

ICA

逢澤正嵩

2016 年 4 月 13 日

本ノートでは、独立成分分析 (Independent Component Analysis) について説明する。ICA は一言でいうと「混ざり合った独立な信号を、データから計算できる何らかの尺度を用いて、独立な成分に分解する」手法である。

基本的な論理の流れは、A Hyvarinen らが書いた Independent Component Analysis という古典的教科書 (2001 著) を参考にしている。教科書も原論文をそのまま書いてる感じなので、興味のある人は教科書を参照のこと (Independent Component Analysis pdf で検索すれば、最初に出る)。

目次

1	基本編	2
1.1	導入: 独立成分分析とは?	2
1.2	中心極限定理	2
1.3	独立性と無相関	4
1.4	Whitening	5
1.5	非ガウス性を最大にする	5
1.6	複数の独立な成分を見つける	6
1.7	まとめ	7
2	試運転編	8
2.1	ノイズなし	9
2.2	ノイズあり	11
3	非ガウス性の定式化編	12
3.1	キュムラントとガウス分布	13
3.2	非ガウス性の指標としての kurtosis	14
3.3	非ガウス性の指標としての negentropy	15
3.4	確率分布関数の多項式展開を用いた情報エントロピーの近似	15
3.5	確率分布関数の非多項式展開を用いた情報エントロピーの近似	16

1 基本編

1.1 導入: 独立成分分析とは?

例えば、あなたが新宿駅の雑踏の中で誰かと待ち合わせをしていたとする。あなたを見つけた友人は、手を振りながら声 $s(t)$ であなたに呼びかけた。しかし、残念なことに新宿駅はとてもうるさい。往来する人々の声 $n_i(t)$ によって (i は整数)、友人の声は遮られてしまうのだ。

$$obs(t) = s(t) + \sum_i n_i(t) \quad (1)$$

これじゃ友人の声は、新宿の雑踏の中に埋もれてしまう。しかし、心配する必要はない。あなたはサイボーグで、人工脳の中に友人の声に特化した超高精度友人フィルターを持っていたのだ。

$$brain(t) = w_0 s(t) + \sum_i w_i s_i(t) \quad (2)$$

ここでフィルターは、 $w_0 \simeq 1$ 、 $|w_i| \ll 1$ を満たす。こうして、めでたく $brain(t) \simeq s(t)$ となり、友人の声はあなたに届いたのであった。

上の例では、超高精度友人フィルターというブラックボックスを通して w_i が解った。実は、人間の脳も同じように処理して、聞き分けを可能にしてくれているらしい。しかし、それもブラックボックスであることにはかわりがない。 w_i を抜き出すにはどのようにしたらよいだろうか? これを可能にしてくれる一つの手法が独立成分分析である。

1.2 中心極限定理

実験をした時には興味のある信号 (の和) s_i に付随して誤差 $n(t)$ がある。

$$x(t) = \sum_i s_i(t) + n(t) \quad (3)$$

理論曲線とデータ点の誤差 $n(t)$ のヒストグラムは、実験とモデル化がうまくいっていると、経験的にはガウス分布に従う

$$p(n) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{(n-\mu)^2}{2\sigma^2}\right) \quad (4)$$

このことは、観測している信号が物理的であろうと、非物理的であろうと関係なく成り立つことが知られている。その背後には中心極限定理がある。

中心極限定理を見るために、一様分布 ($-1 \sim 1$) の信号を N 回サンプルして、その和を計算する。

$$sum = \sum_i a_i \quad (5)$$

$N = 1$ であれば、それは一様分布と同じになる。では $N = 2$ にするとどうなるだろうか? 図 1 に $N = 1, 2, 4, 50$ として、10000 回サンプルしたものをヒストグラムにした表示した。ただし、分散の大きさは 1 になるようにスケールしてある。図を見て、 N の値を増やしていくにつれ心持ちガウス分布に近づいていくことがわかる。図の下にはガウス分布でフィットした時の残差が示してあ

る。このように、統計的に独立な信号を足しあわせることで、ガウス分布に近づいていくことが知られている。これが中心極限定理である。今はまったく同じ性質の信号を足し合わせたが、一般に別の性質の信号の足し合わせをしても同様のことが起きることが知られている。簡単な証明は後ほど紹介する。

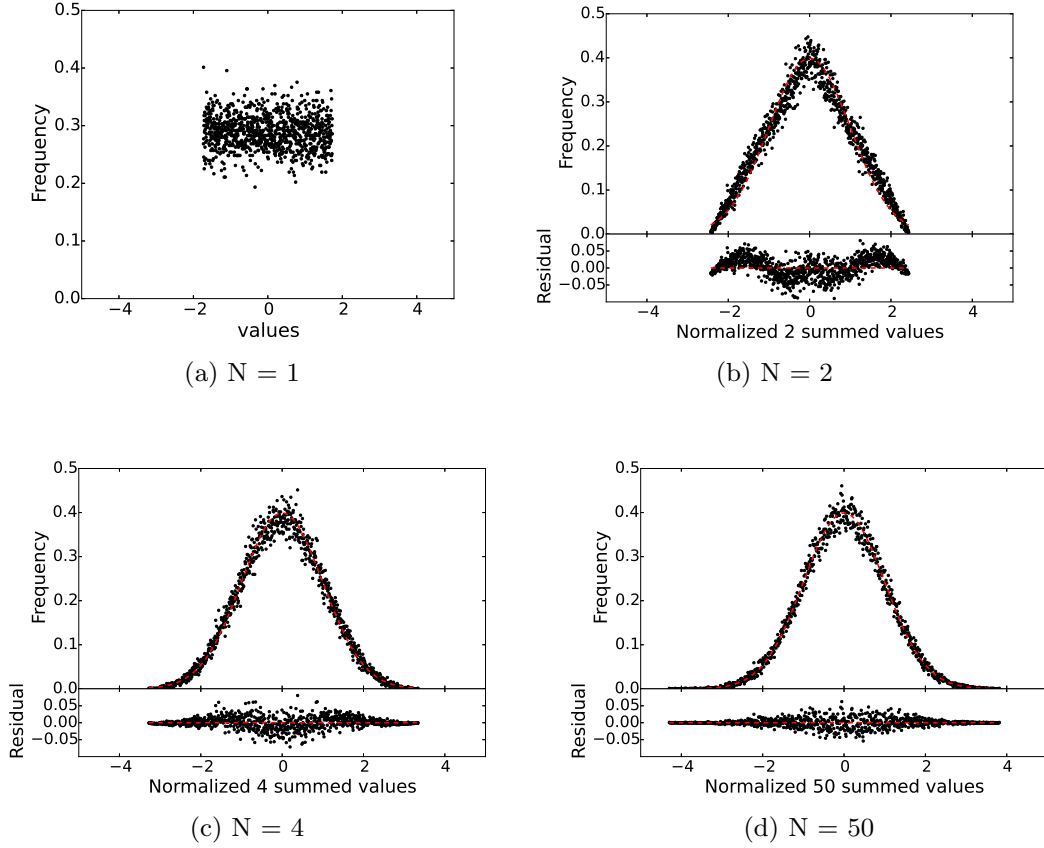


図1 [-1,1) の一様分布の乱数を (a)1 個 (b)2 個 (c)4 個 (d)50 個 足したものをそれぞれ 100000 個作り、 $\sigma = 1$ となるようにスケーリングして、ヒストグラムを作成した。赤線はガウス分布によるフィッティングで、図の下に残差を示した。bin 数は 1000 個。足す個数 (N の値) が大きくなるほど、ガウス分布に近づくことがわかる。

さて、よく分からない信号を足し合わせるとガウス分布っぽくなるということは、逆にいうと和を適切に分離してあげれば、ガウス分布っぽく無くなるだろう。この「ガウスっぽくなければ分離が成功している」という指導原理のもと、信号の分離が行えないだろうか？

まず簡単な例として、2 入力 ($a_1(t), a_2(t)$)、2 出力 ((x_1, x_2)) の線形系を考える。それらは行列 A で結ばれているとする。

$$\begin{aligned} x_1(t) &= A_{1,1}a_1(t) + A_{1,2}a_2(t) \\ x_2(t) &= A_{2,1}a_1(t) + A_{2,2}a_2(t) \end{aligned} \quad (6)$$

すると、 $x_1(t)$ と $x_2(t)$ の分布は、 $a_1(t)$ や $a_2(t)$ に比べて、異なる信号が足されているので、ガウシアンっぽくなっているはずである。もとの信号を抜き出すには、行列 A を知っていれば、その逆行列を x にかければよい。ただ一般に、実験では A は推定できないだろう。そこで、ガウシアンっぽさ (あるいはガウシアンっぽくないさ) を用いて、 a を推定する。

分布 $p(z)$ のガウシアンっぽさ (あるいはガウシアンっぽくなさ) を測定する関数を $G(p(z))$ とする。係数 w_i を導入して、 $z = w_1x_1 + w_2x_2$ として、 $G(p(z = w_1x_1 + w_2x_2))$ を計算しよう。直感的には G を最大 (あるいは最小) 化する w_i をもってくれば、そのときの z の組み合わせがもとの信号 a_1, a_2 のいずれかを再現しているはずである。なので、信号の分離の問題は、結局 $G(w)$ の極値を探す問題に帰着する。 $G(w)$ の具体的な関数については、必要になった時に導入する。

1.3 独立性と無相関

先の節までで、「非ガウス性を用いて、信号を分離する」という ICA のおおよその方針を追った。さて、今まで、自然に使ってきた独立性を数学的に定義する。二つの変数 (x, y) が独立であるとは、その分布関数が直積でかける

$$p(x, y) = p_X(x)p_Y(y) \quad (7)$$

ことと同値である。このことは変数の数が増えても変わらない。また独立性が成り立つと、任意の関数 $g(x), h(y)$ の積の期待値がそれらの期待値の積になる。

$$\begin{aligned} E_{X,Y}\{g(x), h(y)\} &= \int g(x)h(y)p(x, y) \\ &= \int p(x)p_X(x) \int h(y)p_Y(y) \\ &= E_X\{p(x)\}E_Y\{h(y)\} \end{aligned} \quad (8)$$

逆に任意の関数 $g(x), h(y)$ について、 $E_{X,Y}\{g(x)h(y)\} = E_X\{p(x)\}E_Y\{h(y)\}$ となる時信号は独立である。つまり、任意の期待値について、上の式 (8) がなりたつようにすれば、独立にできるのだ。

無相関性 (2 次まで独立) は独立性の必要条件である。例えば、二つの信号 x, y が平均 0、分散 1 であるとする。 x, y が独立であれば、 $E_{X,Y}\{x^2\} = 1$, $E_{X,Y}\{xy\} = E_X\{x\}E_Y\{y\} = 0$, $E_{X,Y}\{y^2\} = 1$ となる。なので独立ならば、無相関である。実はこの 2 次の相関までは適当な線形変換を施せば、無相関 (式 (8) を満たすようにできる) にできる。なので、通常は前処理として無相関化 (実際は Whitening といって、二次の相関を単位行列にするとこまでやる) をやっておいて、線形変換の自由度を消しておくようだ。ちなみに、Whitening をして相関行列を単位行列にしておくと、直交変換を施しても相関行列は変化しないので、直交変換の自由度が線形変換に残っている。

ということで、ICA の手順としては以下になる。今、入力には信号 s_i が入ってきていて、行列 A によって変換を受け、我々の手元に出力 x_i が入ってきているとする。

- 出力の平均値をゼロにしておく $x_i \rightarrow x_i - E\{x_i\}$ 。
- 適当な線形変換 $z = Bx$ をかけて、whitening しておく ($C(xx^T) \rightarrow C(zz^T) = I$)
- z について、適当な重み $y = \sum_i z_i w_i$ で足し合わせ、非ガウス性を与える関数 G を使って s の非ガウス性を最大化するような係数 w_i を探す。

という流れで解析すればよいことがわかる。1 番目の中心化については、自明の操作なので、次に Whitening の具体的な操作を説明する。

1.4 Whitening

Whitening とは、相関関数の非対角成分を消し (無相関化) し、対角成分を 1 となるように x に線形変換を掛ける操作を表す。信号 x を適当な線形変換 B でもって、 $z = Bx$ と変換すると、 z の相関関数は以下ようになる。

$$C(z) = E\{zz^T\} = E\{Bxx^TB^T\} \quad (9)$$

$$= BE\{xx^T\}B^T = BC(x)B^T \quad (10)$$

ここで、 $C(x)$ は対称行列なので直交行列 X と対角行列 $D = \text{diag}(d_1, d_2, \dots)$ を用いて $C(x) = XDX^T$ とかける。

今 $D^{-1/2} = \text{diag}(d_1^{-1/2}, d_2^{-1/2}, \dots)$ とすると、 B としては、 $B = XD^{-1/2}X^T$ と選べば、

$$C(z) = BC(x)B^T = XD^{-1/2}X^T XDX^T XD^{-1/2}X^T = I \quad (11)$$

となるので、Whitening 成功していることがわかる。実用的には、 $C(x)$ を対角化すれば、 X (固有ベクトルが並んでいて、正規直交基底) と D が得られるし、あるいは $C(x)$ を特異値分解 (SVD) をすると、自動的に $C(x) = XDX^T$ の形になる。適当に python の linear algebra とかで実装しよう。

1.5 非ガウス性を最大にする

最後に、 z について、適当な重み $y = \sum_i z_i w_i$ で足し合わせ、分布の非ガウス性を計算する関数 $G(s = \sum_i z_i w_i)$ を最大化しよう。いま G の微分を g 、その微分を g' と書こう。いま、独立な信号 y 間の相関関数も単位行列で表されているとすると、

$$E\{y^2\} = \sum_{i,j} w_i w_j E\{z_i z_j\} = \sum_i w_i^2 = 1 \quad (12)$$

となり、 w_i の大きさが 1 であるという制限がつくことに注意する。まず、 $G(w^T z)$ を最大化するために、素朴に勾配法で最適な w を探そう。 G の w 微分は

$$\frac{\partial G(w^T z)}{\partial w} = zg(w^T z) \quad (13)$$

なので、なんらかの学習パラメータ γ を用いて

$$\Delta w = \gamma E\{zg(w^T z)\} \quad (14)$$

$$w \rightarrow \frac{w + \Delta w}{|w + \Delta w|} \quad (15)$$

と更新しまくれば、 G を最大化にする w が得られる。 G には、例えば

$$G_1(y) = (E\{\frac{1}{a_1} \log \cosh a_1 y\} - \text{const})^2 \quad (16)$$

$$G_2(y) = (E\{-\exp(-y^2/2)\} - \text{const})^2 \quad (17)$$

などがよく用いられるようだ (a_1 は 1 ~ 2 くらいの定数)。これは、のちの計算するが、非ガウス性を評価する尺度 "negentropy" を近似的に評価したものである。"const" が邪魔に見えるが、 G_i

を微分してわかるように、結局中の関数 $\tanh$ とかを最小にすればよい。なので、 G としては単に $G = \log \cosh y$ 、 $\exp(-y^2/2)$ を選べばよい。また、上の関数展開は、いかなる関数でもって同じような展開になるので、結局任意の関数 (定数とかではだめだが) をもってきて、その極値をもとめればよい事が知られている。

さて、上のアルゴリズム式 (14)、式 (15) の大きな問題点は学習率 γ の値を、何らかの形で与えなければいけないことにある。しかし、次に紹介する fast fixed-point algorithm (不動点アルゴリズム) を用いれば、その心配はない。まず、先ほどと同じように勾配法を用いるが、今度は $\|w\|^2 = 1$ という制限も一緒に入れて、未定乗数法を用いてアルゴリズムを作成する。まず次のラグランジアンを用意する。

$$L = E\{G(y)\} + \beta(\|w\|^2 - 1) \quad (18)$$

$$\frac{\partial L}{\partial w} = E\{zg(y)\} + \beta w \quad (19)$$

非ガウス性が最大になるのは $\partial L / \partial w = 0$ となる時であり、これを Newton 法で解こう。ということ

$$F = E\{zg(y)\} + \beta w \quad (20)$$

$$\frac{\partial F}{\partial w} = E\{zz^T g'(y)\} + \beta I \quad (21)$$

としておく。今 $E\{zz^T g'(y)\} \simeq E\{zz^T\}E\{g'(y)\} = E\{g'(y)\}I$ と近似すると F は単位行列に比例するので

$$w_{i+1} = w_i - \left(\frac{\partial F}{\partial w}\right)^{-1} F \quad (22)$$

$$= w_i - \frac{E\{zG(y)\} + \beta w_i}{E\{g'(y)\} + \beta} \quad (23)$$

$$= \frac{E\{zg(y)\} - E\{g'(y)\}w}{E\{g'(y)\} + \beta} \quad (24)$$

w_{i+1} の規格化は後でやればよく、大事なのは方向である。上式をよく見ると $E\{zG(y)\} - E\{g'(y)\}w$ が方向を定めているが、これは β に依存していないため、めでたく学習則に依存しないアルゴリズムができた。ということ

$$\begin{aligned} w_{i+1} &= E\{zg(y)\} - E\{g'(y)\}w_i \\ w_{i+1} &\rightarrow w_{i+1}/|w_{i+1}| \end{aligned} \quad (25)$$

という操作を繰り返していけばよい。

1.6 複数の独立な成分を見つける

前節では、ある w の初期値に対応する、ある一つのベクトル w を見つけるアルゴリズムを紹介した。実際は、独立な成分は他にもあるわけで、全てのベクトルを見つけれられるかどうか少し心配である。まず、異なる解 y_i と y_j 、それに対応する w_i と w_j が見つかったとする。元の信号が $E\{y_i y_j\} = \delta_{i,j}$ を満たしていると仮定すると、データは whitening されているので、 $E\{xx^T\} = I$ を満たし、

$$E\{y_i y_j\} = \sum_{l,m} E\{z_l w_{i,l} z_m w_{j,m}\} = w_i^T w_j = \delta_{i,j} \quad (26)$$

の関係がある。よって独立な信号に対応する、重みベクトル w は正規直交化されている。ということで、一個ずつ順番に成分を見つけていく方法をまず思いつく

1. 独立な成分の個数を m に決めておく。 $p = 1$ とする。
2. w_p を適当に選んでくる。
3. 式 (25) の操作を 1 回行い、 w_p を更新する。
4. w_p から、 $w_i (i = 1 \sim p - 1)$ の成分を引く。つまり

$$w_p \leftarrow w_p - \sum_i^{p-1} (w_p^T w_i) w_i \quad (27)$$

として、今までに見つかった成分とは直交しておくようにする。

5. norm を 1 にしておく。
6. もし w_p がまだ収束していなかったら、3 に戻る。
7. p を 1 増やし、2 に戻る。もし $m < p$ となったら、アルゴリズムを終了する。

独立な成分数が少なければこれでも問題ないが、式 (27) を見てわかるように p が大きくなるにつれて誤差が伝播してしまうので、成分が多いとちょっと微妙である。ということで、初めから m 本の w のベクトルを用意して行列 W としておいて、 W が直交行列になるように (w が正規直交系を張っているため) 更新するほうがよさそうである。ということで、まず $w_i (i = 1 \sim m)$ をランダムに選んで、各 w について式 (25) で更新して、最後に W が直交行列になるように変換する、というアルゴリズムを W が収束するまで繰り返すと、望みの結果が得られるだろう。行列の直交化は sec4 whitening の節で定義した $A^{-1/2}$ を用いて

$$W \leftarrow (WW^T)^{-1/2}W \quad (28)$$

とすれば、直交に近づいていく。

1.7 まとめ

とりあえず、ここまでの議論をまとめよう。用いる仮定を以下に並べる

仮定、前準備

1. m 個の信号 s_i が互いに統計的に独立である。
2. 入力 s 及び出力 x の数は等しく m 個ある。
3. 信号 s の相関行列は、 $E\{ss^T\} = I$ を満たすとする。
4. データ x は簡単な線形モデル $x = As$ に従っているとする。
5. 非ガウス性を教えてくれる G を用意しておく

もし $E\{ss^T\} \neq I$ なら、適当な線形変換 $s = Bs'$ をして、 $A \leftarrow AB$ とすれば、問題ない。観測できるのは x だけなので、極端な話 $x = \alpha A(s/\alpha)$ というスケール変換の自由度がある。3 はそういう自由度を消したり、計算が楽になる目的で入れてある。基本的には独立成分分析では、信号の大きさには興味なく、異なる信号を分離することに興味があることに気をつける。

そして、具体的な独立成分分析のアルゴリズムは以下の通りになる。

独立成分分析のアルゴリズム

1. $x \rightarrow x - E\{x\}$ として、 x の平均を 0 にしておく。
2. $z = Bx$ と線形変換して、Whitening をかけておく。具体的には観測 x の相関行列を $E\{xx^T\} = XDX^T$ と対角化して、 $z = XD^{-1/2}X^T$ とすればよい。
3. 大きさ 1 のベクトル $w_i (i = 1 \sim m)$ を用意しておく。それらを並べた行列が W である。
4. w_i にそれぞれ、式 (25) の操作を施しておく。
5. $W \leftarrow (WW^T)^{-1/2}W$ として、 W を直交行列に近づける
6. W が収束していれば終わり、しなければ step 4 に戻る。

これで独立成分分析の基本的な部分は説明し終わったはずである。関数 G を定めるところ以外には、一つの不定性もないはずである。次に一応数式的におっいていこう。

$$x = As \quad (29)$$

両辺の平均値を引いておくと

$$x - E\{x\} = A(s - E\{s\}) \quad (30)$$

$s - E\{s\}$ の相関行列は対角化されている。今その標準偏差を行列 $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots)$ であらわすとする

$$x - E\{x\} = A\Sigma(\Sigma^{-1}s - \Sigma^{-1}E\{s\}) \quad (31)$$

と書き直せる。ここで、 $\Sigma^{-1}s - \Sigma^{-1}E\{s\}$ が相関行列が単位行列で独立な信号を表している。次に、Whitening で構成した行列を B 、ICA で構成した行列を W とすると

$$WB(x - E\{x\}) = WBA\Sigma(\Sigma^{-1}s - \Sigma^{-1}E\{s\}) \quad (32)$$

とかける。 $WBA\Sigma$ は、きちんと言っていないが、単位行列に一致するわけではなく、信号の順番の自由度 (置換行列の自由度)、そして符号の自由度 (+1 or -1) を持つ。なので実際に解析して得られる $WB(x - E\{x\})$ のある行は $\Sigma^{-1}s - \Sigma^{-1}E\{s\}$ の中のある信号 $\times(+1, -1)$ に一致する。これは事前にはわからないので、何が自分が欲しい信号かは、解析した後に信号をみて確認すればよい。また、元々の信号 s を推定することはできない ($x = Aa(1/a)s$ という不定性があるため) が、

$$WBA\Sigma(\Sigma^{-1}s) = WB(x - E\{x\}) + WBE\{x\} \quad (33)$$

とすれば、offset を足せて、元の信号の定数倍の不定性のところまで求めることができる。なので、アルゴリズム 1 で平均化した時に、ちゃんと平均値を記録しておくことは重要である。

2 試運転編

さて、前節までの内容で線形な場合の独立成分分析の枠組みは完成した。次は実際に使ってみて、その性能を試してみたい。そのために、一番簡単なケースとして 2 信号の入力について、適当に線形で混ぜ合わせて、元の信号を抜きだすことが出来るのか計算してみよう。また、

2.1 ノイズなし

まずは、ノイズなしでやってみよう。 n を整数として、次のような周波数の異なる信号を二つ用意する。

$$x_1(n) = \sin(0.04n + 0.59) + 0.3 \quad (34)$$

$$x_2(n) = 0.3 \sin(0.07n + 1.2) \quad (35)$$

という信号を用意する。

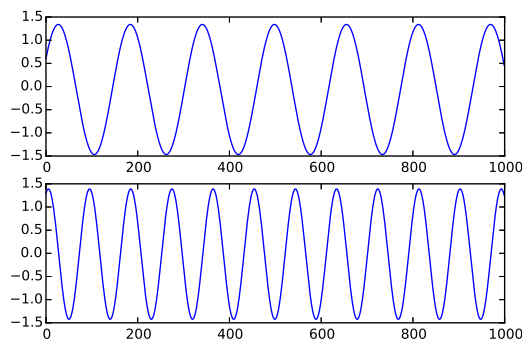


図2 ノイズなし信号。上が式 (34)、下が式 (35)

これを線形に混ぜて、信号をつくる

$$d_1(n) = x_1(n) + 2x_2(n) \quad (36)$$

$$d_2(n) = x_1(n) - x_2(n) \quad (37)$$

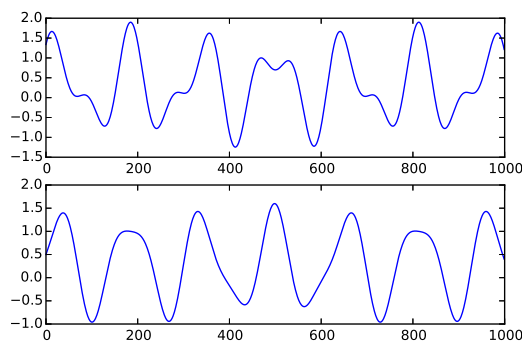


図3 ノイズなし信号を線形で混ぜて作成した擬似データ。上が式 (36)、下が式 (37)

これで、データを作り終わったので、ICA の準備をする。まず、平均を計算すると $E\{d_1\} = 0.35290595$ 、 $E\{d_2\} = 0.3419007$ となるので、これは d_1, d_2 から引いておく。次に、Whitening をかける。Whitening の行列は以前説明したように特異値分解すれば、自然に計算できる (図 4)。次に ICA をかけて、独立な成分抜きだそう。ここでは、非線形な関数として、 $G(x) = \tanh x$ を用いて、式 (25) で、ループを回した。ループは、ICA の行列の最大値の相対誤差 (一回まえとの差

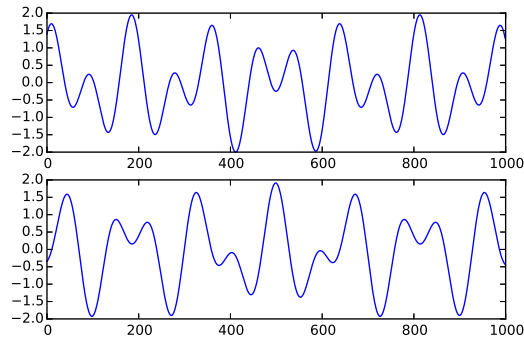


図4 データに Whitening をかけた。

を自身の値で割ったもの) が 10^{-6} を下回ったら止まるようにした。以下に抜き出した信号と元の信号 (の平均を 0、分散を 1 にしたもの)、そしてそれらの差をプロットする (図 5)。これは式 (32) に対応する。

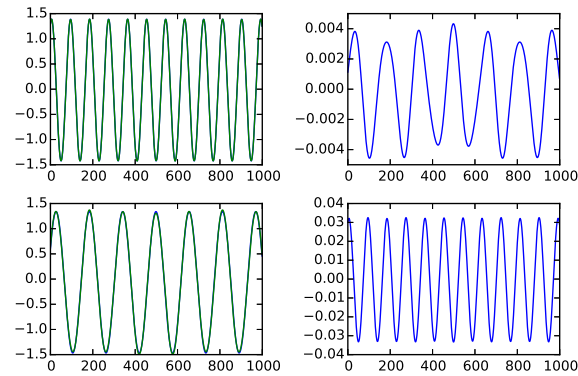


図5 左列:元信号と ICA で推定した信号。右列:正しい信号と推定信号の差。

最後に、オフセットを乗せて、きちんと等倍の不定性の除いては一致することをみよう。これは式 (33) に対応する。

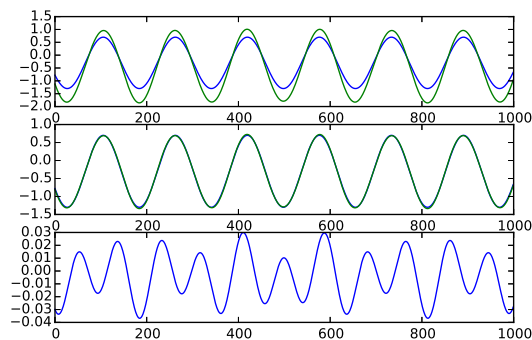


図6 上:元の信号とオフセットを乗せた信号。スケール倍の自由度を除いて一致。中:スケール倍を計算して、スケールした。下:中で描いた二つの信号の差

2.2 ノイズあり

次に、適当にノイズをいれてシミュレーションしなおす。今回は、ある星 X が食を起こして減光している様子をたまたま観測している状況を仮定する。ただし、食とは関係ない別の星 Y が近くにあるとする。ノイズはラプラス分布に従わせた。信号の混ぜ合わせ方は

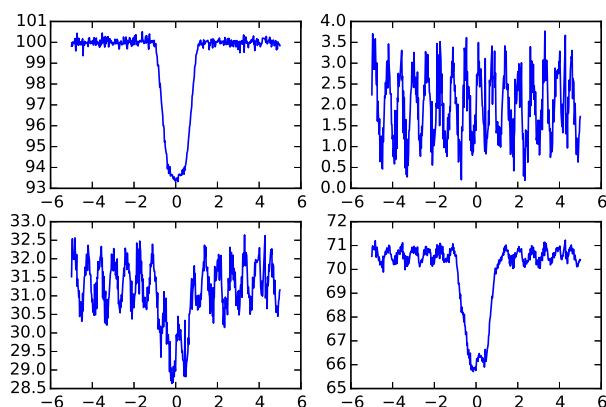


図7 上段:元の信号。左が食の光度曲線で右が擬似変光星。

下段:信号を線形に混ぜ合わせた。右が式 (38) で左が式 (39)

$$d_1(n) = 0.7x_1(n) + 0.3x_2(n) \quad (38)$$

$$d_2(n) = 0.3x_1(n) + 0.7x_2(n) \quad (39)$$

とした。これは、例えば、隣同士の CCD の取得した光度を観測したような状況である。X の中心が CCD1 に近く、Y の中心が CCD2 に近いとかそういう感じである。上図の下側が、混ぜ合わせた信号に対応する。

次に、これから、平均を除いて Whitenning をかけると、以下ようになる。

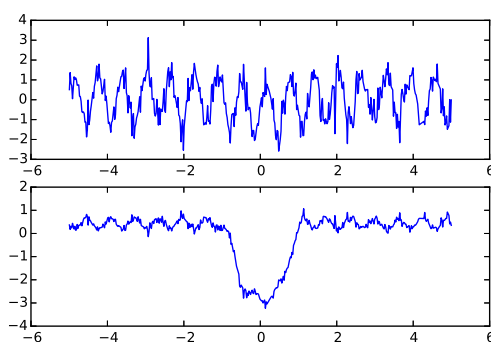


図8 Whitenning をかけた後の信号

そして ICA をかけよう。収束の条件は先ほどと同じである。

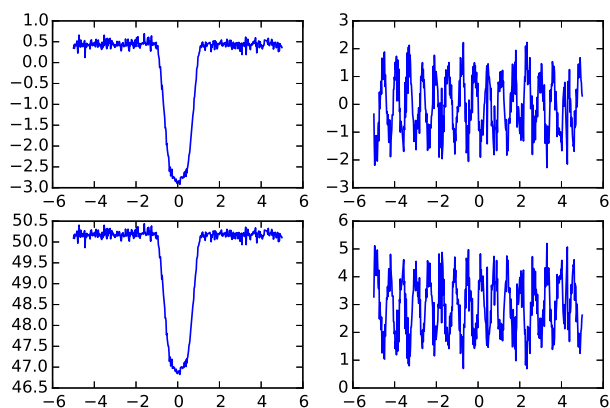


図 9 上段:ICA で分離した信号。下段:元の信号。

結構いい感じに抜き出せていることがわかる。最後に、オフセットをのせて等倍を調整して、元の信号と比較した。以上でテストは終わりである。もっと systematic に調べてもいいかもしれない

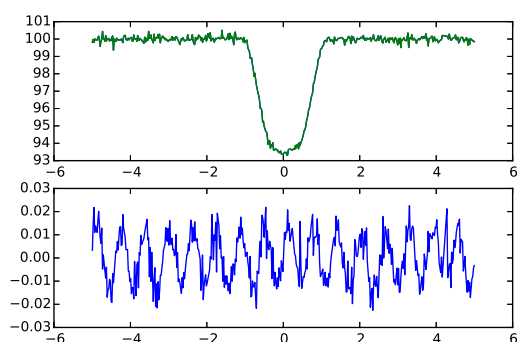


図 10 上段:ICA で分離した信号にオフセットを乗せた信号を適当にスケール変換したものと、元の信号。下段:二つの差。

が、とりあえず試運転としては十分な感じである。

3 非ガウス性の定式化編

ずっと先送りにしてきたが、ここでは関数 G の選び方の基準である非ガウス性を定式化する。分布はキュムラントという量で完全に決まる。そこでキュムラントを導入しつつ、議論を進めよう。最終的には、情報エントロピーを持ち出して非ガウス性を議論する。

3.1 キュムラントとガウス分布

まず、ある確率変数の分布 $f(x)$ に対応した特性関数を以下のように定義する

$$\Phi(\xi) = \int_{-\infty}^{\infty} e^{i\xi x} f(x) dx \quad (40)$$

$$= \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(i\xi)^n}{n!} x^n f(x) dx \quad (41)$$

$$= \sum_{n=0}^{\infty} \frac{(i\xi)^n}{n!} \langle x^n \rangle \quad (42)$$

たとえば平均 μ 、標準偏差 σ のガウス分布の特性関数は、上式に代入して

$$\Phi(\xi) = \exp(im\xi - \frac{\sigma^2}{2}\xi^2) \quad (43)$$

となる。ここで、任意の n 次のモーメント

$$\langle x^n \rangle = \int_{-\infty}^{\infty} x^n f(x) dx \quad (44)$$

が有限の値を持っているとすると、特性関数は次のように展開できる。

$$\Phi(\xi) = \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} \frac{(i\xi)^n}{n!} x^n f(x) dx = \sum_{n=0}^{\infty} \frac{(i\xi)^n}{n!} \langle x^n \rangle \quad (45)$$

ここで、キュムラント関数 $\Psi(\xi)$ とキュムラント $\langle x^n \rangle_c$ を以下のように定義する。

$$\Psi(\xi) = \ln \Phi(\xi) = \ln \sum_{n=0}^{\infty} \frac{(i\xi)^n}{n!} \langle x^n \rangle = \sum_{n=0}^{\infty} \frac{(i\xi)^n}{n!} \langle x^n \rangle_c \quad (46)$$

キュムラントとモーメントは互いに移り変わることができる。以下にその一例を紹介する。

$$\langle x \rangle_c = \langle x \rangle \quad (47)$$

$$\langle x^2 \rangle_c = \langle x^2 \rangle - \langle x \rangle^2 \quad (48)$$

$$\langle x^2 \rangle = \langle x^2 \rangle_c + \langle x \rangle_c^2 \quad (49)$$

$$\langle x^3 \rangle_c = \langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3 \quad (50)$$

$$\langle x^3 \rangle = \langle x^3 \rangle_c + 3\langle x \rangle_c \langle x^2 \rangle_c + \langle x \rangle_c^3 \quad (51)$$

特に 2 次のキュムラントは分散に対応する。たとえば、ガウス分布のキュムラントは式 (43) から簡単に計算できて

$$\langle x \rangle_c = \mu, \quad \langle x^2 \rangle_c = \sigma^2, \quad \langle x^n \rangle_c = 0 \quad (n \geq 3) \quad (52)$$

ガウス分布の場合、モーメントだと偶数次の項が残るが、キュムラントを用いれば 3 次以上の項が全て落ちる。逆にいうと、3 次以上のキュムラントが強ければ強いほどガウス分布から離れていく。なので、キュムラントが非ガウス性の一つの指標となりそうである。

少し横道にそれるがキュムラントを用いて、中心極限定理を証明しておこう。今 X_n が、独立な信号 ΔX_i の足し合わせであるとする

$$X_n = \sum_i \Delta X_i \quad (53)$$

ΔX_n が独立であれば、その特性関数はそれぞれの ΔX_i の特性関数の直積になる。

$$\Phi_n(\xi) = \langle e^{i\xi X_n} \rangle = \prod_i^n \langle e^{i\xi \Delta X_n} \rangle \quad (54)$$

すると、キュムラント関数は $\log$ なので掛け算を足し算にできて

$$\Psi_n(\xi) = \sum_j \psi_j(\xi) = \sum_{j=1}^n \sum_{k=0}^{\infty} \frac{(i\xi)^k}{k!} \langle \Delta X_j^k \rangle_c \quad (55)$$

今、 ΔX_i の平均が 0 で分散が σ_i^2 であるとする。そこで分散の総和として

$$s_n^2 = \sum_i \sigma_i^2 \quad (56)$$

と定義しておく。すると、キュムラント関数は次のように書き直せる

$$\Psi_n(\xi) = -\frac{\xi^2}{2} s_n^2 + \frac{(i\xi)^3}{3!} \sum \langle \Delta X_j^3 \rangle_c \dots \quad (57)$$

今 X_n の代わりに、 $Y_n = X_n/s_n$ の特性関数を考えると、 X_n の特性関数の ξ を $\eta = \xi/s_n$ とすればよいので、

$$\Psi_n(\eta = \xi/s_n) = \exp \left(-\frac{1}{2} \eta^2 + \frac{(i\eta)^3}{3!} \sum_j \frac{\langle \Delta X_j^3 \rangle_c}{s_n^3} \dots \right) \quad (58)$$

となる。 ΔX_j のモーメントがすべて同程度で有限であるとし、 s_n^2 が n の増大とともに、 n のオーダーで増えていくことを考慮すると m 次のキュムラントの項は

$$\frac{\mathcal{O}(n)}{\mathcal{O}(n^{m/2})} \rightarrow 0 \quad (59)$$

となる。なので、 $m \geq 3$ の項は落ちていき、信号を足し合わせると非ガウス性が失われていく。これが、中心極限定理の大雑把な証明である。

3.2 非ガウス性の指標としての kurtosis

キュムラントの中で、3 次のキュムラント (skewness) は分布の非対称性を表す。また、4 次のキュムラント (kurtosis) は分布の中心集中度を与える。これらの絶対値の大きさが大きければ非ガウス性は強くなる。 G として、これらの値を選んでもよい。特に重要なポイントとしては、式 (55) を見てわかるように、キュムラントは加法的であるので $y = q^T s$ とすると、

$$F(q) = kurt(y) = \sum_i q_i^4 kurt(s_i) \quad (60)$$

とかける。この値の絶対値の大きさが非ガウス性を与える。ちなみに、平均が 0、分散 1 の信号の 4 次のキュムラントは $kurt(x) = \langle x^4 \rangle - 3$ である。一般に $F(q)$ を $\|q^2\| = 1$ という条件の下で最大化すると、その極値は q_i のいずれかが 1 で他が 0 である時に得られる事が証明できる。(ICA textbook 参照のこと)。このように、kurtosis を用いると、一応厳密に非ガウスな信号を分離することができる。しかし、kurtosis は outlier に弱いことに知られている。たとえば、ある平均 0 分散 1 の乱数を 100 回サンプルして、その内一つが 10 だとする。10 は一つの要素でしかない

が、kurtosis に $10^4/100 = 100$ の影響を与える。10 が一つあるだけで、100 個の乱数と同じだけの影響力を持ってしまう。このように、実際の観測では避けられない、突発的な outlier に対して kurtosis は強くないため、あまり実用されていない (と教科書で書いてるが、宇宙系の論文を見ると結構使われていますね。)。

3.3 非ガウス性の指標としての negentropy

次に、outlier に弱い kurtosis に変わる他の非ガウス性に指標を探そう。唐突だが、ある確率密度を持つ $p(x)$ に対して以下の量を定義する

$$H(x) = - \int p_x(\eta) \log p_x(\eta) d\eta \quad (61)$$

これは、情報エントロピーと呼ばれる量で、その名の通り情報理論では基本的な概念である。実は分散が 1 である分布の中で、ガウス分布は $H(x)$ を最大にする。なので、 $H(x)$ を非ガウス性の尺度として用いることができそうである。ということで、ある確率変数 x について、分散 1 のガウス分布の情報エントロピーを $H(x_{\text{gauss}})$ とすると、以下の量が、 $p(x)$ とガウス分布との距離を示してくれそうである。

$$J(x) = H(x_{\text{gauss}}) - H(x) \quad (62)$$

この量は negentropy と呼ばれ、これを最大化することが非ガウス性を最大化することと対応する。あとは、 x に対応する $H(x)$ を評価すればよいのだが、式 (61) を直接評価するのは困難らしい。そこで、 $p(x)$ がガウス分布に近い、ということを仮定して、分布を多項式あるいは非多項式で近似し、最低次で $H(x)$ を評価しよう。

3.4 確率分布関数の多項式展開を用いた情報エントロピーの近似

確率分布関数の多項式展開としては、Edgeworth 展開と Gram-Charlier 展開が有名である。ここでは後者を使う。今、ガウス分布を $p_N(x)$ と書き、

$$\frac{p(x)}{p_N(x)} = \sum_{s=0}^{\infty} c_s He_s(x) \quad (63)$$

と展開できたとしよう。ここで He_s は s 次のチェビシェフ多項式である。 $He_s(x)$ は次の性質を持っている

$$\int p_N(x) He_i(x) He_j(x) dx = \delta_{i,j} \quad (64)$$

なので、関数展開の係数は c_s は $He_s(x)$ の期待値に対応する。

$$c_s = \int p(x) He_s(x) dx \quad (65)$$

$He_s(x)$ は s 次の多項式なので、これはキュムラント (モーメント) の期待値の和で書くことができ、たとえば x の平均が 0 で、分散が 1 であるとして、4 次までの展開を書き下すと

$$p(x) \simeq p_N(x) \left(1 + \langle x^3 \rangle_c \frac{He_3(x)}{3!} + \langle x^4 \rangle_c \frac{He_4(x)}{4!} \dots \right) \quad (66)$$

となる。。実は He_s に対応する係数は $\langle x^s \rangle_c$ になるのである。

上に示した展開を用いて、情報エントロピーを近似的に導出しよう。今 $\langle x^3 \rangle_c$ や、 $\langle x^4 \rangle_c$ が 1 より十分小さいと仮定すると、 $\log(1+x) = x - x^2/2\ldots$ を用いて $H(x)$ を計算して

$$H(x) \simeq \int p_N(x) \left(1 + \langle x^3 \rangle_c \frac{He_3(x)}{3!} + \langle x^4 \rangle_c \frac{He_4(x)}{4!} \right) \quad (67)$$

$$[\ln p_N(x) + \left(\langle x^3 \rangle_c \frac{He_3(x)}{3!} + \langle x^4 \rangle_c \frac{He_4(x)}{4!} \right) - \left(\langle x^3 \rangle_c \frac{He_3(x)}{3!} + \langle x^4 \rangle_c \frac{He_4(x)}{4!} \right)^2 / 2] \quad (68)$$

$$= - \int p_N(x) \ln p_N(x) dx - \frac{\langle x^3 \rangle_c^2}{2 \times 3!} - \frac{\langle x^4 \rangle_c^2}{2 \times 4!} \quad (69)$$

となる。ここで、チェビシェフの多項式の直交性を用いた。よって、negentropy は

$$J(x) = - \int p_N(x) \ln p_N(x) dx - H(x) \simeq \frac{1}{12} E\{x^3\}^2 + \frac{1}{48} \text{kurt}(x)^2 \quad (70)$$

とかける。このように、kurtosis と negentropy には密接な関係がある。しかし、前項でも説明したように、kurtosis は outlier に非常に弱いので、何か別の尺度を見つける必要がある。

3.5 確率分布関数の非多項式展開を用いた情報エントロピーの近似

次に outlier に強い展開をするために、非多項式を用いて確率分布関数を展開しよう。まず、関数 $G^i(x)$ の期待値が得られているとする。

$$\int p(x) G^i(x) = a_i \quad (71)$$

上の条件 Eq(71) を満たす分布は無数にある。束縛条件以外に情報を持っていないのなら、その分布候補の中でもっとも不確かさ（エントロピー）が大きい分布が確からしいとする標語で攻めるのが最大エントロピー法である。エントロピーは式 (61) で定義されている事を思い出そう。

さて、束縛条件の下である量を最大化するには、Lagrange 未定乗数法を用いればよい。ということで、未定乗数を λ_i とすると

$$\int -p(x) \ln p(x) + \sum_i \lambda_i p(x) G^i(x) dx \quad (72)$$

を最大にする、分布を求める。このとき上式について、 $p(x)$ について変分をとると

$$\int (-\ln p(x) - 1 + \sum_i \lambda_i G^i(x)) \delta p(x) dx \quad (73)$$

となるので、 $\delta p(x)$ の左側がゼロになる分布であれば、最大値をとることがわかる

$$p(x) = e^{-1} (1 + \sum_i \lambda_i G^i(x)) \quad (74)$$

なので、あとは λ_i を Eq(71) に従って求めれば分布がもとまる。たとえば、束縛条件として、分散が σ 平均が μ であると要請すると、式 (74) から分散 σ 平均 μ のガウス分布が求まる。

さて、一般の分布について、式 (74) を用いて考察しよう。ただし、分布が Gauss 分布に近いと仮定して、関数の組 $(G^i(x))$ (これは非多項式でも構わない) でガウス分布からのズレを表現する。ガウス分布に近いので λ_i は x^2 の係数以外 1 より十分小さいだろう。ということで、ガウス分布の分

の係数だけ外にだしといて、 $|\lambda_i| \ll 1 (i = 1, \dots, n+2)$ を満たす。ここで $G^{n+1} = x$ 、 $G^{n+2} = x^2$ とした。すると、分布は次のようになる。

$$\begin{aligned} p(x) &\simeq A \exp \left(-x^2/2 + \lambda_{n+1}x + (\lambda_{n+2} + 1/2)x^2 + \sum_{i=1}^n \lambda_i G^i(x) \right) \\ &\simeq \bar{A} p_N(x) \left(1 + \lambda_{n+1}x + (\lambda_{n+2} + 1/2)x^2 + \sum_{i=1}^n \lambda_i G^i(x) \right) \end{aligned} \quad (75)$$

今分布が分散が 1 で平均が 0、束縛条件 (71) を満たしていると仮定する。また、 G^i については、正規直交な関係を満たしているとする。これは、 G^i を適当な線形結合で直交になるようにすれば実現できる (Gram Schmit とか)。

$$\int p_N(x) G^i G^j = \delta_{i,j} \quad (76)$$

すると、条件としては以下のものが要請される。

$$\int p(x) dx = \bar{A}(1 + \lambda_{n+2} + 1/2) = 1 \quad (77)$$

$$\int p(x) x dx = \bar{A} \lambda_{n+1} = 0 \quad (78)$$

$$\int p(x) x^2 dx = \bar{A}(1 + 3(\lambda_{n+2} + 1/2)) = 1 \quad (79)$$

$$\int p(x) G^i(x) dx = \bar{A} \lambda_i = a_i \quad (80)$$

以上を解いていくと、 $\bar{A} = 1$, $\lambda_{n+1} = 0$, $\lambda_{n+2} = -1/2$, $\lambda_i = a_i$ であることがわかる。このように、期待値がそのまま展開係数になる。

ということで、まとめておくと、ある非線形関数の組み G^i の期待値が a_i で、分布がガウス分布に近いとすると、分布は最大エントロピー法を援用して次のように近似できる

$$\begin{aligned} p(x) &= p_N(x) \exp \left(\sum_i a_i G^i \right) \\ &\simeq p_N(x) \left(1 + \sum_i a_i G^i \right) \end{aligned} \quad (81)$$

さて、次にこの式を用いて、negentropy を計算しよう。 $J(x)$ は今

$$H(x) = \int -p(x) \ln p(x) \quad (82)$$

$$= - \int p_N(x) \left(1 + \sum_i a_i G^i \right) \left(\ln p_N(x) + \ln \left(1 + \sum_i a_i G^i \right) \right) \quad (83)$$

$$\simeq - \int p_N(x) \left(1 + \sum_i a_i G^i \right) \left(\ln p_N(x) + \sum_i a_i G^i - \frac{1}{2} \left(\sum_i a_i G^i \right)^2 \right) \quad (84)$$

$$\simeq - \left[\int p_N(x) \ln p_N(x) + p_N(x) \ln p_N(x) \sum_i a_i G^i + p(x) \sum_i a_i G^i - p(x) \frac{1}{2} \left(\sum_i a_i G^i \right)^2 \right] \quad (85)$$

$$= H(\nu) - \left(0 + \sum_i a_i^2 - \frac{1}{2} \sum_i a_i^2 \right) = H(\nu) - \frac{1}{2} \sum_i a_i^2 \quad (86)$$

今、ガウス分布の変数を ν とした。 $H(\nu)$ はガウス分布のエントロピーを表す。そして、negentropy は以下ようになる。

$$J(x) \simeq \frac{1}{2} \sum_{i=1}^n E\{G^i(x)\}^2 \quad (87)$$

となる。この結果は、前節の多項式による展開の結果を包含した結果 [式 (70)] になっている。

以上で枠組みは完成した。実際にある関数系で展開したいときは、適当な関数系 G^i と x, x^2 を組みあわせて、それぞれが正規直交な関係になるように、線形結合で関数系を新しく F^i と取り直して、その期待値

$$\int p(x) F^i(x) dx \quad (88)$$

を計算すれば、その期待値の二乗和が negentropy の近似になっている。このように、関数の期待値は negentropy と密接に関係して、適当な関数の期待値を最大化するようにすれば、非ガウス性を大きくする方向に向かうのだ。

さて、ある非線形関数 $G(x)$ をもってきて、それで展開するとする。関数系が正規直交系をみたすには G, x, x^2 を適当に組み合わせればよいが、 x, x^2 は既に正規直交なので、 $G \rightarrow (G + ax + bx^2)c$ とかに変えて、正規直交となるように係数 a, b, c を定めればよい。すると、結局 negentropy は少しの計算ののち

$$J(x) = (E\{G(x)\} - E\{G(\nu)\})^2 \quad (89)$$

と近似されることが分かる。今 $E\{G(\nu)\}$ は、分布がガウス分布のときの $G(x)$ の期待値である。これは、式 (16)、式 (17) に対応する。あるいは、一つの関数だけでなく、たとえば適当な奇関数 $G^1(x)$ と偶関数 $G^2(x)$ のペアを持ってきて、 G^1, G^2, x, x^2 の組み合わせで正規直交系を作って、 $J(x)$ を近似してもよい。たとえば多項式のときの近似式 (70) はこの形になっている。

以上で本文で端折った非ガウス性のところは、ある程度説明できただろう。