スペクトル解析のための統計的機械学習

志賀元紀*,**,***

1. はじめに

物質・材料科学において新たに合成された試料の構造や特性を評価するために，電子顕微鏡などの様々な計測機器が用いられる．例えば，試料の微細構造を評価するために，走査透過型電子顕微鏡の計測のHAADF−STEM（High−Angle Annular Dark−Field−Scanning Transmission Electron Microscopy）法を用いれば，原子番号1.4〜2乗に比例する強度の実空間濃淡画像が観測される．この計測法では，2次元の実空間画像が得られるため，研究者が目で見て判断するには非常に優れている．しかしながら，試料を透過した電子線をある一定の領域で積分して得たものであるために，その他の情報が損失している．これに対して，電子線の試料透過によるエネルギー損失の分光計測であるEELS（Electron Energy−Loss Spectrum）法を2次元の試料平面上で計測するSTEM−EELS法は，平面上の各点で多数の帯域からなるスペクトルを直接計測できるため，より多くの情報を得られる．しかしながら，そのデータ量のため研究者が直接チェックすることが難しかったり，雑音との明瞭な区別がつかないなどの問題がある．この問題を解消するためには，データ解析の自動化や定量評価が必要不可欠であり，昨今，研究のみならず産業界で注目が集まっている人工知能・統計的機械学習[1][2]によるアプローチが期待される．ところで，機械学習と人工知能の明確な言葉の区別が不明瞭な記事が多いように思われる．本稿では，データから規則発見やタスク遂行のためにシステムを学習する手法を（統計的）機械学習といい，一方，学習後にタスク遂行のために利用するシステムを人工知能ということにする．

人工知能は，膨大な数のデータを利用でき，かつ，ゴールが明確なタスクに対して優れた性能を発揮することがいくつかの実応用で実証されている．それでは，研究者が何も考えることなく，人工知能のみを用いて，科学研究や製品開発をできるものだろうか？　これに対する前向きな回答は現状難しい．主な理由は2つある．第1の理由は，研究対象は新規・未知の事象であり，利用可能なデータが限られることである．第2の理由は，目標が必ずしも明確でなく，目標や問題設定自体の創発が研究活動に求められることである．このような理由で，人工知能に全てを任せることはできないものの，研究者自身が優れたセンスをもって，背後にあるモデルおよび仮定を構築し，その検証を達成するための論理構築およびデータ収集すれば，有益なシステムとして利用できる．これは素晴らしい研究を行うために必要なことと同じである．つまり，昨今データ解析技術が発達しているものの，研究者自身の理解がその性能を最大限に発揮するために非常に重要なポイントでもある．本稿で扱うスペクトル解析[3]−[6]の問題においても，解析対象の背後にある物理・化学的な理論とデータ解析で仮定されるモデルを理解することが，適切なデータ解析を行うために非常に重要である．

本解説では，電子顕微鏡等の分光計測スペクトルを対象としたデータ解析法をいくつか紹介する．一言にスペクトル解析といっても，膨大な数のスペクトルを網羅的に計測できる状況もあれば，限られた実験条件における少数のスペクトルしか得られない状況もある．また，理論的または経験的にスペクトル形状に関するモデルを知らない場合があれば，それらのモデルを知っており解析に利用できる場合もある．このように，計測条件および解析対象によって様々な状況が存在するものの，全ての状況に対して最適なデータ解析・機械学習法は存在しない．そこで，研究者自身が状況に応じて適切な手法を選択することが，スペクトル解析によって適切な結論を導くために必要不可欠である．もし，この選択を誤ると，せっかく取得したデータが台無しになる恐れがある．本稿では，スペクトル解析アプローチを状況に基づき分類し，代表的な解析手法とその背後にあるモデルを述べる．

* 岐阜大学工学部；准教授（〒501−1193　岐阜市柳戸1−1）
** 科学技術振興機構さきがけ研究者
*** 理化学研究所革新知能統合研究センター；客員研究員
Statistical Machine Learning for Spectral Data Analysis; Motoki Shiga*,**,***(*Faculty of Engineering, Gifu University, Gifu. **Japan Science and Technology Agency, Saitama. ***Center for Advanced Intelligence Project, RIKEN, Tokyo)
Keywords: *statistical machine learning, multivariate analysis, unsupervised learning, spectrum imaging, scanning transmission electron microscopy with electron energy loss spectroscopy*
2018年9月3日受理[doi:10.2320/materia.58.23]

2. 計測スペクトルのピーク関数フィッティング

本節では，スペクトルのピーク形状が理論的あるいは経験的に知られている場合の解析法を述べる．この状況は，例えば，ラマン分光計測スペクトルのピーク形状が単峰性の関数（ローレンツ関数やガウス関数）で近似できる場合である．ピークの個数が複数である場合，複数のピーク関数の線形和によって，計測スペクトルがモデル化される（**図 1**）[7][8]．スペクトル帯域 t での計測スペクトル $y(t)$ のモデル $\hat{y}(t, \boldsymbol{\theta})$ は，ピーク数 K, k 番目のピークの中心位置 μ_k と幅 s_k の単峰性関数 $f_k(t, \mu_k, s_k)$ およびその高さ w_k としたとき，

$$\hat{y}(t, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k f_k(t, \mu_k, s_k) \qquad (1)$$

と記述される．ただし，$\boldsymbol{\theta}$ は全パラメータ（全てのピークの中心，幅，高さ）をまとめたベクトルである．ピーク個数が既知の場合，K を固定して，モデル $\hat{y}(t, \boldsymbol{\theta})$ の計測スペクトル $y(t)$ へのフィッティング誤差（平均二乗誤差）

$$J(\boldsymbol{\theta}) = \sum_{t=1}^{T} \{\hat{y}(t, \boldsymbol{\theta}) - y(t)\}^2 \qquad (2)$$

の最小化によってパラメータ $\boldsymbol{\theta}$ を学習する．ただし，T はスペクトル帯域数である．一般に学習問題の解析的な解を得られないので，勾配法などの逐次最適化アルゴリズムによって学習される．様々な実験条件・時刻で計測されたスペクトルが複数あれば，各スペクトルに対してパラメータ $\boldsymbol{\theta}$ を学習して比較することで，物理量の変動や推移を調べることができる．

ところで，多数の実験条件でスペクトル計測する状況では，ピーク数 K が条件によって変動することがあり，その場合には計測データからピーク数 K を推定する必要がある．しかしながら，ピーク数 K の推定のために平均二乗誤差 $J(\boldsymbol{\theta})$ を最小化するアプローチを使えない．なぜなら，ピーク数 K を大きくすると雑音成分にモデルが過剰フィッティングし，平均二乗誤差をいくらでも小さくできるからである．この問題に対して，ベイズ学習が有効であることが知られる[7]．

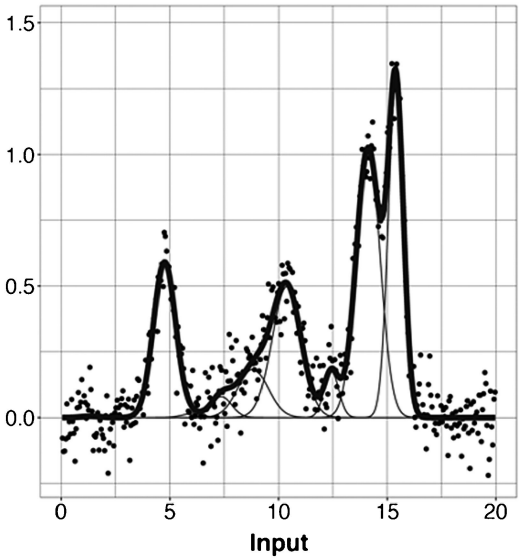


図 1　スペクトルのピーク関数フィッティング（点：観測スペクトル，細線：学習後の単峰性関数，太線：学習後のモデル）．

スペクトル解析において，ピーク形状以外にも，既知の観測過程をモデル化できる場合がある．例えば，注目点だけでなく，実空間の隣接位置の構造特性が計測スペクトルに混合される過程のモデルである（**図 2**）．この状況では，計測スペクトルをピーク関数で単純にフィッティングするのではなく，混合前の純粋な特性スペクトルを推定する逆問題を解くことで，より正確な解析結果を導ける．具体的に，著者らは，試料の深さ方向に焦点を変えながら計測する 3D ラマン分光スペクトル計測に対する解析法を開発した[9]．問題の単純化のために，単峰性ピークのスペクトル（$K=1$）を仮定する．深さ方向の焦点距離を離散的に D 個に区切れば，深さ d での観測スペクトル $y_d(t)$ は以下のように

$$\hat{y}_d(t, \boldsymbol{\theta}) = \sum_{i=1}^{D} c_{i,d}\, w_i f_i(t, \mu_i, s_i) \qquad (3)$$

とモデル化される．ただし，係数 $c_{i,d}, d=1, ..., D, i=1, ..., D$ は，深さ i でのスペクトルが深さ d でのスペクトルに混合する大きさである．また，ピーク位置 μ_i および幅，s_i の関数 $f_i(t, \mu_i, s_i)$ は深さ i での純粋な成分スペクトルである．ここで，混合係数が理論的あるいは実験的に決定されているものとする．この場合，2 乗誤差

$$J(\boldsymbol{\theta}) = \sum_{d=1}^{D} \sum_{t=1}^{T} \{\hat{y}_d(t, \boldsymbol{\theta}) - y_d(t)\}^2 \qquad (4)$$

を最小にするようにパラメータ $\boldsymbol{\theta}$ を学習することによって，直接計測できない層ごとのスペクトルを推定できる．

この節を締めくくるにあたり，ピーク形状が既知の状況に関数モデルを用いる利点を述べる．1 つ目の利点は，無数にあるスペクトル形状から当てはめるべき形状をあらかじめ絞り込めることである．この絞り込みによって，フィッティング問題が格段に易しいものとなり，少数あるいは 1 つのスペクトルしか観測できない状況でもデータ解析が可能となる．ただし，誤った関数モデルを用いると誤った結論を導く

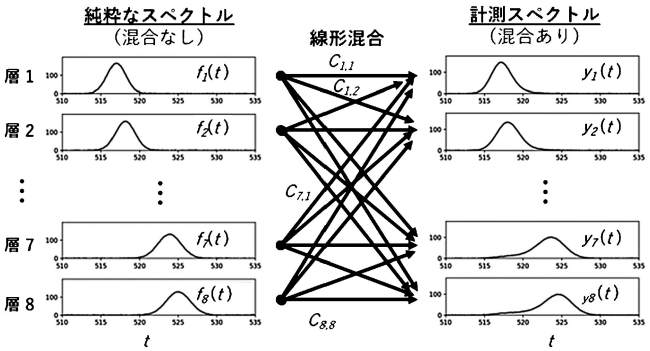


図 2　3D ラマン分光計測における線形混合モデル．

ので，フィッティング結果をプロットして注意深く確認することを忘れずに行うべきである．2つ目の利点は，ピーク関数のパラメータが形状と対応するため，結果を解釈しやすいことである．通常のピーク関数は，ピークの中心位置，幅，高さの3つのパラメータで特徴付けられる．フィッティングによってこれらのパラメータが定まるため，実験条件ごとに学習されたパラメータを比較すれば，解析で知りたい情報を得られる．

3. スペクトル解析のための教師なし学習法

前節と異なりピーク形状に関する知識がない，あるいは，単純な関数モデルで記述できない場合の解析法を述べる．この場合，無数のスペクトル波形の候補からデータに最も適合するスペクトル波形を推定する必要があるため，前節のフィッティング問題より格段に難しい問題を扱うことになる．したがって，少数のスペクトルしか計測できない状況では適合するスペクトル波形を絞り込むことは困難であるが，膨大な数のスペクトルを計測できれば，関数モデルがない場合でもフィッティング問題を解くことは可能である．幸いにも，走査透過型電子顕微鏡などの機器を用いれば，注目領域をグリッド状に細分化した各地点(x, y)でスペクトルを計測できて膨大な数のスペクトルを得られる．この計測データはスペクトラムイメージとよばれる．例えば，STEM-EELSやSTEM-EDXの標準的なデータのサイズは，スペクトル帯域数2,000に対して，試料平面グリッド上の点数$100\times100=10{,}000$という非常に大きな本数のスペクトルである．こうした状況では，関数モデルがなかったとしてもスペクトル解析が可能であり，有限個の成分(化学成分，電子状態)のスペクトルおよび成分濃度の実空間分布を推定できる．本節で紹介する手法は，化学成分に関する情報を使用せず，計測データのみを用いて解析するため，機械学習における教師なし学習のアプローチに分類される．

ところで，スペクトラムイメージ計測データは，実空間の計測地点に番号をつけて$n=1,\cdots,N$として計測スペクトルを並び替えると，サイズ$N\times T$のデータ行列$\boldsymbol{X}$に全データを格納できる．つまり，行列$\boldsymbol{X}$の各行は1地点での計測スペクトルを表す．以降では，スペクトラムイメージがデータ行列$\boldsymbol{X}$で与えられているものとする．

(1) クラスタ解析

クラスタ解析とは，記述子で特徴づけられた観測標本をクラスタ(グループ)に分ける解析である(**図3**)．計測スペクトルの記述子を各スペクトル帯域での強度とすれば，各スペクトルは，帯域数Tの次元の記述子空間中の1点に対応付けられる．この記述子空間上で近いデータ点の集まりがクラスタとなるため，同じ帯域でピークのあるスペクトルは同じクラスタに含まれる．そこで，各クラスタはピーク位置が類似する物理的に意味のある成分の集合とみなせる．クラスタ内の中心点を計算すれば，クラスタの代表スペクトルが得られ，対応する成分のスペクトルとみなせる．

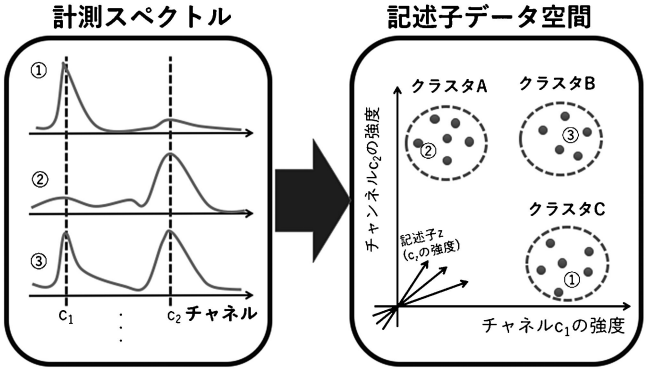


図3　記述子データ空間の構成およびクラスタ解析の概念図．

クラスタ解析の代表的な手法はk-平均法(*k-means*)である．k-平均法のアルゴリズムは，各スペクトルのクラスタ割当と代表スペクトルの計算を収束するまで交互に実行するものである．クラスタの代表スペクトルは，通常，クラスタに属するスペクトルの平均値によって計算される．また，各スペクトルは最も類似する代表スペクトルに対応するクラスタに割当てられる．k-平均法は，各スペクトルを1つのクラスタに割り当てるため，ハードクラスタリングと呼ばれる．

(2) 成分分析

前述のクラスタ解析では，各計測スペクトルに1つのみの成分が含まれることが仮定される．これに対して，成分分析では，複数の成分が含まれる状況を仮定しており，計測スペクトルは成分スペクトルと成分濃度の線形結合でモデル化される．つまり，計測地点n，スペクトル帯域tでの計測スペクトルは，k番目の成分濃度C_{nk}と成分スペクトルS_{tk}を用いて，

$$\hat{X}_{nt}=\sum_{k=1}^{K} C_{nk}S_{tk} \qquad (5)$$

と記述される．ただし，Kは成分数である．一見すると，式(1)と類似したモデル式であるが，スペクトル波形が特定の関数となっておらず，帯域毎に自由に値を設定できるのが本モデルの特徴である．このモデルは関数形が定まっていないため，ノンパラメトリックモデルとよばれる．

サイズ$N\times K$の成分濃度行列$\boldsymbol{C}$，サイズ$T\times K$の成分スペクトル行列$\boldsymbol{S}$とすれば，式(5)は，

$$\hat{\boldsymbol{X}}=\boldsymbol{C}\boldsymbol{S}^{T} \qquad (6)$$

と行列で記述される．一般的に$K\ll\min(T,N)$が成り立つ．この解析は，データ行列を要素数の少ない2つの行列に分解することに相当するため，行列分解とも呼ばれる．計測データが与えられたときに行うべきことは，各成分のスペクトル$\boldsymbol{S}$と各計測点での各成分の濃度$\boldsymbol{C}$を推定することであり，この推定は再構成誤差を最小にするように行われる．$\boldsymbol{S}$と$\boldsymbol{C}$に何も制約がない場合，解の自由度が高すぎて，解が定まらなかったり，物理的に不自然な解析結果を導く恐れが

あったりするため，通常，何らかの制約が課せられる．課せられる制約はもちろん解析結果に影響を与えるため適切に選ぶ必要がある．本節では，スペクトル解析に使用される成分分析の代表的な手法を紹介する．

(a) 主成分分析

主成分分析(Principal Component Analysis)は，多変量解析(複数の観測変数を統計的に扱う分野)において最も有名な手法の一つであり，高次元データを低次元空間に射影しデータ次元を削減する手法として知られる．主成分分析は，データ分布を多次元正規分布でモデル化してフィッティングするものであり，分散が大きい方向ベクトルで張られる低次元部分空間を重要な成分からなる空間と推定し，ここにデータが射影される．一方，残りの分散の小さいマイナーな部分空間は雑音成分からなる空間と推定され除外される．射影される低次元空間の座標軸は互いに直交するように選ばれるので，この直交制約が主成分分析の制約である．主成分分析をするためには，まず，全スペクトルの平均ベクトル$\boldsymbol{\mu}$が原点と一致するようにデータ行列の各行(各観測スペクトル)を$\bar{\boldsymbol{X}}_n = \boldsymbol{X}_n - \boldsymbol{\mu}$と平行移動する．そして，直交制約の下で平均2乗誤差

$$J(\boldsymbol{C}, \boldsymbol{S}) = \sum_{n=1}^{N} \sum_{t=1}^{T} (\boldsymbol{X}_{nt} - [CS^T]_{nt})^2 \qquad (7)$$

を最小する行列$\boldsymbol{C}$と$\boldsymbol{S}$を行列$\bar{\boldsymbol{X}}$の特異値分解$\bar{\boldsymbol{X}} = \boldsymbol{UDV^T}$によって求める．特異値分解後の行列を$K$列目まで取り出した$(\boldsymbol{U})_K, (\boldsymbol{V})_K$によって，成分濃度行列$\boldsymbol{C} = (\boldsymbol{U})_K$および成分スペクトル行列$\boldsymbol{S} = (\boldsymbol{VD})_K + \boldsymbol{M}$が得られる．ただし，行列$\boldsymbol{M}$は平均ベクトル$\boldsymbol{\mu}$を$K$列並べたサイズ$T \times K$の行列である．特異値分解の制約より行列$\boldsymbol{U}$と$\boldsymbol{V}$は正規直交である．つまり，内積が単位行列になる($\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$および$\boldsymbol{VV}^T = \boldsymbol{I}$が成立する)ため，主成分分析で得られる行列$\boldsymbol{C}$も直交行列となることに注意されたい．上記の別解法として，共分散行列$\Sigma = \boldsymbol{U}^T\boldsymbol{U})N$の固有値問題を解くアプローチがある．その詳細については紙面の都合上省略するが，興味のある方は文献(1)の12章を参考にしていただきたい．

主成分分析は，局所解に陥ることなく2乗誤差を最小にする行列$\boldsymbol{C}$と$\boldsymbol{S}$を一意に求められるメリットがある．しかしながら，行列$\boldsymbol{C}$と$\boldsymbol{S}$の値に範囲の制約がないため，その要素が負の値となる物理的に不自然な解析結果を導く可能性がある．また，分解後の行列に直交制約を課すため，同定すべきデータが直交しない場合に誤った結果を導く．EELSスペクトルの解析では，ある電子状態に起因するスペクトルは，特定帯域のピークのみでなく高エネルギー側に強度が継続するポストエッジを含むため，スペクトルが成分間で直交せず，正しい分解結果を得られない可能性が高い．こうしたデメリットがあるため，主成分分析のみを用いて成分濃度と成分スペクトルを同定する解析は難しい．そこで，解析対象に由来するスペクトル成分と雑音成分を分離し，雑音成分を除去する処理に主成分分析を使用するのが適切と考えられる．

(b) 頂点成分分析

主成分分析を用いれば，高次元の記述子空間の中に分布するデータの低次元部分空間を同定できる．しかしながら同定された低次元空間の各軸は，前述したように同定したい成分と対応するとは限らない．この問題を解決するために，低次元空間に計測データを射影した後に純粋な成分スペクトルを推定する頂点成分分析(Vertex Component Analysis)が提案されている(10)．

頂点成分分析は，i番目の空間位置での計測スペクトル$\boldsymbol{X}_{i\cdot}$と純粋な成分スペクトル$\boldsymbol{S}_{\cdot k}, k = 1, \cdots, K$との間に以下の関係式

$$\boldsymbol{X}_{i\cdot} = \sum_{k=1}^{K} C_{ik} S_{\cdot k}^T = \gamma \sum_{k=1}^{K} A_{ik} S_{\cdot k}^T \qquad (8)$$

が成り立つことを利用する．ただし，A_{ik}はゼロ以上かつ$\Sigma_{k=1}^{k} A_{ik} = 1$が成り立つ成分濃度比を表し，また，$\gamma$はスケール因子である．式(8)より，計測スペクトル$\boldsymbol{X}_{i\cdot}$はデータ空間において純粋な成分スペクトルから構成される凸錐(convex cone)に含まれている．スケール因子γを取り除くために，平均ベクトルとの内積が1となるようにデータを正規化すれば，純粋な成分スペクトルを頂点とする単体(simplex)に各計測スペクトル$\boldsymbol{X}_{i\cdot}$が含まれることになる．頂点成分分析は，この単体の頂点を探索することで純粋な成分スペクトルを同定する．具体的な頂点探索アルゴリズムや雑音が多い計測データへの対策などの工夫の詳細は文献(10)を参考にしていただきたい．ところで，頂点成分分析は，純粋な成分スペクトルが観測データに含まれることが正確な成分同定のために必要である．したがって，原子分解能のスペクトル計測のように，成分が空間的に大きく重なる場合に正確な解析が保証されないことに注意されたい．この問題解決のために，部分空間サンプリングに基づく手法が提案されている(11)．

(c) 非負値行列分解

前述したとおり，主成分分析で同定される成分濃度行列$\boldsymbol{C}$と成分スペクトル行列Sは負の値になりうる．物理的な意味合いを考えれば，これらの行列の要素値はゼロ以上であり，負の値となるべきでない．非負値行列分解は，あらかじめ行列$\boldsymbol{C}$と$\boldsymbol{S}$の要素値を非負に制約して，再構成誤差を最小化するように行列分解する手法である．データ行列$\boldsymbol{X}$にも非負値性を仮定するが，バックグラウンド成分除去などの前処理に起因して負値が現れる場合が発生する．その場合には，通常，行列$\boldsymbol{X}$の負の値をゼロに置き換えて解析される．非負値行列分解における最適化は，主成分分析の場合と異なり解析解を得られないため，再構成誤差を小さくするように逐次最適化アルゴリズムが用いられる．その最適化アルゴリズムは行列$\boldsymbol{C}$と$\boldsymbol{S}$を交互に更新するのが一般的であり，代表的なものとして，行列積による更新則によって行列単位に更新する手法(12)，二乗誤差関数の勾配をゼロとする線形方程式を解いた後に負値をゼロとおき更新するAlternating Least Squares (ALS)法(13)，列ベクトル毎に行列を更新するHierarchical ALS (HALS)法(14)などが提案されている．

基本的な非負値行列分解は単純な制約しか課さないので，

実用的にいくつか問題がある．著者らはSTEM–EELSデータを解析する場合に実用的に重要な2つの拡張を施した手法を提案している[15]．1つ目の拡張は，非負値行列分解で得られる行列$\boldsymbol{C}$と$\boldsymbol{S}$がスパースになる傾向があり，不自然な分解結果になる問題を回避するためのものである．EELSスペクトルの解析において，成分濃度行列$\boldsymbol{C}$は成分間で重なりが少なくスパースになる傾向があるものの，本来のEELSスペクトルはスパースでないため，行列分解によって得られるスペクトルの形が物理的に不自然になる傾向がある．この問題を解決するため，我々の提案した非負値行列分解(SO–NMF)[15]では，成分濃度行列$\boldsymbol{C}$が直交(スパースかつ成分重なりが少ない)になるようにソフト直交制約の罰則項を加えた目的関数を新たに導入し，この最適化によって行列分解を行う．この結果として得られる成分濃度行列$\boldsymbol{C}$がよりスパースになり，一方で，スペクトル行列$\boldsymbol{S}$が非スパースに最適化されることによって自然な分解結果を得られる．

2つ目の拡張は，成分数Kの自動選択である．成分数は間違った値に設定されると誤った結論を導くために，計測データから自動的に決定できることが望ましい．我々の論文[15]では，Automatic Relevance Determination (ARD)事前分布[16]に基づく罰則項を導入し，成分数を計測データから学習するアプローチをとった．これら2つの拡張を施した行列分解法(ARD–SO–NMF)は，通常の平均2乗誤差に加えて，ソフト直交制約の罰則項およびARDに関する罰則項を加えた目的関数の最小化によって，成分濃度行列$\boldsymbol{C}$と成分スペクトル行列$\boldsymbol{S}$を最適化する．この最適化を具体的に実行するために，HALS法に基づくアルゴリズムを新たに導出した．本開発法の実行ソフトウェアは著者のGitHubレポジトリ[17]において公開されているので，興味のある方は参考にしていただきたい．

(3) STEM–EELSデータの解析

実際のSTEM–EELSスペクトラムイメージを解析した．解析したデータは半導体メモリ素子の断面試料に対し，1ステップ5 nm，各点から0.1 eV/チャンネルでEELSを取得した後に，試料バックグラウンド成分を除去し，シリコンの$L_{2,3}$吸収端ELNESを取り出したものである．**図4**は，この試料の計測領域における成分の参照分布と参照スペクトルである．このデータには3つの成分が含まれており，それらの空間分布は成分ごとに明確に分かれている．したがって，解析結果として得られる成分濃度行列$\boldsymbol{C}$が直交行列になることが期待される．一方，参照スペクトルは成分毎に特徴的なピーク形状を示すが，多くの帯域で非ゼロ値となりスパースでも直交でもない．

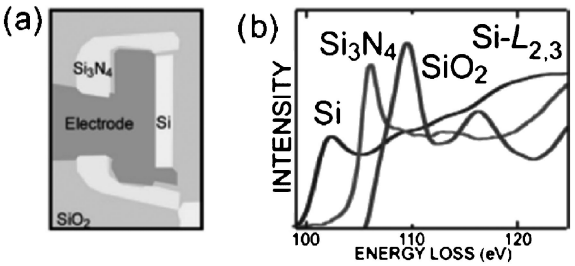


図4　半導体メモリ素子STEM–EELSデータにおける(a)成分の参照分布と(b)参照スペクトル．

図5―図8に主成分分析，頂点成分分析，非負値行列分解，ソフト直交制約付き非負値行列分解(SO–NMF)による行列分解の結果をそれぞれ示した．主成分分析の結果である**図5**(a)から空間的に成分が分離されていないことが分かる．また，得られた成分スペクトルは参照スペクトルと異なり不自然に落ち込む形状をしている．**図6**に示された頂点成分分析の結果は，得られた成分スペクトルが自然な形状をしており，成分1(Si_3N_4)，成分2(Si)，成分3(SiO_2)の全てが主成分分析の結果より正確に同定できていることが分かる．計測されたスペクトラムイメージ中に純粋な成分スペクトルが存在しており，これらを正確に同定できたといえる．また，**図7**より，非負値行列分解の結果は，主成分分析の結果

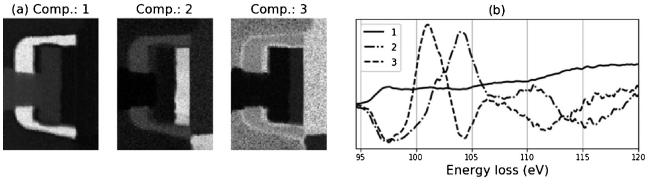


図5　主成分分析による分解結果(a)成分濃度分布と(b)成分スペクトル．

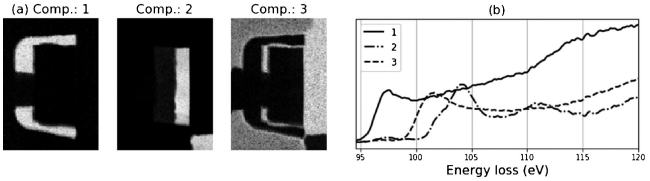


図6　頂点成分分析による分解結果(a)成分濃度分布と(b)成分スペクトル．

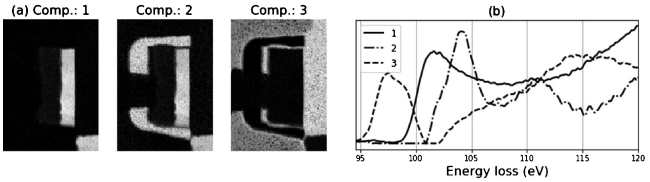


図7　非負値行列分解による分解結果(a)成分濃度分布と(b)成分スペクトル．

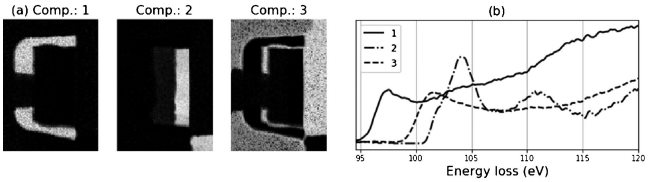


図8　ソフト直交制約付き非負値行列分解(SO–NMF)による分解結果(a)成分濃度分布と(b)成分スペクトル．

よりも正確なスペクトルを同定できているものの，不自然な落ち込みが残っており，また，成分が空間的に分離されていないことが分かる．一方，**図 8** に示される我々の開発法(SO-NMF)は，成分の空間分離性能が他手法より優れており，また，得られた成分スペクトルが参照スペクトルと類似していることが分かる．したがって，従来の非負値行列分解より改善が見られる．

5. ま と め

本稿では，スペクトル解析における統計的機械学習法を紹介した．最後にしばしば質問される事項に関して，私なりの考えを述べたい．まず，1 つ目によくある質問は「どの手法を選択すればよいか？」である．これ対する答えは，データの取得状況・性質・仮定とよく適合した手法を選べば良いといえる．本稿で紹介した全ての手法は教師なし学習である．そこで，明確な目標数値を改善する教師あり学習と異なり，数値目標による判断が難しい．また，本来データには含まれていなかったとしても，解析手法の出力が偶然にも研究者が見たかったものと一致することが起こりうる．こうした誤りを引き起こさないためにも，解析手法の性質・モデルをよく理解し，また，解析結果を異なる計測データと照らし合わせて判断することが必要となる．これは，本稿が伝えたかったことの一つでもある．

また，「正確にデータ解析するために必要なスペクトル本数はいくつか？」という問いをよく投げかけられる．残念ながら，この問いへの答えは一般的に存在しないと考えている．スペクトルのピーク関数フィッティングは，モデルの制約が比較的厳しいために，スペクトル本数が少なくても良い．一方，成分分析は，関数モデルを仮定しないので比較的多くのスペクトル本数を必要とする．また，雑音量や成分数，比較対象の成分の変化度合いなどによって，正確な結論を導くためのサンプル数が異なり，一概にどれくらいの数があれば良いと言い切ることが難しい．

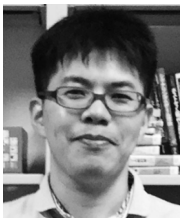
最後によく聞かれる質問は「前処理は必要か？」である．これに対しては，スペクトル解析に適切な前処理が重要と考えている．例えば，STEM-EELS 解析では，バックグラウンド成分を正確に除去しなければ，本稿で述べた解析手法がうまく機能しない．観測スペクトルには，通常，化学成分と強く関係する信号だけでなく，興味のない成分も含まれ，場合によっては興味のない成分の方が振幅が大きい．この観測スペクトルをそのまま解析すると振幅の大きな成分に解析手法が引っ張られるので，正しい結果を得られないとしても不思議でない．興味のない成分の規則性をある程度分かっているようであれば，その規則を用いて興味のない成分を除去し，興味のある成分のみを取り出せば，解析精度が格段に改善される．しかしながら，前処理の副作用として，興味のある成分の一部が除去される場合もあるので，この点を注意深く調べる必要がある．

最後に，データ解析はデータさえあれば必ず適切な結果を導けるとは限らず，適切なモデル・手法・前処理を選択することが必要不可欠であることを強調したい．

名古屋大学の武藤俊介教授には，初稿に対して有益なコメントをいただいたことに深く感謝申し上げます．本研究は，JST さきがけ JPMJPR16N6，新学術領域研究「ナノ構造情報のフロンティア開拓」16H00736，基盤研究(B)16H02866 の支援を受けて行われた．

文　　献

(1) C. M. ビショップ：パターン認識と機械学習，丸善出版，(2012).
(2) T. Hastie, R. Tibshirani and J. Friedman：統計的学習の基礎―データマイニング・推論・予測―，共立出版，(2012).
(3) M. Shiga and S. Muto: Nanoinformatics., Springer, Singapore (2018), 179–203.
(4) H. F. Grahn and P. Geladi: Hyperspectral Image Analysis, Wiley, (2007).
(5) M. Bosman, M. Watanabe, D. T. L. Alexander and V. J. Keast: Ultramicroscopy, **106**(2006), 1024–1032.
(6) N. Dobigeon and N. Brun: Ultramicroscopy, **120**(2012), 25–34.
(7) K. Nagata, S. Sugita and M. Okada: Neural Networks, **28** (2012), 82–89.
(8) H. Sudou, M. Shiga, T. Omodaka, C. Nakai, K. Ueda and H. Takaba: J. Korean Astronomical Society, **50**(2017), 157–165.
(9) H. Wang, H. Zhang , B. Da, M. Shiga, H. Kitazawa and D. Fujita: J. Physi, Chem. C, **122**(2018), 7187–7193.
(10) J. M. Nascimento and J. M. Dias: IEEE Trans. Geoscience and Remote Sensing, **43**(2005), 898–910.
(11) J. Spiegelberg, S. Muto, M. Ohtsuka, K. Pelckmans and J. Rusz: Ultramicroscopy, 182, (2017), 205–211.
(12) D. D. Lee and H. S. Seung: Adv. Neural Inform. Process. Syst., **13**(2001), 556–562.
(13) M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca and R. J. Plemmons: Comput. Stat. Data Anal., **52**(2007), 155–173.
(14) A. Cichocki: IEICE Trans. Fund. Electron. Commun. Comput., **92**(2009), 708–721.
(15) M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, T. Mori and T. Tanji: Ultramicroscopy, **170**(2016), 43–59.
(16) V.Y.F. Tan and C. Fevotte: IEEE Trans. Pat. Anal. Mach. Intell., **35**(2003) 1592–1605.
(17) https://github.com/MotokiShiga



志賀元紀

★★★★★★★★★★★★★★★★★★★★★★
2006年 3 月　岐阜大学大学院工学研究科博士後期課程修了
主な略歴　2006年 4 月京都大学化学研究所・博士研究員，2008年 4 月同研究所・助教
2017年 4 月–現職．2016年10月–JST さきがけ研究者を兼務
専門分野：機械学習，マテリアルズインフォマティクス
◎分光スペクトルなどの微細構造計測データ解析におけるデータ解析法の開発，統計的機械学習を用いた材料探索の研究を中心に活動している．
★★★★★★★★★★★★★★★★★★★★★★