

独立成分解析：多変量データの新しい分解法

甘利 俊一

理化学研究所脳科学総合研究センター

amari@brain.riken.go.jp

カクテルパーティーの会場では、多数の音源が混ざり合う。一般に多変量のデータは、独立の要因(信号源)を合成した混合物である場合が多い。独立成分解析(ICA)は、音源分離の問題に端を発し、ニューラルネットワークの学習法を取り入れ、混合信号を独立の成分に分解する新しい手法として発展してきた。これは従来の統計的手法である主成分分析や因子分析を高次に発展させたものとも考えることもできる。

独立成分解析は、音源分離問題のもとより、画像信号の分解、脳などの測定データの分解と脳内信号の分離など、多くの問題に応用されようとしている。本稿は、独立成分解析の理論的基礎とアルゴリズムおよび応用、さらに将来の動向について、平易に解説することを試みる。

カクテルパーティー効果と信号分離

カクテルパーティー会場を考えよう。ここでは n 人の人がそれぞれ勝手に会話をしている。 i 番目の人の話し声を $s_i(t)$ としよう。 t は時間で、ここでは簡単のため、 $t=1,2,3,\dots$ の離散時間を考える。会場には m 個のマイクが設置され、この会話を記録していたとしよう。 j 番目のマイクの拾う音を $x_j(t)$ としよう。どのマイクにも n 人の会話を混合したものが入ってくる。信号分離とは $\mathbf{x}(t)=(x_1(t), \dots, x_m(t))^T$ をもとに、源信号 $\mathbf{s}(t)=(s_1(t), \dots, s_n(t))^T$ を復元することである。

j 番目のマイクにはすべての音が入るが、その強度は音源とマイクの距離に関する。 i 番目の音の減衰係数を a_{ji} とすれば

$$x_j(t) = \sum_{i=1}^n a_{ji} s_i(t)$$

と書ける。 $A=(a_{ji})$ という行列を用いれば、上式は

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

と記せる。問題は観測データ $\mathbf{x}(t)$ から $\mathbf{s}(t)$ を復元することである。

話を簡単にするために、話者の数 n とマイクの数 m とが等しいとしよう。 $m > n$ の場合は測定情報が多いからもっと簡単である。 $m < n$ の場合は不良設定性が本質的に残る。これについては後で論ずることにして、しば

らくの間は $m=n$ の場合を取り扱う。式(1)は線形であるから、 $n \times n$ 行列 A が分かれば、 $W=A^{-1}$ として

$$\mathbf{y}(t) = W\mathbf{x}(t) \quad (2)$$

を計算すればよい。正しい W を用いれば、 $\mathbf{y}(t) = \mathbf{s}(t)$ である。

問題は A が未知なことである。 A が未知でも、 $\{\mathbf{s}(1), \mathbf{x}(1)\}, \dots, \{\mathbf{s}(t), \mathbf{x}(t)\}$ の組が n 個以上利用できれば A は容易に推定できる。しかし、カクテルパーティーの場合、このようなデータは利用できないのが通常である。まして、 \mathbf{s} を脳内に発生する信号、 \mathbf{x} を頭皮上の脳電図(EEG)の測定データと考えれば、 \mathbf{s} の情報はまったく利用できない。この問題はblind source separationとも呼ばれ、測定データ $\mathbf{x}(1), \dots, \mathbf{x}(t)$ だけを手がかりに原信号 $\mathbf{s}(1), \dots, \mathbf{s}(t)$ を推定しなければならない。

ただ1つ使える手がかりは、異なった情報源からの信号は確率的に独立であるという仮定である。 i 番目の情報源からの時系列信号 $\{s_i(t)\}$ はエルゴード的な確率信号で、 $s_i(t)$ と $s_j(t')$ とは $i \neq j$ ならば独立であるとする。ここではさらに、 $s_i(t)$ と $s_i(t')$ との時間相関を無視し、 $s_i(t)$ 、 $t=1,2,\dots$ は独立で同一分布に従うと仮定しよう(この仮定の成立しないときは A の推定は実はもっと容易になる)。このとき、各時点での信号 \mathbf{s} の確率密度関数 $r(\mathbf{s})$ は、

$$r(\mathbf{s}) = r_1(s_1) \cdots r_n(s_n) \quad (3)$$

と積に分解できる。ただし、各確率変数 s_i の密度関数 r_i は、各人の声の質によるから、通常は未知である。また、話を簡単にするため直流成分（バイアス）は差し引いて、 s_i の期待値は 0、つまり E を期待値として

$$E[s_i] = 0 \quad (4)$$

としておく。

カクテルパーティー効果をどう解くか—歴史的いきさつ

混合信号の分離問題は雑音除去とも関連して、古くから研究されていた。しかし、マイクを多数使った上記の問題設定は、フランスの Jutten と Herault が 10 年ほど前に考えたもので、彼らはニューラルネットワークの学習を利用した。候補となる $W = (w_{ij})$ を学習により逐次的に変更していく。時間 t での候補を W_t 、このときの復元した信号の候補を

$$y(t) = W_t x(t) \quad (5)$$

とする。このとき、 $W_{t+1} = W_t + \Delta W_t$ の各要素を、 η を小さい定数として

$$\Delta w_{ij}^t = -\eta y_j^3(t) y_i(t) \quad (i \neq j) \quad (6)$$

により変更していこうというのである。ニューロの学習のアイデアである。

この方法はうまくいくこともあれば、駄目な場合もある。 n が 3 より大きくなるとなかなかうまくいかない。初期の成功例に、妊婦の心電図データから胎児と母体に由来するものを分離したのがある。

式 (6) でなぜ分離が行えるのか、直観的に説明してみよう。学習によって η 程度の変動を除いて W_t が収束したとしよう。このとき、 $y_j^3(t) y_i(t)$ は $i \neq j$ のときは平均 0 に収束しているはずである。 y_j が互に独立になっていれば

$$E[y_j^3 y_i] = E[y_j^3] E[y_i] = 0 \quad (7)$$

となるから、 y_i と y_j は独立、したがって元の信号が復元したと考えてよい。しかし、もっと簡単な

$$\Delta W_t = -\eta y_i(t) y_j(t) \quad (8)$$

ではなぜいけないのかと、読者はいぶかるかもしれない。式 (8) では、 y_i と y_j の相関がなくなるように、

$$E[y_i y_j] = 0 \quad (9)$$

を達成するように学習が進む。これは主成分分析 (PCA,

principal component analysis) としてよく知られている統計的手法の学習版である。PCA ではなぜ信号分離ができないのか、これについては節を改めて調べてみよう。

独立成分の研究はここ 10 年間ほどで急速に進んだ^{1), 2)}。理論とアルゴリズムについて、大きな発展があり、フランス、アメリカ、フィンランド、そして日本が研究の中心と目されている。日本での研究は、情報幾何を基礎に、リー群や非ホロノーム拘束などを用い、学習の安定論や誤差解析³⁾ で世界に一步先んじている。一方、高速アルゴリズムや種々の新しい分野への応用は世界でもっと盛んである。特に、医療データ解析への応用、画像データの新しい基底としての使用方法や雑音除去法など、これから発展する興味ある主題が多い。

信号処理の分野では、通信路の周波数特性に対応して、受信側で逆フィルタをかける。これを等化器という。周波数特性を時間軸で考えれば、原信号 $s(t)$ をコンボリューションして

$$x(t) = \int A(t-\tau) s(\tau) d\tau \quad (10)$$

とする時間軸上での信号の混ぜ合わせである。適応等化器は、信号処理の分野で古くから研究されていたが、携帯電話の普及によって重要な技術的課題になってきた。この場合は異なるチャンネル間の干渉があるから、時空間的な混合を分離することになる。これらはこれまで説明してきた信号分離問題と密接に関係しているので、背景の違う両者が協同して研究を進めるようになってきている。

昨年、フランスの Aussois で第 1 回の ICA 国際会議が開かれ、150 名ほどの研究者が結集した。Helsinki で開催された本年に続いて、明年は San Diego での開催が予定されており、その次は日本での開催が要望されている。

ICA と PCA

ここで主成分分析 (PCA) の手法を説明し、ICA とどう違うかを明らかにする。いま、多変量の確率変数 \mathbf{x} があり、確率密度関数 $p(\mathbf{x})$ に従って分布しているとする。簡単のため \mathbf{x} の期待値は 0 とする。この分布に従うデータ $\mathbf{x}(1), \dots, \mathbf{x}(t)$ が測定されたとしよう。 \mathbf{x} の各成分の相関は分散共分散行列

$$V = E[\mathbf{x}\mathbf{x}^T] = \int \mathbf{x}\mathbf{x}^T p(\mathbf{x}) d\mathbf{x} \quad (11)$$

で表されているが、これはデータから

$$\hat{V} = \frac{1}{t} \sum_{k=1}^t \mathbf{x}(k)\mathbf{x}^T(k) \quad (12)$$

で測定できる。

分散共分散行列 V の固有値と固有ベクトルを λ_i, \mathbf{e}_i とする。

$$V\mathbf{e}_i = \lambda_i \mathbf{e}_i \quad (13)$$

である (簡単のため固有値の縮退はないものとする)。このとき、基底 $\{\mathbf{e}_i\}$ は正規直交系をなす。信号 \mathbf{x} を各固有ベクトル方向の成分に分解すれば

$$\mathbf{x} = \sum y_i \mathbf{e}_i \quad (14)$$

と書ける。 \mathbf{e}_i を縦ベクトルとする行列を U とおくと、上式は

$$\mathbf{x} = U\mathbf{y} \quad (15)$$

と書ける。 U は直交行列だから

$$\mathbf{y} = U^T \mathbf{x} \quad (16)$$

である。

\mathbf{y} の分散共分散行列 V_y は

$$V_y = E[\mathbf{y}\mathbf{y}^T] = UVU^T \quad (17)$$

である。しかし、 U は V を対角化する直交行列であるから、 λ_i を対角成分とする対角行列を Λ とすれば

$$V_y = \Lambda, \quad (18)$$

つまり $i \neq j$ ならば

$$E[y_i y_j] = 0 \quad (19)$$

であって、 y_i と y_j は無相関である。データから U を求め、 \mathbf{x} を \mathbf{y} に変換するのが PCA である。

データ \mathbf{x} は $\mathbf{x} = A\mathbf{s}$ による独立な信号 \mathbf{s} の混合であったとしよう。 PCA を行えば、 $\mathbf{y} = U^T \mathbf{x}$ の各成分は無相関になるから、これで \mathbf{s} が復元できそうである。ところがカクテルパーティーのデータをこの方法で解析しても多くの場合うまくいかない。第 1 に、独立と無相関とは違うことに注意したい。独立ならば無相関であるが、逆は真ではない (ガウス分布のときのみ真である)。第 2 に、カクテルパーティーの混合行列 A は一般に直交行列ではないのに、 PCA では直交行列のみを用いて復元しようとしているから変である。

混合行列 A は一般に 2 つの直交行列 O, U と対角行列 D を用いて

$$A = ODU \quad (20)$$

と分解できる。特異値分解である。 PCA を用いれば D と U は求めることができるが O は求まらない。さらに、 PCA で

$$\tilde{\mathbf{y}} = \Lambda^{-1/2} \mathbf{y} \quad (21)$$

とおけば、任意の直交行列 O に対して

$$E[O\tilde{\mathbf{y}}(O\tilde{\mathbf{y}})^T] = I \text{ (単位行列)} \quad (22)$$

であるから、 $O\tilde{\mathbf{y}}$ の成分は O として何を用いようが互いに無相関である。これは、 2 次モーメントである共分散行列からでは A ないし W は求まらず、高次のキュムラントなど、通常の線形代数を超える手法が必要であることを意味する。

ICA と情報幾何

信号 \mathbf{x} を W によって $\mathbf{y} = W\mathbf{x}$ と変換してみよう。このとき

$$\mathbf{y} = W\mathbf{A}\mathbf{s} \quad (23)$$

である。 A は固定して、 W を変えながら \mathbf{y} の確率分布がどう変わるかを見よう。 \mathbf{s} は $r(\mathbf{s})$ の分布を持つから、 \mathbf{y} の確率密度関数を $p(\mathbf{y}; W)$ と書くと、

$$p(\mathbf{y}; W) = |W\mathbf{A}|^{-1} r(\mathbf{A}^{-1}W^{-1}\mathbf{y}) \quad (24)$$

である。 $W = A^{-1}$ ならば $p(\mathbf{s}) = r(\mathbf{s})$ となる。

上式を W をパラメータとする統計モデルとみて、データ $\mathbf{x}(1), \dots, \mathbf{x}(t)$ から W を推定するという通常の統計的手法に頼ったらどうであろう。困ったことに、 $p(\mathbf{y}; W)$ は未知関数 $r = r_1 \dots r_n$ によっている。つまり、この統計モデルは n 個の未知関数 r_1, \dots, r_n を含んでいる。この種のモデルはセミパラメトリック (またはノンパラメトリック) と呼ばれ、困難である。近年、情報幾何をもとに推定関数法を用いた解法が発表された。もう 1 つ、仮に独立な成分 y_1, \dots, y_n が復元したとしても、その順番のつけ方は元の順序と一致するとは限らないことと、各信号のスケールは未定のまま残ることを指摘しておこう。これは他の物理的条件から求めなければならない。

情報幾何は確率分布族のなす多様体から導かれた新しい双対微分幾何構造であるが⁴⁾、それには深入りせずこの問題を直観的に考えてみよう。 S として、ベクトル確率変数 \mathbf{s} のあらゆる確率分布 $p(\mathbf{y})$ 全体のなす空間 (関数空間) を考えよう。この中に、独立な確率分布 $q(\mathbf{y})$ 、すなわち

$$q(\mathbf{y}) = q_1(y_1) \dots q_n(y_n) \quad (25)$$

の形の分布のなす空間 E が部分空間として入っている (図-1 参照)。真の分布 r は (未知) であるが、 E の中に入っている。さて、式 (24) で指定される確率分布の全体は、やはり S の中に入っている。 $W = A^{-1}$ のときはこれは $r(\mathbf{y})$ に等しいから、 W を自由パラメータとするこの集合 M は、 E とは r 点で交わる n^2 次元の空間である。 W はその座標系とみなせる。

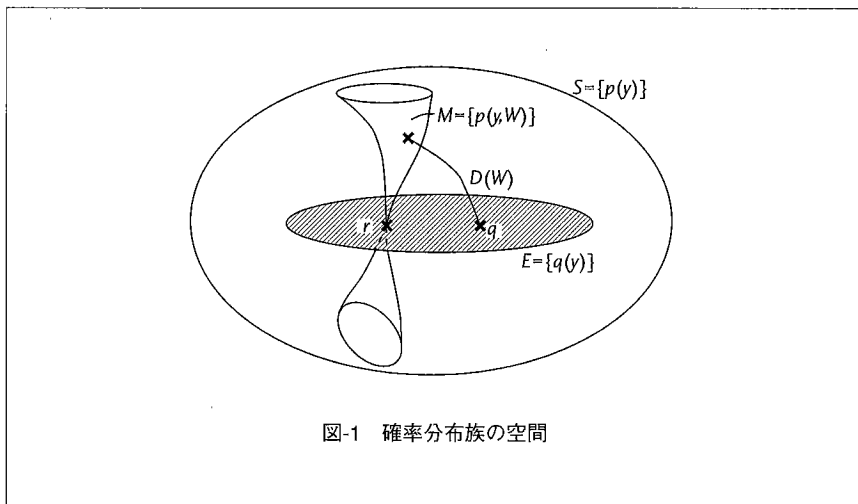


図-1 確率分布族の空間

現在の候補 W を評価するために、 $p(\mathbf{y}; W)$ が独立分布とどの位離れているのかを見よう。2つの確率分布の離れ具合は、Kullback-Leiblerのダイバージェンスで見るとよい。 $p(\mathbf{y}; W)$ と1つの独立分布 $q(\mathbf{y})$ とのダイバージェンスは

$$D[W] = \int p(\mathbf{y}; W) \log \frac{p(\mathbf{y}; W)}{q(\mathbf{y})} d\mathbf{y} \quad (26)$$

で計れる。 q として $r(\mathbf{y})$ を用いればよいが、 r が分かっていないので、とりあえず適当に q を選ぶことにしよう。

図-1 で見るように、 M と E が直交していれば、 $W=A^{-1}$ のときに $D[W]$ が最小になる。もっとも、 M が交点 r で曲がっていれば $W=A^{-1}$ で極小ではなくて極大になることもある。このときは、 q を曲率の反対側、もしくは曲率半径の内側にとらなければならない。直観的にはこの通りであるが、 S はユークリッド空間ではなくてリーマン幾何構造を持ち、さらに双対アフィン平坦接続の空間である。情報幾何学を用いてこうした解析ができる。曲率は、 q を用いて $D[W]$ を最小化しようというアルゴリズムの安定性に関係する。

学習アルゴリズムと分離例

確率分布 $q(\mathbf{y})$ を適当に選び、 $D[W]$ が減少する方向に W を変化させる確率降下法を考えよう。

$$\begin{aligned} D[W] &= -H - E[\log q(\mathbf{y})] \\ &= -H - \sum E[\log q_i(\mathbf{y}_i)] \end{aligned} \quad (27)$$

と書ける。ただし、 H は \mathbf{y} のエントロピー、 E は \mathbf{y} についての期待値で、 $\mathbf{y} = W\mathbf{x}$ という関係を通じて上式は W に依存している。細かい計算は省くが、上式のグラジエント $\partial D / \partial W$ は

$$\partial D / \partial W = W^{-T} - E[\varphi(\mathbf{y})\mathbf{x}^T] \quad (28)$$

という行列である。ここに、 W^{-T} は W の逆行列の転置で、 $\varphi(\mathbf{y})$ は

$$\varphi_i(\mathbf{y}_i) = -\frac{d}{d\mathbf{y}} \log q_i(\mathbf{y}_i) \quad (29)$$

を成分とするベクトルである。この式は、ニューラルネットワークとエントロピー最大化を用いて Bell と Sejnowski によって導出され、

$$\Delta W_t = \eta \{ W_t^T - \varphi(\mathbf{y}_t) \mathbf{x}_t^T \} \quad (30)$$

という学習アルゴリズムが提味された。

これは行列 W の作る空間での最急降下法である。しかし、行列 W の全体は $Gl(n)$ というリー群をなし、ここからリーマン空間の構造を持つことが分かる。リーマン空間での最急降下方向は $\partial D / \partial W$ ではなくて、これにリーマン計量の逆行列を掛けなければいけない。甘利らはこれを自然勾配法と呼び

$$\Delta W_t = \eta [I - \varphi(\mathbf{y}_t) \mathbf{y}_t^T] W_t \quad (31)$$

というアルゴリズムを提唱した⁵⁾。これは、 A や W が悪条件に近いときでもうまく働く効果的なアルゴリズムである。

アルゴリズムを導くには、いろいろな考え方があり、いくつかの web でプログラムも公開されている。また理論として、統計的な誤差解析や学習の安定論も確立している。ここでは、コンピュータシミュレーションによる結果を挙げておこう。図-2 (a) は3個の音源からの信号を混合し4個のマイクで集音した混合信号である。ここでは非ホロノーム束縛アルゴリズムを用いて、音源数を知ることなく学習を進め、3個の音がうまく分離されていることが分かる(図-2 (b))。

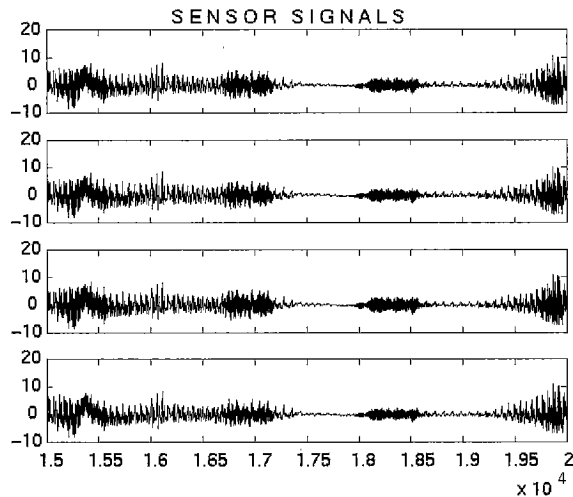


図-2 (a) 4個のマイクからの混合音の時間波形

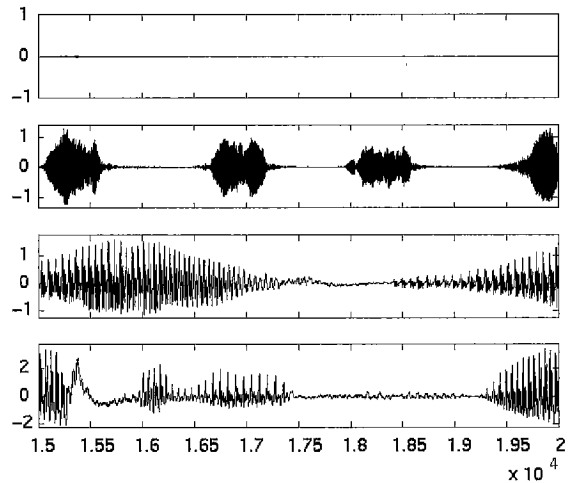


図-2 (b) 分離した3個の独立な音源信号

ICAは実データの解析に有効か？

ICAは従来の多変量解析の枠を超えるものである。これまでPCAを用いていたデータにICAをほどこしてみると、いろいろ興味あることが分かる。特に、EEG、MEG、fMRIなどのデータは、脳の内部にある信号源の信号を混合したものを外部から測定したもので、これらのデータから脳内部の活動を推定することは、不良設定問題である。独立性という手がかりでこれが分離同定できれば、脳科学にとって福音であろう。こうして、実データをもとに、多くの試みが行われている。

しかし、本当の脳内部の信号は分からないから、ICAで分解したものが果たして意味のある信号かどうかを見極めることは難しい。また、脳内部の信号は独立ではないであろうし、定常信号でもなく、測定時の雑音も大きい。もちろん、脳の外部からの雑音は脳活動とは独立であり、雑音除去には有効であることは分かっているが、それ以上にこれを用いることはまだ難しく、挑戦しがいのある研究課題である。

過完全基底とスパース表現

多変量の信号をICAで分析することについて、画像データ処理に最近新しい発展があった。多変量のデータ \mathbf{x} を考えよう。 \mathbf{x} は多数のマイクからの信号でもよいが、1枚の画像をスキャンしてベクトル化した画像信号でもよい。画像信号の場合、これをフーリエ変換したり、ウェーブレット変換したりして、その特徴を調べ、また次の処理を容易にしようとする。これらはいずれも基底変換で、

$$\mathbf{x} = \sum_{i=1}^n y_i \mathbf{a}_i \quad (32)$$

によって、 \mathbf{x} から \mathbf{y} へ変換する。フーリエ変換の場合は直交変換で、 $\{\mathbf{a}_i\}$ は正規直交基底であるが、ウェーブレットの場合は必ずしも直交変換ではない。いずれにせよ、基本となる要素画像 \mathbf{a}_i の重ね合わせから \mathbf{x} ができていると考える。

独立成分解析は、重ね合わせの係数 y_i がなるべく互いに独立になるように \mathbf{a}_i を選ぶとするものである。ところが、基本要素画像として n 個のものを選ぶだけではなく、これよりはるかに多数の m 個 ($m \gg n$) の \mathbf{a}_i を考えたらどうであろう。ICAでは信号源の数 m が n よりはるかに大きいと考えることになる。このような基底系を過完全系という。この場合 $\mathbf{x} = \mathbf{A}\mathbf{y}$ と書けば、 $n \times m$ に逆行列は存在しない。しかし、多数のデータ $\mathbf{x}(1), \dots, \mathbf{x}(t)$ をもとに、 \mathbf{y} の成分が独立であるという仮定のもとで \mathbf{A} の推定値 $\hat{\mathbf{A}}$ を得ることはできる。学習方式も求まる。1枚の画像の場合は、これを小画面に分割してこれを $\mathbf{x}(i)$ として用いる。

\mathbf{A} が推定されたときに、ここから \mathbf{y} をどうやって求めたいだろう。従来の考えは、 $\hat{\mathbf{A}}$ の一般逆行列 \mathbf{A}^{\dagger} を用いて、

$$\mathbf{y} = \mathbf{A}^{\dagger} \mathbf{x} \quad (33)$$

とするものであった。これは、今の場合 $\mathbf{x} = \hat{\mathbf{A}}\mathbf{y}$ を満たすもののうちで、2乗ノルム $\|\mathbf{y}\|^2 = \sum y_i^2$ を最小にするものである。このとき、多くの y_i は0でない値を持つ。この代わりに、 $\mathbf{x} = \hat{\mathbf{A}}\mathbf{y}$ を満たす \mathbf{y} のうちで、1乗ノルム $\sum |y_i|$ を最小にするものを求めたらどうなるだろう。この問題は線形計画法を用いて解け、解 \mathbf{y} の成分はできるだけ多く

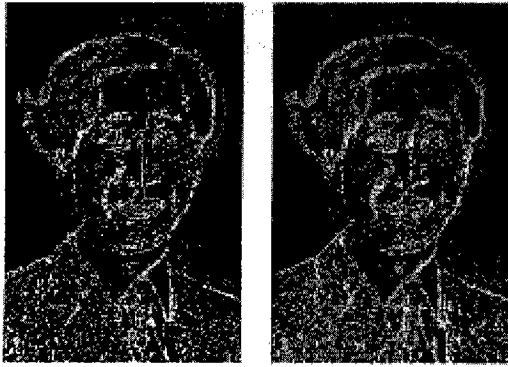


図-3 (a) 3枚の画像を重ね合わせた2枚の画像



図-3 (b) 分離した3枚の画像

が0になって、最小の個数の y_i のみが非零であるようになる。これをスパース表現という⁶⁾。

独立成分解析で、話者の数がマイクの数よりも多いときでも、瞬間の時間では話をしている人数は限られているだろう。このときは $m > n$ でもスパース表現として解が求まる。ここでは、コンピュータでの実験例として音声の代わりに3枚の画像を重ね合わせた2枚(図-3 (a))の混合画像から、スパース法を用いて復元した例(図-3 (b))を示す。

自然画像を表現する過完全系を求めてみると、これは脳内のニューロンの特性に類似していることが分かる。ここから、脳は画像処理に独立成分解析を用いているのではないかという可能性が生じる。また、スパース表現は画像のノイズ除去にきわめて有用であることが報告されている。これは非線形のフィルタリングである。

おわりに

独立成分解析は、多変量データを解析する新しい統計的手法である。これは、カクテルパーティー効果など、混合信号を分離する問題として現れたが、より深

い統計的構造を持つことが明らかになってきた。本稿では、数学的定式化に深入りすることなく、その基本的な考え方を分かりやすく説明することを心がけた。

独立成分解析は、多方面の現実データに応用されようとしている。そのためには、手法そのものもさらに現実に合ったものに改良していかなければならない。現在世界中で研究が進んでおり、これからの発展が楽しみである。

参考文献

- 1) Lee, T.-W.: Independent Component Analysis, Kluwer Academic, Boston (1998).
- 2) Amari, S. and Cichocki, A.: Adaptive Blind Signal Processing — Neural Network Approach, Proceedings of IEEE, 86, pp.2026-2048 (1998).
- 3) Amari, S., Chen, T.-P. and Cichocki, A.: Stability Analysis of Adaptive Blind Source Separation, Neural Networks, 10, pp.1345-1351 (1997).
- 4) 甘利俊一, 長岡浩司: 情報幾何の方法, 岩波応用数学講座 (1993).
- 5) Amari, S.: Natural Gradient Works Efficiently in Learning, Neural Computation, 10, pp.251-276 (1998).
- 6) Lewicki, M.S. and Sejnowski, T.: Learning Nonlinear Overcomplete Representation for Efficient Coding, In Kearns, M., Jordan, M. and Solla, S. (Eds.), Advances in Neural Information Processing Systems, 10, pp. 815-821, Cambridge, MA: MIT Press (1998).

(平成12年8月1日受付)