2P129

階層構造をつくる多変数確率過程の数理

（北大 電子研[1],北大 理学院[2], 北大 生命院[3]）
○宮川尚紀[1,2], 寺本央[1,3], 李振風[1,2], 小松崎民樹[1,3]

【序】一分子解析のような複数個の要素が複雑なふるまいをみせる時系列データから、どのようにしてその背景に存在する情報が抜き出せるのだろうか。ひとつの方法として、時系列データから要素のふるまいを確率分布として抽出し、情報論的な解析を行うことができる。ここでは特に、元の時系列データからどのように要素が結びついているのか、その要素の相関を解析する方法を紹介し、一般化する。Scheidman らは、系を構成する要素の振る舞いの中から、情報理論に基づきk体相関の量を与える連結情報量(Connected Information)と呼ばれる量を提案し、サンショウウオの網膜神経節細胞において解析を行った。彼らは、サンショウウオに自然な映像を見せ、そこでの神経細胞の発火のパターンから確率分布を求め、そこに細胞が何次の相関を持って挙動しているのかを解析し、その細胞間のふるまいが 2 体の相関のみで記述するに十分であることを明らかにした[1,2]。しかし、この連結情報量は、系の要素のふるまいが例えば 2 体の相関によるものと示すだけで、系の中のどの 2 体の要素が相関し合っているか？という問いには答えない。そこで我々は、元の連結情報量が、各要素間の相関とどのように関連し合っているかを議論する。

【方法・結果】簡単のため、全変数が 0 か 1 かのバイナリーな値を取る 3 変数(x_1, x_2, x_3)とし、2次の連結情報量を導入する。3 体の確率分布$P(x_1, x_2, x_3)$が与えられたときに、その中の2変数の情報は次のように与えられる。

$$-H[P^{(2)}] \coloneqq -\max_{Q \in M^2} H[Q]$$

ここに、Hはエントロピー

$$H[P] \coloneqq \sum_x P(x) \log P(x)$$

であり、確率分布Pの不確かさを表す。$-H[P^{(2)}]$は、エントロピーを空間

$$M^2 \coloneqq \{Q(x_1, x_2, x_3) | \sum_{x_i} Q(x_1, x_2, x_3) = \sum_{x_i} P(x_1, x_2, x_3), \quad i \in \{1,2,3\}\}$$

において最大化した量の負符号である。$\sum_{x_i} P(x_1, x_2, x_3)$ は元の確率分布を1変数で足したものであり、2体の確率分布となる（これを2次のMarginalsと呼ぶ）。すなわち、この空間M^kは、元の確率分布と同じ2次のMarginalsを持つ確率分布$Q(x_1, x_2, x_3)$の集合である。これはいわば、元の確率分布から2変数のみに圧縮された情報を持つ確率分布の集合であり、エントロピーを最大に取るということは情報を最小に取ることに等しいことから、この$H[P^{(2)}]$は、元の確率分布の2体のみの情報であることが分かる。しかし、2次の情報を持つということは、同時に1体の情報も持つために、純粋な2次の情報は、一体の寄与を消した

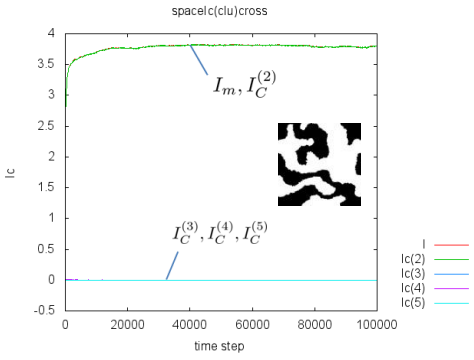
$$I_C^{(2)} \coloneqq \left(-H\left[P^{(2)}\right]\right) - \left(-H\left[P^{(1)}\right]\right)$$

であり、これを2次の連結情報量と呼ぶ。これは、確率分布$P(x_1, x_2, x_3)$に対する2変数の連結の度合いを表す情報量である。ここで、1次の情報$-H[P^{(1)}]$は2次の情報と同様に定義されるが、エントロピーを最大に取る空間は、

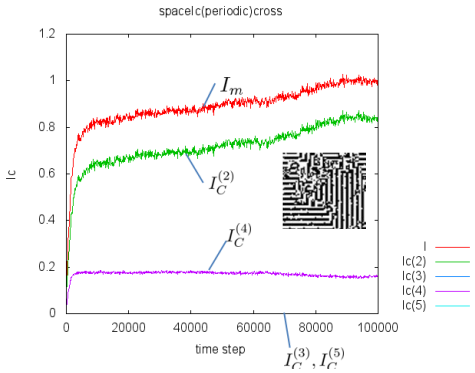
$$M^1 := \{Q(x_1, x_2, x_3) | \sum_{x_i, x_j} Q(x_1, x_2, x_3) = \sum_{x_i x_j} P(x_1, x_2, x_3), \qquad i < j \in \{1,2,3\}\}$$

のように、2変数で足し合わせた1変数の確率分布に関する条件となる。

これを実際に、Nonlinear voter modeと呼ばれる確率的に時間発展する2次元格子モデル[3]において計測した結果が下図である。横軸が時間、縦軸の$I_C^{(k)}$がk次の連結情報量（つまり変数のk次相関の量）を表す。I_mはmulti informationと呼ばれる連結情報量すべての和である。このモデルは、系のパラメータに依存するいくつかの大域的なパターンを持つ。ここではクラスターと準周期的と呼ばれる系を紹介する。クラスターの場合においては、2次の情報のみであるが、準周期的な場合においては、2次に加え4次の相関が加わっている。この4次の相関こそが、クラスターのパターンと準周期的なパターンの違いを特徴付けていると考えることができる。



クラスターにおける連結情報量



準周期的な状態における連結情報量

また、2次の相関を考えると、系の中には2変数のペアは無数に存在している。しかし、序で述べたように、この連結情報量が与える情報は、系の中に2体の相関があるということのみである。各変数間の持つ局所的な2体の相関の値と、系全体の大域的な2体の相関の値がどのように関連しているかを議論する必要がある。しかし、この局所的な相関の和は大域的な相関と一般には一致しないことが分かる。例えば、全変数が3体(x_1, x_2, x_3)の場合には、$(x_1, x_2), (x_2, x_3), (x_1, x_3)$の3つの2体のペアが存在するが、これらの連結情報量の和は3体での連結情報量の和に一致しない。当日には、大域的な連結情報量が局所的な相関に分解できるケースと、その一般的な法則について説明する。

[1] E.Schneidman, S.Stoll, M.J.Berry II and W.Bialek, *Phys.Rev.Lett.* **91**, 238701 (2003)

[2] E.Schneidman, M.J.Berry II, R.Segev and W.Bialek, *nature* **440**, 1007 (2006)

[3] J.Molofsky, R.Durrett, J.Dushoff, D.Griffeath and S.Levin,*Theor.Popul.Biol.* **55**, 270 (1999)