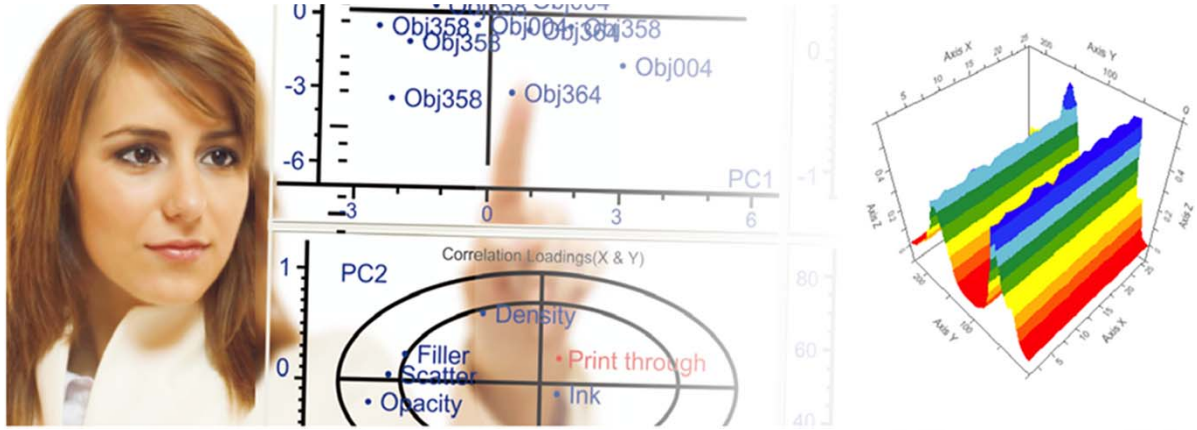


多変量統計解析とは

株式会社クオリティデザイン
〒612-8374 京都市伏見区治部町105番地 301
TEL: 075-605-3270 / FAX: 075-320-3678
E-mail: ask@q-dsn.co.jp
http://www.q-dsn.co.jp

1

AGENDA

01 主成分分析
02 多変量スペクトル分解法
03 各回帰式の考え方
04 3Dデータの取扱い
05 トレーニングセミナーのご案内

2

主成分分析
(PCA: Principal Component Analysis)

01

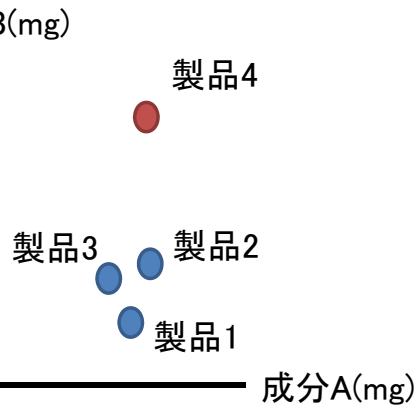
3

多変数のデータを可視化する

データは多くの変数を持つために数値を眺めて直接的に解釈することは極めて困難である
プロット、グラフ等を使って可視化すれば解釈は非常に容易になる

説明変数

試料	成分A (mg)	成分B (mg)
製品1	0.19	0.12
製品2	0.21	0.14
製品3	0.18	0.13
製品4	0.20	0.24

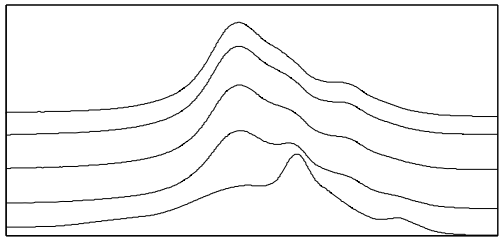


- どの製品が特徴的か？　⇒　製品4は他よりも離れた位置にある
- どうして特徴的なのか？　⇒　縦軸の値（成分B）の量が他よりも多い

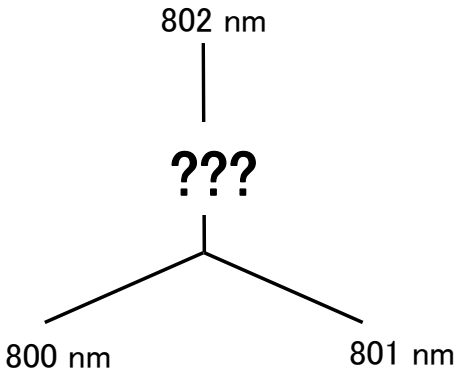
4

スペクトルのように変数が膨大なデータでは・・・

スペクトル



	800 nm	801 nm	------	2500 nm
製品1				
製品2				
製品3				
製品4				

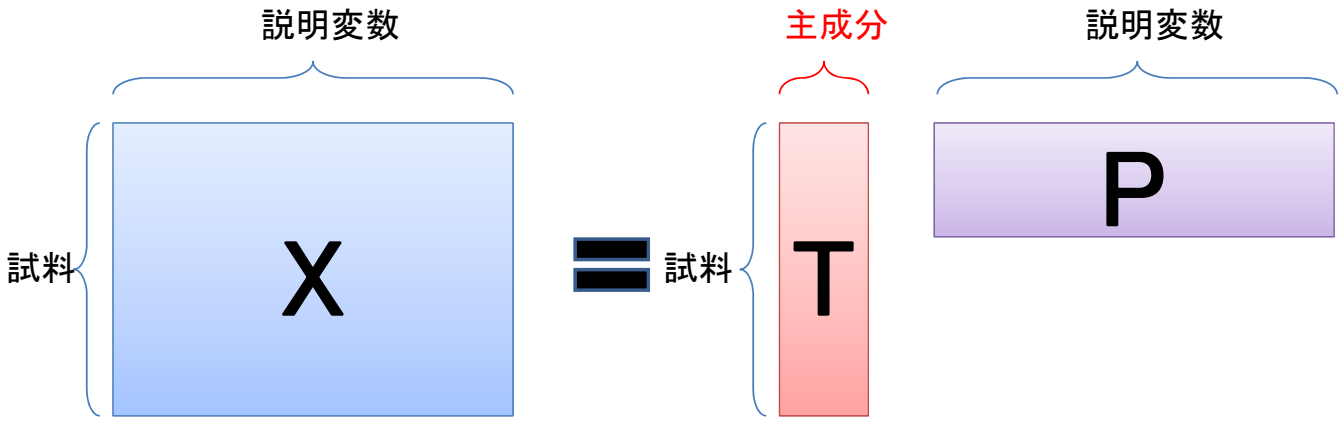


- スペクトルのように説明変数の数が多いデータでは可視化することが困難

⇒情報を失わないように変数の数を少なくすることが必要

情報を失わないように変数の数を減らすには？

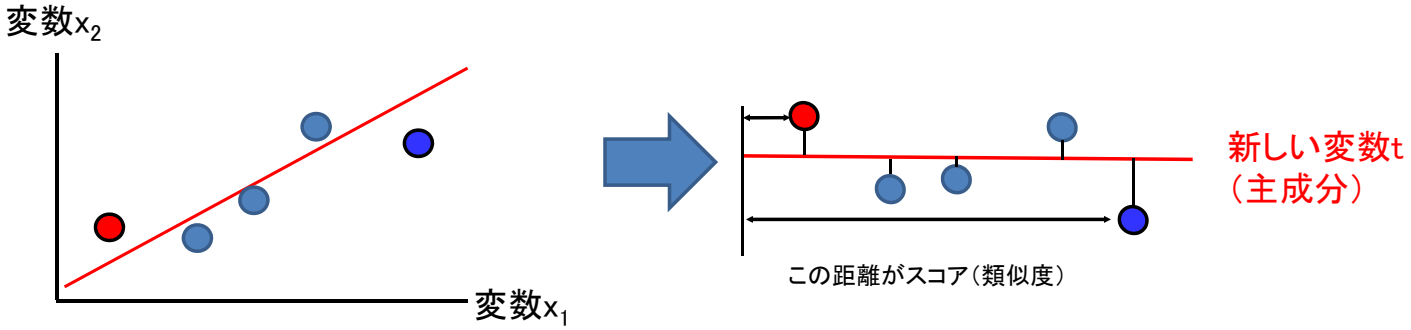
説明変数を数学的に変換して別の説明変数（主成分）にしてしまう



T:スコア　試料間の関係を表す
P：ローディング　スコアの値がどのような説明変数に関連しているのかを表す
主成分：変換によって新しく作られた説明変数

主成分とは？

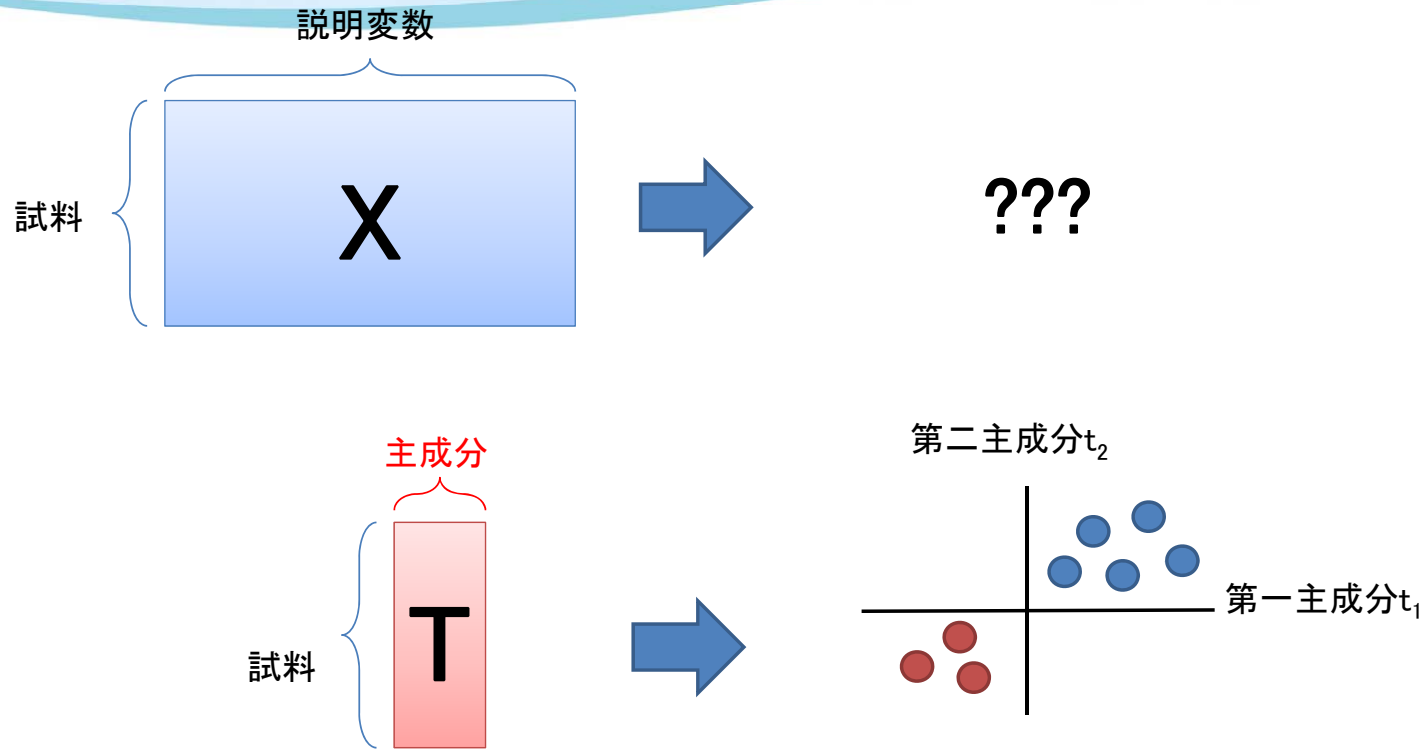
互いに似たような情報をもつ変数を「新しい変数」にまとめてしまう



x_1の値が増えればx_2の値も増える。これらの変数は互いに同じような情報を持つ
⇒2つの軸でなくても1の軸（赤の直線）を引けばサンプル間の関係は十分に表現できる

- 主成分とは同じ情報を持つ説明変数をまとめた「新しい変数」
- このために主成分の数は元の説明変数の数よりもはるかに少ない
- 主成分間には情報の重複がない

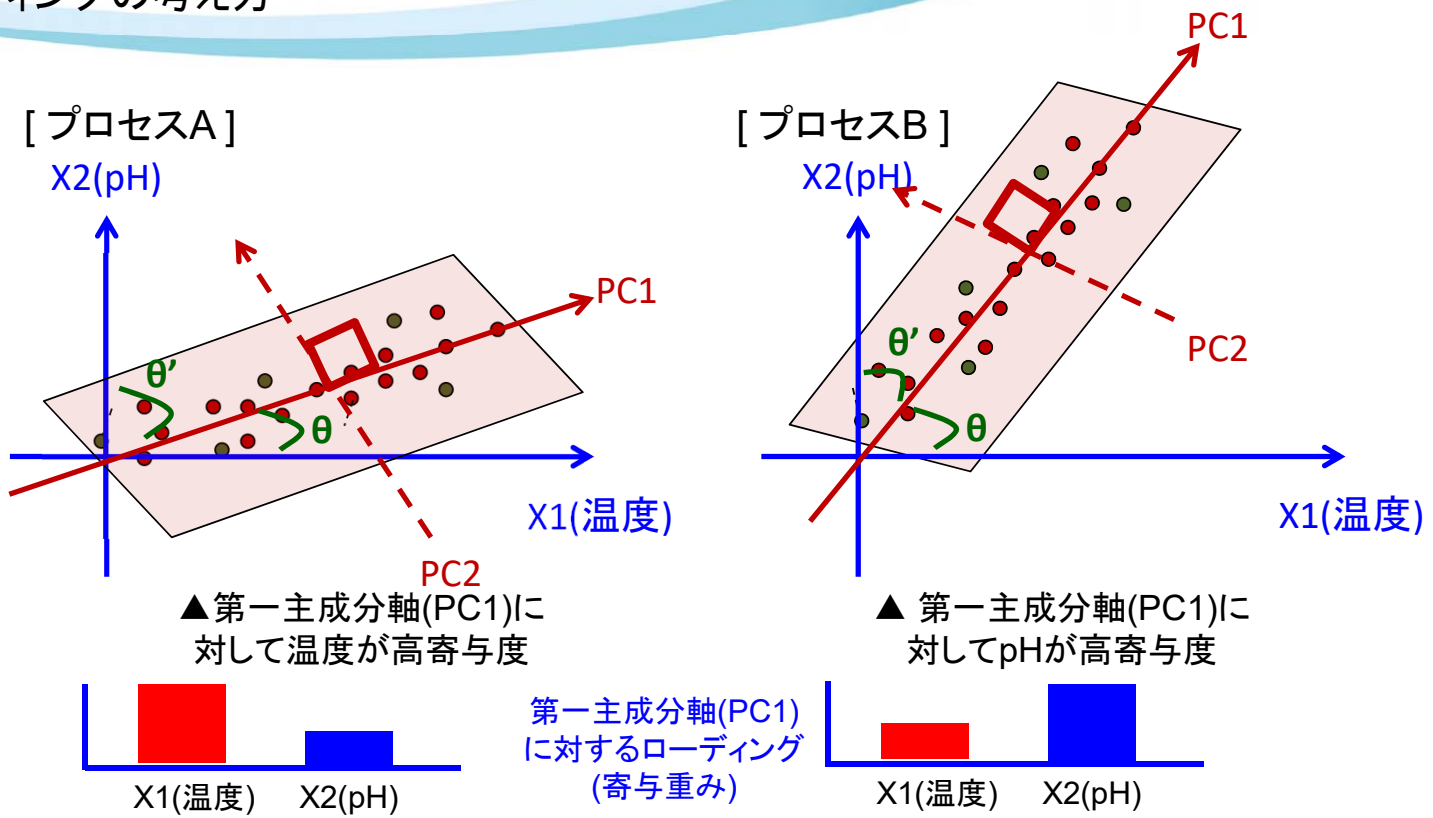
主成分を使って可視化



データの情報が縮約され、主成分の数が十分に少なくなれば各試料のスコアの値を使って可視化が可能になる

8

ローディングの考え方



寄与度＝ローディング(重み)は角度θより計算される。

9

主成分分析(PCA)用語集

主成分

データの変化を記述する因子 ”latent variables”、”factors”、 ”eigen vectors”

スコア, T

サンプルの写像: 元の変数空間から, 主成分によって表現される新しい空間上にサンプルをプロット

ローディング, P

変数の写像: 主成分に寄与する変数

残差, E

主成分によって表現することが出来ない要素: $X = X_{struct} + E$

分散

残差分散 – Eに残る分散(情報)

因子寄与 – X_{struct}によって記述された分散

モデル式:

$$X = TP^T + E$$

structure residual

多変量スペクトル分解法
(MCR: Multivariate Curve Resolution)

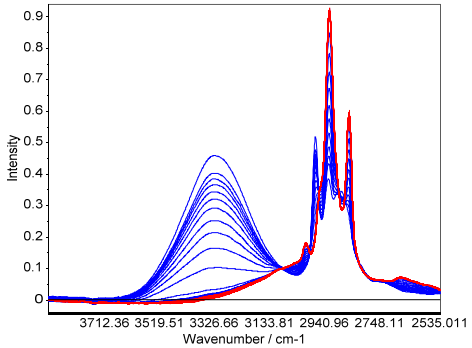
02

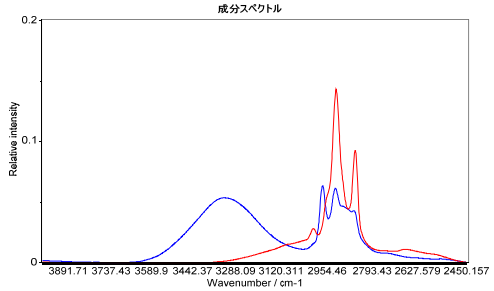
11

多変量スペクトル分解法

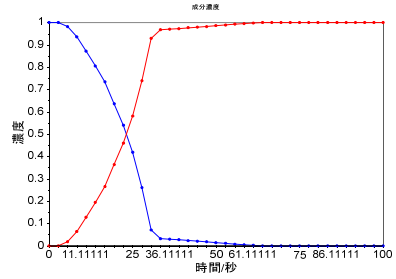
目的:

重なりあったスペクトルのピークを分離する。
ソフトウェアクロマトグラフ。





どの成分
(化学種)が？

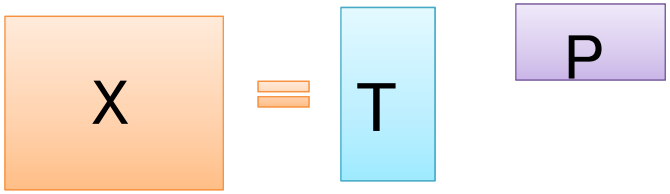


どれだけ？

12

主成分分析 vs多変量スペクトル分解

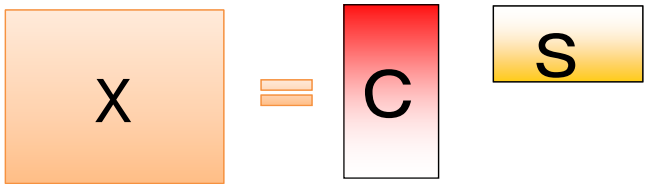
主成分分析



スコアT
　　サンプル間の類似度

ローディングP
　　変数の化学的寄与

多変量スペクトル分解



濃度プロファイルC
　　サンプル間の相対濃度

純成分スペクトルS
　　各化学成分の純スペクトル

13

計算手順

$$X = C\,S^t$$

$$C = X\,S\,(S^t\,S)^{-1}$$

$$S = X^t\,C\,(C^t\,C)^{-1}$$

Alternating Least-Squares (ALS)

適当な濃度Cかスペクトルsの情報を与えて、何度も繰り返し計算をすればC、Sは近似解に収束。

C、Sの初期値は主成分分析等から計算できるが、既知のスペクトルや濃度情報を導入する方がよい。

14

多変量スペクトル分解

化学種

試料の中に含まれる化学成分

濃度プロファイル, C

化学種の量

純成分スペクトル, S

化学種が単一ではかられた時の純スペクトル

残差, E

モデルによって表現することが出来ない要素: $X = X_{struct} + E$

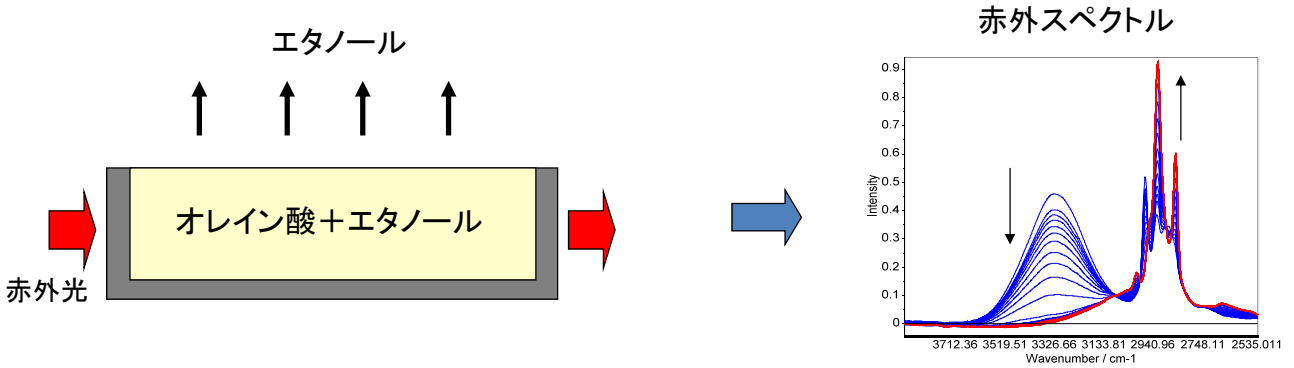
モデル式:

$$X = CS^T + E$$

structure residual

15

MCR事例: オレイン酸−エタノール 混合溶液のモニタリング 1/4

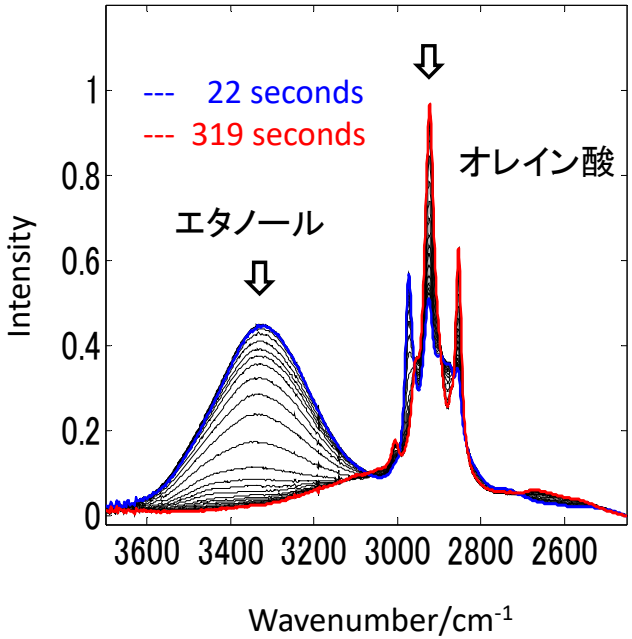


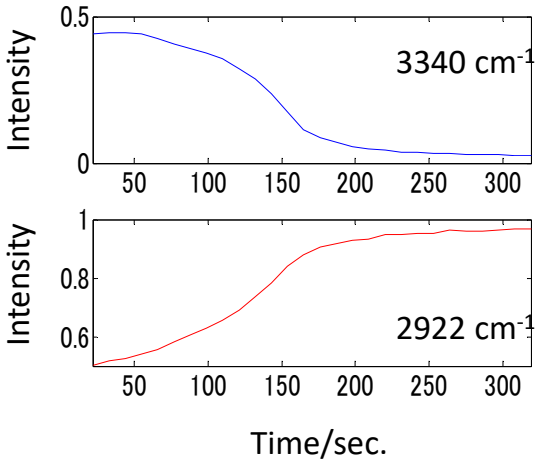
揮発によってエタノールの濃度は減少する，オレイン酸の濃度は増加する
⇒測定中にオレイン酸とエタノールの量がどのように変化するのか？

16

MCR事例: オレイン酸−エタノール 混合溶液のモニタリング 2/4

オレイン酸-エタノール溶液の時間変化をATR-IRで測定
時間の経過に伴いエタノールが揮発し、オレイン酸濃度が高くなる

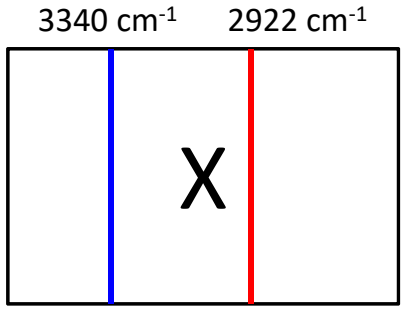




これらのバンド強度の変化はエタノールとオレイン酸の大まかな量的変化を表している
⇒初期値として適している

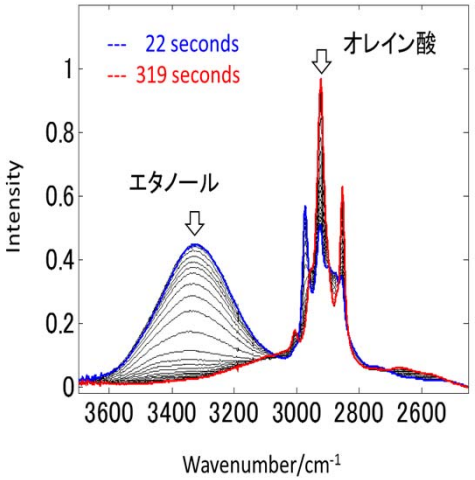
17

MCR事例: オレイン酸–エタノール 混合溶液のモニタリング 3/4



データシートの
2列の値だけを
取り出す

3340 cm^{-1}	2922 cm^{-1}
0.439446	0.503756
0.445404	0.515698
0.444961	0.524694
0.438475	0.53911
0.42484	0.55565
0.407304	0.58275
0.389819	0.606698
0.375	0.627466
0.355164	0.653967
0.322372	0.690764
0.285562	0.737861
0.236851	0.780861
0.173458	0.841757
0.112871	0.878212
0.086219	0.903949
0.069348	0.915652
0.057299	0.928937
0.049136	0.93209

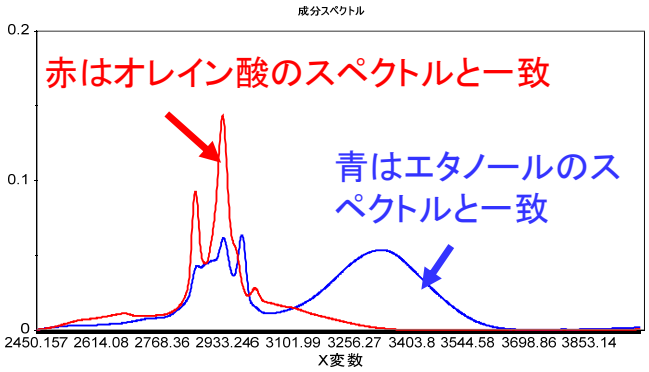


結晶多形の転移などピークシフトが起こる際の変局点考察、存在比考察、プロセス終点考察などへの応用が可能。

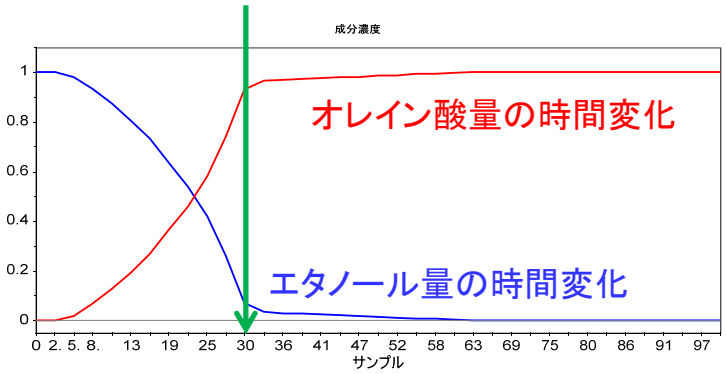
18

MCR事例: オレイン酸–エタノール 混合溶液のモニタリング 4/4

純成分スペクトル

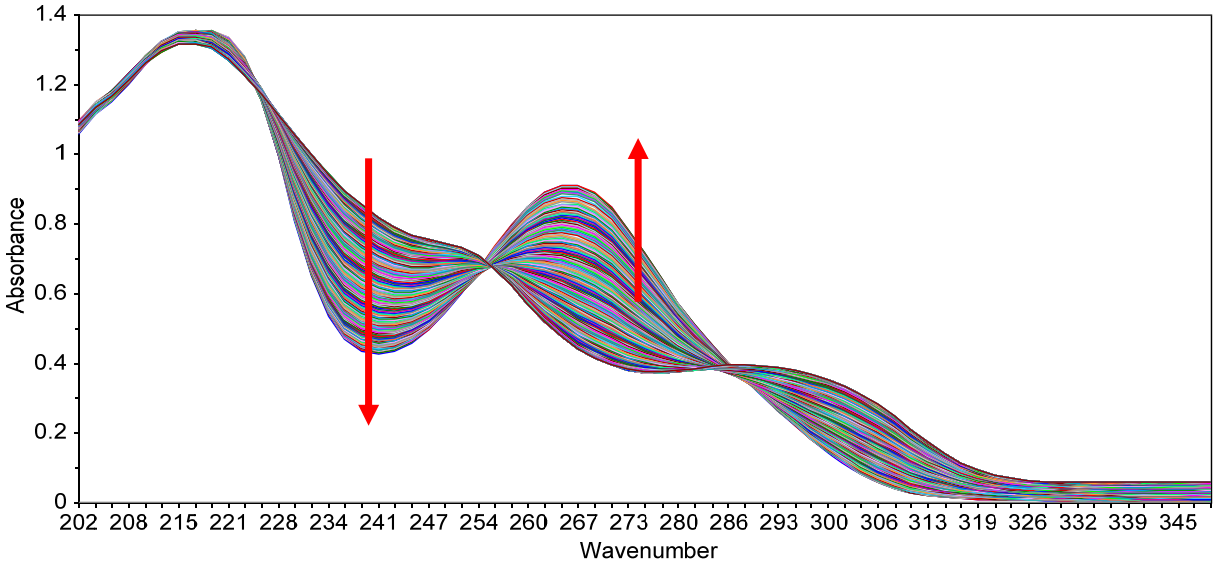


濃度プロファイル



- この2つのスペクトルはオレイン酸とエタノールの純スペクトルに非常によく一致。
- 濃度プロファイルを見ると、測定開始t直後にエタノールの蒸発が急激に始まり30秒程度で大部分が蒸発してしまっていることが分かる。

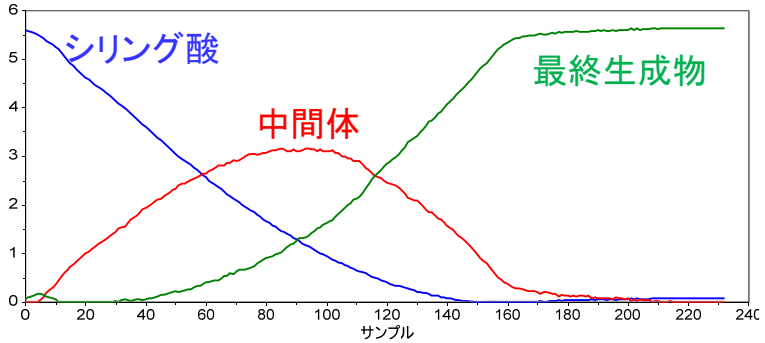
MCR事例:シリングの酵素分解 1/2



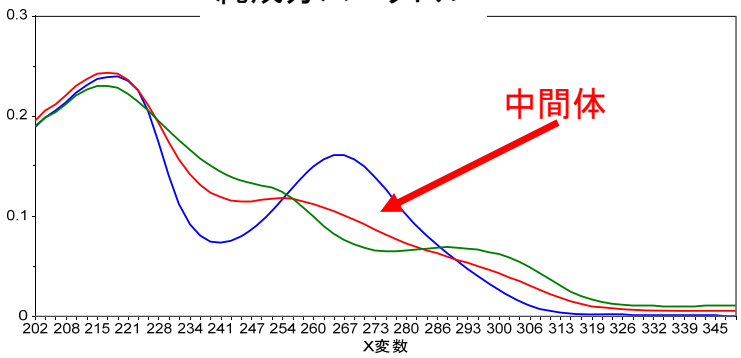
シリングに酵素を加えて分解
分解では中間体を経て最終生成物に
⇒反応途中でどのような成分が発生しているか？

MCR事例:シリングの酵素分解 2/2

濃度プロファイル



純成分スペクトル



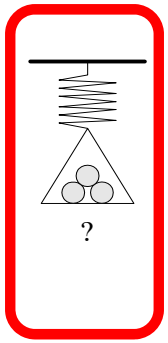
21

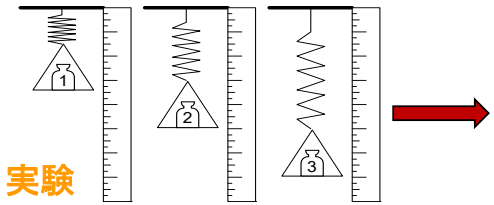
各回帰式の考え方

03

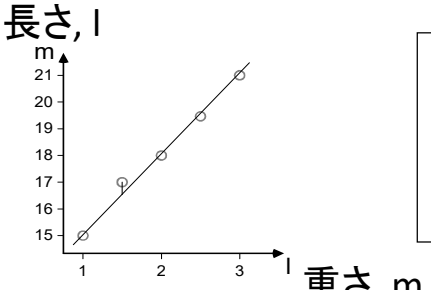
- ULR(単回帰)
- MLR(重回帰)
- PCR(主成分回帰)
- PLSR(部分最小二乗回帰)

回帰モデルとは？





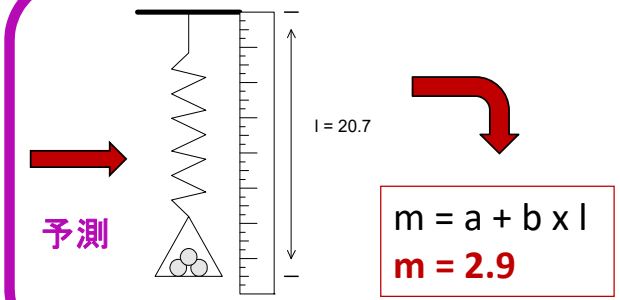
重さ, m	長さ, l
1.0	15.0
1.5	17.1
2.0	18.0
2.5	19.5
3.0	21.0



$$m = a + b \times l$$

$a = -4$

$b = +0.33$

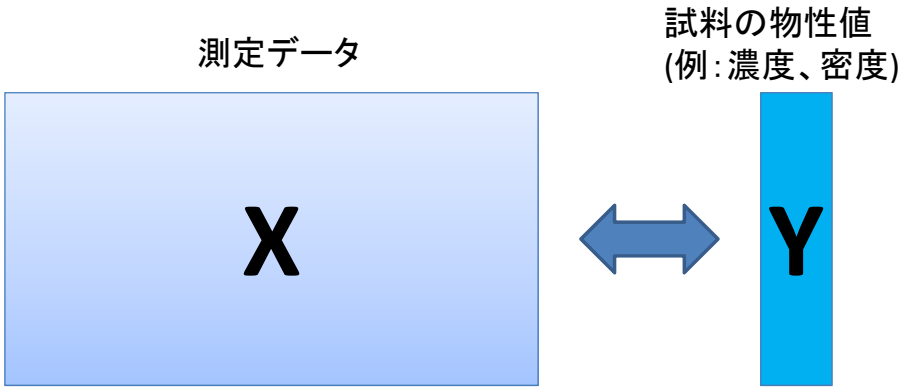


$$m = a + b \times l$$

$$m = 2.9$$

23

回帰分析



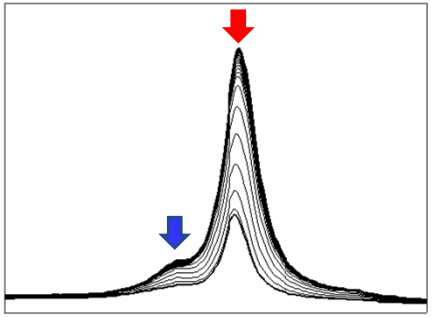
$$x_1 b_1 + x_2 b_2 + \text{-----} + x_n b_n = y$$

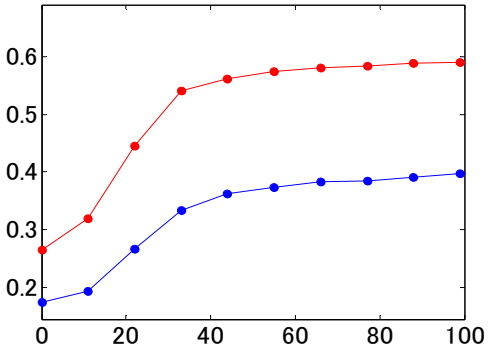
もし測定したデータが対象の性質を上手く捉えていれば、上式を満たすような
b（回帰係数）が存在

24

共線性とは

2つの説明変数が互いに似たような情報をもつこと





試料の中の成分の濃度が増加するとスペクトル中のピークの高さも増加する

2つの変数、赤と青におけるスペクトルの値の変化は非常によく似ている

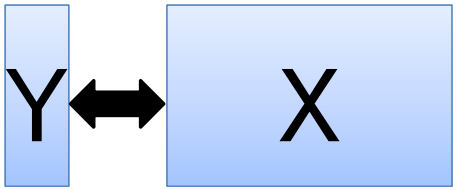
濃度 $y = x_1 b_1 + x_2 b_2$

x_1：スペクトル値：赤　　x_2：スペクトル値：青

このような性質を持つ変数を使って得られた回帰係数は非常に不安定で精度が低くなる⇒情報が重複していない変数を選んで回帰式を計算する必要がある

25

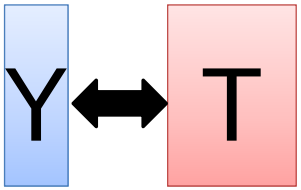
重回帰分析



$$y = x_1b_1 + x_2b_2 + \text{-----} + x_nb_n$$

実際には互いに情報の重複がないxを選ぶのは難しい・・・

主成分帰分析

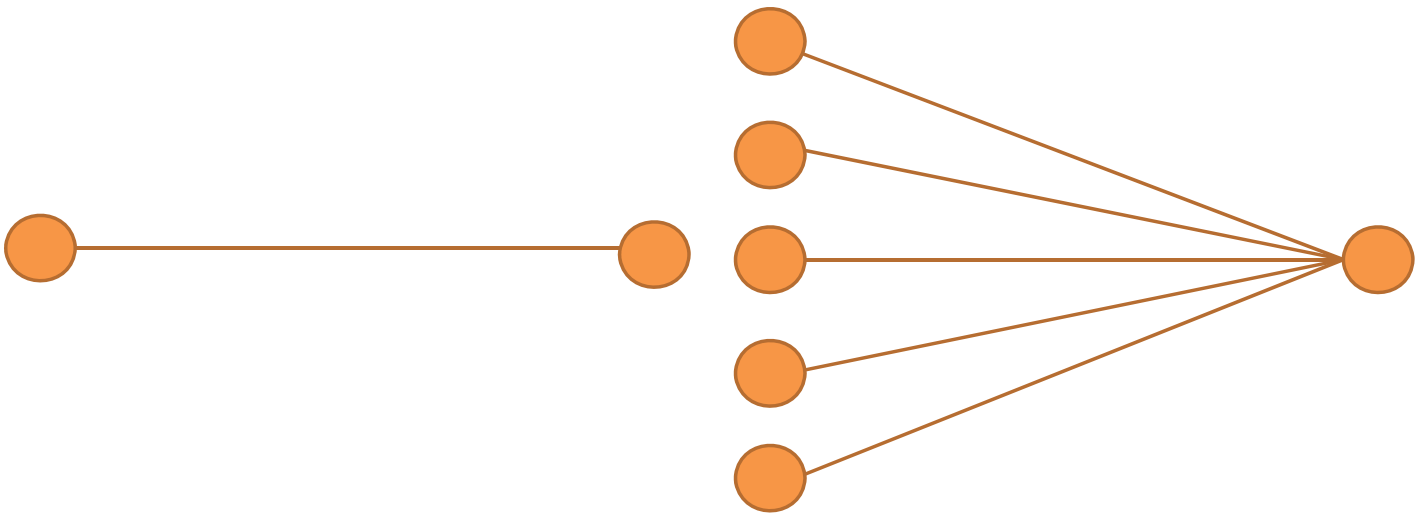


$$y = t_1b_1 + t_2b_2 + \text{-----} + t_Ab_A$$

主成分tの間には情報の重複がない（直交）のでxの代わりにスコアの値を使えばよい

26

ＵＬＲ（単回帰）　ＭＬＲ（重回帰）

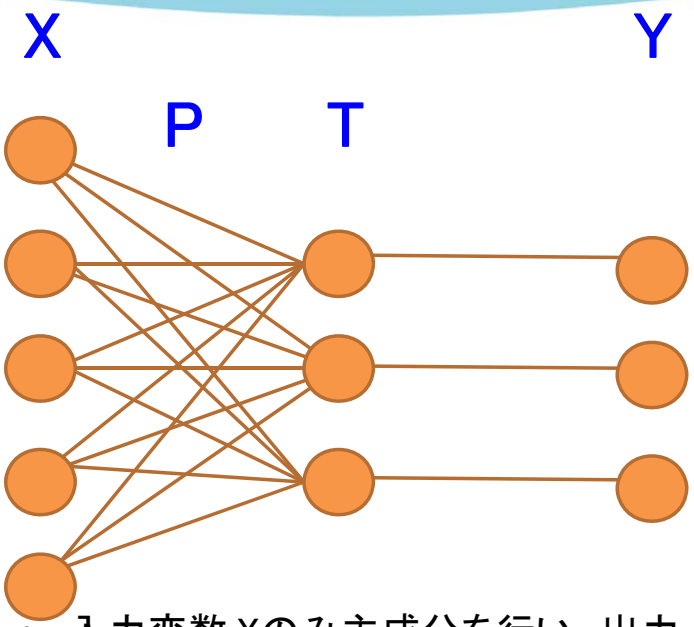


- 入力変数:X、出力変数Yとも単変数のみ設定可能。
- スコア(T、U)、ローディング(P、Q)は考慮せず。

- 複数の入力変数:X、単一の出力変数Yが設定可能。
- 入力変数:Xに共線性が無いことが使用の絶対条件。
- スコア(T、U)、ローディング(P、Q)は考慮せず。

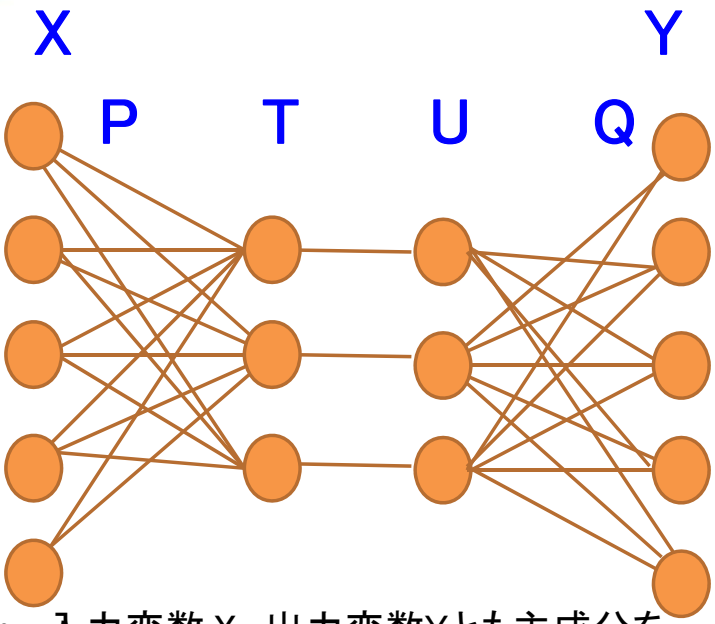
27

ＰＣＲ（主成分回帰）



- 入力変数:Xのみ主成分を行い、出力変数Yに対して回帰。
- 入力変数:X、出力変数Yとも複数の変数が設定可能。
- 共線性の回避が可能。

ＰＬＳＲ（部分最小二乗回帰）

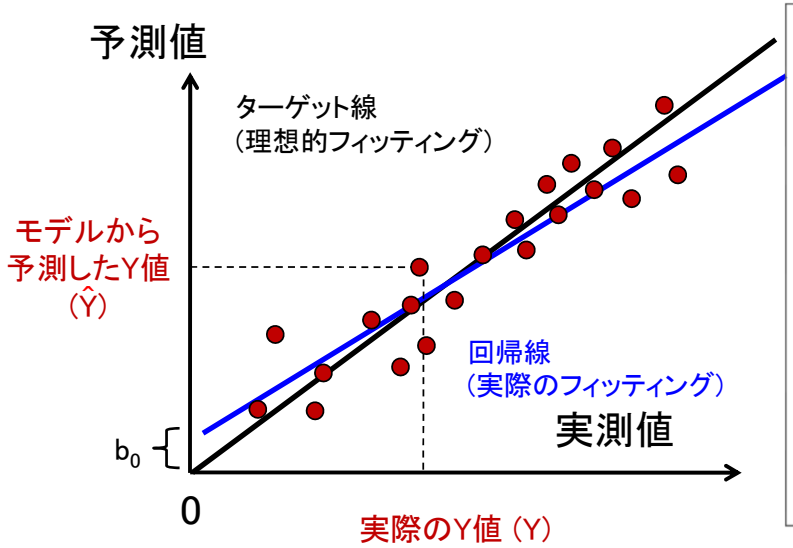


- 入力変数:X、出力変数Yとも主成分を行い、互いのスコア値に対して回帰。
- 入力変数:X、出力変数Yとも複数の変数が設定可能。
- 共線性の回避が可能。

28

予測vs実測プロット

回帰モデルによる予測値と実際の値との差



統計値の概要

- 傾き(Slope): 回帰線の傾き
- 切片(Offset): 回帰線のb_0値
- 相関係数(Correlation): YとYの相関値
- 決定係数(R^2):相関係数の二乗値
- RMSEP (RMSEC):予測(またはキャリブレーション)の実測値単位から計算される誤差値
- 予測標準誤差(SEP):計算結果の標準誤差。残差の標準偏差に等しい
- バイアス(Bias): 残差の平均値

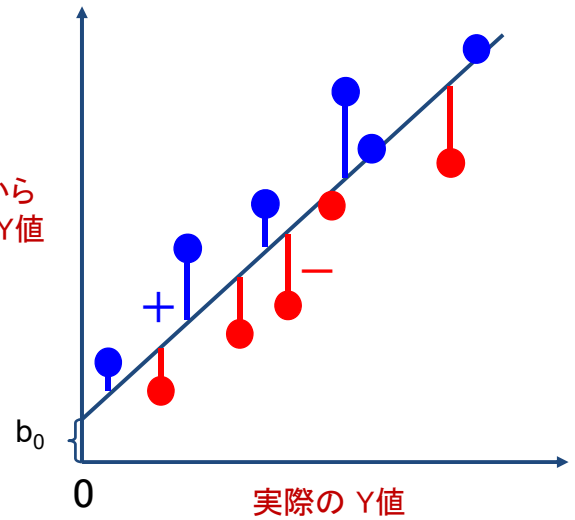
29

誤差の考え方

残差 = $y_i - \hat{y}_i$

残差分散 = $\dfrac{\sum_{i=1}^{N} (y - \hat{y}_i)^2}{N}$

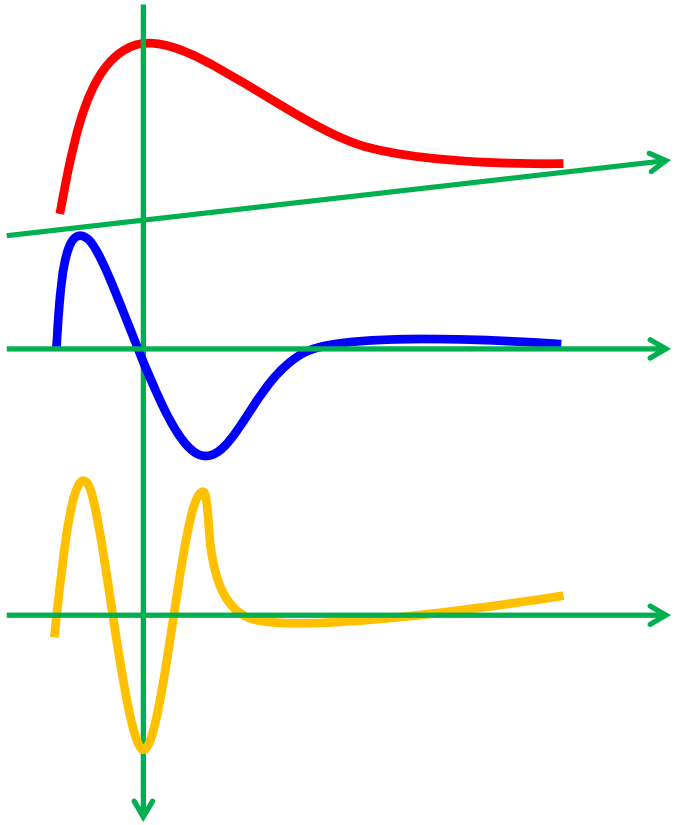
RMSEP = $\sqrt{\dfrac{\sum_{i=1}^{N} (y - \hat{y}_i)^2}{N}}$



予測値標準偏差(Root Mean Squared Error of Prediction)

30

スペクトル解析における微分処理の考え方



微分処理のベースライン補正効果

吸光度：Y=A+BX
ベースラインがサンプルにより変動
吸光度は主に粒形、厚みなどの物理的情報解析に向くが、ベースラインの変動を抑えるなど何らかの前処理は考慮する必要がある。

一次微分：Y=B
基準が光散乱情報Bに揃う
一次微分は粒形、厚みなどの物理的情報を残しながら併せて成分由来の帰属情報による解析も行いたい時に良好な結果が一般的に得られる。

二次微分：Y=0
ベースラインを理論上完全補正
二次微分は粒形、厚みなどの物理的情報を理論上除去し、成分由来の帰属情報による解析を行いたい時に良好な結果が得られる。

31

クロスバリデーションの考え方

キャリブレーションのみ(全サンプルスコア値利用)の場合、
データのバラつきがあっても見かけ上検量線精度が高くなる。

1) ×2 3 4 5 6 7 8 9

2) 1×3 4 5 6 7 8 9

|

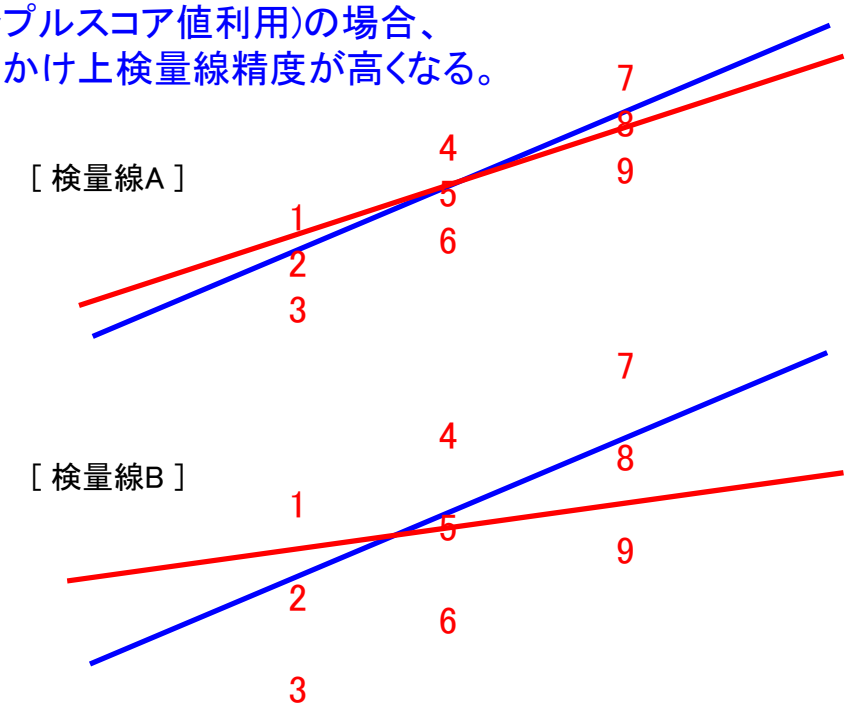
|

|

7) 1 2 3 4 5 6 ×8 9

8) 1 2 3 4 5 6 7 × 9

9) 1 2 3 4 5 6 7 8 ×



見かけ精度を回避するため任意のサンプルを入れ替え、検量線精度の
バリデーション(検証)を行なう。実際の精度に近い値が得れる。

32

PLSR(部分最小二乗回帰法)用語集

スコア: (X-スコア: T, Y-スコア: U) 主成分空間でのサンプルの写像

ローディング: (X-loadings: P, Y-loadings: Q) 主成分の
科学的な寄与度の意味

残差: (X-residuals: E, y-residuals: F) モデルとのズレ

分散: 残差 / 自由度

モデル式: $X = TP^T + E, \quad Y = UQ^T + F'$

回帰係数: $Y = B_0 + X_1*B_1 + X_2*B_2 + \ldots + X_N*E_N$

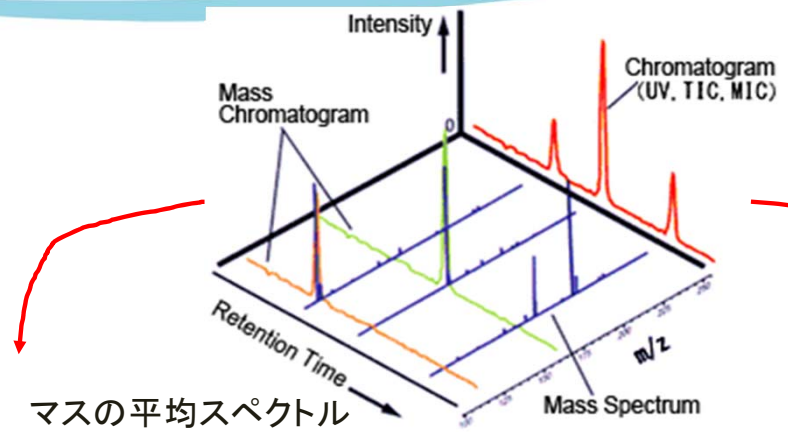
33

3Dデータの取扱い

04

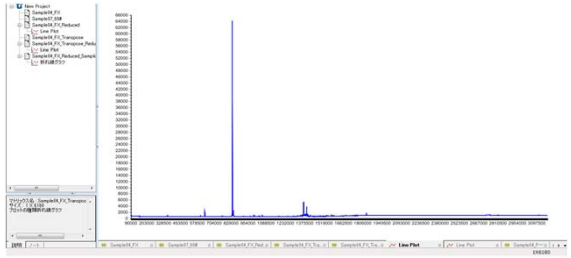
34

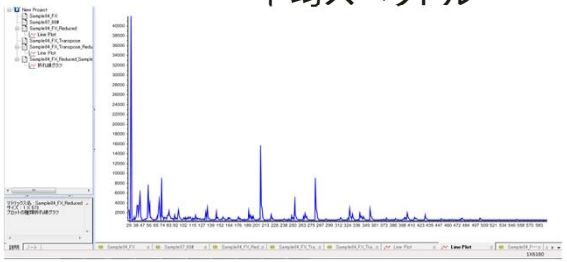
各スペクトルの平均値を計算



マスの平均スペクトル

クロマトグラフの
平均スペクトル



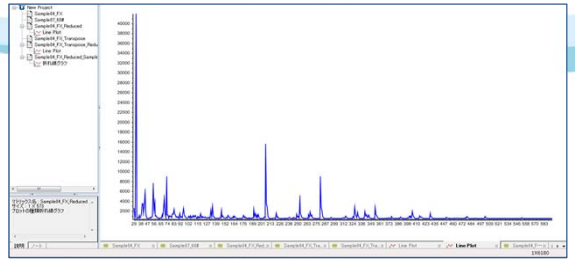


Sample A | Data Set

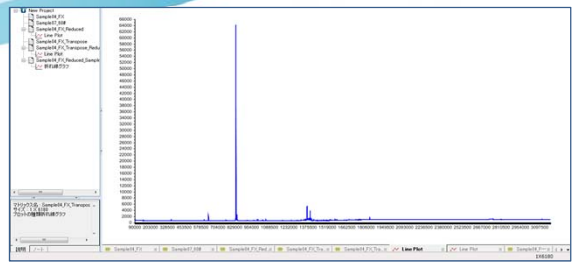
Sample A | Data Set

35

株式会社クオリティデザイン: 多変量統計解析とは

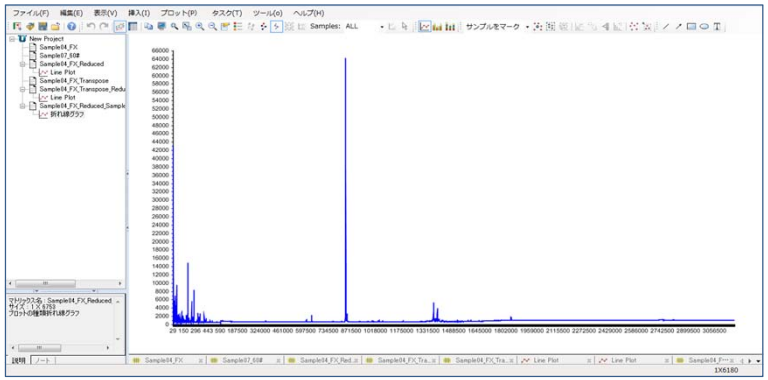


結合



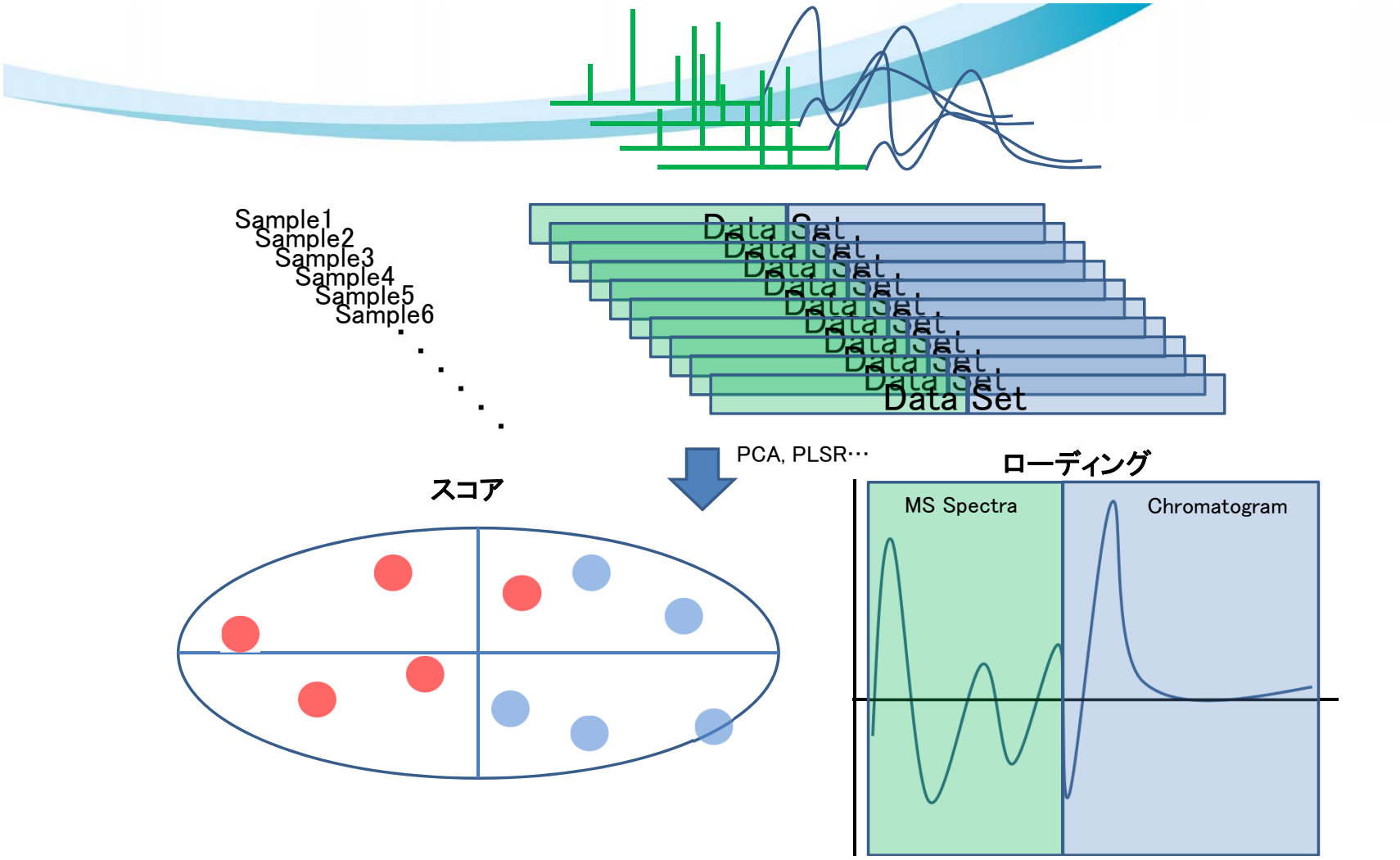
Sample A | Data Set

Sample A | Data Set



Sample A | Data | Set

36


Sample1
Sample2
Sample3
Sample4
Sample5
Sample6
Data Set
PCA, PLSR…
スコア
ローディング
MS Spectra
Chromatogram

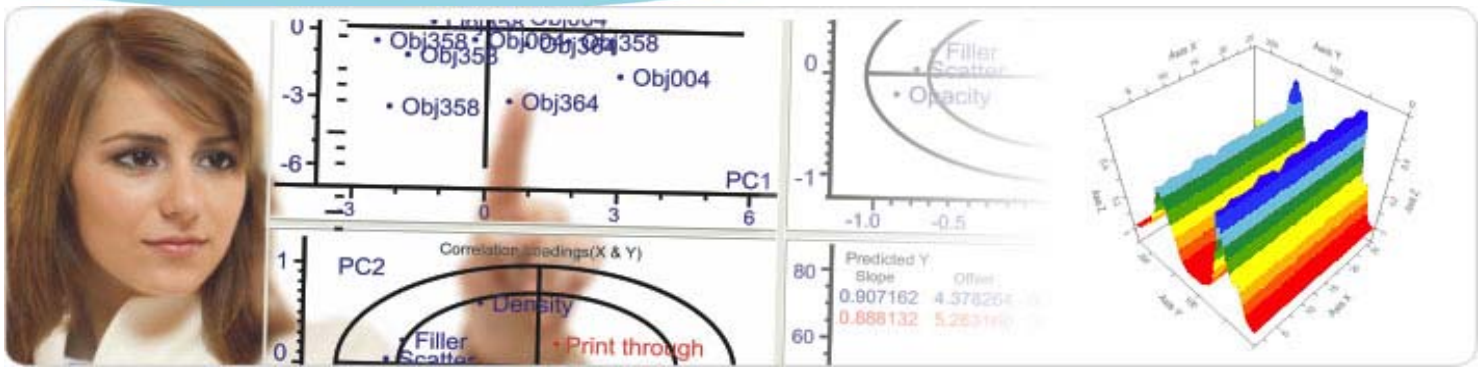
37

トレーニングセミナーのご案内

05

38

多変量統計解析トレーニングコースワークショップ



- コース名
  「 多変量統計解析の理論と実例 レベル1、レベル2 」

- トレーニングコースの内容
  本コースは基礎や原理の土台をしっかりと作るのに役立ちます。理論の講義の他(約4割)、実データを用いた解析(約6割)も行います。講義は産総研所属の多変量統計解析専任講師が担当。

- 開催場所、日程
  東京、大阪にて年3、4回の開催予定。

39