多変量解析

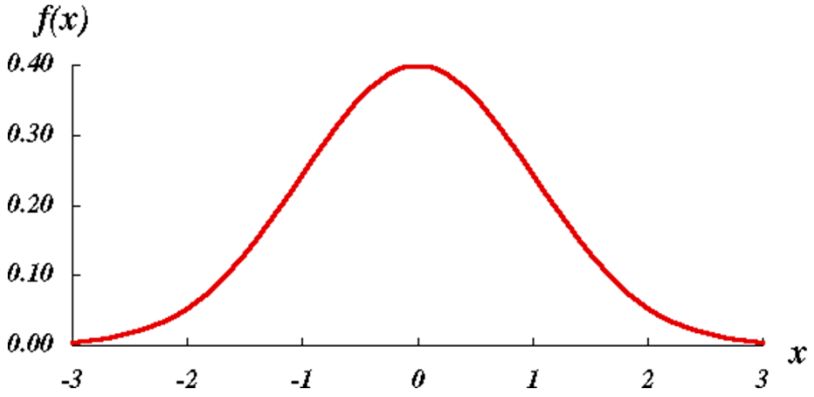
学ぶこと

1. 相関、共分散という考え方
2. 行列の基礎 （おさらい）
3. 回帰分析
4. 主成分分析
5. 主成分回帰

データの計測

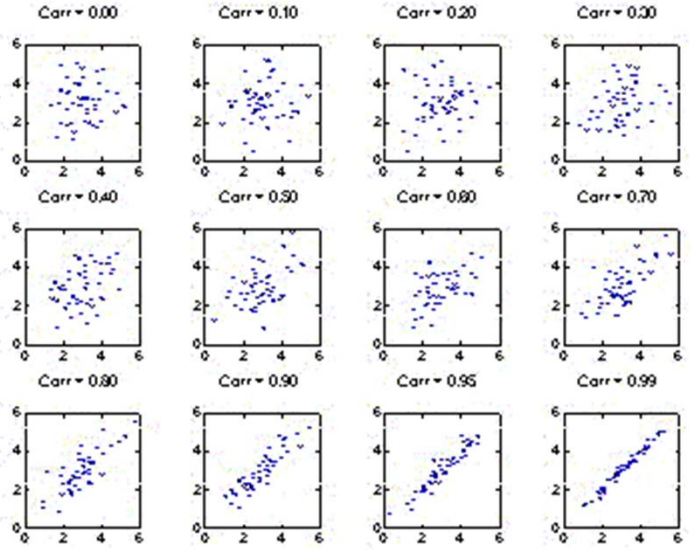
１種類のデータ

複数回計測すると、分布が生じる。



２種類のデータ

複数回計測すると、何らかの関係が見出される。　＝　相関



相関

平均、偏差、分散の関係

- 平均
  $m = \Sigma\, x_i / n$
- 偏差
  個々のデータと平均の差
  $(x_i - m)$
- 分散
  偏差を2乗した平均
  $\sigma^2 = \Sigma\, (x_i - m)^2/n$
- 標準偏差
  分散の平方根 σ

共分散 (covariance) と相関

- 共分散
  分散を2変量に拡張したもの
  $S = \Sigma\, (x_i - m)(y_i - m)/n$
  この値は x, y の単位が異なれば変化してしまう

- 相関
  共分散を標準化
  $R = \Sigma\, (x_i - m)(y_i - m) / (\sigma_x\ \sigma_y)$

相関　＝　（標準化した）共分散

共分散

複数の計測値はベクトル

- １次元のベクトルは数直線上の点で表せる
- ２次元のベクトルは平面で向きと長さを持つ
- 標準化した共分散とはベクトル間の cosΘ にあたる

共分散も相関も、多次元データの「角度」をみている

ベクトルの内積

- 内積 = コサイン

$s_1 = [1, 1]$　$s_2 = [3, -1]$

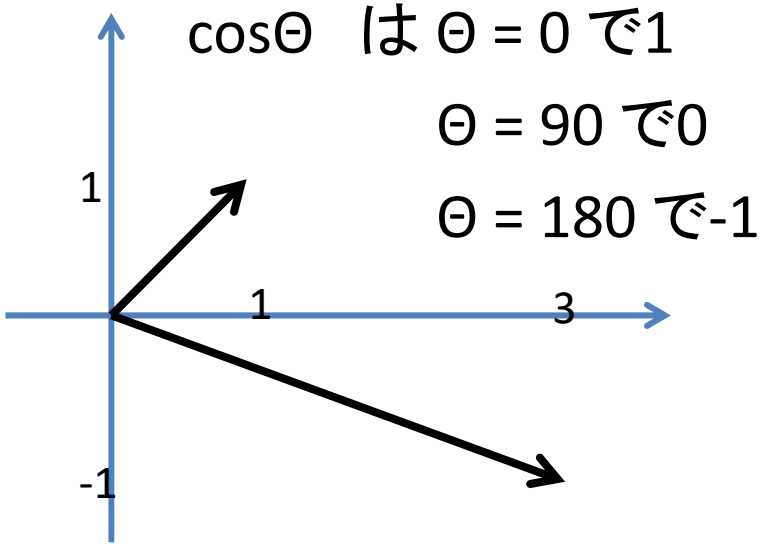
$|s_1|\ |s_2| \cos\Theta = s_1 \cdot s_2$

$= 1x3 + (1x-1)$

cosΘ　は Θ = 0 で1

Θ = 90 で0

Θ = 180 で-1



データ行列の表現

解析データの記述

2	5	3	サンプル1
8	9	2	サンプル2
3	4	1	サンプル3
9	8	2	:
:			

変量1　変量2　変量3　…

行方向に
　サンプル中の変量　や
　ピーク情報（波長）
を記述する。
列方向には
　時系列　や
　特徴
をとる。 つまり列が座標軸。

n x p 行列で表したとき、たいてい n >> p となる。
（例外がゲノム解析）

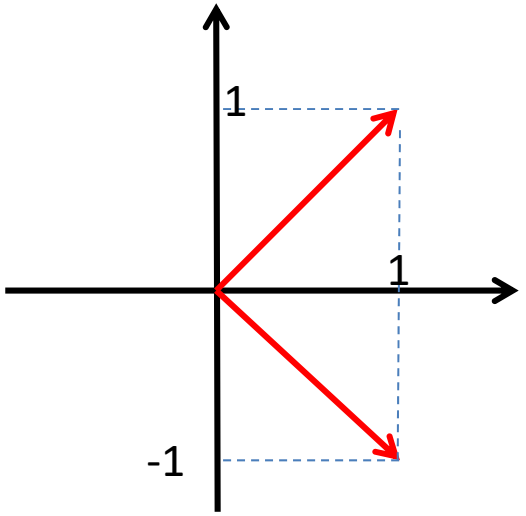
行列　＝　射影

行列は、データであると同時に関数でもある。

例　$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$



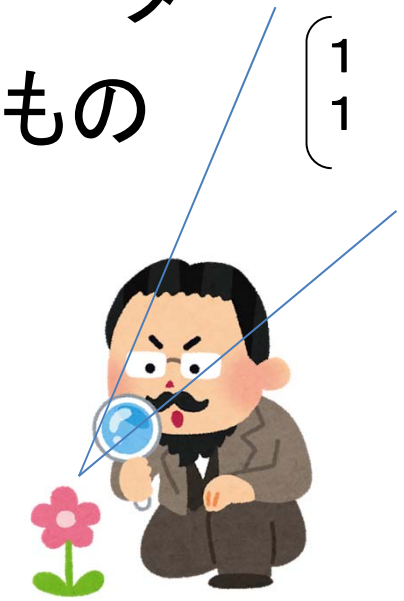
つまり、現在の座標系を上の赤い軸方向に射影する内容。

データ行列の意味

植物など対象を観察した際のデータ

＝　対象の本質を「射影」したもの

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$



- 似た行　＝　類似のサンプル
- 似た列　＝　類似の形質

似ている要素は、いずれも行列の複雑度を下げる

Rによる実習

アンダーソン、フィッシャー (1936) が用いたアヤメ
50ずつ*Iris setosa*，*Iris versicolor*，*Iris virginica*　の3種

- help(“iris”)　または ?iris

Fig. 5. Outline drawings of petal and sepal from plants of *Iris setosa* (left), *I. versicolor* (center), and *I. virginica* var. *Shrevei* (right).

Anderson E. 1936. The Species Problem in Iris.
Annals of the Missouri Botanical Garden 23:457-509.
http://biostor.org/reference/11559

http://biostor.org/reference/11559

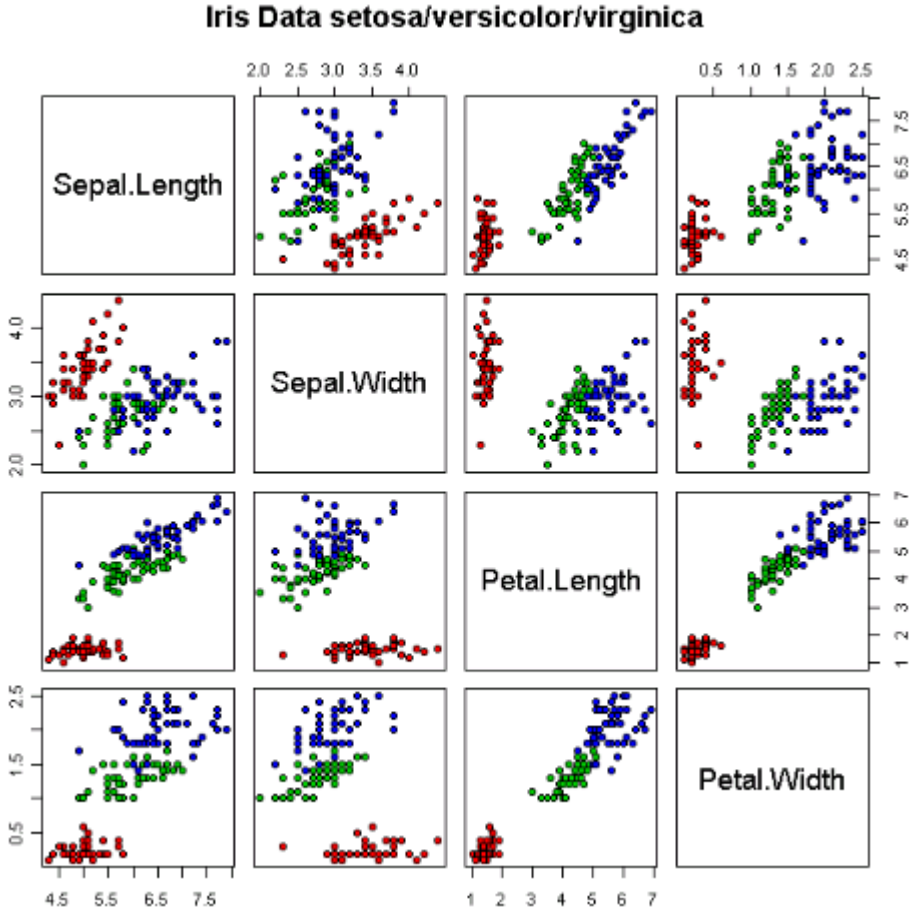
https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x

Irisのデータ

福岡大学工学部図学教室　梶山喜一郎さんのサイトがよくまとまっています。

（「R　IRISデータ」 で検索）

- plot(“iris”)



参考：スピアマンの順位相関

順位が一致しているときと、逆順のときを両端として、-1 から 1 の間に正規化したもの

1 2 3 4 5 6	1 2 3 4 5 6
1 2 3 4 5 6	6 5 4 3 2 1
1 4 9 16 25 36	6 10 12 12 10 6
最大値は Σk^2	最小値は $\Sigma\ k(n-k+1)$

最大と最小の中間値は$\Sigma k(n+1)/2$、幅は $\Sigma\ (k^2 - nk - 1)$

参考：ケンドールの順位相関

n 要素からランダムに2個選んだペアの順序関係が両者で一致している確率のこと

$$\tau = 2P / {}_nC_2 - 1$$

2をかけて1 引くのは [-1, 1] の範囲にするため。

逆行列　＝　逆変換

n次元正方行列 **A** に別の正方行列 **A’** をかけて単位行列 **I** になるとき、**A’** を **A** の逆行列と呼ぶ。

$$\mathbf{A}\,\mathbf{A'} = \mathbf{I}$$

これを $\mathbf{A}^{-1}$ と書く。

逆行列が存在するには、**A** がn次元の基底を成分に持つ(regular つまり正則である)　ことが必要。

これは、変換にロスがないことに相当。

固有値

$Zx = \lambda x \ (x \neq 0)$

スカラー λ を行列 Z の固有値という。

$[Z - \lambda I]x = 0$

つまり$[Z - \lambda I]$に逆行列は存在しない。行列を特異にする（逆行列がない）λ の値が固有値。

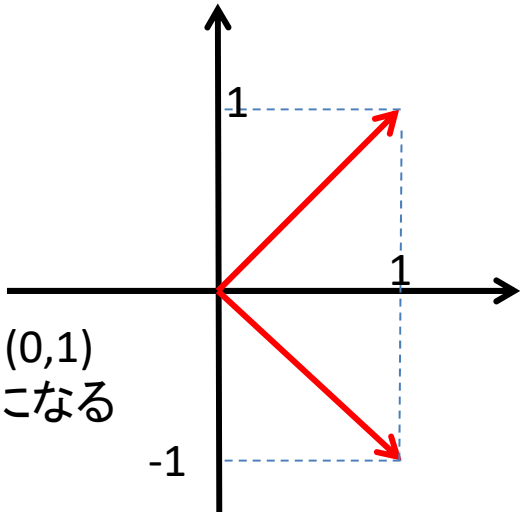
固有値の数はrankに同じ（重複許す）

例

$Z = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ のとき

$|Z - \lambda I| = ((1-\lambda) * (-1-\lambda)) - (1 * 1) = \lambda^2 - 2 = 0$

$\lambda = \pm\sqrt{2}$



Z で単位ベクトル (1,0), (0,1) を射影すると長さが√2になる

固有ベクトル

座標変換において方向が変わらないベクトルが固有ベクトル x。

$$Z x = \lambda x \quad (x \neq 0)$$

普通はノルム（長さ）を1にする。

固有値とは、変換における固有ベクトルの引き伸ばし率。

例

$Z = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$ のとき

$\lambda = \pm\sqrt{2}$ より

- $+\sqrt{2}$ のとき
  $\alpha\,(1 \quad (\sqrt{2}-1))^t$
- $-\sqrt{2}$ のとき
  $\beta\,(1 \quad -(\sqrt{2}+1))^t$

固有ベクトル

固有ベクトル

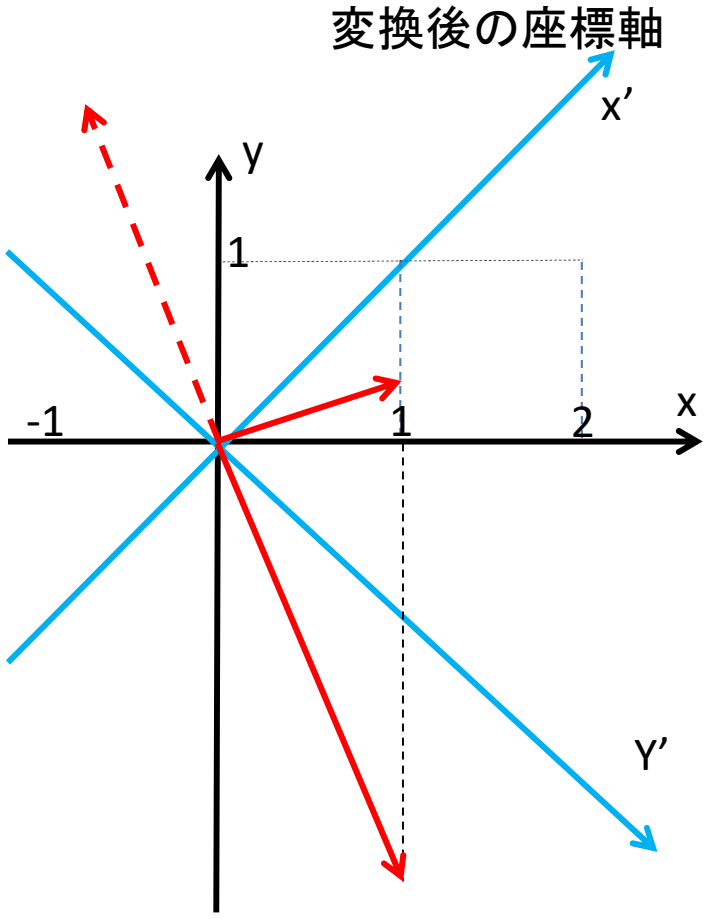
$(1 \quad (\sqrt{2}-1))^t$

$(1 \quad -(\sqrt{2}+1))^t$

は座標変換前と座標変換後の座標が変わらない方向を示す。

固有ベクトルは一次独立。

（基底ベクトルになる。）



注： 2x2の回転行列の場合は向きが変わらないベクトルが存在せず、固有値が虚数になる。3x3の回転行列の場合は、回転軸が固有ベクトル、対応する固有値が 1。

行列の固有値分解

Z が異なるn個の固有値

$$\lambda_1、\ldots、\lambda_i、\ldots、\lambda_n$$

を持ち、対応する固有ベクトルを

$$\mathbf{P} = [\mathbf{x}_1、\ldots、\mathbf{x}_i\ldots、\mathbf{x}_n]$$

とすると

$$\mathbf{P}^{-1}\mathbf{Z}\mathbf{P} = \begin{bmatrix} \lambda_1 & \ldots & 0 & \ldots & 0 \\ \vdots & & \vdots & & \\ 0 & \ldots & \lambda_i & \ldots & 0 \\ \vdots & & \vdots & & \\ 0 & \ldots & 0 & \ldots & \lambda_n \end{bmatrix}$$

行列 Z による一次変換

Z x = y における**x、y**を、**P** の座標系で表現する

$$\mathbf{Z}\,\underline{\mathbf{P}\ \mathbf{x}'} = \underline{\mathbf{P}\,\mathbf{y}'}$$

$$\mathbf{P}^{-1}\mathbf{Z}\,\mathbf{P}\,\mathbf{x}' = \mathbf{y}'$$

xが **Z**の固有ベクトル なら

$$\mathbf{Z}\,\mathbf{P}\ \mathbf{x}' = \lambda\,\mathbf{P}\,\mathbf{x}'$$

$$\mathbf{P}^{-1}\mathbf{Z}\,\mathbf{P}\,\mathbf{x}' = \lambda\,\mathbf{x}'$$

となり、**x'** も固有ベクトル。

つまり、$\mathbf{P}^{-1}\mathbf{Z}\,\mathbf{P}$ は固有値を成分とする対角行列。

固有値分解の例

例

$$Z = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$ のとき

固有値は$\pm\sqrt{2}$

$$Z \begin{pmatrix} 1 \\ \sqrt{2} - 1 \end{pmatrix} = \sqrt{2} \begin{pmatrix} 1 \\ \sqrt{2} - 1 \end{pmatrix}$$

$$Z \begin{pmatrix} 1 \\ -(\sqrt{2}+1) \end{pmatrix} = -\sqrt{2} \begin{pmatrix} 1 \\ -(\sqrt{2}+1) \end{pmatrix}$$

$$P = \begin{pmatrix} 1 & 1 \\ \sqrt{2} - 1 & -(\sqrt{2}+1) \end{pmatrix}$$

基底ベクトルをなす変換

$$P^{-1} = (1/2\sqrt{2}) \begin{pmatrix} 1 & 1 \\ \sqrt{2} - 1 & -(\sqrt{2}+1) \end{pmatrix}$$

基底ベクトルをなす変換

$$P^{-1} Z P = \ldots = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{2} \end{pmatrix}$$

確かに、固有値の対角行列になっている。

つまり縮退していない線形変換は、固有ベクトルの座標系に移動して、基底の長さを固有値倍し、逆変換することに相当する。

固有値の解法

べき乗法

$$\mathbf{Z}^k = \lambda_1^k \mathbf{x}_1 \mathbf{x}_1^t + \ldots + \lambda_i^k \mathbf{x}_i \mathbf{x}_i^t + \ldots + \lambda_n^k \mathbf{x}_n \mathbf{x}_n^t$$

$$= \lambda_1^k \{ \mathbf{x}_1 \mathbf{x}_1^t + \ldots + (\lambda_i/\lambda_1)^k \mathbf{x}_i \mathbf{x}_i^t + \ldots + (\lambda_n/\lambda_1)^k \mathbf{x}_n \mathbf{x}_n^t \}$$

ここで2項目以降を無視すると

$$\mathbf{Z}^k \fallingdotseq \lambda_1^k \{ \mathbf{x}_1 \mathbf{x}_1^t \}$$

転置行列

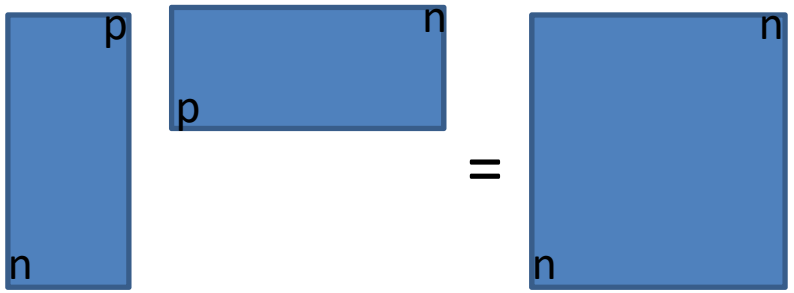
行列の行方向と列方向を入れ替えたものを転置行列という。

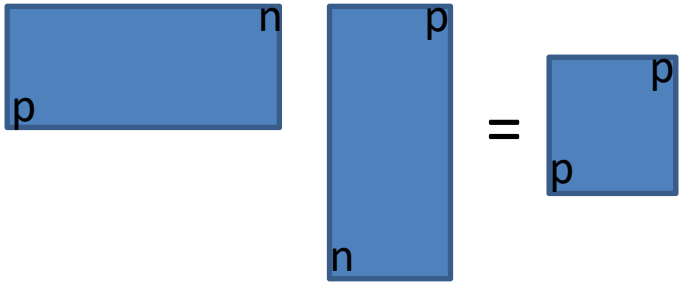
Z を転置したら ${}^tZ$ と書く。

行列を自身の転置行列と掛け合わせると、正方の対称行列になる。

いずれの対称行列も、ランクは等しい。

例





転置行列と逆行列

正規直交基底を列ベクトルとする直交行列と、
その転置行列は逆行列の関係

行列の掛け算を考える。

A = [基底を列方向に持つ行列]

${}^{t}A$ = [基底を行方向に持つ行列]

${}^{t}A\ A$ = [各要素は基底の内積]

= [同じ基底（軸）なら内積が1, 残りは0]

= [単位行列]

対称行列

Z を転置しても等しい、つまり $Z = Z^t$ となるのが対称行列

実対称行列の固有値は全て実数。

対称行列の固有ベクトルは必ず直交する。（一次独立だけではない。）

$Z^t Z = Z^2 = Z Z^t$ となるので、正規行列 (normal) でもある。

なぜ固有ベクトルが直交?

内積を考える。Zは固有ベクトルの方向を変えないので

$Z x_1 = \lambda_1 x_1$ 、 $Z x_2 = \lambda_2 x_2$ が成立。

対称行列なので

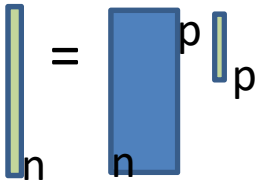
$\lambda_1 (x_1, x_2) = (\lambda_1 x_1)^t x_2 = (Z x_1)^t x_2 = x_1^t (Z^t) x_2 = x_1^t Z x_2 = x_1^t (\lambda_2 x_2) = (x_1, (\lambda_2 x_2)) = \lambda_2 (x_1, x_2)$

固有値が等しくない限り、内積は0。

一般逆行列

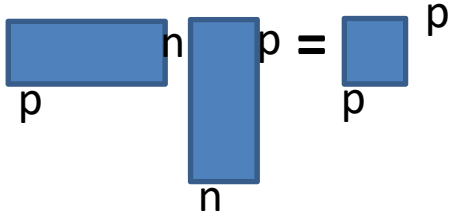
正方で正則の行列なら逆行列が存在する。

$y = \mathbf{X}\, b$ なら $b = \mathbf{X}^{-1} y$



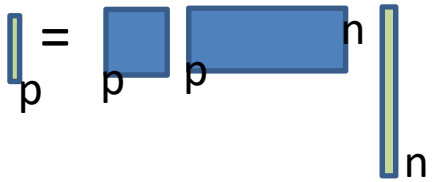
しかし普通のデータ行列は正方でない。そこで**X**の転置をかけて正方にする。

$\mathbf{X}^T\mathbf{X}$



$\mathbf{X}^T y = (\mathbf{X}^T\mathbf{X})\, b$

$b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$



この値を $y = \mathbf{X}\, b$ に代入すると

$\hat{y} = \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$

つまり y の推定値をだすことができる（回帰分析）。

回帰分析 (regression)

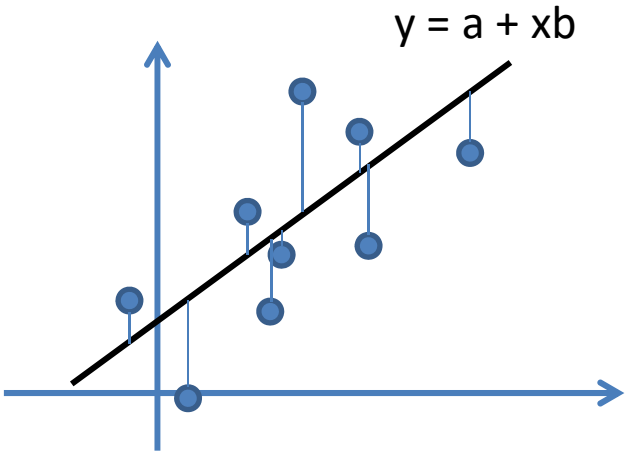
$$y = b_0 + xb_1$$

y: 従属（目的）変数　　x: 独立（説明）変数

説明変数が1つ：　単回帰分析 (univariate linear regression)

説明変数が2つ以上：　重回帰分析 (multivariate LR)

単回帰



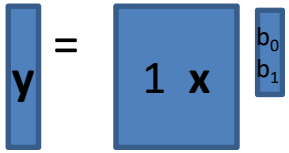
y からのズレ（残差）の二乗和を最小化するように係数 **b** を定めたい (最小二乗法)。

残差の和 $S = \sum (y - b_0 - x\, b_1)^2$

この値を b_0 , b_1 で偏微分して 0 とおけば良い。

$\partial S/\partial b_0 = \sum -2(y - b_0 - x\, b_1) = 0$

$\partial S/\partial b_1 = \sum -2(y - b_0 - x\, b_1)x = 0$

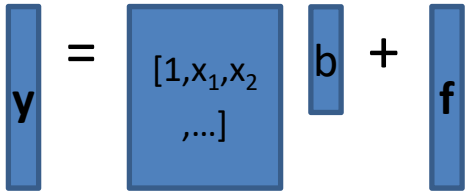


$\mathbf{y} = \mathbf{X}\mathbf{b}$ なので

$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$　と計算

重回帰分析 (MLR)

基本は単回帰に全く同じ



$y = \mathbf{X} b + f$ (残差)

f を最小化するため、

$b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$ で係数を推定。

でも、大体はうまくいかない。

共線性(colinearlity) の問題

X に相関が高い列成分があると、b の計算ができない。

行列式 $|\mathbf{X}^T\mathbf{X}|$ が0に近いことが原因 （ランク落ち）。

対策:

固有値が小さい、行列式が0に近い、特大な値がある等。

線形モデルの例

植物の成長を、種のスコア、土のスコア、気温のスコアで近似

Plant = F1 * seed + F2 * soil + F3 * temp + error

Find best parameters F1~F3. (= minimize error)

$$\begin{bmatrix} plant1 \\ plant2 \\ plant3 \\ \vdots \\ \vdots \end{bmatrix} = \begin{bmatrix} seed1 & soil1 & temp1 \\ seed2 & soil2 & temp2 \\ seed3 & soil3 & temp3 \\ & \vdots & \\ & \vdots & \end{bmatrix} \begin{bmatrix} factor1 \\ factor2 \\ factor3 \\ \vdots \\ \vdots \end{bmatrix} + \begin{bmatrix} error1 \\ error2 \\ error3 \\ \vdots \\ \vdots \end{bmatrix}$$

$$Y = \mathbf{X} \quad b + f$$

Plant = 10 * seed + 5 * soil + 0.2 * temp ± 3

Rによる実習（共線性）

ふたたび iris データ

- s <- lm (Sepal.Length ~ Sepal.width + Petal.Length + Petal.Width, data=iris)
- summary(s)

目的変量が Sepal.Length 。 説明変量として Sepal.Width, Petal.Length, Petal.Width を使う。

2部　主成分分析

スペクトル分解

対称行列 Z が異なるk個の固有値

$$\lambda_1、\ldots、\lambda_i、\ldots、\lambda_k$$

を持ち、対応する固有（**正規直交**）ベクトルを

$$\mathbf{P} = [\mathbf{x}_1、\ldots、\mathbf{x}_i\ldots、\mathbf{x}_k]$$

とすると

$$\mathbf{P}^t \mathbf{Z} \mathbf{P} = \begin{bmatrix} \lambda_1 & \ldots & 0 & \ldots & 0 \\ : & & : & & \\ 0 & \ldots & \lambda_i & \ldots & 0 \\ : & & : & & \\ 0 & \ldots & 0 & \ldots & \lambda_k \end{bmatrix}$$

対称行列で、固有値が全て正のとき(0 も含まない) 正定値行列と呼ぶ。

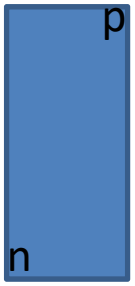
対称行列 **Z** が正定値になる必要十分条件は

$$\mathbf{Z} = \mathbf{X}^t \mathbf{X}$$

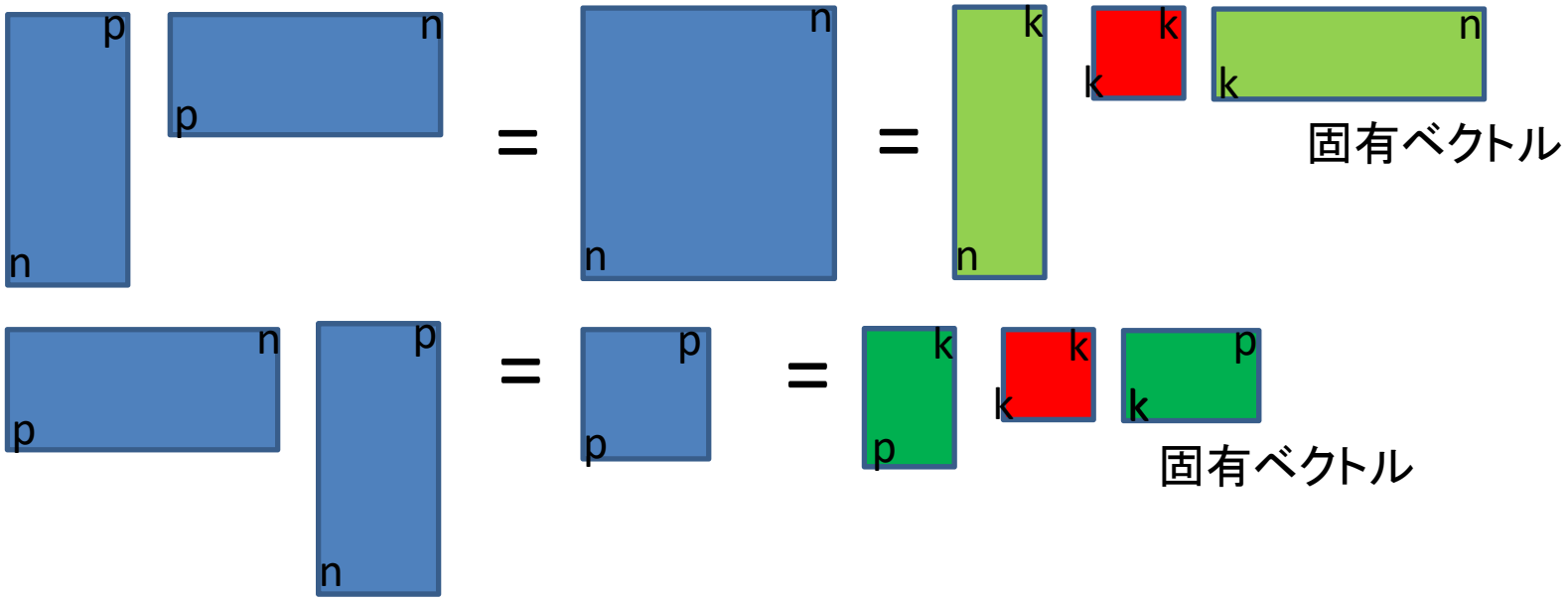
となる正則な正方行列 **X** が存在すること。

スペクトル分解の図示

データはｎｘｐ行列で表す



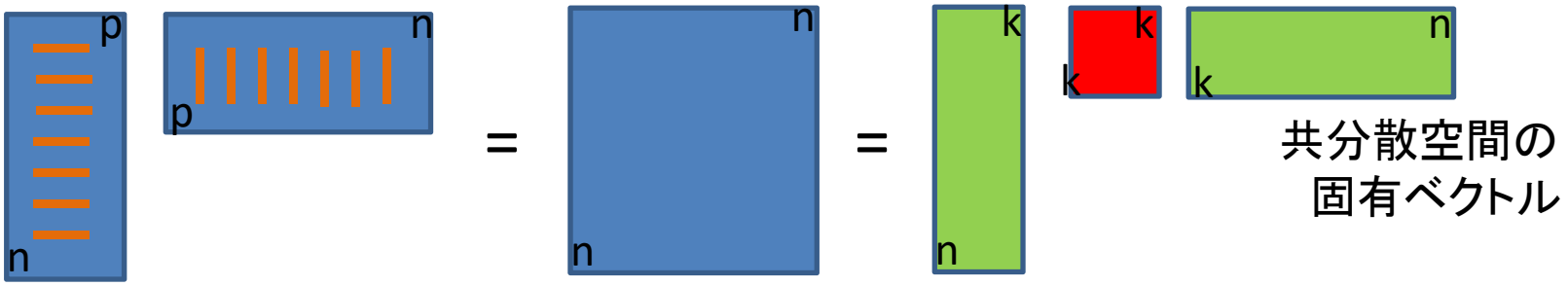
データを転置行列とかけて対称行列にすれば、
スペクトル分解できる。



共分散行列

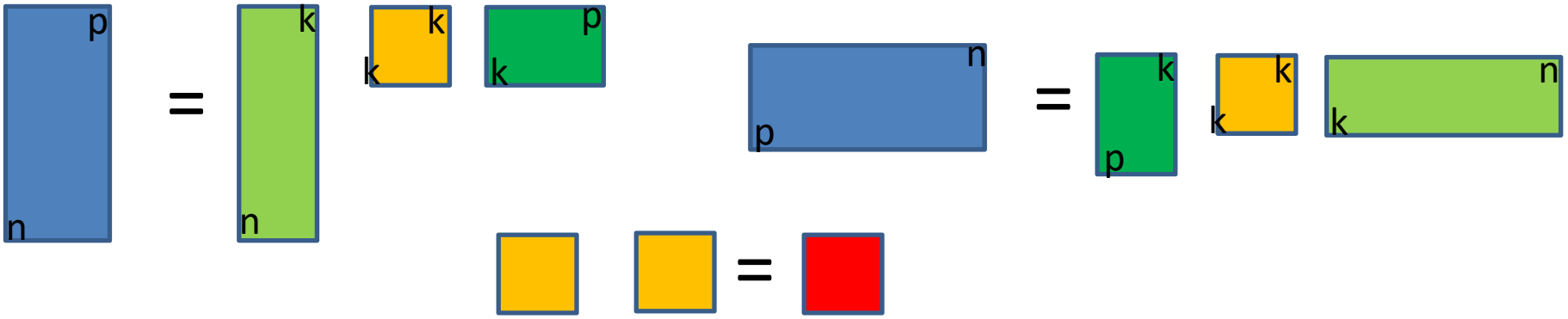
転置してかけ合わせた行列は、共分散行列（自分自身との相関）になっている。

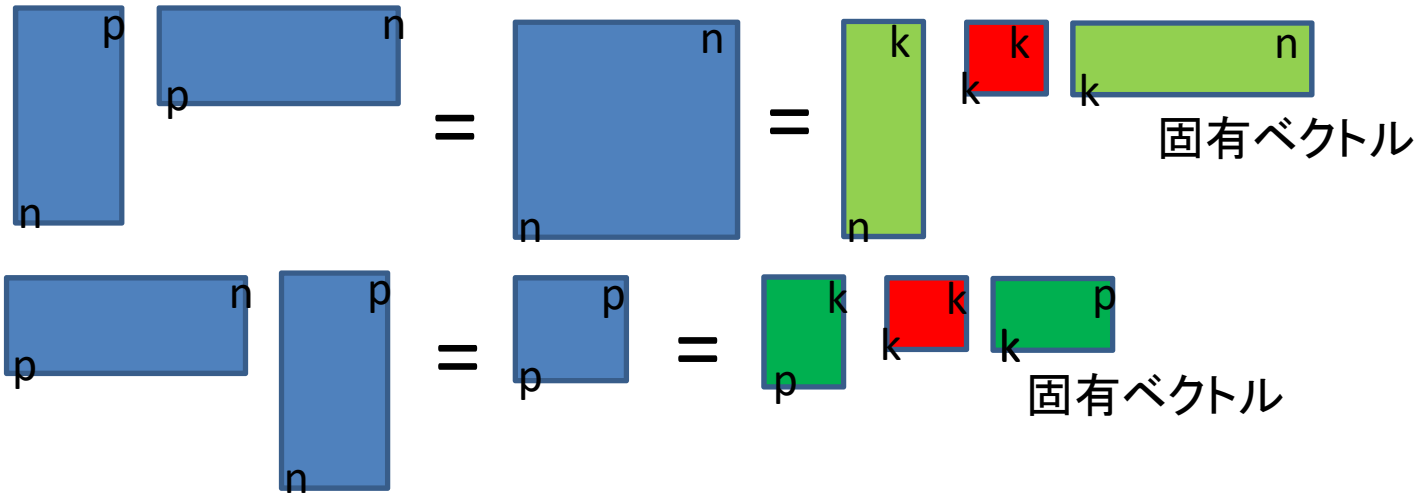
スペクトル分解



行列の特異値分解 (SVD)

スペクトル分解をよく眺むと





行列の特異値分解(2)

SVD … singular value decomposition (特異値分解)

任意の行列 A は

$$\mathbf{A} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$$

に分解できる。ここで **U**, **V** は正規直交行列

$$\mathbf{U}^T\mathbf{U} = \mathbf{I},\ \mathbf{V}^T\mathbf{V} = \mathbf{I}$$

D は **A** のランク分だけ要素を持つ対角行列。
これを特異値という。

特異値と固有値の関係

$$\mathbf{A}\mathbf{A}^T = (\mathbf{U}\,\mathbf{D}\,\mathbf{V}^T)\,(\mathbf{U}\,\mathbf{D}\,\mathbf{V}^T)^T$$

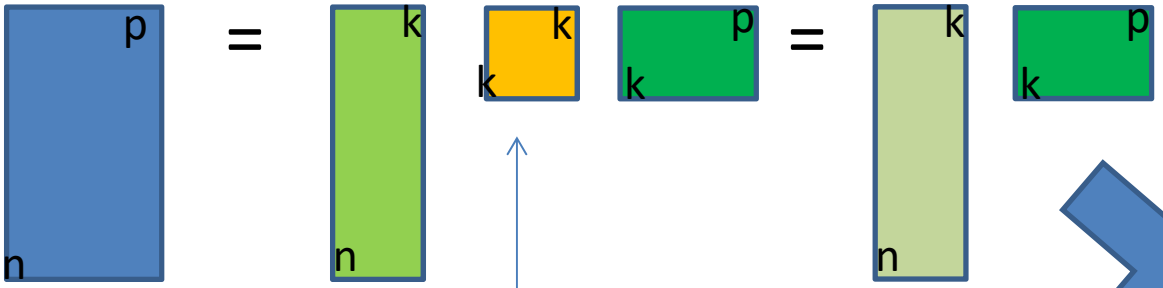
$$=(\mathbf{U}\,\mathbf{D}\,\mathbf{V}^T)(\mathbf{V}\,\mathbf{D}^T\,\mathbf{U}^T)=\mathbf{U}\mathbf{D}^2\mathbf{U}^T$$

$$\therefore \mathbf{A}\mathbf{A}^T\mathbf{x}_i = \mathbf{d}_i^2\,\mathbf{x}_i$$

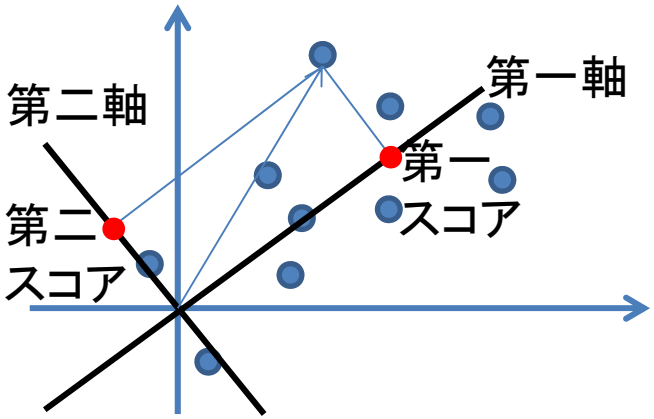
特異値の2乗が$\mathbf{A}\mathbf{A}^T$の固有値に対応している。

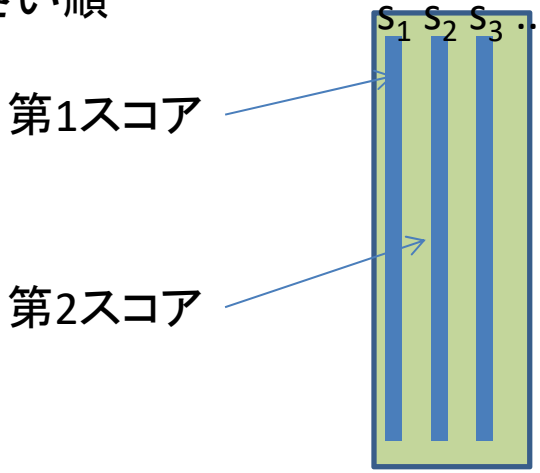
主成分分析 (PCA)

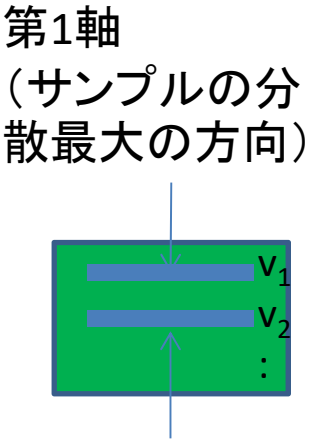
SVD 分解　$\mathbf{A} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T$ を　$\mathbf{U}\,\mathbf{D}$ 部分（スコア）と、
$\mathbf{V}^T$ 部分（ローディング＝共分散行列の軸）に分ける



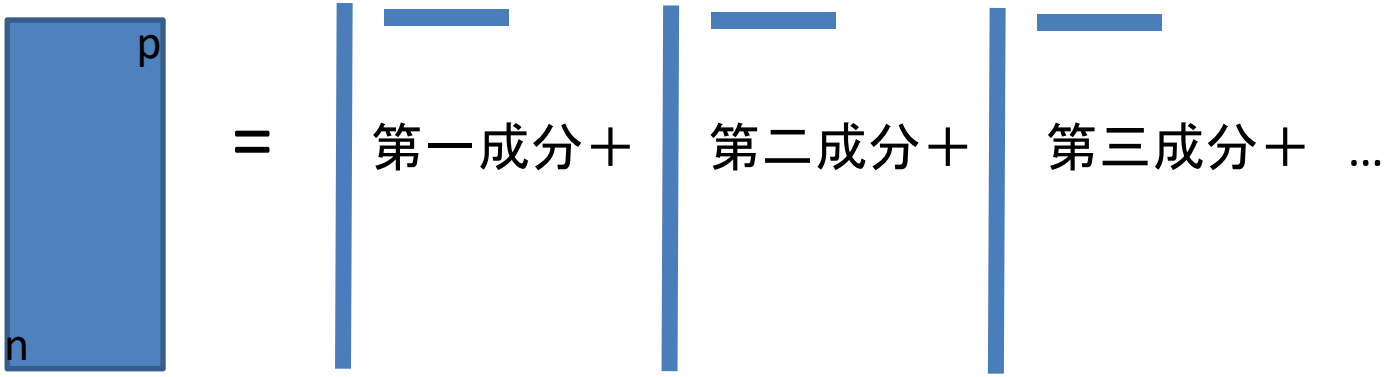
ここで、固有値は大きい順に並べる





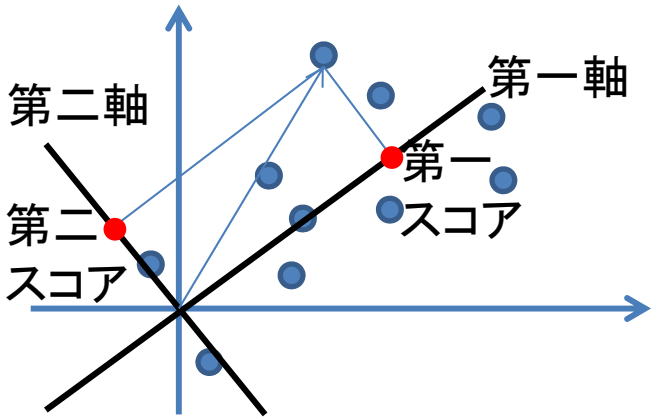


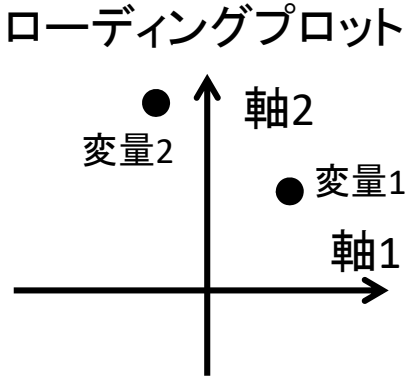
主成分分析 (2)

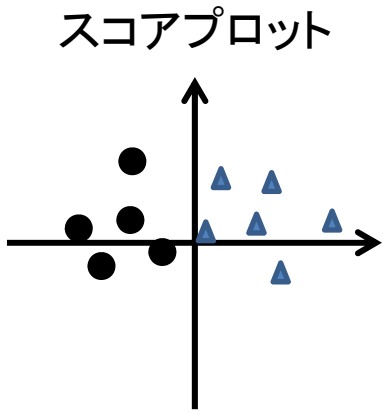


行列を2～3成分でモデルする

各軸において変量の寄与を考慮

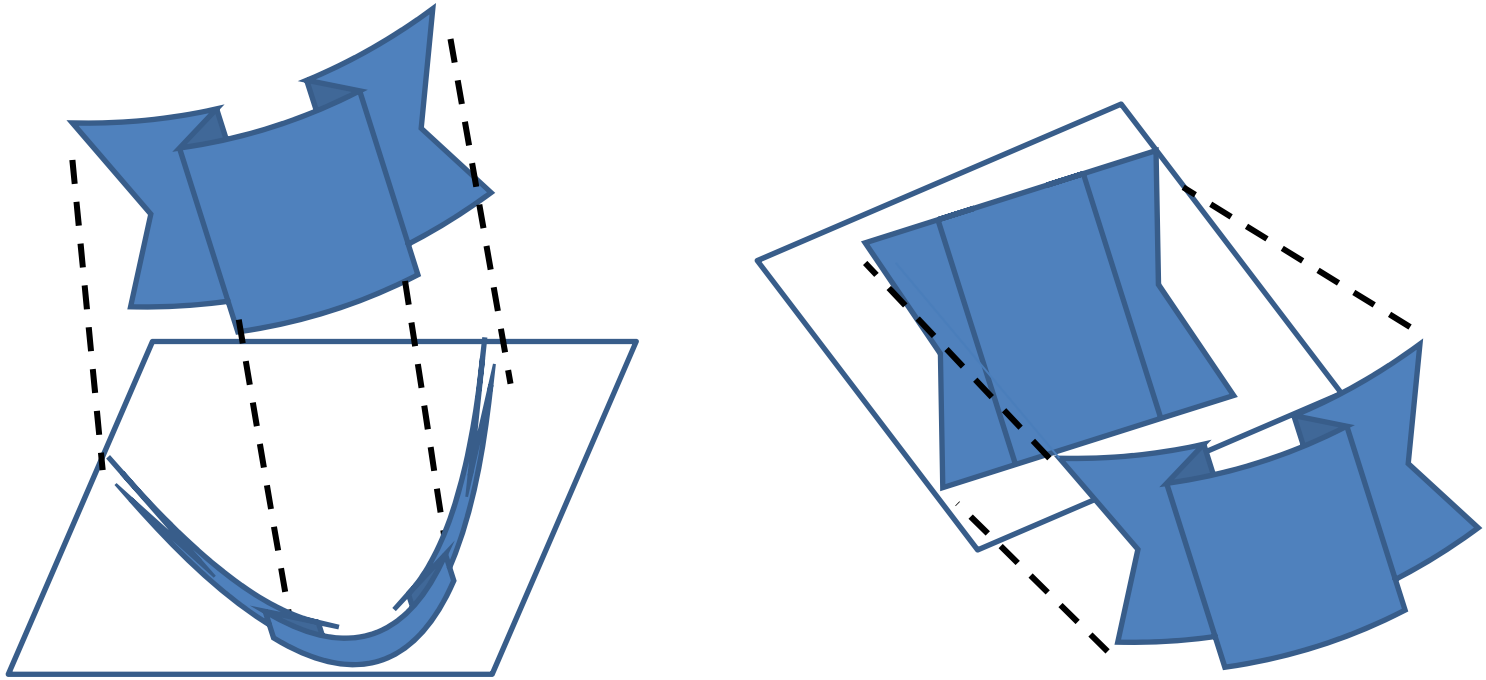






主成分分析の幾何的解釈

PCAは、多次元空間を2次元の平面に射影



一番適した射影とは何か？

PCA は 「面積」 を最大化しているだけ

PCAの利点

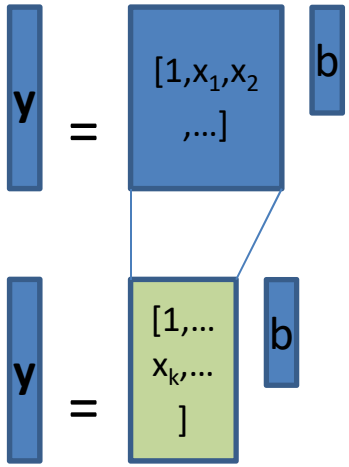
- SVD分解なので逆行列計算より楽
  （このへんは数値計算の教科書を参照。
  Numerical Recipe in C など。）
- 共線性の問題が生じない

つまり、どんな行列でも安心して処理できる。

主成分回帰 (PCR)

主成分のみで回帰分析

決まった数の主成分を選んで回帰分析

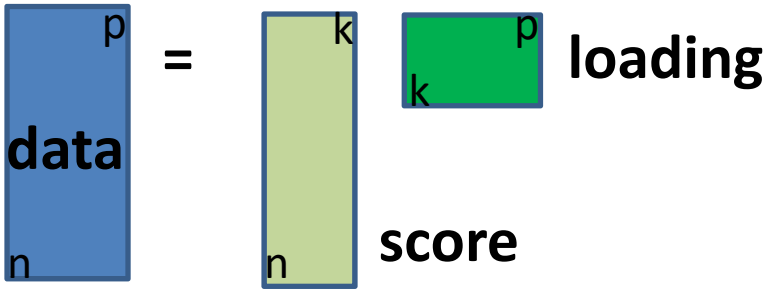


「主成分」として、軸を変換している点に注意

なぜ主成分にするのか

相関が高いものだけを除くと、本来関心のある説明変数を省いてしまうリスク

→ 回帰係数は、変数そのものの効果ではなくなる。

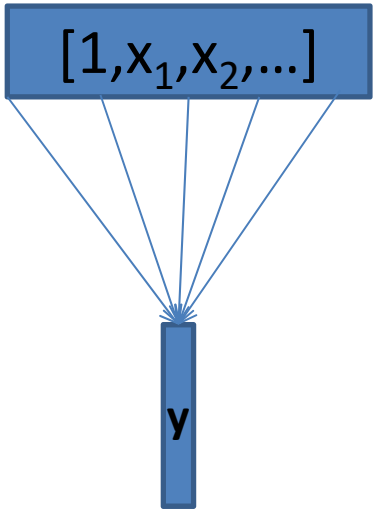


線形回帰分析のまとめ

単、重回帰分析

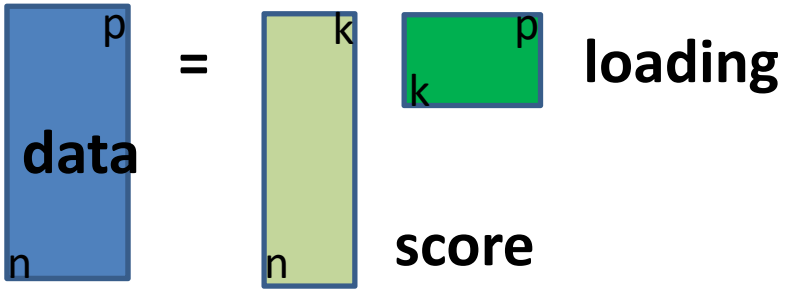
データの軸をそのまま使う

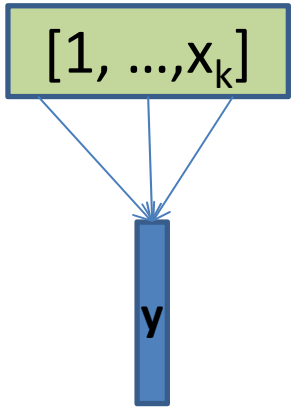
- ノイズが入りやすい
- 変数の選択が必要



主成分回帰

データの分散を最大化する軸を使う





Rによる実習

-
- s <- prcomp(iris[1:4], scale=FALSE)
- label <- iris[,5]
- biplot(s, xlabs = label)
- s
- summary(s)

Biplot はローディングを表示
Prcomp に関するウェブページは多数