1. 主成分分析

1.1. 主成分分析概要

主成分分析(Principle Component Analysis)とは多次元のデータを次元圧縮（データは減らない）する方法である。主成分分析とは直接は関係ないが，次元圧縮の一例として、プログラマのスキルとして python,Java,給料でプログラマの熟練度を測るために Github 上での star⭐の数でスキル(2 次元とする)を評価できると仮定しよう。それはつまり３次元のデータを 2 次元に要約（圧縮）したことになる。

次に３次元から２次元への写像 $\left(f:\mathbb{R}^3\rightarrow\mathbb{R}^2\right)$ 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）するようなものである。

この時，第 n 主成分を分散の大きい順に，$\boldsymbol{l}$ を第１主成分，$\boldsymbol{m}$ を第２主成分と呼ぶ。

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ある。

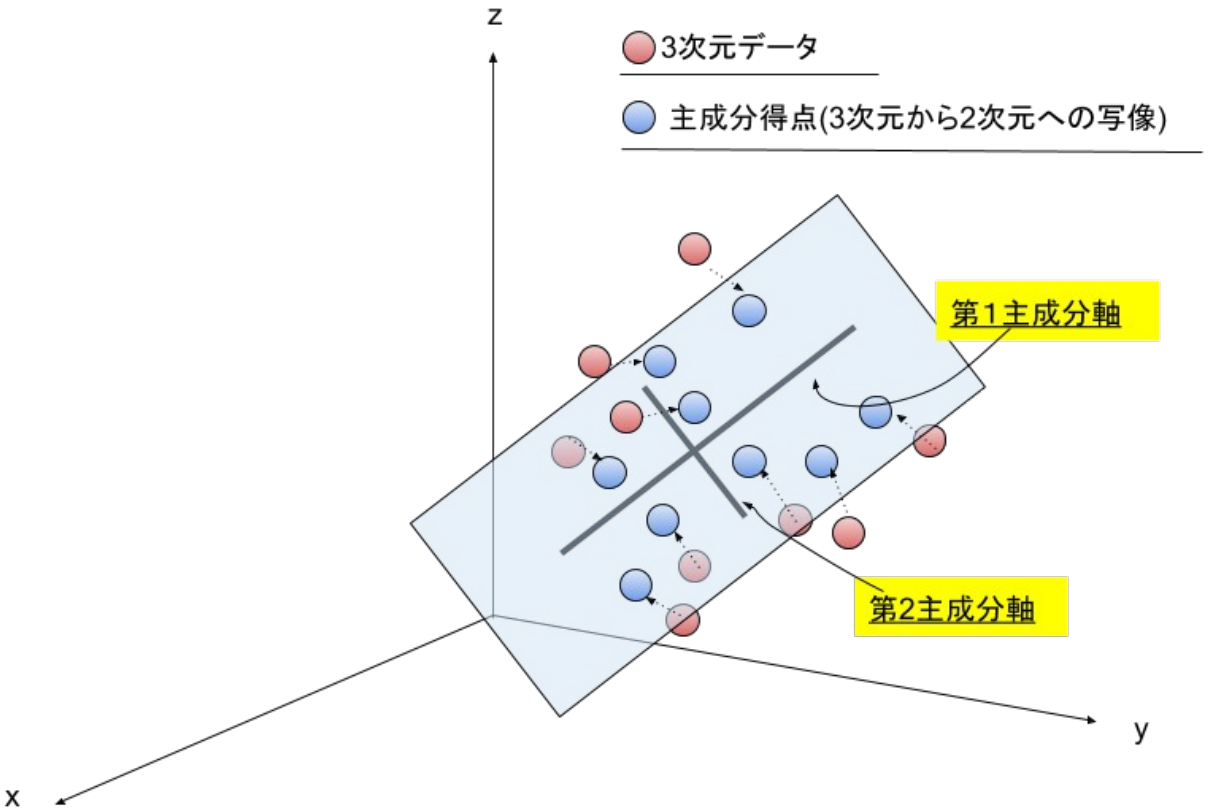

図 1. 3 次元データから 2 次元データ要約図

出典：「意味がわかる主成分分析」URL:https://qiita.com/NoriakiOshita/items/460247bb57c22973a5f0

1.2. 主成分分析と共分散行列

共分散行列とは各成分同士の分散を考慮し、行列にしたものことである。例として、数学と英語のテスト点数があるとする。数学と英語の点数の関係を知りたい、という場合には、複数のデータ群を扱う必要がある。例えば、生徒の「数学の点数」と「英語の点数」がどのような関係にあるか知りたい。数学ができる生徒はやはり英語ができるのか（正の相関）、それとも数学ができる生徒は英語が苦手なのか（負の相関）。

そこで、数学の点数（x の値）と英語の点数（y の値）という、2 つのデータ群を考慮した分散を「共分散」と呼び、この共分散 S_{xy} は次の式で表される。

$$S_{xy}=\frac{1}{n}(x-\bar{x})(y-\bar{y})$$

受験生に対して、「数学の出来具合（数学の点数-数学の平均点）」と「英語の出来具合（英語の点数-英語の平均点）」を掛け合わせた値、の平均を求めている。これが、数学の点と英語の点の共分散で、2 つの科目の点の関係を表す 1 つの指標となる。数学が得意な生徒は英語も得意で、数学が苦手な生徒はやっぱり英語もダメ、という場合には正の値になる。数学が得意だと英語が苦手という傾向がある場合には負の値になる。

式の形から、「x の分散」は「x と x の共分散」と同じなので S_{xx} と表す。同様に、「y の分散」は「y と y の共分散」と同じなので S_{yy} で表す。すると、「共分散行列」が次のように表される。

$$\boldsymbol{S}=\begin{pmatrix} S_{xx} & S_{xy} \\ S_{xy} & S_{yy} \end{pmatrix}$$

1.3. 主成分分析と固有値問題

固有値問題とはある行列に対して、固有値を算出する際に用いる固有値方程式を解くことである。主成分分析するには対象データの共分散行列を算出し、固有値問題を解くことで行える。共分散行列から求めた固有値が大きい順に第一主成分、第二主成分と呼び、各固有ベクトルで線形写像を行うことで基底変換が行える。

以下に主成分分析がなぜ分散共分散行列を対角化する固有値問題となるかを説明する。

訓練データ $x_i=(x_{i1},\ldots,x_{id})^T(i=1,..,N)$ の分散が最大になる方向を求める。データ行列、平均ベクトルを以下のように定義する。

N 個のデータからなるデータ行列: $\boldsymbol{X}=(x_1,\ldots,x_N)^T$

N 個の訓練データの平均ベクトル: $\bar{x}=(\bar{x_1},..\bar{x_d})^T$

平均ベクトルを引き算したデータ行列: $\bar{X}=(x_1-\bar{x},..,x_N-\bar{x})^T$

とすると、平均化訓練データ行列の分散は

$$Var(\bar{X})=\frac{1}{N}\bar{X^T}\bar{X}$$

で定義される。

単位ベクトル： $\boldsymbol{a}=(a_1,\ldots,a_d)^T$ とすると、$\boldsymbol{N}$ 個のデータ $x_i-\bar{x}$ の単位ベクトル $\boldsymbol{a}$ への射影は

$$\boldsymbol{s}=(s_1,..,s_d)^T=\bar{X}\boldsymbol{a}$$

となる。この変換後のデータの分散は、

$$Var(s)=\frac{1}{N}\boldsymbol{s^T s}=\frac{1}{N}(\bar{X}\boldsymbol{a})^T(\bar{X}\boldsymbol{a})=\frac{1}{N}\boldsymbol{a^T}\bar{X^T}\bar{X}\boldsymbol{e}=\frac{1}{N}\boldsymbol{a^T}\bar{Var}(\bar{X})\boldsymbol{s}$$

となる。この分散が最大となる単位ベクトル $\boldsymbol{a}$ は、係数ベクトル a のノルムを 1 となる制約があることを利用して、ラグランジェの未定乗数法を使って求める。

$$E(a)=a^T Var(\bar{X})a-\lambda(a^T a-1)$$

を最大にする a を見つければよい、 λ はラグランジェ未定定数である、a で微分して 0 としておけば、

$$\frac{\partial E(a)}{\partial a}=2\,Var(\bar{X})a-2\lambda a=0$$

より、

$$Var(\bar{X})a=\lambda a$$

となる。 この式は元のデータの共分散行列に関する固有値問題を解くことに等しいので、 分散最大となる単位ベクトル a は固有値問題を解いて求めた固有値・固有ベクトルの中で、 最大固有値に対応する固有ベクトルを a とすればよい。

2．その他、今回の授業で学んだこと

　主成分分析を用いると次元数が多く理解し難いデータも次元を要約することができ、違った味方ができること知りました。また、固有値、固有ベクトルが主成分分析の算出する際に重要となるので、改めて、固有値、固有ベクトルが重要だと再認識しました。