住みつかぬ旅の心やキーボード　＜主成分分析　次元圧縮＞

2020 年 10 月上旬　白田由香利

肩こりが酷くなってきた．一日中コンピュータに向かい仕事をしているのはロックダウンでなくてもいつも同じなのだが，何か改善点はないかと思い，キーボードを富士通のハッピーハッキングに変えてみた．3 月下旬の突然のロックダウンの知らせで，急ごしらえで家庭内に仕事場を作ったが，廉価なキーボードとマウスでは，どこか居心地が悪く，実際に体に影響がでた．また，研究室には科研 B で購入したゴージャスな GPU マシンがあるというのに，家で軽量ノート PC にディスプレイをつなぎ仕事をしているのだから，自分の HOME から離れて異国の地で隔離されている感がある．空港のラウンジで電源を 4 個くらいつなぎ充電しながら仕事をしているような感じである．WiFi ルータ 2 個，携帯電話 1 個，ノート PC1 個の計 4 個である．しかしロックダウン中でも新たな楽しみの発見はある．学生から講義で分からなかったことや質問を受けて，教材を作成することである．グラフィクスを作るため，Mathematica でプログラムを組み，そのグラフィクスを使ってｐｐｔを作り，説明ビデオを作り，MP4 に変換して WEB に掲載する．そして，ここにありますから，とお知らせを出す．リクエストに応えて，オーダーメードの教材を作成する，という Work が自分の性に合っている．とても楽しい．それは，これで理解してにっこりしてくれるであろう学生の存在がネットの向こうにあるからだ．学生の皆様，リクエストお待ちしております．

これから書く内容も学生からのリクエストへの応答である．

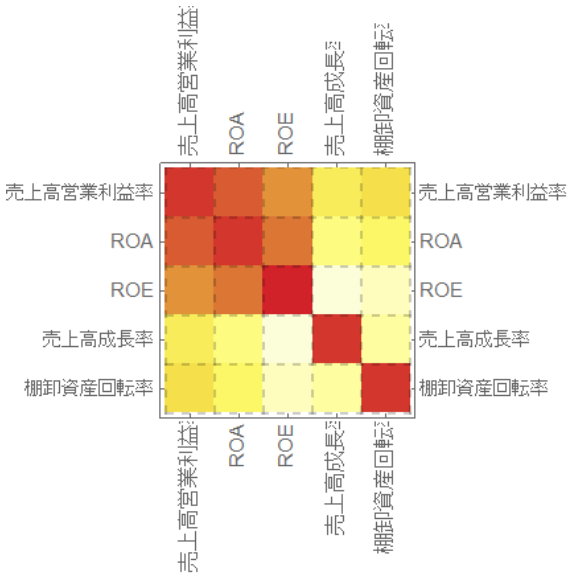
質問：クラスタリングするのに，主成分ではなく，主要な 2 説明変数で散布図をプロットするのではだめですか？

答え：2 つの説明変数には一般に相関があるので，直交する主成分で散布図を描かないとクラスターが発見しにくいので，だめです．

5 次元から 2 次元へ次元圧縮の問題：2019 年度の日本の製薬会社 34 社(残念ながら抗ウイルス薬「アビガン」の富士フィルムは，業種違いで入っていない)の経営説明変数 5 個のデータ（売上高営業利益率, ROA, ROE, 売上高成長率，棚卸資産回転率）を有価証券報告書のデータベース EOL から取ってきた．このデータを入力データとして，Kmeans 法を使って，4 つの会社のクラスターを作りなさい．また，主成分分析を用いて第 1 主成分と第 2 主成分の 2 次元散布図を描き，Kmeans 法によって得られたクラスターが, 主成分分析の散布図上でもクラスターになっていることを確認せよ．

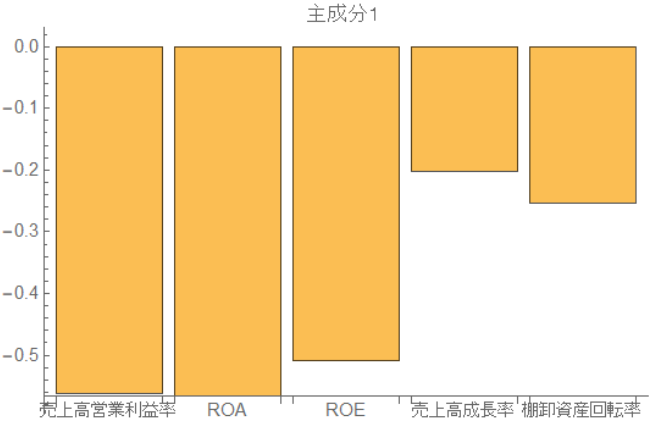
Kmeans 法は機械学習のクラスタリングで最も広く使われている手法である．これを用いて，5 次元データのクラスタリングをする. 入力データは経営指標ごとに標準化を予め行う．その標準化されたデータから，分散共分散行列を作ると下図のようになる．

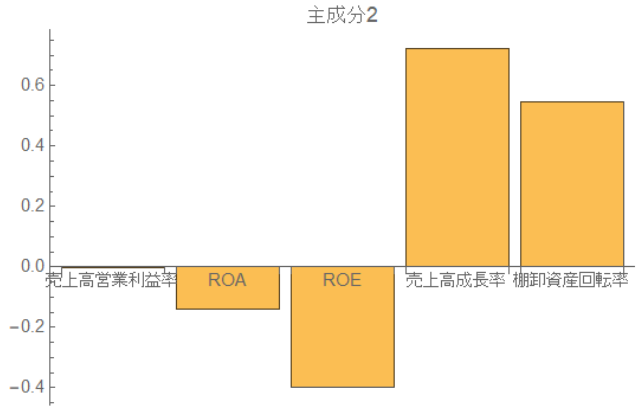
1.	0.93177	0.80018	0.336186	0.377642
0.93177	1.	0.89528	0.249381	0.305716
0.80018	0.89528	1.	0.0285335	0.135281
0.336186	0.249381	0.0285335	1.	0.200112
0.377642	0.305716	0.135281	0.200112	1.



標準化されているので，５個の対角線上の分散は１となる．共分散は，例えば，*Cov(売上高営業利益率, ROA)＝0.93* となる．標準化されているので，共分散の値がそのまま，ピアソンの相関係数となる．右のヒートマップ図は，分散共分散行列の値を色で表現した．赤が１で，白が０である．これを見ると，売上高営業利益率(以下, 利益率)と ROA,ROE の３つの説明変数同士の相関係数が大きいことが分かる．つまり５個の説明変数の中でも，相関の高い説明変数からなる，説明変数のクラスターができている．会社のクラスターと，説明変数のクラスターを混同しないように．主成分分析でも回帰分析においても，説明変数間の相関性を分析することは重要である．

次に主成分分析を行い，上記の分散共分散行列の主成分（固有ベクトル）を求める．主成分は５個あった．第１主成分（固有ベクトル）の要素を見ると，以下のようになる．全ての要素がマイナスの値であることに注意せよ．利益率と ROA,ROE の３つの要素のマイナス値が大きい．





第２主成分も見る．売上高成長率と棚卸資産回転率が大きな正の値となっている．

主成分の５個の要素値は，「どのように説明変数の値をブレンドすれば，主成分値になるか」を示している．以下にその式を示す．掛け算した５項を足している．

第１主成分値＝-0.561631 売上高成長率 -0.566035 ROA -0.509348 ROE -0.202024 売上高成長率 -0.252837 棚卸資産回転率

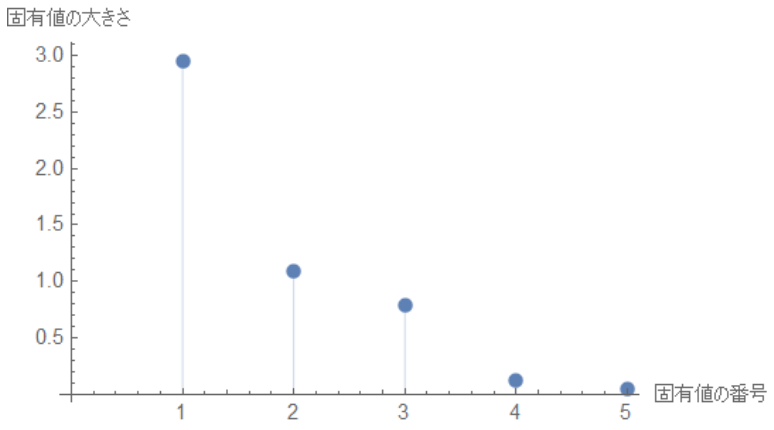
第２主成分値＝-0.00317015 売上高営業利益率 -0.140752 ROA -0.398158 ROE + 0.723353 売上高成長率 + 0.546269 棚卸資産回転率

上記の式に，各社の５個の説明変数値を代入すれば，その会社の各主成分値が求まる．同様に，第５主成分値まで計算できる．

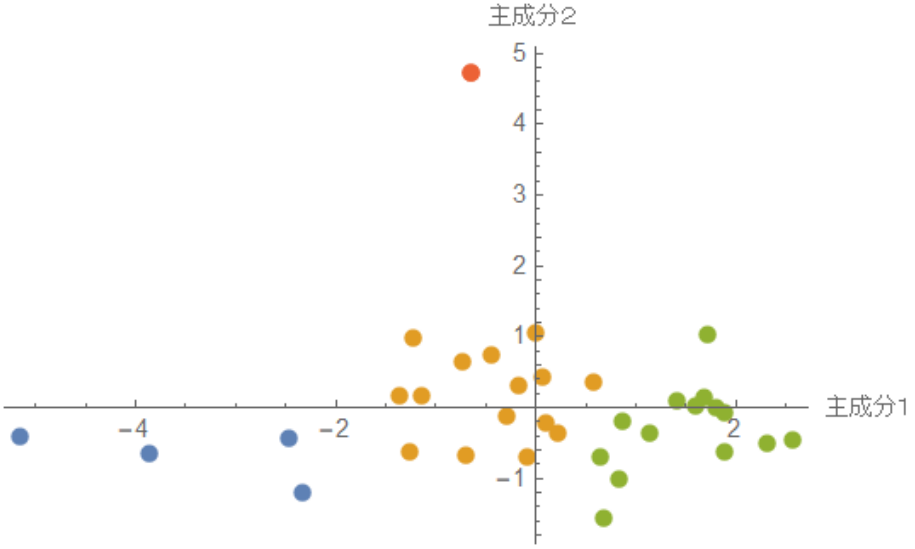
得られた主成分の直感的な意味づけをしたい．第１主成分は**利益率，ROA,ROE** が大きく負に貢献している．先ほど説明変数のクラスターが発見されたが，そのクラスターに対応していると考えてよいだろう．そして，意味に関しても３個とも利益率に関する説明変数である．よって，**第１主成分値は，利益率に関する主成分である**，と解釈できる．利益率は高いほうがハイパフォーマンスな会社であるので，高い値の方が良いことである．係数がマイナスなので，反対に主成分値が低いほうが良い会社，となる．

第 2 主成分を見ると，最も貢献しているのは，売上高営業利益率で，2 番目は棚卸資産回転率である．ROE は負に貢献している．この第 2 主成分に関しては，適当な概念が思いつかない．主成分は，明確な概念と結び付けられる場合もあるが，一般的に概念の名前づけは難しい．「このブレンド比で足し合わせた値」とそのまま解釈するしかないことも多いので，主成分が予想していた概念にぴたりと適合したときは，分析者としてとても嬉しい．

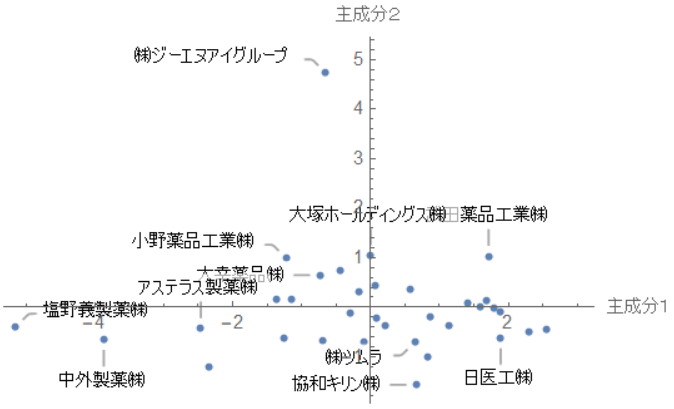
主成分の貢献の度合いは，固有値で見ることができる（右図参照）．主成分分析の項のエキスパンダーで引っ張る力の話を思い出そう．固有値は，第 1 主成分が大きく，第 2 主成分と第 3 主成分の大きさは同程度であることが分かる．



次に，クラスターを見つけるために，第 1 主成分と第 2 主成分を 2 軸として散布図を描く（左図参照）．第 2 主成分方向に，1 社だけ離れている会社があるが，他の会社はどこがクラスターなのか判然としない．そこで，この散布図に，Kmeans 法でクラスタリングした結果で色付けする．



Kmeans 法の結果の色により，4 つのクラスターの分布が分かった．3 つのクラスター青，黄色，緑の主成分 2 の値はあまり変化はなく，主に主成分 1 の値でクラスターが分けられている．主成分 1 は値が負のほうが利益率が高い会社であった．以下に会社名入りの散布図を示す．この図から塩野義や中外，アステラスが，利益率が高いクラスターのメンバーであることが分かる．



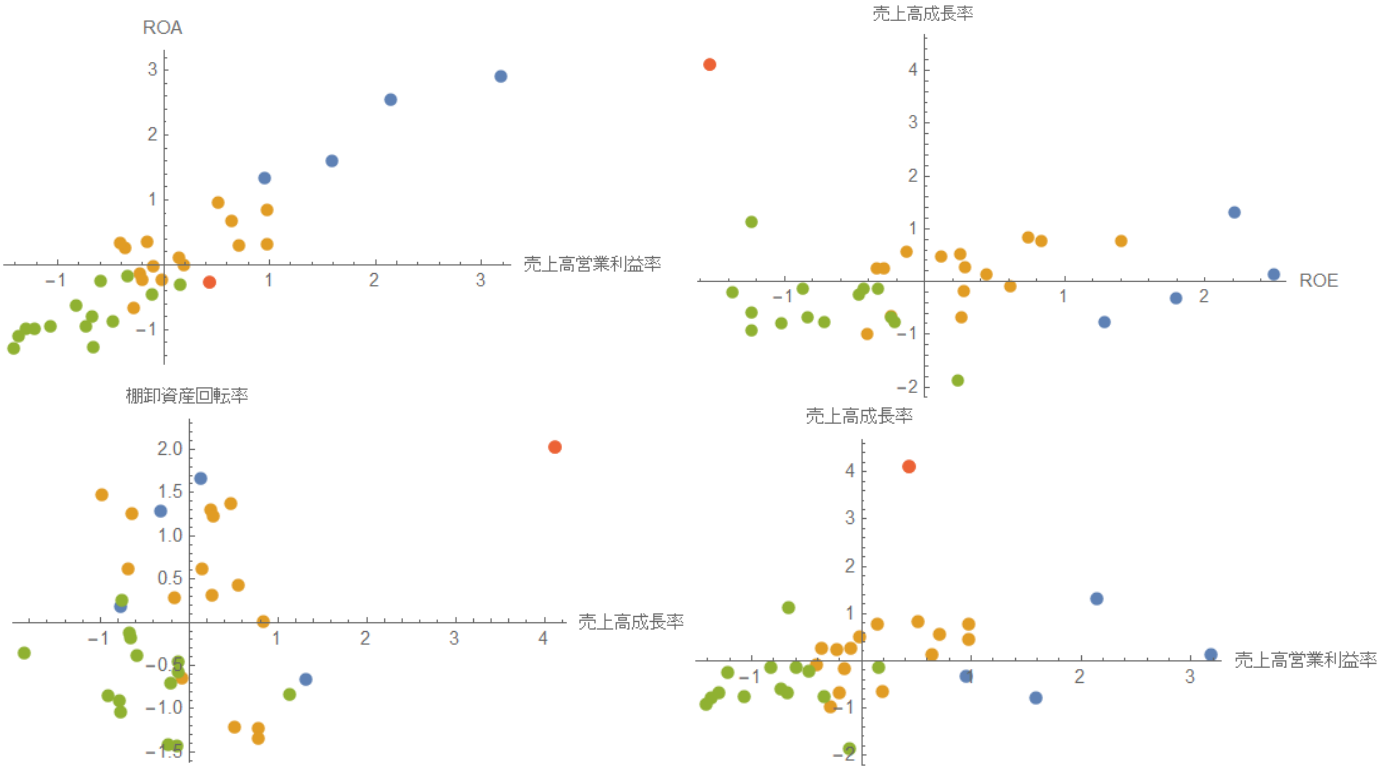
今やっていることは，5 次元のデータを使って Kmeans 法でクラスタリングした，その結果を 2 次元にプロジェクションマッピングした．2 軸に主成分 1，2 を使ったところ，クラスターが分かれていることが確認できた．

これを **5 次元から 2 次元への次元圧縮**と言う．5 次元空間の散布図は見ることができないが，主成分軸の 2 次元にマッピングすることでクラスターが分かれて見えた．

クラスタリングの際は，主成分を軸として散布図を描く，これが基本の一歩である．時により，密度の

高いクラスターがクリアに出現するときもあれば，今回のように Kmeans 法の結果に助けられて，やっと判別できる，密度が小さいクラスターの場合もある．しかし，今回の例でも，クラスターはきちんとクラスターとしてまとまっていた． 次にお見せするのは，適当に 2 変数を選んで散布図を描いても，クラスターにならないという例 4 つである．

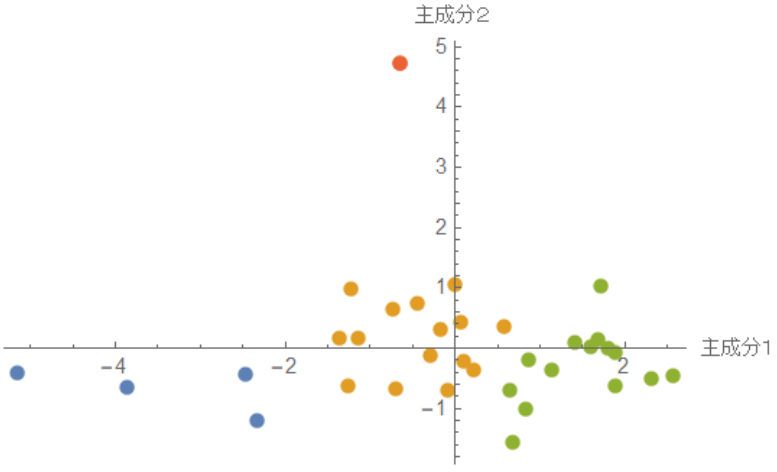
先ほどのヒートマップを見ると，最も相関の高い説明変数は，利益率と ROA であった．それでは，この 2 説明変数の 2 次元空間に 5 次元クラスタリング



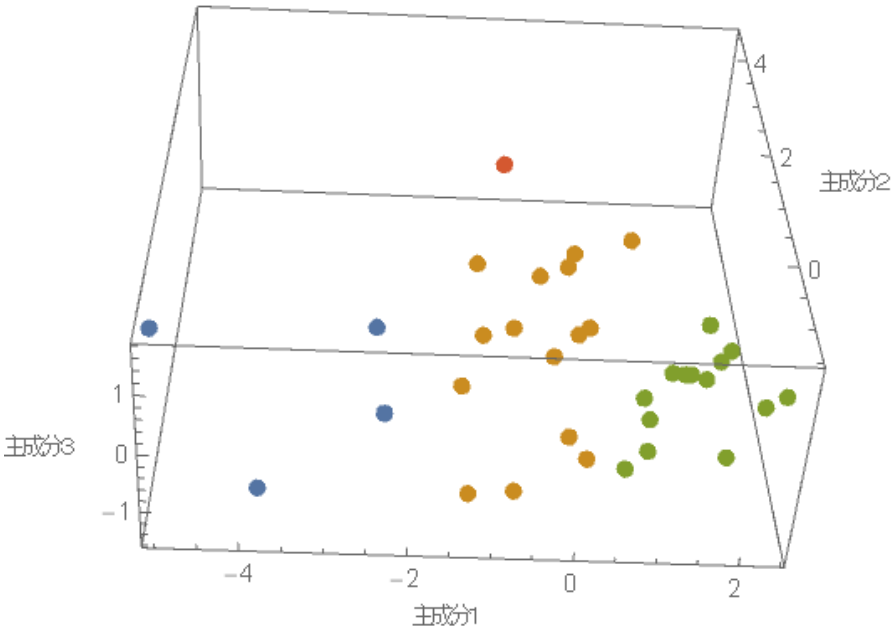
の結果をマッピングしてみよう. 黄色と緑が入り混じってしまった． 2 つの類似した関係の説明変数で散布図を描けば，図のように直線に乗ってくる．

次に，一番相関係数が低い，ROE と売上高成長率でクラスタリングしてみよう(上左図参照)．ここでも黄色と緑が入り混じっていて，クラスターの区別がつきかねる．

上図下段の 2 枚の散布図は， 2 つの説明変数の散布図に 5 次元クラスタリングの結果をマッピングしたものであるが，どちらもクラスターが入り混じってしまう．やはり，主成分を軸にとらないと，5 次元データのクラスタリングの結果をクラスターとして表現できていないことが分かった．



では改めて，主成分 1 と 2 の 2 次元への圧縮，と主成分 1， 2， 3 の 3 次元への圧縮の散布図を見てみよう．いずれも，きれいにクラスターが分かれている．



主成分分析の結果，主成分軸は直交するので，各主成分の相関は無くなる．主成分分析がやってくれていることは，説明変数を 1 次式で足し合わせて再構成して，相関のない主成分という変数を新たに作り出すことだ．

主成分分析にも課題は残る．例えば，真の主成分は，1 次式の結合では表せないような性質をもっていたら(例えば ROA は 2 乗で効く，等)，その場合，1 次式では表現できない．しかし，そこまで考える前に，まずは主成分分析をして散布図を描いてみる，が基本のアプローチである．そこでキュッと固まった密度の高いクラスターが発見できれば Kmeans 法無しでもクラスターの存在が分かる．

終わり

引用元：住みつかぬ旅の心や置炬燵　　芭蕉

主成分分析での多重共線性への対処の課題についての参考文献：

＊López de Prado, M. (2020). *Machine Learning for Asset Managers* (Elements in Quantitative Finance). Cambridge: Cambridge University Press. doi:10.1017/9781108883658　(Section6.5.1)

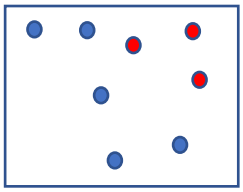
身にしみてプリッキーヌからし秋の風　＜代替効果　多重共線性＞

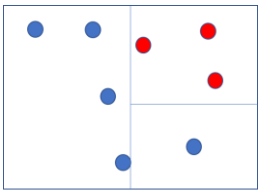
2020 年 10 月上旬　白田由香利

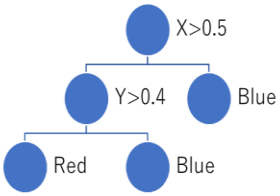
インドネシアにまだ慣れていなかった時のことである．ジャカルタで開かれた国際会議で，日本からの研究者とランチを取っていた．話に夢中で青唐辛子プリッキーヌを避けずに食べてしまった．あまりの辛さに涙が止まらず，水を飲んでも全くとまらず，テーブルでずっと泣いているのも変に見られると思い，席を立った．息が止まりそうに辛い．そうでなくても辛いものは苦手で，インドネシアの先生方と食事を取るときでも，「喉が痛くなると明日の講義に差しさわりがありますので」と，辛いものは避けている私である．昨今激辛といえば，10 倍激辛カレーとか，激辛ラーメンが有名である．お金を出して激辛を楽しんでいる人は唇も喉も，消化器系も丈夫なのだなぁ，と羨ましい限りである．

機械学習の中で, 回帰や分類(クラシフィケーション)をするとき, 複数の変数の中のどれが重要なのか, 変数の重要度を測ることが重要である．相関度の高い変数が複数あると，ある変数の重要度が他の関連する変数によって減じられてしまうことがある．これを代替効果 substitution effect と言う．例えば，喉の痛みの理由の説明変数として，（1）激辛ラーメンを食べる頻度，（2）激辛カレーを食べる頻度，（3）口呼吸をしている頻度，（4）その他喉に悪そうなこと諸々の頻度，を考えてみる．そして，激辛ラーメンを頻繁に食する人は激辛カレーも頻繁に食する，という（1）と（2）に高い相関関係があったとしよう．仮に，1000 人で調べたところ，その相関係数が 0.99 だったとする．相関係数が大きい変数を説明変数に入れて回帰をすると，（1）の重要度が，（2）の存在によって相対的に小さくなってしまう．

分類を使ってこれを説明する．以下の左のような赤と青のデータを分類したいとしよう．







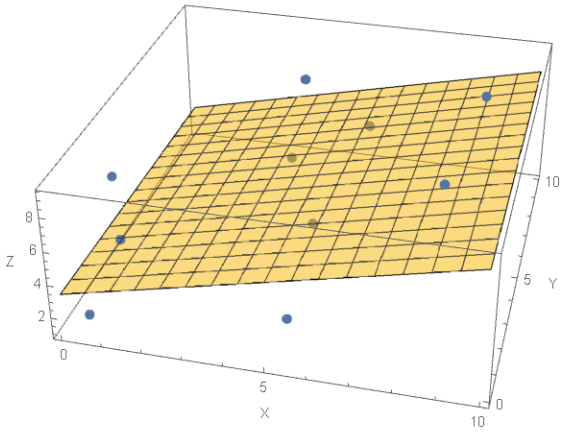
どの説明変数を使って分類したら，上手に分類ができるか考える．上手という意味は，いかに領域分割によって出来た新たな領域の不純度を下げるか，である．私ならば，まず x が 0.5 より大きいか小さいかで分類の線を引く．線の左側は 100%純粋に青だけになった．左側はこれで完了．右側は y が 0.4 よりも大きいか小さいかで分類の線を引く．これによって，赤と青を分類できた．この手続きを決定木で表すと上右図のようになる．ランダムフォレスト法などの決定木に基づく分類においては，どの説明変数によって不純度を減らすことができたかを見て説明変数の重要度を測る．その説明変数によって不純度を大きく下げることができた回数が多い変数が，重要な変数である．機械学習プログラムにおいて，どの説明変数を分割に使うかは，ランダムに抽出してやってみて一番不純度を下げられる変数とその分割の値を決めていく．相関係数が 1 に近い 2 つの説明変数が存在すると，その 2 つが同じ確率で選択されるので，その説明変数の相対的重要度が少なくなってしまうのである．類似している説明変数に半分もっていかれてしまうからである．

回帰における相対的重要度も同様に減じる．激辛の事例に戻って考えると，激辛ラーメンか激辛カレーの頻度のどちらかだけを使って，激辛料理を食べる頻度とするのがよいと思われる．

機械学習においては，代替効果が出ないように，存在する多数の説明変数の中から，適切な説明変数を選択することが大切である．この特徴量の抽出/選択手法は，機械学習のアルゴリズムの発展とともに，

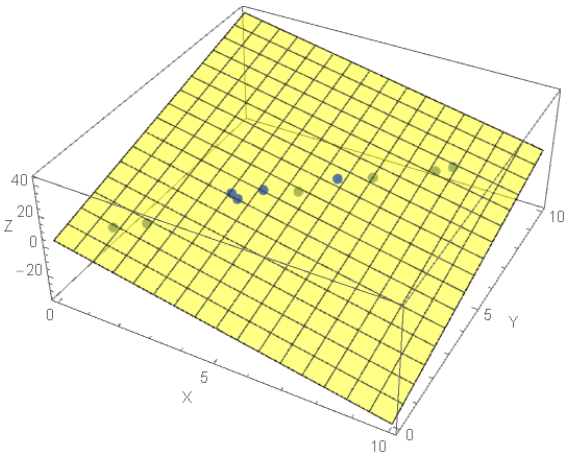
今も次々と新たな手法が作られている．

まずは，説明変数間の相関係数を計算して見て，考えることが重要である．もうひとつの対策として，説明変数の主成分を使う，という方法がある．主成分 1 の軸と主成分 2 の軸は直交する．そして全ての主成分軸は互いに直交する．互いに相関はない．与えられた説明変数の代わりに，主成分値を変数として使えば，代替効果はなくなる．しかし，主成分の概念がぴたりと一言で言えない場合，分析結果の解釈が難しくなる，という欠点は残る．



次に伝統的な線形重回帰分析における多重共線性について説明する．説明変数間に高い相関がある場合，多重共線性（multi-collinearity)があると言う．

まず，多重共線性のない線形重回帰分析の例を示す．2 変数 x，y の線形重回帰分析の例である（左図参照)．与えられたデータの間に回帰平面を求める．その条件としては，実測値 z と予測値 Z の偏差の平方和が最小となるような平面を回帰平面とする．x-y 平面の広い部分に万遍なくデータが散らばっているので，回帰平面がゆるぎない感じがする．多少，データの点の位置がずれても，回帰平面の変動は少ない．



一般に，説明変数間に高い相関(多重共線性 multi-collinearity)がある場合，回帰分析の結果得られた，各変数の係数が信頼できないことがよくある．

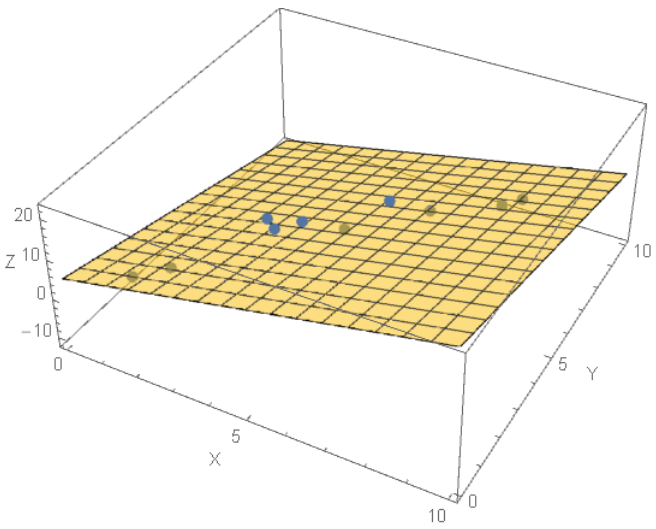
以下に，x と y の間に多重共線性がある場合を示す．与えられたデータに対して回帰平面を求めたところ，

$$z = 2.8 \quad -3.8x + 4.1y$$

となった（左図参照)．

次にこのデータのうちの 1 つのデータを若干動かした．すると，回帰平面は以下のように大きく変わってしまった．

$$z = 4.8 \quad +1.8x - 1.6y$$



なんと，x の係数はマイナスからプラスに変わった(-3.8→+1.8)．そして，y の係数はプラスからマイナスに変わった．どうしてこのようなことが起こるのであろうか？

それは，x と y の相関係数が 1 に近いため，x-y 平面上では，ほぼ一直線になるからである．ちょっとデータが変動しただけで，その直線を軸にして回帰平面が回転してしまうからである．

多重共線性の有無を判別するには，ある変数が他の説明変数の線形回帰式でどの位上手に説明されてしまうかを調べればよい．激辛ラーメンの頻度が激辛カレーの頻度の一次式で殆ど表現できるのであれば，激辛ラーメン頻度は外してしまうのがよい．多重共線性を判別する指標として VIF Variance Information Factor がある．残念ながら EXCEL には VIF を計算してくれる機能はないが，数学ツールの多くは VIF を計算して，多重共線性の有無について分析し，自動的に重要な変数だけを選別してくれる．例えば IBM の SPSS のステップワイズ法は変数の数を増やしたり減らしたりしながら，多重共線性の殆どない変数の集合を選択して，その線形回帰式を提示してくれる．

変数間に多重共線性のあった場合でも，どうしても変数を全部使って線形重回帰を行いたいときは，主成分分析をして，主成分値を説明変数にして線形重回帰をする，という方法がある．しかし，機械学習の代替効果のときと同じく，主成分の概念がぴたりと一言で言えない場合，分析結果の解釈が難しくなるという欠点は残る．

終わり

引用元：身にしみて大根からし秋の風　　芭蕉

代替効果と多重共線性についての参考文献：

＊López de Prado, M. (2020). *Machine Learning for Asset Managers* (Elements in Quantitative Finance). Cambridge: Cambridge University Press. doi:10.1017/9781108883658　(Section6.5)

＊マルコス・ロペス・デ・プラド（著)：ファイナンス機械学習―金融市場分析を変える機械学習アルゴリズムの理論と実践，きんざい，2019 年（8.3 章)．

＊栗原伸一，丸山敦史：統計学図鑑，オーム社，2017 年(97 説明変数間の問題―多重共線性―)．