統計数理 (2001)
第 49 巻 第 1 号 133–153

特集「地図を描く・風景を眺める
── 主成分分析・多次元尺度法とその周辺 ──」
［研究詳解］

非計量多次元尺度構成法への期待と新しい視点

中央大学* 田口善弘
慶應義塾大学** ／イリノイ大学*** 大野克嗣
農業技術研究機構 北海道農業研究センター**** 横山和成

（受付 2000 年 10 月 5 日；改訂 2001 年 5 月 1 日）

要　　旨

遺伝情報や脳神経情報の解析など大量のデータを縮約し特徴を抽出する必要性はますます高まりつつある．われわれはこのような大量情報の圧縮のための手法として，主に社会科学分野で使われている多変量解析の一種である非計量的な多次元尺度構成法に注目している．非計量的な多次元尺度構成法は，ある空間（例えばユークリッド空間）に，対象間の「差違」の大小関係を保存する条件の下で，対象を埋め込む手法であり，特徴抽出にできるだけ先入観を入れない方法と考えることができる．われわれは順序統計量を用いた非計量多次元尺度構成法の新しいアルゴリズムを提案するとともに，得られた布置がもとの距離関係をどのくらい満たしているか判定する方法を提案する．従来の非計量多次元尺度構成法の手法は大量のデータを処理することを眼目とはしてこなかったように見受けられるが，この手法は大規模データの処理にも十分対応可能である．この手法を具体的に土壌微生物の炭素源利用能の表，脳の電位の時系列データ，そして，シダやコケ／緑藻の分類（表現形質の表と塩基配列データ）に適用し，いずれも 2 次元程度の低次元に情報を圧縮できることを見出した．他の多変量解析の手法で同じことを試みた結果とも比較した．

キーワード：非計量多次元尺度構成法，多量データ，情報圧縮，当てはめの良さ．

1. はじめに

ゲノム解析や金融データの解析をまつまでもなく，大量のデータからその特徴を解析したり，情報を縮約したりする必要は増加する一方である．物理学であれば頭を使って何がデータのなかの本質的な量かを考えるのが本筋であるかもしれないが，無限のバリエーションを持つ大規模データの各々について頭を使って本質的な量を見出す，というのは事実上不可能であり，この不可能性こそいわゆる「複雑系」の特徴の一つかも知れないのである．あるいはこのような系のモデルを作るとき，注目する諸量が再現出来たとしても，それが「現象」を再現しているかどうかはよくわからない．注目している量にとらわれず，現象そのものをモデルで再現できているかどうかを確認したい，というのは本質的な要求であろう．そのためには，ややこしい

*理工学部物理学科／理工学研究所：〒112–8551 東京都文京区春日 1–13–27.
**理工学部数理科学科：〒223–8522 神奈川県横浜市港北区日吉 3–14–1.
***物理学科：Loomis Laboratory of Physics, 1110 W Green, Urbana, IL 61801, U.S.A.
****畑作研究部 環境制御研究チーム：〒082–0071 北海道河西郡芽室町新生.

系の挙動全体を一挙にくらべることは非現実的だから，ここでも大量情報の効果的縮約が必須となる．

大量情報の効果的縮約の一助として，われわれは主に社会科学の分野で使われている多変量解析の手法に注目した．多変量解析は得られた多自由度のデータから何らかの「本質」を自動的に取り出そうと試みる．典型的な場合には L 項目のアンケートが N 個の対象（例えば，被験者）に対して行なわれ，被験者を L 項目のアンケートの解答によって分類して集団内の傾向を読む，というような使われ方をする．本稿では，従来あまりこのような方法が使われていなかった分野への応用を念頭に置いている．

多変量解析にはいろいろな手法がある．われわれに特に興味があるのは

$$x_n(i) \quad (n = 1, \ldots, N, i = 1, \ldots, L) \tag{1.1}$$

という型の二元データである．ここで n は記述対象（あるいは分類対象 Operational Taxonomic Units（OTU）．そもそもの研究動機の一つに分類学的データの解析があるので，本稿では記述対象をすべて OTU と呼ぶことにする）の番号であり，i は「アンケートの設問」番号である．このようなデータを扱うには，判別分析，因子分析，主成分分析，クラスター分析，など，非常に多くの手法が知られている．本稿では多次元尺度構成法（Multidimensional Scaling; MDS と略記）（林（1976), 比較的最近の解説書として Borg and Groenen（1997)），特に，非計量多次元尺度構成法（Non–metric Multidimensional Scaling; NMDS と略記）(Guttman (1968), Kruskal (1964a, 1964b)) に注目しその応用可能性を探ることにする．その大きな理由は判別分析，因子分析，主成分分析などの手法がいずれも線形性を仮定した手法だからである．(1.1) 式のようなデータは基本的に L 次元の空間内の N 個の点の配置を表現しているとみなせるが，これらの手法はその「座標軸」を線形変換することで「意味のある軸」を見出そうとする．つまり，「真に」意味のある軸がこれらの軸の非線形な関数であれば決してその軸を見出すことはできない．クラスター分析はこの問題からは自由だが一方で OTU の数が非常に多い場合解析結果に意味を見出しにくくなる欠点がある．

多次元尺度構成法は与えられたデータから OTU 間の距離を計算し，その距離を満たすように何らかの空間に OTU を埋め込む（得られた OTU の配置を布置という）ので，基本的に線形性の呪縛から自由である（距離の計算の際に非線形な変換を考慮すればよい）．本稿で扱う NMDS では，非線形な変換をあらかじめ考慮しなくても自動的に都合のよい変換をしてくれるという利点がある．このため非線形かつ複雑な現象にも応用しうる可能性がある．

本稿では，第 2 節でわれわれが提案する新しい NMDS のアルゴリズムと結果の良さを判定する方法を導入する．第 3 節でいくつかの情報縮約的な応用例（土壌微生物の多様性，脳の EEG，植物の大分類）を示し，第 4 節で同じ例を他の多変量解析手段（主成分分析，計量的な MDS，従来からある NMDS）で情報圧縮した結果と比較する．第 5 節は締めくくりである．

2. 順序情報のみによる NMDS のアルゴリズムと当てはめの良さの判定

NMDS のアルゴリズムとしては多くの提案があり，それぞれに実績があるので既存の NMDS を用いても構わないが，われわれは敢えて NMDS の基本的発想に立ち戻ってアルゴリズムから考え直す．そのひとつの理由は，容易に得られるパッケージにある NMDS プログラムは 100 以上の対象を相手にすることを想定していないものが大多数であるのに対し，われわれは大規模データへの応用を狙っていることにある．結果としてわれわれが提案するアルゴリズムの主たる特徴は

1. 順位相関のみを考えるので明白に非計量的である，

2. 解の良さを判定する方法と密接に結び付いている，

である．

1 について：既存の NMDS では，データとして与えられた OTU 間の距離 $\delta(n,n')$（観測距離，proximity）と布置された点の間の対応する距離 $d(n,n')$（再現距離，distance）の中間に $\hat{d}(n,n')$ という量（disparity と呼ばれる）を想定し，$\hat{d}$ と d の差が最小限になるように布置を決める．ここで $\hat{d}$ は δ のなんらかの関数で

$$\delta(n_1,n_2) > \delta(n_3,n_4) \Rightarrow \hat{d}(n_1,n_2) \geq \hat{d}(n_3,n_4) \tag{2.1}$$

となるようなものである．disparity $\hat{d}$ の決め方や，「差が最小」ということの解釈によって，具体的な手順にはいろいろなバリエーションがある．たとえば，Kruskal の方法では，

・布置から計算した $d(n_1,n_2)$ に対して (2.1) 式に従い，かつ，

$$RS = \sqrt{\sum_{n_1,n_2} (\hat{d}(n_1,n_2) - d(n_1,n_2))^2}$$

を最小にする $\hat{d}(n_1,n_2)$ を求める（この操作を単調回帰という）．

・ストレスと呼ばれる量 $S = RS/\sqrt{\sum_{n_1,n_2} d(n_1,n_2)^2}$ を小さくする方向に布置を動かす．

を（適当な布置の正規化操作を含めて）交互に収束するまで繰り返す．しかし，この $\hat{d}$ の導入は本来，不要なはずである．そこでわれわれのアルゴリズムでは $\delta(n,n')$ と $d(n,n')$ の大小順序の差の自乗

$$\Delta = \sum_{n,n'} (T_k - k)^2$$

を考える．ここで T_k は $\delta(n,n')$ が k 番目に大きい時に $d(n,n')$ が何番目に大きいか，という順序を表す．量 Δ を計算することは Spearman の順位相関係数を計算するのと等価である． Δ を最小化する（順位相関を最大にする）という直接的要請によって布置を決める，というのがわれわれの提案するアルゴリズムの基本である．なお，過去にも Spearman の順位相関係数を最適化基準として NMDS を行なうことが試みられた形跡がある（たとえば，Guttman (1968) にそれらしい言及がある）が，現在主流の方法とはなっていない． NMDS の理念の再検討と大規模データの処理効率の向上という両面で，この方向を再考する意義があるとわれわれは考える．

2 について： 通常，NMDS では布置がどれくらい「よい」かを $\hat{d}$ と d の「差」で判断する．しかし，これは非常に間接的な指標であり，相対的な評価しか出来ない．さらに NMDS では「差」の定義が相互に異なっているので，異なった手法同志を比べるのも困難である．そこで布置の良さを判定する一手段として，われわれのアルゴリズムと密接に関連した方法を提案する．この方法は既存の NMDS の結果（＝布置）に対しても適用可能である．

2.1 アルゴリズム

われわれのアルゴリズムは次のとおり．

1. $\delta(n,n')(n,n' = 1,\ldots,N)$ を

$$\cdots \leq \delta(n_1,n_2) \leq \delta(n_3,n_4) \leq \cdots$$

となるように並べる．

2. N 個の点をある空間 $\mathcal{R}$ にランダムに置いて初期布置とする．
3. 点の位置ベクトルを $\sqrt{\sum_n |\vec{r}_n|^2} = N$ となるように規格化する．
4. $d(n, n')$ を計算し
$$\cdots \leq d(n_1, n_2) \leq d(n_3, n_4) \leq \cdots$$
となるように並べる．
5. $\delta(n, n')$ が k 番目に大きく，$d(n, n')$ が T_k 番目に大きい場合，$C_{nn'} = T_k - k$ とする．
6. OTU n に対して，$C_{nn'}$ が正（負）であれば，$\mathcal{R}$ において n を n' に近づくように（から遠ざかるように），$s|C_{nn'}|$ だけ動かす．ここで s は適当な小さな数である．各 OTU i の移動量をベクトルで具体的に書くと
$$s \sum_{n'} C_{nn'} \frac{\vec{r}_n - \vec{r}_{n'}}{|\vec{r}_n - \vec{r}_{n'}|},$$
となる．ここで $\vec{r}_n$ は OTU n の $\mathcal{R}$ における位置ベクトルである．
7. 解（布置）が収束するまで上の 3 に戻る．

このアルゴリズムはきわめて素直に NMDS の発想を実現しようとしたものである．（従来の NMDS 諸アルゴリズムの延長上にこのアルゴリズムはあるわけではない．）原データが真に非計量的であれば順序以外の情報はまったく与えられていないのだから，それを距離 (disparity) に翻訳した後に埋め込み空間での布置からえられる距離とくらべる旧来の手法には計量的夾雑物が入っているとわれわれは考える．このアルゴリズムが順序のみを使う唯一のアルゴリズムというわけではないが，$\delta(n, n')$ を距離に翻訳しないという点で旧来の NMDS アルゴリズムと一線を画する．

このような単純なアルゴリズムが本当に正解がある場合にそれを与えること（例えば，地球上の 1000 個の都市間の距離を $\delta(n, n')$ として与え，現実の都市の配置を求める，など：図 1）はわれわれも確認している．しかし，当然の疑問は，T_k が区分的定数であることからわかるように，少し布置を変化させても距離の順位は変わらないはずだから得られる布置が真の布置と一致するはずはないのではないか，というものである．その通りなのだが，OTU の布置が一般的 (generic) であればこの「遊び」は非常に小さい（だいたい N^{-2} のように挙動する）ので実用上解は一義に決まるとみてよい．

Δ の最小化との関係では，Δ を減らす方向を heuristic に決めているので精密に Δ が最小にできる保証は一般にはない．しかし，このアルゴリズムは，Δ の最適化をその勾配をもとに行う方法の近似であると解釈でき（注 1），また，実例で見たように普通問題はないようである（$\Delta = 0$ を与える真の解はアルゴリズムの不動点である）．

このアルゴリズムでもっとも時間がかかるのは d の並べ換えの部分である．しかし，これはごく普通の高速ならべかえアルゴリズムを用いれば $N^2 \ln N$ のオーダーで実行できる．これに対し，単調回帰でよく使われる up–and–down block algorithm では，$d(n, n')$ から $\hat{d}(n, n')$ を計算する際，最悪の場合，N^4 のオーダーの計算を必要とする．なぜなら，このアルゴリズムでは行と列に nn' のペアを書いた $N(N-1)/2$ 行，$N(N-1)/2$ 列の行列を用意して，1 行ずつ，上三角の領域に属する要素を計算しなければならないからである．N が十分大きいところで，少なくとも up–and–down block algorithm よりわれわれのアルゴリズムははるかに速いはずである．大規模データに対してわれわれのアルゴリズムは単調回帰を用いるより有利であることが期待できる．

更に，単調回帰には別の問題も存在する．例として以下のような場合を考えよう．まず，$\delta = (1, 2, 3, 4, 5)$ であったとする．これに対して，例えば $d = (5, 2, 3, 4, 1), (5, 4, 3, 2, 1), (4, 5, 3, 1, 2)$

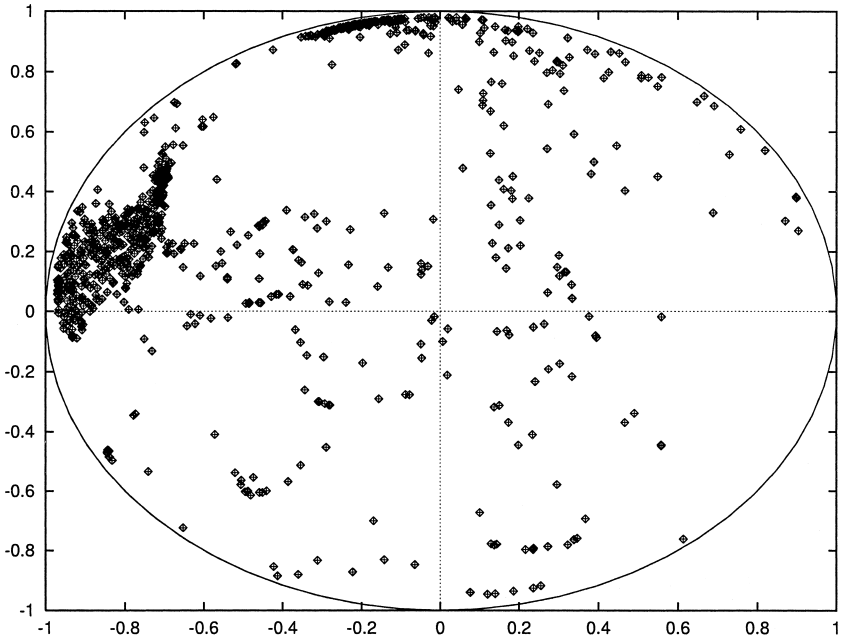

図 1. 地球上の 1000 個の都市の再現．◇ と ＋ がもとの配置と再現をそれぞれ表す．

という 3 つの場合を考えよう．単調回帰のアルゴリズムでは $\hat{d} = (3, 3, 3, 3, 3)$ となってしまう．つまり，ストレス S はみな同じ値となってしまう．しかし，順序の差の二乗であれば，順に $32, 40, 36$ となる．これが直観的な「良さ」(つまり，もとの $(1, 2, 3, 4, 5)$ という順序をどれくらいよく再現しているか) と良く一致しているのは明らかだろう．つまり，残差の評価と言う点でも順位差の方に一日の長があるのである．

一方，Guttman (1968) の rank image 法は発想がここで提案されている方法に近く，計算の手間もソーティングの手間と考えられる．Rank image 法では δ の順にならべた d と大きさの順に並べた d (これが $\hat{d}$ として使われる) との自乗誤差を最小化するので，正しい距離順序へと点配置を最適化することになりそうだが，実際には最適化が目指す目的点が最適化途上の点配置に依存しているため「正解」の認識に問題が生じうる (収束が保証されていず，その問題点はまさに最適化すべき関数の定義にある (Borg and Groenen (1997))．これに対してわれわれの場合は目的とすべき順序が最適化途上の点配置に依存することなく明確に与えられているところが大きく異なる．

2.2 当てはめの良さの判定

さて，得られた布置の良さはどう判定すべきだろうか．上述の Δ が一番もっともらしい統計量であるが，この統計量は独立な量の和ではない．実際，和の項数は $N(N-1)/2$ あるが，データの自由度はデータの個数 N のオーダーであり，個々の項は独立な量とはなり得ない．もちろん，今後，Δ を統計量として研究すべきでこれに直接基づいた統計的検定法を開発すべきである．ここでは，便法として「各 OTU ごとの Δ」とでも呼ぶべき量 $\Delta(n)$ を導入する．これは以下のようにして求める．

1. ある OTU n に注目する．

2. $N-1$ 個の配列である $\delta(n,n')$ と $d(n,n')$ を各々大きさの順に並べる．
3. この 2 つの配列に対して得られた大きさの順序 k', T'_k（$1 \le k', T'_k \le N-1$; T'_k は $\delta(n,n')$ が k' 番目に大きい時に $d(n,n')$ が何番目に大きいか，という順序を表す）から $\Delta(n) = \sum_{n'}(T'_k - k')^2$ を計算する．

この $\Delta(n)$ はそれぞれが異なった OTU の位置の関数である $N-1$ 項の和なのでお互いに独立な量の和とみなせる．この統計量に対して,「2 つの配列が無相関である」という帰無仮説のもとでは N が大きいとき平均が $E[\Delta(n)] = ((N-1)^3 - (N-1))/6$ で分散が $\mathrm{Var}[\Delta(n)] = (N-1)^2N^2(N-2)^2/36$ の正規分布であることが知られている (Lehman (1975))（ただし，タイデータが含まれている場合には表式はもっと複雑になる）．これを用いると，適当な有意水準を決めれば，2 つの配列 $\delta(n,m)$ と $d(n,m)$（ただし $m \neq n$）の相関が有意かどうか判定することが出来る．また，着目している OTU の与えられた有意水準での誤差範囲（配置の遊び）を決めることもできる（ただし本稿ではこれはしていない）．各 n についてこれを見ることで，布置全体の良さが判定できる．

この判定法のよいところは，われわれの NMDS が最小化を目指す Δ と類似の量を判定の基準として用いることができること，また，伝統的な MDS に対してもこの判定を行なうことでわれわれのものをも含め複数の異なったアルゴリズムで得られた結果を比較できることである．

ここでわれわれの方法を既存の方法の一つと比べよう．比較するためのルーチンとしては VisTa (Vista (2000)) に含まれる MDS モジュールを用いた．この MDS モジュールは単純な計量的な（線形を仮定した）MDS であり，SMACOF アルゴリズムを使用している (Young (2000))．データは VisTa 付属の colardism.lsp である．これはコーラ飲料を分析したものである．VisTa では 3 次元未満の埋め込みはできないので，3 次元で行なった．われわれの結果と VisTa の結果は一致しなかった．どちらが「良い」かは難しい問題だが，われわれの判定法ではわれわれの結果は OTU 2 が有意水準 1%を越えてしまうが，VisTa の MDS では全ての OTU が 0.5%基準を満たしたので，VisTa の MDS がわれわれよりややよい，と言えるだろう．判定基準の良否を比較するために，

$$D(n) = \frac{E[\Delta(n)] - \Delta(n)}{\sqrt{\mathrm{Var}[\Delta(n)]}}$$

を図 2 に載せる．

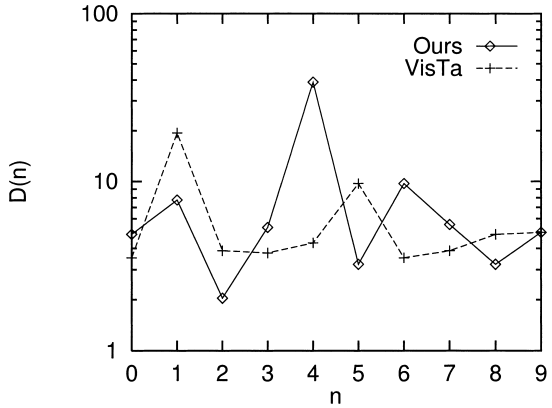

図 2. コーラのデータの布置の判定結果の比較.

$D(n)$ が大きい（つまり $\Delta(n)$ が小さい）ほど δ と d が無相関であるという帰無仮説の有意確率が下がり，そのような n の個数が多いほど得られた布置が良いことになる．全体としてみればほぼ同じ適合度であることがわかるだろう．これは簡単なテストに過ぎないが，われわれの NMDS は機能しているといえよう．

3. 情報圧縮的な応用例

以下に実際にわれわれの NMDS を「情報の圧縮」という観点から用いた例をいくつかあげよう．これらはあくまで NMDS でどのくらいデータの冗長性を除去できるかという観点からなされたものであり，どのように有効な結果が出せるかというところまでは立ち入っていない．

(1.1) 式のようなデータが与えられた場合，もっとも素直な観測距離 $\delta(n, n')$ の定義は

$$\delta(n, n') = \sqrt{\sum_i (x_n(i) - x_{n'}(i))^2} \tag{3.1}$$

ということになろう．もちろん，こうでなくてはならないという理由はない．NMDS を用いるかぎり，$\delta(n, n')$ の大きさの順序を変えねば結果は距離の定義によらず同じである．この意味で NMDS にはある程度の頑健さが期待できる．

さて，(1.1) 式のようなデータを L 次元の線形空間内の N 個の点とみなすことは前に述べた．したがって，上記のような δ の定義を用いれば，L 次元のユークリッド空間に布置できるのは自明である．また，一般に N 個の点は $N-1$ 次元の空間が与えられればどんな $d(n, n')$ を与えられても，必ずその距離関係を満たすように埋め込むことが出来る（例：任意の 3 辺よりなる三角形は三角不等式に反しないかぎりかならず 2 次元平面に描ける）．よって， NMDS で情報を圧縮するときは，L や N よりどのくらい小さい空間次元で適当と判定される布置が見付かるか，が焦点になる．

3.1 土壌微生物データ：もっとも一般的な「アンケート」の場合

土壌には多くの微生物が存在している．それらはなんらかの生態系を構成しており，土壌の状態に多くの影響を与えているであろう．しかし，土壌中の微生物の多くは同定されていないのが現実であり，生態系と土壌の状態の関係はよくわかっていない．横山 (1996) は土壌微生物を区別するために各々の微生物の炭素源利用能を調べ，土壌の状態を定量化しようとした．詳しい解説は省くが，要するにわれわれは (1.1) 式でいうと

n：微生物の種類（未知）
i：炭素源
$x_n(i)$：微生物 n が炭素源 i を利用できれば 1，出来なければ 0

というデータを持っていることになる．炭素源の数は 96 であり，これはこの分析システムを開発したメーカーによって選択されたものである．微生物のサンプル数は 47 であるが，これは無数に存在する土壌微生物（の培養可能な部分集合）からのランダムな選択であるとしてよく，同じ種が二度選ばれている可能性もある．

ここで (3.1) 式の様な距離を使ってもいいのだが，どうせ二値しか取らないので，δ としてハミング距離（この場合は (3.1) で単に平方根を取らない場合と一致）を採用して NMDS で 2 次元のユークリッド空間に埋め込んで布置を得，埋め込みの適切さを判定した．その結果，47 個の OTU のうち，1%以上 5%以下の帰無仮説の有意確率をもつものが 1 つ，0.5%以上，1%以下の有意確率をもつものが 1 つあったが，他はすべて 0.5%以下の有意確率を持つ解を得たので，これを有意な布置として採用した．

さて，この結果を以下のように処理した．得られた布置は

$$(x_n, y_n), n = 1, \ldots, N$$

という N 個の 2 次元ベクトルである．この布置の「意味」を判定するため，この 2 次元の布置の「主軸」と「副軸」をもとめた．主軸とは，

$$y = ax$$

という式で書ける直線で，かつこの直線と点 (x_n, y_n) との距離を δs_n としたとき $\sum_n (\delta s_n)^2$ が最小となるような直線である．簡単に言うと，データのバラツキが楕円形をしていた場合には長軸の方向になるような軸である．共分散を用いた主成分分析（Principal Component Analysis；PCA と略記）は

$$(x_1, x_2, \ldots, x_N)$$
$$(y_1, y_2, \ldots, y_N)$$

というふうに布置を読みかえた 2 本のベクトルから

$$(X_1, X_2, \ldots, X_N)$$
$$(Y_1, Y_2, \ldots, Y_N)$$

という新たな（変換された）座標としてこのような軸を求める多変量解析である．得られた座標 (X, Y) が主成分である．Y は X と直交しているという条件の中で $\sum_n (\delta s_n)^2$ が最小となる軸である．以後，高次元の場合は既に求まっている主成分と直交するように次々と主成分を求めて行く．この PCA を用いれば第 1 及び第 2 主成分として求めたい主軸と副軸が求まる．

その結果，得られた布置の主軸の寄与率は 85%となった．つまり，NMDS を用いることで 2 次元に落ち，しかもそのうちの「主軸」がほとんどの情報を担っているとわかったわけである．事実上，1 次元に落ちたといっても過言ではない．このように NMDS は高次元のデータの情報を圧縮するのに大きな力を発揮するのである．では，この軸の意味はなんだろうか？ NMDS で情報が圧縮されて軸が求まったとしてもその意味はそれほど簡単には求まらない．試行錯誤の末，この軸は

「利用出来る炭素源の数＝炭素源利用能」

と解釈できる軸であることがわかった．個々の微生物が利用できる炭素源の数 C_n は

$$C_n = \sum_i x_n(i)$$

で求まるが C_n を主軸座標 X_n の関数としてプロットしたのが図 3 である．極めて直線性がよく，相関（相関係数の 2 乗）も 0.88 と大きい．つまり，このデータから土壌微生物は「炭素源利用能」という多様度空間の中にきれいに分布していることが分かる．このことの生態学的な意味はともかくとして，心理学などへの応用例で示されているように，NMDS は与えられたデータから秩序を「自動的に」見出す能力があることが確認された．

原データに PCA を直接適用するだけでも炭素源の数の重要性はわかる．しかし，あとで見るように，それがどのくらい決定的な（あるいは重要な）因子であるかは，NMDS での 2 次元への縮約抜きには主張できない．

3.2 EEG：時系列データの場合

前節の NMDS の応用相手は典型的な多変量解析用のデータであり，目新しいことはない．しかし，NMDS には他にもいろいろ使い道がある．たとえば，n チャンネルの時系列データが与

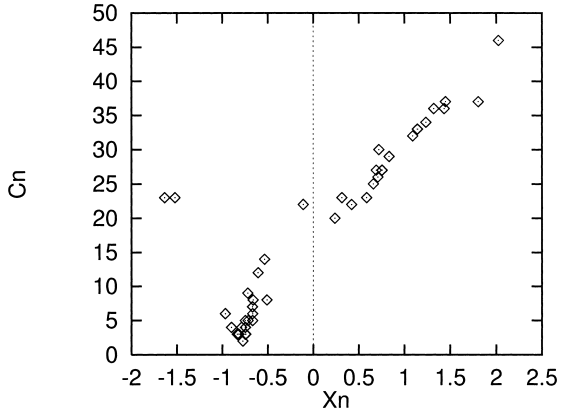

図 3. NMDS で得られた主軸座標と炭素源利用能の総和との関係.

えられたとしよう．このデータからなんらかの「特徴」を抽出することが NMDS で可能だろうか？ ここではそのような時系列データへの NMDS の応用の一例として Electroencephalogram (EEG) を取りあげる．これは脳の複数の部位の電位の時系列的な記録であり，被験者にいろいろな作業を行なわせて，シグナルの変化から脳の機能を解明するためなどに使われる．

Zhang et al. (1995) は次のような実験を行なった．EEG を測定される被験者は 2 つの図形を 0.3 秒間隔で提示される．彼らは 64 チャンネル（一個一個が脳の部位に対応）の EEG 時系列データを多数得，それらをチャンネルごとに単純平均し，同じ図形の提示と異なった図形の提示では EEG 時系列のある点に統計的に有意な差がある，と結論した．

さて，このデータを NMDS では次のように解析した．

n：チャンネル
i：時刻
$x_n(i)$：チャンネル n の時刻 i での電位

とし，(3.1) のように δ を定義して NMDS で 2 次元に埋め込んでみた．つまり，脳の部位ごとの時系列データの差を距離とみなし，64 箇所の部位を布置するのである．図形の認知が検出できるならば，脳の部位のどこか特定の場所の時系列が変動し，その部位の他の部位との相対的な位置が同じ図形を提示したか異なった図形を提示したかで異なると期待できる．なお，δ は 10 回の独立な試行すべての和をとって求めた (EEG (2000))．

結果は 64 個の OTU のうち，同じ図形を提示した場合で 1%以上 5%以下の OTU が 2 個で他の OTU はすべて 0.5%以下の帰無仮説の有意確率，異なった図形を提示した場合はすべての OTU が 0.5%以下の帰無仮説の有意確率，となり，ほぼ 2 次元で埋め込むことが出来た．これらの布置の主軸と副軸を求めると同じ図形の場合の主軸の寄与率は 92%，違う図形の主軸の寄与率は 83%となり，ともにほとんど 1 次元的な縮約が出来ることがわかった．PCA によるさらなる解析を行なったが，この二つのデータに有意の差は見いださなかった．原論文の結果を注意深く見ると統計的に有意の差と言っても非常にわずかであり，大規模な縮約に耐えて判然と見えるような差はないということであろう．

この布置の意味を理解するため，主軸と副軸の座標値を頭蓋上の各チャンネルの位置に配置し等高線を描いてみたところ，主軸の座標値は眼球運動などに対応するいわば「ノイズ」信号であり，副軸の信号だけが，視覚野である左右後頭部に極在した信号を持つ意味のあるデータ

であることがわかった．このことから少なくとも NMDS はノイズと意味ある信号の分離には使えることが判明した．

3.3 分類学：2 種のデータの比較

最後に 2 つの異なったデータを比較するために NMDS を使う方法を紹介しよう．具体的には生物の種とその遺伝子塩基配列および表現形質表が与えられた場合，後二者の相関を議論する道具として NMDS を用いるのである．

n：種
i：表現形質，または，塩基
$x_n(i)$：n 番目の種の i 番目の表現形／塩基の「値」

距離 δ の定義はやや面倒である．例えば，表現形の場合，値が二値とは限らない．ある形質（例えば，角の数）などに二値以上の値があることがある．その場合，一番単純には異なった形質はすべて異なっているという意味で同等に扱わねばならず (3.1) 式のような単純な距離の定義はできない．その意味で，δ の定義にある程度任意性がはいるが，それでも NMDS の場合，δ の大小関係が（ほとんど）変わらなければ，結果の布置も変わらないという頑健さが期待できるので，結果は無意味ではないだろう．

3.3.1 シダデータの解析

51 種のシダの表現形質と塩基の配列データが与えられている（対象となるシダの学名は表 1）

表 1. 分類に使われている 51 種のシダなどの学名.

Anemia mexicana	Cephalomanes thysanostomum
Asplenium filipes	Lygodium japonicum
Azolla caroliniana	Angiopteris evecta
Blechnum occidentale	Marsilea quadrifolia
Cheiropleuria bicuspis	Matonia pectinata
Cyathea lepifera	Metaxya rostrata
Blotiella pubescens	Botrychium strictum
Dennstaedtia punctilobula	Osmunda cinnamomea
Histiopteris incisa	Ceratopteris thalictroides
Lindsaea odorata	Plagiogyria japonica
Lonchitis hirsuta	Loxogramme grammitoides
Microlepia strigosa	Polypodium australe
Monachosorum henryi	Psilotum nudum
Pteridium aquilinum	Acrostichum aureum
Calochlaena dubia	Adiantum raddianum
Dicksonia antarctica	Coniogramme japonica
Dipteris conjugata	Platyzoma microphyllum
Davallia mariesii	Pteris fauriei
Elaphoglossum hybridum	Taenitis blechnoides
Nephrolepis cordifolia	Salvinia cucullata
Onoclea sensibilis	Actinostachys digitata
Rumohra adiantiformis	Thelypteris beddomei
Diplopterygium glaucum	Vittaria flexuosa
Stromatopteris moniliformi	Lycopodium digitatum
Micropolypodium okuboi	Equisetum arvense
	Cycas circinalis

(形質データ(2000a)). この表にはいくつか外群的な(シダでない)種が含まれている．表現形質の種類数は 77 で形質により 0, 1, 2, 3, 4, 5 までの値をとる．ある形質に対して 2 種類の値をとるような種もある．この例では，距離 δ は，表現形質の場合，基本的に

$$\delta(n, n') = L - \sum_i \delta[x_n(i), x_{n'}(i)]$$

とする．ただし，$\delta[x_n(i), x_{n'}(i)]$ は以下の通りとする．

$$\delta[x_n(i), x_{n'}(i)] = \begin{cases} 1 & x_n(i) = x_{n'}(i) \\ 0 & x_n(i) \neq x_{n'}(i) \\ 0.5 & x_n(i) \text{ か } x_{n'}(i) \text{ が複数の値をとり，部分的に一致している場合} \end{cases}$$

他方，塩基配列の場合は単純に

$$\delta[x_n(i), x_{n'}(i)] = \begin{cases} 1 & x_n(i) = x_{n'}(i) \\ 0 & x_n(i) \neq x_{n'}(i) \end{cases}$$

とした．2 次元に埋め込んだ結果は

表現形質

帰無仮説の有意確率が	
5%以上の OTU	1 個
1%以上 5%未満の OTU	1 個
0.5%以上 1%未満の OTU	3 個

塩基

帰無仮説の有意確率が	
5%以上の OTU	1 個
1%以上 5%未満の OTU	2 個
0.5%以上 1%未満の OTU	2 個

などとなった．いつものように共分散の PCA を使って主軸と副軸を求める．表現形質では主軸の寄与率は 69%，塩基では 72%であり，いままでの 2 例に比べると主軸の寄与率は低かった．つまり，比較的 2 次元性が高い，ということになるだろう．こうして得られた結果の主軸同士の相関は 0.8 程度であり，副軸同士には相関が無かった．このことから，NMDS を使うことで全く異なった種類のデータの一致の程度についてある程度のことが言えることがわかる．今の例では，塩基配列と表現形質の並行関係が見えている．もちろん，よく知られているように，表現形質と塩基による系統解析の結果は必ずしも一致しない．しかし，われわれの知るかぎり両者の一致度を見る方法は一度も提案されたことがないようである． NMDS によりこれがある程度できる可能性が示唆されたといってよい．

3.3.2 緑藻とコケ植物

59 種の緑藻とコケ植物の表現形質と塩基の配列データが与えられている(対象となる緑藻・コケ植物など(外群を含む)の学名は表 2)(形質データ(2000b)). 表現形質の種類数は 110 種類で形質により a, b, c, d, e までの値をとる．種によってはある形質に対して 2 種類の値をとる．この場合，距離 δ の与え方は基本的に前小節と同じであるが，こちらは不明の項目(値が不明の形質や塩基)が多い．特に塩基の場合は種によっては非常に多数の塩基の個数が不明なので，L' =[n 種と n' 種でともに塩基が判明している数] を用いて

$$\delta(n, n') = \frac{L' - \sum_i \delta[x_n(i), x_{n'}(i)]}{L'}$$

表 2. 分類に使われている 59 種の緑藻・コケ植物などの学名.

Glycine max	Cephaleuros parasiticus
Oryza sativa	Characium vacuolatum
Zamia pumila	Dunaliella parva
Psilotum	Chlamydomonas reinhardt
Equisetum arvense	Volvox carteri
Atrichum	Chlorococcopsis min
Notothylas breutellii	Draparnaldia plumosa
Phaeoceros laevis	Uronema belkae
Porella pinnata	Chlamydomonas moewusii
Conocephalum conicum	Stephanosphaera pluvial
Asterella tenella	Carteria radiosa
Riccia	Gonium pectorale
Klebsormidium flaccidum	Chlorella kessleri
Coleochaete nitellarum	Chlorella vulgaris
Fissidens taxifolius	Prototheca wickerhamii
Plagiomnium cuspidatum	Chlorella prototheco ide
Micromonas pusila	Chlorella minutissima
Mantoniella squamata	Neochloris aquaticus
Nephroselmis pyriformis	Neochloris vigenis
Pedinomonas minutissima	Pediastrum duplex
Tetraselmis carteriifor	Scenedesmus obliquus
Enteromorpha intestinal	Characium hindakii
Ulva fasciata	Chlorella fusca
Ulothrix zonata	Ankistrodesmus falcatu
Cymopolia barbata	Pseudotrebouxia gigante
Batophora oerstedtii	Pleurastrum terrestre
Codium decorticatum	Characium perforatum
Cladophoropsis membrano	Parietochloris pseudoal
Blastophysa rhizopus	Friedmannia israelensis
Trentepohlia sp.	

と定義しなおして使った．これがいい選択であるかどうかはわからないが，もとの定義では $x_n(i)$ と $x_{n'}(i)$ の一致数が多いほど，距離が近くなるので，このままでは不明の塩基数が多いと非常に距離が遠くなってしまう．種によっては不明の塩基数が多いので，このままでは距離の大小が不明塩基数の多寡で決ってしまい適当ではない（逆に言うとそれほど種によって塩基の不明数に巾がある）．このように距離を決めて 2 次元に埋め込んだ．結果は

表現形質： 帰無仮説の有意確率が 0.5%以上 1%未満の OTU：2 個
塩基：帰無仮説の有意確率が 1%以上 5%未満の OTU：1 個

とシダ類よりはずっとよい埋め込みである．これらを各々，共分散の PCA で解析して主軸・副軸を求めると，表現形質では主軸の寄与率は 94%，塩基では 76%であり，表現形質の方はほとんど一次元的な多様性しかないことが示唆される．

それぞれの軸の相関をとってみると，表現形質の主軸と相関があるのは，塩基の副軸で相関は 0.66 となり，塩基の主軸とは 0.25 程度の相関をもっていた．つまり，主軸同士の相関は大きくはなかった．そこで，塩基の主軸と副軸を回転し，表現形質の主軸との相関が最大になるように選んだところ（このためには表現形質の主軸を従属変数，塩基の主軸と副軸を従属変数とする線形重回帰分析をすればよい），相関は 0.5 まで上昇した．この軸がおそらく，塩基と表

現形質での「共通軸」となるのだろう．主軸同士が相関しなかったことについての考察は線形MDS の結果と比較するところで再び触れる．

3.4 応用例のまとめ

この節では，土壌微生物の炭素源利用能，脳の EEG データ，植物の表現／塩基形質についてNMDS を使用した解析例を示した．いずれも 2 次元にほぼ縮約することが出来，一部は事実上1 次元に落すことも出来た．もとのデータの高次元性から考えると驚くほど，低次元に落すことが出来る能力を NMDS は持っていることがわかった．また，得られた布置の「軸」も無意味なものではないことがある程度理解できたであろう．NMDS の情報の圧縮性能があまりに強力なのでここまで圧縮して大事な情報が本当に落ちていないかどうか，やや不安でさえある．

情報を圧縮してもその「意味」は相変わらず考えねばならない．土壌微生物の場合は，「主軸」が「炭素源利用能」を表現していると述べたが，これは試行錯誤でしかわからなかった．しかし，正準相関分析や重回帰分析などを用いれば座標の大凡の意味付けは案外簡単かも知れない．一方，EEG については副軸の座標値の頭蓋上での分布を見ることでその意味がわかったが，原データの取り方に立ち返ればこれもかなり自然な方法で意味がわかったことになる．植物分類群についてはまだ軸の意味を明確に出来ていず，単に 2 種のデータ（表現形質と塩基データ）の一致を見るにとどまっている．軸の意味をいかに見出して行くかが今後の課題であろう．

4. 既存の手法との比較

前節では，NMDS の情報圧縮能力についてわれわれの提案したアルゴリズムを使っていくつか応用例を扱った．しかし，実用上の問題を議論するならば，公平を期すためにも，いくつかの他の手法（NMDS 以外の手法，またわれわれとはことなったアルゴリズムを使う NMDS）がどれくらい情報圧縮に有用か検討すべきだろう．

比較すべき相手として PCA，（線形）計量的な MDS，従来の NMDS，の 3 つを取り上げる．PCA と MDS については前述の VisTa のモジュールを，従来の NMDS については ALSCAL (ALSCAL (2000)) の ordinal モジュールを採用した．

4.1 土壌微生物

土壌微生物データの場合，主軸が求まり，その軸が「炭素源利用能の総和」を表現する軸だということがわかっている．そこで，他の手法で同じように「主軸」を求めて，同じような解析が可能かどうかを見ることで，比較が可能であろう．

4.1.1 PCA

まずは PCA と比較しよう．

$$(x_1(1), x_2(1), \ldots, x_n(1), \ldots, x_N(1))$$

$$(x_1(2), x_2(2), \ldots, x_n(2), \ldots, x_N(2))$$

$$\cdots\cdots\cdots\cdots$$

$$(x_1(i), x_2(i), \ldots, x_n(i), \ldots, x_N(i))$$

$$\cdots\cdots\cdots\cdots$$

$$(x_1(L), x_2(L), \ldots, x_n(L), \ldots, x_N(L))$$

のようにデータを L 本の N 次元ベクトルであるとみなし，共分散で PCA を行なった．共分散で行なったことに特に意味は無く，相関で PCA を行なっても構わないであろう．

第 1 主成分と炭素源利用能との相関は 0.90（われわれの NMDS では相関 0.88）なので，PCA でも炭素源利用能が重要であることはわかる．しかし，第 1 主成分の寄与率は 52%で，第 10 主成分までの累積寄与率でも 90%に達しなかったことからわかるように，PCA の縮約能力は悪く，結果として，炭素源利用能がどのくらい決定的か判然としない．

4.1.2 線形計量 MDS

距離 $\delta(n, n')$ は NMDS を上で実行したときと同じものを使って，MDS を行なった．VisTa では 3 次元以上の MDS しかできないので，3 次元で行なった．得られた結果を同じように共分散の PCA で回転し，主軸，副軸を求めた（3 次元なので副軸は 2 つ求まる）．この軸と炭素源利用能の総和との相関をみると 0.97 となっており，むしろわれわれの NMDS による 2 次元の埋め込みよりよい結果となった．

この差がどこから来たのか確認するため，われわれの NMDS でも同様に 3 次元の布置を求めて，主軸を求め，炭素源利用能の総和との相関を計算したところ，0.96 という高い相関を得た．結局のところ，計量 MDS がわれわれの NMDS よりすぐれているわけではなく，3 次元の布置を 2 次元に押し込んだため，主軸に無理が来て軸がゆがみ相関がやや落ちたのであり，基本的に計量 MDS とわれわれの NMDS は同程度であり，PCA よりやや優れている，ということになるだろう．

4.1.3 なぜ，計量 MDS はよく PCA はやや悪いのか?

しかし，それにしても，PCA や線形計量的な MDS で同等，あるいは，それほど劣らない結果が得られるなら，わざわざ NMDS を導入する意味はあるのだろうか？ 大体，なぜ，線形計量的な MDS が NMDS と同等なのだろうか？

これを理解するために 3.1 節で得られた 2 次元の布置の場合の $d(n, n')$ と $\delta(n, n')$ のグラフを載せる（これは一般に Shepard Plot と呼ばれている）．NMDS は本来， d と δ の非線形な関

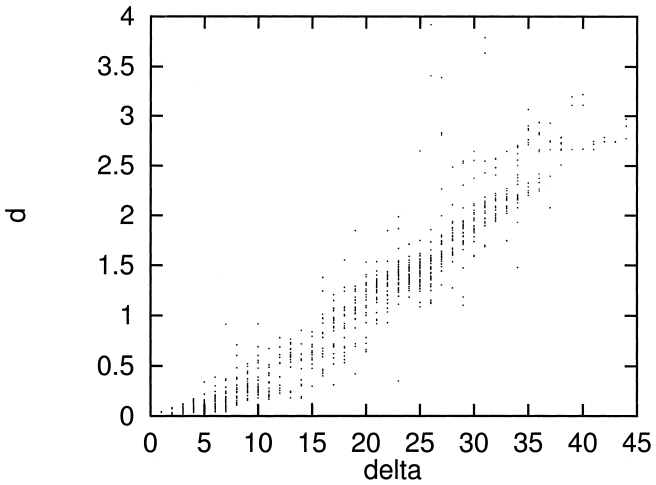

図 4. 土壌微生物データの 2 次元布置の Shepard Plot.

係まで考慮して両者の大小関係がなるべく変わらないように d を求める．しかし，図 4 を見ればわかるようにこの場合，明らかに両者の間に線形的関係が成り立っている．このような場合には線形変換しか考えなくてもよい結果が得られるのは明らかだろう．もちろん，結果として線形法でよかったから NMDS は不要，とは言えない．いつも線形とは限らないからだ．むしろ，線形でよい場合には線形の MDS と同じ結果を出すと言う意味でいつも NMDS を使っておいた方が無難である．

なお，PCA による布置は，線形 MDS でハミング距離の代りに距離 (3.1) を取った場合に近いと思われる．これが PCA が線形 MDS より悪い理由だとすると，線形 MDS がうまくいったのは，偶然 δ_{ij} としてハミング距離を取ったのがよかったのに過ぎず，距離 (3.1) をとれば，結果は悪くなったと予想される．こう考えると，NMDS はやはり必要である．

4.2 EEG

前小節で述べたように δ と d の関係が線形的であれば，NMDS を使う意味はない．EEG の場合はどうだったのだろうか．図 5 に 3.2 節で得られた同じ図形を提示した場合を 2 次元に布置した場合の Shepard Plot を示した．明らかに線形的関係が成立している． VisTa の線形計量 MDS モジュール（3 次元）で得られた布置とわれわれの NMDS で得られた布置を比較するため

$$
\begin{gathered}
(X_1^a, X_2^a, \ldots, X_N^a) \\
(Y_1^a, Y_2^a, \ldots, Y_N^a) \\
(Z_1^a, Z_2^a, \ldots, Z_N^a) \\
(X_1^b, X_2^b, \ldots, X_N^b) \\
(Y_1^b, Y_2^b, \ldots, Y_N^b) \\
(Z_1^b, Z_2^b, \ldots, Z_N^b)
\end{gathered}
$$

という 6 本のベクトル（a はわれわれの結果，b は VisTa の結果，X, Y, Z は各々，主軸，1 本目の副軸，2 本目の副軸，の座標である）を共分散の PCA で解析したところ，第 1 主成分（主

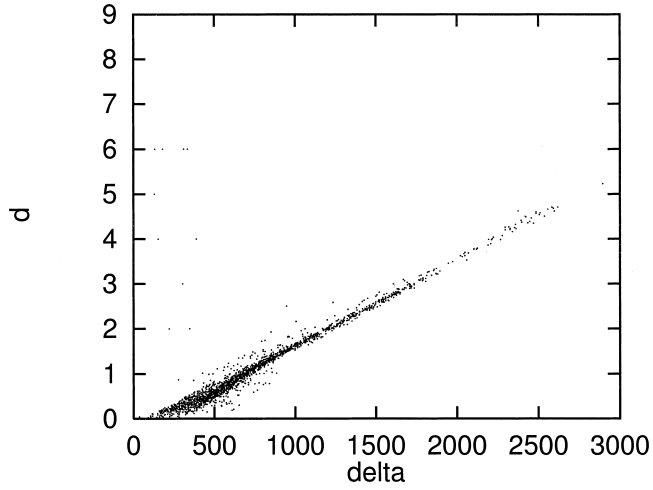

図 5. EEG（同図形提示）の 2 次元布置の Shepard Plot.

軸）の寄与率が 87%，第 2 主成分（1 本目の副軸）までの累積寄与率が 95%，第 3 主成分（2 本目の副軸）までの累積寄与率が 98%であり，つまり，この 6 本のベクトルはほとんど 2 本のベクトルに縮約できるのである．これはもともと，2 次元に縮約できるデータをわざわざ 3 次元で縮約しているのだから当然だろう．いずれにせよ，VisTa の結果とわれわれの NMDS の結果とは違いがないのは明らかである．

ここで最後にもう一点だけ確認しておく．確かにわれわれの NMDS は線形の範囲内で布置がもとまることを示したが，本当は非線形な変換でよりよいものがあるのに，われわれのアルゴリズムでは見逃しているという可能性はないだろうか？

この点を明確にするため，ALSCAL の ordinal モジュールを使って，2 次元の布置を求めた．$\delta(n, n')$ の定義は同じである．ALSCAL の ordinal モジュールはよく使われている非計量 MDS のモジュールである．詳細は省略するが，同じ図形を提示する場合も，そうでない場合も，われわれの手法で得られた Shepard Plot と同じようなものしか得られず，われわれの NMDS が見逃している非線形な変換がもしあるとしても，それは ALSCAL でも見出せないようなものであることがわかった．この意味ではわれわれの NMDS は既存の NMDS モジュールである ALSCAL の ordinal モジュールと同じような結果を出していると言えよう．

4.3 分類学

4.3.1 シダの塩基データ

図 6 にシダの 3.3.1 節で得られた 2 次元の布置の塩基データの場合の Shepard Plot を載せる．δ と d の関係があきらかに非線形である．この場合は NMDS が有効である可能性がある．

そこで，δ は同じままで VisTa の MDS モジュール（線形計量 MDS）を適用した．その結果，51 種のうち

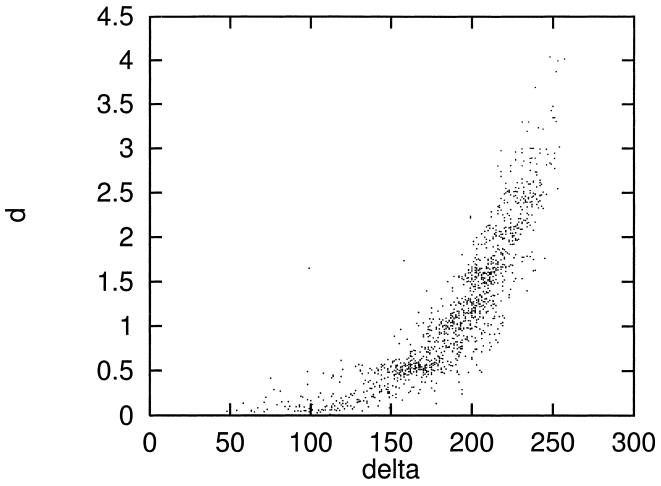

図 6. シダの塩基データの 2 次元布置の Shepard Plot.

塩基

帰無仮説の有意確率が	
5%以上の OTU	6 個
1%以上 5%未満の OTU	11 個
0.5%以上 1%未満の OTU	5 個

となってしまい，意味のある布置が得られなかった．つまり，実際に非線形な関係が出て来れば，線形計量の MDS は役に立たない．このような場合があるので，NMDS は必要なのである．

それでは，既存の NMDS とわれわれの NMDS では求まる δ と d との非線形関係は異なるのだろうか？ このことを確認するため，3 次元の布置を ALSCAL の ordinal モジュールとわれわれの NMDS で求めてみた．結果は，非線形関係のみならず，求まった布置も同じであった．つまり，同様に ALSCAL の 3 次元の布置とわれわれの MDS で求まった 3 次元の布置を並べて 6 本のベクトルとし，共分散の PCA にかけると，第 3 主成分までで 95%の累積寄与率となった．このことからわかるように，われわれの NMDS は非線形的な関係が実際にある場合にも既存の NMDS と同等の結果を出しうることが例示された．

4.3.2 緑藻／コケ植物の塩基データ

最後に緑藻／コケ植物の塩基データを再び取り上げよう．図 7 に 3.3.2 節で得られた 2 次元の布置の場合の Shepard Plot を描いた．非常に乱れていて単純な関数関係がないことがみてとれる．このようになった原因はもとの塩基データに欠損が多く，不完全なデータだったためであろう．われわれの経験では塩基データの埋め込みを行なうと δ と d の間に前節で見たような強い非線形のベキ的な関数関係を得るのが普通である．そうならないことからもこのデータにはかなり問題があるといってよいだろう．

このデータに VisTa の MDS モジュールで 3 次元の線形計量 MDS を試した．驚いたことに，その種に関連する δ と d が無相関であるという仮説が，すべての種について棄却できる，つまりきわめていい布置なのである．δ と d のきれいな関数関係がない場合には，NMDS を使う意味はあまりないのだろう．非線形変換を許容したところで結果はそれほど良くはならないのだ．

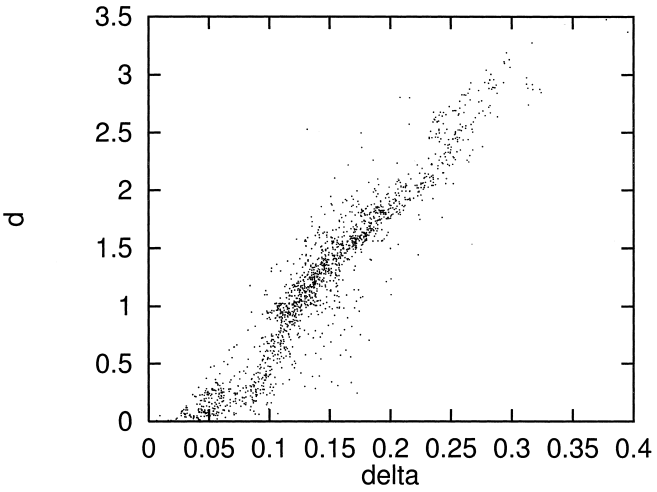

図 7. 苔／緑藻の塩基データの 2 次元布置の Shepard Plot.

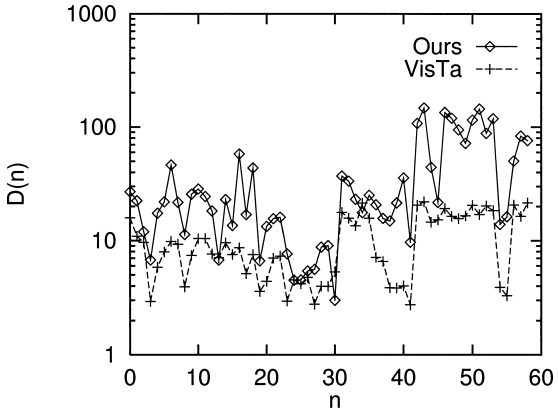

図 8. 苔／緑藻の塩基データの布置の適切さの比較.

3.3.2 節で，主軸同士の相関（塩基 vs 表現形）が低かったのはおそらく，この乱れた塩基データのせいであろう．

参考までに 3 次元の VisTa とわれわれの NMDS で得られる 3 次元布置について $D(n)$ を図 8 に描いておいた．もちろん，われわれの方が圧倒的に大きな $D(n)$ の値を示す．しかし，劣っている VisTa の布置も十分適切である．

5. おわりに

既存の方法との比較でわかったことは情報圧縮に NMDS が必須な局面は多くないかも知れないが，明らかに NMDS が MDS にまさる場合があるということである．

確かに，一番素直な (3.1) 式のような距離の定義が実世界の構造を反映していて線形変換で十分であったり，あるいは最後の例（緑藻／苔の塩基データ）で見たようにデータがあまりにばらついていて非線形性を取り入れるほどのことがない場合は多いだろう．多変量解析というとき NMDS が PCA などと比較するとあまり使われないのはこのためだろう．しかし，シダの塩基データに見たように，本当に非線形な関係がある程度きれいに見えると NMDS は威力を発揮しうる．そこで，実践的には，まず，線形計量 MDS で処理し，われわれが導入した当てはめの良さを見る方法が（従来のストレスと異なり）MDS を実現するアルゴリズムへの依存性がなく線形計量 MDS にも有効なことを利用して処理結果を判定し，満足な結果と見られない場合に NMDS を使う，ということも考えられる．

われわれの NMDS が ALSCAL という標準的アルゴリズムを越える結果を出すような例にここでは出会わなかった．しかし，$\hat{d}$ という余計な介在物を導入していないわれわれの NMDS の方がより自然な結果を出す場合は多くないにせよ想定できるだろう．

いずれにせよ，$x_n(i)$ があるのにわざわざ (3.1) 式のようなことをして，距離 δ を求め，NMDS で布置を求める，という「迂遠な」作業をしてまでデータに含まれる情報を縮約しようとすることはわれわれの知るかぎりされて来なかった．だが，土壌微生物データの例に見たように，NMDS の情報縮約能はきわめて大きく，数十次元の高次元データを 2，3 次元に縮約するということがしばしば可能になる．そもそも，人間が研究する気になるデータは，大抵この程度の単純なデータなのかも知れない．この意味でも，NMDS を使う情報縮約法はさらなる探究開発

に値するのではないか.

注 1)
Δ の最適化をその勾配をもとに実行するには，(−) 勾配ベクトル

$$-\frac{\partial}{\partial \vec{r}_m}\sum_{n,n'}(T_k-k)^2$$

の方向に各点 m を逐次移動させていけばよい．連鎖律を使うと

$$=-\sum_{n,n'}2(T_k-k)\sum_{pair}\frac{\partial d_{pair}}{\partial \vec{r}_m}\frac{\partial T_k}{\partial d_{pair}}$$

となる（ここで d_{pair} はペア距離であり，和は点 m を含むペアについてのみとる）．これがゼロになるところが Δ が極小の位置である．一方，われわれのアルゴリズムでは

$$-\sum_{n,n'}2(T_k-k)\frac{\partial d_k}{\partial \vec{r}_m}=0$$

を満たすように $\{\vec{r}_m\}$ を決めよ，ということになっている．ここで d_k はデータ中の順位が k であるペアの距離である．つまり

$$\sum_{pair}\frac{\partial d_{pair}}{\partial \vec{r}_m}\frac{\partial T_k}{\partial d_{pair}}\to\frac{\partial d_k}{\partial \vec{r}_m}$$

の置き換えがわれわれのアルゴリズムではなされている．この意味は，（真の順位が）k 番目のペアが対象 m を含むときのみその T_k の変化への寄与を取り入れ，それ以外のペアの距離が変わったことによる T_k の変化を無視することである．T_k は $\{d_{pair}\}$ の関数として区分的に定数だからその微分はきわめて特異である．力学系の用語で言えば，正直に勾配法を使うと，デルタ関数的な力（撃力）で駆動される散逸力学系を取り扱うことになる．当然，特異性を均す操作を施した系を考えるべきである．その際 T_k は特に d_k に依存するが，対象の総数 N が十分に大きければ，撃力の生じるところは空間的にかなり隣接して存在するので $\partial T_k/\partial d_k$ を均して一定とおいても構わないだろう．他方，T_k の他のペア距離への依存性は弱い（そのための変化は頻繁ではない）ので無視するのが尤もらしい．これがわれわれのアルゴリズムの本質である．したがって，われわれのアルゴリズムは順序統計量 Δ「だけ」では決まらない最適化基準を布置に課している．だが，その誤差は実用上問題が無い程度であると期待される．

謝　辞

本稿の執筆にあたり，編集委員の伊庭幸人氏から貴重なコメントを多数頂いた．ここに感謝する．

参 考 文 献

ALSCAL (2000). ALSCAL: http://forrest.psych.unc.edu/research/alscal.html

Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*, Springer, New York.

EEG (2000). データは http://kdd.ics.uci.edu/databases/eeg/eeg.html の smni_eeg_data.tar.gz の中の c_m_co2c0000337（同図形提示）と c_n_co2c0000337（異図形提示）を用いた.

Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika*, **33**, 469–506.

林知己夫 (1976).『多次元尺度解析法：その有効性と問題点』，サイエンス社，東京.
形質データ (2000a). データは
表現形質：http://ucjeps.berkeley.edu/bryolab/GPphylo/ferndata/P95morph.nex
塩基配列：http://ucjeps.berkeley.edu/bryolab/GPphylo/ferndata/P95rbcl.nex
から得た.
形質データ (2000b). データは
表現形質：http://ucjeps.berkeley.edu/bryolab/GPphylo/Nexus/GPMorph.txt
塩基配列：http://ucjeps.berkeley.edu/bryolab/GPphylo/Nexus/GPMolecular.txt
から得た.
Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetirc hypothesis, *Psychometrika*, **29**, 1–27.
Kruskal, J. B. (1964b). Multidimensional scaling: A numerical method, *Psychometrika*, **29**, 115–129.
Lehman, E. L. (1975). *Nonparametrics*, Holden–Day, San Francisco.
Vista (2000). http://forrest.psych.unc.edu/research/index.html
横山和成 (1996). 土壌微生物群集の多様性評価，土と生物，**47**, 1–8.
Young, F. (2000). 私信.
Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W. and Litke, A. (1995). Event related potentials during object recognition tasks, *Brain Research Bulletin*, **38**, p.531.

Proceedings of the Institute of Statistical Mathematics Vol. 49, No. 1, 133–153 (2001) 153

New Possibility of Non-metric Multidimensional Scaling

Y-h. Taguchi

(Department of Physics, Chuo University)

Y. Oono

(Loomis Laboratory of Physics, University of Illinois at Urbana-Champaign)

Kazunari Yokoyama

(Upland Agriculture Research Center, Hokkaido National Agricultural Experiment Station)

Needs for extracting features from large scale data of, e.g., genetic systems, nonlinear dynamical systems, etc., are increasing these days. Non-metric multidimensional scaling (NMDS), a kind of multivariate analysis used mainly in social sciences, may be of use for such feature extraction. NMDS tries to imbed objects into a certain metric space, e.g., a Euclidean space, so that the rank order of distances between objects is maximally preserved. The method may be regarded as the most unprejudiced way to extract features, but conventional approaches do not seem to have paid particular attention to large scale data. We propose a new algorithm of NMDS that is efficient for large scale data and introduce a statistical test to evaluate how well the resultant configuration explains the data.

Key words: Multidimensional scaling, large scale data, information compression, goodness of fit.