

以下、雑誌、レスピレーションリサーチファンデーション「呼吸」（Surrogate data法を用いた生体信号解析、
鱒岡直人、網崎孝志）から引用。

文献はメディカルオンライン <http://www.medicalonline.jp/> から購入可能。

時系列データ解析は生体の生理学的現象から生じた生体信号を解析するための重要な手法である。

時系列データは等間隔にすると解析が容易になる。

時間間隔が等間隔な時系列データ $u(i)$, ($i = 1, 2, 3, \dots, n$) を考える。 $u(i)$ には信号値が入る。

等間隔時系列データから Takens の埋め込み定理を用いた m 次元空間用の再構成ベクトルセット X_i , ($i = 1, 2, 3, \dots, n - (m - 1)\tau$) は以下ようになる。

$$X_i = \{u(i), u(i + \tau), u(i + 2\tau), \dots, u(i + (m - 1)\tau)\}, (i = 1, 2, 3, \dots, n - (m - 1)\tau),$$

時間遅れ座標系の時間遅れ (time lag) の τ は正の整数の定数である。

τ は時系列データの自己相関関数が最初に $1/e$ に減衰する時間を参考にすることが多い。

時間遅れの実際の時間は $\tau \times \Delta t$ となる。 Δt はサンプリング時間である。

時間遅れの τ を導入して拡張した ApEn も報告されている。

以下に式を示す。

$$\begin{aligned} \text{ApEn}(n, m, r, \tau) &= \Phi^m(r) - \Phi^{m+1}(r) \\ &= (n - (m - 1)\tau)^{-1} \cdot \sum_{i=1}^{n-(m-1)\tau} \ln C_{m,i}(r) - (n - m\tau)^{-1} \cdot \sum_{i=1}^{n-m\tau} \ln C_{m+1,i}(r), \\ C_{m,i}(r) &\text{は以下で定義される。} \end{aligned}$$

$$\begin{aligned} C_{m,i}(r) &= (n - (m - 1)\tau)^{-1} \cdot \sum_{j=1}^{n-(m-1)\tau} \Theta(r - |X_i - X_j|) \\ |X_i - X_j| &\text{はベクトル } X_i \text{ とベクトル } X_j \text{ の距離である。} \end{aligned}$$

Pincus は ApEn 考案時、 $\tau = 1$ の場合を考えた。 $\tau = 1$ のとき時系列データ $u(i)$ から作製した再構成ベクトルは、

$$X_i = \{u(i), u(i + 1), u(i + 2), \dots, u(i + (m - 1))\}, (i = 1, 2, 3, \dots, n - (m - 1)).$$

ApEn は以下のように計算される。

$$\begin{aligned} \Phi^m(r) &= (n - (m - 1))^{-1} \cdot \sum_{i=1}^{n-(m-1)} \ln C_{m,i}(r), \\ \text{ApEn}(n, m, r) &= \Phi^m(r) - \Phi^{m+1}(r) \\ &= (n - (m - 1))^{-1} \cdot \sum_{i=1}^{n-(m-1)} \ln C_{m,i}(r) - (n - m)^{-1} \cdot \sum_{i=1}^{n-m} \ln C_{m+1,i}(r), \end{aligned}$$

$C_{m,i}(r)$ は以下で表される。

まず、 $d[X_i, X_j] = \max_{k=1,2,\dots,m} (|u(i + (k - 1)) - u(j + (k - 1))|)$ を定義する。

$d[X_i, X_j]$ はベクトル X_i の要素、 $u(i + (k - 1))$ とベクトル X_j の要素、 $u(j + (k - 1))$ の差の絶対値のうち、埋め込み次元 m 以下の場合で最大の値を意味する。これを用いて、

$$C_{m,i}(r) = (n - (m - 1))^{-1} \cdot \sum_{j=1}^{n-(m-1)} \Theta(r - d[X_i, X_j]),$$

n は時系列データのデータ数、 m は埋め込み次元 (embedding dimension)、 r は定数、 $\ln$ は自然対数 (natural logarithm) である。

$d[X_i, X_j]$ は、ベクトル X_i とベクトル X_j 間の距離に関連する。

$\Theta(t)$ はヘビサイド関数で、 $t \geq 0$ ならば $\Theta(t) = 1$ とし、 $t < 0$ ならば $\Theta(t) = 0$ とする。

$\sum_{j=1}^{n-(m-1)} \Theta(r - d[X_i, X_j])$ はベクトル X_i を中心として X_i から一定の定数 r 内に $d[X_i, X_j] \leq r$ を満たす他のベクトル X_j ($j = 1, 2, 3, \dots, n - (m - 1)$) が何個含まれているかを意味している。
 もっと砕けて説明すると、 $C_{m,i}(r)$ は、ベクトル X_i を中心として、距離 r 内に他のベクトル X_j ($j = 1, 2, 3, \dots, n - (m -$

$1)$ が何個含まれるか ($\sum_{j=1}^{n-(m-1)} \Theta(r - d[X_i, X_j])$) を確率 (比率とも解釈可能) で表したものである ($n - (m - 1)$ で割っている)。

距離 r はユークリッドの距離を用いても本質は変わらない。

通常、 $m = 2$ 、 $r = 0.1 \sim 0.25 \times SD$ (SD は対象とした時系列データの標準偏差) の定数にして計算する。

$ApEn(n, m, r)$ の変数を定数とすることで計算速度が飛躍的に短縮でき、解釈もしやすくなった。

埋め込み次元 m を 2 にする、距離 r を定数にする。これらが Pincus が工夫した重要な点である。

さらに、詳細な説明。

定義式の意味として再構成ベクトルを m 次元の時間遅れ座標系の中に描写された点とすると、ある点 (ベクトル X_i) を中心として、一定の距離 r 内にどれくらい他の点 (ベクトル X_j , ($j = 1, 2, 3, \dots, n - (m - 1)$)) が含まれているかを確率 $C_{m,i}(r)$ として計算し、 $\ln C_{m,i}(r)$ とエントロピーの形にしている ($\ln$: 自然対数)。

$m = 2$ ならば半径 r の円の中に含まれる点の数 (ベクトル X_j , ($j = 1, 2, 3, \dots, n - (m - 1)$)) のうち、ベクトル X_i を中心とした半径 r の円内に何点存在するか)、 $m = 3$ ならば半径 r の球の中に含まれる点の数 (ベクトル X_j , ($j = 1, 2, 3, \dots, n - (m - 1)$)) のうち、ベクトル X_i を中心とした半径 r の球内に何点存在するか) を用いて、各々 $\ln C_{2,i}(r)$, $\ln C_{3,i}(r)$ を計算する。

符号はマイナスであるが選択情報量を計算したことになる。

さらに、 m 次元の時間遅れ座標系に埋め込まれた点 ($n - (m - 1)$ 個) すべてに対して同様の操作を行い ($\sum_{i=1}^{n-(m-1)} \ln C_{m,i}(r)$)、平均を計算した統計量 ($n - (m - 1)$ で割り算する) が $\Phi^m(r)$ である、この意味として情報量に関連した統計量の平均値を計算したことになる。

$$\Phi^m(r) = (n - (m - 1))^{-1} \cdot \sum_{i=1}^{n-(m-1)} \ln C_{m,i}(r).$$

式の解釈として、任意の再構成ベクトルを一点選んだとき、それを中心として一定の距離 r 内にどれくらい他の点が含まれているか (分布確率) をエントロピーの形として計算したら平均値としてどの程度なのかを示している。

次いで、エントロピーの符号を正にするためマイナス符号をつける ($-\Phi^m(r)$)。

そして、時間遅れ座標系の埋め込み次元 m が 1 のみ大きい場合 ($m+1$) と比較した値が $ApEn$ である

$$ApEn(n, m, r) = -\Phi^{m+1}(r) - (-\Phi^m(r)) = \Phi^m(r) - \Phi^{m+1}(r)$$

周期的で規則正しい時系列データならば、 m あるいは $m+1$ 次元で作成された再構成ベクトルは時間遅れ座標系での分布が同じ性質なので、

$ApEn(n, m, r) = \Phi^m(r) - \Phi^{m+1}(r) \cong 0$ となり規則正しい状態を判定できる。

逆に乱雑な時系列データの場合、値は大きくなる。

具体的に、生体信号からApEnを計算したとする。

ApEnは下限値と上限値が存在するので生体信号を定量化するのに便利な統計量である。

ApEnは下限が存在し下限値は0である。サインカーブやトーラスのような規則正しい時系列データならば0に近い値をとる。

全く乱雑な場合が上限値となる。ApEnの上限値は時系列データの個数 n に依存する。従って、時系列データの個数 n が異なった条件での解析は意味をなさないことに注意が必要である。

従って、生体信号をA/Dコンバーターで時系列データ化する長さ（時間）は同じでなければならない。

Matlabで発生させた一様乱数を時系列データとするとデータ数が600個でApEnは1.4前後、2000個でApEnは1.9前後となる。

カオスは乱雑でApEn値は大きいという誤解があるが、決定論的カオスから作成した1変数の時系列データ（x軸に正射影）で計算するとLorenzモデルApEn値は0.135、RösslerモデルでApEn値は0.218と小さい値であり、初期値には敏感であるが、決して乱数のように乱雑なわけではない。カオスは背景に明確な決定論をもつことを考えれば値が小さい結果に納得がいくと思われる。

カオスは極めて特殊な状態であり、このような数学的に特殊なものを生体に安易にあてはめると誤謬が生じる可能性がある。

参考文献

1. Pincus SM. Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA 88: 2279-301, 1991.
2. Burioka N, Cornélissen G, Halberg F, Kaplan DT, Suyama H, Sako T, Shimizu E. Approximate entropy of human respiratory movement during eye-closed wake and different sleep stages. Chest 123: 80-6, 2003.
3. Burioka N, Cornélissen G, Maegaki Y, Halberg F, Kaplan DT, Miyata M, Fukuoka Y, Endo M, Suyama H, Tomita Y, Shimizu E. Approximate entropy of the electroencephalogram in healthy awake subjects and absence epilepsy patients. Clin EEG Neurosci. 36: