可視化手法による非線形多変量データ解析

可視化手法による非線形多変量データ解析

林　恭平*，大西　章**

Nonlinear Multivariate Data Analysis by Visualizing Methods

Kyohei HAYASHI* , Akira ONISHI**

要　旨

多変量解析法は化学分野においても応用できる問題は多い．しかし，多変量解析手法は基本的に線形手法であり，報告されている事例もその範囲内のものがほとんどである．一方，化学分野問題には非線形特性を有するものも少なくない．階層型ニューラルネットワーク手法は非線形相関が可能であるが，多変量データの場合については信頼性のある相関面を得るには膨大なサンプル数を必要とするために，その実用性は限定的である．したがって，非線形多変量データの相関分析には次元の低減化が不可欠である．本稿では，多次元データの次元圧縮法として，従来法である主成分分析手法に加えて新しいニューラルネットワーク手法の一つである自己組織化マップを併用して，非線形多変量データの可視化による相関解析を行い，その実用性を示した．

キーワード：多変量解析法，非線形相関，可視化，主成分分析，自己組織化マップ

1．はじめに

相関分析，モデリング，パターン認識などの多変量解析手法の化学分野問題への応用に関する研究分野「ケモメトリックス」が Kowalski らによって 1974 年に提唱され，近年では新しい知識情報処理手法の導入も意欲的に進められている．我が国では 1990 年代になって参考書が 2,3 発刊されているものの，初心者向きのいわゆる「オールインワン」タイプの教科書は見あたらず，実際的に取り組むには複数の参考書とフリーソフトを入手する必要があり，普及にとって障害となっている．そこで，著者らは，ケモメトリクス技術の普及を目的に，統一されたエクセルシート様式で各種多変量解析手法を手軽に体験できるエクセルマクロを開発し，これを利用して知識情報処理手法を含む統計解析法の有用性を実感できる例題を検討した．

多変量解析法は化学分野においても，化学製品の特性分析，機器分析データの自動処理，プロセス操業データ解析と最適化，プロセス異常診断など応用範囲は広く，その有用性を示す事例集も発刊されている．しかし，体系化されている多変量解析手法は基本的に線形手法であり，報告されている事例もその範囲内に限られたものがほとんどである．一方，化学分野問題には非線形特性を有し，データが統計学的解釈に必要な条件を満たしていない場合も少なくない．階層型ニューラルネットワーク手法は非線形相関が可能なことから，その応用が一時注目されたが，多変量データの場合については信頼性のある相関面を得るには膨大なサンプル数を必要とするため，モデリング手法としての実用性は低い．したがって，非線形多変量データの相関分析には次元の低減化が不可欠である．本稿では，多次元データの次元圧縮法として，従来法である主成分分析手法に加えて新しいニューラルネットワーク手法の一つである自己組織化マップを併用して，非線形多変量データの可視化による相関解析を行い，その実用性を示した．

2010 年 8 月 20 日受理

* 2008 年度工業化学科卒業生
(Graduation of Industrial Chemistry)
** 総合工学システム学科　物質化学コース
(Dept. of Industrial Systems Eng. :Materials Chemistry Course)

2．多変量データの相関解析手法

2.1 重回帰分析

多変量データにおいて，サンプル数を n，説明変数の数を p，目的変数を y_i (I=1,2,…,n)，説明変数を x_{ij} (i=1,2,…,n)(j=1,2,…,p)とすると，目的変数と説明変数間の相関関係を次式で表現するのが重回帰分析法である．

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i \quad (i = 1,2,\cdots,n) \tag{1}$$

ここで，$b_1,b_2,\ldots,b_p$は偏回帰係数，e_iは残差である．偏回帰係数の値は残差の二乗和Σe_i^2を最小にする条件から，行列・ベクトル表現で次式で求められる．

$$\mathbf{b}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2}$$

回帰式相関においては，偏回帰係数が説明変数の種類や単位に左右されないように，各変数をその平均と標準偏差を用いて平均値0，分散1に規格化して計算を行うのが普通であり，この場合の偏回帰係数を標準偏回帰係数という．なお，説明変数間に強い相関（共線性）がある場合は信頼性のある偏回帰係数が求められない．

2.2 主成分分析

主成分分析法は，p 個の特性値のもつ情報を，より小数の m(m≦p)個の総合特性値（主成分）に要約する手法である．すなわち，総合特性値 z_k (k=1,2,…,m)を重み係数l_{kj} (k=1,2,…,m)(j=1,2,…,p)を用いてp個の特性値の一次式で表現し，重み係数が下記の条件を満たすように定めるものである．

$$\left.\begin{aligned} z_1 &= l_{11}x_1 + l_{12}x_2 + \cdots + l_{1p}x_p \\ z_2 &= l_{21}x_1 + l_{22}x_2 + \cdots + l_{2p}x_p \\ &\cdots\cdots \\ z_m &= l_{m1}x_1 + l_{m2}x_2 + \cdots + l_{mp}x_p \end{aligned}\right\} \tag{3}$$

$$l_{k1}^2 + l_{k2}^2 + \cdots + l_{kp}^2 = 1 \quad (k = 1,2,\cdots,m) \tag{4}$$

条件：

1) 第1主成分 z_1 の係数 l_{1j} (j=1,2,…,p)は，(4)式を満たし，z_1 の分散が最大と成るように定める．
2) 第2主成分以後の主成分 z_k の係数 l_{kj} は，(4)式を満たし，$z_1,z_2,\ldots,z_{k-1}$ と無相関となる条件で，z_k の分散が最大となるように定める．

具体的な計算は，説明変数を x_{ij} (i=1,2,…,n)(j=1,2,…,p)とすると，説明変数の分散・共分散行列の固有値と固有ベクトルを求めると，固有値の大きい方から対応する固有ベクトルを主成分 $z_1,z_2,\ldots$ の重み係数とすればよい．

固有値 λ_k は主成分 z_k の分散に等しく，主成分 z_k の分散の総分散に対する割合を寄与率，主成分と元の特性値との相関を因子負荷量という．

主成分分析の手法を用いれば，多変量データの次元を縮約することができるため，比較的少ないサンプル数であっても非線形モデリングが可能となる．また，2次元または3次元への写像によって，多変量データの特性を視覚的にとらえることができる．

2.3 自己組織化マップ(Self-Organizing Maps : SOM)

主成分分析手法による線形写像法は縮約された主成分のいくつかの次元軸を取り出して，残りは切り捨てて視覚化するのに対し，自己組織化マップは元のデータ間のつながり・距離を維持しながら平面上の類似したニューロンの位置にマッピングする方法であり，多次元データの持つすべての情報を2次元に集約するものである．

SOM アルゴリズムの概要を以下に示す．

元の多変量データは，サンプル数を n とし，特徴数 p を持つ p 次元ベクトル $\mathbf{x}_k$ (k=1,2,…,n)とし，サイズ L×L のノード（格子点，神経細胞に相当）にマッピングするものとする．

1)各ノードのコードベクトル $\mathbf{m}_{11},\ldots,\mathbf{m}_{ij},\ldots,\mathbf{m}_{LL}$ をランダムに設定して初期化する．

2)逐次 $\mathbf{x}_k$ を提示し、最も類似したコードベクトル $\mathbf{m}_{ij}$ を探し、そのノードの位置を c とする．

ここで，類似度の測度はユークリッド距離とする．

$$\| \mathbf{x}_k - \mathbf{m}_c \| = \min \| \mathbf{x}_k - \mathbf{m}_{ij} \|$$

3)コードベクトルを次式にしがって更新する．ただし，t は繰り返し回数，r はノードの格子上の位置を表すベクトルである．

$$\mathbf{m}_{ij}(t) = \mathbf{m}_{ij}(t-1) + (1/t)\Phi(\| \mathbf{r}_{ij} - \mathbf{r}_c \|)(\mathbf{x}_k - \mathbf{m}_{ij})$$

ここで，Φ(p)は近傍関数で，

$$\Phi(p) = \exp\left(-\frac{p^2}{2\sigma^2(t)}\right)$$

あるいは，

$$\Phi(p) = \begin{cases} 1, & p \le d(t) \\ 0, & p > d(t) \end{cases}$$

などの関数が用いられる．σ(t)と d(t)は格子上での近傍の広さを定義する時変のパラメータで，学習の初期には格子全体とし，学習が進むにつれて序々に縮小して，最終的には隣接するノードの値になるようにする．

2),3)の操作を数千回程度繰り返すことによってマップが完成する．

2次元マップ上には類似度に応じたクラスターが形成されるが，クラスターの境界では隣接したノードで

あってもコードベクトルの差異は大きい．これを視覚化して確認するには，次式で求められる関数 f(i,j)を計算して等高線表示させた密度マップが有用である．

$$f(i,j)=\sum_{k,l}\left\| \mathbf{m}_{ij}-\mathbf{m}_{kl} \right\|^2$$

(k,l : i,j に隣接する格子添字)

3．検討に用いたデータ

本検討で使用したデータは文献4)から引用したもので，清涼飲料びん用アルミ合金キャップの生産ラインにおける200ロットの割れ特性評価データである．このデータは各ロットごとに含有成分割合，製造温度条件，不良個数 r，生産個数 n が記録されているものである．データの一部を表1に示す．表1で，Si～Ti は合金構成元素の含有割合%，HRIN,HTIN,HTOUT はそれぞれ圧延機 A の開始温度，圧延機 B の開始温度と終了温度，不良率は，不良個数を生産個数で割って1万倍して ppm 単位で表したものである．

表1　使用データ

Si	Fe	Cu	Mn	Mg	Cr	Zn	Ti	HRIN	HTIN	HTOUT	不良率
0.132	0.545	0.135	0.405	0.427	0.001	0.063	0.008	520	399	205	0
0.12	0.526	0.144	0.396	0.417	0.003	0.063	0.012	499	418	205	0
0.128	0.534	0.138	0.377	0.445	0.002	0.063	0.015	516	432	205	32.4
0.131	0.567	0.135	0.402	0.466	0.001	0.046	0.01	481	402	205	0
0.134	0.562	0.135	0.402	0.436	0.006	0.144	0.013	475	397	206	0
0.134	0.547	0.151	0.391	0.444	0.005	0.044	0.012	509	411	206	0
0.129	0.603	0.146	0.368	0.441	0.004	0.045	0.011	529	443	206	81.0
…	…	…	…	…	…	…	…	…	…	…	…

4．検討結果と考察

4.1 重回帰分析

目的変数を不良率，説明変数を含有成分割合と製造温度条件として，説明変数の逐次選択法により得られた重回帰式の結果を表2と図1に示す．図1より明らかなように非線形性が強く，本回帰式を予測式として使用することはできない．

表2　標準偏回帰係数

HTIN	0.398	Cr	0.074
Mn	−0.261	Fe	−0.081
Zn	−0.174	Ti	−0.058
Si	0.200	HRIN	0.023
Cu	−0.124	HTOUT	0.017
Mg	−0.035		

自由度調整寄与率 R^{*2} =0.311

元の文献4)では，非線形性を緩和するために，目的変数を次のロジット変換して重回帰分析を行っている．

$$z=\ln\frac{r+0.5}{n-r+0.5}$$

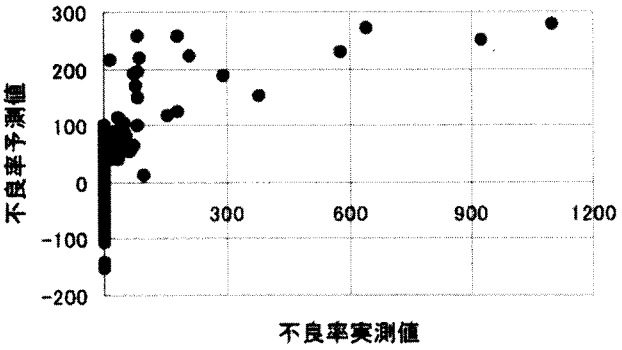

図1　重回帰式による不良率予測値

ここで，n,r　は各ロットごとの製造個数と不良個数である．この相関結果を表3と図2に示す．

表3　標準偏回帰係数（ロジット変換後）

HTIN	0.500	HTOUT	−0.022
Mn	−0.397	Cr	0.007
Si	0.290	Ti	−0.015
Zn	−0.086	HRIN	0.010
Cu	−0.048	Mg	0.015
Fe	−0.039		

自由度調整寄与率 R^{*2} =0.595

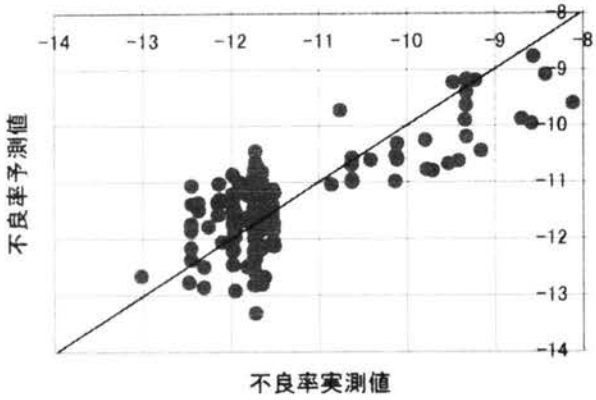

図2　重回帰式による不良率予測値
（ロジット変換後）

この変換後においても非線形性は残るが，元の報告ではこの結果をもとに不良率を下げるには，(a) HTIN を下げる，(b) Mn 量を上げる，(c) Si 量を下げる知見が得られたと結論している．しかし，(a)，(b)，(c) が単独で効果があるのかどうか，複合効果はどうか，コスト的に高価な Mn をどの程度にすればよいのかなどについては，試行錯誤的に追加の実験的検討を行っている．

4.2 自己組織化マップ

説明変数のすべてを用いた SOM によるパターン分類結果を図３に示す．不良ありのロットデータは左上から右下への対角線の左方に大まかには位置しているが，不良なしデータもランダムに存在し，分類は不十分である．

次に，元の説明変数には分類に不要なものが含まれるとの立場から，先の重回帰分析結果とパターン分類における各特徴のクラス識別力を示す Fisher 比[8]を勘案して，説明変数を HTIN, Mn, Si, Zn にしぼった SOM 検討結果を図４，５に示す．図４より，不良ありの２ロット（サンプル 74, 86）を除いて良好にパターン分類されており，４個の説明変数には良－不良に関する情報が十分に保持されていると解釈できる．図５は図４のクラスタリングに対応した密度マップである．SOM 学習では目的変数情報は使用しないが，右下に不良率の特に大きなロットデータが正しく分類されていることが判明した．図５左方は不良なしの領域であるが，高い等高線を示しており，ここに位置するサンプルが良－不良とは別な情報を持っていることを示唆している．

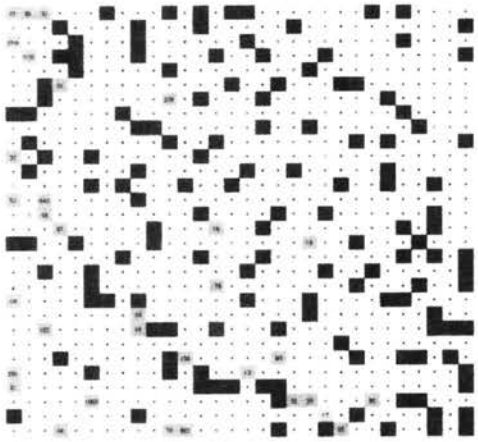

図3　SOMによるパターン分類結果
（説明変数：全データ）
■不良なし　□不良あり

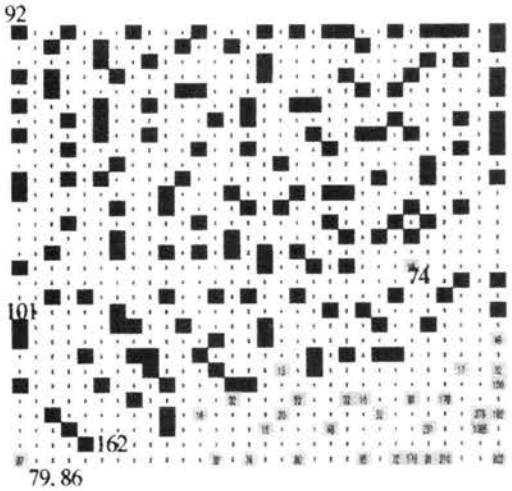

図4　SOMによるパターン分類結果
（説明変数：HTIN, Mn, Si, Zn）
■不良なし　　不良あり

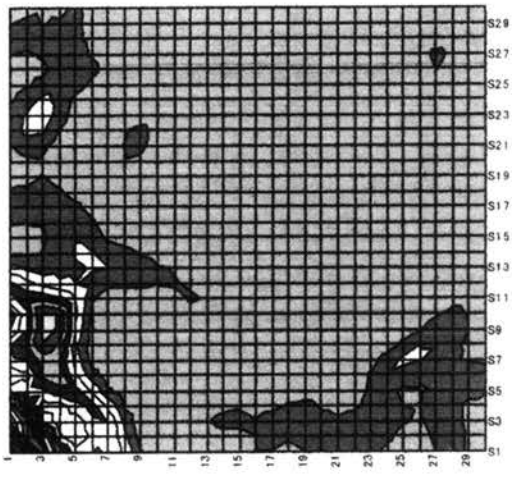

図5　SOMによる密度マップ
（説明変数：HTIN, Mn, Si, Zn）

4.3 主成分分析

主成分分析の結果を用いた Karhunen-Loeve プロットを図6，7に示す．先の SOM 検討で不良ありデータ領域から離れた位置にあったサンプル 74 は主成分分析では不良ありデータ領域の端，かつ，不良なしとありの判別が困難な位置にあることがわかる．また，サンプル86も主成分スコアZ1,Z3の端に位置していることがわかる．

一方，不良なしデータのうち，SOM による密度マップ上で左側に帰属されたものも，主成分分析ではクラスターの端に位置し，サンプル数の少ない領域であり，データの信頼性が高くないサンプルと解釈できる．

図8，9に主成分スコアに対応する不良率を3Dプロット図で示す．第1主成分スコア Z1 が 0〜3 の領域では比較的なだらかな曲面を示しているが，3 付近で急激に湾曲しており，線形手法では十分に解析できないことがわかる．

本主成分分析において，第1主成分から第3主成分までの累積寄与率[%]は，それぞれ36, 61, 85 であった．因子負荷量の各値は表4に示した．第1主成分の寄与率は特に高くはないが，不良率はほぼ第1主成分で決まり，第2主成分は Si の効果，第3主成分は Zn の効果を表しているといえる．

表4　因子負荷量

	Si	Mn	Zn	HTIN
Z1	−0.116	−0.776	0.354	0.827
Z2	0.989	−0.039	0.114	0.054
Z3	−0.088	0.358	0.914	−0.068

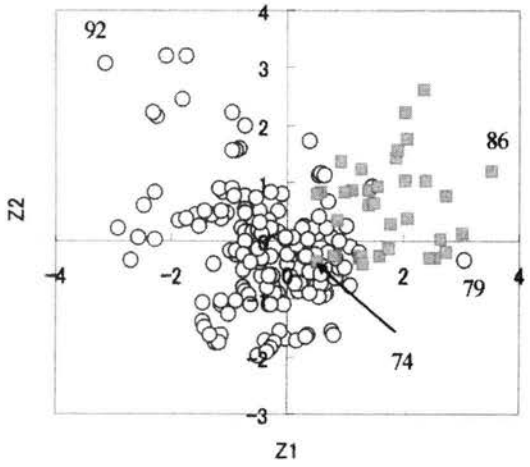

図6　K−L プロット(Z1−Z2)
○ 不良なし　　不良あり

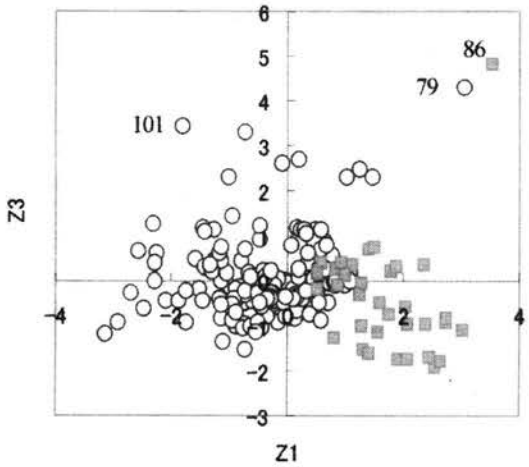

図7　K−L プロット(Z1−Z3)
○ 不良なし　　不良あり

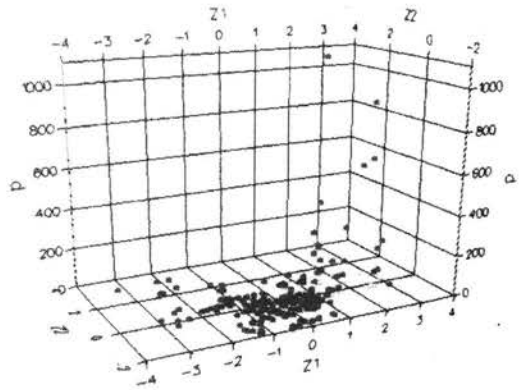

図8　3D プロット(Z1−Z2−不良率)

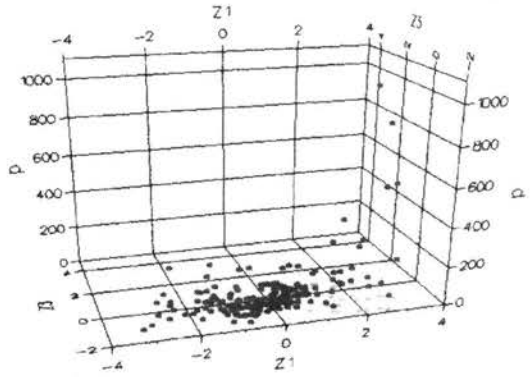

図9　3D プロット(Z1−Z3−不良率)

図１０－１３は各説明変数の主成分スコアZ1, Z2への影響を図示したものである．具体的には注目説明変数値を実測値の最大－最小間で十等分した値とし，他の説明変数値は実測値の平均値を用いて計算したものである．図１０，１１より，HTINを下げることとMnを増すことは，ほぼ同じ効果があり，製造コストを下げるには，HTINの設定値の検討が重要であることがわかる．また，図１２からZnを減らすことも検討の余地がある．Siの増減については信頼性の低い領域に入るので，実測値の平均の値が好ましいと結論される．

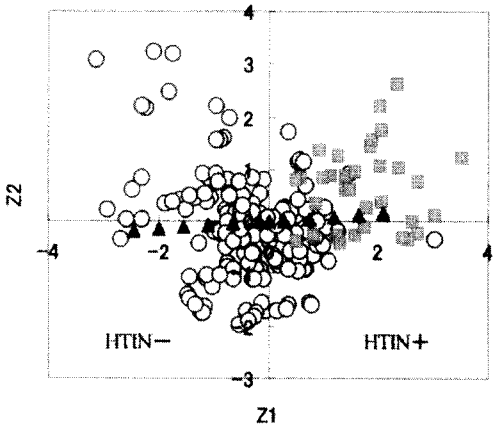

図１０　HTINのZ1, Z2への影響

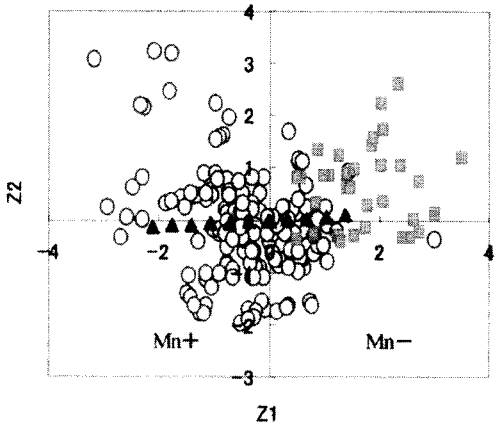

図１１　MnのZ1, Z2への影響

５．おわりに

重回帰分析では十分に分析できない非線形データの解析を可視化手法によって試みた結果，重回帰分析よりもより深い知見を得ることができた．可視化手法は非線形を有する多変量データの分析に有用であることが確かめられた．

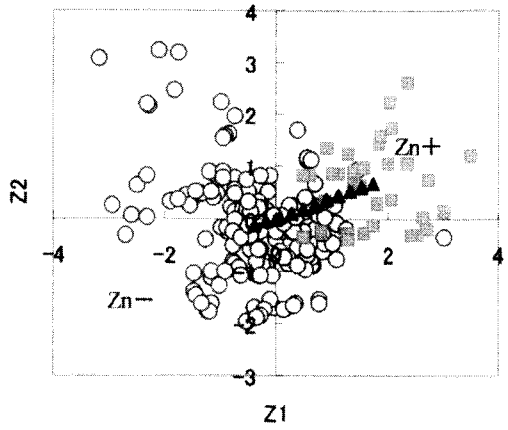

図１２　ZnのZ1, Z2への影響

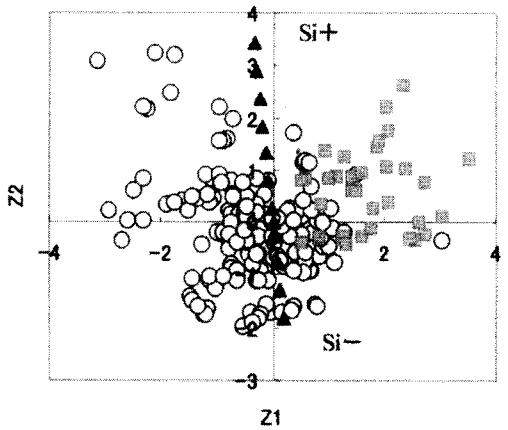

図１３　SiのZ1, Z2への影響

参考文献

1) 相島鐵郎：ケモメトリックス－新しい分析化学－，丸善(1992)
2) 宮下芳勝，佐々木慎一：ケモメトリックス　化学パターン認識と多変量解析，共立出版 (1995)
3) 佐藤寿邦，佐藤洋子：Excel VBAによる化学プログラミング，培風館 (2002)
4) 吉澤正，芳賀敏郎：多変量解析事例集第２集，日科技連 (1997)
5) Jure Zupan et al 著，田辺和俊監訳：化学者のためのニューラルネットワーク入門，丸善 (1996)
6) 徳高平蔵ほか：自己組織化マップの応用　多次元情報の２次元可視化，海文堂 (1999)
7) 坂和正敏，田中雅博：ニューロコンピューティング入門，森北出版 (1997)
8) 佐々木慎一ほか：化学者のためのパターン認識序説，東京化学同人 (1984)
9) 奥野忠一ほか：続多変量解析法，日科技連 (1976)