独立成分分析

1．はじめに

気象学が扱う流体運動は，非線型であるために，理解することは簡単ではない．そこで，問題を簡単にするために，流れ場を背景とする場（時間平均した場であることが多い）と，それに重なる比較的小規模あるいは短周期で変動する場とに分割しようとする試みが行われてきた．北極振動のように比較的時間スケールが長い低周波振動現象についても，半ば期待をこめて，同様な分割が試みられている．

ところが，実際問題として，物理学的な実態を伴ったどのような現象が，どのような時空間分布を持ち，いくつ重ね合わさっているかについては，事前にわからない．そこで分離の方法が問題になってくる．よく用いられるのは，多変量解析の典型的な手法である主成分分析（PCA；Principal Component Analysis，気象学・気候学ではEOF解析；Empirical Orthogonal Function Analysisとして馴染みが深い）である．PCAは，多数の変数を組み合わせ，互いに無相関な変数を作り，その中から振幅の大きなものを抽出する手法である．これまで，低周波振動を定義する際に有効な方法であると考えられてきた．しかし，PCAを使うべきかどうか，また，PCAを使った結果が何を意味するのかは，慎重に考えるべきである．例えば，ホワイトノイズに対しても規模の大きな変動を拾ってしまう傾向があることは，結果を吟味する際に最初に考えなければならない点の一つである．

こうした手法に対する改善策は以前から検討され，実用化されてきた．しかし，1990年代後半に独立成分分析（ICA；Independent Component Analysis）が知られるようになって，状況は一変することになる．

2．独立成分分析誕生の経緯

ICAの歴史は，「混合信号の分離問題」という信号処理問題の発展としてとらえることができる．例えば，パーティ会場のようにさまざまな音が重なり合った喧騒の中を考えてみよう．そのような中でも，私たちは周囲の雑音に煩わされることなく会話を楽しむことができる．人はどのようにして同時に届く複数の音声を聞き分けることができるのであろうか．信号源に関する情報が乏しい状況下で，観測された信号から必要な源信号を分離・抽出する問題一般を，混合信号の分離問題という．とくに西洋ではカクテルパーティが一般的であることから（また，カクテルが混合を意味することもあって），このような問題を「カクテルパーティ問題」ともいっている．

気象学と同様に，混合信号の分離問題では，信号空間の計量に基づくPCAが用いられてきた．しかしこの方法による信号の分離は互いに無相関であるような信号を抽出するが，抽出された信号は必ずしも互いに統計的に独立な信号とは限らないことも弱点の一つとして指摘されてきた．ICAは，観測された複数の信号から，統計的に独立な信号成分を個々に抽出する技術として近年登場したのである．

3．独立成分分析の考え方

ICAは，情報理論では，「相互情報量」「情報エントロピー」「ネゲントロピー」などで説明されている．しかし，その基本的な考え方は難しくない．以下では，その基本的な考え方について紹介する．

3.1 モデル

はじめに問題を定式化しよう．n箇所（地点）の観測信号（時系列データ）を，n個の成分を持つ列ベクトル$\mathbf{x}(t)$で表す．同様に源信号をn個の成分を持つ列ベクトル$\mathbf{s}(t)$で表す．一般に，観測信号の数（あるいはセンサーの数）と源信号の数は一致しない．しかし，本稿では簡単のために，両者は一致し，共にnであるとする．通常，時刻について離散的なデータが得られるが，表記を簡単にするために時刻についてはtで表すこととした．観測信号$\mathbf{x}(t)$は源信号$\mathbf{s}(t)$の線型和になっていると仮定する．

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{1}$$

ここで $\mathbf{A}$ は正則（$\det\ \mathbf{A}\neq 0$）な $n\times n$ の正方行列である．問題は，観測信号 $\mathbf{x}(t)$ が与えられた時に，どのようにして源信号 $\mathbf{s}(t)$ を見出すか，ということになる．

まず，観測信号 $\mathbf{x}(t)$ の時間平均 $\overline{\mathbf{x}(t)}$ からの偏差を扱うようにする（これを「中心化」という）．(1)式について時間平均からの偏差を考えると，線型性から次のような関係が得られる．

$$\mathbf{x}(t)-\overline{\mathbf{x}(t)}=\mathbf{A}(\mathbf{s}(t)-\overline{\mathbf{s}(t)})$$

以下では，偏差 $\mathbf{x}(t)-\overline{\mathbf{x}(t)}$，$\mathbf{s}(t)-\overline{\mathbf{s}(t)}$ を改めて $\mathbf{x}(t)$，$\mathbf{s}(t)$ とし，これらについて考えることとする．

3.2 主成分分析

次に，主成分分析の方法を簡単に復習してみよう．観測信号 $\mathbf{x}(t)$ の（標本）分散共分散行列 $\overline{\mathbf{x}^t\mathbf{x}}$ は対称行列である．ここで，${}^t\mathbf{x}$ は $\mathbf{x}$ の転置を表すとする．対称行列の性質として，その固有ベクトルは互いに直交する．そこで，分散共分散行列は規格化された固有ベクトルを並べて作る（正規）直交行列 $\mathbf{V}$ によって次のように対角化される．

$$\mathbf{VD}=\overline{\mathbf{x}^t\mathbf{x}}\mathbf{V}$$

ここで，$\mathbf{V}$ が直交行列であることから，${}^t\mathbf{VV}=\mathbf{I}$ が成り立つ．そこで，左側から ${}^t\mathbf{V}$ をかけることにより，次式を得る．

$$\mathbf{D}={}^t\mathbf{V}\overline{\mathbf{x}^t\mathbf{x}}\mathbf{V}=\overline{{}^t\mathbf{V}\mathbf{x}^t({}^t\mathbf{V}\mathbf{x})}=\overline{\mathbf{x}_p{}^t(\mathbf{x}_p)}$$

ここで，$\mathbf{x}_p$ は $\mathbf{x}_p={}^t\mathbf{V}\mathbf{x}$ と定義される新たな時系列データで，その成分 $x_{pi}(t)$ は主成分と呼ばれる．$\mathbf{x}_p(t)$ の分散共分散行列は対角行列 $\mathbf{D}$ になるので，主成分 $x_{pi}(t)$ は互いに無相関となっている．また，主成分の各々の分散が対角行列 $\mathbf{D}$ の対角成分に現れることになる．

3.3 統計的独立性

独立成分分析の核心部分に入る前に，統計的な独立性についても復習しておこう．確率変数 x，y が統計的に独立であるとは，両者が同時に実現する確率密度分布 $f_{XY}(x,y)$ が，それぞれの確率密度関数 $f_X(x)$，$f_Y(y)$ の積で書けることである．

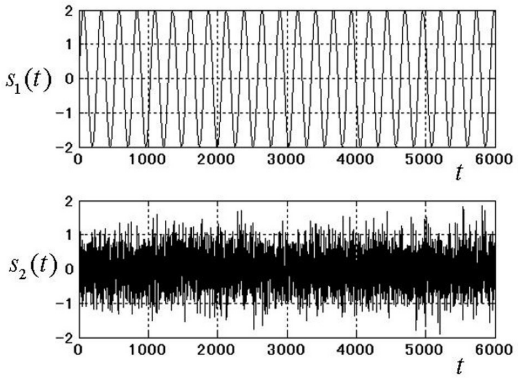
$$f_{XY}(x,y)=f_X(x)f_Y(y) \tag{2}$$

例えば，源信号 $\mathbf{s}(t)=(s_1(t),\ s_2(t))$ として $s_1(t)$ に正弦波，$s_2(t)$ に正規乱数（正規分布をする乱数）を考える．第1図に時系列を示し，第2図に散布図を示す．第2図を見ると，$s_1(t)$ の値が決まっても同時刻の $s_2(t)$ の値についての情報を得ることができず，逆に $s_2(t)$ の値が決まっても同時刻の $s_1(t)$ の値についての情報を得ることができないことがわかる．したがって，$s_1(t)$ と $s_2(t)$ とは統計的に独立である．

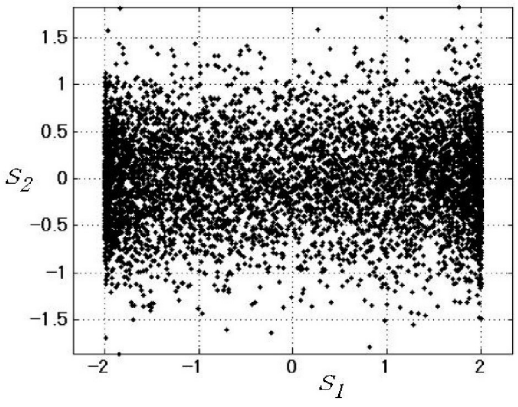
次に，これらのデータの線型和として人工的な観測データ $\mathbf{x}(t)$ を作成する．

$$\begin{pmatrix}x_1(t)\\x_2(t)\end{pmatrix}=\begin{pmatrix}3 & 9\\3 & -2\end{pmatrix}\begin{pmatrix}s_1(t)\\s_2(t)\end{pmatrix} \tag{3}$$

これに PCA を施した結果，得られた主成分 $x_{p1}(t)$，$x_{p2}(t)$ を第3図に示す．第3図では，$x_{p1}(t)$ と $x_{p2}(t)$ とは無相関ではあるが，独立ではないことに注意する．例えば，$x_{p2}(t)=0$ の場合，$x_{p1}(t)$ の値は $-12\leq x_{p1}\leq 12$



第1図 人工的な源信号（$s_1(t)$：正弦波，$s_2(t)$：正規乱数）．



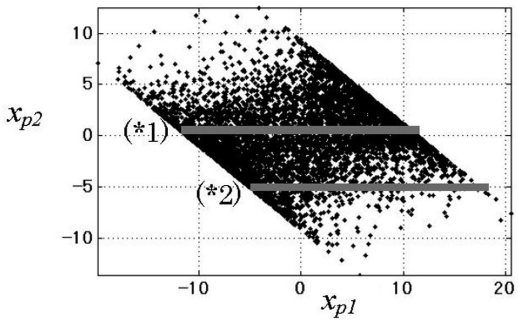
第2図 源信号 $s_1(t)$，$s_2(t)$ の散布図（同時確率密度分布）．

と限られてしまう（図中*1）．あるいは，$x_{p2}(t)=-5$の場合，$x_{p1}(t)$の値は$-6\leq x_{p1}\leq 18$となる（図中*2）．このように$x_{p2}(t)$の値によって$x_{p1}(t)$の確率密度が変化するため，PCAによって得られた両者は統計的に独立ではない．

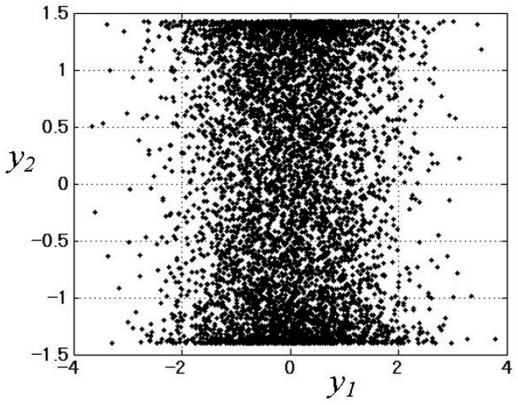
3.4 中心極限定理と独立成分分析

それでは，どのようにすれば統計的に独立であるような信号を分離できるだろうか．その鍵は中心極限定理にある．中心極限定理は，おおまかには，複数の信号の線型結合は元の信号よりも正規分布（ガウス分布）に近い，と表現される．これを手がかりに，正規分布からずれている信号を源信号とみなして分離を試みてみよう．

具体的な手続きとして，まず，主成分分析の結果を基にして，主成分を規格化し，分散を1にする（白色化という）．$\mathbf{x}_p$を白色化した$\mathbf{x}_n$は次のように定義される．



第3図　PCAによって得られた主成分$x_{p1}(t)$，$x_{p2}(t)$の散布図．



第4図　ICAによって推定された源信号（独立成分）の散布図．

$$\mathbf{x}_n \equiv \mathbf{D}^{1/2}\mathbf{x}_p \tag{4}$$

次に，$\mathbf{x}_n$を基にして統計的に独立な変数を得ることを念頭に，新たな変数$\mathbf{y}(t)$を次のように定義する．

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}_n(t) \tag{5}$$

ここで，$\mathbf{W}$を（正規）直交行列に限定しておけば，$\mathbf{y}(t)$の各成分は互いに無相関で分散が1のままになる．問題は，$\mathbf{y}(t)$の各成分が正規分布からできるだけずれるような直交行列$\mathbf{W}$を見つけることになる．

正規分布からの偏差を定量化するためには，変数$\mathbf{y}(t)$の成分$y_i(t)$について，ある関数$G(y_i(t))$の時間平均$\overline{G(y_i(t))}$を用いるとする．$G(x)$の選び方については後述の資料をご覧いただきたい．$G(x)=\log(\cosh(x))$，$G(x)=x^4$などがよく用いられる．そして，$\mathbf{W}$を変化させて，この値を正規分布をする確率変数μを用いた$\overline{G(\mu)}$から大きく外れる（極大値や極小値をとる）ように$\mathbf{W}$を決定する．こうして得られた$\mathbf{W}$を用いれば，統計的に独立であるような$\mathbf{y}(t)$が推定できるであろう．以上のような操作によって源信号を推定することがICAの基本的な方法である．

実際に，先に示した人工的な混合データについて，ICAによって推定された源信号の散布図を第4図に示す．標本数が有限であるという制限の下で，統計的に独立であるような信号が分離され，散布図は源信号の散布図（第2図）と明確に対応している．図には示さないが，$y_2(t)$を時系列データとして表示すると，正弦波に極めて近い．第2図と第4図を比較すると，符号，信号の順番が源信号とは異なることがわかる．これは，直交行列$\mathbf{W}$が座標軸の反転や座標軸の交換を許すことに対応している．また，振幅が異なるのは白色化されていることによる．

4．独立成分分析についてのその他の情報

ICAの総合的な解説は，Hyvärinen *et al*.（2005）や村田（2004）がある．また，本稿に関連して，正規分布からの偏差を評価することが統計的独立性の必要条件になることはMori *et al*.（2006）にコンパクトに記してある．インターネット上では，A. Hyvärinenによる Webサイト（http://www.cs.helsinki.fi/u/ahyvarin/whatisica.shtml，最終閲覧日：2009年12月25日）にICAのチュートリアルやいくつかの計算機言語によるソースコードが公開されている．

独立成分分析の工学的な応用例は，冒頭に述べた混合信号の分離問題の他に，雑音の除去，脳波解析など多岐にわたり，それぞれ成果をあげている．一方，地球物理学関連分野におけるICAの応用例は，意外にも，現状では多くない．今後の有効な活用が期待される．と，同時に，このことは，地球物理学関連分野のデータの複雑さを改めて強く示唆しているように思われてならない．

参 考 文 献

Hyvärinen, A., J. Karhunen and E. Oja（根本　幾，川勝真喜　訳），2005：詳解　独立成分分析―信号解析の新しい世界．東京電機大学出版局，532pp.

Mori, A., N. Kawasaki, K. Yamazaki, M. Honda and H. Nakamura, 2006：A reexamination of the Northern Hemisphere sea level pressure variability by the Independent Component Analysis. SOLA, 2, 5–8.

村田　昇，2004：入門　独立成分分析．東京電機大学出版局，246pp.

（桜美林大学自然科学　森　厚）
（筑波大学附属高校　川﨑宣昭）
（学芸大学自然科学　山崎謙介）