データ分析基礎
主成分分析

京都大学 国際高等教育院 附属データ科学イノベーション教育研究センター

せきど ひろと
關戸 啓人

sekido.hiroto.7a@kyoto-u.ac.jp

主成分分析

主成分分析の概要

★ **主成分分析**（**Principal Component Analysis, PCA**）

- ★ **次元の縮約の観点から，新しい座標を構成する**もの

- ★ 例えば，世界500都市の1時間おきの気温20年分のデータが有るとする
- ★ データ数は500×175320程度
- ★ 都市の緯度，経度，人口密度，内陸度，…，などの別の座標を導入することで，全ての気温のデータを保存しなくても良い？

★ このような新しい座標の導入をデータのみから自動的に算出

- ★ 筋の良い座標の取り方がわかる
- ★ データ容量，計算量削減，**ノイズ除去**

主成分分析の概要

★ 登場する確率変数は全て平均が0になるように，定数を足したり引いたりしているとする

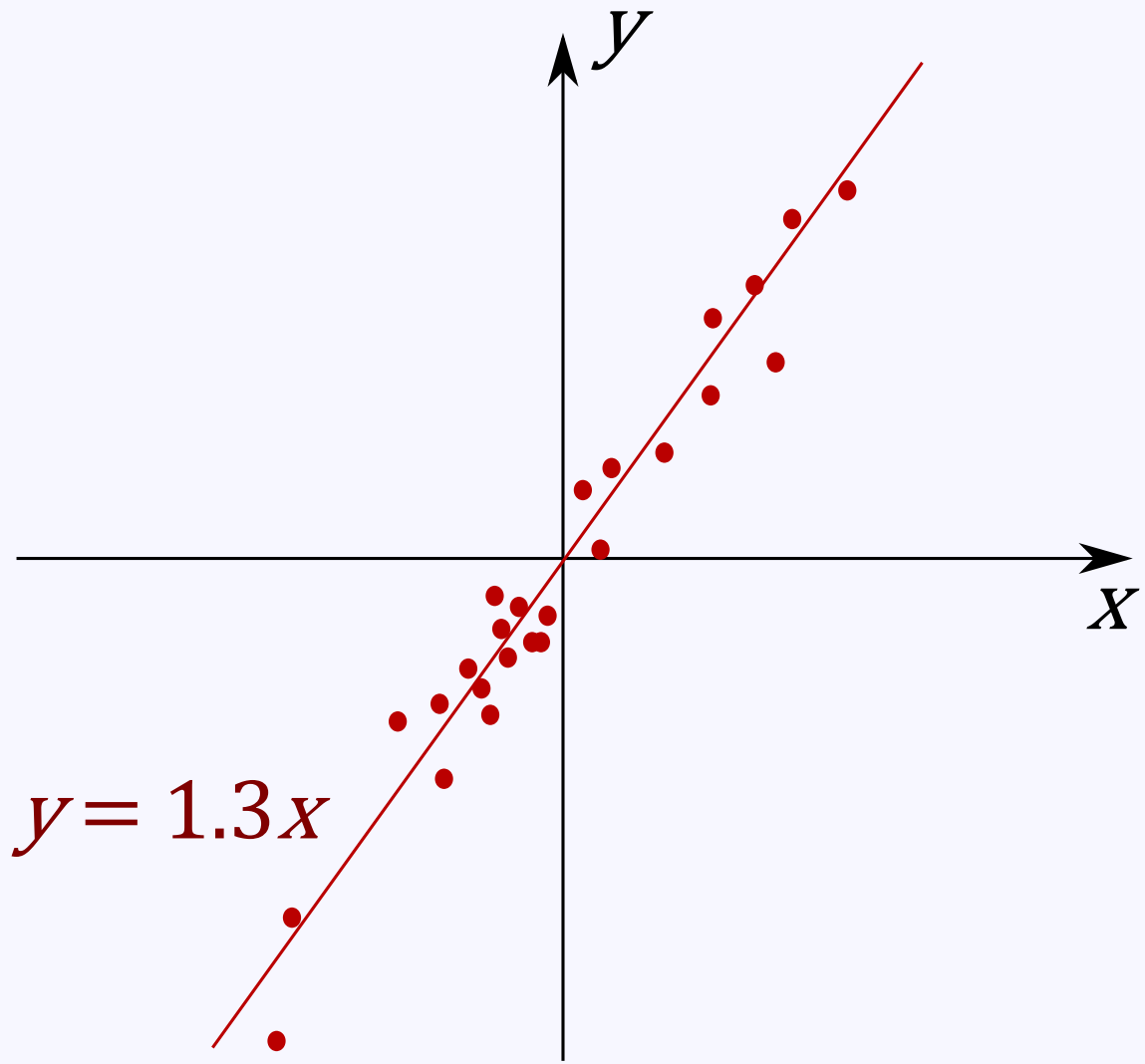
★ データは中心化されている

★ 最初は，2変数の簡単な例で主成分分析の考え方を述べる

★ 確率変数 X は体重，Y は身長を表すとし，データ (x_k, y_k) が与えられたとしよう

★ X と Y には関係があって，近似的に $Y = 1.3X$ ぐらいであるとする

主成分分析の概要



主成分分析の概要

★ データ (x_k, y_k) は直線 $y = 1.3x$ の付近に散らばっている

★ $(\alpha + \varepsilon, 1.3\alpha + \delta)$ で α を適当に取ると，ε や δ は小さいことが多い

★ そこには，（理論的に説明できるかどうかはわからないが）何らかの力が働いていると考えることができる

★ その何らかの力は，以下の確率変数 Z で表されるであろう．

$$Z = \frac{X + 1.3Y}{\sqrt{1^2 + 1.3^2}}$$

★ 確率変数 X, Y を直交変換で Z, U に移すとしたら，U は以下のようになる

$$U = \frac{1.3X - Y}{\sqrt{1.3^2 + (-1)^2}}$$

★ 主成分分析は，このように，確率変数を**直交変換**することである

主成分分析の概要

★ この例では，Z は体の大きさ，U は肥満度を表しているように思える

★ どちらが重要な確率変数かは置いておいて，データの散らばりをより良く説明している確率変数は Z である

- ★ $z_k = x_k + 1.3y_k$ の値を見れば，大体の x_k, y_k の値がわかるという意味である

★ このように，直交変換した後の確率変数で，元のデータを1番良く説明している変数を第1主成分，2番目により良く説明している変数を第2主成分，などと呼ぶ．

★ **主成分分析は，次元の縮小**に用いられる

- ★ 直交変換した後の全ての確率変数を用いれば，元のデータは完全に復元できる
- ★ しかし，それなりに小さい k について，第1主成分から第 k 主成分までのみを用いても，ほぼデータは復元できるようになる
- ★ よって，いくつかの主成分のみを考えても支障がなくなる（支障が出ないように次元を減らす）

主成分の定義A

★ 主成分の定義は（ここで紹介するのは）2種類あるが，1つ目の定義を述べる

★ 1つ目の定義での考え方は，**ばらつきとは，分散である**

★ そして，**ばらつきをより良く説明する，とは分散が大きいこと**と考える

★ 元々の確率変数を $X_1, X_2, \ldots, X_n$ とする

★ 第1主成分 Z_1 は

$$Z_1 = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n,$$
$$w_1^2 + w_2^2 + \ldots + w_n^2 = 1$$

と書けるものの中で，分散が最も大きいものである

★ また，第k主成分は，上の形で書け，第$k-1$主成分までと直交するものの中で，分散が最大となる確率変数

主成分の定義B

★2つ目の定義では，第1主成分Z_1を元々の変数との相関係数の2乗和を最大化する確率変数と取る

★第k主成分は，同様に，第$k-1$主成分までと直交する中で，元々の変数との相関係数の2乗和を最大化するように取る

★これは**前処理としてデータの標準化を行った後に定義Aで主成分を定義している**とも思える

★データの標準化：平均が0，分散が1になるように定数を足したりかけたりする

★前処理の問題なので，スライドでは以下定義Aで説明する

★どちらの定義も行列の固有値問題に帰着されるが，定義Aは分散共分散行列，定義Bは相関行列の固有値問題になる

分散共分散行列と相関行列

★ 分散共分散行列の (i,j) 成分は，X_i と X_j の共分散

$$\mathrm{Cov}(X_i, X_j) = \mathrm{E}[(X_i - \mathrm{E}[X_i])(X_j - \mathrm{E}[X_j])]$$

である．

★ 相関行列の (i,j) 成分は，X_i と X_j の相関係数

$$\frac{X_i \text{と} X_j \text{の共分散}}{(X_i \text{の標準偏差})(X_j \text{の標準偏差})}$$

である．相関係数の絶対値は1以下となる．

★ 確率変数 X_1 は分散が大きいが，確率変数 X_2 は分散が小さい，となれば全体の結果は確率変数 X_1 の影響が強くなる．これを防ぐため，各変数を標準化して考えたものが，相関行列を用いたものだと考えることができる．

- ★ 実際にはほとんどの場合において標準化を行う定義Bで主成分分析を行う
- ★ ただし，データの「標準化」の仕方を考えた方が良い場合も多い

主成分

★ 以降，定義 A で述べる

★ 共分散行列の固有値を大きい順に以下とする

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$$

★ また，λ_k に対応する固有ベクトルを以下とする

$$(w_{1,k}, w_{2,k}, \ldots, w_{n,k})^{\mathrm{T}}$$

★ 第 k 主成分は

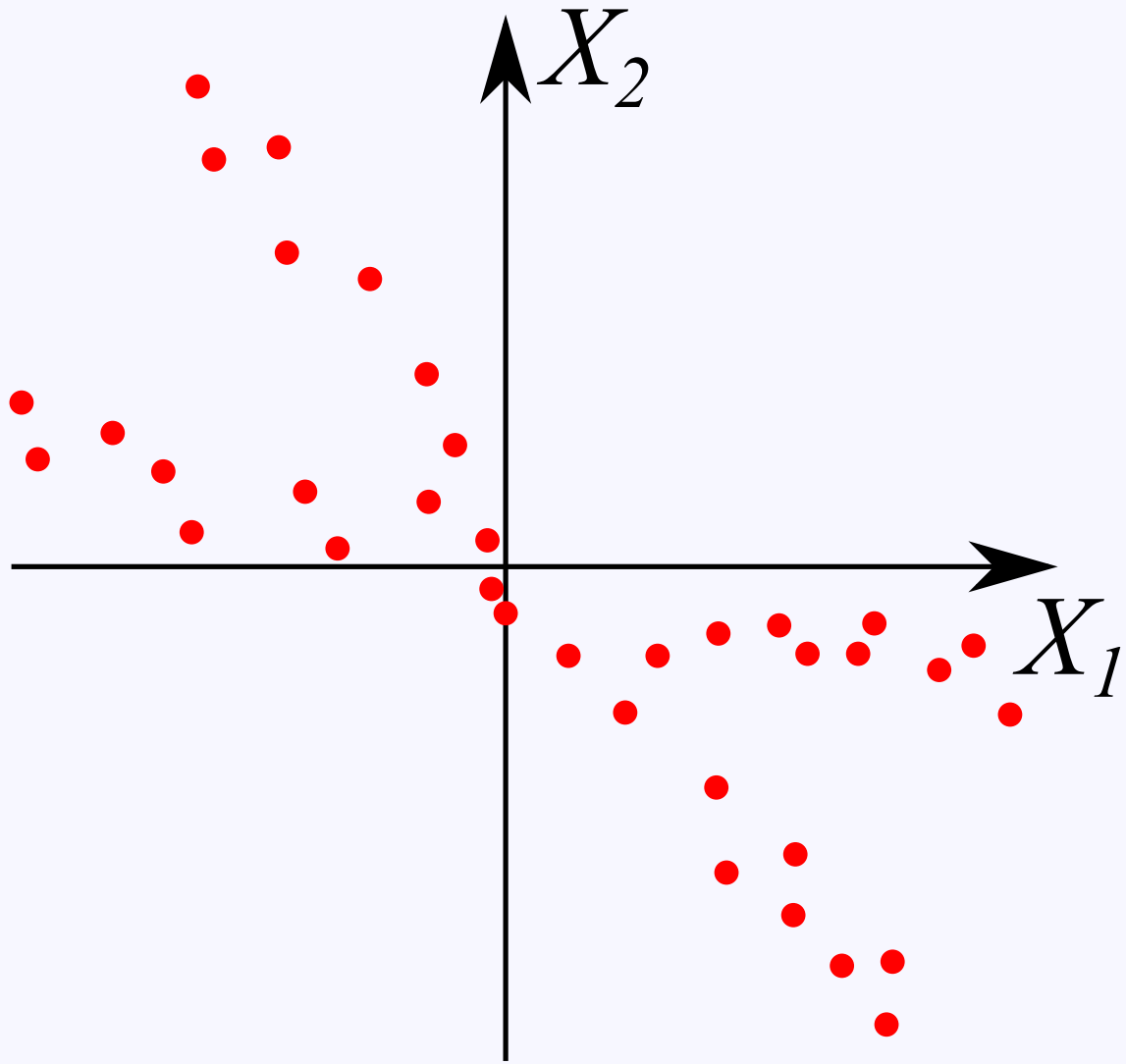
$$Z_k = w_{1,k}X_1 + w_{2,k}X_2 + \cdots + w_{n,k}X_n$$

となり，その分散は $\mathrm{V}[Z_k] = \lambda_k$ となる

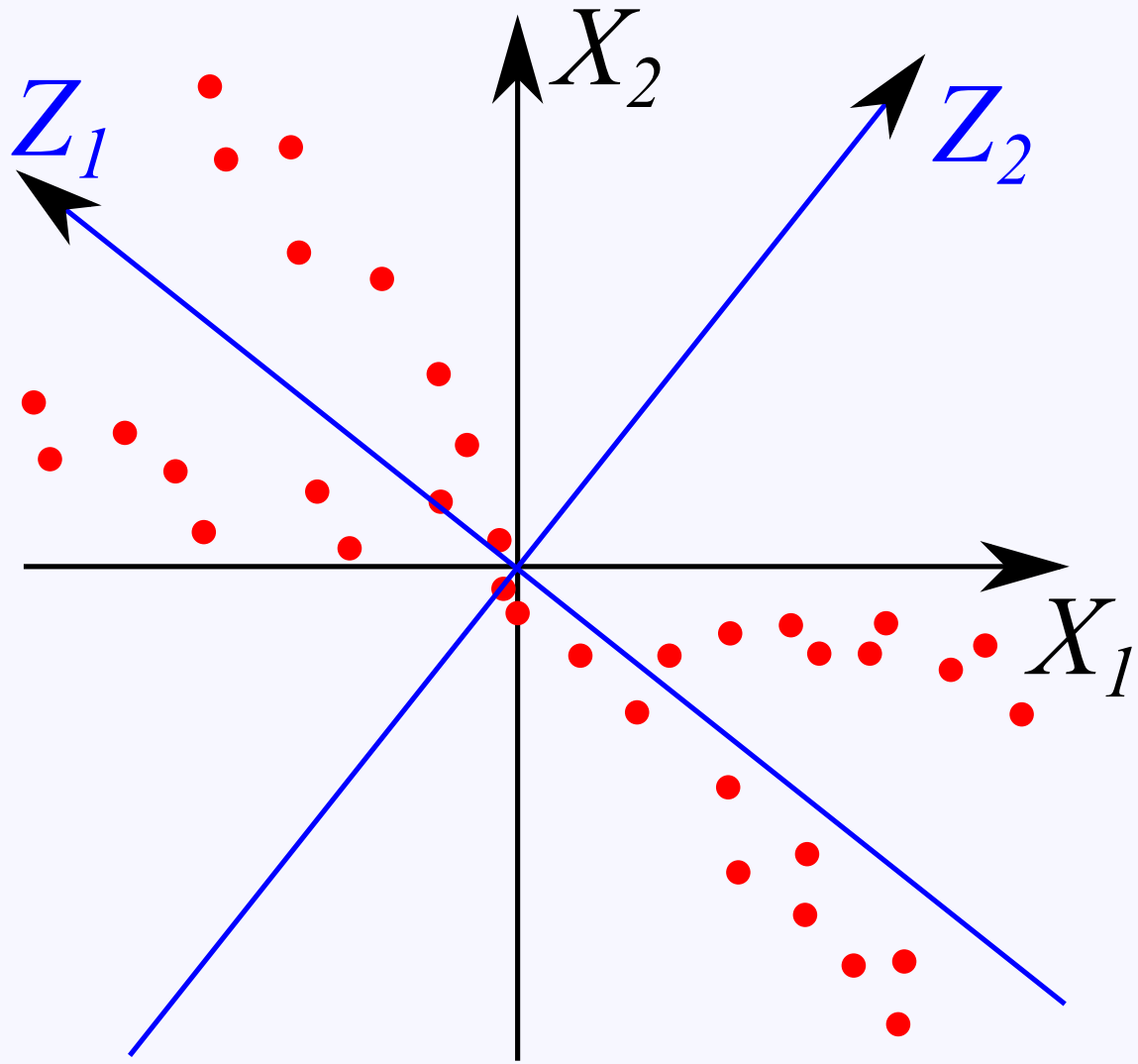
★ 証明は例えばラグランジュの未定乗数法を用いる（ここでは略）

★ データ行列が $A = UDV^{\mathrm{T}}$ と特異値分解されているとき，AV の第 k 列目が第 k 主成分

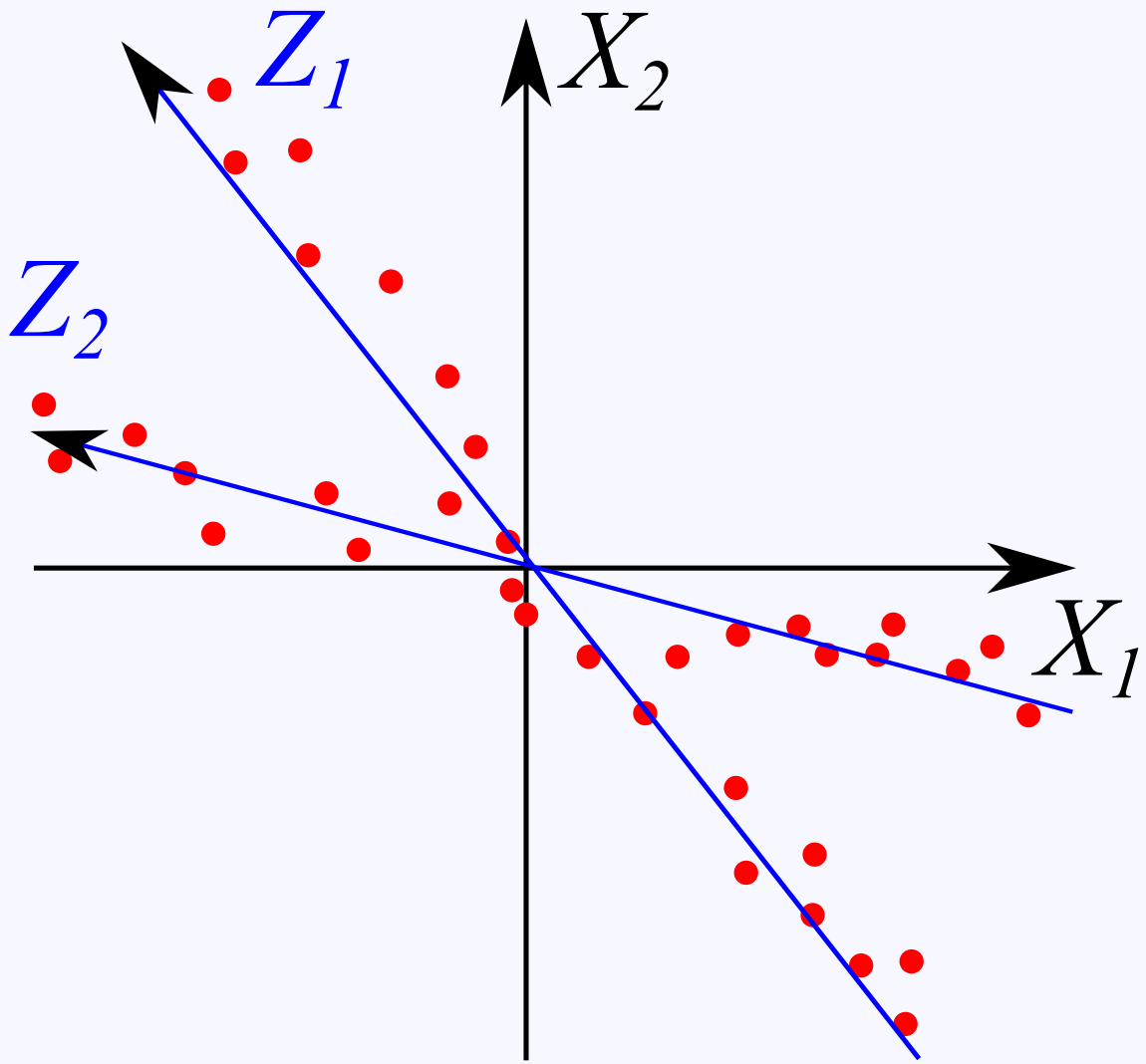
例：データ



例：主成分分析の結果



例：他の手法（因子分析，独立成分分析など？）を用いると



例に対する補足

★ 直交変換という制約が故に，隠れた要因を発見できないかもしれない

★ データについて，個々の構成要素を得ようとする方法として，別の方法で，因子分析がある

- ★ 因子分析のやり方は，いろいろな定義があり，それぞれ結果も一致しない

★ 独立成分分析では，各確率変数ができるだけ独立になるように定める

- ★ これも，いろいろな定義がある

★ 対して，主成分分析は，少ない主成分でデータを説明する，データの総合的なスコアを定める，ということに特化している

データで表す

★ 確率変数の数を n，標本サイズを m とする

★ データ行列を以下とする

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix} \in M_{m,n}(\mathbb{R})$$

★ X_i の標本平均は $\displaystyle\sum_{k=1}^{m} x_{k,i} = 0$

★ X_i と X_j の不変共分散は $\dfrac{1}{m-1}\displaystyle\sum_{k=1}^{m} x_{k,i}x_{k,j}$

★ 不偏共分散行列は $\dfrac{1}{m-1}A^{\mathrm{T}}A$

主成分分析を計算機で行うには

★ 主成分を求めるには

★ **共分散行列の（大きい方から数十個の）固有値と固有ベクトルを求める**

★ または

★ **データ行列の（大きい方から数十個の）特異値と右特異ベクトルを求める**

★ 特異値：$\sqrt{AA^{\mathrm{T}}}$ の固有値

★ 右特異ベクトル：$A^{\mathrm{T}}A$ の固有ベクトル

★ 左特異ベクトル：AA^{T} の固有ベクトル

★ 主成分分析は行列の特異値分解そのもので数学的に性質が良く知られており，ある意味で自然な分析

データと分散

★ 寄与率：分散の割合

★ 元々の確率変数での分散の和と，主成分での分散の和は等しい（直交変換だから）

$$\sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2 = \sum_{i=1}^{n}\sum_{k=1}^{m} z_{k,i}^2 = \mathrm{tr}(A^T A)$$

★ 主成分 Z_i の**寄与率**

$$\sum_{k=1}^{m} z_{k,i}^2 \Big/ \sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2$$

★ 主成分 $Z_1, Z_2, \ldots, Z_s$ の**累積寄与率**

$$\sum_{i=1}^{s}\sum_{k=1}^{m} z_{k,i}^2 \Big/ \sum_{i=1}^{n}\sum_{k=1}^{m} x_{k,i}^2$$

★ 累積寄与率が，ある程度大きくなるように，使用する主成分の数を決めることが多い

（補足）無相関

★ $A = UDV^{\mathrm{T}}$ とする（特異値分解）

★ $U^TU = I$，$V^{\mathrm{T}}V = I$ を満たし，D は対角成分以外0

★ すると，主成分からなるデータ行列は $AV = UD$ で，以下が成り立つ

$$\text{const} \times \mathrm{Cov}(UD) = (UD)^{\mathrm{T}}UD = D^{\mathrm{T}}U^{\mathrm{T}}UD = D^{\mathrm{T}}D$$

★ $D^{\mathrm{T}}D$ は対角行列：つまり主成分同士は無相関

★ ここでは，分散を最大化する方針で主成分を定義したが，「直交変換である，かつ，無相関にする」という方針でも同じ結果が得られる

（補足）残差最小化

★ 第k主成分のみを用いてデータ行列を復元したとき

- ★ 復元されたデータ行列を$\tilde{X} = (\tilde{x}_{i,j})$とすると
- ★ $\sum(x_{i,j} - \tilde{x}_{i,j})^2$が最小化されている

★ 主成分分析はデータ行列をできるだけ良く近似するように次元の縮小を行っている

- ★ これを定義と思っても主成分分析が得られる
- ★ **低ランク近似**

因子負荷量

★ 元々の各変数と主成分との相関係数

★ 定義Bで行った場合（データの正規化を行った場合）は

$$\mathrm{cor}(X_i, Z_j) = \sqrt{\lambda_j} w_{i,j}$$

★ この主成分は，元々のどの変数の影響を強く受けているか？ という指標

★ 主成分の意味を理解するのに使うことがある

主成分得点

★ 各データを主成分で表したときの値を主成分得点という

★ i個目のデータの第j主成分得点は$(AV)_{i,j}$

主成分分析の概略：まとめ

★ **主成分分析**とは，数学的には**確率変数の直交変換（データの回転）**

- ★ $X_1, X_2, \ldots, X_n$ から $Z_1, Z_2, \ldots, Z_n$ に変換
- ★ $Z_1, Z_2, \ldots, Z_n$ は**無相関**，分散は $V[Z_1] \geq V[Z_2] \geq \cdots \geq V[Z_n]$

★ 主成分分析は，**次元の縮約**に用いられる

- ★ データのばらつきを説明するには，大きい方からいくつかの主成分だけで十分かもしれない．累積寄与率を参考にする

★ 主成分は，データの裏に隠れた要素，要因を表しているかも

- ★ 希望するものが得られているかはわからない，説明がつくとも限らない
- ★ そういうものの解析をしたいのなら，因子分析，独立成分分析なども視野にいれる

★ 計算方法は，共分散行列の固有値分解，または，**データ行列の特異値分解**

- ★ 実際には，固有値，あるいは，特異値の大きい方から数個だけ必要