周期性のあるデータにおける独立成分分析の研究

M2008MM016 金森弘晃
指導教員: 松田眞一

1 はじめに

近年、独立成分分析 (Independent Component Analysis：ICA) の研究が盛んになってきている。独立成分分析とは複数の混合されたデータから統計的に独立な成分を導き出す手法である。この手法は特に音声分離や脳科学の分野で積極的に利用され、活発な議論がなされている。一方でデータ解析としての独立成分分析はまだ未熟である。独立成分分析のデータ解析への適用は金融データへの適用（Hyvärinen et al.[3]）やアクセスログデータへの適用（宮本ら [4]）があるが、周期性の取り扱いについては解析者の経験によるところが大きく、曖昧なままである。周期性が既知である場合、その情報を取り入れて解析した方がより多くの情報を得られる可能性がある。特に独立成分分析で扱うデータは時系列データであることが多い。データ解析の分野では時系列データであれば何らかの周期性が存在していることが多く、できればこの周期性を既知の成分とし解析に活かすべきである。本研究では周期性のあるデータにおいてどのようにデータ解析を進めるべきかを論ずる。

2 独立成分分析

いま、n 次元のデータが観測されたとする。1 つ目の次元データをスカラー量 $x_1(t)$、2 つ目を $x_2(t)$ のように表し、縦に n 個並べたベクトル $\boldsymbol{x}(t)$ を、

$$\boldsymbol{x}(t) = (x_1(t), x_2(t), ..., x_n(t))^T, (t = 1, 2, ..., l) \quad (1)$$

と書き表すことにする。ここで t は時刻を表し、それぞれ離散値をとるものとする。l はそのデータ数である。一方、未知の m 次元の統計的に独立な成分を

$$\boldsymbol{s}(t) = (s_1(t), s_2(t), ..., s_m(t))^T, (t = 1, 2, ..., l) \quad (2)$$

とする。そのときこの間には混合行列 A を用いて

$$\boldsymbol{x}(t) = A\boldsymbol{s}(t) \quad (3)$$

という関係があることを仮定する。つまり独立成分分析とは、観測されたデータ $\boldsymbol{x}(t)$ より未知の独立成分 $\boldsymbol{s}(t)$ と混合行列 A を求める問題になる。A の逆行列を W とすれば、その成分 w_{ij} を用いて

$$s_i(t) = \sum_j w_{ij} x_j(t) \quad (4)$$

となる。

2.1 制約条件

しかしながら独立成分 $\boldsymbol{s}(t)$ と混合行列 A の二つが未知であるため、通常この問題の解は無限に存在する。独立成分分析（Hyvärinen et al.[3], 甘利・村田 [1], 村田 [5], 甘利 [2] を参照）ではこの問題を解決するため、次のような条件を設定している。

1. 成分は互いに独立である。
2. 独立成分の分布は非正規分布に従う。
3. 独立成分データの次元は観測データの次元と同じか、それよりも小さくなくてはならない。
4. 混合行列は時間に関わらず不変と仮定する。
5. 混合行列のランク rank(A) はフルランクでなければならない。

2.2 導出過程

独立成分分析の手順は次のとおりである。

1. 観測データを中心化する
2. データを無相関化する
3. 各成分が互いに独立となる方向へ直交回転する

まずはじめに観測データの中心化と無相関化を行う。この無相関化には主成分分析や因子分析がよく使用されている。無相関化の最大のメリットは問題が簡単化されることである。また次元の多い場合には次元縮約として使用することもできる。無相関化を行った後は、成分を互いに独立な方向に回転する必要がある。成分の回転において、どの方向にどれだけ回転すればよいかという問題が存在する。独立成分は混合後よりも非正規的なため、何らかの非正規測度によって探索的に独立成分を探す方法がとられる。非正規性の基準としては、分布を用いるものとして Kullback-Leibler 情報量、最尤法、相互情報量、エントロピーなどがある。それ以外にはモーメント、キュムラント、特性関数、非線形相関など多数あるが、それぞれよい部分もあれば悪い部分も持ち合わせている。非正規性の評価基準を選択したら、あとは何らかの探索法によって回転方向を求める。探索法としては勾配法、不動点法、ヤコビ法が挙げられる。今回の解析では主に Hyvärinen et al.[3] の fastICA というアルゴリズムを用いる。この方法は評価関数にネゲントロピー、探索法に不動点法を利用した方法である。なおこのアルゴリズムは初期値に乱数を用いているため同じデータであっても解析ごとに結果が異なる。特に係数が少し異なるだけでなく全く意味づけの異なる成分が得られてしまうことがあり、本論文ではこの現象を独立成分の不安定性と呼ぶことにする。

3 使用するデータについて

3.1 気温データ

気象庁の公表する 2008 年の気温データを使用する。また 1 時間ごとの比較的細かいデータを収集した。観測地点は、札幌、東京、松本、名古屋、大阪、津山、福岡、那覇の 8 地点である。周期性という観点では、1 年を通した大きな周期性のほかに、1 日周期で変動する短期間の周期性が存在している。

3.2 シミュレーションデータ

シミュレーションでは気温データモデルを参考に、長期間周期性、短期間周期性、ノイズの組み合わせの 9 次元データを作成した。特に断らない限り設定値は以下のようにする。

1. サンプル数は N=5000 とする。
2. 長期間周期性の周期は 1666 回とする。
3. 短期間周期性の周期は気温データと同じ 24 回周期とする。
4. 長期間周期性（振幅）の大きさは 15,12,9 とする。
5. 短期間周期性（振幅）の大きさは 10,8,6 とする。
6. ノイズは平均 0、標準偏差 3 の正規ノイズとする。

4 解析結果

4.1 独立成分分析における寄与率

解析結果を見たときにどの成分が観測データにどれだけの影響を与えているのか一目で把握できる指標があれば便利である。実際に解析において寄与率を算出している例（宮本ら [4]）もあるようである。本研究では、解析によって得られた混合係数行列 $A_{ij}(i=1,...,n, j=1,...,m)$ を用いて、

$$P_j = \sum_{i=1}^{n} A_{ij}^2, (j=1,...,m) \tag{5}$$

を各独立成分の強さとする。これは独立成分の観測データにおける分散を計算している。独立成分分析によって得られた独立成分はいずれも分散が 1 に統一されているため、その混合係数の 2 乗和をとれば観測データにおける分散となる。この強さの総和に対する比を寄与率とする。

4.2 周期性の定量化

今回は自己相関係数を用いて短期間周期性の定量化を行う。自己相関係数関数を $f(x)$、短期間周期性の周期を T とすればその周期性 C を、

$$C = f(T) + f(2T) - f\left(\frac{T}{2}\right) - f\left(\frac{3T}{2}\right) \tag{6}$$

と定義する。相関係数は -1 から 1 の範囲の値をとるため、最終的な周期性の最高値は 4 となる。

4.3 気温データの独立成分分析結果

気温データに独立成分分析を行った結果の 1 つが表 1 である。得られた 8 つの独立成分のうち寄与率の高い成分のみを抜き出した。気温データには独立成分の不安定性が存在するため別の結果が出ることがあるが、何も前処理をしないとき順序は不定だが多くの場合この結果が出る。一番寄与率の大きい成分は第 6 独立成分である。その成分は大きく値が変動していることから 1 年周期の季節トレンドであることが分かる。次に大きな成分は第 5 独立成分である。表よりこの成分は 1 日の周期性が含まれていることがわかる。他の独立成分は特に大きな寄与率を持っていないが、3 つ目に大きい成分として札幌と松本の係数が大きい第 1 独立成分が現れることが多い。

	寄与率	周期性	名古屋	東京	大阪
1	0.027	0.003	0.739	0.886	0.929
5	0.096	0.901	2.881	2.297	2.395
6	0.837	0.021	8.209	7.368	8.023

福岡	札幌	那覇	松本	津山
0.443	3.319	−0.557	1.457	0.598
2.140	1.584	0.225	4.071	3.743
7.840	8.743	4.837	8.678	8.504

表 1 気温データの独立成分分析結果

4.4 解明したい問題点

本論文において解明したい点は以下の通りである。

1. 周期性を解析前に除いたらどうなるのか
2. なぜ解析ごとに結果が大きく異なる不安定性が存在するのか
3. 独立成分数は指定すべきか
4. ノイズの影響はどの程度か
5. 得られた独立成分の分布に特徴はあるか
6. 外れ値のある場合にはどういった解析が有効か

次章からは以上についてより深く取り扱う。

5 既知の周期性を除く方法

周期性の周期が既知の場合、その周期性をあらかじめ除いて解析することが考えられる。

5.1 トレンド成分を除く

気温のデータやシミュレーションデータでは一番大きな独立成分としてトレンド成分が出る。気温データの場合それは 1 年周期の季節性の周期であり、この季節効果を除きたい。平年値を引いた場合は強い独立成分の不安定性があるため、結果が毎回変わるがそのうちの代表的な一つを示す。表 2 がその結果である。1 番大きな成分は第 8 独立成分である。この成分は周期性の値が高く、1 日内の周期性を含んでいることが分かる。また周期性の値に関しては 2 倍近くに増えており、平年値を除くことで純度の高い周期性を検出できている可能性が高い。2 番目に大きな成分は第 3 独立成分であり、これは札幌のみ異符号の成分で平年値を除かない場合の 3 番目に近い。3 番目に大きな成分は札幌と福岡の値が大きい成分が現れているが、この成分は現れないこともある。なお得られた独立成分をプロットしてみると長期的に変化している成分が存在しないことから、季節トレンド成分は正常に除去できていることが分かる。

5.2 特定の周期性を除く

同様に 1 日内の短期間周期性を除くことも可能である。気温データの場合には 24 の時間帯ごとに平均をとりそれを除けばよい。その他に、時期によって周期性の振幅が異なる場合には期間ごとの平均をとってその相対平均を除くことができる。

	寄与率	周期性	名古屋	東京	大阪
1	0.086	−0.062	−0.602	−0.352	−0.803
3	0.105	0.088	−1.203	−0.822	−0.909
8	0.645	1.963	2.879	2.328	2.466

福岡	札幌	那覇	松本	津山
−1.764	−1.720	−0.820	−0.112	−0.541
−1.066	1.606	−1.244	−0.547	−1.216
1.955	2.586	0.817	4.166	3.588

表 2 平年値を引いた気温データの独立成分分析結果

5.3 外れ値のある場合

外れ値のあるデータにおいてはデータ全体の対数をとる方法が有効である。また、より独自性の強い高域成分を除くために低域通過フィルタを併用することもできる。

6 独立成分の不安定性

独立成分分析は探索法によって独立となる成分を探し出すため、全く同一のデータだとしても解析ごとに毎回結果が異なる。大体の場合は係数が多少違うだけで済むが、何らかの条件で結果が大きく変わってしまう現象が見られることがある。今回この不安定性が優ガウス的ノイズや、周期性の不安定さから起こることがわかった。この章ではノイズのあるシミュレーションにおける不安定性の再現と、周期性の不安定さによって起こる場合にクラスタ平均を除去する方法を提案する。

6.1 ノイズによる不安定性

周期性の上にノイズとして μ が 0 で ϕ が 2.5 のラプラス分布を混合した。表 3 はこのルールで作成された同一のデータに独立成分分析を 5 回連続して行った寄与率の結果である。表を見ると、大きな成分が 1 つになったり 2 つになったり毎回不安定な結果であることが分かる。特に 2、3、4 回目には寄与率の大きな成分が 1 つしか無くなっており、周期成分まで影響のある不安定性が起こった。この現象はロジスティック分布においても同様に起こり、優ガウス的ノイズが独立成分の不安定性の直接原因になっていることが分かった。またノイズとしてではなく、共通の成分としてラプラス分布をのせた場合にも不安定性が存在することがあり、そもそも独立成分分析に優ガウス分布は適していない可能性がある。

6.2 クラスタ平均の除去

気温データの 1 日内周期性は日ごとに不安定であり、それが独立成分の不安定要因となっている可能性がある。5.2 節では期間ごとの周期性を計算する方法に触れたが、気温データにおいて本当に分類すべきものは日照の有無ではないかと考えた。しかし日照の有無は 1 日ごとに変化するため、飛び離れた日付ごとに平均をとる方法を新たに考案する必要がある。そこで周期性の除去時にクラスター分析の導入を考えた。クラスター分析とはいくつかの対象を似たもの同士のグループに分類する方法である。手順は以下のとおりである。

	1 回目	2 回目	3 回目	4 回目	5 回目
1	0.636	0.031	0.047	0.039	0.011
2	0.012	0.824	0.028	0.011	0.011
3	0.011	0.011	0.014	0.015	0.012
4	0.011	0.032	0.011	0.013	0.011
5	0.011	0.011	0.028	0.034	0.635
6	0.011	0.020	0.813	0.013	0.283
7	0.282	0.016	0.018	0.018	0.011
8	0.011	0.033	0.012	0.056	0.011
9	0.011	0.017	0.025	0.796	0.011

表 3 ラプラスノイズを乗せたシミュレーション

1. ある 1 つの観測地点の長い時系列データをデータ幅（気温データの場合は 24 回）で区切る
2. 区切ったデータごとにそれぞれ中心化する（分散の標準化は行わない）
3. クラスター分析の方法を選び、実行する
4. 得られたクラスターごとに平均を算出する
5. 元データより対応するクラスタ平均を除く
6. 他の観測地に対しても同様の操作を行う

手順 3 のクラスター分析法であるが、McQuitty 法が一番独立成分の不安定性に強いことが分かった。次に McQuitty 法でクラスター数を 3 とした結果を載せる。

	寄与率	周期性	名古屋	東京	大阪
5	0.894	−0.017	−8.119	−7.227	−7.908
8	0.043	−0.131	−1.317	−1.509	−1.685

福岡	札幌	那覇	松本	津山
−7.619	−8.869	−4.761	−8.674	−8.601
−1.066	−2.989	0.652	−2.124	−1.328

表 4 クラスタ平均を除いた結果 (McQuitty 法)

結果を見ると周期性の値が大きい成分が無く、周期性については正常に除去できているようである。第 5 独立成分にはトレンド成分が現れている。また第 8 独立成分に札幌と松本の値が大きな成分が現れており、その他の成分への影響もそこまで大きくはないようである。焦点となっている不安定性であるが、このクラスタ平均除去後のデータは第 5 独立成分の分離を起こすことは全く無く、安定した解析結果が得られている。気温データにおいては McQuitty 法が一番結果が安定しているが、その他の方法もデータによってはかなりうまくいく可能性がある。気温データにおいて安定性の高かったものはそのほかにウォード法、メディアン法、K 平均法が挙げられる。注意すべき点は 2 日周期や 1 週間周期の成分が存在した場合、その成分がクラスター数の増加につれて除去されてしまう可能性があることである。実際にシミュレーションデータにおいてクラスター数が多い場合にその他の成分が減衰しているのを確認している。使用の際にはクラスター数を出来る限り低く設定することが大切である。

	2 成分	3 成分	4 成分	6 成分	8 成分
1	−1.177	−1.230	−1.207	−1.224	−1.207
2	−1.177	−1.230	−1.207	−1.209	−1.205
3	−1.177	−1.230	−1.206	−1.207	−1.177
4	−1.177	−1.230	−1.208	−1.230	−1.205
5	−1.177	−1.230	−1.205	−1.208	−1.206

表 5 気温データにおける最大成分の尖度

7 独立成分数

実際のデータにおいて独立成分数は未知であることが多い。独立成分を指定して解析を行った場合、結果にどのような影響があるのだろうか。

7.1 気温データにおける成分数指定

気温データにおいて成分数を指定して解析したところ、4 から 8 成分ではいずれも独立成分の不安定性があることが分かった。3 成分以下では不安定性が無いが、2 成分にしてしまうとそもそも長期間周期性と短期間周期性が別の成分に分かれない結果となった。このことから一番安定して結果が得られたのは 3 成分であり、もし成分数を指定するならば 3 が良いのではないかという結論に至った。

7.2 尖度による成分数推定

ところでそもそも独立成分の推定に独立性の評価基準を使っていることから、独立成分数の推定にもその評価基準（例えば尖度）を参考にすることを考える。まず気温データにおける成分数とその得られた成分の尖度について確認する。表 5 は気温データを独立成分分析にかけて得られた最も寄与率の高い成分（多くの場合これは長期間周期性である）の尖度である。各成分数につき 5 回の実験を行った。まず尖度が最も 0 から離れている成分数は、7.1 節で議論した成分数と同じ 3 成分である。3 成分よりも成分数を増やすと、尖度の値は少しだけ 0 に近づいている。逆に 2 成分とすると 2 種類の周期性が分かれず、その結果尖度は 0 に近づいている。このように気温データにおいては、尖度がある程度妥当な成分数の評価基準として使えそうである。

8 独立成分の分布

解析で得られた独立成分はどのような分布に従うのだろうか。そこで気温データにおいて得られた独立成分がどのような分布に従うのかを調べた。今回、ガンマ分布、正規分布、t 分布、ロジスティック分布、ラプラス分布についてそれぞれ適合度検定を行った。その結果、ほとんどの分布はすべての成分について棄却されたが、ロジスティック分布のみはたまに 1 つの成分が棄却されないことがあった。そこで何度か fastICA を実行し、ロジスティック分布（尖度 1.2）との適合度検定を繰り返した。10 回ほど繰り返したところ、毎回必ずある 1 つの成分の p 値が 0.01 前後であることが分かった。表 6 は気温データにおける高次統計量をまとめたものである。ロジスティック分布との適合度検定で最も p 値が大きかった第 5 独立成分の尖度は 1.051 と確かに近い値を示している。全体的に見ると、観測データは全ての尖度が負で劣ガウス的分布だったのに対し、解析後の独立成分はその尖度が負の成分（周期性）は一つに集約され、他の成分が全て優ガウス的な分布になっている。気温データにおける寄与の小さい成分は正規分布には従っていないことが分かった。なお正弦波の尖度は −1.5 で、二山の分布になる。

元データ	得られた独立成分				
尖度		寄与率	周期性	歪度	尖度
−1.003	1	0.004	0.085	−0.009	2.324
−1.008	2	0.840	0.025	−0.118	−1.205
−1.090	3	0.009	−0.133	0.309	2.267
−1.097	4	0.011	0.172	0.383	0.767
−1.132	5	0.007	−0.000	−0.203	1.051
−1.105	6	0.026	0.009	0.311	0.663
−0.933	7	0.008	−0.048	−0.709	1.997
−1.123	8	0.094	0.890	0.201	0.585

表 6 気温データの高次統計量

9 まとめ

本研究での主な研究成果としては以下のとおりである。

1. 周期性を事前に除去する方法を提案した。またその除去により推定精度を上げることができた。
2. 独立成分の不安定性の原因を調べた。優ガウス的ノイズと混合行列の非定常性が原因として挙げられる。
3. クラスタ平均除去によって不安定性を取り除いた。
4. 独立成分数の評価基準として尖度を提案した。
5. 気温データにおける独立成分の分布は優ガウス的で少し歪んでいる。

10 おわりに

今回は数種類のデータのみしか用いることができなかったが、実用のためにはより多種のデータを試す必要があるだろう。しかし今回の研究がデータ解析における独立成分分析の利用を促すものになることを期待している。

参考文献

[1] 甘利俊一, 村田昇：独立成分分析、多変量データ解析の新しい方法, サイエンス社, 2002.

[2] 甘利俊一：統計科学のフロンティア５、 多変量解析の展開、 独立成分分析とその周辺, 岩波書店, 2002.

[3] Aapo Hyvärinen, Juha Karhunen, Erkki Oja 著, 根本幾, 川勝真喜 訳：【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, 2005.

[4] 宮本友介, 清水昌平, 西川康子, 狩野裕：Analysis of Web access data with ICA, 日本行動計量学会大会発表論文抄録集 30 pp.208-211, 2002.

[5] 村田昇：【入門】独立成分分析, 東京電機大学出版局, 2004.