信号処理

～第 5 部 独立成分解析～

Matlab 対応版

横田 康成

平成 15 年 9 月 26 日

目 次

1 独立成分解析の基礎 **3**
1.1 問題設定 . . . 3
1.2 最尤推定法 . . . 4
1.3 平均相互情報量最小化法 . . . 6
1.4 Super-Gaussian と Sub-Gaussian . . . 8
1.5 独立成分解析の性能の評価 . . . 9

2 プリ・ホワイトニングを伴う学習高速化 **10**
2.1 プリ・ホワイトニングと歪対称化法 . . . 10
2.2 射影追跡法 . . . 12
2.3 Fast ICA . . . 13
2.4 非正規性最大化法 . . . 14

3 BSD **15**
3.1 問題設定 . . . 15
3.2 z 変換を用いた表現 . . . 16
3.3 周波数領域での BSD . . . 17
3.4 時間領域での BSD . . . 17

4 ICA の推定精度 **18**
4.1 ICA とセミパラメトリック推定, 推定関数法 . . . 18
4.2 r-スコア関数, nuisance 接空間 . . . 19
4.3 A-スコア関数, 有効スコア関数 . . . 20
4.4 ICA における推定関数の空間 . . . 21
4.5 ICA の推定精度と有効性 . . . 22

A 付録 **22**
A.1 確率変数の独立性 . . . 22
A.2 確率変数の正規性 . . . 23
A.3 確率変数のエントロピーとその近似 . . . 23

A.4 自然勾配 24
A.5 主成分解析 25
A.6 行列の直交化 25
A.7 スコア関数と Cramér-Rao の下界, その情報幾何学的解釈 25
A.8 推定関数法 27

B 演習の解答 (Matlab のプログラム) 30

1 独立成分解析の基礎

1.1 問題設定

n 個の信号源があり，それらから信号 $s_i(t),\ i=1,\ldots,n$ が出力されるとしよう．これらの信号を原信号と呼ぶことにする．我々は，m 個のセンサーを用いてこれらの信号の観測を試みるが，原信号が混ざり合って，$x_i(t),\ i=1,\ldots,m$ として観測されるものする．これらの信号を観測信号と呼ぶことにする．混ざり合い方が線形であるとすると，観測信号は，原信号を用いて次式で表現することができる．

$$x_i(t)=\sum_{j=1}^{n} a_{i,j}s_j(t), \qquad i=1,\ldots,m \tag{1}$$

ここで，$a_{i,j},\ i=1,\ldots,n,\ j=1,\ldots,m$ は，混ざり合いの強さを表す係数である．我々は，原信号を知ることが目的であるから，観測された観測信号から原信号を推定しなければならない．この時，混ざり合いの割合を表す係数 $a_{i,j},\ i=1,\ldots,n,\ j=1,\ldots,m$ が既知であれば，これらが適当な条件を満たせば，観測信号から原信号を推定することは容易である．それでは，この係数が未知である場合は，どうなるのであろうか？実は，$n\leq m$ で，原信号 $s_i(t),\ i=1,\ldots,n$ が互いに独立，原信号のうち 2 つ以上が正規分布に従わないならば，観測信号から原信号を推定することが可能になる．そのための手法を独立成分解析 (Independent Component Analysis, 以下，ICA) という．

ICA は，更に，BSS(Blind Source Separatoin) と BSD(Blind Source Deconvolution) に分類される．BSS は，式 (1) で示されるように観測信号が表現され，BSD は，一般に，

$$x_i(t)=\sum_{j=1}^{n}\sum_{\tau=0}^{T} a_{i,j}(\tau)s_j(t-\tau), \qquad i=1,\ldots,m \tag{2}$$

として表されるように，原信号が時間遅れを持って混ざり合って観測信号となる場合を対象とする．本書では，1 章，2 章で BSS について述べ，3 章で BSD について述べる．

BSS を対象にする限り，原信号，観測信号ともに，時刻を表す変数 t の値は意味を持たなくなるため，各時刻の信号は，単なる標本と考えても良い．そこで，原信号と観測信号をそれぞれ確率変数 $\boldsymbol{s}_i,\ i=1,\ldots,n$, $\boldsymbol{x}_i,\ i=1,\ldots,m$ とおき，それらの標本が各時刻 t での $x_i(t),\ y_i(t)$ であると考える[1]．また，列ベクトルを用いて，$\boldsymbol{s}=(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n)^T$，$\boldsymbol{x}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m)^T$ とおき，$a_{i,j}$ を i 行 j 列要素に持つ行列 $\boldsymbol{A}$ を定義すれば，式 (1) は，次式で書けることになる．

$$\boldsymbol{x}=\boldsymbol{A}\boldsymbol{s} \tag{3}$$

したがって，適当な n 行 m 列の行列 $\boldsymbol{W}$ を用いて表現される

$$\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x} \tag{4}$$

を原信号 $\boldsymbol{s}$ の推定値とする．ベクトル $\boldsymbol{y}$ は，復元信号とも呼ばれ，行列 $\boldsymbol{A}$，$\boldsymbol{W}$ は，その役割から，それぞれ混合行列，復元行列などと呼ばれることが多い．

また，これ以降，観測信号の平均がゼロではなかった場合には，$\boldsymbol{x}-E[\boldsymbol{x}]$ を新たに観測信号 $\boldsymbol{x}'$ とみなし，平均がゼロであるとして解析を行うものとする．このようにしても，平均がゼロであるか否かは，独立性とは無関係であるし，観測信号 $\boldsymbol{x}'$ に対して得られた復元行列 $\boldsymbol{W}$ を用いて，

$$\boldsymbol{W}\boldsymbol{x}'=\boldsymbol{W}(\boldsymbol{x}-E[\boldsymbol{x}])=\boldsymbol{W}\boldsymbol{x}-\boldsymbol{W}E[\boldsymbol{x}]=\boldsymbol{y}-\boldsymbol{W}E[\boldsymbol{x}]$$

であるから，$\boldsymbol{y}=\boldsymbol{W}(\boldsymbol{x}'+E[\boldsymbol{x}])$ として，平均がゼロでないかも知れない本来の原信号 $\boldsymbol{y}$ を得ることができるので，問題はない．

[1]本書では，ベクトルを太字の小文字，行列を太字の大文字で表すことにする．また，確率変数については，特に太字で表記しないものとする．

原信号の推定値 $\boldsymbol{y}$ が真値 $\boldsymbol{s}$ に一致する，つまり $\boldsymbol{y} = \boldsymbol{s}$ であるための必要十分条件は，$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{A}\boldsymbol{s} = \boldsymbol{s}$ なので，$\boldsymbol{W}\boldsymbol{A} = \boldsymbol{I}$ である．ただし，$\boldsymbol{I}$ は，$n \times n$ の単位行列である．しかし，原信号と混合行列の両方が未知なので，ちょっと考えれば，原信号 $\boldsymbol{s}$ の順番と振幅までを求めることが不可能であることがわかる．したがって，$\boldsymbol{y}$ の解を，行列 $\boldsymbol{W}\boldsymbol{A}$ が，適当な対角行列に対し，行を適当に並び替えて得られるすべての行列の範囲に拡大することができる．混合行列 $\boldsymbol{A}$ の階数が n でなければならないことも条件であることがわかるであろう．また，当面は，観測信号の個数 m が原信号の個数 n に一致していると仮定する．つまり，混合行列 $\boldsymbol{A}$，復元行列 $\boldsymbol{W}$ は，ともに正方行列になる．

1.2 最尤推定法

まず，原信号 $\boldsymbol{s}$ の確率密度分布 $r(\boldsymbol{s})$ が既知であるものとして，独立成分解析を最尤推定法により行ってみよう．$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ なる変換において，観測信号 $\boldsymbol{x}$ の確率密度分布 $p(\boldsymbol{x})$ と原信号 $\boldsymbol{s}$ の確率密度分布 $r(\boldsymbol{s})$ の関係は，

$$p(\boldsymbol{x})d\boldsymbol{x} = r(\boldsymbol{s})d\boldsymbol{s}$$

で与えられる．$d\boldsymbol{x}$ は，$d\boldsymbol{s}$ に $\boldsymbol{s}$ から $\boldsymbol{x}$ への変換のヤコビアン

$$\left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{s}}\right| = \begin{vmatrix} \dfrac{\partial x_1}{\partial s_1} & \cdots & \dfrac{\partial x_n}{\partial s_1} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial x_1}{\partial s_n} & \cdots & \dfrac{\partial x_n}{\partial s_n} \end{vmatrix}$$

を乗じたものになるが，今，この変換は行列 $\boldsymbol{A}$ により与えられる線形変換なので，ヤコビアンは行列 $\boldsymbol{A}$ の行列式になる．すなわち，

$$d\boldsymbol{x} = \left|\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{s}}\right| d\boldsymbol{s} = |\boldsymbol{A}|d\boldsymbol{s}$$

となる．したがって，観測信号 $\boldsymbol{x}$ の確率密度分布 $p(\boldsymbol{x})$ は，

$$p(\boldsymbol{x}) = r(\boldsymbol{s})|\boldsymbol{A}|^{-1}$$

と表される．今，$\boldsymbol{A}^{-1}$ を $\boldsymbol{W}$，$\boldsymbol{s}$ を $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ として推定しようとしているので，上の式は，

$$p(\boldsymbol{x}) = r(\boldsymbol{W}\boldsymbol{x})|\boldsymbol{W}|$$

と書ける．そして，$\boldsymbol{x}$ の標本 $\boldsymbol{x}(t),\ t = 1, \ldots, T$ を観測し，その対数尤度

$$L(\boldsymbol{W}) = \sum_{t=1}^{T} \log p(\boldsymbol{x}(t)) = \sum_{t=1}^{T} \log(r(\boldsymbol{W}\boldsymbol{x}(t))|\boldsymbol{W}|) = \sum_{t=1}^{T} (\log|\boldsymbol{W}| + \log r(\boldsymbol{W}\boldsymbol{x}(t))) \tag{5}$$

を最大にする復元行列 $\boldsymbol{W}$ を推定値 $\hat{\boldsymbol{W}}$ としよう．そのためには，上式を $\boldsymbol{W}$ で偏微分してゼロとおいた方程式

$$\frac{dL(\boldsymbol{W})}{d\boldsymbol{W}} = \boldsymbol{0}$$

を解けばよい．

まず，$\log|\boldsymbol{W}|$ の $\boldsymbol{W}$ に関する微分は，$|\boldsymbol{W}|$ の $\boldsymbol{W}$ に関する微分が $\boldsymbol{W}^T$ の余因子行列 $\tilde{\boldsymbol{W}}^T$ になることを利用すれば，

$$\frac{d\log|\boldsymbol{W}|}{d\boldsymbol{W}} = \frac{\tilde{\boldsymbol{W}}^T}{|\boldsymbol{W}|} = (\boldsymbol{W}^T)^{-1} \tag{6}$$

となる．また，$\log r(\boldsymbol{W}\boldsymbol{x})$ の $\boldsymbol{W}$ に関する微分は，

$$\frac{d\log r(\boldsymbol{W}\boldsymbol{x}(t))}{d\boldsymbol{W}} = \frac{d\log r(\boldsymbol{y})}{d\boldsymbol{y}}\bigg|_{y=Wx(t)} \boldsymbol{x}(t)^T$$

と書ける．ここで，各原信号 $\boldsymbol{s}_i$ は互いに独立なので，原信号の確率密度分布 $r(\boldsymbol{y})$ は，各原信号 s_i の確率密度分布 $r_i(s_i)$ の積で表現できることを利用すれば，

$$\begin{aligned}
\frac{d\log r(\boldsymbol{y})}{d\boldsymbol{y}} &= \frac{d}{d\boldsymbol{y}}\sum_{i=1}^{n}\log r_i(y_i) \\
&= \left(\frac{d\log(r_1(y_1))}{dy_1},\ldots,\frac{d\log(r_n(y_n))}{dy_n}\right)^T \\
&= \left(\frac{r_1'(y_1)}{r_1(y_1)},\ldots,\frac{r_n'(y_n)}{r_n(y_n)}\right)^T
\end{aligned}$$

となる．ここで，

$$\boldsymbol{\phi}(\boldsymbol{y}) = (\phi_1(y_1),\ldots,\phi_n(y_n))^T = \left(\frac{r_1'(y_1)}{r_1(y_1)},\ldots,\frac{r_n'(y_n)}{r_n(y_n)}\right)^T$$

とおけば，

$$\frac{d\log r(\boldsymbol{W}\boldsymbol{x}(t))}{d\boldsymbol{W}} = \boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{x}(t))\boldsymbol{x}(t)^T \tag{7}$$

と書ける．ゆえに，式 (5) で与えられる対数尤度 $L(\boldsymbol{W})$ の $\boldsymbol{W}$ による微分は，式 (6),(7) を用いて，

$$\begin{aligned}
\frac{dL(\boldsymbol{W})}{d\boldsymbol{W}} &= \sum_{t=1}^{T}\Big\{\left((\boldsymbol{W}^T)^{-1} + \boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{x}(t))\boldsymbol{x}(t)^T\right)\Big\} \\
&= T(\boldsymbol{W}^T)^{-1} + \sum_{t=1}^{T}\boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{x}(t))\boldsymbol{x}(t)^T
\end{aligned} \tag{8}$$

となる．ゆえに，この式をゼロ行列とおいた方程式

$$(\boldsymbol{W}^T)^{-1} + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{x}(t))\boldsymbol{x}(t)^T = \boldsymbol{0}$$

の根が独立成分解析の解となる．しかし，この方程式を解くことは困難なので，$\frac{dL(\boldsymbol{W})}{d\boldsymbol{W}}$ を大きくするような更新則

$$\begin{aligned}
\boldsymbol{W} &\leftarrow \boldsymbol{W} + \eta\Delta\boldsymbol{W} \\
\Delta\boldsymbol{W} &= (\boldsymbol{W}^T)^{-1} + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}(\boldsymbol{W}\boldsymbol{x}(t))\boldsymbol{x}(t)^T
\end{aligned} \tag{9}$$

を用いて，学習によりその最大値である解を探索するしかない．ただし，η は，学習率と呼ばれる正の実数である．大きな値に設定すると解への収束が早くなる反面，学習が不安定になったり，解の推定精度が悪化する．上記の更新則は，Bell-Sejnowski のアルゴリズムと呼ばれている．

一般に，ユークリッド直交座標系においては，行列 $\boldsymbol{W}$ の関数 $f(\boldsymbol{W})$ の最急勾配方向は，関数 $L(\boldsymbol{W})$ の微分 $\frac{dL(\boldsymbol{W})}{d\boldsymbol{W}}$ に一致する．しかし，パラメータ空間は，常にユークリッド直交座標系であるとは限らず，独立成分解析の場合，リーマン計量空間を持つ．リーマン計量空間では，正則な正方行列 $\boldsymbol{W}$ の場合，最適な勾配方向は，自然勾配方向

$$\frac{dL(\boldsymbol{W})}{d\boldsymbol{W}}\boldsymbol{W}^T\boldsymbol{W}$$

となることがわかっている（付録 A.4 参照）．そこで，式 (9) の代わりに，その式に $\boldsymbol{W}^T\boldsymbol{W}$ を乗じることにより得られる自然勾配を用いた更新則

$$\begin{aligned} \boldsymbol{W} &\leftarrow \boldsymbol{W} + \eta\Delta\boldsymbol{W} \\ \Delta\boldsymbol{W} &= \left(\boldsymbol{I} + \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}(\boldsymbol{y}(t))\boldsymbol{y}(t)^T\right)\boldsymbol{W} \end{aligned} \tag{10}$$

を利用する方が，収束が早く，かつ安定になる．ただし，$\boldsymbol{y}(t) = \boldsymbol{W}\boldsymbol{x}(t)$ である．

この最尤推定法を使うためには，関数 $\boldsymbol{\phi}(\boldsymbol{y})$，つまり，原信号の確率密度分布 $r(\boldsymbol{s})$ を知っている必要がある．もちろん，知らない場合がほとんどであろうから，現実的には，この方法は使えないように思える．しかし，この関数 $\boldsymbol{\phi}(\boldsymbol{y})$ をかなり適当に選択しても，独立成分解析が行えることがわかっている．もちろん，この選択によって，推定精度，収束の速度，安定性が異なってくる．こうした問題については，4 章で推定関数の立場から詳しく述べる．

1.3 平均相互情報量最小化法

復元信号 $\boldsymbol{y}$ を順番と振幅の不定性を除いて原信号 $\boldsymbol{s}$ に一致させることが目標であるが，ここで，$\boldsymbol{s}$ の各要素が独立であることに着目し，問題のすり替えを行ってみよう．つまり，各要素が独立になるように $\boldsymbol{y}$ を定める問題とするわけである．具体的には，$\boldsymbol{y}$ の独立性を測るなんらかの尺度を用いて，この尺度が独立に向かうように復元行列 $\boldsymbol{W}$ を更新して行くことにする．独立性の尺度の一つとして，$\boldsymbol{y}$ の各要素間の平均相互情報量

$$\bar{I}(\boldsymbol{y}) = \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{y})$$

があるので（付録 A.1 参照），これを利用しよう．$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ なる線形変換を行うわけであるから，先ほど述べたように，$\boldsymbol{y}$ の確率密度分布 $q(\boldsymbol{y})$ は，$\boldsymbol{x}$ の確率密度分布 $p(\boldsymbol{x})$ を用いて

$$q(\boldsymbol{y}) = p(\boldsymbol{W}^{-1}\boldsymbol{y})|\boldsymbol{W}|^{-1}$$

と表現できる．また，微小要素については，

$$d\boldsymbol{y} = |\boldsymbol{W}|d\boldsymbol{x}$$

である．したがって，$\boldsymbol{y}$ のエントロピーは，

$$\begin{aligned} H(\boldsymbol{y}) &= -\int q(\boldsymbol{y})\log q(\boldsymbol{y})d\boldsymbol{y} \\ &= -\int (\log p(\boldsymbol{W}^{-1}\boldsymbol{y}) - \log|\boldsymbol{W}|)p(\boldsymbol{W}^{-1}\boldsymbol{y})|\boldsymbol{W}|^{-1}d\boldsymbol{y} \\ &= -\int (\log p(\boldsymbol{x}) - \log|\boldsymbol{W}|)p(\boldsymbol{x})d\boldsymbol{x} \\ &= -\int p(\boldsymbol{x})\log p(\boldsymbol{x})d\boldsymbol{x} + \log|\boldsymbol{W}| \\ &= H(\boldsymbol{x}) + \log|\boldsymbol{W}| \end{aligned}$$

となり，平均相互情報量 $\bar{I}(\boldsymbol{y})$ は，

$$\bar{I}(\boldsymbol{y}) = \sum_{k=1}^{n} H(y_k) - H(\boldsymbol{x}) - \log|\boldsymbol{W}| \tag{11}$$

となる．そこで，平均相互情報量 $\bar{I}(\boldsymbol{y})$ を $\boldsymbol{W}$ で偏微分し，ゼロとおいた方程式を解くことにより，平均相互情報量を最小化する解が得られることになる．右辺第 2 項 $H(\boldsymbol{x})$ については，$\boldsymbol{W}$ には依存しないので，微分しても 0 である．第 3 項 $\log|\boldsymbol{W}|$ については，前節で述べたように，

$$\frac{\partial \log|\boldsymbol{W}|}{\partial \boldsymbol{W}} = (\boldsymbol{W}^T)^{-1}$$

となる．第 1 項の各 $H(y_k),\ k = 1, \ldots, n$ を考えよう．$q(\boldsymbol{y})d\boldsymbol{y} = p(\boldsymbol{x})d\boldsymbol{x} = r(\boldsymbol{s})d\boldsymbol{s}$ を考慮すれば，

$$\begin{aligned} H(y_k) &= -\int q_k(y_k)\log q_k(y_k)dy_k \\ &= -\int q(\boldsymbol{y})\log q_k(y_k)d\boldsymbol{y} \\ &= -E[\log q_k(y_k)] \end{aligned}$$

となり，行列 $\boldsymbol{W}$ の各要素 $w_{i,j}$ で偏微分すれば，

$$\begin{aligned} \frac{\partial H(y_k)}{\partial w_{i,j}} &= -E\left[\frac{\partial}{\partial w_{i,j}}\log q_k(y_k)\right] \\ &= -E\left[\frac{d\log q_k(y_k)}{dy_k}\frac{dy_k}{dw_{i,j}}\right] \\ &= -E\left[\frac{d\log q_k(y_k)}{dy_k}x_j\delta_{i,k}\right] \end{aligned}$$

となり，

$$\varphi_k(y_k) = \frac{d\log q_k(y_k)}{dy_k}$$

とおけば，

$$\frac{\partial H(y_k)}{\partial w_{i,j}} = -E[\varphi_k(y_k)x_j\delta_{i,k}]$$

となる．ゆえに，

$$\sum_{k=1}^{n}\frac{\partial H(y_k)}{\partial w_{i,j}} = -E[\varphi_i(y_i)x_j]$$

となるので，平均相互情報量 $\bar{I}(\boldsymbol{y})$ を $\boldsymbol{W}$ で微分したものは，

$$\frac{d\bar{I}(\boldsymbol{y})}{d\boldsymbol{W}} = -(\boldsymbol{W}^T)^{-1} - E[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{x}^T]$$

となる．ただし，$\boldsymbol{\varphi}(\boldsymbol{y}) = (\varphi_1(y_1), \ldots, \varphi_n(y_n))^T$ である．ゆえに，自然勾配 $\frac{d\bar{I}(\boldsymbol{y})}{d\boldsymbol{W}}\boldsymbol{W}^T\boldsymbol{W}$ を用いた平均相互情報量 $\bar{I}(\boldsymbol{y})$ を $\boldsymbol{W}$ を最小化する更新則は，

$$\begin{aligned} \boldsymbol{W} &\leftarrow \boldsymbol{W} + \eta\Delta\boldsymbol{W} \\ \Delta\boldsymbol{W} &= \left(\boldsymbol{I} + E[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T]\right)\boldsymbol{W} \end{aligned} \tag{12}$$

となる．これは，$\boldsymbol{\phi}(\cdot) = \boldsymbol{\varphi}(\cdot)$ とし，期待値を標本平均に置き換えれば，式 (10) と同一の式である．$\phi(\cdot)$ は，原信号の確率密度分布の対数の微分であり，$\varphi(\cdot)$ は，復元信号の確率密度分布の対数の微分である．すなわち，真値付近，つまり $\boldsymbol{s} = \boldsymbol{y}$ では，$\phi(\cdot) = \varphi(\cdot)$ となるので，対数尤度を最大化する ICA と，平均相互情報量を最小化する ICA は，等しいことをやっていることになる．

[**演習 1**]　原信号の数を $n = 2$ とし，原信号 s_1，s_2 がそれぞれ独立に $[-1, 1]$ の一様分布に従う確率変数であるとしよう．また，混合行列 $\boldsymbol{A}$ を平均 0，分散 1 の正規乱数により設定して，観測信号 $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{s}$ を作成

し，式 (10)，あるいは式 (12) に基づいて独立成分解析を行ってみよう．独立成分解析にあたっては，原信号の振幅までは同定できないので，観測信号 $x_i,\ i = 1, 2$ は，いずれも平均 0，分散 1 に規格化しておこう．また，$\phi_i(y_i)(= \varphi_i(y_i))$ は，適当に $-y_i^3,\ i = 1, 2$ としておこう．また，学習率 η は，$\eta = 0.1$〜0.5 位，標本数 T は，1000 程度でよいであろう．

さらに，s_1，s_2 をそれぞれ独立に平均 0，分散 1 の正規分布に従う確率変数を 3 乗することにより得られる確率変数とした場合にはどうなるか調べてみよう．$\phi_i(y_i) = -\tanh(y_i),\ i = 1, 2$ とした場合についても調べてみよう．さらに，s_1 を $[-1, 1]$ の一様分布に従う確率変数，s_2 を平均 0，分散 1 の正規分布に従う確率変数を 3 乗することにより得られる確率変数とした場合にはどうであろうか．いろいろな組み合わせで調べてみよう．(ica_learn1.m, nor.m)

1.4 Super-Gaussian と Sub-Gaussian

演習 1 を行うことにより気づいたと思うが，すべての原信号が一様分布に従う場合には，$\phi_i(y_i) = -y_i^3$，また，すべての信号が正規信号の 3 乗である場合には，$\phi_i(y_i) = -\tanh(y_i)$ とすれば，独立成分解析は成功する．原信号にこれら 2 種類の信号が含まれている場合には，$\phi_i(y_i)$ をどちらに選んでも上手く行かない．関数 $\phi_i(y_i)$ は，原信号の確率密度分布が未知なので，その代用分布の対数の微分として与えられる関数である．$\phi_i(y_i)$ をかなり適当に選んでも，独立成分解析は上手く行くが，最低限，2 種類を切り替える必要があることがわかっている．具体的には，各復元信号の 4 次キュムラント（付録 A.2 参照）の符号により，異なる関数 $\phi_i(y_i)$ を選択する必要がある．正の 4 次キュムラントを持つ確率密度分布を Super-Gaussian 分布，負の 4 次キュムラントを持つ確率密度分布を Sub-Gaussian 分布という．一様分布の 4 次キュムラントは負になるから，Sub-Gaussian であり，一方，正規信号を 3 乗した信号は正の 4 次キュムラントをもち，Super-Gaussian となる．ゆえに，これらの分布では，本来，異なる関数 $\phi_i(y_i)$ を選択する必要があったのである．

こうしたことから，各反復における各復元信号 $y_i,\ i = 1, \ldots, n$ に対して，それが Super-Gaussian ならば，$\phi_i(y_i) = -(y_i + \tanh(y_i))$，Sub-Gaussian ならば，$\phi_i(y_i) = -(y_i - \tanh(y_i))$ を用いる方法が考案された．つまり，$\phi(\boldsymbol{y})$ を

$$\phi(\boldsymbol{y}) = -\begin{pmatrix} y_1 + \mathrm{sgn}(c[y_1^4])\tanh(y_1) \\ \vdots \\ y_n + \mathrm{sgn}(c[y_n^4])\tanh(y_n) \end{pmatrix} \tag{13}$$

する．y_i が Super-Gaussina か Sub-Gaussian かのどちらの分布であるかを推定する問題が新たに付加されるが，通常は，4 次キュムラント $c[y_i^4] = E[y_i^4] - 3E[y_i^2]^2$ を推定し，その正負により判定する．こうした $\phi_i(y_i)$ を切り替える手法は，Lee, Girolami, Sejnowski により提案された．ここでは，この手法を Extended Informax 法と呼ぶ．

[演習 2]　原信号 s_1，s_2 をそれぞれ以下のように作成し，Extended Informax 法により，独立成分解析を行ってみよう．それ以外の条件は，演習 1 と同様としよう．

- s_1，s_2 ともに一様分布に従う信号
- s_1，s_2 ともに正規信号の 3 乗
- s_1：一様分布に従う信号，s_2：正規信号の 3 乗

(ica_learn2.m, nor.m)

1.5 独立成分解析の性能の評価

独立成分解析法の性能を評価するための尺度を考えよう．もっとも自然な尺度は，原信号 $\boldsymbol{s}$ と復元信号 $\boldsymbol{y}$ の間の平均 2 乗誤差である．復元信号は，その順番と振幅については一意に決定されないから，それらを補正した平均 2 乗誤差を求める必要がある．原信号 s_i と復元信号 y_j の間の振幅を補正した 2 乗誤差 $err_{i,j}$ は，

$$err_{i,j} = E[(s_i - ay_j)^2]$$

を最小にする係数 a を代入した $err_{i,j}$ となる．$err_{i,j}$ を a で偏微分しゼロとおいた方程式を解くことにより，

$$a = \frac{E[s_i y_j]}{E[y_j^2]}$$

が得られるので，これを $err_{i,j}$ に代入すると，

$$err_{i,j} = E[s_i^2] - \frac{E[s_i y_j]^2}{E[y_j^2]}$$

となる．各原信号 $\boldsymbol{s}_i$ に対し，$err_{i,j}$ を最小にする復元信号 $\boldsymbol{y}_j$ を選択し，それらの平均を平均 2 乗誤差 mse とすると，

$$mse = \frac{1}{n}\sum_{i=1}^{n} \min_j err_{i,j}$$

となる．

一方，混合行列 $\boldsymbol{A}$ のパラメータ推定の立場からみれば，復元行列 $\boldsymbol{W}$ の逆行列を混合行列 $\boldsymbol{A}$ の推定値 $\hat{\boldsymbol{A}} = \boldsymbol{W}^{-1}$ と見て，$\hat{\boldsymbol{A}}$ と $\boldsymbol{A}$ の間の平均 2 乗誤差を評価尺度とすることができる．もちろん，復元信号の順番と振幅に関する任意性を考慮する必要があるが，このためには，混合行列の推定値 $\hat{\boldsymbol{A}}$ の各列の順序と振幅に関する任意性を考慮すればよいことになる．ゆえに，mse を導出した場合と同じアナロジーで，$\boldsymbol{A}$，$\hat{\boldsymbol{A}}$ の i 行 j 列の要素をそれぞれ $a_{i,j}, \hat{a}_{i,j}$ とすると，

$$errA_{i,j} = \sum_{k=1}^{n} a_{k,i}^2 - \frac{\sum_{k=1}^{n} a_{k,i}\hat{a}_{k,j}}{\sum_{k=1}^{n} \hat{a}_{k,j}}$$

として，混合行列の推定値 $\hat{\boldsymbol{A}}$ の平均 2 乗誤差は，

$$mseA = \frac{1}{n}\sum_{i=1}^{n} \min_j errA_{i,j}$$

と求められる．

これに対し，Amari らは，もっと簡便に独立成分解析法の性能を評価する尺度：パフォーマンスインデックスを提案した．独立成分解析が成功した際は，復元行列 $\boldsymbol{W}$ と混合行列 $\boldsymbol{A}$ の積で与えられる行列は，各行，各列とも一つの要素しか非ゼロの値を持たないような行列となっている．パフォーマンスインデックスは，それからの誤差を測るようになっており，具体的には，行列 $\boldsymbol{WA}$ の i, j 要素を $p_{i,j}$ とすると，次式で与えられる．

$$PI = \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{|p_{i,j}|}{\max_k |p_{i,k}|} - 1\right) + \sum_{j=1}^{n}\left(\sum_{i=1}^{n}\frac{|p_{i,j}|}{\max_k |p_{k,j}|} - 1\right)$$

上記の mse，PI は，いずれも，独立成分解析法そのものの評価法であり，混合行列 $\boldsymbol{A}$ や原信号 $\boldsymbol{s}$ が既知の場合に用いるものである．では，実際に，未知のデータに対して独立成分解析を行っている場合，独立

成分解析が成功したか否かを判断するためにはどうしたらよいのだろうか？基本的には，復元信号 y_i の間の独立性を評価することになるが，そのためには，復元信号の確率密度分布を推定しなければならず，煩雑である．そこで，復元信号 y_i の間のクロスキュムラントを用いて評価することが便利であろう．例えば，復元信号 y_i がいずれも平均が 0 として，

$$cc = \sum_{i,j(i\neq j)} \left(c[y_i y_j^2]^2 + c[y_i y_j^3]^2 \right)$$

は，クロスキュムラントの性質より，復元信号が互いに独立ならば，ゼロになる．一方，ゼロであるからといって独立であるとはいえない．したがって，この cc を用いて独立成分解析の失敗を判定することは可能である．

[演習 3]　演習 2 で行った ICA において，学習の各反復での mse, $mseA$, cc の変化を描画してみよう．(ica_learn3.m, mseica.m, nor.m)

2　プリ・ホワイトニングを伴う学習高速化

2.1　プリ・ホワイトニングと歪対称化法

主成分解析では，与えられた確率変数ベクトル $\boldsymbol{x}$ に対して，$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x}$ として，確率変数ベクトル $\boldsymbol{y}$ の各要素を無相関にする行列 $\boldsymbol{P}$ を求めた（付録 A.5 参照）．実際，$\boldsymbol{x}$ の分散共分散行列 $E[\boldsymbol{x}\boldsymbol{x}^T]$ のすべての固有ベクトルを行ベクトルに持つ行列 $\boldsymbol{P}$ を選択すればよい．このとき，$\boldsymbol{y}$ の分散共分散行列 $E[\boldsymbol{y}\boldsymbol{y}^T]$ は，$E[\boldsymbol{x}\boldsymbol{x}^T]$ の固有値を対角要素に持つ対角行列 $\boldsymbol{\Lambda}$ となる．さて，ここで，

$$\tilde{\boldsymbol{x}} = \boldsymbol{P}^T \boldsymbol{\Lambda}^{-1/2} \boldsymbol{P}\boldsymbol{x}$$

なる変換を考える．ここで，$\boldsymbol{\Lambda}^{-1/2}$ は，行列 $\boldsymbol{\Lambda}$ の対角要素の $-1/2$ 乗をとった対角行列を表す．$\boldsymbol{P}E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{P}^T = \boldsymbol{\Lambda}$ であることと，$\boldsymbol{P}$ が直交行列であることを利用すれば，この変換により得られる $\tilde{\boldsymbol{x}}$ の分散共分散行列は，

$$\begin{aligned} E[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T] &= E[(\boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}\boldsymbol{x})(\boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}\boldsymbol{x})^T] \\ &= \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P} \\ &= \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P} \\ &= \boldsymbol{P}^T\boldsymbol{I}\boldsymbol{P} \\ &= \boldsymbol{I} \end{aligned} \tag{14}$$

と単位行列になる．ゆえに，こうした変換 $\tilde{\boldsymbol{x}} = \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}\boldsymbol{x}$ を白色化（ホワイトニング）という．

独立成分解析において，観測信号 $\boldsymbol{x}$ に対し，こうした変換を行って得られる白色化された $\tilde{\boldsymbol{x}}$ と原信号 $\boldsymbol{s}$ の関係は，

$$\tilde{\boldsymbol{x}} = \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}\boldsymbol{A}\boldsymbol{s}$$

である．ここで，

$$\tilde{\boldsymbol{A}} = \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1/2}\boldsymbol{P}\boldsymbol{A}$$

とおいて，

$$\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{A}}\boldsymbol{s}$$

とすると，

$$\begin{aligned} E[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T] &= E[(\tilde{\boldsymbol{A}}\boldsymbol{s})(\tilde{\boldsymbol{A}}\boldsymbol{s})^T] \\ &= \tilde{\boldsymbol{A}}E[\boldsymbol{s}\boldsymbol{s}^T]\tilde{\boldsymbol{A}}^T \end{aligned}$$

となる．白色化された $\tilde{\boldsymbol{x}}$ の分散共分散行列 $E[\tilde{\boldsymbol{x}}\tilde{\boldsymbol{x}}^T]$ は単位行列であり，原信号 $\boldsymbol{s}$ の分散共分散行列 $E[\boldsymbol{s}\boldsymbol{s}^T]$ は対角行列であることを考えれば，行列 $\tilde{\boldsymbol{A}}$ は，正規化されていない直交行列ということになる．

前述したように復元信号の振幅は不定であるから，観測信号 $\boldsymbol{x}$ を事前に白色化し，白色化された $\tilde{\boldsymbol{x}}$ に対して独立成分解析を行った場合，復元行列は，直交行列の中から探せばよいことになる．白色化していない場合，復元行列を正則な正方行列の中から探さなければならないことを考えれば，探索範囲が大きく限定されたことになる．これにより，学習が高速化されることが期待される．独立成分解析におけるこうした白色化をプリ・ホワイトニングと呼ぶ．

それでは，具体的に，直交行列の中から復元行列 $\boldsymbol{W}$ を探すためには，どのような更新則を採用したらよいのだろうか？式 (10) で示される更新則は，$\boldsymbol{D} = \boldsymbol{I} + E[\boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T]$ などとおけば，

$$\boldsymbol{W} \leftarrow (\boldsymbol{I} + \eta\boldsymbol{D})\boldsymbol{W}$$

のような形になっている．そこで，$\boldsymbol{W}$ が直交行列であるとして，更新後の $(\boldsymbol{I} + \eta\boldsymbol{D})\boldsymbol{W}$ もまた直交行列となるための $\boldsymbol{D}$ の条件を求めよう．直交行列であるためには，

$$((\boldsymbol{I} + \eta\boldsymbol{D})\boldsymbol{W})((\boldsymbol{I} + \eta\boldsymbol{D})\boldsymbol{W})^T = \boldsymbol{I}$$

となっていなければならないので，左辺を展開すると，

$$\boldsymbol{I} + \eta\boldsymbol{D} + \eta\boldsymbol{D}^T + \eta^2\boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{I}$$

となる．学習率 η が十分小さいものとすれば，η^2 を含む項は無視できるので，最終的に，$\boldsymbol{D}$ は，条件

$$\boldsymbol{D} + \boldsymbol{D}^T = \boldsymbol{0}$$

を満たしていないといけないことになる．適当な行列 $\boldsymbol{B}$ を用いて $\boldsymbol{D} = \boldsymbol{B} - \boldsymbol{B}^T$ となるならば，$\boldsymbol{D} + \boldsymbol{D}^T = \boldsymbol{0}$ を満たすことは容易にわかる．そこで，

$$\boldsymbol{B} = \boldsymbol{I} + E[\boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T]$$

とすれば，

$$\boldsymbol{D} = \boldsymbol{B} - \boldsymbol{B}^T = \boldsymbol{I} + E[\boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T] - \boldsymbol{I}^T - E[\boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T]^T = E[\boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T] - E[\boldsymbol{y}\boldsymbol{\phi}(\boldsymbol{y})^T]$$

となるので，独立成分解析の更新則は，

$$\begin{aligned} \boldsymbol{W} &\leftarrow \boldsymbol{W} + \eta\Delta\boldsymbol{W} \\ \Delta\boldsymbol{W} &= \big(E[\boldsymbol{\phi}(\boldsymbol{y})\boldsymbol{y}^T] - E[\boldsymbol{y}\boldsymbol{\phi}(\boldsymbol{y})^T]\big)\boldsymbol{W} \end{aligned} \tag{15}$$

となる．もちろん，実際に有限個の標本 $\boldsymbol{x}(t),\ t = 1, \ldots, T$ を用いる場合には，期待値を標本平均に置き換えなければならない．

[演習 4]　原信号として，s_1 を一様分布に従う信号，s_2 を正規信号の 3 乗を設定し，式 (15) に基づいて，プリ・ホワイトニング付きの独立成分解析を行ってみよう．その収束の速さと，プリ・ホワイトニングを行わない場合の収束の速さを比較してみよう．ただし，$\eta = 1.0$ 位に設定し，それ以外は，演習 1,2 と同じ条件にしておこう．(ica_learn4.m, prewhitening.m)

2.2 射影追跡法

式 (11) で示した復元信号間の平均相互情報量 $\bar{I}(\boldsymbol{y})$ を，プリ・ホワイトニングが行われたという条件の下で，もう一度，見てみよう．プリ・ホワイトニングが行われているので，復元行列 $\boldsymbol{W}$ は，直交行列に限定することができる．直交行列の場合，その行列式は 1 であるから，平均相互情報量は，

$$\bar{I}(\boldsymbol{y}) = \sum_{k=1}^{n} H(y_k) + \text{Constant} \tag{16}$$

と書ける．ただし，Constant は，復元行列 $\boldsymbol{W}$ に依存しない定数である．したがって，平均相互情報量 $\bar{I}(\boldsymbol{y})$ を最小化することは，各復元信号のエントロピーの総和を最小化することと等価である．そこで，直交な復元行列を正射影方向成分にわけて $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$ と表現し，まずその中の一つの復元信号 $y_1 = \boldsymbol{w}_1^T \boldsymbol{x}$ に着目し，そのエントロピー $H(y_1)$ を最小にする射影方向 $\boldsymbol{w}_1$ を探す問題を考える．

ところで，射影追跡法とは，多次元データ，つまり，多次元空間の中に分布しているデータにおいて，興味ある軸（1 次元線形部分空間）を見つける方法である．主成分解析は，分散を最大にする軸を見つける射影追跡法と位置づけられる．また，興味ある軸をエントロピーを最小化する軸と考えれば，上記の y_1 を見つける問題は，射影追跡法の一つであるといえる．

射影追跡法により，y_1 のエントロピーを最小化する $\boldsymbol{w}_1$ を求めるためには，y_1 のエントロピーを算出するために，その確率密度分布 $p_{y_1}(y_1)$ を求める必要がある．これは，一般に計算コストがかかる．そこで，エントロピーの近似式

$$H(y_1) \simeq -\frac{1}{12} E[y_1^3]^2 - \frac{1}{48} c[y_1^4]^2$$

を利用しよう．これは，正規分布に従う時には 0 となり，非正規分布に対しては 0 よりも小さい値をとる．上式は，キュムラント $c[y_1^4]$ をモーメントを用いて書き下せば，

$$H(y_1) \simeq \frac{1}{48}\Big(-E[y_1^4]^2 - 4E[y_1^3]^2 + 6E[y_1^4]E[y_1^2]^2 - 9E[y_1^2]^4\Big)$$

となる．$H(y_1)$ を $\boldsymbol{w}_1$ で偏微分することにより，以下の最急勾配法による更新則が導かれる．

$$\begin{aligned} \boldsymbol{w}_1 &\leftarrow \text{normal}\big(\boldsymbol{w}_1 - \eta \cdot \text{normal}(\Delta \boldsymbol{w}_1)\big), \\ &\text{where,} \\ \Delta \boldsymbol{w}_1 &= -\frac{1}{6} E[y_1^4] E[\boldsymbol{x} y_1^3] - \frac{1}{2} E[y_1^3] E[\boldsymbol{x} y_1^2] + \frac{1}{2} E[\boldsymbol{x} y_1^3] - \frac{3}{2} \boldsymbol{w}_1 + \frac{1}{2} E[y_1^4] \boldsymbol{w}_1 \end{aligned} \tag{17}$$

ただし，上式において，normal(·) は，ベクトル · のノルムを 1 に規格化することを意味する．

式 (17) は，エントロピーを最小化する一つの軸 $\boldsymbol{w}_1$ を求めるものであったが，$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ を用意し，この式にしたがって同時に更新させれば，いずれも，エントロピーを最小化する軸を獲得することになる．ただし，この場合，同じ軸を獲得しても意味がないし，観測信号にプリホワイトニングを行っていれば，互いに直交していなければならない．そこで，$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ を互いに直交するように更新するようにすれば，$\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$ として，以下の更新則が導かれる．

$$\begin{aligned} \boldsymbol{W} &\leftarrow \text{orthog}(\boldsymbol{W} - \eta \cdot \text{normal}(\Delta \boldsymbol{W})), \\ \Delta \boldsymbol{W} &= -\frac{1}{6} \text{diag}(E[\boldsymbol{y}^4]) E[\boldsymbol{y}^3 \boldsymbol{x}^T] - \frac{1}{2} \text{diag}(E[\boldsymbol{y}^3]) E[\boldsymbol{y}^2 \boldsymbol{x}^T] \\ &\quad + \frac{1}{2} E[\boldsymbol{y}^3 \boldsymbol{x}^T] - \frac{3}{2} \boldsymbol{W} + \frac{1}{2} \text{diag}(E[\boldsymbol{y}^4]) \boldsymbol{W} \end{aligned} \tag{18}$$

ただし，上式において，ベクトルに対する n 乗の記号は，ベクトルの各要素に対する n 乗を意味するものとする．また，normal(·) は，行列 · の各行ベクトルのノルムを 1 に規格化したもの，orthog(·) は，· の各行ベクトルを直交化した直交行列を表すものとする．

[**演習 5**]　原信号として，s_1 を一様分布に従う信号，s_2 を正規信号の 3 乗を設定し，式 (18) に基づいて，プリ・ホワイトニング付きの独立成分解析を行ってみよう．ただし，$\eta = 1.0$ 位に設定し，それ以外の条件は，これまでの演習と同じにしておこう．(ica_learn5.m, prewhitening.m)

2.3 Fast ICA

一度，射影追跡法に戻り，$y_1 = \boldsymbol{w}_1^T \boldsymbol{x}$ のエントロピー $E[q_1(y_1)]$ を最小化する $\boldsymbol{w}_1$ を $\|\boldsymbol{w}_1\| = 1$ の条件下で解いてみよう．ただし，$q_1(y_1)$ は，y_1 の確率密度分布である．ラグランジェの未定乗数法により

$$J = E[q_1(y_1)] - \frac{\lambda}{2}(\boldsymbol{w}_1^T \boldsymbol{w}_1 - 1)$$

を評価関数とし，これを $\boldsymbol{w}_1$ で偏微分すると，

$$\frac{\partial J}{\partial \boldsymbol{w}_1} = E[\boldsymbol{x} q_1'(y_1)] - \lambda \boldsymbol{w}_1$$

となる．これをゼロベクトルとおいた方程式

$$E[\boldsymbol{x} q_1'(y_1)] = \lambda \boldsymbol{w}_1$$

の解を求めればよいことになる．この方程式の両辺に左から $\boldsymbol{w}_1^T$ を乗ずれば，

$$\lambda = E[y_1 q_1'(y_1)]$$

となるので，結局，$\boldsymbol{w}_1$ は，

$$\boldsymbol{w}_1 = \frac{E[\boldsymbol{x} q_1'(y_1)]}{E[y_1 q_1'(y_1)]}$$

として求められることになる．しかし，y_1 の確率密度分布 $q_1(y_1)$ は，$\boldsymbol{w}_1$ に依存しているので，上式で与えられる $\boldsymbol{w}_1$ はその近傍での解に過ぎない．そこで，

$$\boldsymbol{w}_1 \leftarrow \frac{E[\boldsymbol{x} q_1'(y_1)]}{E[y_1 q_1'(y_1)]} \tag{19}$$

とした更新則による学習を行わなければならないことになる．

次に，$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ を用意し，式 (19) にしたがって同時に更新し，かつ直交化することを考える．この場合，式 (19) は，$\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$ とし，また，後で $\boldsymbol{W}$ を直交化するならば，

$$\boldsymbol{W} \leftarrow \mathrm{orthog}\Big(\boldsymbol{\Lambda}^{-1} E[\boldsymbol{q}'(\boldsymbol{y}) \boldsymbol{x}^T]\Big) \tag{20}$$

と書き直せる．ただし，$\boldsymbol{q}'(\boldsymbol{y}) = (q_1'(y_1), \ldots, q_n'(y_n))^T$，$\boldsymbol{\Lambda}$ は，$\lambda_i = E[y_i q_i'(y_i)],\ i = 1, \ldots, n$ を対角要素に持つ対角行列を表す．また，$\mathrm{orthog}(\cdot)$ は，$\cdot$ の各行ベクトルを直交化した直交行列を表すものとする．

次に，別のアプローチ，つまり前述した方程式

$$E[\boldsymbol{x} q_1'(y_1)] = \lambda \boldsymbol{w}_1$$

をニュートン法により解くことを考えよう．ニュートン法は，$f(x) = 0$ の解を，$x \leftarrow x - f(x)/f'(x)$ とした更新則により解くものであるから，方程式

$$\boldsymbol{f}(\boldsymbol{x}) = E[\boldsymbol{x} q_1'(y_1)] - \lambda \boldsymbol{w}_1 = \boldsymbol{0} \tag{21}$$

に適用すると，

$$\boldsymbol{w}_1 \leftarrow \boldsymbol{w}_1 - \boldsymbol{f}'(\boldsymbol{x})^{-1} \boldsymbol{f}(\boldsymbol{x})$$

となる．

$$\begin{aligned} \boldsymbol{f}'(\boldsymbol{x}) &= E[\boldsymbol{x}\boldsymbol{x}^T q_1''(y_1)] - \boldsymbol{I}\lambda \\ &\simeq E[\boldsymbol{x}\boldsymbol{x}^T]E[q_1''(y_1)] - \boldsymbol{I}\lambda \\ &\simeq \boldsymbol{I}(E[q_1''(y_1)] - \lambda) \end{aligned}$$

より，更新則

$$\boldsymbol{w}_1 \leftarrow \boldsymbol{w}_1 - \frac{E[\boldsymbol{x}q_1'(y_1)] - \lambda\boldsymbol{w}_1}{E[q_1''(y_1)] - \lambda} \tag{22}$$

が得られる．更新式 (22) の代わりに，その式の両辺に $\lambda - E[q_1''(y_1)]$ を乗じ，また，$\boldsymbol{w}_1$ は，後でノルムを 1 に規格化するものとすれば，以下の別な更新則が得られる．

$$\boldsymbol{w}_1 \leftarrow \mathrm{normal}\Big(E[\boldsymbol{x}q_1'(y_1)] - \boldsymbol{w}_1 E[q_1''(y_1)]\Big) \tag{23}$$

式 (22),(23) は，ニュートン法を用いており，最急勾配法よりも収束が早いことから，Fast ICA（の 1 次元バージョン）と呼ばれている．

次に，$\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ を用意し，式 (22)，あるいは式 (23) にしたがって同時に更新し，かつ直交化することを考える．式 (22) は，$\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^T$ とし，また，後で $\boldsymbol{W}$ を直交化するならば，

$$\boldsymbol{W} \leftarrow \mathrm{orthog}\Big(\boldsymbol{W} + \boldsymbol{D}(\boldsymbol{\Lambda} - E[\boldsymbol{p}'(\boldsymbol{y})\boldsymbol{y}^T])\boldsymbol{W}\Big) \tag{24}$$

となる．ここで，$\lambda_i = E[y_i q_i'(y_i)],\ i = 1, \ldots, n$ として，$\boldsymbol{D}$ は，$(E[q_i''(y_i)] - \lambda_i)^{-1},\ i = 1, \ldots, n$ を対角要素に持つ対角行列，$\boldsymbol{\Lambda}$ は，$\lambda_i,\ i = 1, \ldots, n$ を対角要素に持つ対角行列，$\boldsymbol{p}'(\boldsymbol{y}) = (q_1'(y_1), \ldots, q_n'(y_n))^T$ を表す．式 (24) で示される更新則による ICA を Fast ICA といい，現在，最もよく利用される独立成分解析法の一つとなっている．

[**演習 6**]　原信号として，s_1 を一様分布に従う信号，s_2 を正規信号の 3 乗を設定し，式 (24) に基づいて，プリ・ホワイトニング付きの独立成分解析を行ってみよう．ただし，$q_i(y_i),\ i = 1, 2$ を y_i^4 で代用しよう．それ以外の条件は，これまでの演習と同じにするものとする．(ica_learn6.m, prewhitening.m)

[**演習 7**]　Web から適当な画像を 3 枚取得し，白黒 8bit 階調，同サイズの画像（128 × 128[pixel] 程度）にした後，それらを，それぞれ適当な方法で走査してベクトルにし，これらを原信号 $s_i,\ i = 1, 2, 3$ とする．この原信号を乱数により作られた適当な正則な混合行列 $\boldsymbol{A}$ により線形変換して観測信号 $x_i,\ i = 1, 2, 3$ とする．この観測信号に対して，様々な手法で独立成分解析を行い，元の原画像が復元される様子を確認しよう．(ica_image1.m, ica_image2.m, ica_image3.m, nor.m, prewhitening.m)

2.4　非正規性最大化法

演習 6 において，$p_{y_i}(y_i) = y_i^4,\ i = 1, \ldots, n$ とおき，結果として上手く独立成分解析を行えることを確認してきた．では，なぜ，$p_{y_i}(y_i) = y_i^4$ として上手く行くのだろうか？また，以前にも，原信号や復元信号の確率密度分布の対数の微分 $\phi_i(y_i)$ を $-y_i^3$，$-y_i - \tanh(y_i)$，$-y_i + \tanh(y_i)$ などとおいて上手く行くことを見てきた．なぜだろうか？本節では，このことについて考えてみよう．

Fast ICA の基礎となっている射影追跡法の節で述べたように，プリ・ホワイトニングが行われた場合には，独立成分解析は，各復元信号のエントロピーの総和

$$J = \sum_{i=1}^{n} H(y_i)$$

を最小化する問題に帰着された．各復元信号のエントロピーはそれぞれの確率密度分布の期待値であるから，上式は，

$$J = \sum_{i=1}^{n} E[q_i(y_i)]$$

とも書ける．Fast ICA では，この確率密度分布 $q_i(y_i)$ を y_i^4 とおき，$E[q_i(y_i)] = E[y_i^4]$ の極値を与えるような直交変換 $\boldsymbol{W}$ を導いたのである．分散が 1 の下での 4 次中心モーメント $E[y_i^4]$ は，正規分布に対しては 3，それ以外の分布の時には，3 よりも大きいか，小さい値をとり，一般に，正規分布から離れるほど，3 から離れた値をとる．また，中心極限定理によって，独立な複数の信号を混合すれば，その分布はより正規分布に近づく，つまり，4 次中心モーメントが 3 に近づくことになる．つまり，$q_i(y_i)$ を y_i^4 とおいた Fast ICA は，正規分布から離れた分布を持つ軸，言い換えると非正規性を最大にする軸を探索していることに他ならない．このように，独立成分解析を非正規性を最大するという立場から見ると，必ずしも 4 次モーメントを用いる必要はないように思える．実際，$\dfrac{1}{a_1}\log\cosh a_1 y_1,\ 1 \le a_1 \le 2$，$-\exp(-y_1^2/2)$ などとしても上手くゆく．

また，Fast ICA の更新則：式 (24) は，$\boldsymbol{D}$ を η を対角部分に持つ対角行列，$\boldsymbol{\Lambda} = \boldsymbol{I}$，$\boldsymbol{p}'(\boldsymbol{y}) = -\boldsymbol{\phi}(\boldsymbol{y})$ として，また，直交化を外せば，式 (10) と全く同じになる．$q_i(y_i) = y_i^4$ なので，$q_i'(y_i) = 4y_i^3$，一方，$\phi_i(y_i) = -y_i^3$ なので，式 (24) と式 (10) は，結局，同じ関数を用いていることになる．しかし，式 (10) では，極値を探索するのではなく，評価関数を最大化する方向に学習を進める．原信号の分布が Super-Gaussian と Sub-Gaussian により，評価関数の大小が反転するため，これらの分布で，学習する方向が逆になる．そのため，これらの分布に応じて，異なる代用関数を用いなければならなかったのである．これに対し，FastICA の更新則，すなわち式 (24) では，4 次キュムラントを推定し，Super-，あるいは Sub-Gaussian により学習の方向を変える働きを対角行列 $\boldsymbol{D}$ が担っていると考えることができる．

3 BSD

3.1 問題設定

1, 2 章で述べてきた BSS(Blind Source Separation) では，n 個の原信号 $s_i(t),\ i = 1, \ldots, n$ が，各時刻で $a_{i,j}$ 倍されて混合され，

$$x_i(t) = \sum_{j=1}^{n} a_{i,j} s_j(t), \quad i = 1, \ldots, m \tag{25}$$

として観測される場合を対象にした．しかし，実際には，信号源からセンサまでの途中経路で，周波数毎に異なる減衰比で減衰する，あるいは信号源とセンサ間の距離に応じて到達時間が遅れる，あるいは原信号が何かに反響してこだまのような信号に変化するなど，式 (25) で表現できない場合も多い．むしろ，式 (25) で表現できるのは，理想化された特殊なケースだけであると言っても良い．原信号がセンサで記録されるまでに受ける変換は，それが前述したような線形変換であるならば，式 (25) に畳み込み演算を含めた次式により一般的に表現できる．

$$x_i(t) = \sum_{j=1}^{n} \sum_{\tau=0}^{T} a_{i,j}(\tau) s_j(t-\tau), \quad i = 1, \ldots, m \tag{26}$$

原信号 $s_i(t),\ i = 1, \ldots, , n$ が空間的位置が異なる信号であるとすると，式 (26) で表される変換は，時間的にも空間的に混合する変換であるといえる．こうした立場から，式 (26) で表現される観測信号 $x_i(t),\ i = 1, \ldots, m$ を時空間混合信号と呼ぶことがある．そして，時空間混合信号 $x_i(t),\ i = 1, \ldots, m$ から，原信号 $s_i(t),\ i =$

$1,\ldots,n$ を次式により $y_i(t),\ i=1,\ldots,n$ として復元する問題を BSD(Blind Source Decomposition) という．

$$y_i(t)=\sum_{j=1}^{m}\sum_{\tau=0}^{L}w_{i,j}(\tau)x_j(t-\tau),\quad i=1,\ldots,n \tag{27}$$

BSD においては，各原信号 $s_i(t),\ i=1,\ldots,n$ が互いに独立であるだけでなく，各信号の各時刻での値が独立であり，一つの確率密度分布に従って生成されていることが要求される．つまり，各時刻 t で $y_j(t)$ が独立になるように，$w_{i,j}(\tau)$ を決める問題ということになる．こうした処理をデコンボリューション，あるいは白色化という．BSD は，ICA の一つに位置づけられるが，信号処理における古くからの概念である白色化フィルタ（復元フィルタ）とも関係が深い．

3.2 z 変換を用いた表現

式 (27) を z 変換を用いて表現してみよう．$x(t),\ t=0,1,\ldots$ に対する片側 z 変換は，

$$X(z)=\sum_{t=0}^{\infty}x(t)z^{-t}$$

と定義されている．$x(t)$ と $y(t)$ の畳み込み演算は，z 変換を利用すると，$X(z)Y(z)$ と積で表現することができる．さて，$y_i(t)$，$x_i(t)$，$w_{i,j}(t)$ の z 変換を，それぞれ $Y_i(z)$，$X_i(z)$，$W_{i,j}(z)$ とし，これらをベクトルや行列を用いて，

$$\boldsymbol{Y}(z)=\begin{pmatrix}Y_1(z)\\Y_2(z)\\\vdots\\Y_n(z)\end{pmatrix},\quad \boldsymbol{X}(z)=\begin{pmatrix}X_1(z)\\X_2(z)\\\vdots\\X_m(z)\end{pmatrix},\quad \boldsymbol{W}(z)=\begin{pmatrix}W_{1,1}(z)&W_{1,2}(z)&\cdots&W_{1,m}(z)\\W_{2,1}(z)&W_{2,2}(z)&\cdots&W_{2,m}(z)\\\vdots&\vdots&\ddots&\vdots\\W_{n,1}(z)&W_{n,2}(z)&\cdots&W_{n,m}(z)\end{pmatrix}$$

と表せば，式 (27) は，

$$\boldsymbol{Y}(z)=\boldsymbol{W}(z)\boldsymbol{X}(z) \tag{28}$$

と行列やベクトルの積を用いて書き直せる．ただし，行列やベクトルの要素は，いずれも z の負のべきの多項式である．

同様に，$s_i(t)$，$a_{i,j}(t)$ を z 変換を用いて表現すれば，式 (27) は，

$$\boldsymbol{Y}(z)=\boldsymbol{W}(z)\boldsymbol{A}(z)\boldsymbol{S}(z) \tag{29}$$

と書くことができる．ゆえに，行列 $\boldsymbol{W}(z)\boldsymbol{A}(z)$ が単位行列となっていれば復元信号 $\boldsymbol{Y}(z)$ は原信号 $\boldsymbol{S}(z)$ に一致することが分かる．ただし，BSS と同様に，$y_i(t),\ i=1,\ldots,n$ の振幅と順番に関し不定性がある．更に，信号処理の立場から，$A_{i,j}(z)$ が最小位相性[2]を満足していないと，$W_{i,j}(z)$ が存在しないことが分かる．この場合でも，$y_i(t)$ が $s_i(t)$ に対して時間遅れがあることを許容すれば，復元することが可能になる．つまり，$\boldsymbol{P}$ を任意の並び替え行列，$q_i,\ i=1,\ldots,n$ を振幅の不定性を表すスカラーとして，

$$\boldsymbol{W}(z)\boldsymbol{A}(z)=\boldsymbol{P}\ \mathrm{diag}(q_1z^{-\tau_1},\ldots,q_nz^{-\tau_n})$$

と書ければ，解であるということができる．

また，$\boldsymbol{A}(z)$ の次数 T が有限であっても，$\boldsymbol{W}(z)$ の次数は，一般に無限になる．したがって，実際には，有限の長さ L で打ち切らざるを得ない．したがって，あくまでも近似的にしか求められないことになる．また，BSS と同様，原信号の数 n よりも観測信号の数 m の方が多いか等しくなければならない．ここでは，簡単化のため，$m=n$ を仮定する．

さて，次に，具体的に $W(z)$ を推定する手法について説明しよう．

[2] $A_{i,j}(z)=0$ とおいた多項式のすべての根が複素平面における原点を中心とする単位円の内部にある．

3.3 周波数領域での BSD

$x_i(t)$ が，$t = 1, 2, \ldots, N$ で定義されているものとし，式 (28) において，$z = \mathrm{e}^{i2\pi k/N}$ とおけば[3]，離散フーリェ変換の像の間の関係

$$\boldsymbol{Y}(k) = \boldsymbol{W}(k)\boldsymbol{X}(k), \qquad k = 1, \ldots, N$$

が得られる．したがって，これまでに述べてきた BSS を，$k = 1, \ldots, N$ ごとに行うことにより，復元行列 $\boldsymbol{W}(k),\ k = 1, \ldots, N$ が得られることになる．実際には，$x_i(t),\ t = 1, \ldots, N$ を長さ L の複数のブロックに区切り，それぞれを離散フーリェ変換し，各ブロックを標本とみなして BSS を行うことになる．こうした手法は，周波数領域での BSD と言われる．次に，周波数領域での BSD の問題点を考えてみよう．

各 k で，BSS を用いて $\boldsymbol{X}(k)$ に対する復元行列 $\boldsymbol{W}(k)$，復元信号 $\boldsymbol{Y}(k)$ を求める際に，順序と振幅の不定性が存在する．一般的な BSS では，大きな問題とはならなかったが，この場合，得られた復元行列 $\boldsymbol{W}(k)$，復元信号 $\boldsymbol{Y}(k)$ に対して，逆離散フーリェ変換することにより，復元行列 $\boldsymbol{W}(z)$ や復元信号 $\boldsymbol{Y}(z)$ の時間軸表現を得るため，順序，振幅をあわせないと，時間軸での表現は真とは全く異なったものになってしまう．そのため，隣り合う k で $\boldsymbol{Y}(k)$ に相関がある，あるいは k の変化に対して $\boldsymbol{Y}(k)$ が滑らかに変化するなどの仮定を設けて，順序や振幅を揃えることが行われる．したがって，こうした手法を適用する場合には，こうした仮定が含まれていることに注意しておく必要がある．

次に，$x_i(t)$ の各ブロックに対して離散フーリェ変換すると，その像は，加算平均を取った形になるため，正規分布に従いやすくなることに注意する必要がある．BSS は，正規分布に従う場合，適用できないから，正規分布に近づくと原信号の推定精度が悪化することになる．ブロック長を L に設定すると，同じ長さ，つまりその次数までの $\boldsymbol{W}(z)$ が求められる．前述したように，$\boldsymbol{W}(z)$ は，無限の次数を持つから，近似精度の意味で，その打ち切り次数 L は大きい方が望ましい．一方で，L を長く設定すると，離散フーリェ変換の加算平均効果によって正規分布近づき，BSS の推定精度が悪くなる．つまり，周波数軸での BSD では，$\boldsymbol{W}(z)$ の次数がごく短いときにしか実用にならないことになる．

3.4 時間領域での BSD

BSD は，原信号が，互いに独立で，各時刻でも独立であることを利用して，BSS の更新則を導いた際とほぼ同様なアプローチで，時間領域での更新則を導くことができる．ただし，原信号は各時刻でも独立なので，原信号の確率密度分布は，

$$r(\{\boldsymbol{s}(t)\}) = \prod_{i=1}^{n}\prod_{t=0}^{L} r_i(s_i(t)) \tag{30}$$

とおくことになる．復元行列 $\boldsymbol{W}(\tau)$ を

$$\boldsymbol{W}(\tau) = \begin{pmatrix} w_{1,1}(\tau) & w_{1,2}(\tau) & \cdots & w_{1,n}(\tau) \\ w_{2,1}(\tau) & w_{2,2}(\tau) & \cdots & w_{2,n}(\tau) \\ \vdots & \vdots & \ddots & \vdots \\ w_{n,1}(\tau) & w_{n,2}(\tau) & \cdots & w_{n,n}(\tau) \end{pmatrix}, \qquad \tau = 0, 1, \ldots, L$$

とおけば，最終的に，更新則は，

$$\begin{aligned} \boldsymbol{W}(\tau) &\leftarrow \boldsymbol{W}(\tau) + \Delta\boldsymbol{W}(\tau) \\ \Delta\boldsymbol{W}(\tau) &= \eta \sum_{\tau'=0}^{\tau} \{\delta(\tau')\boldsymbol{I} - \boldsymbol{R}(\tau')\}\boldsymbol{W}(\tau - \tau') \\ & \qquad \tau = 0, 1, \ldots, L \end{aligned} \tag{31}$$

[3]指数の肩の i は虚数単位を表す．

となる．ここで，$\delta(\tau')$ は，$\tau'=0$ で 1，$\tau'\neq 0$ で 0 となる関数を表わし，$\boldsymbol{I}$ は，n 行 n 列の単位行列を表す．また，$\boldsymbol{R}(\tau')$ は，

$$\boldsymbol{R}(\tau') = E[\boldsymbol{\varphi}(\boldsymbol{y}(t))\boldsymbol{y}^T(t-\tau')] \tag{32}$$

を表すが，エルゴード性を仮定して，時間平均

$$\boldsymbol{R}(\tau') = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\varphi}(\boldsymbol{y}(t))\boldsymbol{y}^T(t-\tau') \tag{33}$$

として求められる．

[**演習 8**] 適当に時空間混合信号を作り，式 (31) に基づいて BSD を行ってみよう．(tdica1.m, nor.m)

4 ICA の推定精度

4.1 ICA とセミパラメトリック推定, 推定関数法

ところで，これまで紹介してきた ICA の更新則は，いずれも，与えられた観測信号 $\boldsymbol{x}$ に対し，

$$E_{A_0,r_0}[\boldsymbol{F}(\boldsymbol{x},\boldsymbol{A})] = \boldsymbol{0} \tag{34}$$

なる連立方程式を満足する行列 $\boldsymbol{A}$ を混合行列の推定値としていることに気がついたであろうか？例えば，最尤推定法（平均相互情報量最小化法）などは，適当な非線形ベクトル関数 $\boldsymbol{\varphi}(\cdot)$ を用いて，

$$\boldsymbol{F}(\boldsymbol{x},\boldsymbol{A}) = (\boldsymbol{I}+\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T)\boldsymbol{W} \tag{35}$$

としている．ここで，$\boldsymbol{F}$ は行列関数を意味する．また，$E_{A_0,r_0}[\cdot]$ は，$\boldsymbol{x}$ に関する $\cdot$ の期待値であるが，$\boldsymbol{x}$ は真値 $\boldsymbol{A}_0, r_0(\boldsymbol{s})$ に依存するので，正確には $\boldsymbol{A}_0, r_0(\boldsymbol{s})$ を条件とする $\boldsymbol{x}$ に関する条件付期待値となる．実際の問題では，期待値は，n 個の標本 $\boldsymbol{x}(t),\ t=1,\ldots,T$ に対する標本平均

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{F}(\boldsymbol{x}(t),\boldsymbol{A}) = \boldsymbol{0} \tag{36}$$

で代用される．本書で紹介した以外にも，様々な関数 $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{A})$ が発見的に見出され，学習則に採用されている．式 (34) が，任意の真値 $\boldsymbol{A}_0, r_0$ に対して成立するなど，他にもいくつかの条件を満たさないといけないが，こうした関数 $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{A})$ を推定関数といい，推定関数を用いた推定を推定関数法という（付録 A.8 参照）．

最尤推定法（平均相互情報量最小化法）における推定関数，つまり式 (35) がゼロ行列に等しくなる条件を考えよう．この式の期待値は，

$$E_{A_0,r_0}[\boldsymbol{F}(\boldsymbol{x},\boldsymbol{A})] = (\boldsymbol{I}+E_{A_0,r_0}[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T])\boldsymbol{W}$$

であるが，この式がゼロ行列になるためには，復元行列 $\boldsymbol{W}$ か，あるいは右辺の括弧内がゼロ行列にならなければならない．復元行列 $\boldsymbol{W}$ がゼロ行列の場合，正則な行列ではないので，その逆行列として混合行列 $\boldsymbol{A}$ を推定できないので明らかに意味のない解である．したがって，$\boldsymbol{I}+E_{A_0,r_0}[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T]$ がゼロ行列となる場合のみを考えればよい．この非対角部分は，復元信号 $\boldsymbol{y}$ の平均がゼロベクトルなので，$\boldsymbol{y}$ の各要素が独立になったとき，$E_{A_0,r_0}[\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T] = E_{A_0,r_0}[\boldsymbol{\varphi}(\boldsymbol{y})]E_{A_0,r_0}[\boldsymbol{y}]^T = \boldsymbol{0}$ となる．一方，対角部分は，各復元信号 y_i が $E_{A_0,r_0}[\varphi(y_i)y_i] = -1$ を満たすような振幅になったとき，例えば，$\varphi(y_i) = -y_i^3$ を選択した場合，$E_{A_0,r_0}[y_i^4] = 1$ を満足するような振幅になったとき，ゼロになる．つまり，最尤推定法（平均相互情報量最

小化法）では，各復元信号が互いに独立になり，適当な振幅になるように，復元行列 $\boldsymbol{W}$ が定められることになる．前述したように，振幅については，不定なので，適当な振幅の値に推定されてもしょうがない．

ここで，ICA に利用可能な推定関数 $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{A})$ には，他にどのようなものがあるのであろうか？そして，ある推定関数 $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{A})$ が与えられたとき，これを用いたときの混合行列 $\boldsymbol{A}$ の推定精度はどの程度であり，どのような推定関数を用いたとき，推定精度を最大にするのか？という疑問が沸く．そして，推定精度を最大にする推定関数を用いた推定関数法は，Cramér-Rao の下界である推定精度の限界を達成しているのであろうか？ということにも興味があるであろう．

ICA においては，混合行列 $\boldsymbol{A}$ を推定することが目的であるが，その背後に，未知である原信号 $\boldsymbol{s} = (s_1, \ldots, s_n)^T$ の確率密度分布 $r(\boldsymbol{s}) = r_1(s_1) \cdots r_n(s_n)$ の推定問題を含んでいる．この場合，$\boldsymbol{A}$ に対して，$r(\boldsymbol{s})$ を余計なパラメータ (nuisance parameter) という．nuisance パラメータを含む一般的なパラメータ推定における推定精度については，付録 A.7 に簡単にまとめた．ICA の場合，付録 A.7 におけるパラメータ θ が行列 $\boldsymbol{A}$ に，パラメータ ξ が原信号の確率密度分布 $r(\boldsymbol{s})$ に対応する．ただし，ICA の場合，推定したいパラメータである混合行列 $\boldsymbol{A}$ は n^2 次元パラメータであるが，nuisance パラメータである確率密度分布 $r(\boldsymbol{s})$ は，無限次元のパラメータである点が問題となる．無限次元の nuisance パラメータを含むパラメータ推定をセミパラメトリック推定という．セミパラメトリック推定では，nuisance パラメータを含むフィッシャーの情報行列が無限次元になるため，推定精度の限界，つまり Cramér-Rao の下界を直接求めることができない．これに対し，Amari ら[4]は，情報幾何学の立場から，セミパラメトリック推定における推定関数，推定精度の問題に挑んだ（付録 A.7 参照）．統計量の推定精度の限界である Cramér-Rao の下界，スコア関数，推定関数法，情報幾何学についての知識が全くない場合には，事前に，付録 A.7，A.8 を読んでおこう．

4.2 r-スコア関数, nuisance 接空間

nuisance パラメータ $r(\boldsymbol{s})$ に対するスコア関数，つまり r-スコア関数を求めよう．nuisance パラメータ $r(\boldsymbol{s})$ が無限次元パラメータなので，$v(x; \boldsymbol{A}, r) = \frac{\partial}{\partial r} \log p(x; \boldsymbol{A}, r)$ などとすることができない．そこで，まず，関数空間の中において，原信号の確率密度分布 $r(\boldsymbol{s})$ がとりうる空間（多様体）を考え，その中のある点から確率密度分布が微小に変化することにより表現可能な確率密度分布の集合，つまり，その点における接空間を考える．一つの原信号の確率密度分布が $r_i(s_i)$ から

$$r_i(s_i; \alpha_i, \delta) = r_i(s_i)\{1 + \delta\alpha_i(s_i)\}$$

へと変化する場合，この変化を，$\alpha_i(s_i)$ を変化方向とする変化量 δ の変化であると考える．$\alpha_i(s_i)$ は，式 (44) で述べたヒルベルト空間 $H_{A,r}$ の要素である．ICA の場合，$r_i(s_i)\{1 + \delta\alpha_i(s_i)\}$ が，平均 0，分散 1 の確率密度分布[5]でなければならないことから

$$E_{A,r}[s_i\alpha_i(s_i)] = 0, \quad E_{A,r}[(s_i^2 - 1)\alpha_i(s_i)] = 0 \tag{37}$$

という条件があることを覚えておこう．

このようにヒルベルト空間 $H_{A,r}$ を導入することにより，無限次元パラメータである原信号の確率密度分布 $r_i(s_i)$ の微小変化は，変化方向 $\boldsymbol{\alpha} = (\alpha_1(s_1), \ldots, \alpha_n(s_n))^T$ を一つ定めれば，一つの媒介変数 δ により表現できるようになる．変化方向 $\boldsymbol{\alpha}$ に限定した nuisance スコア関数 $v(\boldsymbol{x}; \boldsymbol{\alpha})$ は，

$$r(\boldsymbol{s}; \boldsymbol{\alpha}, \delta) = \prod_{i=1}^{n} r(s_i; \alpha_i, \delta)$$

[4]S. Amari, J.-F. Cardoso, "Blind source separation - semiparametric statistical approach," IEEE trans. on Signal Processing, vol.45, no.11, pp.2692–2700, 1997.

[5]前述したように，原信号の分散は不定なので，適当に 1 であるとしておく．別に，分散 σ_i^2 でもよい．

として，

$$v(\boldsymbol{x};\boldsymbol{\alpha}) = \lim_{\delta\to 0}\frac{d}{d\delta}\log p(\boldsymbol{x};\boldsymbol{A}, r(\boldsymbol{s};\boldsymbol{\alpha},\delta))$$

と書ける．$p(\boldsymbol{x}) = r(\boldsymbol{s})|\boldsymbol{A}|^{-1}$ なので，

$$\begin{aligned}
\frac{d}{d\delta}\log p(\boldsymbol{x};\boldsymbol{A}, r(\boldsymbol{s};\boldsymbol{\alpha},\delta)) &= \frac{1}{p(\boldsymbol{x})}\frac{d}{d\delta}p(\boldsymbol{x};\boldsymbol{A}, r(\boldsymbol{s};\boldsymbol{\alpha},\delta)) \\
&= \frac{1}{p(\boldsymbol{x})}\frac{d}{d\delta}(|\boldsymbol{A}|^{-1}r(\boldsymbol{s};\boldsymbol{\alpha},\delta)) \\
&= \frac{1}{r(\boldsymbol{s};\boldsymbol{\alpha},\delta)}\frac{d}{d\delta}r(\boldsymbol{s};\boldsymbol{\alpha},\delta) \\
&= \frac{d}{d\delta}\log r(\boldsymbol{s};\boldsymbol{\alpha},\delta) \\
&= \frac{d}{d\delta}\log\prod_{i=1}^{n} r_i(s_i,\alpha_i,\delta) \\
&= \sum_{i=1}^{n}\frac{d}{d\delta}\log r_i(s_i,\alpha_i,\delta) \\
&= \sum_{i=1}^{n}\frac{1}{r_i(s_i,\alpha_i,\delta)}\frac{d}{d\delta}r_i(s_i,\alpha_i,\delta) \\
&= \sum_{i=1}^{n}\frac{1}{r_i(s_i,\alpha_i,\delta)}r_i(s_i)\alpha_i(s_i)
\end{aligned}$$

となる．ゆえに，$\delta\to 0$ とすれば，$r_i(s_i,\alpha_i,\delta)\to r_i(s_i)$ となるので，変化方向 $\boldsymbol{\alpha}$ に限定した nuisance スコア関数 $v(\boldsymbol{x};\boldsymbol{\alpha})$ は，

$$v(\boldsymbol{x};\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i(s_i) \tag{38}$$

となる．したがって，nuisance スコア関数全体は，考えうるすべての $\boldsymbol{\alpha}$ に対する $v(\boldsymbol{x};\boldsymbol{\alpha})$ の集合であるから，

$$T^N_{A,r} = \{v(\boldsymbol{x};\boldsymbol{\alpha})\}_{\forall\alpha} = \left\{\sum_{i=1}^{n}\alpha_i(s_i)\right\}_{\forall\alpha_i(s_i)}$$

と書くことができる．nuisance スコア関数全体 $T^N_{A,r}$ は，確率密度分布 $r(\boldsymbol{s})$ の一つの接空間をなしていることから，nuisance 接空間という．

4.3 A-スコア関数，有効スコア関数

次に，推定したい混合行列 $\boldsymbol{A}$ に対するスコア関数 $\dfrac{\partial}{\partial\boldsymbol{A}}\log p(\boldsymbol{x};\boldsymbol{A},r)$ を求めよう．便宜上，行列 $\boldsymbol{A}$ を

$$\boldsymbol{A} = \boldsymbol{A}_0(\boldsymbol{I}-\boldsymbol{E})$$

と表現し，固定した $\boldsymbol{A}_0$ に対し，行列 $\boldsymbol{E}$ の関数と考える．そして，$\boldsymbol{A}_0$ における行列 $\boldsymbol{E}$ に対するスコア関数 $\boldsymbol{U}(\boldsymbol{x};\boldsymbol{A},r)$ を求めると，

$$\boldsymbol{U}(\boldsymbol{x};\boldsymbol{A},r) = \left.\frac{\partial\log p(\boldsymbol{x};\boldsymbol{A},r)}{\partial\boldsymbol{E}}\right|_{E=0} = \boldsymbol{\varphi}(\boldsymbol{s})\boldsymbol{s}^T + \boldsymbol{I} \tag{39}$$

となる[6]．ただし，$\boldsymbol{s} = \boldsymbol{A}^{-1}\boldsymbol{x}$ であり，また，

$$\boldsymbol{\varphi}(\boldsymbol{s}) = (\varphi_1(s_1),\ldots,\varphi_n(s_n))^T,\quad \varphi_i(s_i) = \frac{r_i'(s_i)}{r_i(s_i)}$$

[6]導出は，少々面倒である．

である．ちなみに，$\boldsymbol{A}_0$ における $\boldsymbol{E}$-スコア関数と $\boldsymbol{A}$-スコア関数の関係は，

$$\left.\frac{\partial \log p(x;\boldsymbol{A},r)}{\partial \boldsymbol{A}}\right|_{A=A_0} = -(A_0^{-1})^T \left.\frac{\partial \log p(x;\boldsymbol{A},r)}{\partial \boldsymbol{E}}\right|_{E=0} \tag{40}$$

となっている．

次に，有効スコア関数 $\boldsymbol{U}^E(\boldsymbol{x};\boldsymbol{A},r)$，つまり $\boldsymbol{E}$-スコア関数 $U(\boldsymbol{x};\boldsymbol{A},r)$ を，nuisance 接空間 $T^N_{A,r}$ に直交する線形部分空間に射影した成分を求めよう．まず，$\boldsymbol{E}$-スコア関数 $\boldsymbol{U}(\boldsymbol{x};\boldsymbol{A},r)$ の非対角成分 $\varphi_i(s_i)s_j,\ i\neq j$ は，s_i と s_j が互いに独立であり，平均 0 であることから，

$$\begin{aligned}\left\langle \varphi_i(s_i)s_j,\ \sum_{k=1}^m \alpha_k(s_k)\right\rangle &= E_{A,r}\Big[\varphi_i(s_i)s_j\sum_{k=1}^m \alpha_k(s_k)\Big]\\ &= \sum_{k=1}^m E_{A,r}\Big[\varphi_i(s_i)s_j\alpha_k(s_k)\Big]\\ &= 0\end{aligned} \tag{41}$$

となり，nuisance 接空間 $T^N_{A,r}$ と直交していることがわかる．つまり，非対角成分については，スコア関数が有効スコア関数に等しくなっている．次に，対角成分を考えよう．まず，式 (37) より，任意の $\alpha_i(s_i)$ は s_i と s_i^2-1 のいずれとも，あるいはそれらの線形和とも直交していることがわかる．ゆえに，$\boldsymbol{E}$-スコア関数 $\boldsymbol{U}(\boldsymbol{x};\boldsymbol{A},r)$ の対角成分において，nuisance スコア関数に直交する成分は，c_i, d_i を定数として，

$$c_i(s_i^2-1)+d_is_i$$

と表されることになる．まとめると，有効スコア関数の i 行 j 列要素 $\boldsymbol{U}^E(\boldsymbol{x};\boldsymbol{A},r)_{i,j}$ は，

$$\boldsymbol{U}^E(\boldsymbol{x};\boldsymbol{A},r)_{i,j} = \begin{cases} c_i(s_i^2-1)+d_is_i, & i=j\\ \varphi_i(s_i)s_j, & i\neq j\end{cases}$$

と書けることになる．

4.4 ICA における推定関数の空間

有効スコア関数 $\boldsymbol{U}^E(\boldsymbol{x};\boldsymbol{A},r)$ の非対角要素により張られる線形空間

$$T^E_{A,r} = \{\varphi_i(s_i)s_j\}$$

を考える．さらに，とりうるすべての $r(\boldsymbol{s})$ に対する $T^E_{A,r}$ の集合

$$T^E_A = \{T^E_{A,r}\}_{\forall r_i}$$

を定義する．次に，

$$E_{A,r}[a(\boldsymbol{x})] = 0$$

を満足するすべての $a(\boldsymbol{x})$ からなる集合として，$(\boldsymbol{A},r)$ の点における ancillary 空間 $T^A_{A,r}$ を定義する．これは，nuisance 接空間 $T^N_{A,r}$ や $T^E_{A,r}$ のいずれとも直交している．同様に，とりうるすべての $r_i(s_i)$ に対する $T^A_{A,r}$ の集合

$$T^A_A = \{T^A_{A,r}\}_{\forall r_i}$$

を定義する．

Amari, Kawanabe によって，任意の r, r' に対し，

$$E_{A,r}[\boldsymbol{U}^E(\boldsymbol{x};\boldsymbol{A},r')] = 0$$

となることが示されている．これは，すべてのセミパラメトリック推定に対していえることではなく，少なくとも ICA について成り立つものである．これは，混合行列 $\boldsymbol{A}$ の非対角部分については，s_i と s_j が独立ならば，任意の $\varphi(\cdot)$ に対し，

$$E_{A,r}[\varphi_i(s_i)s_j] = 0$$

になることから分かる．対角部分については，平均 0，分散 1 の条件より，常に，

$$E_{A,r}[c_i(s_i^2 - 1) + d_i s_i] = 0$$

が成り立つ．以上より，ICA における推定関数 $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{A})$ は，有効スコア関数 $\boldsymbol{U}^E(\boldsymbol{x}; \boldsymbol{A}, r)$ の線形結合と $\boldsymbol{T}^A_{A,r}$ に属する適当な行列関数 $\boldsymbol{B}(\boldsymbol{x}; \boldsymbol{A}, r)$ の和

$$\boldsymbol{F}(\boldsymbol{x}; \boldsymbol{A}) = \boldsymbol{C}(\boldsymbol{A}, r)\boldsymbol{U}^E(\boldsymbol{x}; \boldsymbol{A}, r) + \boldsymbol{B}(\boldsymbol{x}; \boldsymbol{A}, r)$$

で与えられ，それがすべてであることが分かる．ただし，$\boldsymbol{C}(\boldsymbol{A}, r)$ は，4 次テンソルであり，有効スコア関数 $\boldsymbol{U}^E(\boldsymbol{x}; \boldsymbol{A}, r)$ の線形結合の係数を表す．つまり，$\boldsymbol{C}(\boldsymbol{A}, r)$ の要素を $\boldsymbol{C}(\boldsymbol{A}, r)_{h,i,j,k}$ で表すならば，$\boldsymbol{C}(\boldsymbol{A}, r)\boldsymbol{U}^E(\boldsymbol{x}; \boldsymbol{A}, r)$ の j, k 要素は，

$$\sum_{h,i} \boldsymbol{C}(\boldsymbol{A}, r)_{h,i,j,k}\boldsymbol{U}^E(\boldsymbol{x}; \boldsymbol{A}, r)_{h,i}$$

と表されることになる．$\boldsymbol{C}(\boldsymbol{A}, r)$ を適当に選ぶことにより，

$$\boldsymbol{I} + \alpha\boldsymbol{\varphi}(\boldsymbol{y})\boldsymbol{y}^T + \beta\boldsymbol{y}\boldsymbol{\varphi}(\boldsymbol{y})^T$$

なども表現できるので，これも推定関数の一つであることが分かる．

4.5 ICA の推定精度と有効性

その際の推定精度，つまり分散共分散行列は，式 (47) で与えられる．真の分布が r_0 であるとして，有効スコア関数 $U^E(\boldsymbol{x}; \boldsymbol{A}, r_0)$ を推定関数 $F(\boldsymbol{x}, \boldsymbol{A})$ に選んだ時，この M 推定量は，漸近的に Fisher 有効推定量，つまり，最適な推定法になる．

A 付録

A.1 確率変数の独立性

確率変数 $y_i,\ i = 1, \ldots, n$ は，y_i の確率密度分布を $q_i(\boldsymbol{y}_i)$ として，$\boldsymbol{y} = (y_1, \ldots, y_n)^T$ の確率密度分布 $q(\boldsymbol{y})$ が

$$q(\boldsymbol{y}) = \prod_{i=1}^{n} q_i(y_i)$$

として表現できるとき，互いに独立であるという．独立であることを評価するためには，確率密度分布を知る必要がある．確率密度分が未知である場合には，確率変数 y_i の具体的な標本から，確率密度分布をヒストグラムの推定などを介して推定する必要がある．これは，一般に大変な作業である．そこで，実際には，独立であるための一つの必要条件を用いて，独立性を評価することが多い．その一つがクロスキュムラントである．y_1, y_2 のクロスキュムラント $c[y_1^{n_1} y_2^{n_2}]$ は，これらの確率変数が独立ならば，$c[y_1^{n_1} y_2^{n_2}] = c[y_1^{n_1}]c[y_2^{n_2}]$ とそれぞれのキュムラントの積として表現することができる．ただし，n_1, n_2 は，任意の整数であり，$n_1 + n_2$

を次数という[7]．3 次以下のクロスキュムラントやキュムラントは，中心モーメント（平均が 0 ならば，モーメントとしてもよい）に一致する．

平均 0 の独立な確率変数 y_1, y_2 のクロスキュムラントにおいて，例えば，$n_1 = n_2 = 1$ の場合，つまり $c[y_1 y_2] = c[y_1]c[y_2]$ は，$E[y_1 y_2] = E[y_1]E[y_2] = 0$ と書ける．この条件は，一般に無相関と呼ばれる．$n_1 = 2,\ n_2 = 1$ とすれば，$E[y_1^2 y_2] = E[y_1^2]E[y_2] = 0$，$n_1 = 2,\ n_2 = 2$ とすれば，$c[y_1^2 y_2^2] = c[y_1^2]c[y_2^2] = E[y_1^2]E[y_2^2]$，$n_1 = 3,\ n_2 = 1$ とすれば，$c[y_1^3 y_2^1] = c[y_1^3]c[y_2^1] = E[y_1^3]E[y_2^1] = 0$ などとなる．こうした条件を満たすことを示すことで，独立性を評価することがある．しかし，これは独立性のための必要条件に過ぎないので，これらの条件を満たしたからといって，独立であるとは限らないことに注意する必要がある．

情報理論の立場では，以下の尺度を用いると便利である．確率変数 $y_i,\ i = 1, \ldots, n$ が独立ならば，それらのそれぞれのエントロピーの和 $\sum_{i=1}^{n} H(y_i)$ と $\boldsymbol{y}$ 全体のエントロピー $H(\boldsymbol{y})$ は一致する．これは，独立性の定義より容易に導かれ，独立であるための必要十分条件である．そこで，$y_i,\ i = 1, \ldots, n$ の間の平均相互情報量を

$$\bar{I}(\boldsymbol{y}) = \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{y})$$

と定義すると，これがゼロになることにより，独立性を判定することができる．

A.2 確率変数の正規性

確率変数 $\boldsymbol{y}$ が（多次元）正規分布に従うとき，この確率変数は正規性を満たすという．正規性を示すためには，やはり確率変数の確率密度分布を知る必要があり，面倒なことが多い．そこで，正規性のための適当な必要条件を評価することにより，正規性の判定を行うことがある．こうした正規性の判定においても，高次キュムラントを用いることが多い．これは，正規分布に従う確率変数に対しては，3 次以上のキュムラントが 0 になることを利用している．例えば，平均が 0 の確率変数 y_1 の 3 次キュムラントは $c[y_1^3] = E[y_1^3]$，4 次キュムラントは $c[y_1^4] = E[y_1^4] - 3E[y_1^2]^2$ などとして表現できるので，これらがゼロになるか否かを調べることにより，正規性の判定を行うことができる[8]．また，4 次中心モーメントを分散の 2 乗で規格化した $E[y_1^4]/E[y_1^2]^2$ は，尖度と呼ばれ，正規分布に従うとき，値 3 をとる．ゆえに，尖度が 3 か否かにより，正規性を判定することもできる．4 次キュムラントと尖度は，本質的には同じものである．

情報理論の立場からは，正規性は以下のように評価することができる．これには，分散，あるいは分散共分散行列を一定値に制約する条件の下で，エントロピーを最大にする確率密度分布は，（多次元）正規分布であることを利用する．したがって，確率変数 $\boldsymbol{y}$ に対し，それと等しい分散共分散行列の正規分布に従う確率変数を $\boldsymbol{y}_{Gauss}$ とすると，

$$H(\boldsymbol{y}_{Gauss}) - H(\boldsymbol{y})$$

は，常に正の値をとり，確率変数 $\boldsymbol{y}$ が正規分布に従うとき，最小値 0 をとることになる．

A.3 確率変数のエントロピーとその近似

確率密度分布 $p(x)$ を持つ連続な確率変数 x のエントロピー $H(x)$ は，

$$H(x) = -\int p(x) \log p(x) dx$$

[7] ここでは，クロスキュムラントとして，$c[y_1^{n_1} y_2^{n_2}]$ といった表現を用いたが，本来は，$c[y_1, n_1\ ;\ y_2, n_2]$ などと表現しなければならない．$y_1^{n_1} y_2^{n_2}$ のキュムラントや，$y_1^{n_1}$ と $y_2^{n_2}$ のクロスキュムラントとは異なるので注意．

[8] ここで，キュムラントとして，$c[y_1^3]$ などと表現したが，本来は，$c[y_1, 3]$ などと表現しなければならない．y_1 の 3 次キュムラントと y_1^3 のキュムラントは異なるので注意．

として定義される．連続な確率変数のエントロピーは，正確には，differential entropy と呼ばれるが，対応する日本語がないので，単にエントロピーと呼ぶことにする．エントロピーは，確率密度分布 $p(x)$ が未知である場合，近似式

$$H(x) \simeq -\frac{1}{12}E[x^3]^2 - \frac{1}{48}c[x^4]^2 \tag{42}$$

を用いて求めることがある．この近似式では，正規分布に従うとき，$H(x) = 0$，それ以外の分布の時，$H(x) \leq 0$ となる．

A.4 自然勾配

正則な（非特異な）正方行列 $\boldsymbol{W}$ を変数とする評価関数 $f(\boldsymbol{W})$ の最大値を探索する非線形最適化問題を考える．$\boldsymbol{W}$ を微小量 $\Delta\boldsymbol{W}$ だけ変化させて $\boldsymbol{W} + \Delta\boldsymbol{W}$ としたとき，評価関数の値 $f(\boldsymbol{W} + \Delta\boldsymbol{W})$ を最大にする方向 $\Delta\boldsymbol{W}$ は，$\Delta\boldsymbol{W}$ の Frobenius ノルム[9]を一定にする条件下では，最急勾配方向 $\Delta\boldsymbol{W} = \dfrac{df(\boldsymbol{W})}{d\boldsymbol{W}}$ になる．しかし，行列を変数に持つ評価関数の場合，ICA のように，行列が線形変換として作用することが多い．この場合，行列 $\boldsymbol{W}$ の各要素が全体的に大きな値を持つ場合と小さな値を持つ場合とでは，微小変化 $\Delta\boldsymbol{W}$ が評価関数に与える影響は異なる．つまり，計量空間がユークリッド計量空間とは異なっているのであり，こうした計量をリーマン計量という．

そこで，単位行列を基本として，そこからの変化分を考える．つまり，

$$\boldsymbol{W} + \Delta\boldsymbol{W} = (\boldsymbol{I} + \Delta\boldsymbol{W}\boldsymbol{W}^{-1})\boldsymbol{W}$$

とし，変化分を $\Delta\boldsymbol{W}\boldsymbol{W}^{-1}$ と考える．これは，$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$，$\boldsymbol{y} + \Delta\boldsymbol{y} = (\boldsymbol{W} + \Delta\boldsymbol{W})\boldsymbol{x}$ とすると，$\Delta\boldsymbol{y}$ は，

$$\Delta\boldsymbol{y} = \Delta\boldsymbol{W}\boldsymbol{W}^{-1}\boldsymbol{y}$$

と表現できることから，$\boldsymbol{y}$ から $\Delta\boldsymbol{y}$ への線形変換を表していると解釈できる．

変化分 $\Delta\boldsymbol{W}\boldsymbol{W}^{-1}$ の Frobenius ノルムを一定値に抑える条件

$$tr((\Delta\boldsymbol{W}\boldsymbol{W}^{-1})^T(\Delta\boldsymbol{W}\boldsymbol{W}^{-1})) = \epsilon$$

の下で，評価関数を最大にする方向を探す．$f(\boldsymbol{W} + \Delta\boldsymbol{W})$ を 1 次のテイラー級数で近似すると，

$$f(\boldsymbol{W} + \Delta\boldsymbol{W}) \simeq f(\boldsymbol{W}) + tr\left(\frac{df(\boldsymbol{W})^T}{d\boldsymbol{W}}\Delta\boldsymbol{W}\right)$$

であるから，ラグランジェの未定乗数 λ を使って，評価関数を

$$J = tr\left(\frac{df(\boldsymbol{W})^T}{d\boldsymbol{W}}\Delta\boldsymbol{W}\right) - \lambda(tr((\Delta\boldsymbol{W}\boldsymbol{W}^{-1})^T(\Delta\boldsymbol{W}\boldsymbol{W}^{-1})) - \epsilon)$$

とし，これを $\Delta\boldsymbol{W}$ で微分してゼロとおけば，

$$\frac{dJ}{d\Delta\boldsymbol{W}} = \frac{df(\boldsymbol{W})^T}{d\boldsymbol{W}} - 2\lambda\boldsymbol{W}^{-1}(\boldsymbol{W}^{-1})^T\Delta\boldsymbol{W}^T = 0$$

となる．ゆえに，先に定義したリーマン計量に基づいて与えられる最適な勾配方向は，

$$\Delta\boldsymbol{W} = \frac{1}{2\lambda}\frac{df(\boldsymbol{W})}{d\boldsymbol{W}}\boldsymbol{W}^T\boldsymbol{W}$$

となる．これを自然勾配（ナチュラルグラジェント）という．

[9]行列の各要素の 2 乗和であり，$tr(\Delta\boldsymbol{W}^T\Delta\boldsymbol{W})$ として求められる

A.5　主成分解析

確率変数ベクトル $\boldsymbol{x}$ を $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ と線形変換することにより，$\boldsymbol{y}$ の各要素を無相関化することを考えよう．$\boldsymbol{x}$ の期待値はゼロ（ベクトル）であるものとする．$\boldsymbol{y}$ の各要素が無相関ということは，任意の異なる 2 要素の共分散がゼロということである．つまり，分散共分散 $E[\boldsymbol{y}\boldsymbol{y}^T]$ を対角行列にするような線形変換 $\boldsymbol{W}$ を見つければよい．そこで，

$$E[\boldsymbol{y}\boldsymbol{y}^T] = E[(\boldsymbol{W}\boldsymbol{x})(\boldsymbol{W}\boldsymbol{x})^T] = \boldsymbol{W}E[\boldsymbol{x}\boldsymbol{x}^T]\boldsymbol{W}^T$$

とすれば，行列 $\boldsymbol{W}$ は，$\boldsymbol{x}$ の分散共分散行列 $E[\boldsymbol{x}\boldsymbol{x}^T]$ を対角化する行列であることがわかる．$\boldsymbol{x}$ の分散共分散行列 $E[\boldsymbol{x}\boldsymbol{x}^T]$ は，対称行列であるから，相異なる正の実数の固有値を持ち，対応する固有ベクトルは互いに直交する．これらの固有ベクトルを行列 $\boldsymbol{W}$ の各行ベクトルにすれば，$\boldsymbol{x}$ の分散共分散行列 $E[\boldsymbol{x}\boldsymbol{x}^T]$ を対角化可能である．対角化された $E[\boldsymbol{y}\boldsymbol{y}^T]$ は，対角部分に $E[\boldsymbol{x}\boldsymbol{x}^T]$ の固有値を並べたものとなる．もっとも大きい固有値に対する固有ベクトルが示す方向を $\boldsymbol{x}$ の主成分方向，その方向の成分を主成分という．こうした解析を主成分解析と呼んでいる．

A.6　行列の直交化

正則な正方行列 $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ の各列ベクトルの直交化は，Gram-Schmidt の直交化法により実現できる．Gram-Schmidt の直交化法は，Matlab などの行列演算パッケージでは，QR 分解を用いると便利である．QR 分解は，正則な正方行列を直交行列 $\boldsymbol{Q}$ と上三角行列 $\boldsymbol{R}$ の積に分解するもので，これにより得られた直交行列は，元の行列 $\boldsymbol{A}$ の各列ベクトルを直交化している．実際には，元の行列の列ベクトルの方向の変化を $[-\pi/2, \pi/2]$ の範囲に抑えるため，上三角行列 $\boldsymbol{R}$ の対角要素の符号と等しい符号を持つ絶対値 1 の値を対角要素に持つ対角行列を $\boldsymbol{S}$ として，$\boldsymbol{Q}\boldsymbol{S}$ として直交化を行う．

A.7　スコア関数と Cramér-Rao の下界，その情報幾何学的解釈

パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\xi}$ を持つ確率密度分布 $p(x; \boldsymbol{\theta}, \boldsymbol{\xi})$ を考えよう．ただし，パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\xi}$ は，いずれも有限次元パラメータであるとする．この確率密度分布に従う確率変数を X とする．このとき，その統計的に独立な X の T 個の標本 $D_T \equiv \{x_1, \ldots, x_T\}$ を観測し，パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\xi}$ を最尤推定する問題を考える．最尤推定では，対数尤度

$$L(\boldsymbol{\theta}, \boldsymbol{\xi}) = \log \prod_{i=1}^{T} p(x_i; \boldsymbol{\theta}, \boldsymbol{\xi}) = \sum_{i=1}^{T} \log p(x_i; \boldsymbol{\theta}, \boldsymbol{\xi})$$

を $\boldsymbol{\theta}$，$\boldsymbol{\xi}$ でそれぞれ偏微分したものをゼロに等しいとおいた方程式の解が推定値 $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\xi}}$ となる．ここで，

$$\begin{aligned} \boldsymbol{u}(x; \boldsymbol{\theta}, \boldsymbol{\xi}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log p(x; \boldsymbol{\theta}, \boldsymbol{\xi}) \\ \boldsymbol{v}(x; \boldsymbol{\theta}, \boldsymbol{\xi}) &= \frac{\partial}{\partial \boldsymbol{\xi}} \log p(x; \boldsymbol{\theta}, \boldsymbol{\xi}) \end{aligned}$$

とおけば，

$$\sum_{i=1}^{n} \boldsymbol{u}(x_i; \boldsymbol{\theta}, \boldsymbol{\xi}) = \boldsymbol{0}, \qquad \sum_{i=1}^{n} \boldsymbol{v}(x_i; \boldsymbol{\theta}, \boldsymbol{\xi}) = \boldsymbol{0}$$

の解が推定値 $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\xi}}$ ということになる．推定値 $(\hat{\boldsymbol{\theta}},\ \hat{\boldsymbol{\xi}})^T$ の真値 $(\boldsymbol{\theta},\ \boldsymbol{\xi})^T$ の周りの分散共分散行列は，

$$\boldsymbol{V} = \begin{pmatrix} E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] & E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})^T] \\ E[(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] & E[(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi})^T] \end{pmatrix}$$

となる．$\boldsymbol{\theta}$ と $\boldsymbol{\xi}$ が不偏推定量であれば，この分散共分散行列 $\boldsymbol{V}$ の下限は，Cramér-Rao の下界

$$\boldsymbol{V} \geq \frac{1}{T}\boldsymbol{G}^{-1}$$

で与えられる．だたし，行列 $\boldsymbol{G}$ は，フィッシャーの情報行列

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_u & \boldsymbol{G}_{uv} \\ \boldsymbol{G}_{vu} & \boldsymbol{G}_v \end{pmatrix} = \begin{pmatrix} E[\boldsymbol{u}\boldsymbol{u}^T] & E[\boldsymbol{u}\boldsymbol{v}^T] \\ E[\boldsymbol{v}\boldsymbol{u}^T] & E[\boldsymbol{v}\boldsymbol{v}^T] \end{pmatrix}$$

である．また，行列の不等式は，正定値の意味での大小を意味する．最尤推定量は，漸近的に $(T \to \infty)$，この下限の推定精度を実現し，有効推定量となる．また，概して，対数尤度の微分である $\boldsymbol{u}(x;\boldsymbol{\theta},\boldsymbol{\xi}), \boldsymbol{v}(x;\boldsymbol{\theta},\boldsymbol{\xi})$ が大きいほど，フィッシャーの情報行列は大きくなり，結果として Cramér-Rao の下界は下がり，推定精度が向上することになる．これは，最尤推定が，対数尤度の微分である $\boldsymbol{u}(x;\boldsymbol{\theta},\boldsymbol{\xi}), \boldsymbol{v}(x;\boldsymbol{\theta},\boldsymbol{\xi})$ をゼロとするようにパラメータを決定することから，この $\boldsymbol{u}(x;\boldsymbol{\theta},\boldsymbol{\xi}), \boldsymbol{v}(x;\boldsymbol{\theta},\boldsymbol{\xi})$ の 2 乗の期待値が大きいほど，ゼロの位置をより正確に検出できるようになり，推定精度が向上すると解釈することができる．言い換えると，パラメータ $\boldsymbol{\theta},\boldsymbol{\xi}$ の値を変化させたとき，対数尤度が敏感に変化するほど，対数尤度の最大値をより正確に見つけやすくなるといえる．こうした意味から，$\boldsymbol{u}(x;\boldsymbol{\theta},\boldsymbol{\xi}), \boldsymbol{v}(x;\boldsymbol{\theta},\boldsymbol{\xi})$ をそれぞれ $\boldsymbol{\theta}$-スコア関数，$\boldsymbol{\xi}$-スコア関数，あるいはこれらを区別しない場合には単にスコア関数と呼ぶ．

さて，パラメータ $\boldsymbol{\theta}$ だけに着目すると，$\boldsymbol{\theta}$ の推定値の分散共分散行列

$$V_\theta = E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T]$$

の Cramér-Rao の下界は，

$$V_\theta \geq \frac{1}{T}(\boldsymbol{G}_u - \boldsymbol{G}_{uv}\boldsymbol{G}_v^{-1}\boldsymbol{G}_{vu})^{-1} \tag{43}$$

と書ける．仮に，$\hat{\boldsymbol{\theta}}$ と $\hat{\boldsymbol{\xi}}$ が無相関ならば，$\boldsymbol{G}_{uv}$，$\boldsymbol{G}_{vu}$ はゼロ行列となるので，上式は，

$$V_\theta \geq \frac{1}{n}\boldsymbol{G}_u^{-1}$$

となる．また，$\boldsymbol{\xi}$ が既知である場合にも，

$$V_\theta \geq \frac{1}{n}\boldsymbol{G}_u^{-1}$$

となる．一般に，

$$(\boldsymbol{G}_u - \boldsymbol{G}_{uv}\boldsymbol{G}_v^{-1}\boldsymbol{G}_{vu})^{-1} \geq \boldsymbol{G}_u^{-1}$$

であるので，$\boldsymbol{\xi}$ が未知で，$\hat{\boldsymbol{\theta}}$ と $\hat{\boldsymbol{\xi}}$ に相関がある場合には，$\boldsymbol{\theta}$ の推定精度は，一方のパラメータ $\boldsymbol{\xi}$ の存在によって，悪化することになる．こうした意味で，パラメータ $\boldsymbol{\theta}$ に対し，パラメータ $\boldsymbol{\xi}$ を nuisance パラメータという．

nuisance パラメータ $\boldsymbol{\xi}$ が，例えば確率密度分布などのように，無限次元パラメータである場合を考えよう．この場合，行列 $\boldsymbol{G}_{vv}, \boldsymbol{G}_{uv}, \boldsymbol{G}_{vu}$ を求めることができない．そこで，情報幾何学[10]の立場から式 (43) の意味を考え直してみよう．次式を満たす $w(x)$ の集合 $H_{\theta,\xi}$ を考えよう．

$$H_{\theta,\xi} = \{w(x) \mid E_{\theta,\xi}[w(x)] = 0,\ E_{\theta,\xi}[w^2(x)] < \infty\} \tag{44}$$

ここで，$E_{\theta,\xi}[\cdot]$ は，$p(x;\boldsymbol{\theta},\boldsymbol{\xi})$ に関する期待値を意味する．$w_1(x), w_2(x) \in H_{\theta,\xi}$ の内積を

$$\langle w_1(x), w_2(x)\rangle = E_{\theta,\xi}[w_1(x)w_2(x)]$$

[10]情報幾何学とは，確率密度分布多様体の微分幾何学を意味する．つまり，確率密度分布は，0 以上の値をとり，その積分が 1 であるという制約があるため，すべての関数空間の中の特定の空間（多様体）に制約されている．その多様体の各点，つまり各確率密度分布での接空間内，つまり線形空間近似して，確率密度分布の変化，挙動を理解，解析する学問といえる．

と定義すれば，$H_{\theta,\xi}$ はヒルベルト空間になることがわかる．幾何学的に言えば，$w(x) \in H_{\theta,\xi}$ は，確率密度分布 $p(x;\boldsymbol{\theta},\boldsymbol{\xi})$ から

$$p(x;\boldsymbol{\theta},\boldsymbol{\xi})(1+\varepsilon w(x))$$

への微小な偏差を表す．ただし，ε は，$p(x;\boldsymbol{\theta},\boldsymbol{\xi})(1+\varepsilon w(x))$ がすべての x で負にならないような微小な定数である．さらに，$p(x;\boldsymbol{\theta},\boldsymbol{\xi})(1+\varepsilon w(x))$ は，その積分が 1 となるので，再び確率密度分布となる．考えうるすべての $w(x)$ から構成される集合 $H_{\theta,\xi}$ は，すべての確率密度分布からなる集合における $p(x;\boldsymbol{\theta},\boldsymbol{\xi})$ での接空間とみなすことができる．

さて，スコア関数の各要素

$$u_i(x;\boldsymbol{\theta},\boldsymbol{\xi}) = \frac{\partial}{\partial\theta_i}\log p(x;\boldsymbol{\theta},\boldsymbol{\xi}), \qquad v_i(x;\boldsymbol{\theta},\boldsymbol{\xi}) = \frac{\partial}{\partial\xi_i}\log p(x;\boldsymbol{\theta},\boldsymbol{\xi})$$

は，パラメータ $\boldsymbol{\theta},\boldsymbol{\xi}$ が最尤推定値となっていれば，これらの期待値は 0 になっているはずである．つまり，スコア関数は，接空間 $H_{\theta,\xi}$ の要素である．更に言えば，u_i, v_i は，接空間 $H_{\theta,\xi}$ の中で，パラメータ θ_i, ξ_i をそれぞれ変化させることにより表現可能な確率密度分布への偏差 $w(x)$ を表す．一方，その他の $w(x)$ は，パラメータ $\boldsymbol{\theta},\boldsymbol{\xi}$ を変化させることにより表現不可能な確率密度分布への偏差を表す．$\boldsymbol{\xi}$-スコア関数 $\boldsymbol{v}(x;\boldsymbol{\theta},\boldsymbol{\xi})$ によって張られる線形部分空間を $T^N_{\theta,\xi}$ とする．$\boldsymbol{\theta}$-スコア関数 u_i において，この $T^N_{\theta,\xi}$ に直交している成分 u_i^E は，

$$u_i^E = u_i - \sum_{j,k} E_{\theta,\xi}[u_i v_k](\boldsymbol{G}_v^{-1})_{j,k} v_j$$

と表される．ただし，$(\boldsymbol{G}_v^{-1})_{j,k}$ は，行列 $\boldsymbol{G}_v^{-1}$ の j,k 要素を表す．したがって，$\boldsymbol{\theta}$-スコア関数 $\boldsymbol{u}$ において，この $T^N_{\theta,\xi}$ に直交している成分 $\boldsymbol{u}^E = (u_1^E, u_2^E, \ldots)^T$ は，

$$\boldsymbol{u}^E = \boldsymbol{u} - \boldsymbol{G}_{uv}\boldsymbol{G}_v^{-1}\boldsymbol{v}$$

となる．$T^N_{\theta,\xi}$ に直交している $\boldsymbol{\theta}$-スコア関数 $\boldsymbol{u}^E$ から得られるフィッシャーの情報行列 $\boldsymbol{G}^E = E_{\theta,\xi}[\boldsymbol{u}^E(\boldsymbol{u}^E)^T]$ は，

$$\boldsymbol{G}^E = E_{\theta,\xi}[\boldsymbol{u}^E(\boldsymbol{u}^E)^T] = \boldsymbol{G}_u - \boldsymbol{G}_{uv}\boldsymbol{G}_v^{-1}\boldsymbol{G}_{vu}$$

となる．このフィッシャー情報行列 $\boldsymbol{G}^E$ により与えられる Cramér-Rao の下界は，

$$\boldsymbol{V}_\theta \geq \frac{1}{T}(\boldsymbol{G}^E)^{-1}$$

となるが，これは，式 (43) で導いた Cramér-Rao の下界に等しい．つまり，パラメータ $\boldsymbol{\xi}$ が存在する場合のパラメータ $\boldsymbol{\theta}$ の推定分散の下界は，$\boldsymbol{\theta}$-スコア関数 $\boldsymbol{u}$ の $T^N_{\theta,\xi}$ に直交している成分 $\boldsymbol{u}^E$ から作られるフィッシャー情報行列で与えられる Cramér-Rao の下界に等しいといえる．$\boldsymbol{\theta}$-スコア関数 $\boldsymbol{u}$ の $T^N_{\theta,\xi}$ に沿った成分はスコア関数としての役割を果たさないので，$T^N_{\theta,\xi}$ を $\boldsymbol{\theta},\boldsymbol{\xi}$ における nuisance 接空間という．また，$\boldsymbol{\theta}$ の推定に役立つ nuisance 接空間 $T^N_{\theta,\xi}$ に直交した $\boldsymbol{\theta}$-スコア関数 u^E を有効 $\boldsymbol{\theta}$-スコア関数という．さらに，$\boldsymbol{G}^E = E_{\theta,\xi}[\boldsymbol{u}^E(\boldsymbol{u}^E)^T]$ を有効フィッシャー情報行列という．

A.8 推定関数法

パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\xi}$ を持つ確率密度分布 $p(x;\boldsymbol{\theta},\boldsymbol{\xi})$ を考え，パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\xi}$ を条件とする x に関する条件付期待値を $E_{\theta,\xi}[\cdot]$ と書くことにする．パラメータ $\boldsymbol{\xi}$ に依存しない，かつ $\boldsymbol{\theta}$ と等しい次元のベクトル関数 $\boldsymbol{f}(x,\boldsymbol{\theta})$ は，任意の $\boldsymbol{\theta}$, $\boldsymbol{\xi}$ に対して以下の条件を満たすならば，推定関数と呼ばれる．

$$E_{\theta,\xi}[\boldsymbol{f}(x,\boldsymbol{\theta})] \quad = \quad \mathbf{0},$$

$$
\begin{aligned}
|\boldsymbol{K}| &\neq 0, \\
E_{\theta,\xi}[\boldsymbol{f}(x,\boldsymbol{\theta})\boldsymbol{f}^T(x,\boldsymbol{\theta})] &< \infty, \\
\text{where,} \qquad & \boldsymbol{K} = E_{\theta,\xi}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{f}(x,\boldsymbol{\theta})\right]
\end{aligned} \tag{45}
$$

方程式 $E_{\theta_0,\xi_0}[\boldsymbol{f}(x,\boldsymbol{\theta})] = \mathbf{0}$ を満たす $\boldsymbol{\theta}$ が真値 $\boldsymbol{\theta}_0$ であるから，この方程式を満たす解を求めることにより，$\boldsymbol{\theta}$ を推定することができる．もちろん，実際には，確率密度分布 $p(x;\boldsymbol{\theta},\boldsymbol{\xi})$ に従う独立な標本 $D_T = \{x_1,\ldots,x_T\}$ を利用することになるので，期待値は集合平均に置き換えられて，

$$
\sum_{i=1}^{T}\boldsymbol{f}(x_i,\boldsymbol{\theta}) = \mathbf{0}
$$

を満たす解が推定値 $\hat{\boldsymbol{\theta}}$ となる．こうした推定法を推定関数法といい，推定関数法による推定量 $\hat{\boldsymbol{\theta}}$ を M 推定量という．標本数 $T\to\infty$ では，集合平均は期待値に一致するので，M 推定量は明らかに一致性を持つ．$E_{\theta,\xi}[\boldsymbol{f}(x,\boldsymbol{\theta})] = \mathbf{0}$ が，任意の $\boldsymbol{\xi}$ に対して成立することが，推定関数法のミソであり，$\boldsymbol{\xi}$ がどのような値であっても，M 推定量 $\hat{\boldsymbol{\theta}}$ は，一致性を持つことになる．

次に，M 推定量 $\hat{\boldsymbol{\theta}}$ の推定精度を，その分散共分散行列を求めることにより調べてみよう．定義により，推定量 $\hat{\boldsymbol{\theta}}$ は，

$$
\sum_{i=1}^{T}\boldsymbol{f}(x_i,\hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{46}
$$

を満たす．ここで，真値を $\boldsymbol{\theta}$ として，推定値の誤差を $\Delta\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ と表現しよう．$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \Delta\boldsymbol{\theta}$ を式 (46) に代入して

$$
\sum_{i=1}^{T}\boldsymbol{f}(x_i,\boldsymbol{\theta}+\Delta\boldsymbol{\theta}) = \mathbf{0}
$$

とする．M 推定量の一致性により，標本数 $T\to\infty$ では，$E_{\theta,\xi}[\Delta\boldsymbol{\theta}] = \mathbf{0}$，かつ $\Delta\boldsymbol{\theta}$ のばらつきも限りなく小さくなる．そこで，上式の左辺を $\boldsymbol{\theta}$ の周りで 1 次のテイラー級数で近似すると，

$$
\sum_{i=1}^{T}\boldsymbol{f}(x_i,\boldsymbol{\theta}) + \sum_{i=1}^{T}\frac{\partial\boldsymbol{f}(x_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Delta\boldsymbol{\theta} = \mathbf{0}
$$

となる．ここで，両辺を T で割り，

$$
\frac{1}{T}\sum_{i=1}^{T}\frac{\partial\boldsymbol{f}(x_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \equiv \boldsymbol{K}_D
$$

とおくと，誤差 $\Delta\boldsymbol{\theta}$ は，

$$
\Delta\boldsymbol{\theta} = -\boldsymbol{K}_D^{-1}\frac{1}{T}\sum_{i=1}^{T}\boldsymbol{f}(x_i,\boldsymbol{\theta})
$$

と書ける．ゆえに，M 推定量 $\hat{\boldsymbol{\theta}}$ の分散共分散行列 $\boldsymbol{V}_\theta \equiv E_{\theta,\xi}[\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T]$ は，

$$
\boldsymbol{V}_\theta = E_{\theta,\xi}\left[\boldsymbol{K}_D^{-1}\frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}\boldsymbol{f}(x_i,\boldsymbol{\theta})\boldsymbol{f}^T(x_j,\boldsymbol{\theta})\boldsymbol{K}_D^{-1^T}\right]
$$

となる．ここで，$\boldsymbol{K}_D$ は平均 $E_{\theta,\xi}[\boldsymbol{K}_D] = E_{\theta,\xi}\left[\frac{\partial\boldsymbol{f}(x,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]$ の確率変数であるが，T 個の標本の平均を取っているので，$T\to\infty$ では，$\boldsymbol{f}(x_i,\boldsymbol{\theta})\boldsymbol{f}^T(x_j,\boldsymbol{\theta})$ のばらつきに比較すれば，十分に確定変数とみなせる．ゆえに，分散共分散行列 $\boldsymbol{V}_\theta$ は，漸近的 $(T\to\infty)$ に，

$$
\boldsymbol{V}_\theta = E_{\theta,\xi}[\boldsymbol{K}_D^{-1}]\frac{1}{T^2}\sum_{i=1}^{T}\sum_{j=1}^{T}E_{\theta,\xi}[\boldsymbol{f}(x_i,\boldsymbol{\theta})\boldsymbol{f}^T(x_j,\boldsymbol{\theta})]E_{\theta,\xi}[\boldsymbol{K}_D^{-1}]^T
$$

と近似できる．$x_i,\ i=1,\ldots,T$ は統計的に独立なので，$\boldsymbol{K}=E_{\theta,\xi}[\boldsymbol{K}_D]$ とおくと，結局，

$$\begin{aligned}\boldsymbol{V}_\theta &= \boldsymbol{K}^{-1}\frac{1}{T^2}\sum_{i=1}^{T}E_{\theta,\xi}[\boldsymbol{f}(x_i,\boldsymbol{\theta})\boldsymbol{f}^T(x_i,\boldsymbol{\theta})]\boldsymbol{K}^{-1^T}\\ &= \frac{1}{T}\boldsymbol{K}^{-1}E_{\theta,\xi}[\boldsymbol{f}(x,\boldsymbol{\theta})\boldsymbol{f}^T(x,\boldsymbol{\theta})]\boldsymbol{K}^{-1^T}\end{aligned}\tag{47}$$

となる．M 推定量 $\hat{\boldsymbol{\theta}}$ は，$\boldsymbol{\xi}$ に依存せず一致推定量であるが，その推定分散 $\boldsymbol{V}_\theta$ は，漸近的にも $\boldsymbol{\xi}$ に依存することに注意する必要がある．

ところで，パラメータ $\boldsymbol{\theta}$ は有限次元であるが，$\boldsymbol{\xi}$ は関数次元，すなわち無限次元パラメータである場合がある．こうした場合のパラメータ $\boldsymbol{\theta}$ の推定をセミパラメトリック推定という．セミパラメトリック推定では，推定関数法が特に有用であるが，一般に，推定関数を見つけ出すことは容易ではない．

B　演習の解答 (Matlab のプログラム)

```
function ica_learn1(x,eta,maxIter)
% 元祖 y^3 を使う ICA
% x: 観測信号
% eta: 学習係数
% maxIter: 最大学習回数

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% 復元行列を単位行列初期化（乱数で初期化しても良い）
W=eye(m);

% 各観測信号を，平均 0，分散 1 に規格化する
x=nor(x);

% 学習の反復
for i=1:maxIter
    y = W*x;
%    DeltaW = (W')^(-1) + phi(y)*x'/n; % 最急勾配方向に学習
    DeltaW = (eye(m)+phi(y)*y'/n)*W; % 自然勾配方向に学習
    W = W + eta * DeltaW;
    plot(y(1,:),y(2,:),'r.');
    axis([-3 3 -3 3]);
    drawnow;
end

function y = phi(x)
y=-(x.^3);
% y=-tanh(x);

------------------------------------------------------------
function ica_learn2(x,eta,maxIter)
% Extended Infomax 法
% x: 観測信号
% eta: 学習係数
% maxIter: 最大学習回数

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% 復元行列を単位行列初期化（乱数で初期化しても良い）
W=eye(m);

% 各観測信号を，平均 0，分散 1 に規格化する
x=nor(x);

% 学習の反復
for i=1:maxIter
    y = W*x;
%    DeltaW = (W')^(-1) + phi(y)*x'/n; % 最急勾配方向に学習
    DeltaW = (eye(m)+phi(y)*y'/n)*W; % 自然勾配方向に学習
    W = W + eta * DeltaW;
    plot(y(1,:),y(2,:),'r.');
    axis([-3 3 -3 3]);
    drawnow;
end

function y = phi(x)
y=-(x+diag(sign(cumulant(x')))*tanh(x));

function y=cumulant(x)
y=mean(x.^4)-3*mean(x.^2).^2;

------------------------------------------------------------
function ica_learn3(x,eta,maxIter,s,A)
% Extended Infomax 法で，誤差を見る
% x: 観測信号
% eta: 学習係数
% maxIter: 最大学習回数
```

```
% s : 原信号（誤差評価用）
% A : 混合行列（誤差評価用）

close

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% x の各行を規格化する，誤差評価用に，s の各行についても規格化しておく．
x=nor(x);
s=nor(s);

% x の補正を行ったので，混合行列も変更しておく
A=x/s;

err=[];

% 復元行列を単位行列初期化（乱数で初期化しても良い）
W=eye(m);

% 学習の反復
for i=1:maxIter
    y = W*x;
    err=[err mseica(s,A,y,W)];
    DeltaW = (eye(m)+phi(y)*y'/n)*W;
    W = W + eta * DeltaW;
    plot(y(1,:),y(2,:),'r.'); axis([-3 3 -3 3]);
    drawnow;
end

figure
K=1:size(err,2);
semilogy(K,err(1,:),'b-',K,err(2,:),'r-',K,err(3,:),'g-');
legend('mse','mseA','cc');

function y = phi(x)
y=-(x+diag(sign(cumulant(x')))*tanh(x));

function y=cumulant(x)
y=mean(x.^4)-3*mean(x.^2).^2;

-----------------------------------------------------------
function ica_learn4(x,eta,maxIter)
% プリホワイトニング法
% x: 観測信号
% eta: 学習係数
% maxIter: 最大学習回数

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% 観測信号をプリホワイトニングする
[x,V,D]=prewhitening(x);

% 復元行列を単位行列初期化（乱数で初期化しても良い）
W=eye(m);

for k=1:maxIter
    y = W*x;
    dW = ((phi(y)*y'-y*phi(y)')/n)*W;
    W = W + eta * dW;
    plot(y(1,:),y(2,:),'r.');
    axis([-3 3 -3 3]);
    drawnow;
end

function y = phi(x)
y=-(x+diag(sign(cumulant(x')))*tanh(x));

function y=cumulant(x)
y=mean(x.^4)-3*mean(x.^2).^2;
```

```
------------------------------------------------------------
function ica_learn5(x,eta,maxIter)
% 射影追跡法による ICA
% x: 観測信号
% eta: 学習係数
% maxIter: 最大学習回数

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% 観測信号をプリホワイトニングする
[x,V,D]=prewhitening(x);

% 復元行列を単位行列初期化（乱数で初期化しても良い）
W=eye(m);

for k=1:maxIter
    y = W*x;
    dW = (x*(y.^3)')*diag(mean(y.^4,2))/(n*6)+(x*(y.^2)')*diag(mean(y.^3,2))/(n*2) ...
        -(x*(y.^3)')/(n*2)+(3/2)*W'-W'*diag(mean(y.^4,2))/2;
    W = W + eta * dW';
    [Q,R] = qr(W');
    W = Q';
    plot(y(1,:),y(2,:),'r.');
    axis([-3 3 -3 3]);
    drawnow;
end

------------------------------------------------------------
function ica_learn6(x,eta,maxIter)
% Fast ICA
% x: 観測信号
% eta: 学習係数（Fast ICA では不要なので，使わない）
% maxIter: 最大学習回数

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% 観測信号をプリホワイトニングする
[x,V,D]=prewhitening(x);

% 復元行列を単位行列初期化（乱数で初期化しても良い）
W=eye(m);

for k=1:maxIter
    y = W*x;
    L = diag(mean(y.*gg(y),2));
    B = (diag(mean(ggg(y),2))-L)^(-1);
    W = W + B*(L-(gg(y)*y')/n)*W;
    [Q,R] = qr(W');
    W = Q';
    plot(y(1,:),y(2,:),'r.');
    axis([-3 3 -3 3]);
    drawnow;
end

function y=g(y)
y=y.^4;
%y=log(cosh(y));
%y=-exp(-y.^2/2);

function y=gg(y)
y=4*y.^3;
%y=sinh(y)./cosh(y);
%y=y.*exp(-y.^2/2);

function y=ggg(y)
y=12*y.^2;
%y=1./cosh(y).^2;
%y=(1-y.^2).*exp(-y.^2/2);

------------------------------------------------------------
```

```
function ica_image1
% extended infomax 法で画像の ICA を行う

close
figure
set(gcf,'Position',[10 550 800 200]);

% 原信号（画像）の読み込み
load pics_128x128
s1=images(:,:,1);
s2=images(:,:,2);
s3=images(:,:,3);
s=[s1(:) s2(:) s3(:)]';

m=size(s,1); % 信号数
n=size(s,2); % 標本数

viewer(s)
pause

% 観測信号の設定
A=randn(m);
x=A*s;
x=nor(x);

viewer(x)
pause

eta=0.4;
W=eye(m);

for k=1:70
    y = W * x;
    DeltaW = (eye(m)+phi(y)*y'/n)*W;
    W = W + eta * DeltaW;
    viewer(y)
end

% 画像を描画する
function viewer(x)
n=size(x,1);
for i=1:n
    y=reshape(x(i,:),128,128);
    subplot(1,n,i)
    % Image processing toolbox がない場合, image() を使う.
    imshow(y,[min(y(:)) max(y(:))])
    drawnow
end

function y = phi(x)
y=-(x+diag(sign(cumulant(x')))*tanh(x));

function y=cumulant(x)
y=mean(x.^4)-3*mean(x.^2).^2;

-----------------------------------------------------------
function ica_image2
% プリホワイトニング法で画像の ICA を行う

close
figure
set(gcf,'Position',[10 550 800 200]);

% 原信号（画像）の読み込み
load pics_128x128
s1=images(:,:,1);
s2=images(:,:,2);
s3=images(:,:,3);
s=[s1(:) s2(:) s3(:)]';

m=size(s,1); % 信号数
n=size(s,2); % 標本数
```

```
viewer(s)
pause

% 観測信号の設定
A=randn(m);
x=A*s;
[x,V,D]=prewhitening(x);

viewer(x)
pause

eta=1.0;
W=eye(m);

for k=1:50
    y = W * x;
    dW = ((phi(y)*y'-y*phi(y)')/n)*W;
    W = W + eta * dW;
    viewer(y)
end

% 画像を描画する
function viewer(x)
n=size(x,1);
for i=1:n
    y=reshape(x(i,:),128,128);
    subplot(1,n,i)
    % Image processing toolbox がない場合, image() を使う.
    imshow(y,[min(y(:)) max(y(:))])
    drawnow
end

function y = phi(x)
y=-(x+diag(sign(cumulant(x')))*tanh(x));

function y=cumulant(x)
y=mean(x.^4)-3*mean(x.^2).^2;

-------------------------------------------------------------
function ica_image3
% FastICA で画像の ICA を行う

close
figure
set(gcf,'Position',[10 550 800 200]);

% 原信号（画像）の読み込み
load pics_128x128
s1=images(:,:,1);
s2=images(:,:,2);
s3=images(:,:,3);
s=[s1(:) s2(:) s3(:)]';

m=size(s,1); % 信号数
n=size(s,2); % 標本数

viewer(s)
pause

% 観測信号の設定
A=randn(m);
x=A*s;
[x,V,D]=prewhitening(x);

viewer(x)
pause

W=eye(m);

for k=1:20
    y=W*x;
```

```
    L=diag(mean(y.*gg(y),2));
    B=(diag(mean(ggg(y),2))-L)^(-1);
    W=W+B*(L-(gg(y)*y')/n)*W;
    [Q,R]=qr(W');
    W=Q';
    viewer(y)
end

% 画像を描画する
function viewer(x)
n=size(x,1);
for i=1:n
    y=reshape(x(i,:),128,128);
    subplot(1,n,i)
    % Image processing toolbox がない場合，image() を使う.
    imshow(y,[min(y(:)) max(y(:))])
    drawnow
end

function y=g(y)
y=y.^4;
%y=log(cosh(y));
%y=-exp(-y.^2/2);

function y=gg(y)
y=4*y.^3;
%y=sinh(y)./cosh(y);
%y=y.*exp(-y.^2/2);

function y=ggg(y)
y=12*y.^2;
%y=1./cosh(y).^2;
%y=(1-y.^2).*exp(-y.^2/2);

-------------------------------------------------------
function y=nor(y)
% 信号 y の平均を 0，分散を 1 に規格化する

y=y-repmat(mean(y,2),1,size(y,2));
y=diag(mean(y.^2,2).^(-1/2))*y;

-------------------------------------------------------
function err=mseica(s,A,y,W)

n=size(s,1);
N=size(s,2);

y=diag(mean(y.^2,2).^(-1/2))*y;

% PI を求める
% P=abs(W*A);
% err0=sum(sum(P*diag(1./max(P)))-1)+sum(sum(P'*diag(1./max(P')))-1);

% 復元信号の mse を求める
Q=repmat(mean(s.^2,2)',n,1)-(y*s'/N).^2;
err1=mean(min(Q));

% 混合行列の mse を求める
A=A*diag(mean(A.^2).^(-1/2));
We=inv(W);
Q=repmat(mean(A.^2),n,1)-((We'*A/n).^2)./repmat(mean(We.^2)',1,n);
err2=mean(min(Q));

% クロスキュムラント cc を求める
err3=cc_sub(y*y')+cc_sub(y.^2*y')+cc_sub(y.^3*y')+cc_sub(y.^4*y');
err3=err3/size(y,2);

err=[err1 ; err2 ; err3];

function p=cc_sub(P)
n=size(P,1);
P=abs(P);
```

```
PD=diag(P);
PND=P-diag(PD);
p=sum(PND(:))/sum(PD(:))*(n-1);

-------------------------------------------------------
function [y,v,d]=prewhitening(x)

x=x-repmat(mean(x,2),1,size(x,2));

[v,d]=eig(x*x'/size(x,2));
d=d^(-1/2);
y=v*d*v'*x;

-------------------------------------------------------
function [y,W]=tdica1(x,eta,maxIter,MaxTau)
% Extended Infomax 法による BSD(TDICA)
% x: 観測信号
% eta: 学習係数
% maxIter: 最大学習回数
% MaxTau : 復元フィルタの次数

m=size(x,1); % 観測信号数
n=size(x,2); % 標本数

% 復元行列の初期化
W=zeros(m,m,MaxTau);
W(:,:,1)=eye(m);

% x の各行の平均を 0 に，分散を 1 に規格化する
x=nor(x);

% 学習の反復
for i=1:maxIter
    y = zeros(size(x));
    for j=0:MaxTau-1
       y = y + W(:,:,j+1)*delay(x,j);
    end
    for k=0:MaxTau-1
        cory(:,:,k+1)=phi(y)*delay(y,k)'/n;
    end
    DeltaW=zeros(m,m,MaxTau);
    for j=0:MaxTau-1
        for k=0:j
            DeltaW(:,:,j+1) = DeltaW(:,:,j+1)...
                + ((k==0)*eye(m)+cory(:,:,k+1))*W(:,:,j-k+1);
        end
    end
    W = W + eta * DeltaW;
%     plot(y(1,:),y(2,:),'r.');
%     K=5;
%     axis([-K K -K K]);
%     axis square
    K=30;
    plot(-K:K,xcorr(y(1,:),y(1,:),K),'bo-',-K:K,xcorr(y(2,:),y(2,:),K),'rx-',...
        -K:K,xcorr(y(1,:),y(2,:),K),'gs-')
    axis([-K K -300 2500]);
    drawnow;
end

function y = phi(x)
y=-(x+diag(sign(cumulant(x')))*tanh(x));

function y=cumulant(x)
y=mean(x.^4)-3*mean(x.^2).^2;

function y=delay(x,n)
N=size(x,2);
y=x(:,[N-n+1:N 1:N-n]);
```