エントロピーを頑張るドキュメント ver.0.9.1

情報量やエントロピーといった概念は、少し進んだ統計学を勉強し始めると途端に顔だしてくる。ところが、突然情報量を $I(x) = -p(x)\log x$ と定義すると言われても、なんでそれが「情報」と言えるのか、そもそも情報量って何を測りたいのか、さっぱりわからないと感じた人も多いと思う。当のこのドキュメントの作者は結構長いこと、情報量って概念がとらえどころがなく感じられて、苦労した。とはいえ、情報理論を一から勉強するのは億劫だし、ネットで探してもいまいち理解できなかったり、情報が散らばっていてなかなか飲み込めない、というのがありがちな状況だろう。そこでこのドキュメントは、情報理論の基礎的な内容を、あまりこの手のものが得意ではない噛み砕いて説明することを目的とした。想定する読者は、「一応一通り統計学は習ったし、普段から研究で使ってはいるけど、情報理論とか言われてもよくわからないです」みたいな平均的な心理学者を考えている。いくつかの内容や説明は Stone (2018) を参考にしてある。とりあえず目標としては、「情報量なにそれおいしいの？」というところから始めて、時系列データの因果関係の推定に用いられる移動エントロピーを理解するところまで頑張りたい。

本題に入る前に表記について先に書いておくと、統計学の教科書は確率変数を大文字 X、その実現値を x と書いたりするが、このドキュメントではそのような厳密さはすっ飛ばして、すべて小文字で書く。確率と確率密度もあまりこだわらずに書いてしまう。前提とする数学レベルは $\sum$ 記号がわかれば一応読めると思う。一節だけ $\int$ が出るが、複雑な積分計算なんかはしない。

ただ、ベイズの定理と期待値の定義だけは知っているものとしている。使う等式だけ提示しておくと、x と y という確率変数があるとき、以下が成り立つ。

$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

$$p(x, y) = p(x \mid y)p(y)$$

当然、y についても成り立つので、

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

$$p(x,y) = p(y|x)p(x)$$

が成り立つ。これらについて知らなかったら、ググればわかりやすい資料が見つかるはずだ。

期待値の定義も簡単に述べておこう。x の確率分布を $p(x)$ として、関数 $f(x)$ の期待値を $p(x)$ で取るというのは

$$\mathbb{E}[f(x)] = \sum_x p(x)f(x)$$

と定義される。$\mathbb{E}[]$ は「期待値をとる」という意味でよく使われるが、このドキュメントではこれ以降出てこない。簡単な例として歪みのないサイコロを投げたときに得られる値の期待値を求めてみよう。この場合、サイコロは歪んでないので、どの目が出る確率も等しい。よって x をサイコロの目とすれば

$$p(x) = \frac{1}{6}$$

また、サイコロの出た目そのもの値の期待値を取りたいので、$f(x) = x$ だ。期待値の定義通りに計算すると

$$\mathbb{E}[f(x)] = \sum_x \frac{1}{6}x$$

$$= \frac{1}{6}(1+2+3+4+5+6) = 3.5$$

と、期待値を求めることができた。

1. 情報量を定義する

情報とはそもそもなんだろうか？「情報」という概念は、Shannon (1948) 以前は曖昧な言葉であった。Shannonの論文で初めて明瞭に定義されたらしい。実際Shannonによると

“A basic idea in information theory is that information can be treated very much like a physical quantity, such as mass or energy.” Caude Shannon, 1985.
とのこと。

どんな定義なんだろうか？まずは言葉を使って議論しよう。Shannonの情報量の定義は「確率が等しい2つの選択肢から1つを選び取れたこと」である。また、この情報を1単位として、「ビット (bit)」という単位が定義される。

定義したはいいが、今のままでなにを言っているのか判然としないだろう。これはどういう意味か？例えば、裏表が出る確率が等しいコインを投げたとしよう。投げる前、あるいは出た面を観測する前は、裏か表かわからない。そのとき、どっちの面が出たかを知ることができたとき、それは情報を得たと言えるだろう。1ビットという情報量は、それが定義だ。では、コインではなくて、歪みのないサイコロを投げたとき、どの目が出たかを知るというのは、どれくらい情報を得たことになるのか？はたまた、サイコロが歪んでいたときは？いずれも確率分布する確率変数から、実際に得られた実現値を知ったときにどの程度情報を得たと言えるのか、ということについて考えている。それらを共通のモノサシで考えるための数学的な道具が情報量というわけだ。

ついでに言えば、情報量は「驚きの量 (サプライズ)」でもある。例えば今日、朝起きたら太陽が東から昇ってきた。これは驚きがあるだろうか？あるいは情報があるだろうか？おそらく、普通に生活している人にとってはほとんど驚きはないだろう。一方、「今日、太陽が西からのぼってきたんだ！」ということを知るのは、とても情報があるし、驚くべきことだろう (ただし、そんなことを言う人がいたら、おそらく寝ぼけていることを疑ったほうがいいだろう)。つまり、事象の情報量が高いというのは、おおよそ起こり得なさそうなことが起きたということだ。それは言い換えれば、サプライズのある出来事である[1]。

さて、滔々と言葉で語ってきたが、ここで数式で定義しよう。情報量は次のように定義される。

---

[1] 1つ注意しておくと、驚きといっても何も主観的にどう感じたか、あるいはどう感じるべきか、ということを言っているわけではない。情報理論で定義される「サプライズ」がそういう定義であるというだけだ。

$$I(x) = -\log p(x) \cdots (1)$$

$p(x)$ は x の確率分布で、例えば、コインだったら裏が出る確率と表が出る確率である。ただし、対数の底は2とする。別に自然対数でも常用対数でも、対数の底は簡単に変換できるので底を何にとっても議論に支障はないのだが、底を2で取ると単位がビットになる。そこで以下では特に断りがない場合、底を2で考える。しかし、たまに思い出したように強調して $\log_2$ と書くこともあるが、あまり気にしなくてもよい。

こういう風に天下りに「こう定義される」と提示されると、途端に何もわかった気がしなくなる心理学者は多いと思う。私はこういう数学の説明スタイルに慣れるのには時間がかかったし、なんなら今も苦労している。初見の人は (1) の情報量の定義を見ると「なんでlogなんて取る必要があるんじゃ？」と思ったに違いない。そこで、なぜこれが先ほど上で言葉で定義した情報量[2]になるかを確認しよう。コインの裏表が等しい確率のとき (つまり、両方0.5のとき)、コインを投げて出てきた面が裏でも表でも、得られる情報量は以下のようになる。

$$I = -\log \frac{1}{2} = 1$$

対数の底をあからさまに書いて、対数の分数が引き算になることを利用して確認すると、$-\log_2 1 + \log_2 2 = 0 + 1 = 1$となり、情報量が1ビットであることがわかる。つまり「確率が等しい2つの選択肢から1つを選び取れる」情報という風に定義した情報量そのものがちゃんと確認できた。歪みのないコインの裏表の結果を知ることには、1ビットの情報量があるのだ。

さらに、こういうのは極端な値を入れてみるとわかりやすい。先ほど、情報量はサプライズだ、と書いた。つまり、起きることが確定している事象には情報量 (サプライズ) はな

---

[2] 「確率が等しい2つの選択肢から1つを選び取れたこと」

いはずだ。起きることが確定しているとは？確率が1の事象だ。早速、確率 $p(x) = 1$ を (1) に代入して計算してみよう。

$$I = -\log 1 = 0$$

上の式が言っているのは、$p(x) = 1$の事象の情報量が0であるということだ。これは直観的な結果だろう。太陽が東から昇るのは、私たち人間にとってほとんど確実に起きる出来事だ。よって、太陽が今日も東から昇ってきたことを確認することは、情報量が0なのだ。

数式を用いて定義した情報量 (1) については、まだ説明に誤魔化しがある。コインが歪んでいて、表が出る確率 $p(x_{head}) = 0.8$、$p(x_{tail}) = 0.2$ と、表の方が出やすいとき、情報量はどうなるのか？このとき、当たり前っちゃ当たり前だが、表が出たときと裏が出たときで、情報量は変わってくる。実際にそれぞれ計算してみよう[3]。

$$I(x_{head}) = -\log p(x_{head}) = -\log 0.8 = 0.32$$

$$I(x_{tail}) = -\log p(x_{tail}) = -\log 0.2 = 2.32$$

結果の値を見ての通り、裏表で情報量は異なる。珍しい事象の方が、情報量が大きいのだ。試しに、いろんな歪み方をしたコインを想像してみて、$p(x_{head})$ を 0.01から0.99まで、0.01刻みでいろんな値を取ったときの情報量を見てみよう。それを図にしたものが図1だ。$p(x_{head})$ が低いほど、情報量は大きく、逆に表が出る確率が高いときほど情報量が下がっていくのが見て取れるだろう。例えば、$p(x_{head}) = 0.8$ のとき $I(x_{head}) = 0.32$ であるが、この値はどういう意味だろう？情報量は単位がビットで、確率0.5の2つの事象から1つの選択ができたことを示すのだった。$I(x_{head}) = 0.32$は、そのような選択の0.32回分の情報が、このコインを投げて表が出たときに得られたということを示している。また、裏が出たときは $I(x_{tail}) = 2.32$ であったわけだが、これは、裏表平等に出るコインを2回投げたときの結果を知る以上の情報が得られているということを示している。

---

[3] 一番右の等号は小数点第3位以下を切り捨てていて、本来 ≃ とでも書くべきところだが、面倒なので = で結んでしまっている (以降の例でもそうする)。

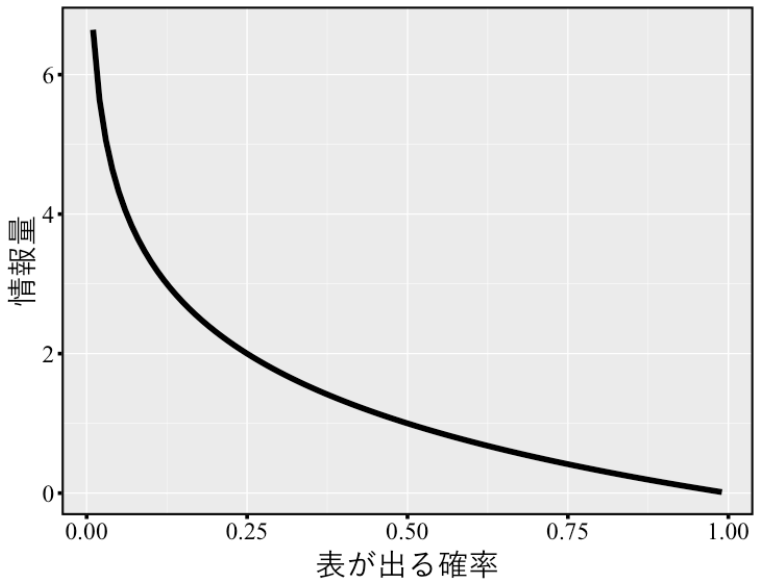

図1

2. エントロピーを定義する

上の例では、個別の事象がなんらかの確率で起きるとして、それが起きたときにその事象にどれくらい情報があると言えるということを考えてきた。そのことを改めて書いてみると、事象 x がいろんな値を取るとして、そのそれぞれの x についてなんらかの確率が $p(x)$ が振られるとしていた (例えば、コインの場合は、x は表か裏か、どっちかだ)。その x は、確率が高い事象から低い事象まで様々かもしれない。すなわち得られる情報量は大きいかもしれないし、小さいかもしれない。では、起きた事象を確認することに平均的にどの程度の情報量が期待できるのだろうか？そういう量としてエントロピーを定義しよう。エントロピーは、情報量の期待値である。

$$H(x) = -\sum_i p(x_i)\log p(x_i) \cdots (2)$$

再び、コインの例で考えてみよう。歪みがなくて、裏表が出る確率が等しいコインの場合、$p(x_{head}) = p(x_{tail}) = 0.5$ であるため、エントロピーは次のように計算できる。

$$H(x) = -p(x_{head})\log p(x_{head}) - p(x_{tail})\log p(x_{tail})$$

$$= -0.5 \log 0.5 - 0.5 \log 0.5$$

$$= 0.5 + 0.5 = 1$$

これはある意味当たり前の結果だ。このコインは、裏表どっちが出ても情報量は1ビットであるため、当然その期待値であるエントロピーも1ビットになる (単位が変わらずビットであることに注意しよう)。次に、$p(x_{head}) = 0.8$、$p(x_{tail}) = 0.2$ のとき、情報量はどうなるのかを考えてみよう[4]。

$$H(x) = -p(x_{head})\log p(x_{head}) - p(x_{tail})\log p(x_{tail})$$

$$= -0.8 \log 0.8 - 0.2 \log 0.2$$

$$= 0.72$$

この場合、エントロピーは歪みのないコインより小さいようだ。つまり、歪んでいて表が出やすいコインを投げて得ることが期待できる情報は、歪みのないコインを投げるときより小さい。このような性質を利用すれば、エントロピーを不確実さ (uncertainty) の程度として利用できる。どういうことかというと、エントロピーは、確率分布が平坦で、どんな値が得られるかがわかりづらい分布をほど高くなるのだ。コインの場合で、再び

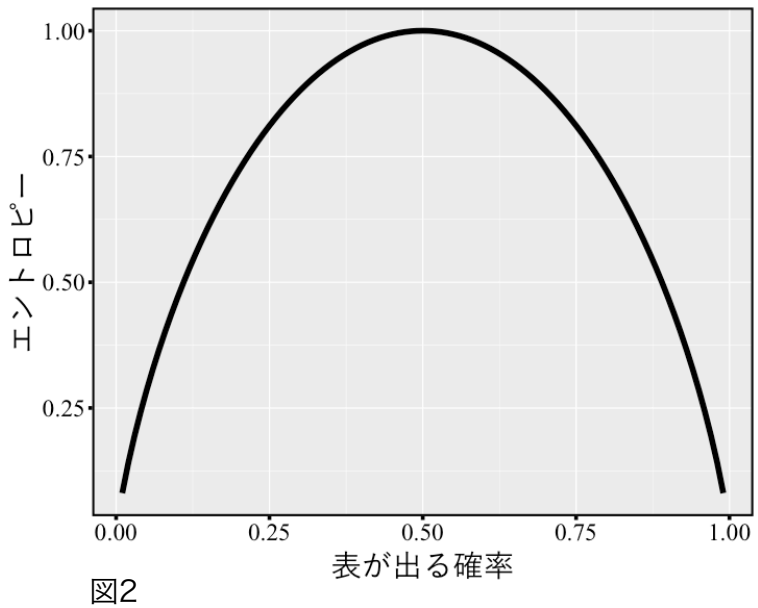

図2

[4] プログラムでも関数電卓でも対数表を見て手計算でもなんでもいいので、実際に計算してみるとよいだろう。

$p(x_{head})$ を 0.01から0.99まで、0.01刻みでいろんな値を取ったときの情報量を見てみよう。その結果を図2に示した。表が出る確率 $p(x_{head})$ が0.5のとき、エントロピーが最大値を取っているのがわかると思う。$p(x_{head}) = 0.5$ というのは、最も得られる結果が不確実なコインだ。一方、$p(x_{head})$ が0や1に近いときというのは、コイン投げの結果が容易に予想がつく。そのような状況は、不確実性が低いと言えるだろう。エントロピーがそのような直観を反映した値になっているのが見て取れると思う。

ここまでで、情報量とエントロピーについて、得られる情報の大きさ、不確かさ、不確実性、サプライズと、色々な言い方をしてきた。しかし、(1) と (2) 式を見れば、いずれも同じことを別の言い方をしているのにすぎないことがわかったと思う。

科学研究をしていると、このエントロピーという量は様々な場面で出てくるし、実際使うことができる。例えば、行動指標の分布のエントロピーを従属変数にして分析したり、統計学で「最大エントロピー」を持つ分布が重要な役割を持っていたりする。また、後ほど出てくる相互情報量や移動エントロピーといった量は神経科学や動物行動学でもよく使われる。さらに、Karl Fristonの自由エネルギー原理は、生きているエージェントがエントロピーを環境に排出することこそが、生命が環境内で持続するための条件である、なんて主張をする (Friston, 2010)。熱力学、統計力学でも、利用不可能なエネルギーの量としてエントロピーが出てきて、実はここで定義したエントロピーと同じ形をしている。しかしこれは心理学者にはあまり馴染みがないし、関係する機会は少ないだろう[5]。と、この先には色々と興味深い世界が広がっているわけだが、そういう楽しげな話は一旦置いておいて、連続分布での情報量やエントロピーについて見ていこう。

3. エントロピーを連続量に拡張する

これまでのところ、扱ってきたのはコインという、裏か表しか事象がない例だ。これは確率分布としては $x = x_{head}$ か $x = x_{tail}$ しかとらない離散分布である[6]。もちろんサイコロの

---

[5] 無関係とは言わない！むしろホットなトピックとして、このドキュメントの作者はとても興味を持っている。例えば Collell and Fauquet (2015)、del Prado Martín (2011)、Zeron et al. (2019) などなど。

[6] ようはベルヌーイ分布だ。

ように事象のパターンが増えても同じ計算で情報量やエントロピーは計算できるし、動物の行動の選択回数でも、行動の生起回数でも、なんでもいいんだが、とにかく離散分布で議論をしてきた。エントロピーを連続変数まで拡張しよう。

と言っても、定義は見た目上ほとんど変わらない。連続確率分布の場合、確率の代わりに確率密度関数 $p(x)$ で事象の起こりやすさを関数で表し、確率はその積分で表示されるのは、どこかで習ったことはあるだろう。x をなんらかの連続変数としたら、エントロピーは次のように定義される。

$$H(x) = -\int p(x)\log p(x)dx \cdots (3)$$

離散変数のエントロピー (2) 式との違いは、総和記号 $\sum$ が積分記号 $\int$ に置き換わっているだけであることがわかる。また、確率分布は離散分布の例と同じように $p(x)$ と書いてしまっているが、ここでは連続分布を考えている。このドキュメントでは、特に区別して書いたりはせず、分布が離散か連続かはその都度空気を読んでもらうことにするが、特にそこで悩む部分はないと思われる。

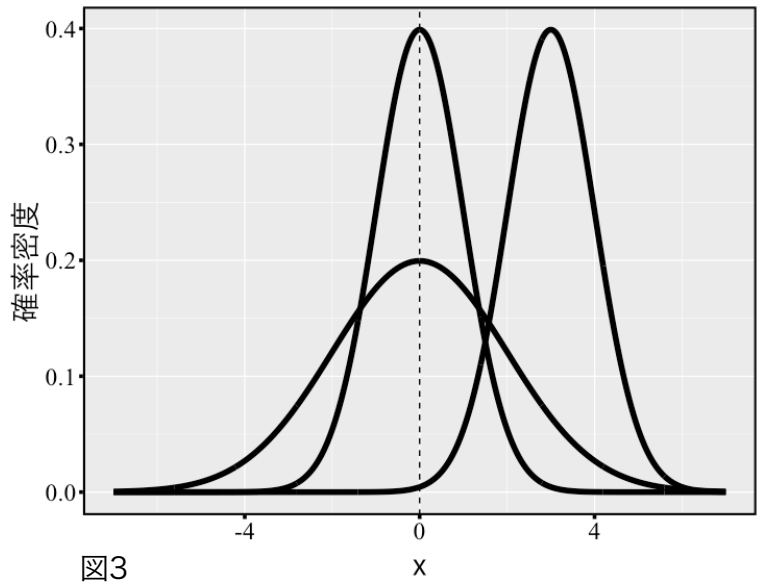

図3

```
# 正規分布のエントロピーを計算するRコード
fx = function(x, mu, sig){
  res = dnorm(x, mu, sig)  * log(dnorm(x, mu, sig))
  res[!is.finite(res)] = 0
  # 正規分布は本来-∞から+∞までの値をとるが、
  # PC上ではとても小さい値は0になってしまう
  # log 0 = -∞ でエラーが起きるので、その部分を0にしている
  return(-res)
}
# 計算する
integrate(function(x) fx(x, 0, 1), -Inf, Inf)
integrate(function(x) fx(x, 0, 2), -Inf, Inf)
integrate(function(x) fx(x, 3, 1), -Inf, Inf)
```

ここでは例として正規分布のエントロピーを測ってみよう。図3に3つの正規分布の図を示した。正規分布は平均と分散2つのパラメータを持つ分布である。そこで、Normal(μ, σ) と表記することにしよう。3つの分布が表示されているが、それぞれ、Normal(0,1)、Normal(0,2)、Normal(3,1)である。後者2つは、標準正規分布から分散を広げたものと、平均をシフトしたものだ。このエントロピーを (3) 式に従ってコンピュータ上で計算する

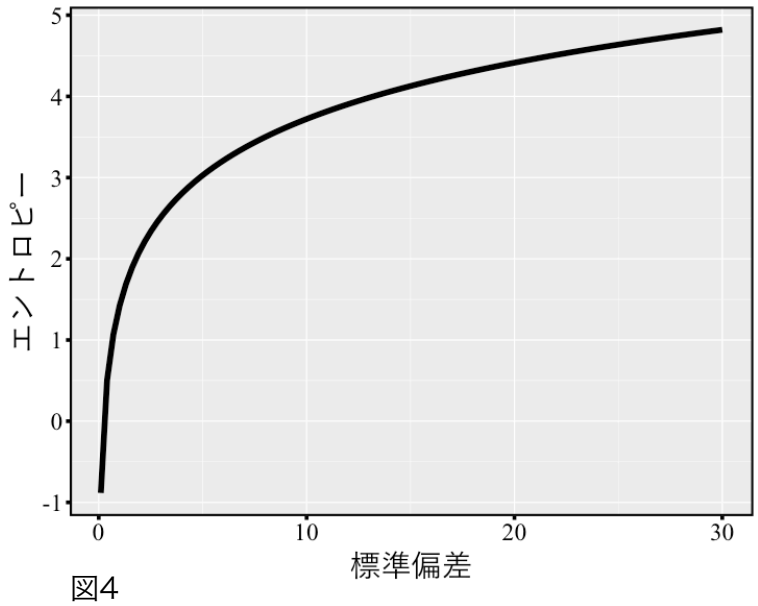

図4

と、Normal(0,1) は1.42、Normal(0,2) は2.11、Normal(3,1) は1.42程度になる。Rコードをつけておいたので、興味があれば実行するとよいだろう。値を見れば、分散が広がるとエントロピーが大きくなることがわかる。一方、平均値をずらしてもエントロピーは変わらないことがわかる。これは直観的な結果だ。平均値がずれても分散が等しければ、分布としての不確かさは同程度だ。つまり、どんな値が得られそうか、という予測は同程度にできてしまう。なので、エントロピーは等しくなる。一方、分散が大きいときというのは、どんな値が出るか予想がつかない状況である。そんな不確かな状況では、値を観測するのは情報として価値が高いはずだ。エントロピーは得られる情報量の平均値なので、実際そのように挙動していることがわかる。このことをもう少し確かめてみよう。平均値は固定した下で、正規分布の標準偏差を細かく刻んで増やしていったときにエントロピーがどういう挙動になるのかを確かめてみる。その結果が図4だ。見ての通り、分散を大きくしていくほど、エントロピーが高くなっているのが見て取れるだろう。

最大エントロピー分布

閑話休題。何か確率変数と思しき変数を得たが、それについて特に情報がないとき、どんな分布で考えるのがよいだろうか？今のところその確率変数についてはほとんど何も知らないのだから、エントロピーが最大になる分布を選ぶのがよいだろう、というのが1つの方針になるだろう。これを最大エントロピー原理という。そのような分布を定める最大エントロピー分布 (maximum entropy distribution) とは、文字通りエントロピーが理論的に最大になる分布である。統計学では、最大エントロピー分布が様々なところで重要になってくるので、どんなものかくらいは知っておくと後々便利だろう。確率変数について何も知らないと言っても、「この値はどう考えても負の値は取らないだろう」と言ったことはわかることが多いだろう。そのような条件を「制約」と呼ぶ。どの分布が最大エントロピー分布になるのかは、着目している変数にかけられる制約によって異なる。ここでは以下の3つをここでは紹介しよう。

・正規分布は：変数の分散が固定されているときの最大エントロピー分布
・指数分布：非負で平均が固定されているときの最大エントロピー分布
・一様分布：上限と下限だけが決まっているときの最大エントロピー分布

これらの分布は、上記の制約の下では最も大きいエントロピーを持つ分布であることが証明されている (証明は大変なのでしない)。

4. 様々なエントロピーや情報量に慣れ親しもう

4.1. 結合エントロピー (joint entropy)

ここまでのところ、常に変数は x 1つだけであった。だが、変数が複数あってもエントロピーは定義できる。x と y という2つの確率変数があって、その同時分布 $p(x,y)$ があるとする。$p(x,y)$ は、x と y の組について、何か1つの確率が割り振られる確率分布だ。このとき、x と y を同時に知ることができたとして、どの程度の情報を平均的に得られるか？これが結合エントロピー $H(x,y)$ である。定義は以下の通りだ。

$$H(x,y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y) \cdots (4)$$

変数が1つだったのが2つになった分、総和記号が2重になっているのと、$p(x)$ と今まで書いていた部分が $p(x,y)$ になっているだけなので、そこまで変な形には見えないだろう。一応、1つだけ試しに計算してみてどんなものかだけは見ておこう。

状況設定としてはこういう場面を想像してみよう。カレーの具材として、あなたは肉には豚肉か牛肉を使う。野菜には、ニンジンかジャガイモを使う。つまり、$x = \{$豚肉, 牛肉$\}$、$y = \{$ニンジン, ジャガイモ$\}$ というわけだ。ただ、あなたは妙なこだわりを持っていて、具材はそれぞれどっちかしか入れないし、豚肉のときはジャガイモと合わせることが多くて、牛肉はニンジンと合わせることが多いとしよう。で、その同時確率が表1だとする。この結合エントロピーを計算しよう。ここからは豚肉、牛肉をそれぞれ x_1、x_2、ニン

表1	豚肉 (x_1)	牛肉 (x_2)
ニンジン (y_1)	0.18	0.32
ジャガイモ (y_2)	0.42	0.08

ジンとジャガイモは y_1、y_2 と書く。例えば豚肉とジャガイモを合わせる確率は $p(x_1, y_2) = 0.42$ だ。あとは、結合エントロピーを (4) 式に従って計算する。

$$
\begin{aligned}
H(x, y) &= -\sum_x \sum_y p(x, y) \log p(x, y) \\
&= -p(x_1, y_1) \log p(x_1, y_1) \quad \text{愚直に} x_1、x_2、y_1、y_2 \text{ を入れている} \\
&\quad -p(x_1, y_2) \log p(x_1, y_2) \\
&\quad -p(x_2, y_1) \log p(x_2, y_1) \\
&\quad -p(x_2, y_2) \log p(x_2, y_2) \\
&= 0.45 + 0.53 + 0.53 + 0.29 = 1.79
\end{aligned}
$$

と、結合エントロピーが1.79ビットであることが計算できた。1例でも計算した経験があれば、慣れ親しみ方は段違いだろう。自分で適当に値を入れて計算すればすぐわかるが、結合エントロピーもまた、分布が平坦 (この場合だと、どの組み合わせでも $p(x, y) = 0.25$ に近い分布) になるほどエントロピーは高くなる。

4.2. 条件付きエントロピー (conditional entropy)

再び、2つの確率変数 x、y があったとき、条件付きエントロピーは $H(y|x)$ と書く。縦棒は「ギブン」と読むので「エイチ x ギブン y」と読めばよい。これはいわば、x について知ることで、どの程度 y がわかるかということを意味する量である。もちろん、x、y の間には関連があるかもしれないし、ないかもしれない。関連があれば、x について知ることができたら、y についての不確かさは下がるだろう (好きな異性の好みを知れば、自分にチャンスがあるかどうかの不確かさは多少なりとも下がるはずだ)。両者に関係なければ、x について知ったところで、y についてはちっとも知ることはできない (このドキュメントを読んでエントロピーについてわかっても、あなたに明日恋人ができるかどうかかは

全くわからない。両者に関連がないからだ)。条件付きエントロピー $H(y|x)$[7] はそんな量でなければならない。その定義は次の通りだ。

$$H(x|y) = -\sum_{x}\sum_{y} p(x,y)\log\frac{p(x,y)}{p(y)} \quad \cdots (5)$$

また、$p(x|y) = \frac{p(x,y)}{p(y)}$ より、$p(x,y) = p(x|y)p(y)$ であるので、(5) 式は以下のようにも変形できる。

$$H(x|y) = -\sum_{x}\sum_{y} p(x|y)p(y)\log p(x|y) \quad \cdots (6)$$

なぜ、こう定義するのか。それは (5) 式を変形するとわかりやすいかもしれない。対数の割り算は引き算になることを使って変形してみよう。

$$H(x|y) = -\sum_{x}\sum_{y} p(x,y)\log p(x,y) + \sum_{x}\sum_{y} p(x,y)\log p(y)$$

ここで、右辺第二項に注目しよう。$\log p(x)$ は x には依存せず、$\sum_{y} p(x,y) = p(x)$ と周辺化ができる。よって

$$H(x|y) = H(x,y) - H(x) \quad \cdots (5)'$$

が成り立つ。条件付きエントロピーの定義式は、x と y を同時に知ったときの情報量から、x 単体を知ったときの情報量を引いたものであるということだ。つまり、y を知った

---

[7] 当然 $H(y|x)$ も考えられる

上で、それでもなお残る x の不確かさが $H(x|y)$ である。さらに、(5)' から当然次の関係も成り立つ。

$$H(x,y) = H(x) + H(x|y) \text{・・・(5)''}$$

これで、条件付きエントロピーと、ここまでのところで、その意味を押さえた。再び、表1のカレーの例で考えてみよう。ここではあえていろんな表記になれる意味合いも込めて、$H(y|x)$ を求めてみよう。お肉の方の食材 (x) を知ったとき、野菜 (y) についてはどの程度不確かさが残っているのか。まず条件付き確率を x_1、y_1 を例に計算すると以下のようになる。

$$p(y_1|x_1) = \frac{p(x_1,y_1)}{p(x_1)} = \frac{0.18}{0.18+0.48} = 0.3$$

他の条件付き確率も同様に計算すればよい。すると、(6) 式を使えば条件付きエントロピーは次のようにな値になる。

$$H(y|x) = -\sum_x \sum_y p(y|x)p(x)\log p(y|x)$$

$$= -p(y_1|x_1)\log p(y_1|x_1) \quad \text{再び愚直に}x_1\text{、}x_2\text{、}y_1\text{、}y_2\text{ を入れている}$$

$$-p(y_2|x_1)\log p(y_2|x_1) \quad \text{根気よく1つ1つ計算して足し上げればいい}$$

$$-p(y_1|x_2)\log p(y_1|x_2)$$

$$-p(y_2|x_2)\log p(y_2|x_2)$$

$$= 1.60$$

4.3. 各エントロピー同士の関係を考える

ここまでのところ、単一変数のエントロピー $H(x)$、複数の変数の同時分布の結合エントロピー $H(x,y)$、ある変数で条件づけたときの条件付きエントロピー $H(x|y)$ というのを定義し、試しに計算してみた。エントロピー $H(x)$ というのは、事象を観測することでどの程度平均的に情報を得られるか、言い換えれば出来事が平均的にどの程度予測できないかといういわば不確かさを表す尺度であった。結合エントロピー $H(x,y)$ はそれが複数の変数になっただけなので、直観的にもわかるやすいだろう。問題は条件付きエントロピー $H(x|y)$、こいつである。突然「変数なんとかで条件づけたときの」なんて言われてもなんのこっちゃと思う人もいるだろう。そこで、各エントロピーの関係を示すことで、その引っかかりを解消しようと思う。

早速これまで出てきたエントロピーの関係を示そう。各エントロピーには次のような関係がある。まず、x と y が独立のとき、結合エントロピーと1変数のエントロピーには次の関係がある。

$$H(x,y) = H(x) + H(y) \text{・・・(7)}$$

この関係は、x と y が独立のとき同時確率分布が $p(x,y) = p(x)p(y)$ と、各変数の分布の積となることから確かめられる。実際にやってみよう。

$$H(x,y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

$$= -\sum_x \sum_y p(x)p(y) \log p(x)p(y)$$

$$= -\sum_x \sum_y p(x)p(y)(\log p(x) + \log p(y)) \text{ (対数の積は和になる規則)}$$

$$= -\sum_x \sum_y p(x)p(y) \log p(x) - \sum_x \sum_y p(x)p(y) \log p(y)$$

ここで、$p(x)$ は y には依存しないし、$p(y)$ は x に依存しない。なので、x に関係する総和と y に関係する総和の部分を別個に計算してもよい。また、確率分布の定義から、

$\sum_x p(x) = 1$ かつ $\sum_y p(y) = 1$ である。これらの性質を使う。

$$= -\sum_x p(x)\log p(x)\sum_y p(y) - \sum_y p(y)\log p(y)\sum_x p(x)$$

$$= H(x) + H(y)$$

と、めでたく (7) の関係が導けた。$H(x,y)$ とは、x と y を同時に知ったときに得られる情報量であるが、x と y が互いに独立のとき、その情報量は個別に x と y について知った時の情報量 $H(x)$ と $H(y)$ の和であると、(7) 式は言っている。これはある意味当たり前の関係だろう。

また、x と y が独立のとき次のような関係がある。

$H(x) = H(x\,|\,y)$・・・(8)

x と y が独立であるというのは、$p(x) = p(x\,|\,y)$ という状況だ。つまり、y がわかっても x の確率分布は一切変わらず、両者には関係がないということだ。このとき、y がわかっても x については何も情報は得られない。つまりエントロピーは変わらない。(9) 式はそのようなことを言っている。実際そうなるか確かめてみよう。(5) 式の定義を使って

$$H(x\,|\,y) = -\sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(y)}$$

$$= -\sum_x\sum_y p(x,y)\log p(x\,|\,y)$$

ここで、x と y が独立であるため、$p(x,y) = p(x)p(y)$、$p(x\,|\,y) = p(x)$ である。

$$= -\sum_x \sum_y p(x)p(y)\log p(x)$$

$$= -\sum_x p(x)\log p(x) \sum_y p(y) = -\sum_x p(x)\log p(x) = H(x)$$

最後の変形には、$\sum_y p(y) = 1$ を使った。これで (8) 式を示すことができた。x と y が独立のとき、y を知っても x についての不確かさは一向に減らないのである。

逆に x と y の間になんらかの相関関係があったとしたらどうだろう？y について知ることができれば、x の不確かさは多少なりとも下がるだろう。つまり、以下の不等式が成り立つはずだ。

$$H(x) > H(x|y)$$

よって独立な場合と合わせて、条件付きエントロピーは次の性質があると言える。

$$H(x) \geq H(x|y) \cdots (9)$$

等号が成り立つは x と y が独立なときで、そうでないときは不等号が成り立つ。

4.4. 相互情報量 (mutual information)

前節では条件付きエントロピーを定義し、その意味は x と y という確率変数において、y を知ったときの x についての平均的な情報量 (不確かさ) であるということを述べた。では、y について知っているときと知らないときで、どの程度不確かさが減るのだろうか？これには相互情報量という特別な名前がある。定義式を以下に示す。

$$I(x,y) = H(x) - H(x|y) \cdots (10)$$

情報量という名前だが、エントロピー同士の引き算であるため、「平均的にどの程度不確かさが減るか」という量であることには注意しておきたい。例えば、気になっている異性の好きなタイプ y について知るとしても “y_1 = 高身長” と “y_2 = 経済力”があるでは、”x = 自分にもチャンスがある” の不確かさがどの程度変わるかは異なるだろう。(10) 式で測るのは「いろいろな y がありうるけど、平均的には不確かさが下がるのだろうか」ということだ。「いろいろな y 」と言ったが、それはもちろん確率分布として着目している対象によって決まる (反応時間なら0以上の連続量だし、選択率なら0から1の間を取り、脳波なら負の値まで取るだろう)。ちなみに、証明はしないが相互情報量は $I(y,x) = H(y) - H(y\,|\,x)$ も成り立ち、$I(x,y) = I(y,x)$ となるので安心してほしい。つまり、どっちの変数で条件づけても相互情報量は同じ値になる。相互情報量は方向性を持たない量なのだ。一方、次に紹介する移動エントロピーは方向性を持つ量である。

また、相互情報量には別の表し方もある。式を変形させながら見ていこう。

$$I(x,y) = H(x) - H(x\,|\,y)$$

$$= -\sum_x p(x)\log p(x) + \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(y)}$$

ここで $p(x) = \sum_y p(x,y)$ を使えば、

$$= -\sum_x \sum_y p(x,y)\log p(x) + \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(y)}$$

$$= \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(y)p(y)}$$ (対数の引き算は割り算になるルールを使った)

と変形できる。書き直すと、

$$I(x,y) = \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(x)p(y)} \quad \cdots (10)'$$

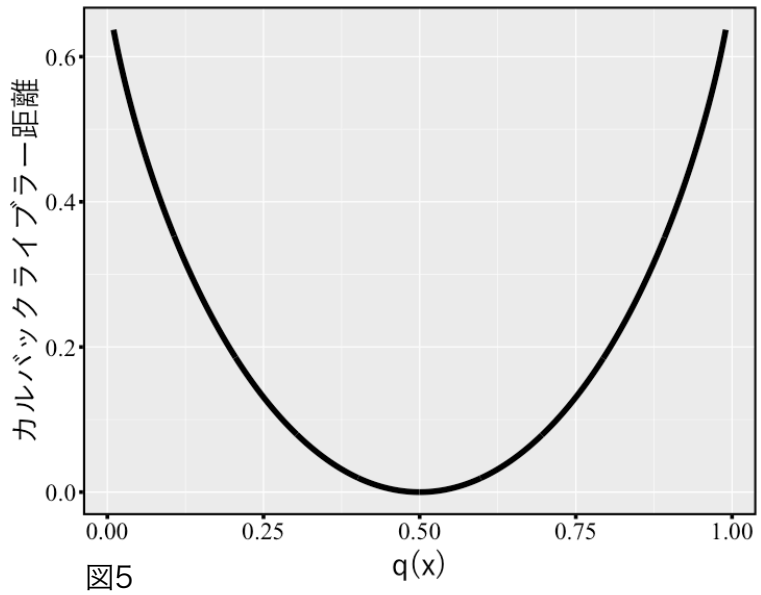

図5

資料によっては (10)' が相互情報量の定義として紹介されていることもある (むしろ、そっちの方が多いかも)。(10)' をさらに変形すると

$$I(x,y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_x \sum_y p(x,y) \log p(x,y) - \sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y)$$

$$= H(x) + H(y) - H(x,y)$$

という風に相互情報量を表すこともできる。

5. カルバック・ライブラー情報量 (Kullback–Leibler divergence)

カルバック・ライブラー情報量は、相対エントロピー (relative entropy)とも呼ばれる量で、2つの分布の類似性を測る指標として使われる。統計学でも、例えばモデル選択を行うときに用いる情報量基準の理論的に裏付けに使われる。なので、カルバック・ライブラー情報量も知っておいた方が色々捗るだろう。いきなり定義から提示しよう。$p(x)$ と $q(x)$ という2つの分布があるとして、カルバック・ライブラー情報量は次のように定義される。

$$KL(q(x)\|p(x)) = \sum_x q(x)\log\frac{q(x)}{p(x)} \cdots (11)$$

この値の性質について考えてみよう。$p(x)$ と $q(x)$ が完全に一致しているとき、つまり $p(x) = q(x)$ であれば、$\log\frac{q(x)}{p(x)}$ の部分が0になる。$\log 1 = 0$ であるためだ。逆に、$p(x) \neq q(x)$ のとき、カルバック・ライブラー情報量は0より大きい値をとる。もっと言えば、$p(x)$ と $q(x)$ の形が違えば違うほど、カルバック・ライブラー情報量は大きくなる。これが、カルバック・ライブラー情報量が「バックライブラー距離」とも呼ばれる理由だ[8]。

試しに最も簡単なカルバック・ライブラー距離を計算してみよう。コインが二枚あって、1枚は歪みがないので表が出る確率が $p(x_{head}) = 0.5$ であるとしよう。もう一枚は歪んでいて、$q(x_{head}) = 0.7$ であるとする[9]。当然 $p(x_{tail}) = 0.5$ かつ、$q(x_{tail}) = 0.3$ である。(11) 式の定義に従ってカルバック・ライブラー情報量を計算すると

$$\begin{aligned} KL(q(x)\|p(x)) &= \sum_x q(x)\log\frac{q(x)}{p(x)} \\ &= q(x_{head})\frac{q(x_{head})}{p(x_{head})} + q(x_{tail})\frac{q(x_{tail})}{p(x_{tail})} \\ &= 0.7\log\frac{0.7}{0.5} + 0.3\log\frac{0.3}{0.5} = 0.12 \end{aligned}$$

計算できたはいいが、これだけでは挙動がわかりづらいと思うので、$q(x_{head})$ の値を0.01から0.99まで細かく刻んで、カルバック・ライブラー距離がどう変わっていくのか見てみよう。$p(x_{head})$ は0.5に固定して、様々な　$q(x_{head})$ について $KL(q(x)\|p(x))$ を計算したの

---

[8] ただし、「距離」と言っても、一般には $\sum_x q(x)\log\frac{q(x)}{p(x)} \neq \sum_x p(x)\log\frac{p(x)}{q(x)}$ である。なので、カルバック・ライブラー情報量のことをカルバック・ライブラー擬距離と呼ぶこともある。

[9] ようは、ここでは2つのベルヌーイ分布のカルバック・ライブラー情報量を測ろうとしている。

が図5である。$p(x) = q(x)$ のときカルバック・ライブラー情報量が最小になっているのがわかると思う。

さらに、(10) で定義した相互情報量は、カルバック・ライブラー距離で表すこともできる。早速関係を見てみよう。まず、相互情報量が (10)' 式で表せることを思い出そう。

$$I(x,y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(y)p(y)}$$

ここで、$p(x|y) = \frac{p(x,y)}{p(y)}$ なので、$p(x,y) = p(x|y)p(y)$ となる。それを代入すると

$$I(x,y) = \sum_x \sum_y p(x|y)p(y) \log \frac{p(x|y)}{p(x)}$$

$$= \sum_y p(y) \sum_x p(x|y) \log \frac{p(x|y)}{p(x)} \quad (p(y) \text{ は } x \text{ に依存しないので外に出した})$$

$$= \sum_y p(y) KL(p(x|y) \| p(x)) \quad ((11) \text{ 式を見たら、こうなるのがわかるだろう})$$

上式は $p(y)$ で $p(x|y)$ と $p(x)$ の間のカルバック・ライブラー情報量の期待値を取っていることに相当する。これはどういうことだろうか？ ある y で条件づけたとして、条件付き分布 $p(x|y)$ が考えられる。その $p(x|y)$ と $p(x)$ はどの程度似ているだろうか？それがカルバック・ライブラー情報量 $KL(p(x|y) \| p(x))$ だ。しかし、y は確率分布 $p(y)$ に従って、いろんな値を取りうる。そこで、いろんな y がありうるが、平均的には $KL(p(x|y) \| p(x))$ がどんな値になるのか、期待値を計算する。それが相互情報量 $I(x,y)$ になるというわけだ。もともと相互情報量は $I(x,y) = H(x) - H(x|y)$ と定義するところから出発したが、こうやって別の見方もできることがわかると面白い (と思う)。

ここまでのところでいろんな値を定義してきた。一通りまとめて載せておこう。

情報量：$I(x) = -\log p(x)$

エントロピー：$H(x) = -\sum_x p(x)\log p(x)$

結合エントロピー：$H(x,y) = -\sum_x \sum_y p(x,y)\log p(x,y)$

条件付きエントロピー：$H(x\,|\,y) = -\sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(y)}$

$$= -\sum_x \sum_y p(x\,|\,y)p(y)\log p(x\,|\,y)$$

相互情報量：$I(x,y) = H(x) - H(x\,|\,y)$

$$= H(x) + H(y) - H(x,y)$$

$$= \sum_x \sum_y p(x,y)\log \frac{p(x,y)}{p(x)p(y)}$$

$$= \sum_y p(y) KL(p(x\,|\,y)\|p(x))$$

カルバック・ライブラー情報量：$KL(q(x)\|p(x)) = \sum_x q(x)\log \frac{q(x)}{p(x)}$

6. 移動エントロピー (transfer entropy)

最後に発展的な話として、移動エントロピーを紹介しよう。移動エントロピーとは、複数の時系列データから互いの因果関係を推定するために用いる情報量である (Schreiber, 2000)。具体的な応用例としては、神経活動の領域間の影響関係 (Gao et al., 2020)、動物の移動のフォロー・フォロワー関係 (Orange, & Abaid, 2015)、ソーシャルメディアでの情報の流出入の推定 (Ver Steeg, & Galstyan, 2012) など幅広く用いられている。

時系列データとは、文字通り時間で並んだデータのことで、ここでは等間隔に並んだ $x_t = \{x_1, x_2 \dots x_T\}$ となっているデータを考えよう。時間的に推移しているデータというのは、前の時点の影響を受けると考えるのが自然だろう。例えば、昨日の気温が25度だったら、通常の気象条件なら今日も似たような値を取り、突然-30度になったりはしないはずだ。つまり、時系列データというのは以前のデータの履歴に依存するはずである。t 時点での x の値 x_t は x_{t-1} に依存すると考えれば、その確率分布は $p(x_t|x_{t-1})$ と表せる[10]。当然、条件付きエントロピー $H(x_t|x_{t-1})$ も考えられる。もちろん、$H(x_t|x_{t-1})$ は t_{t-1} 時点でのx_{t-1} が観測された上での x_t の不確かさだ。

ここで、もう一つ時系列データ $y_t = \{y_1, y_2 \dots y_T\}$ があると考えよう。この y が x に何か因果的な影響を与えていたら、 y_{t-1} が x_t に影響を与えていると仮定できる。もし、その仮定が正しければ、y_{t-1} を知ることができれば x_t の不確かさを下げることができるはずだ。移動エントロピーは「過去の y を知ったときに現在の x の不確かさをどれだけ下げられるか」という量として定義される。

$$T_{y \to x} = H(x_t|x_{t-1}) - H(x_t|x_{t-1}, y_{t-1}) \cdots (12)$$

左辺を見ると、2つの条件付きエントロピーの差で定義されているのがわかる。$H(x_t|x_{t-1})$ は、x の過去だけから現在の x を予測したときの不確かさである。$H(x_t|x_{t-1}, y_{t-1})$ はそれに加えて y が付け加わっている。y があるときとないときの不確かさの程度を引き算しているので、「y を知ることでどれくらい x についての不確かさを下げられたか」という量になっているわけだ。例のごとく、こういうのは極端な値を入れるとわかりやすい。x と y の間で全く因果関係がない場合どうなるか？それはつまり、この場合 x_t と y_{t-1} が独立

---

[10] 「$t-1$ 時点より前の影響は考えないのだろうか？」と疑問思ったかもしれない。当然 $p(x_t|x_{t-1}, x_{t-2}, \dots x_{t-k})$ とさらに過去に遡った影響を考えることもできる。が、ここでは最も簡単な移動エントロピーを紹介すべく、$t-1$ 時点の影響だけを考えている。少し後で、それより前も考慮する移動エントロピーを考える。

であるということなので、 (8) 式より$H(x_t|x_{t-1}) = H(x_t|x_{t-1}, y_{t-1})$ となり、$T_{y\to x} = 0$ である。両者に因果関係がないとき、移動エントロピーが0になることがわかる[11]。

当然、逆に x から y への逆側の影響も考えることができて、

$$T_{x\to y} = H(y_t|y_{t-1}) - H(y_t|x_{t-1}, y_{t-1})$$・・・(13)

と定義できる。さらに、これはよく見たら、自分自身の過去のデータで条件づけているところ以外は相互情報量の定義 (10) と全く同じであることがすぐにわかる。移動エントロピーは、相互情報量の拡張版なのだ。

しかし、相互情報量が向きを持たない量であった一方、移動エントロピーは方向性を持つ量だ。つまり、通常 $T_{x\to y} \neq T_{y\to x}$ である。例えば、x から y への影響よりも、y から x への影響が大きければ、移動エントロピーは $T_{x\to y} < T_{y\to x}$ となる。

ここまで扱ったのは、一時点前の影響を扱った移動エントロピーだ。しかし、現実世界では常に一時点前とは限らない。例えば1000Hzで取った脳波データの一時点前というのは、1ms前だ。影響関係が常に1ms後に出ると考えるのはだいぶ無理がある。そこで $t-1$ 時点より前の影響を考慮するよう、移動エントロピーを改変する。

---

[11] ここまで読んで、こう思った人もいるかもしれない。「なんでこれで因果関係を定量化できたことになるのか？確かに y から x に因果的な影響があれば、移動エントロピーは上がるだろう。しかし、逆は真とは限らないはずだ。すなわち、移動エントロピーが0以上であることは x と y の間に因果関係があることを含意しない」と。これを書いている人の理解では、それはその通り。因果関係とは何か、統計的にどう定義し、推論するかはそれだけで一大問題だ。あくまで情報理論ではこう定義して議論を進めていくという話で、ここでは「因果関係とはこういうものである」ということを言っているわけではない。とはいえ、それだとあまりに歯切れが悪い。実践的には、実験や観察の統制をきちんととったり、ドメイン知識から「これは因果と考えていいはずだ」と判断したり、得られた因果関係から予測されるさらなる現象を検証することで知見の確度を高めたりと、実験科学者らしいプラグマティックな態度が必要になるだろう (なんかこれもこれで歯切れが悪いな、でもそうなので仕方ない)。

$\bar{x}^k = \{x_{t-1}, x_{t-2}, \ldots x_{t-k}\}$ を、t 時点から見て k 時点前までの x を集めたものとしよう。当然、$\bar{y}^k = \{y_{t-1}, y_{t-2}, \ldots y_{t-k}\}$ である。前述の要求を満たす移動エントロピーは次のように定義される。

$$T_{y \to x} = H(x_t | \bar{x}^k) - H(x_t | \bar{x}^k, \bar{y}^k) \cdots (14)$$

定義 (12) との違いは、単に $t-1$ より前のデータで条件づけているか否かである。何時点前までのデータを参照するかは k で表記してある。実際のデータ解析では、k は自身で非現実的にならない範囲で適当な値を入れる必要がある。

ちなみに、Rで移動エントロピーを実行するにはRTransferEntropyというパッケージがある。ただし、これは2変量までしか受け付けない。3変量以上の移動エントロピーを計算できるパッケージとしてはPythonのIDTxlやMatlabのTRENTOOLがあるようだ。例えば脳波のような多変量時系列の場合、各チャネルの活動が他のチャネルにどの程度相互に影響しているかを見たいわけなので、3変量以上の移動エントロピーを計算する必要がある。

最後に余談だが、時系列の因果関係となると「グレンジャー因果 (Granger causality)」という名前を聞いたことがあって「それと移動エントロピーはなにが違うの？」と思った人もいるかもしれない。グレンジャー因果については、このドキュメント作者も勉強中で詳しく知らないので詳細は述べないが、簡単に両者の関係を指摘しておく。グレンジャー因果は背景になんらかのモデルを仮定する[12]。一方、移動エントロピーは特定のモデルを仮定しないで時系列間の関係を測るモデルフリーな手法である。ただし、確率分布がガウス分布のとき移動エントロピーとグレンジャー因果は等価であることが示されているらしい (Barnett et al., 2009)。さらに、一定条件の下で両者の等価性がさらに一般化できるらしいが、このあたりについては、このドキュメントの筆者は全くわかっていないので、とにかくそういう話題があるということだけ挙げておく (Hlaváckova-Schindler, 2011)。

---

[12] ノンパラメトリックモデルでも構わないらしいんだが、通常はベクトル自己回帰などのパラメトリックモデルを使うことが多いと思われる。

Reference

Barnett, L., Barrett, A. B., & Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. Physical Review Letters, 103(23), 238701.

Collell, G., & Fauquet, J. (2015). Brain activity and cognition: a connection from thermodynamics and information theory. Frontiers in psychology, 6, 818.

del Prado Martín, F. M. (2011). Macroscopic thermodynamics of reaction times. Journal of Mathematical Psychology, 55(4), 302-319.

Friston, K. (2010). The free-energy principle: a unified brain theory?. Nature Reviews Neuroscience, 11(2), 127-138.

Gao, Y., Wang, X., Potter, T., Zhang, J., & Zhang, Y. (2020). Single-trial EEG emotion recognition using Granger Causality/Transfer Entropy analysis. Journal of Neuroscience Methods, 346, 108904.

Hlavácková-Schindler, K. (2011). Equivalence of granger causality and transfer entropy: A generalization. Applied Mathematical Sciences, 5(73), 3637-3648.

Orange, N., & Abaid, N. (2015). A transfer entropy analysis of leader-follower interactions in flying bats. The European Physical Journal Special Topics, 224(17), 3279-3293.

Schreiber, T. (2000). Measuring information transfer. Physical Review Letters, 85(2), 461.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system Technical Journal, 27(3), 379-423.

Stone, J. V. (2018). Information theory: A tutorial introduction. arXiv preprint arXiv:1802.05968.

Ver Steeg, G., & Galstyan, A. (2012). Information transfer in social media. In Proceedings of the 21st international conference on World Wide Web, 509-518.

Zenon, A., Solopchuk, O., & Pezzulo, G. (2019). An information-theoretic perspective on the costs of cognition. Neuropsychologia, 123, 5-18.