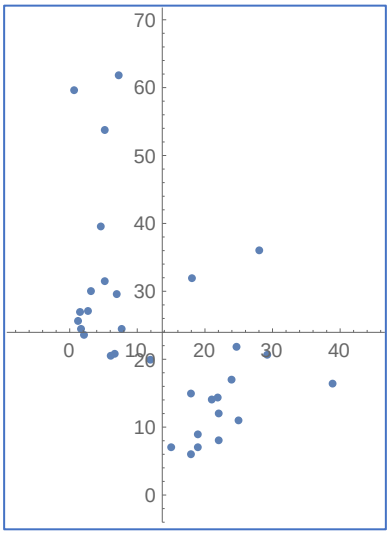
ハイウェイ我を絵に見る夏野かな　　＜主成分分析＞

2020 年 10 月上旬　白田由香利

2016 年 9 月上旬，ANA 深夜便でインドネシア，ジャカルタ空港に飛ぶ．5 時に空港ホテルに到着，7 時にユサンチ氏が迎えに来てくれるまで仮眠をとる．そして，半分寝ぼけまなこで，ユサンチ嬢との再会を喜んだ．ユサンチ氏は，インドネシア国立大学工学部の学生で，2016 年 7 月に 2 週間ばかり，私の研究室に来て一緒に論文を書いたことがある．ユサンチ氏の大型車に荷物を積み込み，助手席へ入り込む．”Prof Shirota, I wonder if you are hungry. Please enjoy this if you would like to have something.” とユサンチ氏がワニの形をした細長い大きいパンを手渡してくれた．それからユサンチ氏の母上の選んでくれたというルアック・コーヒー豆も頂いた．ルアックとは，ジャコウネコのことで，美味しいコーヒーで有名である．2 週間，研究生活をともにしていると，この人は動物が好き，コーヒーをよく飲む，等という私の趣味嗜好がいやでも分かってしまう．今回は，インドネシア国立大学とボゴール農科大学の 2 か所で，“Visually See Text Mining Math Processes on LSA, SVD, and Gibbs Sampling” という特別講義を行なう．テーマは，テキストマイニングで使う数学である．

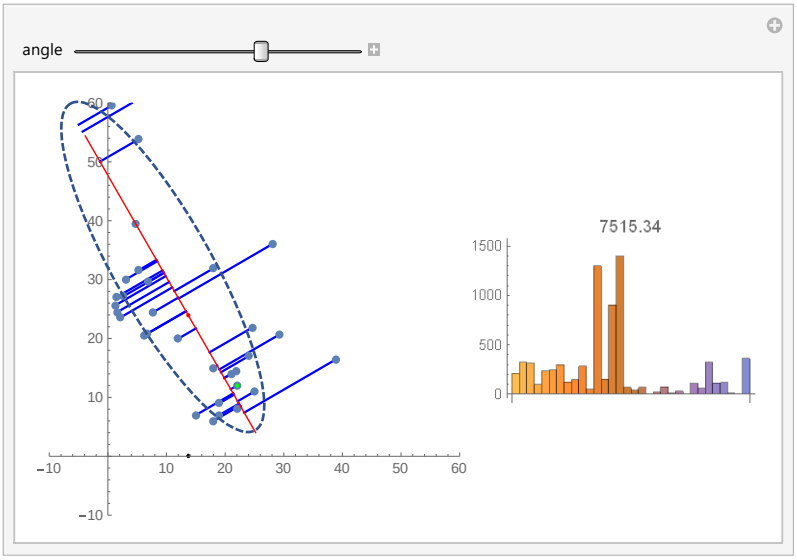
赤道直下の熱い太陽のもと，ユサンチ氏の車でボゴールまでのハイウェイを飛ぶように走り抜ける．両側の眼下に広がるはジャングルの樹海．しかし時折，道路わきにはスターバックスの看板も見られるのが面白い．ユサンチ氏の猫と，ビュンビュン走るトヨタの大型車，その助手席でパンを食べる私．ともに絵にみるような珍道中である．私にとっては，講義に向かう，気の置けない知り合いとのこうしたジャングル・ハイウェイ・ドライブは至福の時である．人に依って人生最高の時間は違うが，動物好きの私にとってはこの一瞬が永遠のように感じられる．



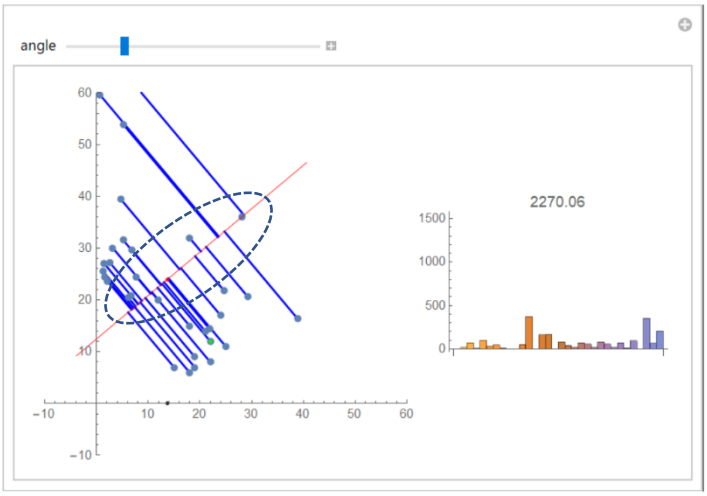
この時の講義でも使った，主成分分析(Principal Component Analysis, PCA)の説明をする．主成分分析は機械学習などでも次元圧縮の手法として頻繁に使われる．

データ{x, y}の集合が以下のように与えられていたとする(右の散布図参照)．主成分分析とは，この散布図に軸を 1 本引くが，その軸に関する座標値の分散が最大となるように引きなさい，という問題である．各点の違いが明確になるように，なるべく垂線の足の位置を分散させたい．この軸を主成分軸と呼ぶ．

これを可視化で説明する．右図は，散布図に主成分軸を引いた様子である．各点から主成分軸へ垂線を下す．その垂線の足の点がその点の主成分軸上の値，主成分値となる．主成分軸上の小さい赤い点が，データの平均(重心)を表している．



主成分軸上でのこの平均からの偏差がその点の主成分値である．平均(重心)を中心として軸をくるくる回転させて，その主成分値の分散が一番大きくなる軸の方向を探す．下図の軸の角度では，赤の重心の周りに垂線の足の点が固まっていて，これは求める解ではないことが分かる．

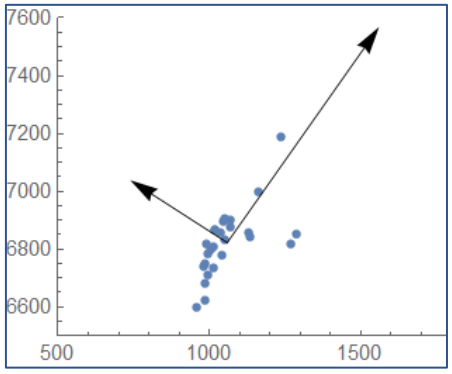


上図のほうが，垂線の足の点が分散して存在しており，分散が大きいことが分かる．点線で囲っている部分が，広い．分散が大きいことは，主成分値が分散していることなので，点と点の違いが明確になる．換言すると，主成分軸の傾きで，データの分布がどちら方向に偏っているのかが分かる．主成分方向に偏って，分布が間延びしている．

これはビッグデータ分析などで，全くデータの分布が不明なときに役立つ．まず主成分分析によって主成分を求めることで，どの方向に分布が伸びているのかが分かるからだ．

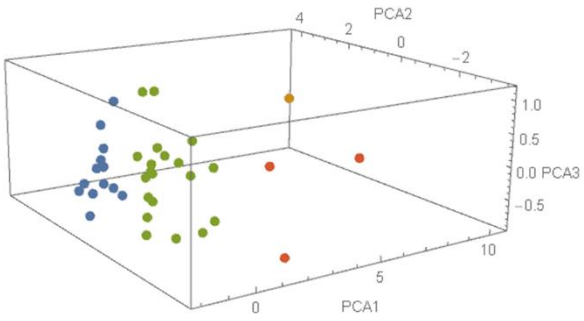
上図中，隣の棒グラフは，各点の主成分値の二乗を表している．これは私が主成分分析の意味を説明するための教材として作ったものである．スライダーを動かして，回転角度を変化させると軸がぐるぐる回る．それにつれて，納豆の糸のように各点の垂線が動く．アニメーションで見ると動きが面白い．

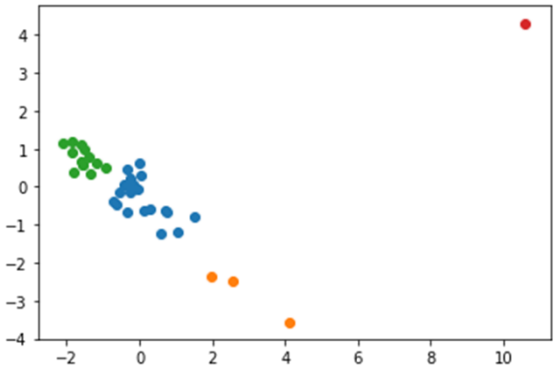
以下に別のデータの散布図を示す．データが x, y のように 2 個の場合，主成分が 2 個ある．長いベクトル方向が第 1 主成分で，短いベクトル方向が第 2 主成分である．主成分の軸は必ず直交する．データが x, y, z の 3 個の場合，散布図は 3 次元となり，主成分は 3 個となる．一般に，変数が n 個あるときは，第 n 主成分まであるが，時には，次元数が落ちて，主成分の数が減る．



機械学習でいう次元圧縮とは，例えば，7 次元のデータ散布図を 3 次元あるいは，2 次元の主成分軸の散布図にプロジェクション・マッピングすることを言う．各データが 7 つの属性値をもつとする．可視化しようとしても 7 次元空間なので，可視化できない．しかしどうしても可視化したい．

その時，主成分分析で，第 1，第 2，第 3 主成分軸を求める．番号が若いほうが，重要な意味をもっているからである．その 3 軸を x 軸，y 軸，z 軸とし，各データの第 1 主成分値，第 2 主成分値，第 3 主成分値を求めて，それを座標として 3 次元プロットする．2 次元に次元圧縮したい場合は，第 1 主成分と第





2 主成分を横軸と縦軸にとり，2 次元プロットする．機械学習のクラスタリングにおいて，クラスターが存在しているか否かを視覚的に見るには，このように主成分分析で次元圧縮したプロットを作ることである．クラスタリングを行うときは，まず，これから初めるとよい．上図では，点に色がついているが，これは Kmeans 法という別のクラスタリング手法によって，4 つのクラスターを生成した結果である．Kmeans 法の結果を主成分分析の主成分軸でプロットすると，確かにクラスターとして固まっていることが見て確認できた．

終わり

引用元：　馬ぼくぼく我を絵に見る夏野かな　　芭蕉

＊ユサンチ氏との共著論文：

Fajrina, Yussanti Nur, Yukari Shirota and Riri Fitri Sar. "Topic Extraction Analysis for Sidoardjo Mudflow Disaster Impacts." Gakushuin Economics Papers 53, no. 3 (2016): 101-114.

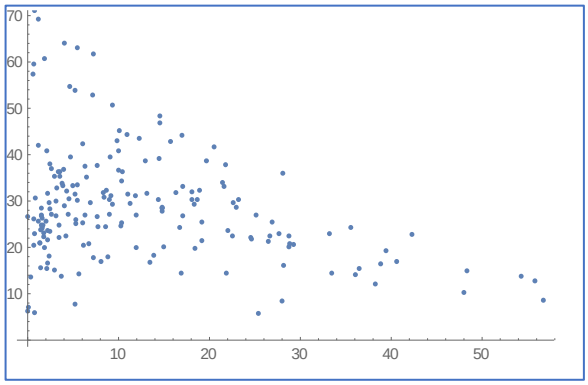
おもてなし　よきもの見せむ猫写真　　＜主成分分析＞

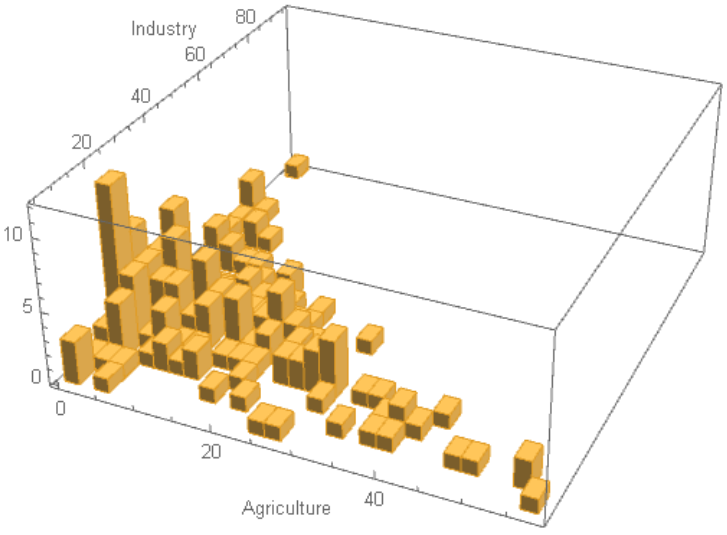
2020 年 10 月上旬　白田由香利

ユサンチ氏が学習院大学に来たのは，インドネシア国立大学との国際交流プログラムの一環で，学習院大の東洋文化研究所がお世話係となってくださっていた．その関係で，東洋文化研が共催する日本文化に関するイベントのパーティに呼ばれ，喜んで参加した．パーティはうちの大学の大学長の乾杯で始まり，宮家の女王様も来賓としていらっしゃるような，カジュアルではあるが格式のあるものだった．服装を心配するユサンチ氏に，白いブラウスとパールのネックレスをお貸しした．「先生，アイロンもお借りできますか？」　ユサンチ氏に，皺のブラウスをお貸ししてしまい，こちらが恐縮してしまった，

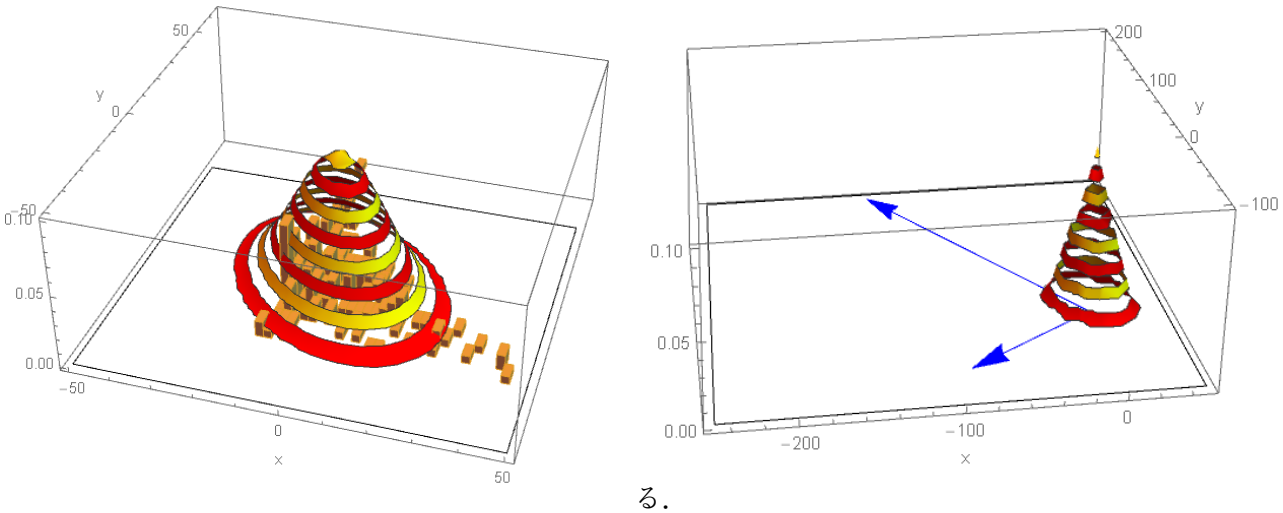
このパーティ開始の前，哲学科の先生（猫がお好き）と，私とユサンチ氏とで，猫写真を見せあった．皆様，携帯電話には愛猫の写真が多数格納されている．ユサンチ氏の実家はバリ島で，兄上の猫は，兄上が家の近くの杜で拾ったインドネシアの純正種の猫だ．早朝，散歩をしていたら，赤ちゃん猫がいたのだそうだ．毛はキジトラで，精悍な顔つきをした，ワイルドな猫だ．ユサンチ氏自身の猫は，白いペルシャ猫で，ブルーの大きな瞳．おとぎ話の中に出てくるような愛くるしい猫である．皆で，可愛い，を連発する．動画もあった．ご実家のリビングルームの大理石の床の上でころころ，ころがっている様子だ．哲学科の先生と院生のかたも，自分の家の猫の写真を見せてくださった．私も，携帯電話を開き，うちのソマリ種の猫の写真を見せる．アーモンド形の瞳で，鳴き声は「ウニャン」と鳴く．これがソマリの特長だ．愛猫家が集まると，国を問わずこのように猫の話で盛り上がる．お互いの猫自慢をし，写真を見せあっていると，ともかく楽しい．　非常に友好的な雰囲気が醸し出された．

さて，主成分分析の話を続ける．以下のような 2 次元のデータが与えられたとする．どのあたりにデータが集中しているかを見るために，3 次元ヒストグラムで表した(右図参照)．





3 次元ヒストグラムは，元の 2 次元データを格子状に分ける．一つのセルに何個の点があるか，つまり頻度を垂直方向に積み重ねていく．点が密のところに高い頻度の塔ができる．さて，この 3 次元ヒストグラムを，2 次元ガウス分布(正規分布)で近似してみる．2 次元ガウス型のお帽子を 3 次元ヒストグラムにかぶせて，形が類似するように，2 方向に引っ張る．その引っ張る方向が第 1 主成分と第 2 主成分であ



る．

主成分軸を数学的に求めるには，データの分散共分散行列を作り，その分散共分散行列の固有ベクトルと固有値を求める．固有ベクトルが主成分軸の方向を表している．そして固有値は，どの位の力で引っ張るかを表している．筋トレでエキスパンダーを使うが，固有値は，えいっとばかりにどれだけの力でその方向に引っ張るかを示している．私は主成分分析の講義ではいつも，エキスパンダーを引っ張る真似をする．

上図のデータの具体的数値を示す．分散共分散行列は以下の 2×2 サイズの行列である．

$$\begin{pmatrix} 146.966 & -57.7582 \\ -57.7582 & 194.497 \end{pmatrix}$$

146.966 が x の分散, 194,497 が y の分散, 共分散が-57.7582 である．第 1 固有ベクトルは$\{-0.556,\quad 0.830\}$.

原点からこの点の方向に伸びるベクトルである．第 1 固有値は約 233．第 2 固有ベクトルは $\{-0.830, -0.556\}$ で，固有値は約 108 である．第 1 固有値のほうが大きい．

主成分分析の講義で，主成分は分散共分散行列の固有ベクトルです，という話を幾度となくしているが，上図の 2 次元ガウス分布のお帽子のグラフィクスをお見せすると，主成分について多くのかたが非常に納得してくださる．

終わり

引用元：　君火を焚け　よきもの見せむ雪まるげ　　芭蕉

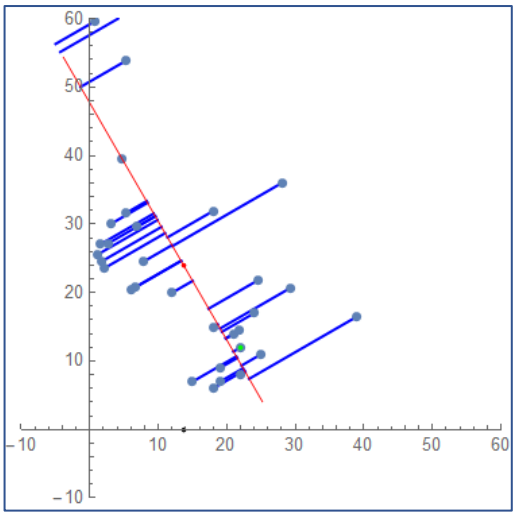
＊共分散については，＜共分散＞の項を参照.

あらたふとラグランジェから固有値が　＜主成分分析＞

2020 年 10 月上旬　白田由香利

主成分分析の話を続ける．

主成分分析とは，散布図中に主成分軸(z 軸と呼ぶ)を 1 本引きたいのであるが，その時の条件として，各点が z 軸へ下ろした垂線の足の点(z 値)の分散が最大となるように引きなさい，という最大値問題である(右図参照)．



データの平均(重心)が図中，小さい赤点で示されている．この点の周りで z 軸をぐるぐる回転させて，分散最大となる傾きを求めるという問題である．

分かりやすくするため，赤い重心を原点に，ずらしておこう．平均点をが原点になるように，予め，全データをずらしておく．

ここで，直線と点の距離についての公式を述べる．

直線 cx+dy=0 があったとする．点(x_0, y_0)（この直線上にはないとする）とこの直線との距離は $\frac{|cx_0+dy_0|}{\sqrt{c^2+d^2}}$ となる．そして，点(x_0, y_0)の垂線の足の点と，原点との距離は $\frac{|dx_0+cy_0|}{\sqrt{c^2+d^2}}$ となる．この証明は本題ではないので省略する．直線の傾きを決める c, d が y と x の係数になっているところに注目してください．

主成分分析は，以下のような，z 値の分散を最大化する最適化問題である．

＜z 値の分散の最大化問題＞

軸上の各点の値 z 値を z = ax+by と置き， z の分散が最大となる a, b を求めたい．制約条件として $a^2 + b^2 = 1$ とおく．

制約条件をこのようにおくのは，もし制約条件がないと，a:b の比率を表すのに無限の組合せができてしまうからである．2:3 = 4:6 = 6:9 というように組合せは無限にある．制約付き最適化問題なので，ラグランジェの未定乗数法で解いていこう．分散 Var(z)の式を，分散についての公式を使って以下のように変形していく．Cov(x, y)とは，x と y の共分散である．

$$Var(z) = Var(ax + by) = a^2Var(x) + 2ab\ Cov(x, y) + b^2Var(y)$$

この関数は，x と y のデータは与えられているので， Var(x), Cov(x,y), Var(y)は計算して具体的な数値が求められる．変数は a と b である．ラグランジェ関数 F(a,b)を以下のように定義し，偏微分をしていく．ラグランジェの未定乗数法の説明は省略する．

$$F(a, b, \lambda) = a^2Var(x) + 2ab\ Cov(x, y) + b^2Var(y) + \lambda(1 - a^2 - b^2)$$

$$\frac{\partial F}{\partial a} = 2Var(x)a + 2\ Cov(x, y)\ b - 2\lambda a$$

$$\frac{\partial F}{\partial b} = 2Var(y)b + 2\ Cov(x, y)\ a - 2\lambda b$$

この 2 式がゼロの値を取るとおく．

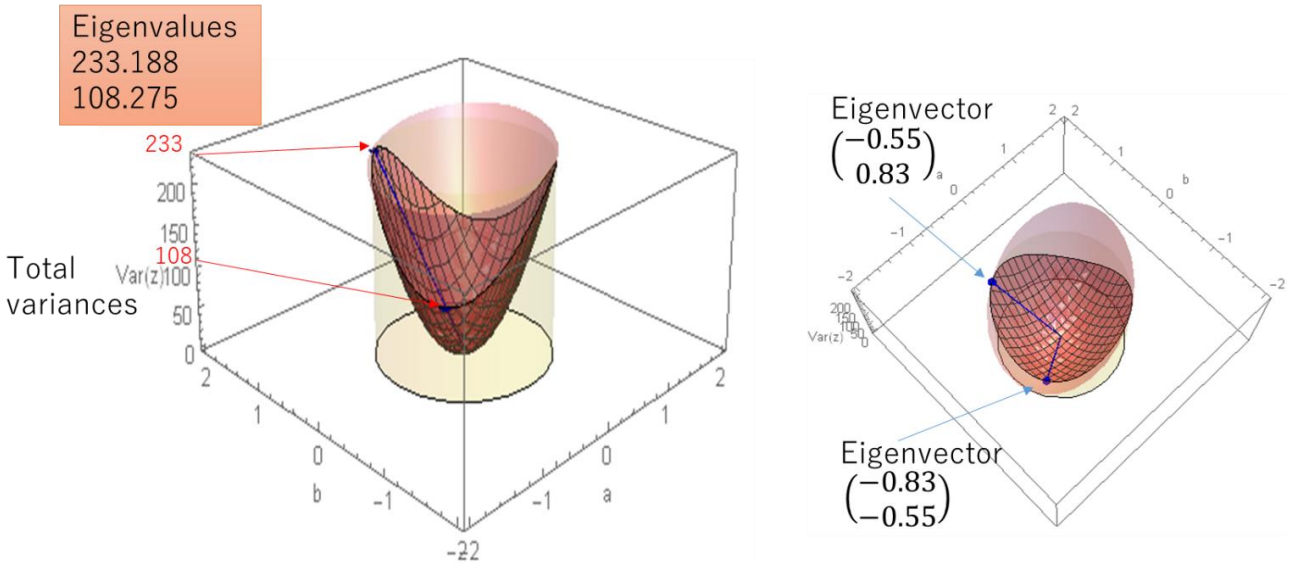
$$Var(x)a + Cov(x,y)\,b = \lambda a$$
$$Var(y)b + Cov(x,y)\,a = \lambda b$$

この連立方程式を行列の形式で表すと以下となる．

$$\begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \lambda \begin{pmatrix} a \\ b \end{pmatrix}$$

何と，共分散行列が出てきた．しかも，a と b は，λ を固有値とする固有ベクトルの要素であったとは，驚きである．

具体的例を可視化でみていこう．下の図では， a 軸と b 軸，そして垂直方向に Var(z)を取った．Var(z)の曲面は縄文式土器のような形状である．そこに，$a^2 + b^2 = 1$の制約を課す．両者のインターセクションは緩やかなカーブを描き，最大点が 2 点，最小点が 2 点ある．



Var(z)の最大値 233.188 が第 1 主成分の固有値であり，Var(z)の最小値 108.275 が第 2 主成分の固有値である．このグラフィクスを天井から眺めると，a-b 平面にプロジェクション・マッピングされた単位円が見える．(右図参照)．原点から最大点へ向かう第 1 固有ベクトルが第 1 主成分の方向を表し，原点から最大点へ向かう第 2 固有ベクトルが第 2 主成分の方向を表している．求めた a, b の値を用いて，主成分軸の式は，bx+ay=0 と表せる．

分散を最大化する制約付き最適化問題を解いていたら，いつの間にか，分散共分散行列の固有値問題になっていた，という何ともいえない妙なる変化(へんげ)を見せてくれる文章題である．

終わり

引用元：　あらたふと青葉若葉の日の光　　芭蕉

＊ラグランジェの未定乗数法を学ぶのであれば，白田由香利：「悩める学生のための経済・経営数学入門―3 つの解法テクニックで数学アレルギーを克服!」，共立出版，2009 年.