尺度量でない量の相関

データは、いつも数直線に乗るものばかりとは限らない。ことば、天気(晴、雨、曇、雪)、色(3原色の混合)、ゲノム(A,T,G,C)など、離散的であったり、多次元であったり、いろんなデータがある。このようなデータも、もしかしたら数直線に乗る形に変換できるかもしれないが、変換の方法は一意的に決められないかもしれない。

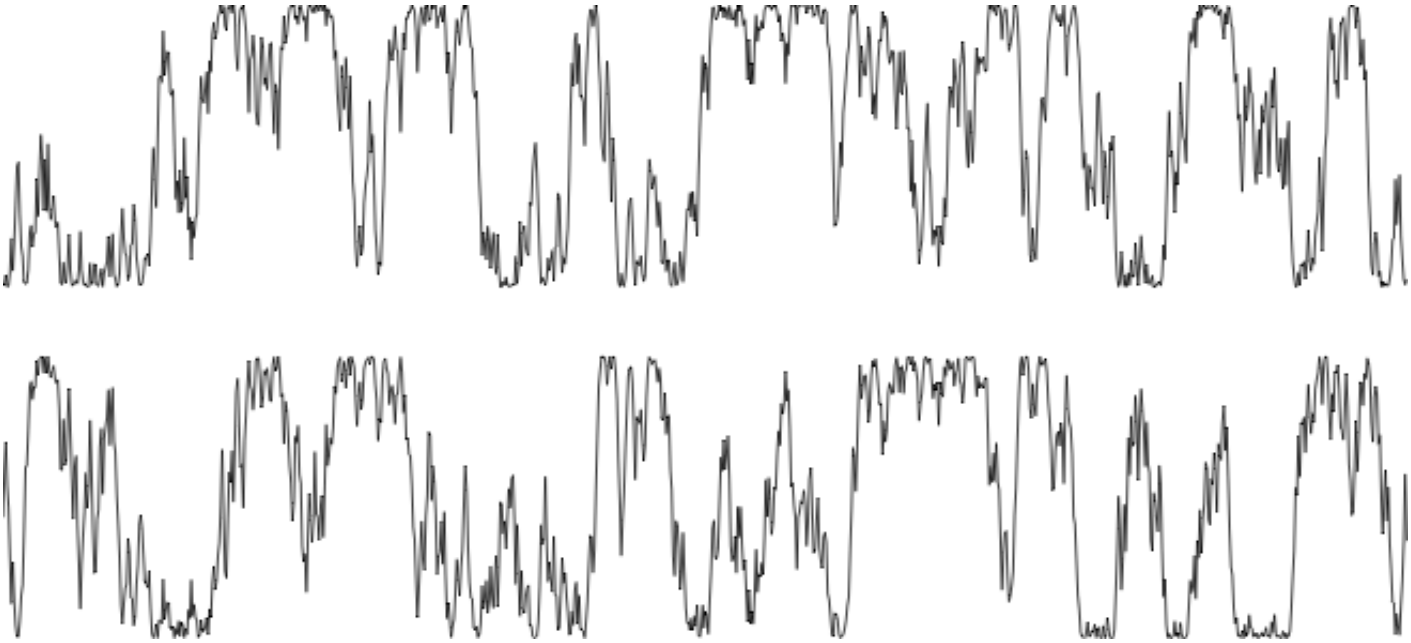
数直線上に表せるデータのことをメトリック(計量、尺度量)データと言い、乗らないものは非計量(nonmetric)データと呼ぶ。非計量データを無理矢理数直線にのせても、まともな解析はできない。非計量データは非計量データのままで扱う必要がある。

例として、次のような文字列の間の関係を捉えたい場合を考える。

```
X:TTACCAAGACCGTCTAACCCAGATTCATCTGATGCTAGTTTGTCCAATCCTAATTGACA
Y:GTTCCTAGTCCATCACTCCCTGGTACTACGGAAGCCTAACGTTCCTTCCCTAGGCGGCC
```

両方ともほぼランダムな文字列だけど、XがCの時は必ずYもCになっている、という関係がある。これを、横軸にてきとうにATCGを並べ、XY平面上に散布させて相関を計算することもできるが、横軸の並べ方に任意性がある。線形相関は、尺度量でないと計算できないし、意味をもたないという点で不便。

また、数直線に乗るデータであっても、相関がいかなる多項式でも表せない場合もある。例えば次のプロットをみてみよう。



上の段と下の段の信号は、密接に関係しながら変化している、どちらもmetricなデータだが、Pearson相関係数は0になるし、多項式でも近似できない。なぜなら、このデータは、円環上をランダムウォークする点の、x座標とy座標を示しているからだ。

このような、一般的なデータに隠れた相関を見付けだすには、情報理論が役にたつが、その前に確率論をおさらいする。

In [ ]: