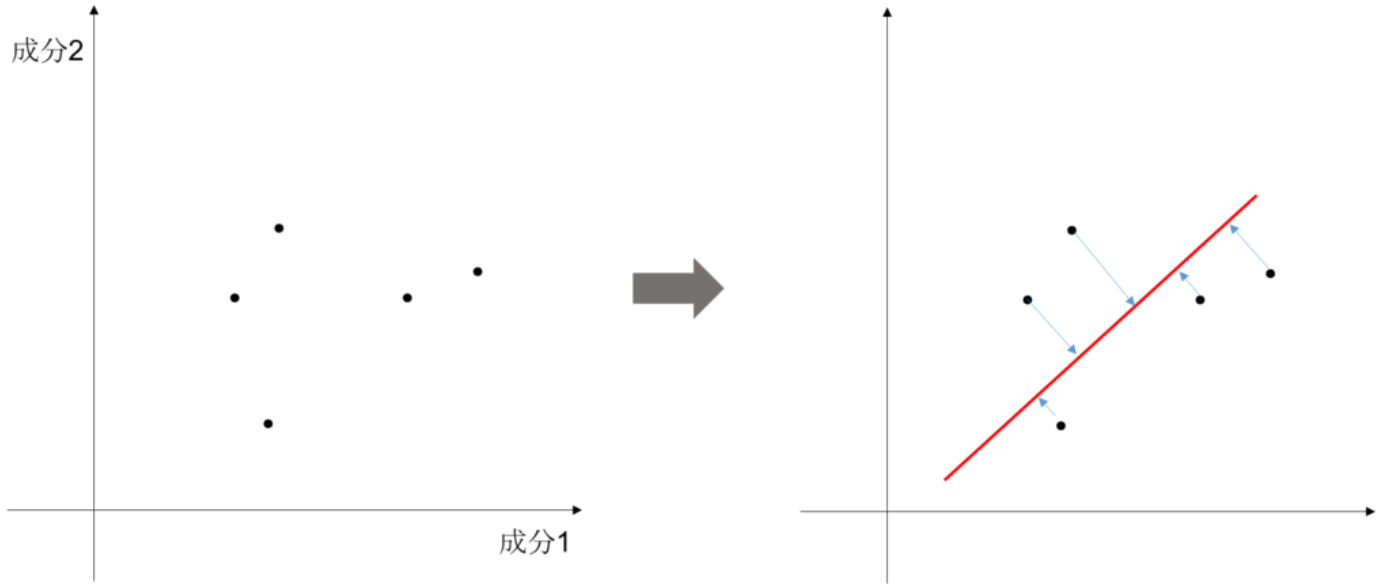
主成分分析についてまとめてください。（自分が理解できていることを採点者に伝えてください。）

●主成分分析とは

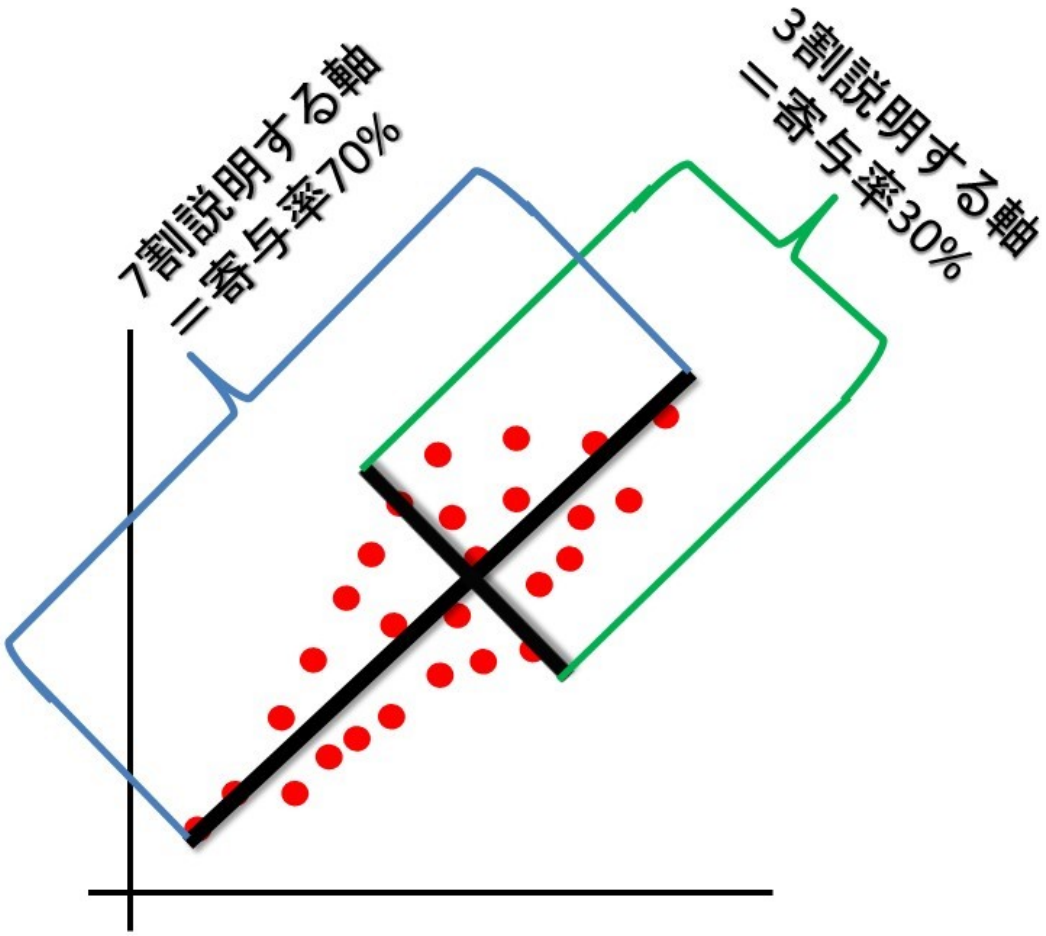
「主成分分析」とは、統計学上のデータ解析手法の一つです。量的な説明変数（因果関係における原因、関数における入力）を、より少ない指標や合成変数（複数の変数が合体したもの）に要約する手法です。この要約を「次元の縮約」と呼ばれることもあります。



具体的に二次元データを一次元データに圧縮する場合、図のようにデータを二次元平面にプロットしたときに何らかの直線に向かって全データを射影するという方法を考えて見ます。射影した結果、直線の垂直方向の情報が完全に失われ、平行方向の情報のみの一次元データが残ります。このときの分散の値が大きいほど各データの点一つ一つの違いをより多く情報として保っていることになり、最も元データの特徴を保存している良い方向になります。逆にバラつき（分散）が少ない方向というのは、各データが共通して持っている自明な情報なので削除しても問題はないと言えます。 まとめると、主成分分析では、データの次元圧縮（縮約）を行う際に、圧縮後のデータの分散が大きくなるような射影をすることで特徴量を自動的に抽出することができるのです。

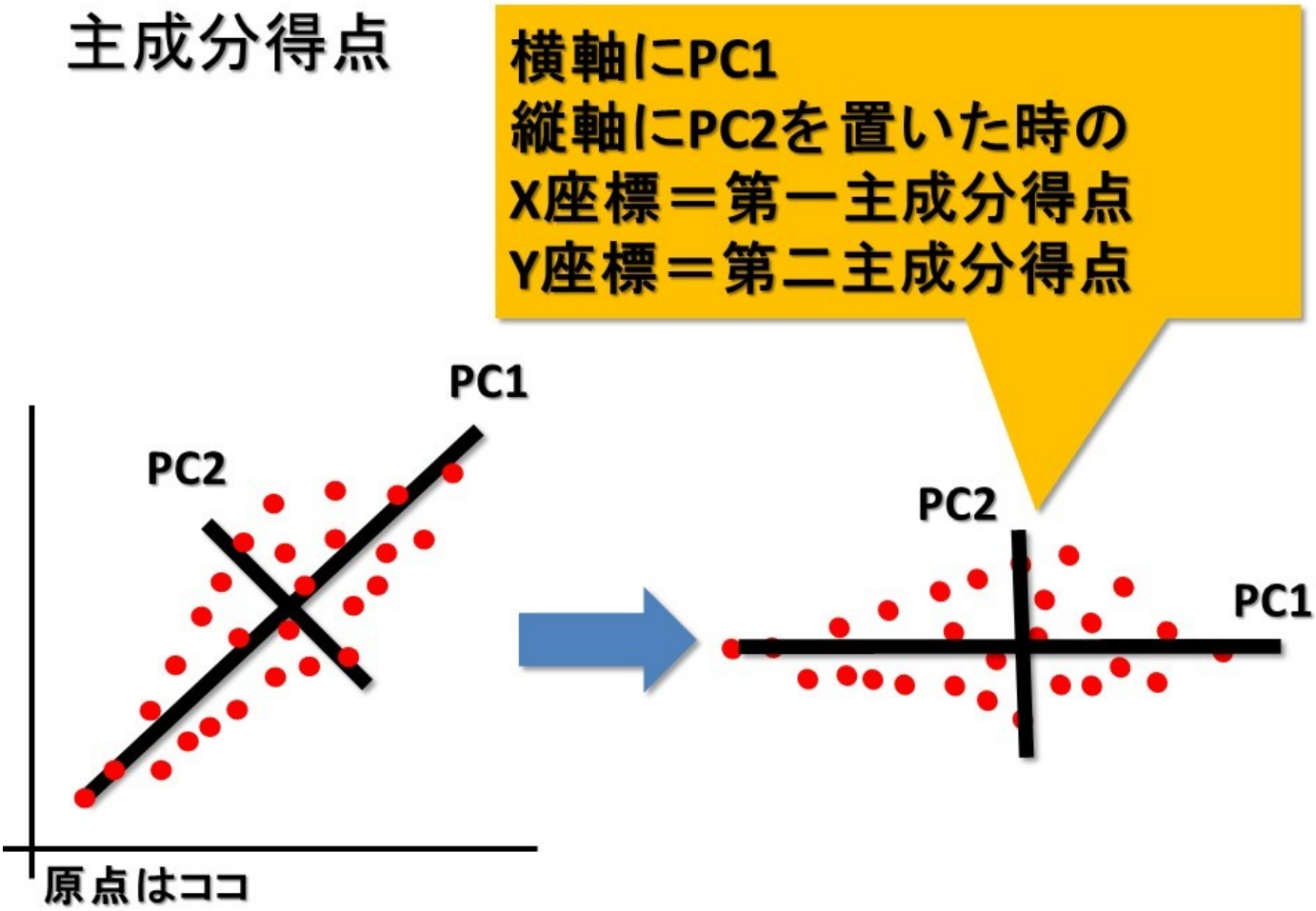
主成分を見つけるためには、分散が最大になるような軸を探します。 数学的には以下の流れで行います。 1.ある方向に射影した時のデータの分散を計算する。 2.分散が最小になるような方向を見つける。 結論的には１を行うと共分散行列が出現し、２を行うとその共分散行列に対する固有方程式が得られます。 実際にデータを主成分分析する場合は、データから共分散行列を生成し、固有ベクトルを計算、全データを射影するという作業を行います。その結果、固有値は分散の値を表していることが導かれ、固有値の大きい方を第１主成分、固有ベクトルを第１主成分軸とし、以下順番に第２、第３・・・というふうに設定していきます。

データの要約という観点からは主成分軸の「寄与率」についての理解が必要になってきます。 寄与率とは、下の図のように「この主成分軸一つで、データの何割を説明できているか」と表したものです。 主成分軸と寄与率の関係は、軸の真ん中にデータを射影した場合、データのばらつきがもう一方の軸の持つ寄与率の分減少します。



データが２次元であれば、２つの主成分軸を使えばデータの100%を説明することができます。しかし、データが３次元以上であれば、２つの主成分軸のみだと全てを説明することは不可能です。データのばらつきの特徴をどのくらい説明できているのかを見るときに寄与率は重要です。 理論上、主成分の数は変数（データ項目）の数だけ定義できます。何番目までの主成分を採用するかは「寄与率」を基準に判断します。

また主成分得点とは、主成分軸を基に、データを回転させた時の座標に相当する値です。第1主成分軸をPC1、第2主成分軸をPC2とすると、下の図のようになります。



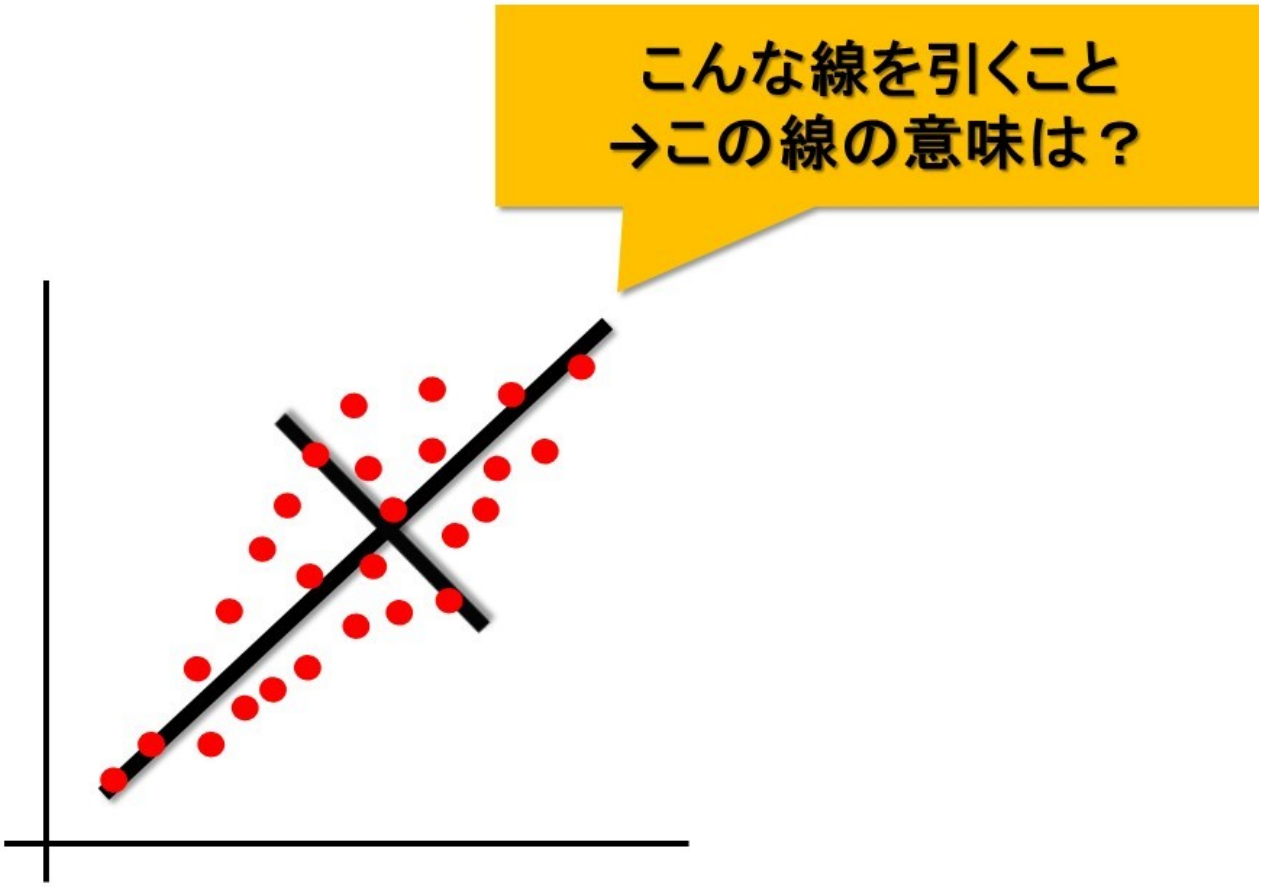
●主成分分析を用いるケース
主成分分析はマーケティングや研究開発など、さまざまな分野で使われています。 例えば以下のような活用法です。 -アンケート調査の結果分析で活用 -メディア企業や商品評価で活用 -研究開発で活用 -画像処理で活用

●主成分分析で何が嬉しいか
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って、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。

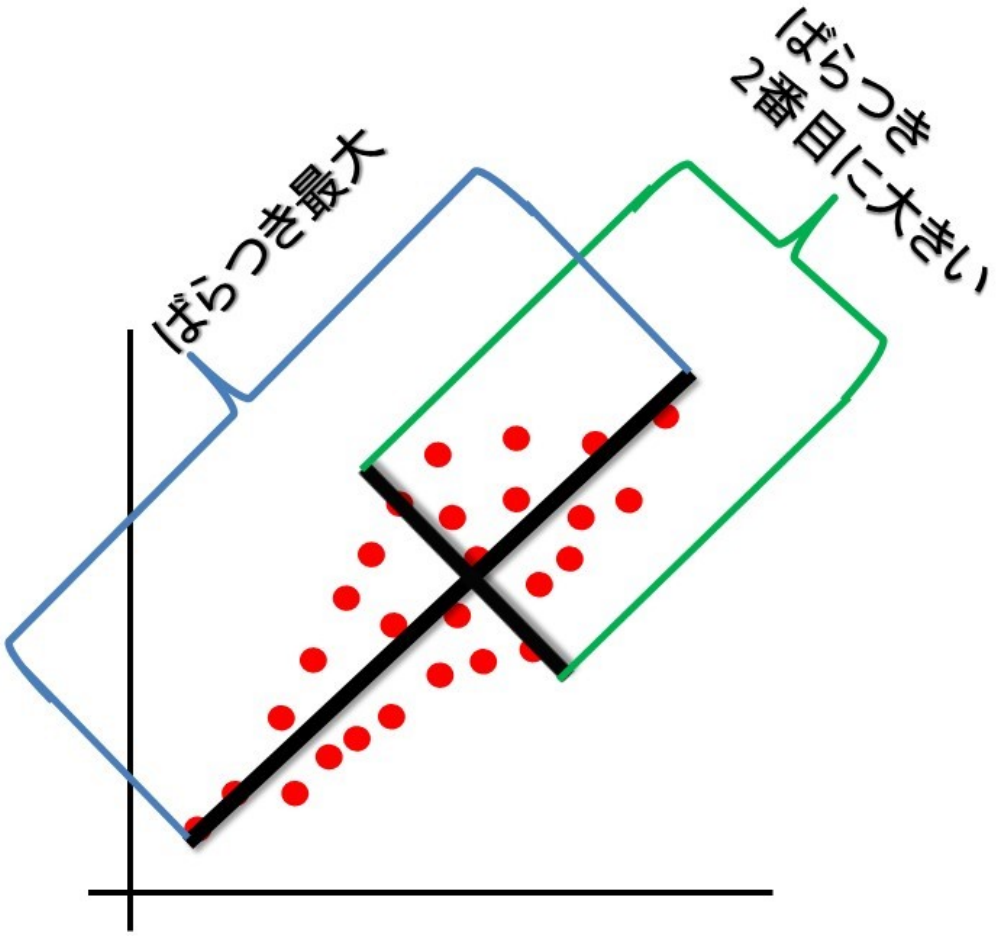
主成分分析について素人にも分かるように簡潔に説明してください。

肥満度を測るためのBMI (Body Mass Index) という指標を考えます。BMIの元のデータは(身長,体重)という2成分を持つデータであったのに対し、BMIはただ1成分の数値となっています。これが次元の圧縮であり、情報を削ぎ落としたにも関わらず、肥満度という特徴を表すのに十分な情報を持っています。このように、データに適切な処理を行えば、情報量の削減と特徴の抽出を同時に行う事ができます。しかし、実際のビックデータでは、BMIの場合のように特徴量があらかじめわかっている場合は少ないです。特徴量とは、データにどのような特徴があるかを数値化したものです。そこで、与えられたデータの傾向から自動的に特徴量を抽出し、その特徴を良く表す低次元データへと次元圧縮を行うのが「主成分分析」です。これは機械学習においては、特に自動的に特徴量を見出すという点において、「教師なし学習」と分類されます。

例えば、下の散布図で表現できる、あるデータが与えられたとします。主成分分析とは、そこにデータのばらつきが最大になるような軸を引くことになります。



主成分分析を通して引かれた黒くて太い線を主成分軸と呼びます。データがよくばらついている、つまり 分散が大きいほど この線はデータのバラつき、すなわち分散が最大になるように引かれています。2本目の線は２番目に幅が広くなるように引かれています。１番目に広いものから順に、第1主成分軸、第2主成分軸・・・と呼ばれています。



主成分分析ができると何が嬉しいかというと、データの要約ができ、それによってデータの特徴を判断しやすくなるという利点があります。そのため、データのカテゴリ分けなどにも応用が可能です。 例えばユーザーの年齢や収入、購買頻度などなど様々な要素がデータとして蓄えられていて、ユーザーのカテゴライズをしたいと思った時を考えてみます。 主成分分析ですと、様々な要素をひとまとめにできるので、例えば「収入が多く購買意欲も高いユーザー」など複数の要素を組み合わせて一つのカテゴリとして扱うことができるようになります。 主成分分析を使ったカテゴリ分けは、視覚的にもインパクトのあるものになります。 以上のように、データの特徴を見てみたい、と思った際に、主成分分析は有力なツールとなります。

主成分分析について数式を用いて説明してください。

二つの変量x、yを持つサンプルデータ（点）が多数あるとする。例えば、x軸が国語の点数、y軸が数学の点数で、点は一人の生徒を表す場合等である。 一人一人の違いを見るとき、x軸の国語の側から見るのと（国語の点数を比較する）、y軸の数学の側から見る事が出来る。 しかし二つの方向から眺めて一眼で違いを判断することは難しい。またサンプルデータの次元が増えればなおさらのことである。もし生徒の散らばり方の情報を最も保持した新しい一つ次元を減らした軸から見れば、一人一人の違い（各点の離れ方）は一目瞭然である。主成分分析は、この様な線を引く。それは座標軸の回転と考える事も出来る。
このように主成分分析とは、新たな座標軸における各点の座標が最もばらけるように（違いが分かるように）、低次元の座標軸を引くことである。そのような座標軸を引くことを計算で求める。 ここで新しい座標軸の単位ベクトルを $(a, b) = (cos\theta, sin\theta)$ とする。 まず、元の座標でデータの中心化を行い分散を求める。中心化とは、其々の軸の平均値に原点を移動する。分散とは、各データの値$(x_1, x_2, \cdots)$から全データの平均値$\bar{x}$を引いた値（偏）差の2乗の総和をデータの個数で割ったもの。従って、中心化させる事で各データの値がそのまま偏差となって計算が簡単になる。偏差の2乗を取るのは、原点を挟んでプラスとマイナスがあるので、相殺されるのを防ぐためである。 そこで、ある点k(中心化後の座標をx_k、y_kとする）と新しい座標軸の単位ベクトルと内積をD_kとすると、$D_k = ax_k + by_k (a = cos\theta、b = sin\theta)$となる。その2乗は以下のようになる。

$$(D_k)^2 = (ax_k + by_k)^2 = a^2 x_k^2 + b^2 y_k^2 + 2abx_k y_k$$

分散（var）を求めるには、全てのデータの２乗の和を求めてデータの個数（n）で割ればいい。以下のように式が展開されるが、a^2, b^2は$(a, b) = (cos\theta, sin\theta)$で定数であるため括り出せる。

$$var = \frac{1}{n}\sum_{k=1}^{n} D_k^2 = \frac{1}{n}\sum_{k=1}^{n}(ax_k + by_k)^2 = \frac{1}{n}\sum_{k=1}^{n}(a^2 x_k^2 + b^2 y_k + 2abx_k y_k)$$

$$= a^2 \frac{1}{n}\sum_{k=1}^{n} x_k^2 + b^2 \frac{1}{n}\sum_{k=1}^{n} y_k^2 + 2ab\frac{1}{n}\sum_{k=1}^{n} x_k y_k \cdots (A)$$

(A)式のa、b以外のところはそれぞれ、回転前の座標軸のx座標の分散（中心化後なので平均0の分散）、y座標の分散、x座標y座標の共分散となっている。これを以下のように定める。

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2 \Rightarrow S_x : x\text{座標の分散}$$

$$\frac{1}{n}\sum_{k=1}^{n} y_k^2 \Rightarrow S_y : y\text{座標の分散}$$

$$\frac{1}{n}\sum_{k=1}^{n} x_k y_k \Rightarrow S_{xy} : x, y\text{座標共分散}$$

また、$a = cos\theta$、$b = sin\theta$から、$a^2 + b^2 = 1$の制約もある。 この制約の中で、分散varの最大値を求めるために、ラグランジュの未定係数法を用いる。 この方法によれば、以下のように関数を作り、Gの最大値を与えるa、b、λを求めれば、Fの最大値を与えるa、bも求まることが分かっている。

$$F(a, b) = S_x a^2 + S_y b^2 + S_{xy} 2ab$$
$$C(a, b) = a^2 + b^2 - 1 = 0$$
$$G(a, b, \lambda) = F(a, b) - \lambda C(a, b)$$

これを解くには、Gをa、b、λで其々偏微分して、＝0と置いた連立方程式を作る。

$$G(a, b, \lambda) = F(a, b) - \lambda C(a, b) = S_x a^2 + S_y b^2 + S_{xy} 2ab - \lambda(a^2 + b^2 - 1)$$
$$\frac{\partial G}{\partial a} = 2S_x a + 2S_{xy} b - 2\lambda a = 0$$
$$\frac{\partial G}{\partial b} = 2S_y b + 2S_{xy} a - 2\lambda b = 0$$
$$\frac{\partial G}{\partial \lambda} = -a^2 - b^2 + 1 = 0$$

上記の偏微分した式をまとめると

$$S_x a + S_{xy} b = \lambda a \quad \cdots (1)$$
$$S_y b + S_{xy} a = \lambda b \cdots (2)$$
$$a^2 + b^2 = 1$$

$$\begin{pmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

この式は共分散行列

$$\begin{pmatrix} S_x & S_{xy} \\ S_{xy} & S_y \end{pmatrix}$$

の固有方程式になっている。 以上から、λは共分散行列の固有値、(a、b)はその固有ベクトルになっているのでそれを求めればよい。 固有ベクトルは通常、a、bの比しか求められないが、ここでは$a^2 + b^2 = 1$の制約があるので、各固有値に対するa、bは一意に決まる。

分散varを最大化するa、b、λはラグランジュの未定係数法により、以下の(1)(2)のように求められた。(1)×a＋(2)×bと置くと、これも＝0となる。 これを整理して、$a^2 + b^2 = 1$の条件を使うと

$$S_x a + S_{xy} b - \lambda a = 0 \quad \cdots (1)$$

$$S_y b + S_{xy} a - \lambda b = 0 \quad \cdots (2)$$

$(1) \times a + (2) \times b$は

$$S_x a^2 + S_y b^2 + 2S_{xy} ab - \lambda(a^2 + b^2) = 0$$

$$\Leftrightarrow S_x a^2 + S_y b^2 + 2S_{xy} ab = \lambda$$

となり、左辺は最大化を目指した主成分得点の分散var（A）に他ならず、varの値は固有値λそのものであることを示してる。このように主成分分析は固有値問題に帰結することがわかる。 各固有値の大きさが、その軸の主成分得点の分散の大きさを表す。それが大きいほど、全体のデータの特徴を、一方向からよく眺められることになり主成分分析の目的に合致する。大きい固有値の固有ベクトルを第１主成分、二番めの大きさの固有値の固有ベクトルを第２主成分と呼ぶ。 対称行列の固有ベクトルは互いに直交する。共分散行列は対称行列なので、その固有ベクトルの第１主成分と第２主成分は直交する。

その他、今回の授業で学んだことを記述してください。

ナイーズベイスを初めて知りました。迷惑メールの分離にも応用されているそうです。もっと詳しい説明を聞きたかったのですが、時間切れで残念でした。 ベイズ理論が応用された手法に興味を持ちました。主成分分析は理解が授業時間内では十分できていませんでしたが、課題をしていく中で腹に落ちてきた感じです。

In [ ]: