主成分分析について説明せよ

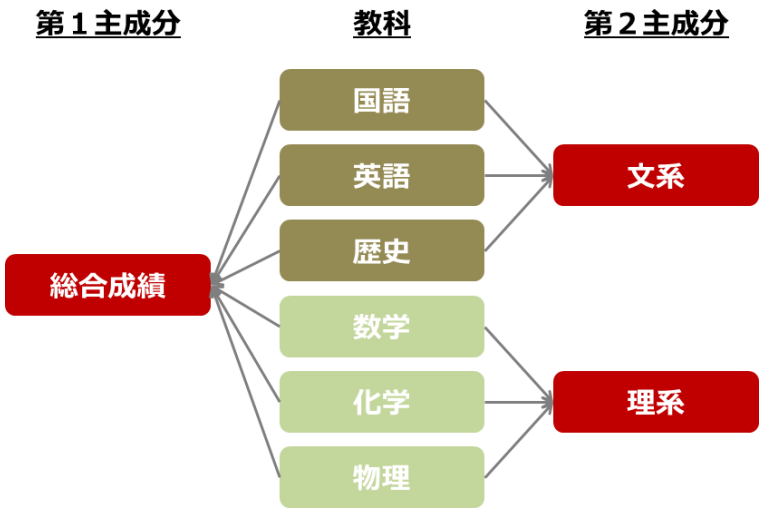
「主成分分析」とは、統計学上のデータ解析手法のひとつです。たくさんの量的な説明変数を、より少ない指標や合成変数（複数の変数が合体したもの）に要約する手法です。この要約は「次元の縮約」という表現で呼ばれることもあります。要約した合成変数のことを「主成分」と呼びます。

わかりやすく言えば、たくさんの次元（指標）のデータから、全体をわかりやすく見通しの良い1～3程度の次元に要約していくことです。たとえば、身長と体重という2次元から、BMI（ボディマス指数）という肥満度を表す1次元の指標に要約するのが主成分分析、と言えばイメージしやすいでしょうか。

ビッグデータは多変量、多次元であるためそのままでは理解しにくいですが、主成分分析を行うことにより、データの持つ情報をできる限り損なわず、かつデータ全体の雰囲気を可視化し、誰もが理解しやすい形にすることが可能です。

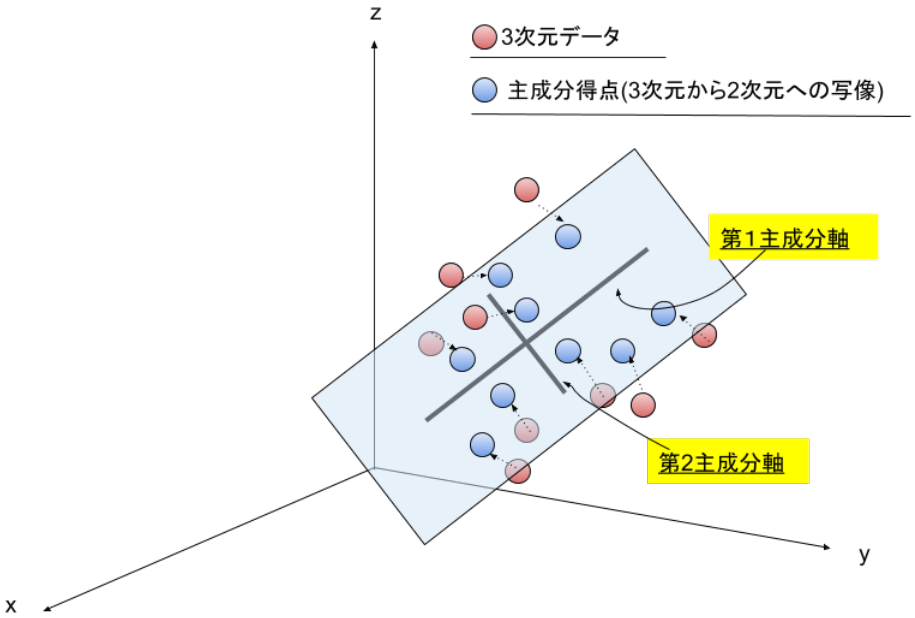
たとえば、10科目のテストを実施している学校があるとしましょう。テスト結果を分析する際、ある教科の点数と別の教科の点数は単純に比較できません。平均点も違えば、点数分布も違うからです。

このとき主成分分析を行えば、第1主成分に総合成績、第2主成分に文系科目／理系科目という指標で、各学生の能力を可視化できます。ある学生の総合的な学力がどのくらいなのか、文系と理系のどちらの能力が高いのかが一目瞭然になるのです。
こちらの例を参考にしたモデル図が、下記になります。



３次元から２次元への写像（$f:\mathbb{R}^3 \to \mathbb{R}^2$）を考えた主成分分析とは、座標で考えると，例えば３次元のデータ（x,y,z座標）を二次元のデータ(l,m座標)に要約（圧縮）するようなものです。

この時，第n主成分を分散の大きい順に，lを第１主成分，mを第２主成分と呼びます。
イメージとしては，三次元空間にある赤い点を主成分軸（この場合第１・第２主成分）にして2次元で表すということです。



主成分分析について数式を用いて説明してください

主成分軸をどのように見つけるのか

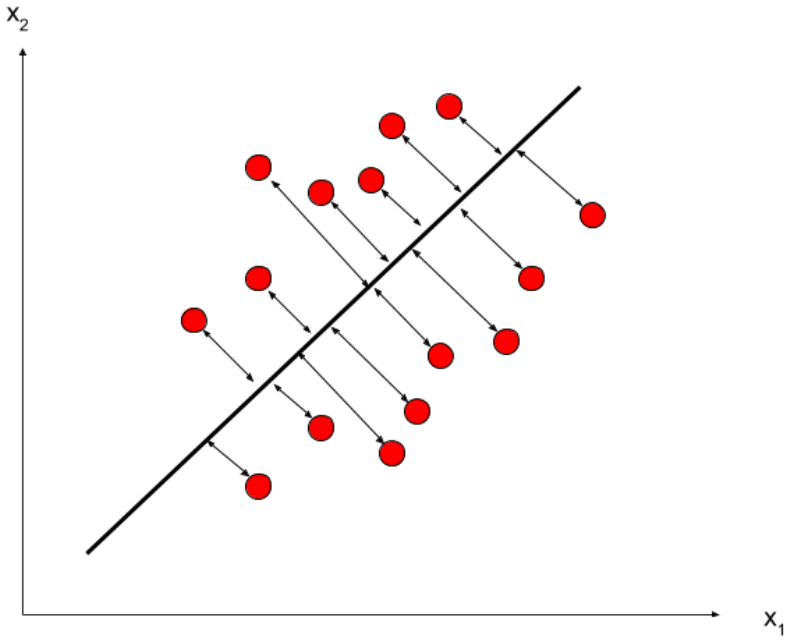
主成分を見つけるためには，分散が最大になるような軸を探します。
分散はどのようなものかといえば以下の式のようにデータxiとデータ平均μの二乗和を平均nで割ったものです。（端的に言ってしまえば，データの散らばり具合を定義していることになります．）

分散の定義

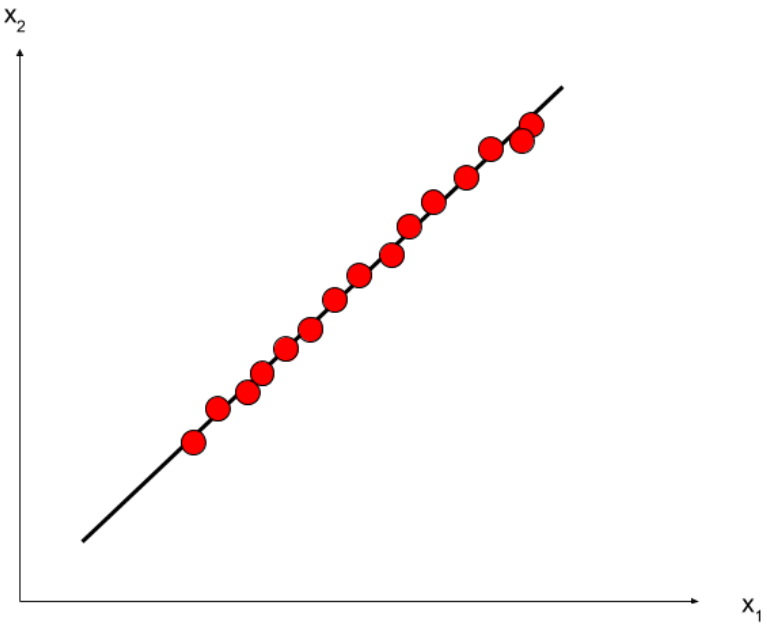
$Var[X]=E[(X-\mu)^2]=1/2\times\Sigma_{(i=1\text{から}n)}=(x_i-\mu)^2$

この時Xを確率変数といいます。右辺のx_iと同じです．定義から$x_i=\mu$の時分散は０になり，$\mu=0$のとき分散は最大となります。（$\mu>0$）
分散が最大となる軸を探すというのことは軸に最も近い確率変数を探すということになります。

下の図のような具合にデータが散りばめられている時、分散は小さくなり情報の損失は大きくなります。



それに対して，ほぼありえないことですが直線にぴったし確率変数が当てはまった場合は分散は最大になり情報の損失は0になります（下の図）。このような軸を求めるのが主成分分析の目標です。



固有値と固有ベクトル

n次正方行列において，　$A\vec{v}=\lambda\vec{v}$ を満たす零ベクトルではないn次元実数ベクトル, $\vec{v}\neq\vec{0}_n$とスカラーλが存在するとき，　λをAの固有値．　$\vec{v}$ →をAの(λに対する)固有ベクトルといいます。

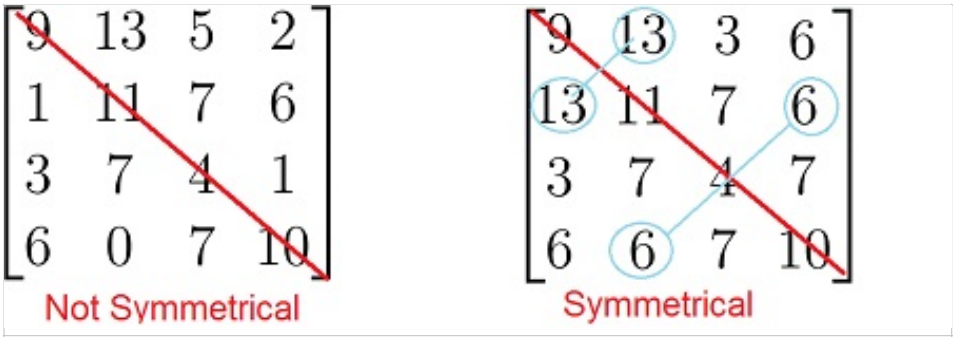
実際には行列に対して固有多項式を定義してやり、それを解くことによって求めます。

主成分分析・因子分析の計算の中身は、行列の変換を行って、不動の軸と、集まり方を集計することです。その軸の呼び名と、集まり方の呼び名が「固有ベクトル」と「固有値」ということです。分析では、固有値の大きさ（と標準化した際の値）から、各変数の成分の大きさを決定しています。

固有ベクトルが直交するのは

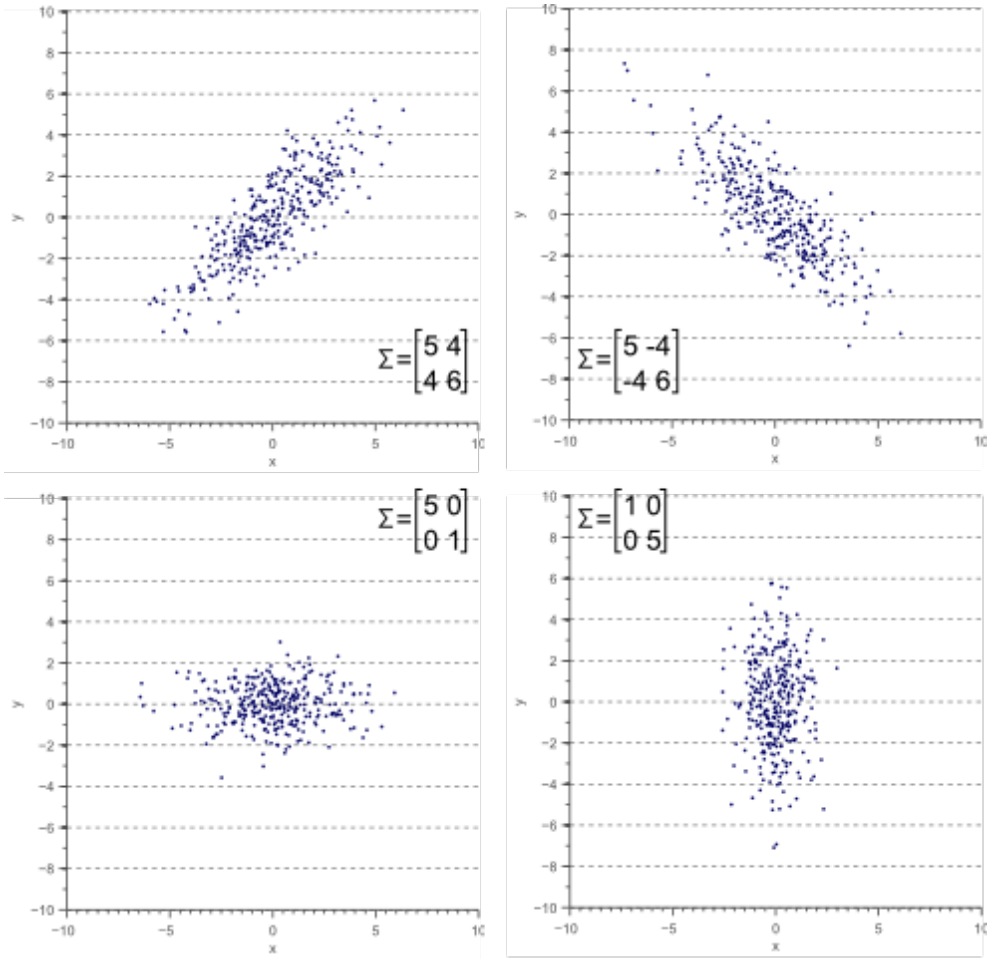
どういったときに、固有ベクトルは直交するのか？答はよく知られていて、「変換行列が対称行列だったとき」です。対称行列とは、行列の左上から右下に引いた対角線を中心に、上三角と下三角が鏡に映したように正反対になっている行列のこと。

もちろん、先の分散共分散行列は対称行列なので、次のような形をとる（synmetric matrix）：



分散共分散行列

データの分散と変数間の共分散を表す数値行列を作成したものが、共分散行列 で、観察されたデータの経験的記述です。 variance「分散」と covariance「共分散」による行列が、「分散共分散行列」。「分散共分散行列」は、行列の構成要素の名称を並べたもので、次のプロットデータと共に示した具体例は下の図で表せます。



上段の分布は、それぞれ「右上がり」「右下がり」なのは、それは分散共分散行列の「共分散」部分である (4, 4), (-4, -4) がそれぞれ「正」「負」のため。下段の (0, 0)の分布の場合、相関は見られない。

まとめると、共分散行列はデータの形を定義します。固有ベクトルに沿った斜め方向のばらつきは、共分散によって示され、x 軸や y 軸に平行したばらつきは分散によって示されます。この解説に沿ってプロットを見ると、下段左では分散 (5, 1) は y 軸のバラツキが小さく、下段右では分散 (1, 5) は x 軸のバラツキが小さい。

主成分分析のアルゴリズム

1）全データの重心（平均値）を算出
2）重心からデータの分散（ばらつき）が最大となる方向（第1主成分）を算出
3）第1主成分と直角に交わる（直交）方向で分散が最大となる箇所（第2主成分）を算出
4）直近の主成分と直交する方向で分散が最大となる箇所（第3主成分）を算出
5）4）をデータの次元分だけ繰り返す