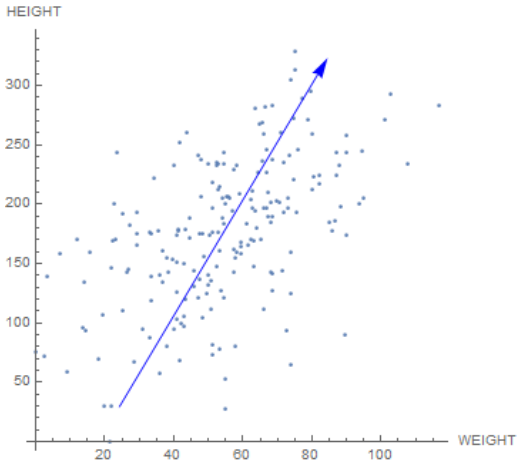
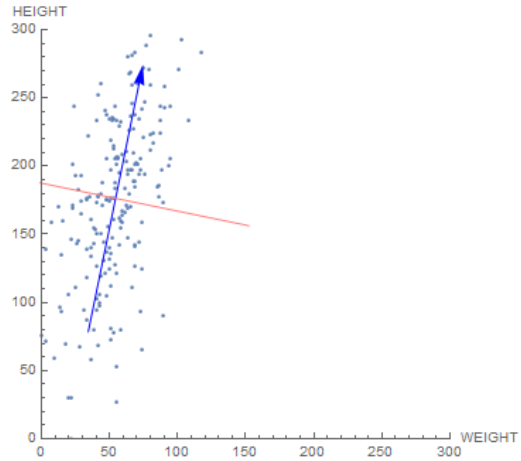
しばらくは数にこもるや秋の夜　＜主成分分析　北北東に進路をとれ＞

2021 年 11 月上旬　白田由香利

滝に打たれる修行があるが、きっと無我の境地に至れるのであろう。鎌倉の禅寺で座禅を組む体験とかはしたことがあるが、死ぬまでに一度は滝に打たれるというのをやってみたいと思う。数学に夢中になって問題を解くというのも、ある意味、雑事を頭から追い出す。複雑な問題を真剣に考えていると、他のことをイメージしている余裕もなくなり、それだけに集中する。コロナ禍で外出できなくても、一心不乱に数学の問題を解くのも心の掃除になるかもしれない。経営数学を教えていると、学会の論文査読で、皆が読みたくない数式の多い（面倒な）論文を回されることがよくある。たいてい、数式だけでなく、それが表わす新しい概念などと組になっているので、その新しいことを理解するまで時間がかかる。AI の手法も日進月歩なので、それを理解するまで、調べものが多くなる。一つの論文を理解するまでに、多くの関連論文を読む必要が出てくる。「この論文、いったい何が言いたいのだろうか？」と考えながら一心不乱に関連論文を読む。

新しいフレームで新しい発見をした論文ほど（つまり新規性のある論文ほど)、こちらの持っている従来パターンからはみ出すので読むのに手間がかかる。読み解くこつは、まず予想を立てて読んでいくことである。図と表をじっと眺めて、要旨と結論を読んで、「こういうことが書いてあるのではないかしらん」と予想を立てることが肝心だ。予想がはずれても、何も考えないで読むよりは理解のスピードが格段に速まる。



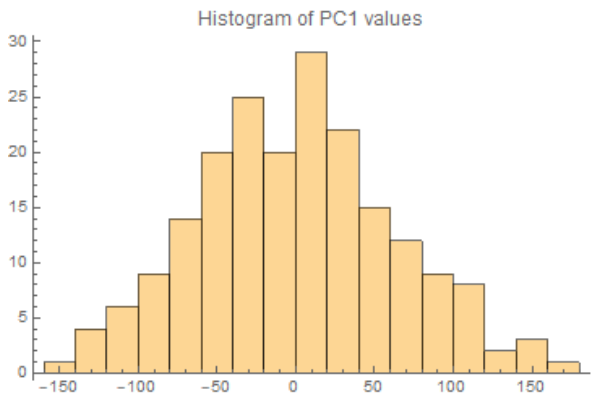

さて、主成分分析の話をする。動物村の住人の体重と身長を測り、散布図を描いた。分布の傾向を見たい。どの方向に分布が伸びているか線を引いてみよう。その線（主成分軸）に沿って、データの値が散らばってみえるような方法に主成分軸を引く。第 1 主成分を PC1 と呼ぶ、第 2 主成分を PC2 と呼ぶ。右図に描いた軸は PC1 の軸である。北北東に進路をとっている。東西南北を使ったのはあくまでも比喩で、分析のときには西南などとは言わないので注意。

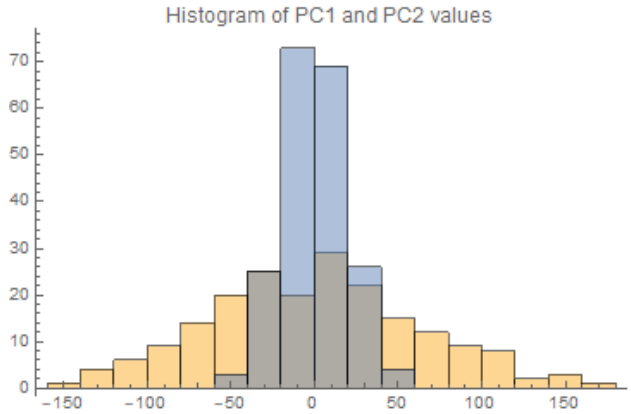
PC1 と PC2 の両軸を描いてみた。軸の範囲を 0 からに変えたので見かけが違っているが、PC1 は同じ方向である。

PC1 軸と PC2 軸は必ず直交する。

この PC 1 軸に各データの点から垂線の足を下したところの値を、各データの PC1 の値と呼ぶ。PC1 の値と PC2 の値はどちらがちらばっているだろうか？　もちろん PC1 値のほうが分散が大きい。

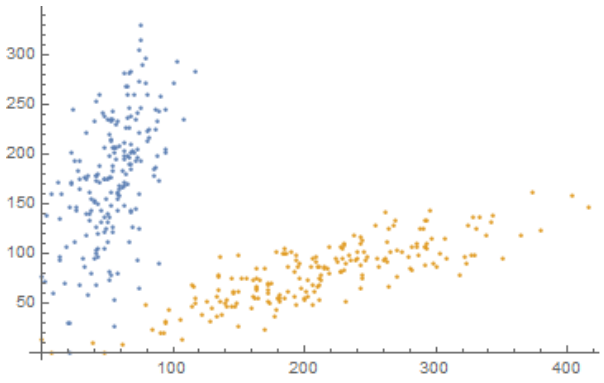
このデータの分布の第 1 主成分は、身長の影響が殆どである。第 2 主成分は、体重の影響が殆どである。

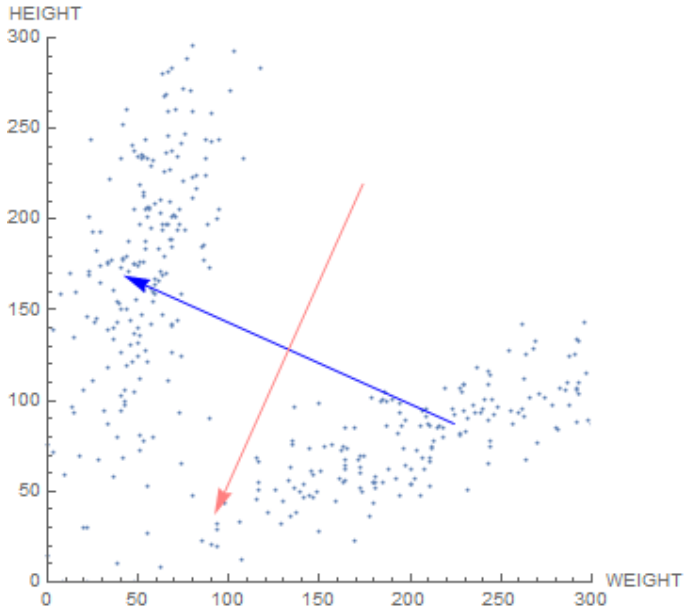



PC 1 の値のヒストグラムを描くと、 0 近くの値が大きく、正規分布のような形状をしている。データの重心（平均と平均）のところを 0 としている。散布図でいうと、PC 1 軸と PC2 軸が交わったところが 0 である。PC1 値の分布は分散が大きい。PC2 値の分布（青）を同じヒストグラムで描くと違いがよく分かる。PC2 値の分布の分散は小さい。

第 1 主成分軸は分散が大きくなる方向に取る。



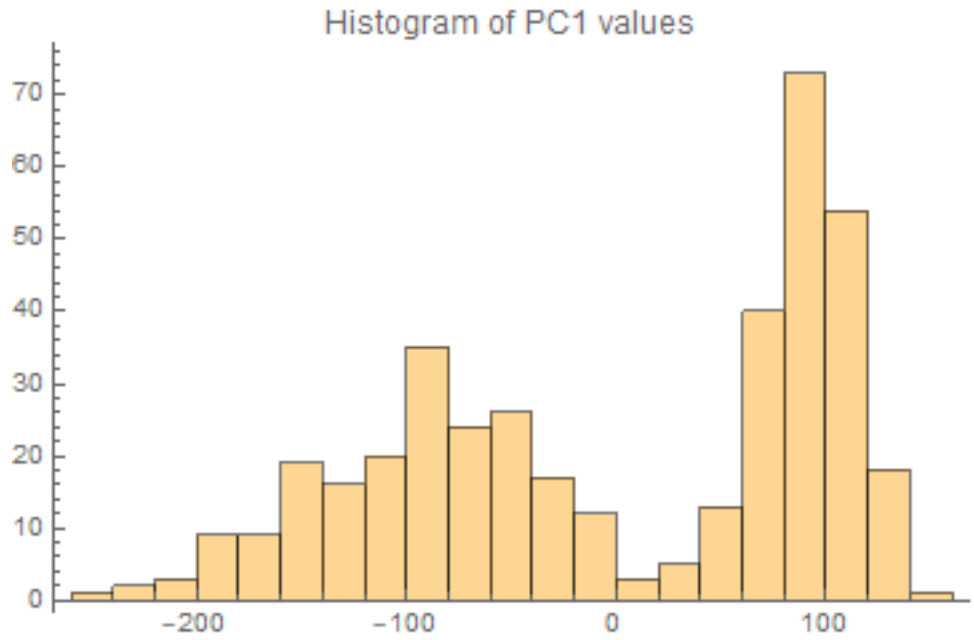




次に、入力データが 2 つのクラスターに分かれていると主成分軸はどうなるかみていこう。動物村には A 種族と B 種族がいて、ゲノム的に大いに違っていたとしよう（このあたり、話はいい加減に書いていますので気にしないで下さい）。入力データとしては、A も B も分からず、混ぜ合わされている。

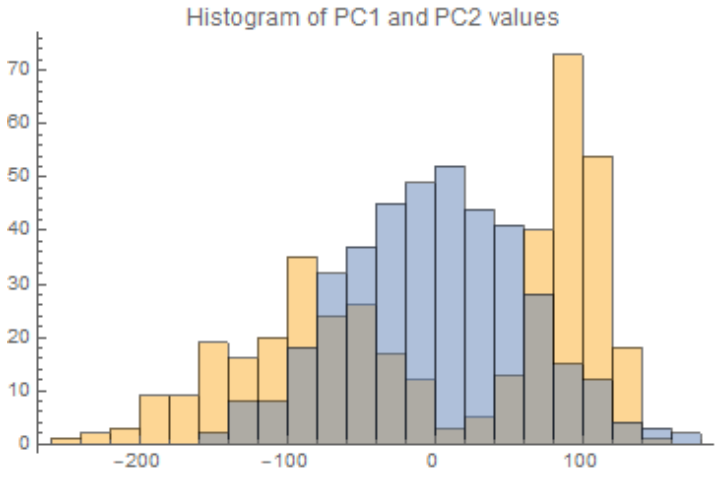
このデータに対して主成分分析を行い、PC1 軸と PC2 軸を求めた。図を見てください。青が PC1 軸、ピンクが PC2 軸です。軸の正方向をどちらにもっていくのかは、問題ではない。こちら方向が正と決めたら、その方針で PC1 値の正負を決めれば、それでよい。どちらを正にするかは、マセマティカなどの統計ツールの計算結果に依存して決まる。変えるのも面倒なので、計算結果をそのまま使っている。

この図では PC 1 値は大きくなるほど、その体重の値は小さくなっていく。このデータの分布の第 1 主成分は、体重の影響が殆どである。第 2 主成分は、身長の影響が殆どである。

それでは、PC1 値のヒストグラムを描いてみる。答えを見る前に、どのようになるのか答えを想像してみよう。




これが答えのヒストグラムです。２つの山があります。これは２つのクラスターに対応している。




PC２値のヒストグラム（青）は、右図に示すように、一山である。

それでは、未知のビッグデータが与えられたとき、上記のようなアプローチでどのように分類をするか考えてみよう。変数としては、サンプルごとに、身長、体重、年齢、収入、摂取カロリー、睡眠時間、等々、多数が与えられたとする。

一体このデータ、どういう分布をしているのだろうか？

困ったときは、初めに主成分分析を行う。PC1 軸と PC2 軸の 2 次元散布図を描いてみる。これで、多くのことが分かる。PC1 軸は、多数の変数のブレンドである。例えば、９５％が体重で、５％が身長をブレンドした座標軸が PC1 軸である、というようなブレンド比が分かる。分散共分散行列の固有ベクトルが主成分軸の方向を表し、その固有値からブレンド比が求められる。

PC1 値のヒストグラムを描くと、クラスターが発見できるかもしれない。

いいですか、ビッグデータの振舞を調べるときには、まず主成分分析にかけてみると、分布の概要が見えてきます。

終わり

引用元：しばらくは滝にこもるや夏の初　　芭蕉