samples extremely to reduce the influence of chance correlation.

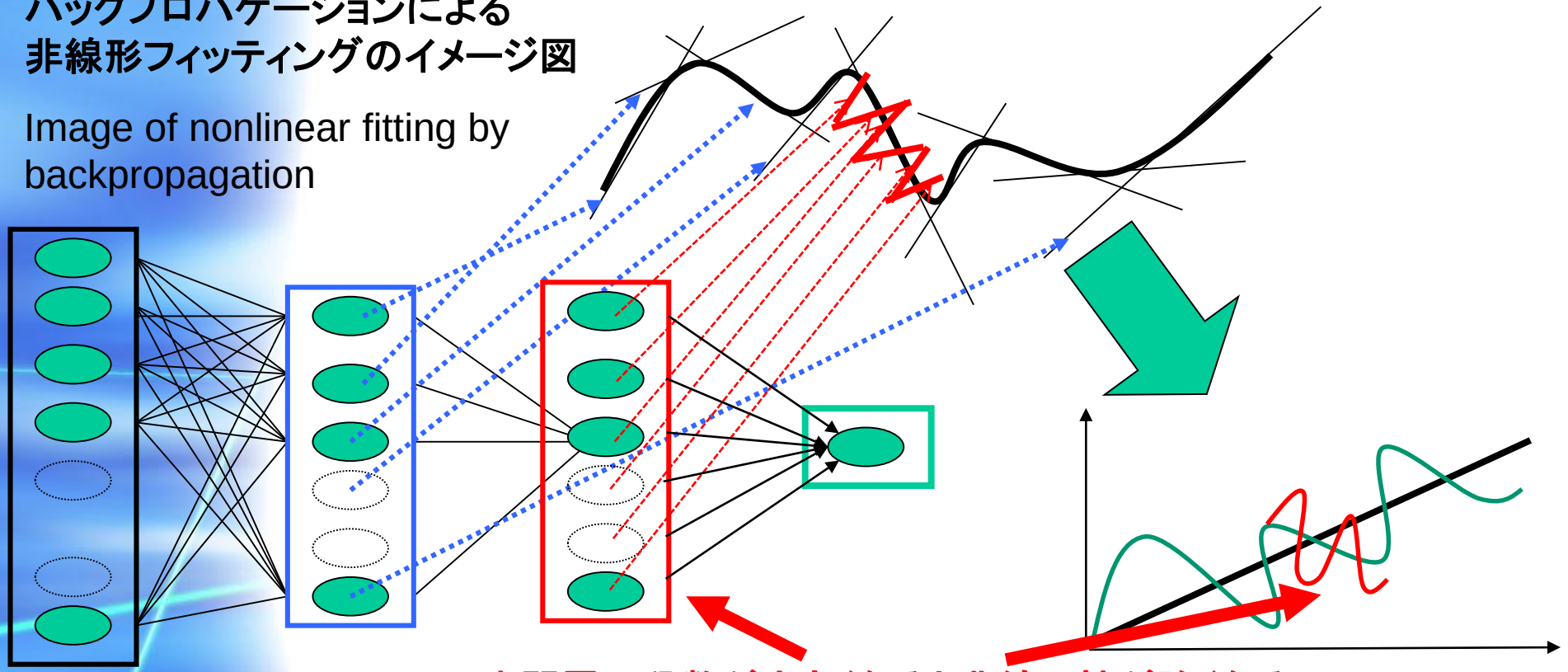
□機械学習型人工知能 Machine learning type artificial intelligence

・パーセプトロンとニューラルネットワーク Perceptron and neural network

ニューラルネットワーク：バックプロパゲーション
Neural network: Backpropagation

バックプロパゲーションによる
非線形フィッティングのイメージ図

Image of nonlinear fitting by backpropagation



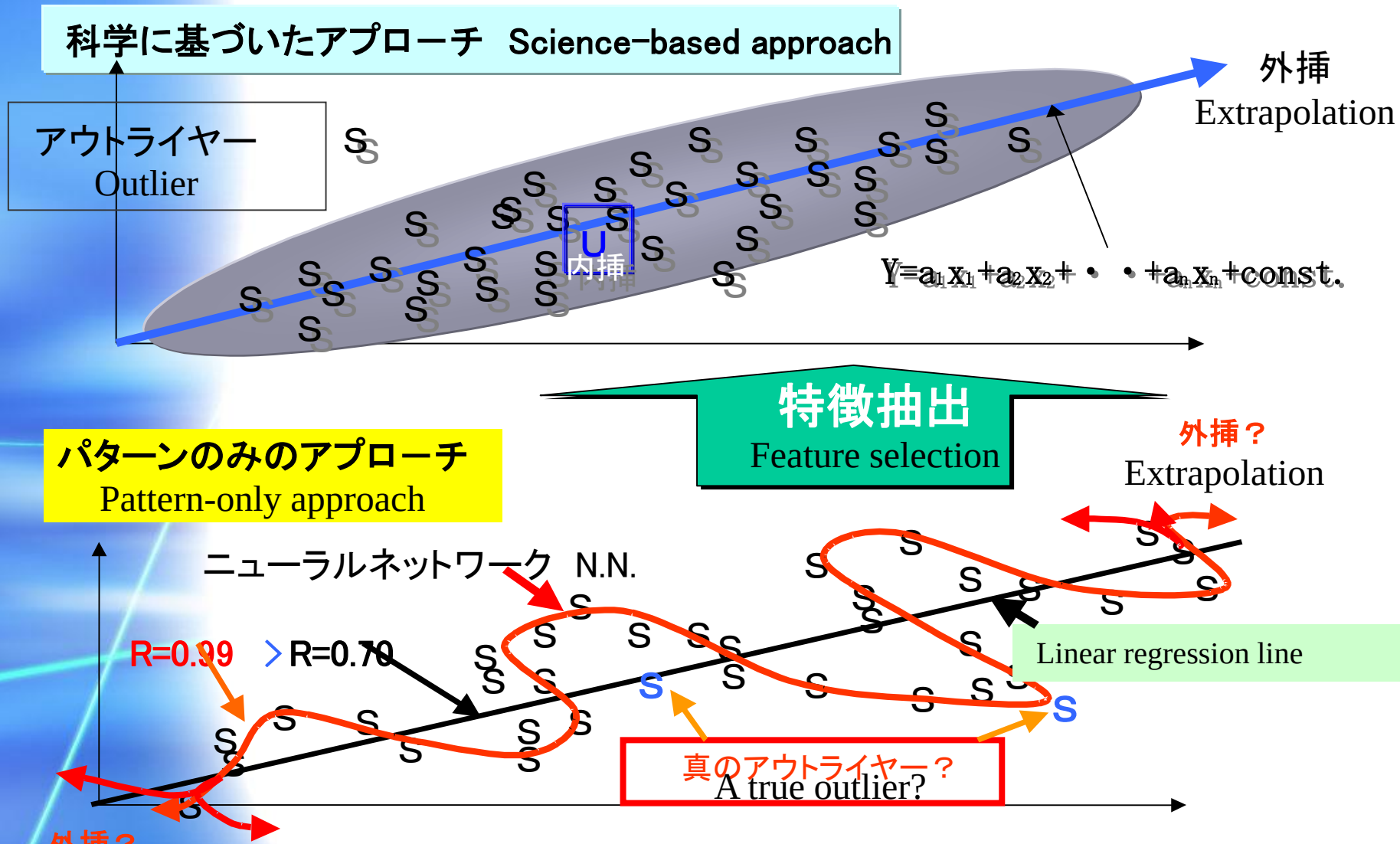
中間層の段数が大きくなると非線形性が強くなる

Nonlinearity increases as the number of intermediate layers increases

中間層の層数は急激な場合の数の拡大を伴う

The number of intermediate layers is accompanied by an increase in the number of sudden cases

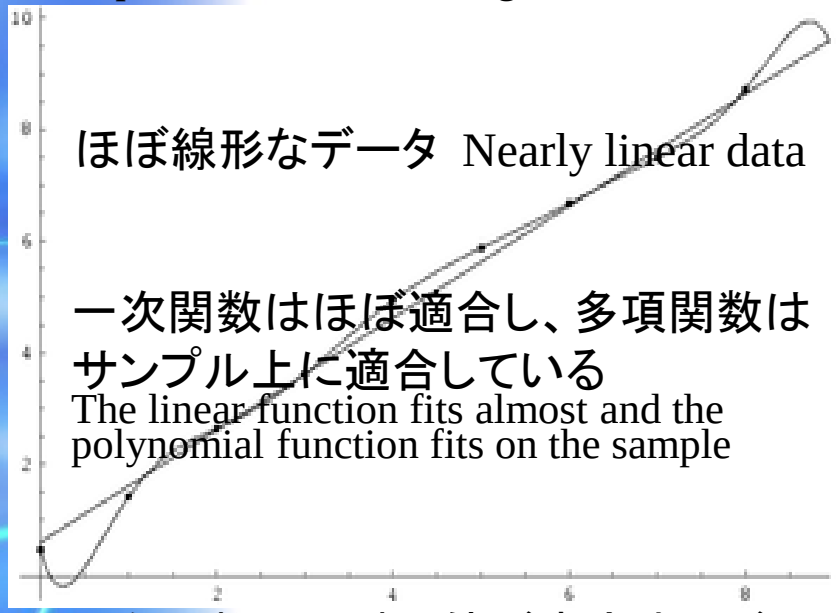
◆線形および非線形フィッテイング
Linear and non-linear fitting

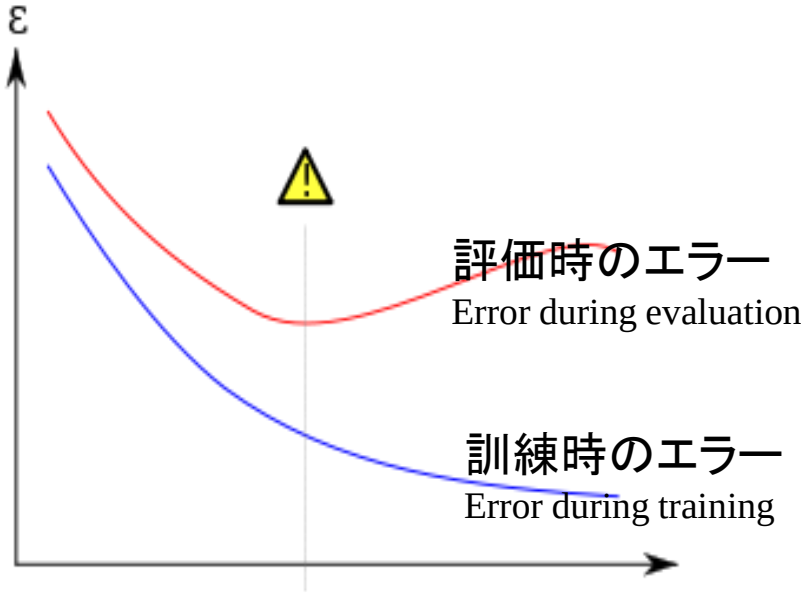


◇過剰適合/過学習
Overfitting / overlearning

過剰適合（かじょうてきごう、英: Overfitting）とは、統計学や機械学習において、訓練データに対して学習されているが、未知データ（テストデータ）に対しては適合できていない、汎化できていない状態を指す。汎化能力の不足に起因する。

Overfitting is a generalization that is learned for training data in statistics and machine learning but cannot be applied to unknown data (test data). Refers to a state that has not been completed. Due to lack of generalization ability.

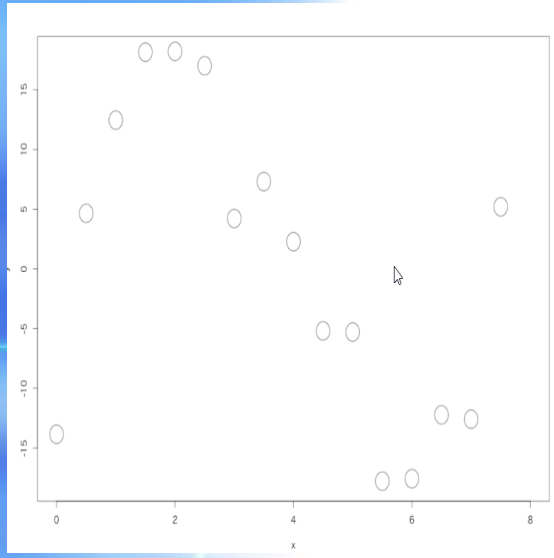




ニューラルネットワークでの過剰適合の状況
Status of overfitting in neural networks

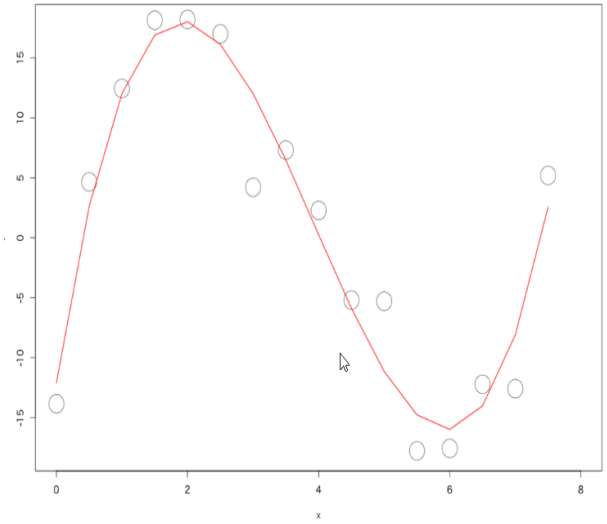
https://ja.wikipedia.org/wiki/過剰適合

◇過剰適合/過学習
Overfitting / overlearning



全サンプル数21
学習用16サンプル
Total number of samples 21
16 samples for learning

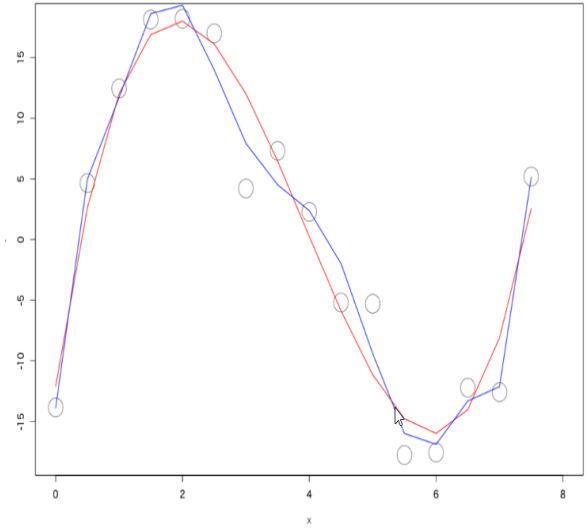
3パラメーターを用いた重回帰
解析信頼性を保ったパラメーター数
Multiple regression using 3 parameters
Number of parameters maintaining analysis reliability



相関係数： 0．968
Correlation coefficient: 0.968

汎化能大
Generalization ability

9パラメーターを用いた重回帰
解析信頼性を伴わない
パラメーター数
Multiple regression using 9 parameters
Number of parameters without analysis reliability

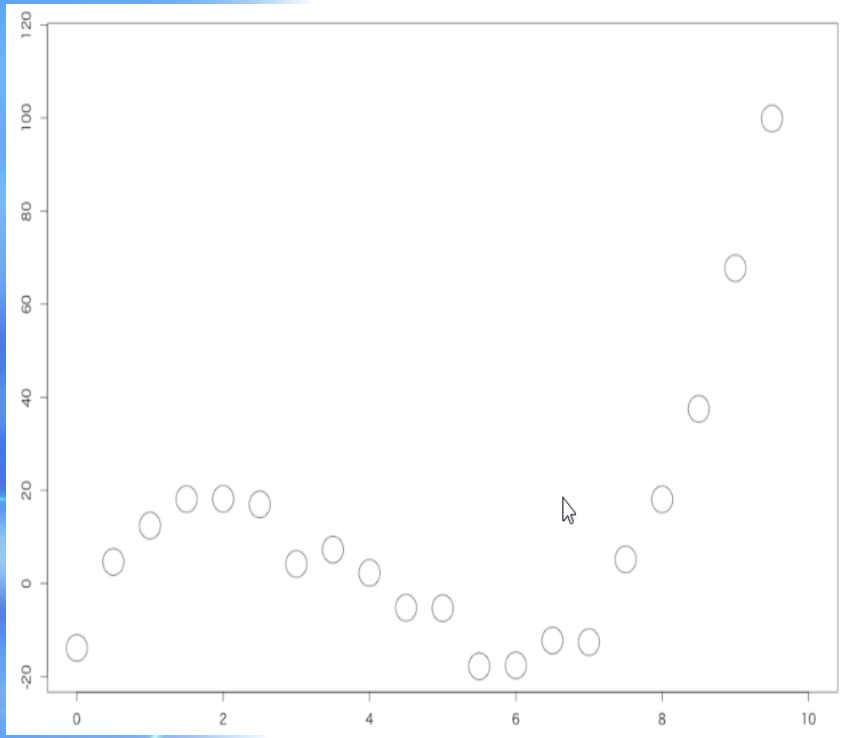


相関係数： 0．986
Correlation coefficient: 0.986

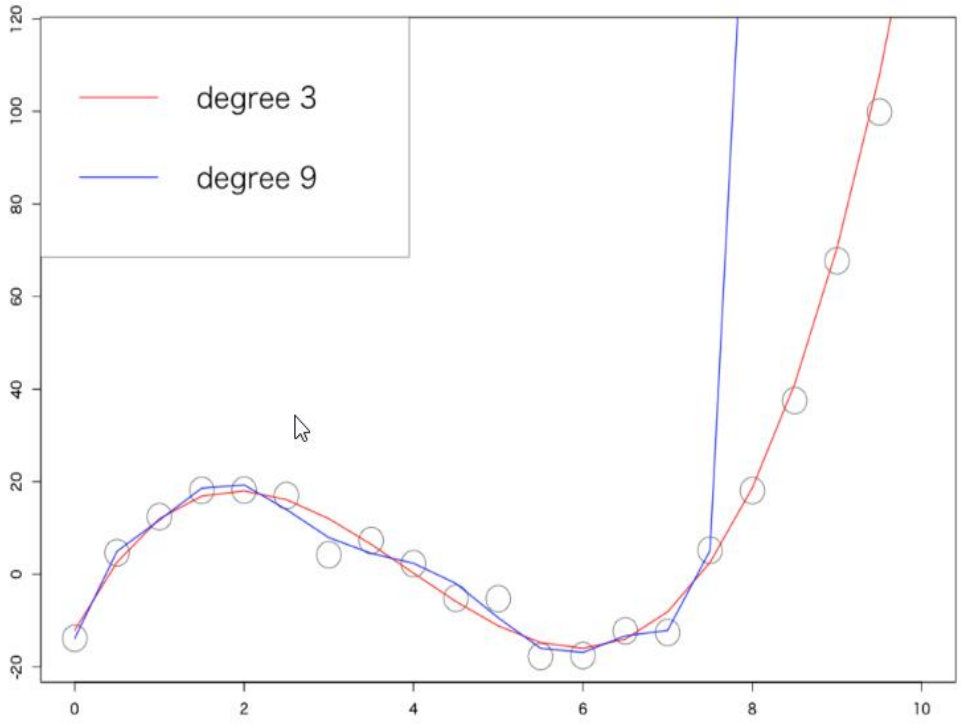
過学習
Over-learning

https://tjo.hatenablog.com/entry/2016/04/14/190000

◇過剰適合/過学習
Overfitting / overlearning



全サンプル21
学習用16＋テストデータ5
Total samples 21
16+ test data for learning 5



パラメーター3の場合と9の場合の回帰図
Regression chart for parameter 3 and 9

https://tjo.hatenablog.com/entry/2016/04/14/190000