Journal of Surface Analysis Vol.26 No.1 (2019) pp. 10 - 21
吉原一紘 COMPRO12の使用法(3)
https://doi.org/10.1384/jsa.26.10

解説

COMPRO12の使用法（3）

吉原 一紘*
シエンタオミクロン（株）
140-0013 東京都品川区南大井6-17-20
** Kazuhiro.Yoshihara@ScientaOmicron.com*
（2019年5月13日受理; 2019年6月24日掲載決定）

COMPRO12の使用法(1, 2)ではCOMPRO12に搭載されているデータ処理法を解説した．本解説では引き続き，基本的なアルゴリズムの解説も含めてCOMPRO12の使用法を紹介する．

The Usage of COMPRO12 (part 3)

K. Yoshihara*
ScientaOmicron, Inc.
6-17-20, Minami-Oi, Shinagawa-ku, Tokyo 140-0013, Japan
** Kazuhiro.Yoshihara@ScientaOmicron.com*
(Received: May 13, 2019; Accepted: June 24, 2019)

The usage of COMPRO12 (part 1 and part 2) introduced the usage of data processing in COMPRO. In this lecture, the usage of COMPRO12 will be continuously explained with the basic introduction of algorithms used in COMPRO.

35. 角度分解データのシミュレーション

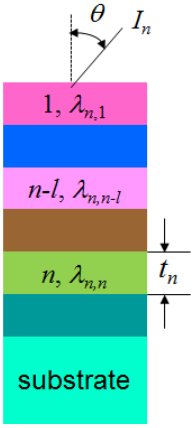

Fig. 131 Layer structure (color online)

Fig. 131に示すような多層薄膜物質のピーク強度の放出角度依存性をシミュレーションで求める．第n番目の層の遷移のピーク強度は次式で与えられる．ここでI_nは第n番目の層の遷移のピーク強度，I^0はn番目の層の遷移の成分濃度が100%の時の信号強度，c_nは成分濃度．t_nは第n番目の層の厚さ，$\lambda_{n,l}$は第n番目の層の電子が第l番目の層を通過するときのIMFP．θは放出角度である．I_nは

$$I_n = I^0 c_n\left[1-\exp\left(-t_n/\lambda_{n,n}\cos\theta\right)\right] \exp\left(-t_{n-1}/\lambda_{n,n-1}\cos\theta\right)\cdots\exp\left(-t_1/\lambda_{n,1}\cos\theta\right)$$

と表せる[1].

メニューバーの[Simulation] - [Simulate ARXPS]を選択すると，Fig. 132に示すように，デフォールトの膜構造を形成する各成分（遷移）の深さ方向の濃度分布（上図）とピーク強度の角度依存性（下図）が表示される．デフォールトの構造は，基板の上に２個の成分（遷移）が均一に混ざった膜厚2.00 nmの薄膜である．薄膜中の各遷移の濃度分布は図中に点で示してある．各遷移の結合エネルギーは全て500 eV，相対感度は1としている．遷移は特定されていないのでIMFPは 2.5 x ([kinetic energy of electron (eV)] / 1000) ^ 0.75 (nm)という簡略式[2]から計算する．デフォールトの測定条件と膜構造に基づくピーク強度の放出角依存性が求められる．測定条件は，線源はAl，

acceptance angleは40度，測定の相対誤差は5%，仕事関数は4.5 eVである．

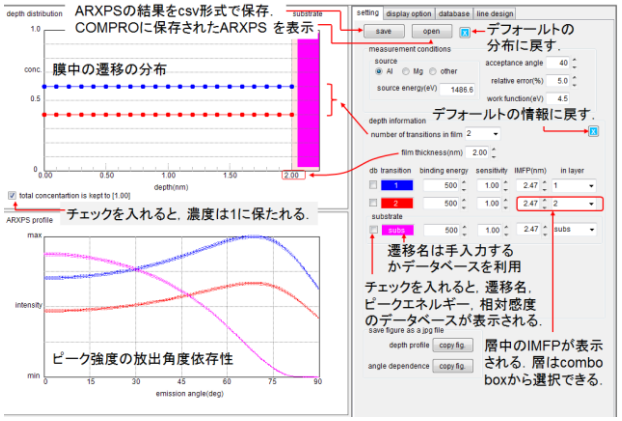

Fig. 132. Angle dependence of intensities of default setting. (color online)

遷移の濃度分布の設定

Fig. 133に示すように，遷移の濃度を示す分布図の点をマウスで囲んで目的の濃度に移動させることにより，分布を設定できる．[total concentration is kept to [1.00]]にチェックを入れると，合計濃度が1.00になるように他の遷移の濃度も自動的に変更される．一つの点だけを動かしたい場合には，その点だけをクリックして移動すると，当該点だけを変更できる．遷移の濃度分布が変更されると，それに対応する角度分解データが表示される．

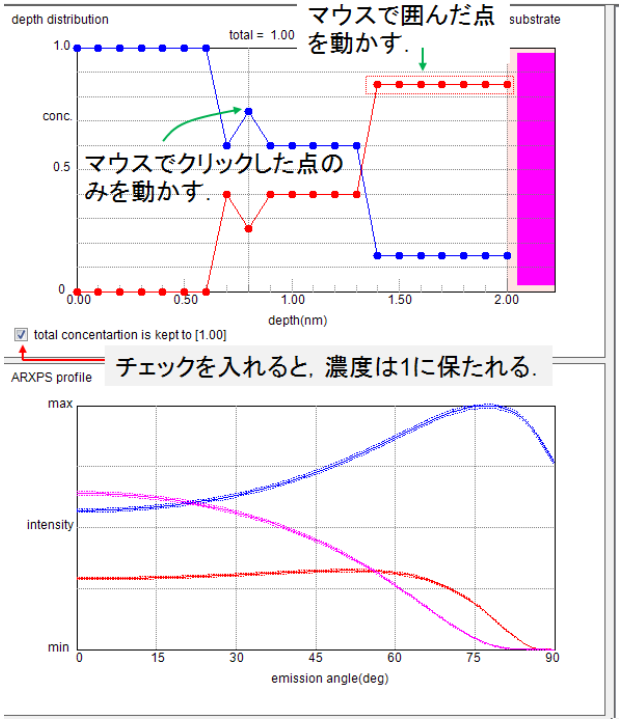

Fig. 133 Change of the distributions of transitions. (color online)

遷移の特定

[transition]のカラーボックス内に遷移名，結合エネルギー，相対感度を記入する．なお，[db]チェックボックスにチェックを入れると，遷移に関するデータベースがFig. 134のように表示される．元素名を選択し，遷移の一つをクリックすると必要情報がチェックした箇所にコピーされる．

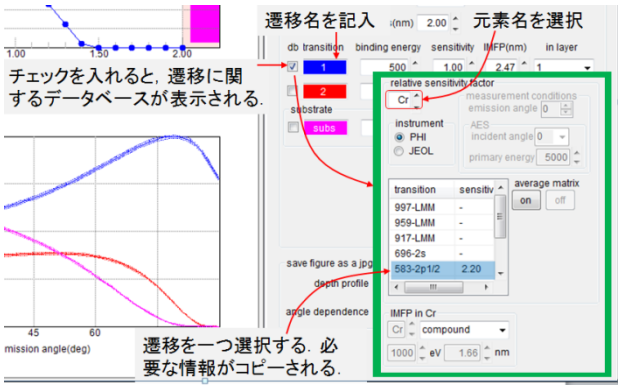

Fig. 134 Database for transition. (color online)

IMFP

IMFPは電子が通過する層の元素によって異なる．[in layer]コンボボックスで通過する層を選択すると，Fig. 135に示すように当該遷移のIMFPが表示される．IMFPはTPP-2M式[3]で計算される．なお，IMFPの値は，スクロールバーを用いて手動で変更出来る．

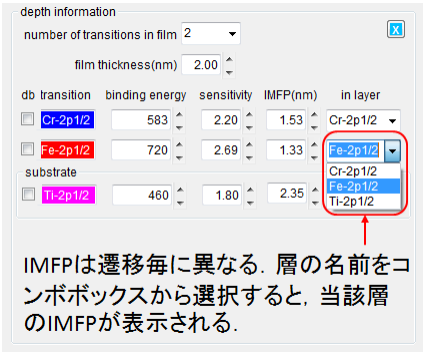

Fig. 135 Setting of IMFP. (color online)

ピーク強度の角度依存性の表示方法

[display option] タブを選択すると表示方法を変更できる．次のようなオプションがある．

[I_element]：遷移のピーク強度を表示

[I_element / sum(I_element)]：遷移のピーク強度の相対変化を表示

[normalized：[I_element]]：遷移のピーク強度を最大値・最小値で規格化して表示

[ratio : I_element / I_reference]：遷移同士のピーク強度比を表示．I_referenceはカラーテーブルから選択

[normalized ratio : [I_elm / I_ref]]：遷移同士のピーク

強度比を最大値・最小値で規格化して表示
Fig. 136に[ratio : I_element / I_reference]の例を示す.

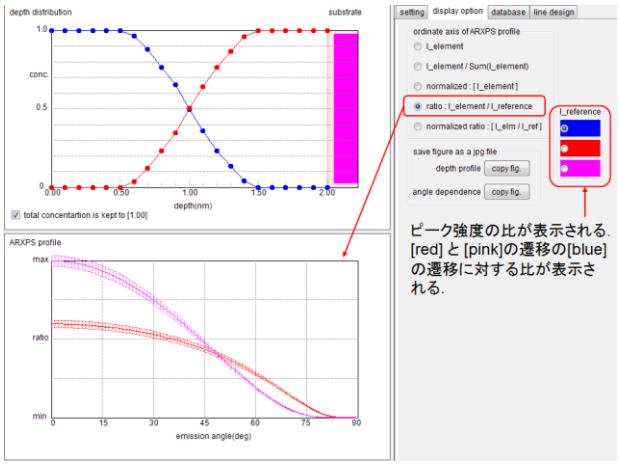

Fig. 136 Option for the display of ARXPS profile (color online)

36. 表面近傍のポテンシャルの曲がりによる光電子ピークの変形のシミュレーション

本シミュレーションはNIMS吉川英樹氏により作成された．半導体の界面近傍の相当の厚さの領域でキャリアの存在しない空乏層が作られるが，その空乏層中ではポテンシャルが緩やかに変化している．このポテンシャルの曲がりをシミュレーションにより求める．メニューバーの[Simulation] - [Band bending analysis]を選択すると，光電子スペクトルがFig. 137のように表示される．

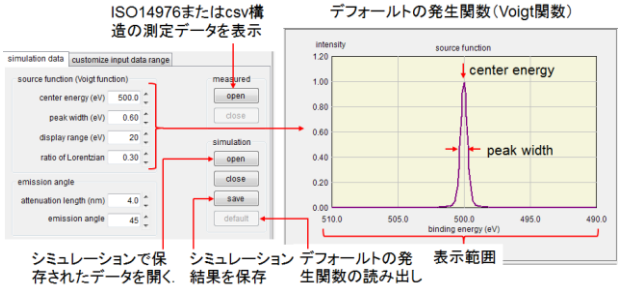

Fig. 137 Source Voigt function. (color online)

発生した（ポテンシャルによる変形を受けていない）光電子スペクトル形状はVoigt関数で表す．デフォールトの関数のパラメータ値は [source function (Voigt function)] グループボックス内のスクロールバーで変更できる．デフォールトでは光電子の減衰長さは4.0nm, 放出角度は45度に設定されている．これらの値は[emission angle]グループボックス内のスクロールバーで変更できる．

ポテンシャルの曲がりの設定

ポテンシャルの曲がりを表す関数は，linear（直線），quadratic（二次関数），exponential（指数関数）の組み合わせで設定される．関数の組み合わせ層数は最大５層である．なお，最表面のポテンシャルは0eV（基準）で，ポテンシャルエネルギーの符号が＋になると光電子の結合エネルギーは小さくなる．ポテンシャル関数のパラメータは [a]，[depth (nm)]，[potential(eV)] である．ここで[a] は quadratic と exponentialのパラメータであり，[depth (nm)]はポテンシャル関数が存在する最大深さ，[potential (eV)]はポテンシャル関数の最大ポテンシャル値である．
各関数形は次のように表せる．

linear: $d = gradient \cdot (p - p_1) + d_1$

$$[gradient = (d_2 - d_1)/(p_2 - p_1)]$$

quadratic: $d = \big(gradient + a(p - p_2)\big) \cdot (p - p_1) + d_1$

exponential; $d = \frac{(d_1 - d_2)(\exp(a \cdot p) - \exp(a \cdot p_1))}{(\exp(a \cdot p_1) - \exp(a \cdot p_2))} + d_1$

各関数形に含まれる変数の意味はFig. 138に示すので，参照されたい．quadraticとexponentialの関数の曲がり具合はパラメータ：aによって決定される．

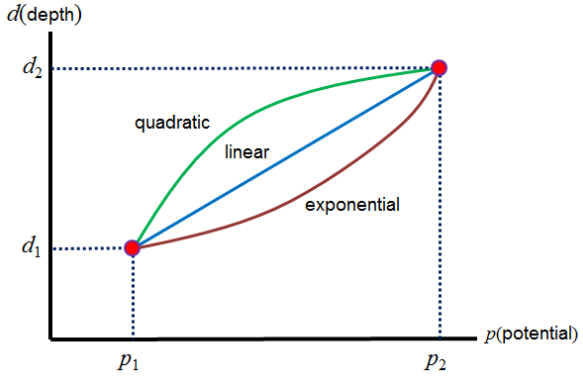

Fig. 138 Potential function. (color online)

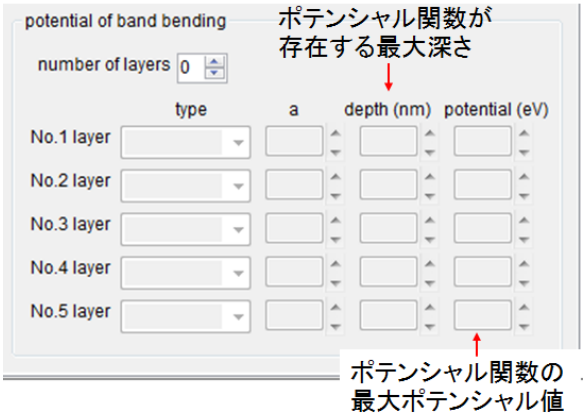

Fig. 139 Setting of potential function. (color online)

ポテンシャル関数の設定画面をFig. 139に示す．層数を[number of layers]コンボボックスから選択し．それぞれの層のポテンシャル関数を[type]コンボボックスから選択し，パラメータを入力する．
ポテンシャル関数の設定例として３個のポテンシャル層が存在するポテンシャル曲がりを示す．最表面から2.0nmまではlinear，2.0nmから3.0nmまではquadratic，3.0nmから5.0nmまではexponentialとする．最大のポテンシャル値は，それぞれ0.2ev，0.5ev，1.0eVとする．パラメータ：aの値は0.2 (quadratic)と0.3 (exponential)とする．これらの値を入力するとポテンシャル関数（Fig. 140）とバンド曲がりにより変形したスペクトル（Fig. 141）が表示される．

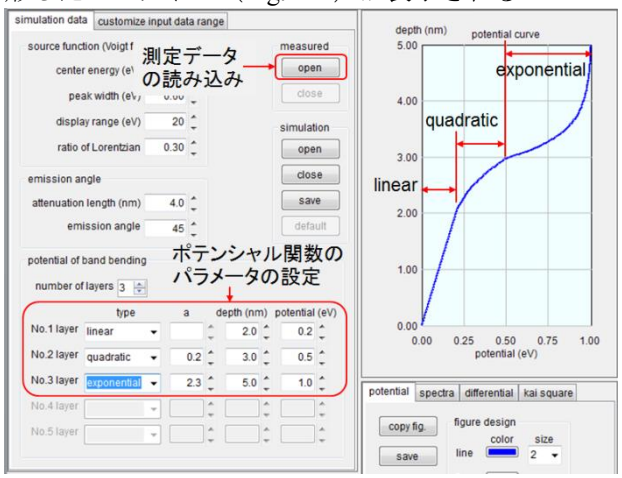

Fig. 140 Example of potential function. (color online)

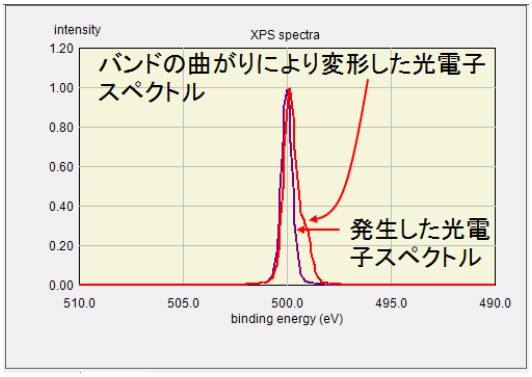

Fig. 141 Photoelectron peak affected by band bending. (color online)

観測値との比較

シミュレーションで得られたスペクトルは測定データと比較することが出来る．[open]ボタンをクリックすると，ISO14976構造かcsv形式で保存された測定データファイルを読み込むことが出来る．ただし，csv形式で保存されたファイルは，第１行目（multi blocksの場合には各ブロックの第１行目）に放出角度が記述されている必要がある．

測定データは発生関数とシミュレーション結果を表示した画面にFig. 142に示すように上書きされる．ピークの中央値と放出角度（もし記録されていれば）は自動的に測定データの値に一致するように表示される．測定スペクトルとシミュレーションスペクトルの差が示される．放出角度が異なる 複数の測定（multi blocks）がなされた場合には，放出角度ごとのカイ二乗値が表示される．

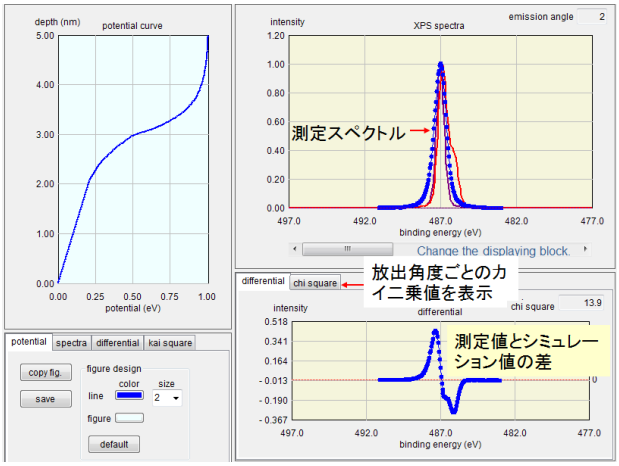

Fig. 142 Comparison with observed spectrum. (color online)

37. 主成分分析

多変量解析の一つである主成分分析（Principal Component Analysis: PCA）は主としてTOF-SIMSデータの解析に用いられる．

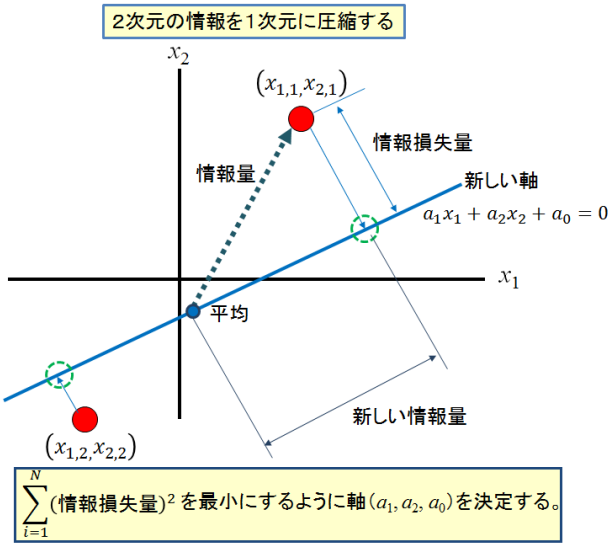

Fig. 143 Concept of PCA. (color online)

ある問題に対して，いくつかの要因が考えられるとき，それらの要因を一つ一つ独立に扱うのではなく，総合的に取り扱う解析手法が主成分分析である．主成分分析は多くの変量$(x_1, x_2, \cdots, x_p)$の値をできるだけ情報の損失なしに，$z = a_1x_1 + a_2x_2 + \cdots$

$+a_p x_p + a_0$のような一次式で結びつけて，互いに独立な新たな少数個の一次式（$z_1, z_2, \cdots$：主成分と称する）を求める解析手法である．すなわち，データの特徴を出来るだけ少ない変数（主成分）で代表することを目的とする．

例えばFig. 143に示すように，N個の二次元のデータが与えられたときに，$\sum_{i=1}^{N}(\text{情報損失量})^2$が最小となるような直線を引き，それを新しい軸（主成分）として，その直線上の値で二次元データを表す．これにより二次元データを一次元データとして扱うことが出来る．二次元データのように変数が少ない場合には情報損失量の総和を求め，その最小値を求めることは容易であるが，多次元データの場合には工夫が必要となる．TOF-SIMSデータのように変数が多い系で情報損失量が最小となる複数個の新しい軸（主成分）を見つけるためにはデータの“ばらつき”を対象とした分散共分散行列を作り，その行列の固有値の大きい順に，対応する固有ベクトルの軸を主成分（第１，第２，・・・）としていけば，情報量の損失が最も少なくなるように軸を選定することができる[1]．変数が多いデータを取り扱っても，固有値の大きな固有ベクトルは通常２ないし３個程度なので，２ないし３個の主成分でデータ群の特徴を表す事が出来る．これが主成分分析である．主成分分析の手順については「**付録1**」で解説する．

ポアソン分布に従う測定では，与えられたデータの不確かさはデータのカウント数に依存する．この不確かさを考慮せずにデータ解析を行うと，カウント数の大きなデータの不確かさが小さなカウント数のデータの情報を隠してしまい，正確な結果が得られないことがある．特にTOF-SIMSのように検出分子種ごとのカウント数が大きく異なるような測定では重要な問題となる．ポアソン分布の誤差を考慮したデータの補正方法がPoisson scaling[4]である．Poisson scalingについては「**付録2**」で解説する．COMPRO12ではPoisson scalingが可能である．

メニューバーの[Multivariate analysis] - [Principal component analysis]を選択すると，Fig. 144に示す二次元のデータ入力テーブルが表示される．分析データをこの入力テーブルに手入力するか，あるいは[open]ボタンをクリックしてcsvかexcelファイルを読み込む．データ入力後も行や列の削除または挿入，データ値の修正は可能である．入力後には入力テーブルの列の定義を行う．試料のデータが列に沿って入力されていれば[sample]ボタン，行に沿って入力されていれば[variable]ボタンを選択する．スケーリング補正はPoisson scalingとnormalizationが可能である．Poisson scalingはカウント数の小さなピークに対して大きなピークのカウント数の不確かさの影響を少なくすること，normalizationは試料ごとに最大ピーク強度を[1]，最小ピーク強度を[0]とし，ピーク強度変動の影響を少なくすること，に有効である．

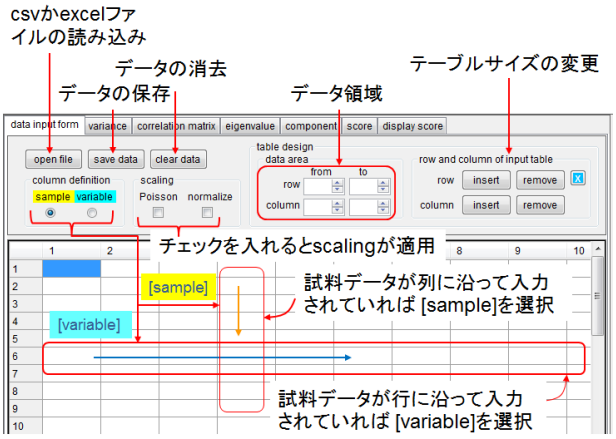

Fig. 144 Data input table. (color online)

例としてcsv形式で保存されたTOF-SIMSデータを読み込み，テーブルに入力するとFig. 145のように表示される．

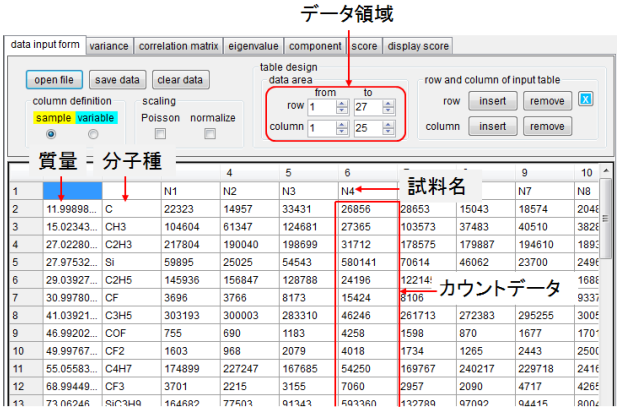

Fig. 145 Example of data table. (color online)

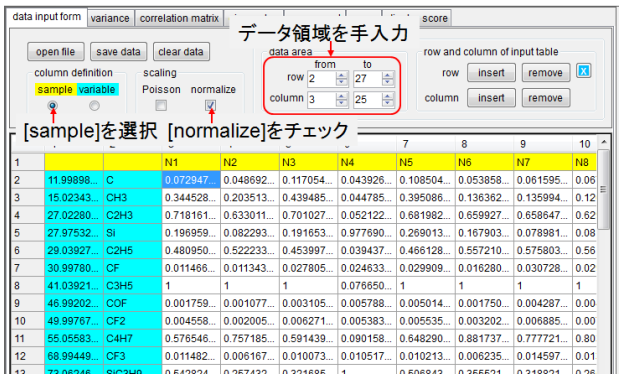

Fig. 146 Definition of data table. (color online)

データ領域（数値領域）を手入力で指定する．データ領域の大きさには制限は無い．データは第２行第３列から始まっているので，Fig. 146に示すように[data area]グループボックス内の[row]と[column]の[from]ボックスの番号をそれぞれ2と3に変更する．列の定義として[sample]ボタンを選択する．データは規格化するために[normalize]にチェックを入れる．

分析結果の表示

入力データを標準化（「**付録1**」参照）して分散共分散行列を作成して固有値解析を行う．固有値の大きな固有ベクトルを２個ないし３個選択する．目安は累積寄与率（cumulative proportion）が0.8以上となる個数である．寄与率（proportion）は当該固有値／固有値の総和，累積寄与率は最大の固有値の寄与率から当該固有値の寄与率までの和を示す．固有値の大きい順に固有ベクトルを第１主成分，第2主成分・・・とする．画面上部のタブをクリックすると，以下に示すようにタブごとに結果が表示される．

[variance] 測定値の平均値，分散，標準偏差の表示

[correlation matrix] 相関行列を表示. COMPROでは入力データを標準化して，分散共分散行列を相関行列にしている．

[eigenvalue] 固有値，寄与率，累積寄与率を表示．寄与率はグラフに表示される．

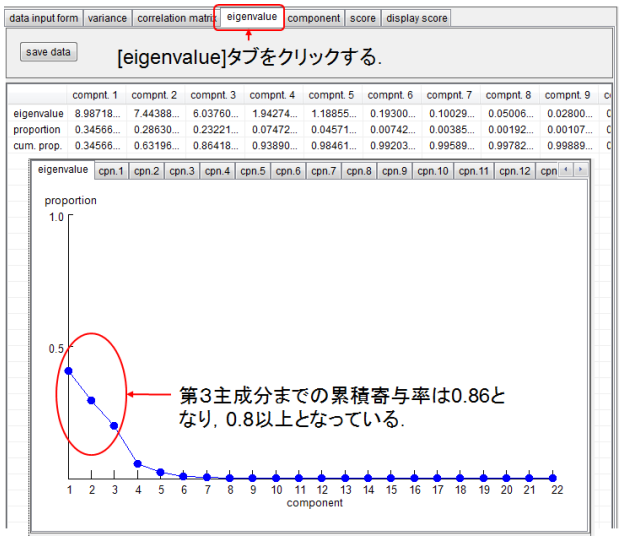

Fig. 147 Display eigenvalue (color online)

Fig. 147に示すように第３主成分までの累積寄与率は0.86となり，目安の0.8以上となっている．したがって，このTOF-SIMSデータの特徴は３個の主成分で表示することができると言える．

[component] 固有ベクトル（主成分）と負荷量（factor loading）が表示される．主成分zは$z = a_1x_1 + a_2x_2 + \cdots + a_px_p + a_0$と表す事ができる．TOF-SIMSの場合には$x_1, x_2$ …は分子種の（標準化された）カウント数であり，a_1, a_2 …はカウント数に掛ける負荷量である．Fig. 143に示すように原点a_0は平均値にとるが，標準化した場合には$a_0 = 0$となり，試料ごとに主成分の値（得点）: zが計算できる．第１主成分（component 1）の得点は，Fig. 148に示される負荷量を使うと

$$z = 0.1910[\mathrm{C}] - 0.1302[\mathrm{CH_3}] - 0.2489[\mathrm{C_2H_3}] + \cdots$$

と表せるので，各試料のCのカウント数，CH_3のカウント数，C_2H_3のカウント数，・・に負荷量を掛ければそれぞれの試料の第１主成分得点が求まる．同様に第２，第３主成分の値も求まる．

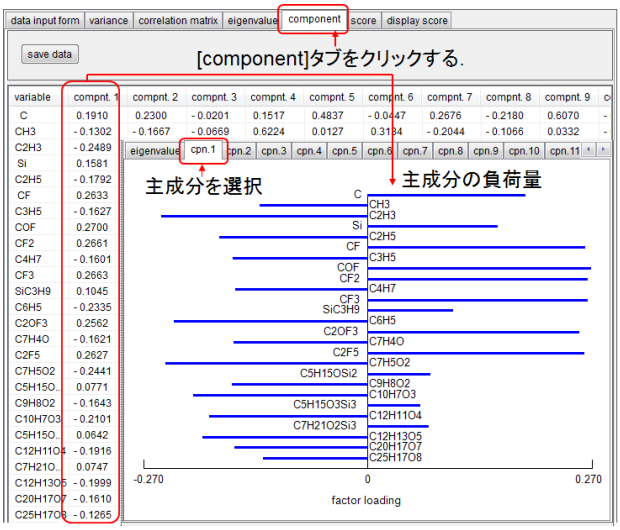

Fig. 148 Display component. (color online)

[score] それぞれの試料データの主成分の得点が表に出力される．例えば試料１の第１主成分の得点が-0.8806，第2主成分の値は-2.9439・・・というように表示される．

[display score] それぞれの試料の得点をFig. 149のように二次元表示する．

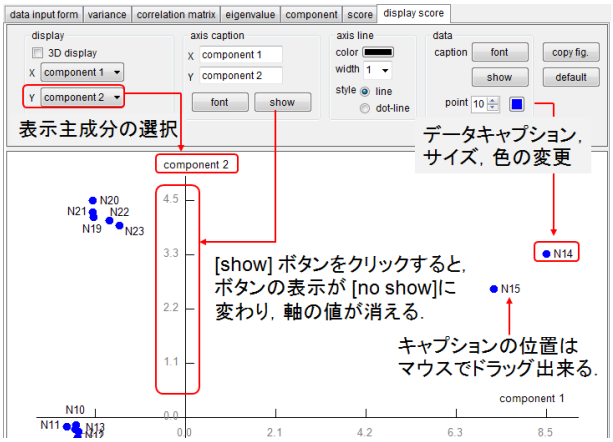

Fig. 149 Display score (2 dimensional). (color online)

デフォールトの軸は第１主成分と第２主成分である．第１主成分は固有値最大の固有ベクトルで，第２主成分の固有値はその次の大きさである．表示する主成分はコンボボックスで変更できる．軸のキャプションは[axis caption]グループボックスで変更できる．試料名の位置はマウスでドラッグすると変更できる．データのキャプションのサイズや色は[data]グループボックスで変更できる．

Fig. 149からデータのグルーピングが出来ることが分かるが，このデータの場合には３個の主成分が大きな固有値を持つことから，より正確には第３主成分までを含めた３次元表示をする必要がある．[3D display]チェックボックスにチェックを入れると，Fig. 150に示すように試料の得点を３次元表示する．３次元の軸の傾きはスクロールバーで変更できる．薄い青い色で表示された点は*x*-*y*面よりも下方にある事を示している．

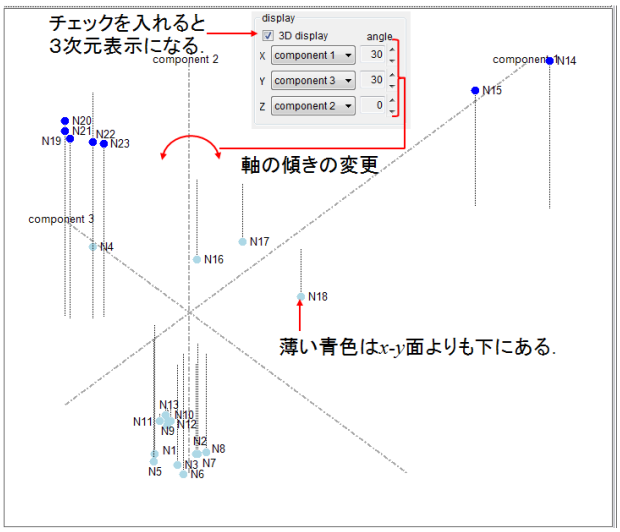

Fig. 150 Display score (3 dimensional). (color online)

38. クラスター分析

多変量解析の一種であるクラスター分析は，与えられたデータ群の中から似たもの同士を集めてクラスターに分類する方法である．クラスター分析に関する解説はJSA誌[5]に掲載されているので参照されたい．

データが似ているか否かはデータ間のユークリッド距離の大小から判定する．データがp個の変量を持つとき，データ（例えばi番目のデータとj番目のデータ）間の相違をデータ間のユークリッド距離：

$$\sqrt{\sum_{k=1}^{p}\left(x_{k,i}-x_{k,j}\right)^2}$$

で代表させる．この距離が小さいもの同士の組み合わせをグループ（クラスター）とする．次にグループ間の距離（グループ間の距離の求め方はWard法と群平均法（group average method））を求めて，同じようにクラスターを形成するという作業をおこなう．Ward法はグループに属しているデータの距離の標準偏差を用い，群平均法はグループに属しているデータの距離の平均を用いる．この作業を進めていき，結果を樹形図（デンドログラム：dendrogram）で表す．

メニューバーの[Multivariate analysis] - [Cluster analysis]を選択すると，Fig. 151に示す二次元のデータ入力テーブルが表示される．分析データをこの入力テーブルに手入力するか，あるいは[open]ボタンをクリックしてcsvかexcelファイルを読み込む．データ入力後も行や列の削除または挿入，データ値の修正は可能である．入力後には入力テーブルの列の定義を行う．試料のデータが列に沿って入力されていれば[sample]ボタン，行に沿って入力されていれば[variable]ボタンを選択する．スケーリング補正はPoisson scaling，normalizationとstandardization（標準化）が可能である．データのstandardizationは「**付録1**」に述べるように単位が異なるデータを比較するときに有効な処理である．グループ間の距離の計算に対してはWard法か群平均法を選択する．

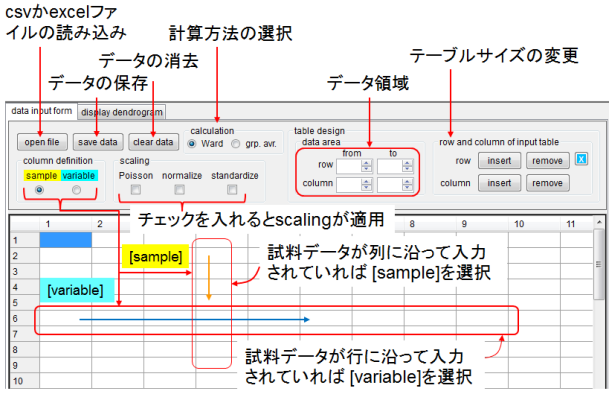

Fig. 151 Data input table. (color online)

Fig. 152 Example of data table. (color online)

例として[37. 主成分分析]に用いたものと同じデータを読み込みFig. 152のように表示させる．

データ領域（数値領域）は手入力で指定する．データ領域の大きさには制限は無い．データは第２行第３列から始まっているので，Fig. 153に示すように[data area]グループボックス内の[row]と[column] の[from]ボックスの番号をそれぞれ2と3に変更する．列の定義として[sample]ボタンを選択する．データは規格化するために[normalize]にチェックを入れる．計算方法は[Ward]を選択する．

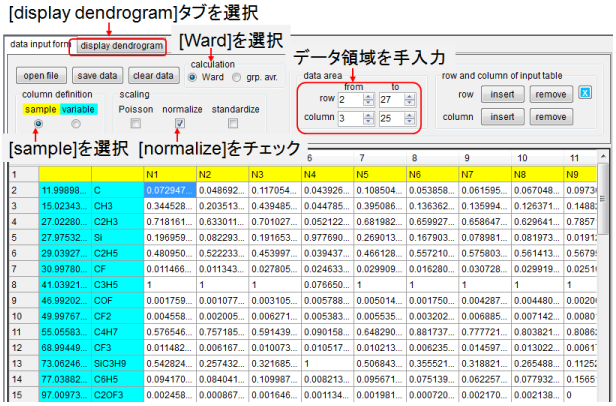

Fig. 153 Definition of data table. (color online)

[display dendrogram]タブを選択すると，クラスター分析結果がFig. 154のように樹形図として表示される．縦軸は試料名，横軸はユークリッド距離である．[log scale]チェックボックスにチェックを入れると横軸がlogスケールになり，小さな距離が強調される．

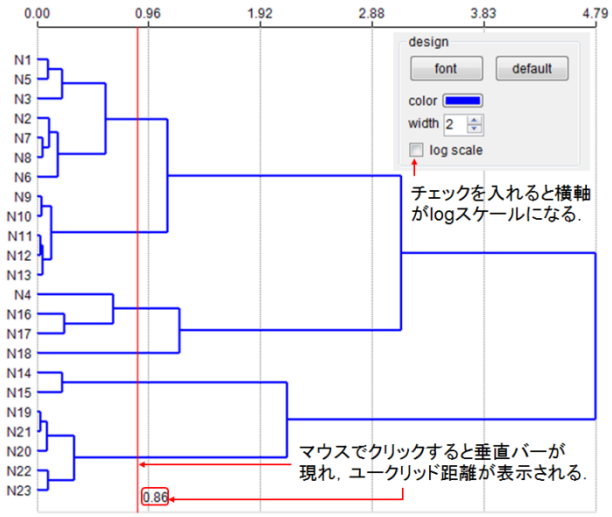

Fig. 154 Display dendrogram. (color online)

ユークリッド距離が近い試料同士がグルーピングされる．Fig. 154の例では，N1, N5, N3が一つのグループを作り，それがさらにN2, N7, N8, N6のグループと一つになることが分かる．Fig. 154中の垂直線で示されるユークリッド距離では23個の試料が，[N1, N5, N3, N2, N7, N8, N6], [N9, N10, N11, N12, N13], [N4, N16, N17], [N18], [N14, N15], [N19, N21, N20, N22, N23]の６グループに分類される事が示される．

クラスター分析は似たもの同士をデータから探しだしてグルーピングをすることが目的の統計手法である．ただし，クラスター分析の場合は“なぜ似ているのか”という判断基準は示されない．一方，主成分分析はデータ群の特徴を出来るだけ少ない数の主成分で表す手法で，グルーピングを目的とする手法ではない．グラフ表示結果を見てデータのグルーピングは出来るが，表示軸として選択する主成分の数によってグルーピングに差異が生じることがある．なお，データが似ているか否かの判断基準は主成分の負荷率によって示される．

参考文献

[1] 吉原一紘，*J. Surf. Anal.* **24**, 175 (2018).
[2] C. J. Powell, *Surf. Sci.* **44**, 29 (1974).
[3] S. Tanuma, C. J. Powell and D. R. Penn, *Surf. Interface Anal.* **21**, 165 (1994).
[4] M. R. Keenan and P. G. Kotula, *Surf. Interface Anal.* **36**, 203 (2004).
[5] 吉原一紘，徳高平蔵，*J. Surf. Anal.* **21**, 10 (2014).

付録1　主成分分析の手順

Table 1のような健康診断結果を使用して，主成分分析の手順を解説する．各検査項目の数値は単位が異なり，このまま分析しては，大きな値の検査項目の影響が大きくなりすぎて正確な分析結果が得られないので，データの標準化を行う．

Table 1 Medical checkup data (color online)

	1	2	3	4	5	6
1		肝指数	血糖値	尿酸値	コレステロール	血圧
2	A	21	98	5.2	128	135
3	B	18	140	8.6	135	180
4	C	38	120	5.1	150	130
5	D	55	85	4.3	82	128
6	E	19	138	9.8	150	190
7	F	48	92	5.8	116	150
8	G	45	110	5.2	90	120
9	H	48	128	4.3	98	100
10	I	32	108	5.4	95	130
	平均値	36.000	113.222	5.967	116.000	140.333
	標準偏差	14.089	19.722	1.919	26.014	28.653

データの標準化（x_{std}）は次式に示すように，データ（x）と平均値（x_{avr}）の差を標準偏差（σ）

で除することで行われる．

$$x_{\mathrm{std}} = {(x - x_{\mathrm{avr}})}/{\sigma}$$

主成分分析ではデータの“ばらつき”だけを考慮すれば良いので，データの値を平均値との差の度合いに変換することにより，単位が異なるデータ間の比較が可能になる．Table 1のデータを標準化するとTable 2のように変換される．

Table 2 Standardized medical checkup data. (color online)

	1	2	3	4	5	6
1		肝指数	血糖値	尿酸値	コレステロール	血圧
2	A	−1.065	−0.772	−0.400	0.461	−0.186
3	B	−1.278	1.358	1.372	0.730	1.384
4	C	0.142	0.344	−0.452	1.307	−0.361
5	D	1.349	−1.431	−0.869	−1.307	−0.430
6	E	−1.207	1.256	1.998	1.307	1.733
7	F	0.852	−1.076	−0.087	0.000	0.337
8	G	0.639	−0.163	−0.400	−0.999	−0.710
9	H	0.852	0.749	−0.869	−0.692	−1.408
10	I	−0.284	−0.265	−0.295	−0.807	−0.361
	平均値	0.000	0.000	0.000	0.000	0.000
	標準偏差	1.000	1.000	1.000	1.000	1.000

データを標準化すると，全ての検査項目の平均値は0となり，標準偏差は1になる．COMPROにおける主成分分析では標準化されたデータを用いて分散共分散行列を作成し，固有値解析を行う．変数xと変数yの共分散σ_{xy}は，変数xと変数yの平均値をそれぞれx_{ave}，y_{ave}とすると，以下のようになる．

$$\sigma_{xy} = \sqrt{\sum_{n=1}^{N}(x - x_{\mathrm{ave}})(y - y_{\mathrm{ave}})}/(N-1)$$

データ点数はN個である．標準化したデータの共分散（σ_{xy}）は元のデータの相関係数（$\sigma_{xy}/(\sigma_x \cdot \sigma_y)$）と同じ値になる．なぜならば，標準化したデータの標準偏差（σ_x, σ_y）は1となるからである．すなわち，分散共分散行列は，データを標準化することにより相関行列に置き換えることができる．相関行列の要素はTable 3のように記述される．

Table 3 Elements of correlation matrix. (color online)

肝指数の分散　　肝指数と尿酸値の共分散

肝指数	肝指数･血糖値	肝指数･尿酸値	肝指数･コレステロール	肝指数･血圧
肝指数･血糖値	血糖値	血糖値･尿酸値	血糖値･コレステロール	血糖値･血圧
肝指数･尿酸値	血糖値･尿酸値	尿酸値	尿酸値･コレステロール	尿酸値･血圧
肝指数･コレステロール	血糖値･コレステロール	尿酸値･コレステロール	コレステロール	コレステロール･血圧
肝指数･血圧	血糖値･血圧	尿酸値･血圧	コレステロール･血圧	血圧

対角要素は各検査項目値の分散で，非対角要素は検査項目値同士の共分散である．例えば肝指数（x_i）と尿酸値（y_i）の共分散は，標準化した値を用いればそれぞれの平均値は0であるから

$$\sum_{i=1}^{9} x_i y_i/(9-1) = -0.7448$$

となる．同様にして共分散を求めるとTable 4のような相関行列（分散共分散行列）が作成出来る．この相関行列（分散共分散行列）の固有値解析を行い，固有値の大きな固有ベクトルを算出する．固有値の大きな固有ベクトルは分散の大きい（すなわち情報損失量が少ない）主成分軸となる．COMPROでは固有値解析はJacobi法を用いている．

Table 4 Correlation matrix.

1.0000	−0.5772	−0.7448	−0.6971	−0.6945
−0.5772	1.0000	0.6694	0.5570	0.4358
−0.7448	0.6694	1.0000	0.6515	0.9446
−0.6971	0.5570	0.6515	1.0000	0.6570
−0.6945	0.4358	0.9446	0.6570	1.0000

選択する固有ベクトルの本数の目安は，固有値の寄与率（当該固有値／固有値の総和）を最大の固有値から順に足していったときに，寄与率の合計（累積寄与率）が0.8以上となる本数である．固有値解析の結果，Table 5に示すように第２主成分までの累積寄与率が0.8539となり，目安の0.8を越えているため，この健康診断結果は２個の主成分により，説明出来ることが分かる．

Table 5 Eigenvalue

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分
固有値	3.6713	0.5981	0.4306	0.2901	0.0099
寄与率	0.7343	0.1196	0.0861	0.0580	0.0020
累積寄与率	0.7343	0.8539	0.9400	0.9880	1.0000

固有値解析から各主成分の負荷量（factor loading）はTable 6のように求まる．

Table 6 Factor loading

	第1主成分	第2主成分	第3主成分	第4主成分	第5主成分
肝指数	−0.4542	0.0066	−0.3001	0.8383	0.0296
血糖値	0.3873	0.8180	−0.3263	0.0956	−0.2555
尿酸値	0.4927	−0.1887	−0.4313	0.0884	0.7265
コレステロール	0.4331	0.0655	0.7424	0.4962	0.1037
血圧	0.4620	−0.5394	−0.2575	0.1846	−0.6287

Table 6の結果は．第１主成分は尿酸値の高い検体には高得点が与えられ，肝指数が高い検体には負の大きな得点が与えられることを，第２主成分は血糖値が高い検体には高得点が与えられ，血圧が高い検体には負の大きな得点が与えられることを示している．例えば検体[A]の第１主成分の得点は，検体[A]の各測定項目の負荷量（Table 6）を各測定項目の数値（Table 2）にかけて，以下のように求める．

$(-0.4542 \times -1.065) + (0.3873 \times -0.772) + (0.4927 \times -0.400) + (0.4331 \times 0.461) + (0.4620 \times -0.186) = 0.1016$

同様に第２主成分の得点も-0.4325と求まる．すなわち，検体[A]は第１主成分と第２主成分を直交軸とした平面上の（0.1016, -0.4325）点に表示される．同じ計算を他の検体について行い，第１主成分，第２主成分上の得点を求め，それらをFig. 155に表示して主成分分析は終了する．

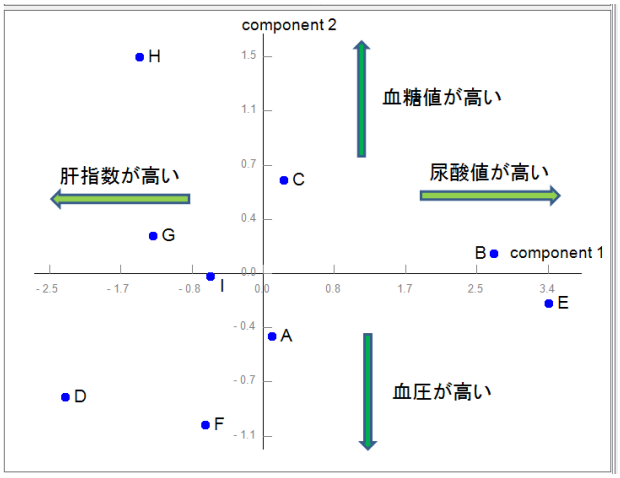

Fig. 155 Display score. (color online)

付録2　Poisson scalingの紹介

電子分光やイオン分光のように，いつ起きるかはランダムであるが，起きる頻度の平均値は一定であるような事象を計測すればポアソン分布で記述できる．ポアソン分布に従う観測値（ν）の分散はν，標準偏差は$\sqrt{\nu}$である．ポアソン分布に従うカウント数の不確かさ（分散）を考慮したカウント数の補正方法がPoisson scaling[4]である．ここではPoisson scalingの概略を紹介する．

n個のデータ点数のスペクトルデータがm本あるスペクトル群を下記のような行列（[D]）の形で記述する．

$$\underbrace{\begin{bmatrix} d_{11} & \cdots & d_{1j} & \cdots & d_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{i1} & \cdots & d_{ij} & \cdots & d_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{m1} & \cdots & d_{mj} & \cdots & d_{mn} \end{bmatrix}}_{n\text{列（データ点数）}} \left.\begin{matrix} \\ \\ \\ \\ \\ \end{matrix}\right\} m\text{行（スペクトル数）} \rightarrow \text{スペクトル}$$

データ行列[D]のカウント数d_{ij}の分散（variance）はポアソン布ならばvar(d_{ij}) = d_{ij}と記述できる．分散は行ファクターと列ファクターに分解できると仮定するとvar(d_{ij}) = $g_i h_j$と表せる．ここで，g_i;行iに共通の値，h_j;列jに共通の値である．

行列の形では，var([D]) = [g][h]$^{\mathrm{T}}$となる．ここで

$$[g]=\begin{bmatrix} g_1 \\ \vdots \\ g_i \\ \vdots \\ g_m \end{bmatrix} \quad [h]=\begin{bmatrix} h_1 \\ \vdots \\ h_j \\ \vdots \\ h_n \end{bmatrix}$$

ポアソン分布を考慮して標準化したデータ行列を$[\tilde{D}]$とすると（標準化に関しては「付録1」を参照）

$$[\tilde{D}] = (a[G])^{-1/2}[D](b[H])^{-1/2}$$

となる．ここで[G]と[H]はそれぞれg_iとh_jを対角成分とする行列（対角行列）で，aとbは任意の定数である．行列[G]と[H]の求め方を以下に記述する．

ポアソン分布では，分散はカウント数と同一である．そこで，g_iとh_jをそれぞれi番目の行とj番目の列のカウント数の出現確率と関連づける．

ここで，E：分散の期待値，　p_i：i番目の行のカウント数の出現確率，q_j：j番目の列のカウント数の出現確率，$d_{..}$：データ行列[D]のカウント数の総和（$d_{..} = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}$）とすると，

$$E[\mathrm{var}(d_{ij})] = E(d_{ij}) = p_i q_j d_{..}$$

と書ける．ここで，

$d_{i.} = \sum_{j=1}^{n} d_{ij}$：データ行列のi行のカウント数の総和

$d_{.j} = \sum_{i=1}^{m} d_{ij}$：データ行列のj列のカウント数の総和

とすると，$p_i = d_{i.}/d_{..}$　$q_j = d_{.j}/d_{..}$ となる．

次のような行列を定義する．

$$[d_m]=\begin{bmatrix} d_{1.} \\ \vdots \\ d_{i.} \\ \vdots \\ d_{m.} \end{bmatrix} \quad [d_n]=\begin{bmatrix} d_{.1} \\ \vdots \\ d_{.j} \\ \vdots \\ d_{.n} \end{bmatrix}$$

この行列を用いると，分散の期待値は以下のように行列の形で記述できる．

$$E[\mathrm{var}([D])] = \frac{1}{d_{..}}[d_m][d_n]^{\mathrm{T}}$$

この式をvar([D]) = [g][h]$^{\mathrm{T}}$と比較すると，

とすれば，一致する．

$$[g]=\frac{[d_m]}{\sqrt{d_{..}}} \quad [h]=\frac{[d_n]}{\sqrt{d_{..}}}$$

$[\tilde{D}] = (a[G])^{-1/2}[D](b[H])^{-1/2}$ で行列[D]を変換する時に，a, bは任意の定数であるから，$a = \sqrt{d_{..}}/n$，$b = \sqrt{d_{..}}/m$とおけば

$$a[g] = \frac{\sqrt{d_{..}}}{n}\cdot\frac{[d_m]}{\sqrt{d_{..}}} = \frac{[d_m]}{n} \quad b[h] = \frac{\sqrt{d_{..}}}{m}\cdot\frac{[d_n]}{\sqrt{d_{..}}} = \frac{[d_n]}{m}$$

と表せる．したがって，対角行列[G]，[H]は

$$(a[G])^{-1/2}=\begin{bmatrix}\left(\frac{d_{1\cdot}}{n}\right)^{-1/2} & 0 & 0 & 0 & 0\\ 0 & \ddots & 0 & 0 & 0\\ 0 & 0 & \left(\frac{d_{i\cdot}}{n}\right)^{-1/2} & 0 & 0\\ 0 & 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & 0 & \left(\frac{d_{m\cdot}}{n}\right)^{-1/2}\end{bmatrix}$$

$$(b[H])^{-1/2}=\begin{bmatrix}\left(\frac{d_{\cdot 1}}{m}\right)^{-1/2} & 0 & 0 & 0 & 0\\ 0 & \ddots & 0 & 0 & 0\\ 0 & 0 & \left(\frac{d_{\cdot j}}{m}\right)^{-1/2} & 0 & 0\\ 0 & 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & 0 & \left(\frac{d_{\cdot n}}{m}\right)^{-1/2}\end{bmatrix}$$

この行列を用いて，データ行列[D]をポアソン分布を考慮して補正したデータ列[$\tilde{D}$]に変換する．

（注）ポアソン分布に従うカウント数の分散は var(d_{ij}) = d_{ij}であり，個々のデータ点毎に独立した値をとる．一方，文献[4]で導入されたPoisson scalingでは，個々のカウント数の分散をカウント数の出現確率に結びつけている．文献[4]では，分散と出現確率を結びつける理由を「分散の見積もりを統計的に集計することにより，個々の分散の見積もりが改善される」としている．詳細は文献[4]を参照してほしい．

査読者との質疑応答

査読者1. 飯田真一（アルバック・ファイ）

本解説記事はCOMPRO12に実装されている，「角度分解データシミュレーション」，「表面近傍のポテンシャルの曲がりによる光電子ピークの変形のシミュレーション」,「主成分分析」,「クラスター分析」の4つの機能について丁寧な解説がなされており，JSA誌に掲載の価値があると考えます．掲載にあたり，以下の点をご検討頂きますようお願い致します．

[査読者1-1]表面近傍のポテンシャルの曲がりによる光電子ピークの変形のシミュレーションについて

このシミュレーションコードを用いる目的を冒頭に追加してはいかがでしょうか？（用途が良く分からなかったので）

[著者]

丁寧に査読いただきありがとうございました．ご指摘の点に基づき，以下のように修正させていただきます．

半導体の界面近傍の空乏層のポテンシャルの曲がりをシミュレーションにより求めることができるという内容の文章を[36]項の最初に付け加えました．

[査読者1-2]主成分分析とクラスター分析について

今回，主成分分析とクラスター分析に同じTOF-SIMSのデータを使って説明されていますが，それぞれの特徴や違いなどについて言及されてはいかがでしょうか？

例えば，主成分分析では第3主成分まででデータの特徴を説明できるとありますが，これはクラスター分析ではどのように対応しているのかなど，読者が，それぞれの手法から導き出された結果をリンクして考えることができるようなコメントを追加頂ければと思います．

[著者]

主成分分析はデータ群の特徴を出来るだけ少ない数の主成分で表す手法で，グループ分けを目的とする手法ではありません．一方，クラスター分析は似たもの同士をデータから探しだしてグループ分けをすることが目的の統計手法です．主成分分析もグラフ表示することによりグループ分けが出来ますが，

累積寄与率の選択（２次元表示にするか３次元まで表示させるか）によってグループの分け方に差異が生じます．また，データが似ているか否かの判断基準は主成分の負荷率によって明示されますが．クラスター分析の場合は“なぜ似ているのか”という判断基準は示されません．この内容の文章を[38]項の最後に付け加えました．

査読者2. 松村純宏（HGSTジャパン）

COMPRO12の使用法の続き(3)ということで, 非常に分かりやすく丁寧に説明されていると思います．更に，使用法だけではなく，行っている計算の詳細に関しても(1), (2)と同様に分かりやすい説明がされていて，JSA誌に掲載されると，COMPRO12を使われている方はもちろん，それ以外でも表面分析に関わられている方とって，参考になるところが非常に多いと思います．是非とも，掲載をお願します．

[査読者2-1]

「Fig. 131」に関して，一番上の層中に「1」と書かれています．これは「1」番目の層ということを示しているのだと思いますが，その下の層では何番目の層かということと一緒にIMFPも表記されているので，同様の表現になっているほうが良いように感じました．

[著者]

丁寧に査読いただきありがとうございました．ご指摘の点に基づき，以下のように修正させていただきます．

Fig. 131の第１層の表示方法を [1, IMFPn, 1] のように変更いたしました．

[査読者2-2]

主成分分析に関して教えてください．COMPROではデータは全て「付録1」の初めに説明されている「データの標準化」を行ってから，主成分分析に必要な分散共分散行列の計算やその後の固有値解析が行われているということです．そうしますと，本文p. 15の「原点をa_0=0となるところを選ぶと」とありますが,「データの標準化」により，自動的に

「a_0=0」

となるように思うのですが, 正しいでしょうか？ 本文p. 13とp. 15の「z」の式に「a_0」が入っていることと「データの標準化」の関連についてお教えいただけるとありがたいです．

[著者]

主成分の得点の計算の原点は平均値に設置しますので，ご指摘の通り，標準化している場合には$a_0=0$となります．Fig. 143の「平均」の表示位置がずれていましたので修正し，p. 15に$a_0=0$となる説明を付け加えました．