周期性のあるデータを中心とした独立成分分析の考察

金 森 弘 晃* 松 田 眞 一†
E-Mail: matsu@nanzan-u.ac.jp

本論文では周期性のあるデータを中心として独立成分分析のデータ解析法について論ずる。まず独立成分分析における前処理法を提案し，その実用性について気温データとシミュレーションデータを使い確かめた。さらに独立成分分析における不安定性などの問題点を列挙し，それらについて構造の考察やの対策法の提案を行った。

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（堤ら [12]）や脳科学（Calhoun et al.[4]）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.[6]）やアクセスログデータへの適用（宮本ら [9]）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。また，独立成分分析にはまだ解決していない多くの問題があり，本論文では不安定性と名付けた現象の解明や独立成分数の推定問題を取り上げる。さらに，独立成分の分布について考察する。

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，

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

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

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

とする。そのときこの間には $n \times m$ 混合行列 A を用いて

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

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

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

*南山大学大学院数理情報研究科数理情報専攻
†南山大学情報理工学部情報システム数理学科

となる。
（Hyvärinen et al.[6]，村田 [10] 参照）

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 2 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。（Hyvärinen et al.[6], 甘利・村田 [1], 村田 [10], 甘利 [2] を参照）

1. 成分は互いに独立である。
2. 独立成分の分布は基本的に非正規分布に従う。
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 $\mathrm{rank}(A)$ 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（Attias[3]，Kano et al.[7] 参照）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，ネゲ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.[6] 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。この解決法としては初期値に単位行列を用いて便宜的に解析結果を固定

している例や，初期値に JADE（Cardoso et al.[5]）の結果を利用している例（堤ら [12]）がある。しかし本論文ではそれらの解決法は使わず，fastICA のデフォルト通り初期値として乱数を用いることとする。そのため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 [8] 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。観測地選択基準は都会である 4 都市（東京，名古屋，大阪，福岡）と南北の都市（札幌，那覇），そして盆地（松本，津山）である。4 都市に比べ，南北の緯度差や，1 日の気温差の激しい盆地の特性がどのように現れるかが焦点となる。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。厳密には周期性として正弦波を利用した。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 の 3 段階とする。
5. 短期間周期性（振幅）の大きさは 10,8,6 の 3 段階とする。
6. ノイズ $\boldsymbol{n}(t), (t = 1, 2, ..., N)$ は基本的に平均 0，標準偏差 3 の正規ノイズとする。すなわち，モデル式は $\boldsymbol{x}(t) = A\boldsymbol{s}(t) + \boldsymbol{n}(t)$ となる。

周期性（正弦波）の分散は a を大きさとすると $a^2/2$ である。つまり大きさを 10 とするとその分散は $10^2/2 = 50$ となる。ノイズは大きさを変化させたものや正規ノイズ以外のものも用いるが，その際はその都度説明する。また，この他に各データ共通の独立成分として一様乱数成分（$U(0, 20)$，分散は 33.333…）を追加する場合もある。

なお，結果に使用する略号の意味は，s（短期間）と l（長期間）の b（大）, m（中）, s（小）である。したがって，sblb の場合は s（短期間）b（大）l（長期間）b（大）となり，短期間周期性も長期間周期性も一番大きい混合データである。また，sslm の場合は s（短期間）s（小）l（長期間）m（中）となり短期間周期性が小さく，長期間周期性が中の混合をした混合データである。

4　新しい指標

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 [9]）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ことになる。（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

解析で得られた独立成分に周期性が含まれているかどうかを確認するため周期性の定量化を行う。特に知りたい周期性は短期間周期の周期性である。長期間の周期を持つ周期性は解析結果のプロットを見れば一目瞭然であるが，短期間周期は見ただけでは分かりづらい。今回は周期性の判定に自己相関係数を使う。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。これはちょうど 2 周期内の自己相関係数の極小値と極大値の差をとっている。2 周期分の極値をとった理由は，1 周期目の極値が存在しないもしくは小さいことがあるからである。本来ならば 1 周期目と 3 周期目など間隔を空けたほうが良いのかもしれないが，今回の研究の主題ではないためそこまで深く追求しないこととする。また注意する点は，これは周期性の量ではなく周期性がはっきりしているかを定量化していることである。いくら周期性のすべてを含む独立成分であっても，過剰なノイズが乗れば周期性の値は低くなることを念頭におく必要がある。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。

5　解析例と問題点

5.1　気温データの独立成分分析結果

気温のデータに独立成分分析を行った結果の一例が表 1 である。得られた 8 つの独立成分数のうち寄与率の高い成分のみを抜き出した。気温の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には多くの場合この結果が出る。ただし，独立成分の順序は常に変化する。

1 番目に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。2 番目に大きな成分は第 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 番目に大きい成分として第 1 独立成分が現れることが多い。この成分は札幌と松本の係数が大きい。

表 1: 気温データの独立成分分析結果

	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

5.2 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
   時系列データには何らかの周期性がある場合が多い。もしもその周期性の周期が分かっていた場合，事前に周期性を除いたら解析結果はどのように変わるのだろうか。

2. なぜ解析ごとに結果が大きく異なるのか
   気温データでは全く同一のデータを解析しても結果が毎回変わることがある。ただ単に係数が少し異なる程度ではなく，解析結果の成分自体が複数に分離してしまうようなことがあり，意味付けさえも異なってしまう。

3. 独立成分数
   一体いくつの独立成分に注目して結果を見ればよいのかが現時点では曖昧である。fastICA には成分数を事前に指定して解析することができるが，指定した場合解析結果への影響はどの程度あるのだろうか。

4. ノイズの影響
   独立成分分析では基本的にノイズの影響を考慮していない。ただし現実のデータには少なからず何らかのノイズがある場合が普通である。ノイズの影響がどの程度あるのだろうか。

5. 得られた独立成分の分布に特徴はあるか
   独立成分は非正規性が最大となるように導出される。では実際に解析で得られた独立成分がどのような分布に従っているのだろうか。

6. 外れ値のある場合
   アクセスログデータのように外れ値のある場合には極端な混合行列が得られる。その場合どのような解析が効果的か。

1. は第 6 章で，2. は第 8 章で，3. は第 9 章で，4. は第 9,10 章で，5. は第 10 章で，6. は第 7 章で扱う。

6　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。

6.1　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 1 番目に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独立成分の強い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例を表 2 で示す。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。

表 2: 平年値を引いた気温データの独立成分分析結果

	寄与率	周期性	名古屋	東京	大阪	福岡	札幌	那覇	松本	津山
1	0.086	−0.062	−0.602	−0.352	−0.803	−1.764	−1.720	−0.820	−0.112	−0.541
3	0.105	0.088	−1.203	−0.822	−0.909	−1.066	1.606	−1.244	−0.547	−1.216
8	0.645	1.963	2.879	2.328	2.466	1.955	2.586	0.817	4.166	3.588

6.2　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。ただし，気温データにおいて周期性を除くことは同時に不安定性を増加させる結果になっている。この特定の周期を除く方法では 2 パターンの異なる結果が得られ，気温データではうまくいかなかった。この不安定性への対処法は 8.2 節で紹介する。

6.3　周期性除去の有効性

周期性除去の有効性を確認するため，標準偏差 10 の正規ノイズを用いたノイズの多いシミュレーションデータを作成した。真の成分数は 3 で，2 つの周期性の他に共通の一様成分 ($U(0, 20)$) をのせた。通常の独立成分分析ではノイズが大きすぎて長周期成分しか結果に現れない。そこで既知のトレンド成分と短期間周期性の両方をデータより除く。その結果，0.322 と寄与率は低いが一様成分が現れた。これまでの方法ではこのようなノイズの多いデータに対処できなかったが，周期性を除くことで対処することができる。

7　その他の前処理法

7.1　外れ値のある場合

値が大きい方に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（Hyvärinen et al.[6] 参照）を併用することもできる。

7.2　次元が少ない時の解析法

データ解析では，収集したデータの次元が極端に少ない場合も考えられる。そのような場合データ次元が真の独立成分の次元よりも少なくなり，正常な解析が行えないことも考えられる。そこで周期性を別次元に分ける方法を提案する。もしもデータに午前と午後の周期性があった場合，1 つの次元を午前データと午後データの 2 つの次元に分ける。これによって便宜上は 2 倍の次元数が稼げるようになる。ただしこの方法は場合によっては不安定性を引き起こすこともあり，解析には注意が必要である。

8　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，初期値の設定が一定でない場合は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。この不安定性と名付けた現象であるが，全くランダムな結果が出るわけではなく，通常は結果が数種類のパターンに分けられる。実際に表 2 の平年値を除いた場合の結果では大きく分けて 3 種類のパターンが見られた。さらにこれらの結果パターンはいくつかの成分が結合と分離を起こして現れているようである。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

表 3: シミュレーションにおける不安定性（ラプラスノイズ）

	寄与率 1 回目	寄与率 2 回目	寄与率 3 回目	寄与率 4 回目	寄与率 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

レーションにおける不安定性を再現し，不安定性が周期性の不安定さによって起こる場合に安定化のためのクラスタ平均を除去する方法を提案する。

8.1 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。パターンとしては，1 回目と 5 回目の周期性が分かれるパターン，そして 2 回目〜4 回目の 1 つだけ大きな成分が現れるパターンの 2 タイプに分けられる。また後者のタイプでは，表 4 のように周期性の値が複数の成分に分離している。これは気温データにおいても周期性が 5 つほどの成分に分離するパターンが見られており，現実データでの解析と非常に良く似た解析結果が得られた。この現象はノイズがロジスティック分布の場合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。一方，正規ノイズや劣ガウス分布ノイズでも不安定性は起こりうるが，優ガウス分布のように比較的寄与率の大きな成分まで変化するような不安定性は起こりにくい。

8.2 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。6.2 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

表 4: 不安定性による周期性の分離（ラプラスノイズ）

	寄与率	周期性
1	0.015	0.027
2	0.816	0.453
3	0.012	0.037
4	0.016	0.109
5	0.013	0.010
6	0.025	0.336
7	0.023	0.460
8	0.031	0.305
9	0.050	1.020

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. 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 として解析した結果を載せる。

表 5: クラスタ平均を除いた結果 (McQuitty 法)

	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

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 0.9 前後の寄与率を持つ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。ところでこの McQuitty 法とはあまりなじみが無いが次のとおりである。（Pedersen et al.[11] 参照）
まず何らかの方法で全ての初期クラスター間の距離を算出する。（今回はユークリッド距離を用いた。）ここで 2 つのクラスター C_k と C_l 間の距離を $D(C_k, C_l)$ と定義する。次にその距離行列の中で 2 つのクラスター間距離 $D(C_k, C_l)$ が最小となる組み合わせ (k, l) を探す。そしてその 2 つのクラスターを結合し，新たなクラスターを C_{kl} と表す。すると距離行列は新たなクラスター C_{kl} とその他のクラスター間の値が更新される。新たなクラスター C_{kl} とそれ以外の任意のクラスター C_i 間の距離を，

$$D(C_{kl}, C_i) = \frac{D(C_k, C_i) + D(C_l, C_i)}{2} \tag{7}$$

と計算する。これは単純に更新前のクラスター間の距離の平均をとっていることになる。除去結果において McQuitty 法が安定した解析結果を得る理由はその性質にある。データに依存するため一概には言えないが，この方法はクラスター数が 3 の場合には主クラスターとなる群が 2 つと外れ値のような少数の群が 1 つできることが多い。つまり不安定な要因となる得る時系列を少数群として個別で処理できるため，その他のデータを晴れの気温変化大群と曇りの気温変化小群に分けることができる。またクラスター数を 2 としたとき，この方法は外れ値を省けないため全くうまくいかない。クラスター数が 4 以上の時には外れ値少数群が優先的に増え，外れ値のパターンが多い場合に対処できる。よって，データによってはクラスター数を増やす必要があるだろう。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。

8.3 クラスタ平均の除去による他の成分への影響

クラスタ平均除去において注意すべき点はクラスター数の増加につれて他の成分に影響を与える可能性があることである。そこで 2 つの周期性と共通の一様成分を加え，正規ノイズを混合したシミュレーションデータにおいて実験を行った。表 6 と表 7 は McQuitty 法においてクラスター数を 3 と 20 とした結果である。いずれも寄与率の大きいほうが長期間周期性，もう一方は共通の一様成分である。短期間周期性はクラスタ平均の除去によって同時に除去されている。結果を見ると，クラスター数 20 としたほうでは一様成分の寄与率が低くなっていることが分かる。このことからクラスタ平均を除くことは，同時に特徴的な他の成分も除いていることが分かった。その影響はクラスター数を増やせば増やすほど大きく，解析ではやはりできるだけ小さなクラスター数を選択したほうが良いだろう。確かに独立成分の不安定性に対する方法としては有効であったが，使用時には出来る限り他の成分への影響を十分に考慮した上で使う必要がある。またデータによっては別のクラスター法の方が良い可能性もある。McQuitty 法は外れ値のみの小さなクラスターを作りやすい方法である。これは不安定性を除く際に有利となる反面、特徴的な成分を減衰させてしまう可能性が特に高い。

表 6: クラスタ平均を除く（クラスター数 3）

	寄与率	周期性	sblb	sblm	sbls
2	0.219	0.016	4.27	4.87	5.96
6	0.706	−0.015	−11.16	−8.75	−6.33

smlb	smlm	smls	sslb	sslm	ssls
4.28	4.76	5.36	4.54	5.03	5.58
−11.15	−8.83	−6.38	−11.09	−8.77	−6.22

表 7: クラスタ平均を除く（クラスター数 20）

	寄与率	周期性	sblb	sblm	sbls
5	0.830	−0.016	−11.40	−9.16	−6.82
7	0.068	−0.026	−1.89	−3.28	−3.32

smlb	smlm	smls	sslb	sslm	ssls
−11.47	−9.17	−6.95	−11.54	−9.25	−6.94
−1.07	−2.14	−2.89	−1.48	−2.15	−4.31

9　独立成分数

独立成分分析では観測成分と独立成分の数が等しいと仮定して解析することになる。独立成分が観測成分よりも少ない場合何らかの次元縮約法によって成分数を合わせるか，直交化法によって逐次的に独立成分を抽出することになる。ところでその独立成分数であるが，通常未知であることが多い。特にデータ解析においては明確な成分数が存在しない場合が多い。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。

9.1　気温データにおける成分数指定

気温データにおける主成分分析では第 1 主成分のみで 0.95 を超える寄与率があり，固有値を用いた成分数の推定では成分数が 1 になってしまう。しかしながら独立成分分析における成分数はこれまで見てきたようにそれだけでは不十分である。周期性のない小さな成分は数に含めないとしても，最低でも長期間周期性と短期間周期性の 2 成分は欲しいところである。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ていたのは 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。ただし，寄与率の小さい成分でも毎回一定の成分が得られており，その他が全てノイズであると言いきれるわけではない。

表 8: 気温データにおける最大成分の尖度

	2 成分	3 成分	4 成分	5 成分	6 成分	7 成分	8 成分
1	−1.177	−1.230	−1.207	−1.210	−1.224	−1.037	−1.207
2	−1.177	−1.230	−1.207	−0.496	−1.209	−1.211	−1.205
3	−1.177	−1.230	−1.206	−1.210	−1.207	−1.210	−1.177
4	−1.177	−1.230	−1.208	−1.210	−1.230	−1.199	−1.205
5	−1.177	−1.230	−1.205	−1.210	−1.208	−1.198	−1.206

9.2 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。尖度の計算式はデータ数を n として

$$\begin{aligned} \mathrm{kurt}(\boldsymbol{x}) &= \frac{\sum_i (x_i - \overline{x})^4}{n} \Bigg/ \left[\frac{\sum_i (x_i - \overline{x})^2}{n-1} \right]^2 - 3 \\ &= \frac{\sum_i (x_i - \overline{x})^4}{(\sum_i (x_i - \overline{x})^2)^2} \frac{(n-1)^2}{n} - 3 \end{aligned} \tag{8}$$

とする。

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 8 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，9.1 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。5 成分とした場合には 5 回のうち 1 回だけ不安定性のため独立成分が分かれ，全く異なる尖度が出ていた。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。

9.3 シミュレーションにおける尖度

同様に，シミュレーションデータに対しても尖度を調べた。その結果，正規分布に従うノイズを仮定した場合は尖度にあまり変化が無く，ノイズが一様分布やラプラス分布などの非正規分布に従う場合には尖度に変化がある場合が多かった。ただし例えばノイズが極端に多い場合などには成分数を絞ると同時に尖度が単調減少するような場合も稀にあり，一概に尖度が 0 から離れていればいいとは言えない結果であった。しかし成分数 2 など真の成分数よりも小さい成分を指定した場合には大幅に尖度が 0 に近づくことがほとんどで，最低限必要な成分数を把握するのに尖度は有効であることが分かった。

10　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独立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 つの独立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 9 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 1 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。なお，周期性が正弦波であるならその尖度は -1.5 であり，二山の形状となる。第 2 独立成分はそのような傾向であるものの完全な正弦波というわけではない。さらに，短周期性を示す第 8 独立成分は尖度が正となり，完全に崩れてしまっている。

一方，正規ノイズを混合したシミュレーションデータにおいて尖度を調べたところ周期成分以外のほとんどの成分が 0 であったため，気温データで見られた周期成分以外の独立成分の優ガウス性は独立成分分析の性質によるものではなく，そもそも寄与の小さい独立成分が正規分布には従っていないことによるものであると分かった。

表 9: 気温データの高次統計量

観測データ			得られた独立成分				
	歪度	尖度		寄与率	周期性	歪度	尖度
名古屋	−0.006	−1.003	1	0.004	0.085	−0.009	2.324
東京	0.006	−1.008	2	0.840	0.025	−0.118	−1.205
大阪	−0.011	−1.090	3	0.009	−0.133	0.309	2.267
福岡	0.047	−1.097	4	0.011	0.172	0.383	0.767
札幌	−0.104	−1.132	5	0.007	−0.000	−0.203	1.051
那覇	−0.230	−1.105	6	0.026	0.009	0.311	0.663
松本	−0.031	−0.933	7	0.008	−0.048	−0.709	1.997
津山	0.039	−1.123	8	0.094	0.890	0.201	0.585

11　まとめ

本研究では独立成分分析における前処理法を提案し，その実用性について気温データとシミュレーションデータを使い確かめた。また独立成分の不安定性，独立成分数，独立成分の分布などに関する未知の疑問点に対してもある程度の回答を出すことができた。主な研究成果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く方法を提案した。
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周期成分以外の独立成分の分布は優ガウス的で少し歪んでいる分布である。

12　おわりに

今後の発展の可能性であるが，今回の研究ではやはりデータ不足の感が否めない。気温データにおける独立成分の分布はある程度つかむことができたが，その他のデータではどうであるかがわかっていない。またクラスタ平均除去に使用するクラスター法も気温データ以外の例が欲しいところである。独立成分数においてももっと多種のデータについて考察をするべきであろう。
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。独立成分分析はこれまでの多変量解析法ではなし得なかった機能を持つ。特に周期性に対する独立成分分析の分離性はとても高く，実用面においても高い効果を発揮するだろう。

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] H. Attias: Independent Factor Analysis，Neural Computation, 1999.

[4] V. Calhoun, T. Adali, G. Pearlson: Independent Component Aanalysis Applied To fMRI Data: A Natural Model And Order Selection, 2001.

[5] J.F.Cardoso, A.Souloumiac: blind beamforming for non gaussian signals, IEE-Proceedings-F, vol.140, no.6, pp.362-370, 1993.

[6] 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.

[7] Y.Kano, Y.Miyamoto, S.Shimizu: Factor Rotation and ICA, 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), pp.101-105, 2003.

[8] 気象庁ウェブページ，2008 年気温データ：http://www.jma.go.jp/.

[9] 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.

[10] 村田昇：【入門】独立成分分析, 東京電機大学出版局, 2004.

[11] T.Pedersen，R.Bruce: Distinguishing Word Senses in Untagged Text，Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, pp.197-207, 1997.

[12] 堤憲亮・半田晶寛・Leandro Di Persia・柳田益造：独立成分分析を用いたブラインド音源分離の実環境に対する有効性の検証, 電子情報通信学会　信学技報 US2003-110, EA2003-140, 2004.