

因子分析法とパターン展開法

醍醐元正

(同志社大学経済学部教授)

1 はじめに

筆者が属する研究グループでは多波長衛星データを解析するパターン展開法 (PDM) [藤原 96] と名付けられた手法を研究開発して来た。PDM は元々 LANDSAT/TM センサーで得られるデータのうち熱赤外域以外の六つの測定波長域を利用した解析手法として開発されたが、その後の研究によって広い範囲の波長について適用可能である事が解っている [Muramatsu 00]

PDM は地上被覆物の反射スペクトルを三つの基本的なスペクトルパターンの合成によって表そうとする解析方法である。その意味では PDM は spectral mixing 法 [Adams 95] と同種の解析手法であると言える。spectral mixing 法との違いは、spectral mixing 法では解析目的やデータ毎に展開の元になる基本スペクトルパターンを変えるのに対して、PDM では常に同じ基本スペクトルパターンを使用する、という点にある。この事によって PDM では解析間の比較が容易になるのである。

一方、見方を変えれば PDM はバンド数だけある衛星データの次元を 3 次元に圧縮しているとも考えられる。次元圧縮は主成分分析 [Gonzalez 77, Merembeck 80] 等でも同様に行えるが、主成分分析では新しく構成された軸に実体的な意味はない。それに対して PDM では三つの基本スペクトルパターンがそれぞれ、水・植生・土壌という意味を持つのである。各軸に意味を持たせる事が可能なのは、PDM では斜交軸を使用しているからである。また斜交軸を使えば座標軸の直交性は考慮しなくて良いから、バンドの増減に対して軸を再構成する必要がなく、同じ座標系を使用出来る。この事を利用すれば、同一の基本スペクトルパターンをバンド数の異なるセンサーに適用する事により統一した視点からそれらのセンサーを比較出来る。

一方 spectral mixing 法では mixing に使用するそれぞれの基本スペクトルパターンに意味があり、また基本スペクトルはバンド数とは無関係に定める事が出来る。即ち PDM にも存在するこれらの性質は、spectral mixing 法の観点からは自然に理解出来る事なのである。しかし同じ手法を多変量解析の観点から解釈しようとするると斜交座標が必要不可欠になって来ると言える。

斜交座標を使用する多変量解析の手法は実は他にも存在する。因子分析では直交座標だけで

はなく斜交座標も利用する。しかもそれは因子を抽出する事、即ち意味を持つ座標軸を構成する為に利用されるのである。この事からも、PDM では斜交軸を使用している事により三つの軸がそれぞれ意味を持ち得る、という我々の主張が正しいと理解出来る。

因子分析法は数学的には必ずしも明瞭に定義されている訳ではない。因子分析法が仮定するモデル自身の一意性の問題等も未だに解決されていないとの事である [奥野 81, p. 327]。一方、それ故に様々な因子即ち座標軸の抽出方法が考案されている。その中には勿論、斜交座標としての因子の抽出方法もある。本稿では因子分析の手法の一つを衛星データに適用して斜交座標を構成し、それを PDM の基本パターンと比較した結果について述べる。先ず次節で因子分析法の概要、3 節で今回衛星データに適用した時の注意点、最後に解析結果について述べる。なお次節以下因子分析法の解説ではアルゴリズムの詳細についてはふれない。因子分析法を使用するに当たっては [芝 79, 奥野 81] 等を参考にしており、詳細についてはそちらを参照されたい。

2 因子分析法

因子分析法の基本は、観測されたデータを幾つかの再構成された変量の一次結合によって次の様に表すものである。

$$z_{ij} = \sum_{k=1}^m a_{jk} f_{ik} + d_j u_{ij}$$

ここで z_{ij} が個体 i ($i=1, N$) に対する j 番目 ($j=1, n$) の観測変数であるが、因子分析法では z_{ij} は標準化されている、即ち

$$\bar{z}_j = \frac{1}{N} \sum_{i=1}^N z_{ij} = 0.$$

$$s_j^2 = \frac{1}{N} \sum_{i=1}^N (z_{ij} - \bar{z}_j)^2 = 1.$$

であるのが一般的である。 f_{ik} ($k=1, m$) 及び u_{ij} 等が再構成された変数即ち因子を表し、 a_{jk} , d_j 等はこの再構成された変数に掛る重み係数である。 f_{ik} は個体 i に対する第 k 因子の値を表すが、この m 個の値は全ての観測変数について共通に用いられるので共通因子と呼ばれる。一方 u_{ij} は n 個の変数に個別に対応する固有の変動を表しているので独自因子 (特殊因子) と呼ばれる。そしてこれらの因子も次の様に標準化して表す。

$$\bar{f}_p = \frac{1}{N} \sum_{i=1}^N f_{ip} = 0. \quad (p=1, 2, \dots, m)$$

$$\begin{aligned} \bar{u}_j &= \frac{1}{N} \sum_{i=1}^N u_{ij} = 0. \quad (j = 1, 2, \dots, n) \\ p &= \frac{1}{N} \sum_{i=1}^N (f_{ip} - \bar{f}_p)^2 = 1. \quad (p = 1, 2, \dots, m) \\ u_j &= \frac{1}{N} \sum_{i=1}^N (u_{ij} - \bar{u}_j)^2 = 1. \quad (j = 1, 2, \dots, n) \end{aligned}$$

そして共通・独自因子間や異なる独自因子間には相関が無く、直交しているとする。

$$\begin{aligned} \sum_{i=1}^N f_{ip} u_{ij} &= 0. \quad p = 1, 2, \dots, m \quad \& \quad j = 1, 2, \dots, n \\ \sum_{i=1}^N u_{ij} u_{ik} &= 0. \quad j \neq k \end{aligned}$$

共通因子間の関係では、その間に直交性を仮定する場合を直交因子モデル

$$\sum_{i=1}^N f_{ip} f_{iq} = 0.$$

直交性を仮定しない場合を斜交因子モデルと呼び、

$$\frac{1}{N} \sum_{i=1}^N f_{ip} f_{iq} = l_{pq}$$

は共通因子間の相関を表す。

主成分分析における主成分の個数と同様に、因子分析においても因子の個数は推定によって定める他はない。因子数を定めると初めの因子負荷行列は相関行列の固有ベクトルから求める事が出来る。しかし因子分析の場合には独自因子があるのでここで使われる相関行列は普通の相関行列から少しだけ異なって来る。独自因子の二乗を独自性と呼ぶが、この独自性を相関行列 R の対角成分から引いた行列

$$R_{ij}^{\dagger} = R_{ij} - d_i^2 \delta_{ij}$$

を固有ベクトル分解して最初の因子負荷行列を求める。

この様に観測データを再構成された変量の一次結合で表す手法は主成分分析法と似ていると言える。しかし似ているだけで基本的に異なる点が幾つかある。異なる点は独自因子の存在と分析結果を表す空間である。

主成分分析では独自因子というものはない。それに対して因子分析では各観測量に独自に変動する因子があると考えられる。この独自因子の大きさは推測するより他になく、いろいろな

推測方法が提案されている。

また主成分分析では一般に観測データを座標軸にした空間を考え、その中に再構成された主成分の軸を表示する。即ち主成分は n 個の観測変量 z_i ($i = 1, n$) の一次結合として表され、主成分が構成する空間は z_i の張る空間内に含まれていた。それに対して因子分析では全体として m 個の共通因子と n 個の独自因子の合計 ($m+n$) 個が独立な変数として ($m+n$) 次元空間を張っており、観測される n 個の変量はその空間内にある。そして因子軸から構成される空間である因子空間は必ずしも z_i の張る空間には含まれない [奥野 81, p. 358]

$A = (a_{ij})$ は行に変量を取り、列に因子を取った因子負荷行列である。そして因子負荷が因子の性質を変量との関係に於いて示しているという事から因子負荷行列を特に因子パターンと呼ぶ事がある。また各因子と各変量との相関の事を因子構造と呼ぶ。因子と変量がいずれも標準化されている事から因子構造は

$$s_{jp} = \frac{1}{N} \sum_{i=1}^N z_i f_{ip}$$

で表される。

直交因子モデルでは因子負荷・因子パターン・因子構造が全て同じ行列になるが、斜交因子モデルでは異なるのでこの区別が重要になって来る。テンソル解析に例えれば因子構造は共変成分を、因子パターンは反変成分を表すと考えれば良いであろう。なお、因子構造と因子パターンの積は

$$SA = R^t$$

の様に相関行列となる。

因子分析では因子空間内での基底即ち因子の取り方は任意であり、共通因子の意味付けが容易になる様に基底を定める事が出来る。その為の基準として因子分析では単純構造というものを考える。因子構造が単純構造であると判断する条件として例えば [Thurstone 47] は以下の様な条件を挙げている、と [芝 79, p. 93] にある。

1. 各行には少なくとも一つ 0 がなければならない。
2. 各列には少なくとも m 個 (m は共通因子の数) の 0 がなければならない。
3. 任意の 2 列をとった時、どちらの列でも 0 となる変量や、いずれか一方の列で 0 となる変量が多くなければならない。

因子を直交変換して単純構造を求める解析手法としてはバリマックス法が良く知られて居る

が、本稿では斜交変換の為のオブリミン法についてのみ述べる。オブリミン法では

$$K = \sum_{p=q}^m \left\{ \sum_{j=1}^n s_{jp}^2 s_{jq}^2 - n \left(\sum_{j=1}^n s_{jp}^2 \sum_{j=1}^n s_{jq}^2 \right) \right\}$$

で定義された K を最小にする解を求める。ここで $\alpha = 0.0$ ならコーティミン基準， $\alpha = 1.0$ ならコバリミン基準，そして $\alpha = 0.5$ としたものをバイコーティミン基準という。これらの基準によって斜交解を求めると，コーティミン基準では因子間の相関に高い値が出る傾向があり，コバリミン基準だと逆に直交性が高くなるという経験則がある。

ここで述べたオブリミン法では因子構造は単純化されるが，因子パターンは単純化されない。一般に因子構造を単純化するとパターンの方は単純化されないし，パターンの方を単純化すると構造の方は単純化されないのは明らかである。単純構造から単純パターンを求めるには相反系を用いる。その為には因子間相関行列 L から

$$D = \{ \text{diag}(L^{-1}) \}^{\frac{1}{2}}$$

で定義された対角行列 D を用いて因子パターン

$$W = SD^{-1}$$

を求めればよい。このときの因子構造は

$$V = PD$$

で与えられる。

これまでの説明から因子分析法で斜交座標を構成する手順を要約すると

1. 変数を標準化する。
2. 独自因子を推定して，それを差し引いた相関行列 R^+ を作る。
3. R^+ を固有ベクトル分解する。
4. 因子数を推定した結果に従って，分解で出来た固有ベクトルから因子負荷行列を作る。
5. 因子負荷行列から斜交変換により斜交因子モデルによる単純構造を求める。
6. 必要ならばその相反系としての単純パターンを求める。

となる。

3 因子分析法の衛星データへの適用

因子分析法はこれまで主に心理学等の分野で利用されて来た。上に述べた概念や手順はそこで使われている手法であり、衛星データに適用する場合にはそれに適した手順に変更する必要がある。

因子分析法では標準化された変数を使うのが普通であるが、これは各測定値の単位が異なっている場合等、測定値の絶対的な大きさや原点を比べる事に意味がない時によく使われる。しかし衛星データでは各測定量は衛星センサーの各バンドに対応しており、それらは各波長での地球からの反射光という同じ意味を持つ。これらをバンド毎に標準化するとそれぞれの画素が持つスペクトルの形という大切な情報が失われてしまう事になる。この様な場合にはデータの標準化は行うべきではない。

一方、この解析では因子分析の結果と PDM の基本スペクトルパターンを比較する事を目的とするのであるから、データの処理方法も出来るだけ PDM と同様に行うのがよい。PDM では、データは太陽光の光量を用いて反射係数に変換し、path radiance として Rayleigh scattering に相当する量を差し引いている。また PDM では、基本スペクトルを各画素毎にバンドの絶対値和で規格化する、即ち

$$\sum_{i=1}^n |A_i| = 1.$$

とする。よって今回の解析においても、データを反射係数に直して Rayleigh scattering 分を引き、各画素毎に絶対値和で規格化した。

独自因子は一意に定まっている量ではなく、推定する他に定め方はない。逆に言えば任意に決められるとも言え、考慮しなくても良いとも言える。今回は独自因子については考慮しないという事にした。

因子分析では因子負荷行列を求めるのに相関行列を用いる。しかし相関行列を用いる事は結局標準化された変数を使用する事と同じである。よって今回の解析では相関行列は使用しない。では分散共分散行列を用いれば良いのかというと、実はそうでもない。分散共分散行列の定義

$$V_{ij} = \frac{1}{N-1} \sum_{k=1, l=1}^N (x_{ik} - \bar{x}_i)(x_{jl} - \bar{x}_j)$$

を見るとデータからその平均値を引いている。勿論、分散を計算する為には当然の事ではあるが、こうする事によってデータの原点という情報が失われてしまう。PDM では原点を保存し

たままのデータの分布から基本スペクトルを定めている。よって今回の解析においても原点を保存する為に

$$U_{ij} = \frac{1}{N-1} \sum_{k=1, l=1}^N x_{ik}x_{jl}$$

という行列を定義し、この固有ベクトルを使って因子負荷を計算した。この際因子数については PDM と同じく 3 に固定している。

ここまでで因子負荷の初期値が得られたが、斜交座標では因子構造と因子パターンが異なるため、どちらを単純化すべきかが次の問題である。PDM では 6 次元のデータ空間に三つのデータ・クラスターが存在する事から、おのおののクラスターから一つずつデータを取りだしてそれをそのクラスターを代表するスペクトルパターンを決定するのに使用している。これはクラスターに属するデータ点の反変座標の一つが大きい値を、それ以外が小さい値を取る事を意味している。よって因子パターンが単純構造を持つ事になる。そこで今回の解析においても因子パターンを単純化する事にする。

最後に単純因子が構成されたとして、その構造かパターンかどちらを PDM の基本スペクトルパターンと比べるべきなのかを考える。基本スペクトルパターンは各バンドの変量が直交していると考えて構成された 6 次元空間の中で表現されている。即ちバンド変量とスペクトルの相関を使って基本パターンが表されている。そこで今回は得られた因子の構造を絶対値和で規格化したものを基本スペクトルパターンと比べる事にした。

以上、今回の衛星データの解析方法を要約すると

1. データを反射係数に直して Rayleigh scattering 分を引き、各画素毎に絶対値和で規格化する。
2. 独自因子は考慮しないで、 U_{ij} を計算する。
3. U_{ij} を固有ベクトル分解する。
4. 因子数を 3 として、分解で出来た固有ベクトルから因子負荷行列を作る。
5. 因子負荷行列から斜交変換により斜交因子モデルによる単純構造を求める。
6. その相反系としての単純因子パターンを求める。
7. その単純パターンに対応する因子構造を求める。
8. 求めた因子構造を絶対値和で規格化して PDM の基本スペクトルパターンと比較する。

となる。

4 解析結果について

今回の解析には LANDSAT-5/TM の 1992 年 3 月 13 日金沢市付近のデータを使用した。図 1 から判る様に、水・植生・土壌（市街地）を含む様に解析の領域を設定した。

そしてこのデータを反射係数に変換して Rayleigh scattering 分を引き、 U_{ij} を計算してその固有値、寄与率と累積寄与率を計算すると表 1、又グラフ化すると図 2 の様になった。これから見ると因子数は 2 で充分だとも考えられるが、前節にある様に PDM に合わせて今回は因子数を 3 とする。

因子数を 3 として因子負荷行列を作り、オブリミン法で斜交変換して因子を求める。この時因子パターンの符号には意味がないので、符号が全て負の場合には正に変換した。そして PDM の標準スペクトルパターンと比較する為にそれを絶対値和で正規化した。オブリミン法では斜交変換での座標軸間の相関を調節する というパラメータが存在する。この値を 0.0 から 1.0 迄変化させた場合の正規化された因子パターンを図示すると図 3 の様になる。比較する為に PDM の標準スペクトルパターンを図 4 に表す。図 3 より $\alpha = 1.0$ 付近で抽出された因子が、PDM の標準パターンをかなりよく再現している事が判る。

PDM の標準スペクトルパターンは、3 つのパターンが持つ 6 個の成分が全て正の符号を持って居る。これは標準パターンが地上被覆物の反射スペクトルから取られている事から当然の事であると言える。しかし因子パターンの符号については全ての成分の符号が同一になる必然



図 1 1992 年 3 月 13 日金沢市周辺の解析領域

表 1 固有値、寄与率と累積寄与率

	固有値	寄与率	累積寄与率
1	0.1733	0.9137	0.9137
2	0.01495	0.07880	0.9925
3	0.001147	0.006047	0.9985
4	1.457 E-4	7.683 E-4	0.9993
5	9.173 E-5	4.835 E-4	0.9998
6	4.028 E-5	2.123 E-4	1.0

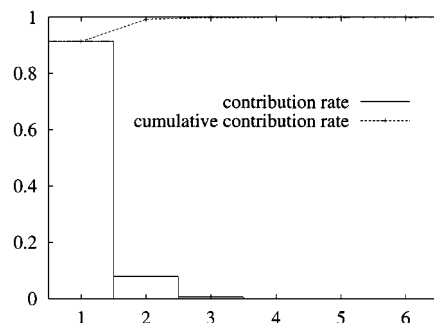


図 2 寄与率と累積寄与率

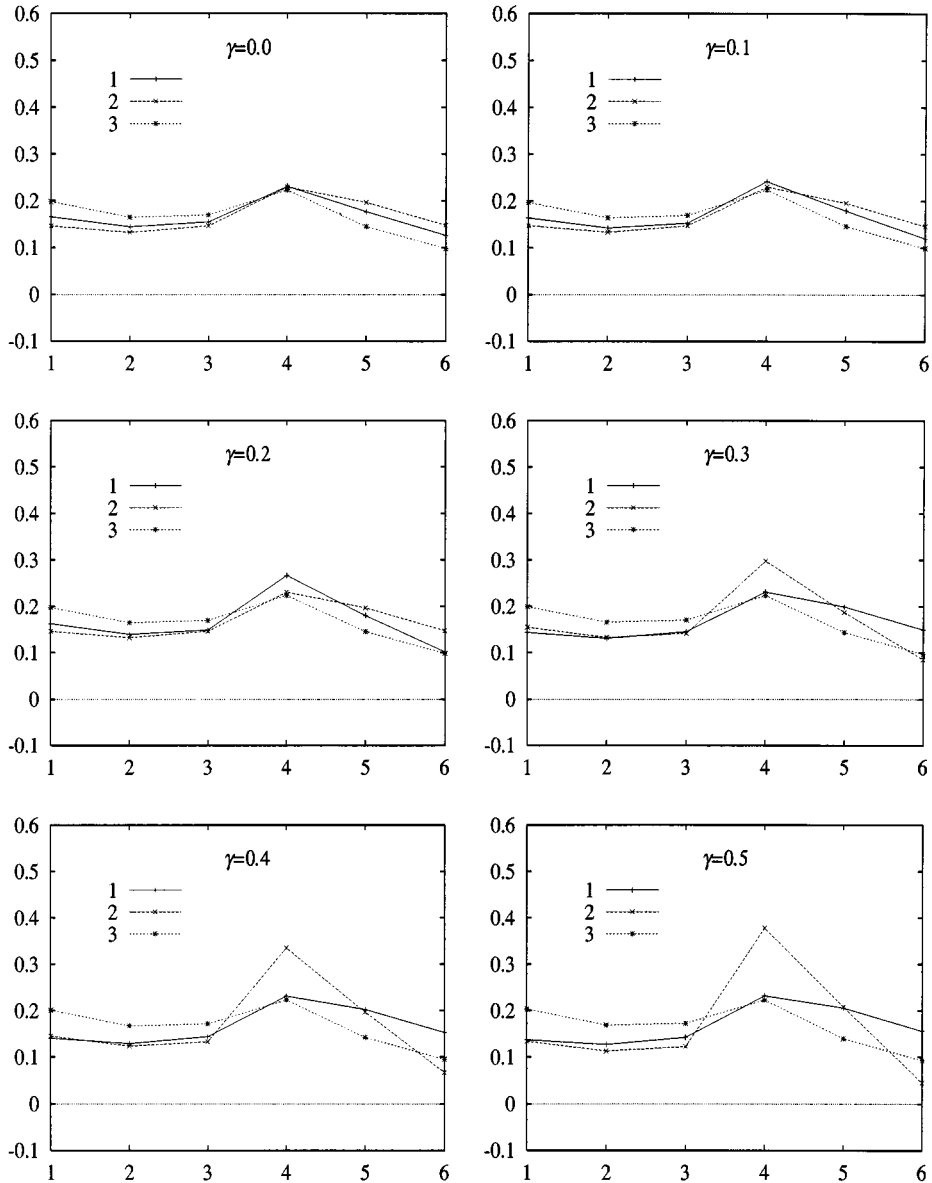


図 3 a U_{ij} を使ったオブリミン変換の結果

性はない。今回抽出した因子パターンを見ても γ が 0.7 から 0.9 の値を取る時に負の値を取る成分がそれぞれたった一つであるが存在している。しかし殆どどのパターンの成分は同じ符号を持つ。

それに対して分散共分散行列 V_{ij} を使って以後の手順を踏んだ場合には、結果は必ずしもそうはならない。 V_{ij} を使い、 γ の値を 0.1, 0.5 と 1.0 にした場合の因子パターンは図 5 の様に正の符号を持つ成分と負の符号を持つそれとがどちらも存在する状況になる。この事からも PDM の標準スペクトルパターンと同様の因子を抽出するには U_{ij} の利用が必須であると言う事が判る。

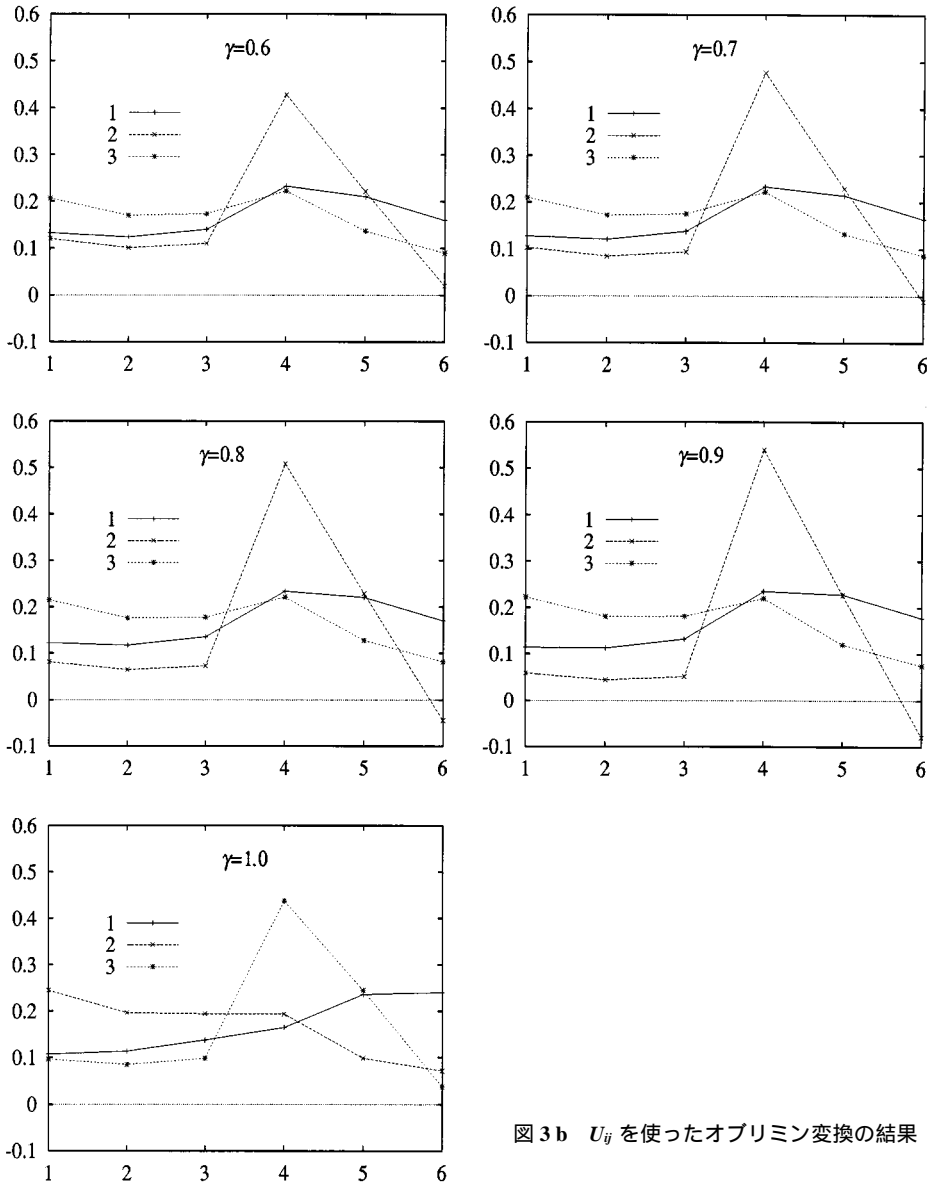


図3b U_{ij} を使ったオプリミン変換の結果

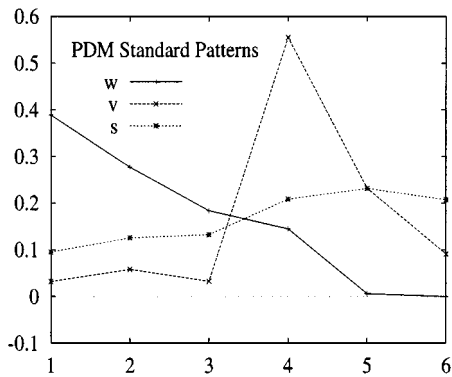


図4 PDMの基本スペクトルパターン

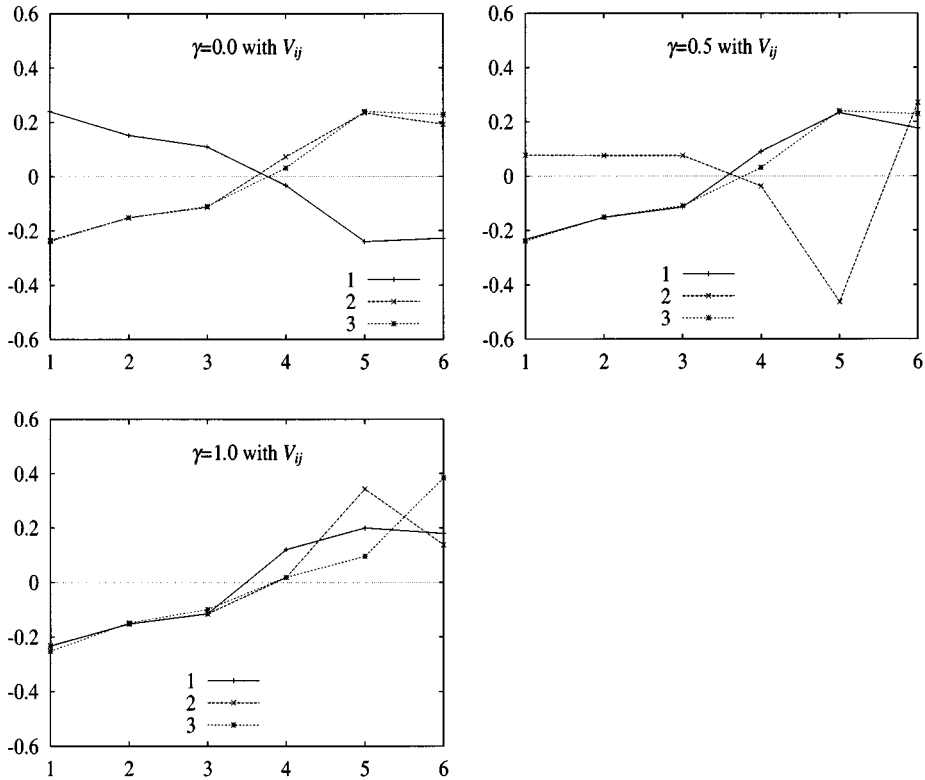


図5 V_{ij} を使ったオブリミン変換の結果

5 ま と め

因子分析法を衛星画像データに適用する事により PDM の標準スペクトルパターンに似た形の因子パターンを抽出する事が出来た。この様に6バンドの TM データで因子分析が巧く働く事が判った事から、次にはより多波長のデータに適用する事が考えられ、放射分光計による地上測定データや将来の ADEOS-II/GLI センサー等に適用してみる事を計画している。

今回の解析過程を見てみると、因子分析の手法の中で使用しているのは結局オブリミン法という斜交変換のみである。この事からも斜交変換の重要性を見る事が出来る。そしてこの斜交変換を衛星データ解析でなく他の分野に応用すれば、面白い結果が得られる可能性がある。主成分分析で得られた座標を更に理解しやすく変換する手法は行われている様であるが [奥野 81, p. 213], それを斜交変換を使って行った方がより意味づけしやすい座標が得られるのではないかと言う事である。

一方、因子分析法では因子を軸として構成される空間の中に変量とその相関に応じて分布している。これは因子分析法と言う手法が考え出された心理学分野での考え方から出て来ている。心理学的なテストでは、多少手を加えるだけで同じ共通因子によって説明され得る多数個

の等質なテストを作る事が出来、この事から分析の対象は無限とも言えるテスト群であって、測定された変量はたまたま抽出されたサンプルであると考えるのである [奥野 81, p. 355]。この状況は多バンドの衛星データの場合にも当てはまると言え、衛星のバンド間の相関を図示し、またそれらの関係を解析するのにこの手法が使えるのではないかとの期待を持っている。

謝辞

本研究で使用した LANDSAT-5/TM データは宇宙開発事業団より研究用として提供されたものである。

参考文献

- [Adams 95] Adams, J. B. et. al. , (1995) " Classification of Multispectral Images Based on Fractions of End-members : Application to Land-Cover Change in the Brazilian Amazon " , *Remote Sens. Environ.*, 52, pp. 137-154.
- [Gonzalez 77] Gonzalez, R. C., and Wintz, P. , (1977) *Digital Image Processing*, Mass. : Addison-Wesley.
- [Merembeck 80] Merembeck, B. F., and Turner, B. J. , (1980) " Directed Canonical Analysis and the Performance of Classifiers under its Associated Linear Transformation " , *IEEE Trans, Geoscience and Remote Sensing*, GE-18, pp. 190-196.
- [Muramatsu 00] Muramatsu, K., Furumi, S., Fujiwara, N., Hayashi, A., Daigo M., and Ochiai, F. , (2000) " Pattern decomposition method in the albedo space for Landsat TM and MSS data analysis " , *INT. J. Remote Sensing*, Vol. 21, No. 1, pp. 99-119.
- [Thurstone 47] Thurstone, L. L. , (1947) *Multiple factor analysis*, Univ. of Chicago Press.
- [奥野 81] 奥野 忠一 , 久米 均 , 芳賀敏郎 , 吉澤 正 , (1981) 『多変量解析法 改訂版』日科技連出版社 .
- [芝 79] 芝 祐順 , (1979) 『因子分析法 第 2 版』東京大学出版会 .
- [藤原 96] 藤原 昇 , 他 (1996) 「衛星データ解析のためのパターン展開法の開発」『日本リモートセンシング学会誌』第 16 巻第 3 号 , pp. 17-34.