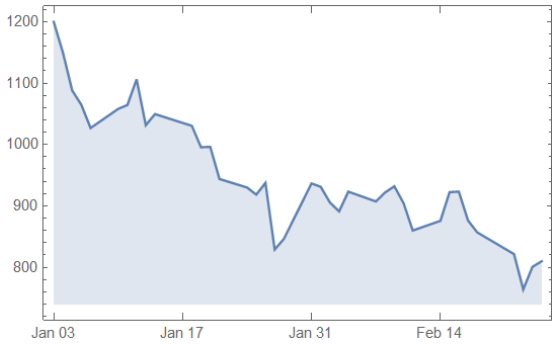
さまざまの距離ためすなりクラスタリング　＜ユークリッド,SBD, DTW＞

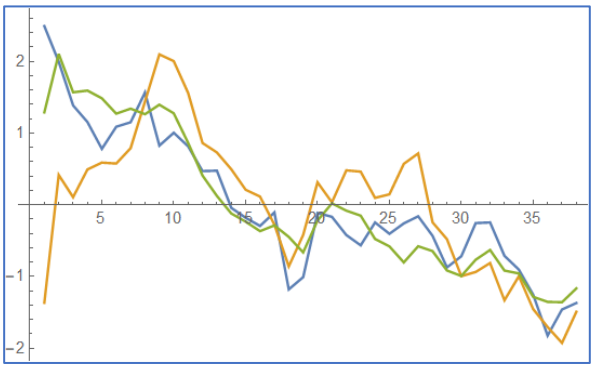
2022 年 2 月下旬　白田由香利

白梅の木に目白が来るのを，リビングに猫と一緒にゴロンと寝転んで眺めていた．3 回目のワクチンはモデルナにしたのだが，予想以上に副作用がきつかった．頭痛がするので，小説も漫画も読めず，テレビも音楽も受け付けない．ただ，猫耳シルエットごしに梅の花と，後ろに広がる青空を眺めていた．無料のワクチン接種は受けさせて頂けたし，家の中は暖かいし，猫は見守っていてくれるし，冷蔵庫の中は苺とプリンとお茶が事前にストック済である．これで頭痛と熱と吐き気さえなければ最高なのだが．

暇なので人生の中で思い出のワンショットを考えてみた．第 1 位は，子供を出産した直後に病院で出してくれたミニミニデコレーションケーキ（どうして，我が子をこの手に抱いたときではないのか不思議だ）．第 2 位は，子供の中学受験の発表を私が見に行き合格リストに名前を見つけたとき．第 3 位は，学習院大に転職が決まって，職場の同僚が開いてくれた会食会で高層ビル窓から見た夜景と頂いた花束．人間はこういう思い出で生きているのだろう．この白梅の景色もまたどこかできっと思い出すのだろう．人生５０年，ではなく９０年はあると思うが，会心の出来といえるような数学グラフィクス教材はあといくつ作ることができるのか，生きているうちにしっかり作ろうと思った．

時系列データクラスタリングの距離の定義について説明する．ユークリッド距離，kShape, DTW（Dynamic Time Warping）の距離の定義の違いを説明する．データは株価とする．以下左図にサンプルを載せる．縦軸は米ドル USD である．あと 2 社，合計 3 社の株価を比較する．比較するために，まずは標準化を行った(下右図参照)．横軸は日時である．



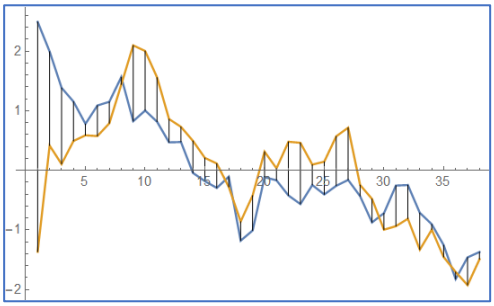


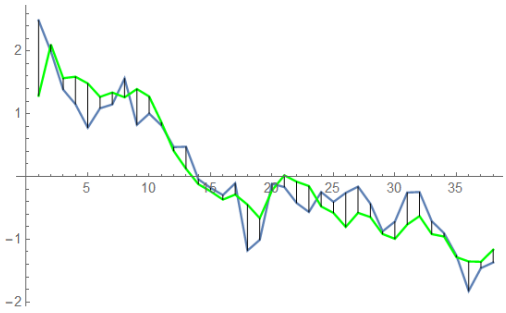
（1）　ユークリッド距離

定義が一番簡単なユークリッド距離から始める．2 次元平面での 2 点間の距離の定義は

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

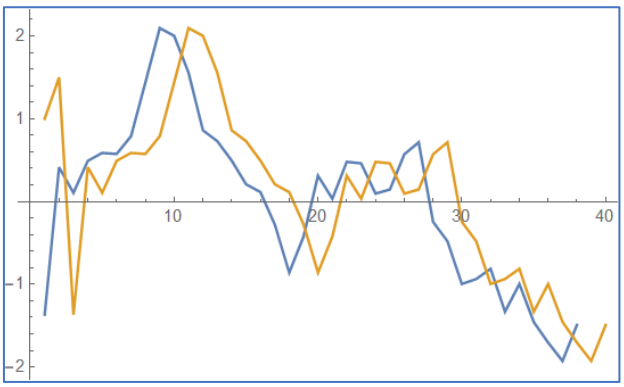
同日時に，株価がいくつ違うのかが距離であり，距離を合計すればよい．右図のように縦に線を引いて，その線の長さの合計で類似度が測れる．3 社の中で最も距離が小さい（類似度が高い）ペアは以下の 2 社であった．





（２） kShape 法で使われている shape based distance(SBD)

時系列データクラスタリングの手法の中で，現在，最も精度がよいと言われて広く使われている kShapez 法の距離は，shape based distance(SBD)である．考え方の要点だと解説する．以下のように，同じ形状のパターンを，2 日分だけずらしたとしよう．この例では黄色いパターンのほうが遅れている．2



つのパターンを関数 f(x), g(x)で表し，f(x)×g(x)の値を当該区間について積分するとする．右図のように w=2 だけずらしたときと，ずらさないとき，この積分値はどちらが大きくなるであろうか？

$$\int f(t) \times f(t-2)dt$$

$$\int f(t) \times f(t-0)dt$$

答えは，ずらさないときのほうが大きくなる．

積分の代わりに離散的に，$\sum f_i \times g_i$ の値を計算して比較すると，に w=2 だけずらしたときは約 19，ずらさないときは約 31 であった．

一般には，f と g は全く形状の違う関数である，そこで，以下のような積分を定義する．f と g の相互相関を計算するのであるが，この積分値が最大となるようにする，ずれ時間 w を見つけて，w だけずらす。上図例では，青パターンを w=2 日分ずらすと最大値となる．

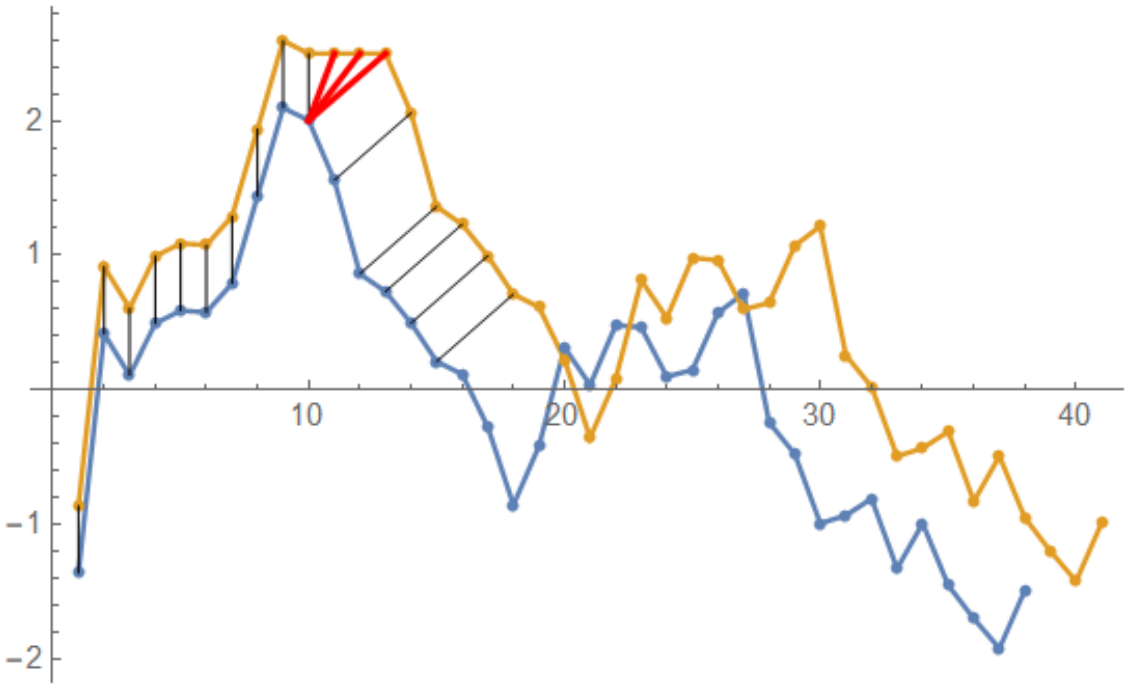
$$(f \times g)(w) = \int f(t) \times g(t+w)dt$$

考え方としては，適当な時間 w だけずらして，相互係数の最大値を計算する．その w のときの相互係数をそれぞれのパターンの自分の距離で割って標準化して，それを値 1 から引いてやる．それが SBD である．1 から引くので，最大値の点は，最小値の点となる．例えば，相互係数が 4.1 で，標準化すると 0.56 で，1−0.56=0.44 が SBD 値となる.こうして，距離最小化の点が求まる．

k Shape の SBD の利点は，ピークや谷が多少ずれていても，同じ点であると認識可能なことである．また，山や谷の大きさ(amplitude)に違いがあっても，同じ点であると認識可能なことである．

（３） DTW

DTW(Dynamic Time Warping)法では，DTW という距離を使っている．



特長は時間軸上の時刻の間隔を伸縮させることである．A パターンが下降を始めも，B パターンがまだ下降を始めないとき，A は時間を止めて足踏みをして，B が下がり始めるのを待つ．上図の赤の線 3 本が示すように，DTW では，青パターンの 1 時刻に対して，黄パターン上の 3 つの時刻が対応していることがある．ユークリッド距離は垂直方向の線分の合計であったが，DTW の距離では，上図のように斜め線になる．DTW のアルゴリズムでは， n 個の時刻点があった場合，n 個の点がもう片方の n 個の点のどれに対応したときに，**全体としての距離が最小になるか**を，総当たり的に調べていく．多対多の関係をすべてチェックしていくのである．そのため，ＤＴＷは非常にＣＰＵ時間がかかる．私のＧＰＵマシンで計算していても，kShape は短時間で終わるが，ＤＴＷは 1 時間以上かかる，ということがよくある．DTW は時間が掛かる．

株価のように日時という時間軸が固定している場合（世界中で 24 時間は 24 時間という意味で），ＤＴＷのように，時間軸を調整するというメリットはあまりないのではないか，と考える．つまり，kShape の SBD のように時間軸は固定で，あとは個々の状況の違いによる比較的少ないずれを w で調整するだけ，という方策のほうが目的に合致していると感じる．それゆえに，株価のクラスタリングにおいては，一般的に言って kShape の方がＤＴＷよりも優れていることが多いと考える．

引用元：さまざまの事思い出す桜かな　　芭蕉

参考文献

[1] Paparrizos, John, and Luis Gravano. “k-shape: Efficient and accurate clustering of time series.” Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM, 2015.
[2] Aghabozorgi, Saeed, Ali Seyed Shirkhorshidi, and Teh Ying Wah. “Time-series clustering–A decade review.” Information Systems 53 (2015): 16-38.